# GUÍA DE CLASES:
# INTRODUCCIÓN A R PARA ANÁLISIS DE DATOS

Corporación Universitaria Comfacauca
Unicomfacauca

ACREDITACIÓN INSTITUCIONAL EN ALTA CALIDAD
Vigilada MinEducación

Sello Editorial
Unicomfacauca

Serie:
**INGENIERÍAS**

Guía de clase

# INTRODUCCIÓN A R PARA ANÁLISIS DE DATOS

**Jhon Alexander Segura Dorado**
**Yesid Ediver Anacona**
Programa de Ingeniería Industrial
Grupo de Investigación
Cadenas de valor

Sello Editorial
Unicomfacauca

Calle 4 # 8-30
Popayán, Colombia
e-mail: selloeditorial@unicomfacauca.edu.co
Teléfono: 57+(2) 8386000 Ext 148.
www.unicomfacauca.edu.co

**Colección:** Material docente
**Serie:** Humanidades

**Rectora (E):** Claudia Milena García Castillo
**Coordinación editorial y edición académica:** Paola Martínez Acosta
**Diseño y diagramación:** Oscar Eduardo Chávarro Vargas
**Editor General de Publicaciones:** Yesid Ediver Anacona Mopán



Primera edición en español
noviembre de 2024

**ISBN:** 978-628-95883-6-1

Texto financiado por la Corporación Universitaria Comfacauca-Unicomfacauca, en el marco de la convocatoria interna de Publicación de Material Docente, coordinada por la Dirección de Ciencia, Tecnología e Innovación.

# CONTENIDO


9 **Presentación**
11 **Introducción**
13 **Fundamentación teórica**
17 **Fundamentación metodológica**
19 **Fundamentación curricular y didáctica**

21 **Capítulo 1**
21 Introducción al analisis de datos con r
21 Acerca de r y RStudio
22 Cómo instalar r y RStudio
25 Estructura de rstudio
27 Instalar paquetes de RStudio
28 Como leer con RStudio
31 Mi primer programa
32 Una primera aplicación de análisis de datos con RStudio
33 Ventajas de utilizar RStudio
34 Desventajas de RStudio

37 **Capítulo 2**
37 El rol de la estadística en el análisis de datos
37 ¿Qué es la estadística?
37 La estadística en la simulación
38 División de la estadística
39 Términos básicos de la estadística
42 Tipos de variable y escalas de medición
44 Variables aleatorias y distribuciones de probabilidad
44 Distribuciones de probabilidad

49 **Capítulo 3**
49 Datos de entrada a un modelo de simulación
50 Preparación de los datos
51 Limpieza de datos
52 Gráfico de dispersión
54 Qué son los datos atípicos

57 **Capítulo 4**
57 Introducción al analisis de datos con R
58 Casos como la estimación

71 **Capítulo 5**
71 Caso de aplicación con R
71 Análisis de los precios de las viviendas
80 Análisis de rotación de los empleados de una empresa

92 **Conclusiones**

93 **Glosario**

96 **Referencias**

# PRESENTACIÓN

Los sistemas de producción se vuelven cada vez más complejos, superando las capacidades de los métodos tradicionales para comprender su comportamiento. En consecuencia, se hace imperativo recurrir a herramientas que puedan representar y evaluar modelos, permitiendo así el desarrollo de estrategias de mejora continua para mantener la competitividad. La simulación, con sus características únicas, se convierte en una solución eficaz para abordar esta necesidad. En la actualidad, existen software especializados como R que permiten desplegar modelos de análisis de datos.

La presente guía tiene como objetivo principal familiarizar a los lectores con los conceptos fundamentales del uso del software R en el análisis de datos, con un enfoque práctico. Los capítulos se estructuran de la siguiente manera:

En el primer capítulo, se realiza una introducción a los conceptos básicos del análisis de datos en R, destacando su importancia. Se expone una estructura metodológica que sirve como guía para llevar a cabo un estudio de análisis de datos, junto con las ventajas y desventajas asociadas. El segundo capítulo aborda en detalle todos los conceptos estadísticos que respaldan los modelos utilizados en el análisis de datos. En el tercer capítulo, se explora el proceso de análisis de datos desde la entrada inicial hasta la preparación de los datos para su posterior modelado. El cuarto capítulo introduce a los lectores en el análisis de datos con R, proporcionando una visión general de la herramienta y sus capacidades. Por último, el quinto capítulo presenta un caso de estudio modelado paso a paso utilizando R, permitiendo a los lectores aplicar los conocimientos adquiridos en un contexto práctico.

**Palabras Claves**: Análisis de datos, estadística aplicada, software R, modelos estadísticos, simulación

# INTRODUCCIÓN

La Corporación Universitaria Comfacauca se encuentra inmersa en un firme compromiso con la mejora continua de las estrategias pedagógicas, buscando constantemente la optimización de la transmisión del conocimiento por parte del docente y su plena asimilación por parte del estudiante, quien debe discernir la aplicabilidad de estos conocimientos en su desarrollo profesional. En esta evolución académica, el análisis de datos emerge como una herramienta imprescindible para validar la información teórica impartida en los variados cursos de Ingeniería Industrial. Dicha herramienta posibilita la percepción de las capacidades y el comportamiento de sistemas complejos sin la necesidad de su reproducción o experimentación en la realidad, lo que se evita debido a los costos, riesgos y limitaciones asociadas.

Este texto introductorio se adentra en la presentación de una guía destinada a la comprensión de los conceptos de simulación y su aplicación en R, un software de gran potencial que habilita la visualización y evaluación de cambios en operaciones, procesos logísticos, gestión de materiales y manufactura de manera eficaz. Todo ello se lleva a cabo sin incurrir en los altos costos, riesgos y prolongados tiempos asociados a la experimentación y el análisis por ensayo y error en el mundo real.

El software R, por su parte, capacita para la exploración de una multiplicidad de escenarios y condiciones, orientando hacia la identificación de soluciones óptimas. Los resultados se presentan con una precisión destacada a través de gráficos, informes y estadísticas generados por la simulación.

Este libro se estructura en cinco capítulos. El primer capítulo abarca los fundamentos esenciales de R, la metodología para la configuración de un estudio de simulación y las ventajas intrínsecas de este software sobre otras herramientas aplicadas en la toma de decisiones.

En el segundo capítulo, se destaca la relevancia de la estadística en la exitosa realización de estudios de simu-

lación, revisando los conceptos de estadística descriptiva e inferencial, y brindando ejemplos ilustrativos para una comprensión más profunda. Además, se abordan los tipos de distribuciones de probabilidad más recurrentes en los modelos.

El tercer capítulo se dedica a la adquisición, preparación y análisis de los datos de entrada en los modelos de simulación, resaltando la importancia de una revisión minuciosa de los datos iniciales, ya que la fiabilidad de los resultados finales depende en gran medida de la integridad de los datos de origen.

El cuarto capítulo introduce el software R, proporcionando una visión general de su interfaz gráfica, explicando el funcionamiento de recursos fijos y ejecutores de tareas, las conexiones entre objetos, las perspectivas del modelo y el uso de la ventana de propiedades.

En el quinto y último capítulo, se presentan tres casos de estudio aplicados a una empresa específica. Se describe detalladamente el proceso de construcción del modelo, con el propósito de poner en práctica los conocimientos adquiridos en el cuarto capítulo. Además, se presentan los resultados a través de indicadores y se incluyen ejercicios complementarios para permitir a los profesores practicar la modelación.

# FUNDAMENTACIÓN TEÓRICA

La ciencia de datos, una disciplina que ha adquirido una importancia sin precedentes en la era digital, tiene sus raíces en una evolución histórica que se ha extendido a lo largo de varias décadas y abarca un amplio espectro de campos de estudio (Liu et al., 2022). Los orígenes de la ciencia de datos se remontan al siglo pasado y se caracterizan por una intrincada mezcla de contribuciones de destacados pioneros en matemáticas, estadísticas, informática y disciplinas relacionadas (Amarasinghe et al., 2020). Si bien la expresión "ciencia de datos" en sí misma es relativamente reciente, las ideas y conceptos que la sustentan se gestaron desde mucho antes, en un contexto en el que la revolución tecnológica y la acumulación masiva de datos comenzaron a transformar la forma en que comprendemos y utilizamos la información (Chong & Xia, 2020).

Uno de los precursores más sobresalientes de la ciencia de datos fue el estadístico y matemático británico Ronald A. Fisher, quien en las décadas de 1920 y 1930 sentó las bases de la estadística moderna (Mölder et al., 2021). Fisher desarrolló técnicas avanzadas que revolucionaron la inferencia estadística, el diseño de experimentos, el análisis de varianza y los métodos de regresión. Sus contribuciones permitieron abordar problemas complejos de manera más precisa y eficaz, allanando el camino para futuros desarrollos en la ciencia de datos (Wang et al., 2022). No obstante, la explosión de la ciencia de datos como un campo de estudio independiente se produjo en las décadas de 1960 y 1970, gracias a la convergencia de múltiples factores. Uno de los aspectos más significativos de esta evolución fue el advenimiento de la informática y la creciente capacidad de procesar grandes volúmenes de datos (Wu et al., 2021). La revolución digital permitió a científicos e ingenieros manipular y analizar datos de maneras que anteriormente eran inimaginables. Esta capacidad dio origen a la gestión de bases de datos como un campo

de estudio independiente, brindando la infraestructura necesaria para gestionar y acceder a grandes conjuntos de datos de manera eficiente (Trinklein & Synovec, 2022).

Simultáneamente, emergieron técnicas de minería de datos que permitieron descubrir patrones y tendencias en conjuntos de datos masivos. La minería de datos se convirtió en un componente esencial de la ciencia de datos, al proporcionar las herramientas necesarias para explorar y extraer conocimiento a partir de la vasta cantidad de información disponible. La ciencia de datos también fue profundamente influenciada por los avances en inteligencia artificial (IA) y aprendizaje automático (Schöneich et al., 2022). En la década de 1950, figuras destacadas como Alan Turing y John McCarthy sentaron las bases de la inteligencia artificial y el concepto de "aprendizaje de máquinas". Estos avances tecnológicos permitieron a los científicos de datos desarrollar algoritmos de aprendizaje automático y técnicas de modelado predictivo, que hoy en día son pilares fundamentales de la ciencia de datos (Ochoa et al., 2023). Otro hito relevante en la historia de la ciencia de datos fue la popularización de la estadística bayesiana, que se desarrolló de manera paralela a las estadísticas tradicionales. La estadística bayesiana se basa en el teorema de Bayes y se convirtió en una herramienta valiosa para modelar la incertidumbre y la probabilidad en los datos. Esto desempeñó un papel esencial en el análisis de datos complejos y en la toma de decisiones informadas.

A medida que las empresas y organizaciones comenzaron a acumular grandes cantidades de datos en la era de la información, la demanda de expertos en análisis de datos aumentó significativamente. Esto llevó al surgimiento de programas académicos especializados en ciencia de datos y al establecimiento de roles profesionales como el científico de datos en la industria. En la actualidad, la ciencia de datos ha evolucionado como una disciplina interdisciplinaria que combina estadísticas, matemáticas, informática y conocimientos de dominio en una amplia variedad de campos. Su aplicación abarca desde la medicina hasta el marketing, la inteligencia artificial, la investigación científica, el análisis financiero, la gestión de la cadena de suministro, y muchos otros sectores. La ciencia de datos se ha convertido en una herramienta fundamental para la toma de decisiones basadas en datos y para descubrir información valiosa a partir de la vasta cantidad de datos disponibles en la era digital (Villanueva & Chen, 2019).

A medida que avanzamos en esta nueva era de información y tecnología, la ciencia de datos continúa desempeñando un papel esencial en la comprensión y optimización de nuestro mundo cada vez más basado en datos. Los profesionales de la ciencia de datos son los arquitectos de la información, encargados de transformar datos en conocimiento y brindar a las organizaciones las herramientas necesarias para tomar decisiones informadas. Su trabajo es clave en campos tan diversos como la atención médica, la investigación científica, la predicción del clima, la ciberseguridad y la logística (Chong et al., 2019).

El papel de la ciencia de datos en la sociedad contemporánea es innegable, y su impacto se extiende a todos los rincones del mundo. A medida que la tecnología continúa avanzando y la cantidad de datos generados crece exponencialmente, la ciencia de datos seguirá desempeñando un papel esencial en la resolución de problemas complejos y en la toma de decisiones estratégicas. Su evolución y desarrollo constante son una muestra de la adaptabilidad de la ciencia y su capacidad para transformar nuestro mundo.

# FUNDAMENTACIÓN METODOLÓGICA

Este libro ha sido meticulosamente diseñado con el objetivo de brindar a los estudiantes un camino de aprendizaje coherente y estructurado, lo que les permitirá adentrarse en el fascinante mundo del software R. Comenzamos nuestro viaje con una introducción profunda a los conceptos fundamentales relacionados con el uso de R, tanto desde una perspectiva teórica como práctica. A lo largo de esta travesía, exploraremos minuciosamente las múltiples ventajas que ofrece esta poderosa herramienta y cómo puede convertirse en un aliado invaluable en la toma de decisiones informadas. Además, abordaremos la cuestión crucial del orden metodológico que debe seguirse al embarcarse en un estudio de análisis de datos, proporcionando una hoja de ruta clara para nuestros lectores.

El papel de la estadística en los modelos de simulación ocupará un lugar central en nuestro recorrido. Entender a fondo los conceptos estadísticos se erige como un pilar fundamental para la realización de un análisis exhaustivo de los datos de entrada a un modelo. Una vez hayas adquirido una comprensión diáfana de estos aspectos, así como de cómo aplicarlos de manera efectiva utilizando el software R, estarás listo para dar vida a la teoría en la práctica. Para ilustrar este proceso, desarrollaremos un caso práctico de simulación que utilizará una empresa manufacturera como estudio de caso. Este enfoque, donde teoría y práctica se entrelazan de manera cohesionada, garantiza que cada concepto se construye sobre una base sólida previamente establecida, proporcionando así una sólida base para tu proceso de aprendizaje.

Con el fin de visualizar la estructura metodológica de este libro de manera más clara, te presentamos un diagrama ilustrativo (ver Figura 1), que te servirá como un mapa detallado para tu travesía de aprendizaje. Este enfoque paso a paso, que se asemeja a una escalera de conocimiento, te brindará la seguridad de que cada peldaño

se funda sobre el anterior, permitiendo así un crecimiento sólido y sostenido de tu comprensión y habilidades en el uso de R y en el análisis de datos. Estamos emocionados de acompañarte en este viaje de descubrimiento y crecimiento, y confiamos en que este libro se convertirá en una valiosa herramienta en tu desarrollo académico y profesional.

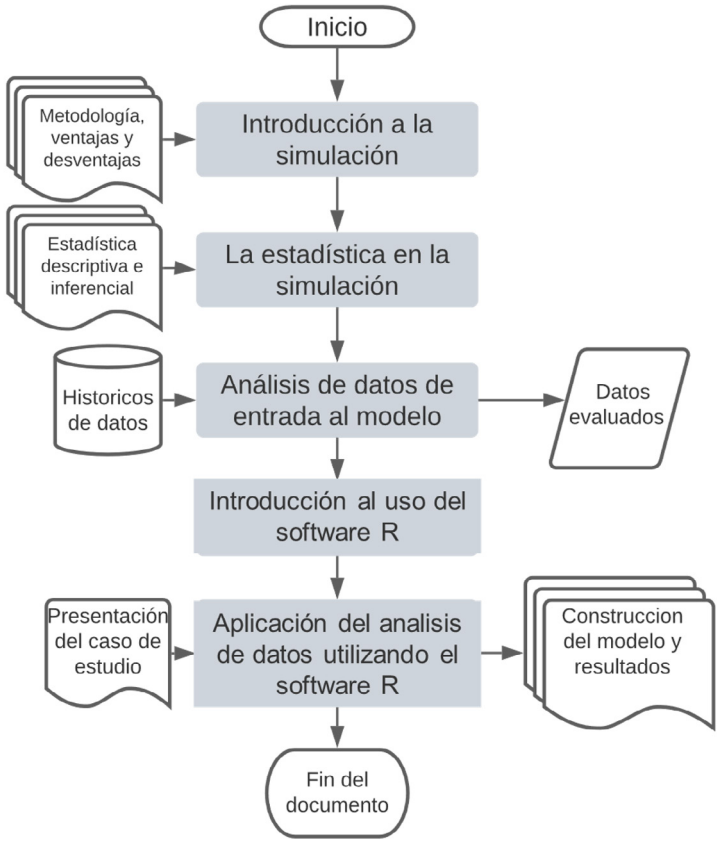



Figura 1. Estructura metodológica del documento

# FUNDAMENTACIÓN CURRICULAR Y DIDÁCTICA

La formación de ingenieros industriales implica la implementación de procesos educativos de naturaleza diversa y variable en cuanto a su nivel de complejidad. Estos niveles de complejidad se ajustan de manera específica a los objetivos individuales de cada curso o materia. La percepción de esta complejidad se encuentra intrínsecamente ligada a la perspectiva de cada participante en el proceso educativo, influida por las materias a tratar y las estrategias de enseñanza aplicadas.

En este contexto, es crucial destacar la importancia de una guía pedagógica como la que se presenta aquí, la cual se erige como una herramienta invaluable para fortalecer y apoyar los planes de estudio relacionados con la asignatura centrada en el análisis de datos a través del uso del software R. Esta guía ofrece una estructura organizada y una serie de recursos diseñados específicamente para facilitar la comprensión de este tema crucial en la formación de ingenieros industriales.

En el siguiente apartado, se identifican y enumeran las asignaturas que, por su naturaleza y contenido, pueden experimentar un impacto positivo aún mayor a través de la integración de técnicas de simulación en sus procesos de enseñanza. Estas asignaturas se beneficiarían especialmente de la incorporación de enfoques basados en análisis de datos para ampliar y enriquecer la experiencia de aprendizaje de los estudiantes.

**Electiva de profundización II (Análisis de datos):** permiten la creación de visualizaciones de datos de alta calidad. Gracias a estas capacidades, los analistas pueden comunicar eficazmente sus hallazgos y resultados a través de gráficos informativos y atractivos.

**Planificación de la producción:** el análisis de datos puede ayudar a mantener un equilibrio adecuado en los

niveles de inventario. Se pueden utilizar modelos de pronóstico para predecir la demanda y, en función de esos pronósticos, se pueden tomar decisiones sobre la cantidad de productos a producir o mantener en stock.

**Diseño y distribución en planta:** el análisis de datos puede ayudar en la optimización de la distribución espacial de las áreas de trabajo, maquinaria, equipos y flujos de materiales en una planta. El análisis de datos geoespaciales y de tráfico puede ayudar a diseñar un flujo de trabajo eficiente y minimizar los desplazamientos innecesarios.

**Logística y cadenas de suministro:** el análisis de datos permite la optimización de rutas de transporte y la gestión eficiente de flotas. Esto ayuda a reducir costos, disminuir los tiempos de entrega y minimizar el impacto ambiental.

# CAPÍTULO 1

# INTRODUCCIÓN AL ANÁLISIS DE DATOS CON R

## ACERCA DE R Y RSTUDIO

El desarrollo de R tuvo sus inicios a principios de la década de 1990, impulsado por George Ross Ihaka y Robert Gentleman, su trabajo inicial se detalla en Ihaka y Gentleman 1996. Con la creación de R, buscaron fusionar las fortalezas de dos lenguajes de programación, S y Scheme (Kronthaler & Zöllner, 2021). Con el paso del tiempo, este proyecto ha experimentado un éxito arrollador, llevando a R a convertirse en el estándar predominante para el análisis de datos en el ámbito universitario, en numerosas empresas y en organismos gubernamentales. Al ser una solución de código abierto, R se beneficia de la ausencia de requisitos de licencia y es plenamente funcional en múltiples sistemas operativos.

R emerge como una plataforma versátil que habilita una variedad inmensa de aplicaciones para el análisis de datos, habitualmente agrupadas en lo que se conoce como "paquetes". Estos paquetes permiten ampliar las capacidades de R para abordar prácticamente cualquier necesidad, ya sea para análisis especializados, representaciones visuales de alta complejidad o demandas específicas. En este contexto, R se presenta como una herramienta poderosa y adaptable. La generación de gráficos aptos para su publicación, en diversos formatos de archivo, es una caracterís-

tica inherente a R. Tanto el propio R como los paquetes de análisis reciben desarrollo continuo de parte de un equipo de programadores y una comunidad activa de usuarios.

Lo que RStudio aporta es una aplicación independiente que opera en base a R, permitiendo el acceso a todas las funcionalidades que ofrece este lenguaje. No obstante, a diferencia de la versión pura de R, RStudio se distingue por presentar una interfaz moderna y atractiva. Este entorno brinda comodidades para el análisis y visualización de datos, al proporcionar una ventana para la creación de scripts, soporte para la entrada de comandos y una gama de herramientas visuales. La interfaz de RStudio está compuesta por cuatro secciones que muestran simultáneamente datos, comandos, resultados y gráficos generados, otorgando así una panorámica integral. La filosofía de RStudio, desarrollada y ofrecida por RStudio, Inc., radica en empoderar a los usuarios para que utilicen R con eficacia y productividad.

En la actualidad, los datos son el recurso fundamental del siglo XXI. En este contexto, RStudio se erige como la herramienta primordial para aprovechar y dar forma a esta invaluable materia prima.

## CÓMO INSTALAR R Y RSTUDIO

Para instalar RStudio, instale R primero. Podemos encontrar R en el sitio web https://www.r-project.org/ (ver Figura 2).

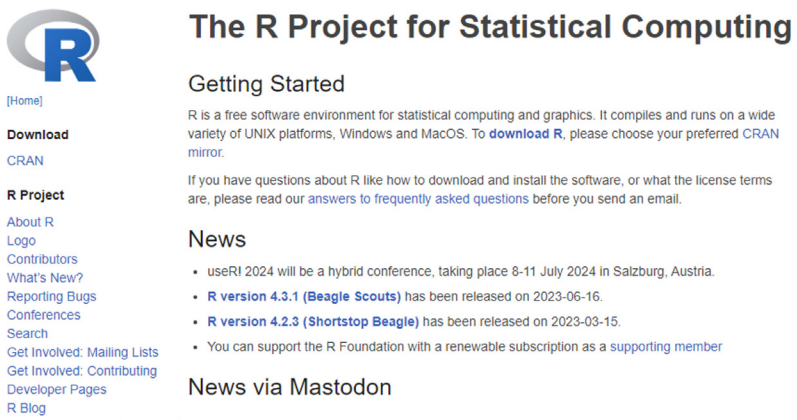



Figura 2. Instalación de R

Haciendo clic en el enlace “Descargar R”, nos dirigimos hacia una página denominada “Cran Mirrors”. Estos espejos son plataformas que ofrecen acceso a R y sus correspondientes paquetes. Al examinar esta página, se observa que R suele ser distribuido por universidades de diversos países. Ejemplos notables incluyen la Universidad de Münster en Alemania, la ETH Zurich en Suiza y la Universidad de Economía de Viena en Austria. En general, optamos por seleccionar el país en el que nos encontramos y accedemos a la sección de descargas. Aquí, podemos obtener versiones de R para Linux, (Mac) OS X y Windows. Al seleccionar la opción de descarga, procedemos a seguir las indicaciones de instalación para finalizar la implementación de R en nuestro sistema. El sitio web principal para la extensa comunidad de usuarios de R es https://www.r-project.org, y es altamente recomendable explorar este recurso. En este sitio, abunda una rica variedad de información acerca de R, su proceso de desarrollo, los contribuyentes en su creación, manuales de uso, una revista y una sección de Preguntas Frecuentes (FAQ).

Una vez completada la instalación de R, se encuentra a nuestra disposición la “R Console”. En este punto, R se encuentra completamente funcional. Sin embargo, al abrir la “R Console”, es evidente que presenta una interfaz rudimentaria y poco atractiva. En este entorno, se nos proporciona información esencial sobre R, como su naturaleza como software de código abierto. Además, se observa una línea de entrada en la cual podemos ingresar comandos (ver Figura 3).

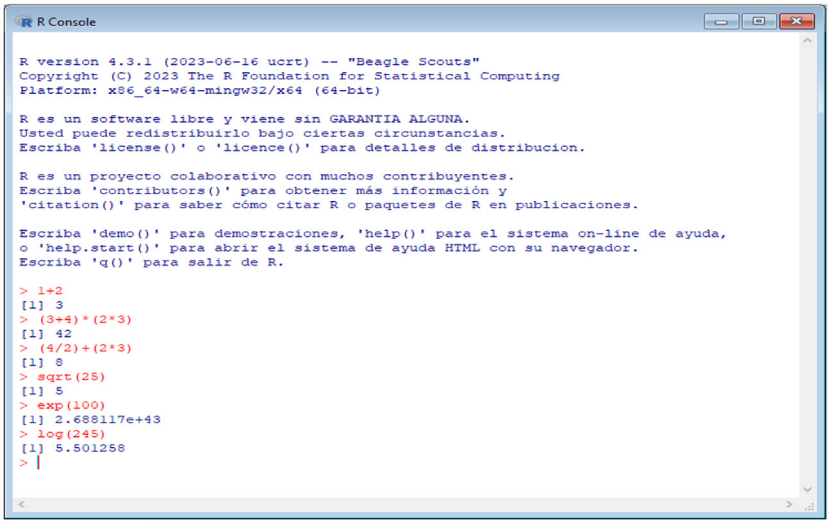



Figura 3. Interfaz de R

A pesar de que, por lo general, no empleamos la Consola directamente para realizar análisis de datos, tal como mencionamos previamente, podemos aprovechar su uso de manera sumamente cómoda al utilizarla como una calculadora en nuestro ordenador. Este planteamiento representa un primer paso introductorio, el cual recomendamos enfáticamente explorar. En la Tabla 1 se detallan diversas funciones esenciales de esta calculadora:

**Tabla 1. Funciones esenciales para realizar cálculos en el software R**

| Función | Descripción | Ejemplo |
|---|---|---|
| + | Adición de valores | 5 + 3 = 8 |
| – | Resta de valores | 2 – 4 = –2 |
| * | Multiplicación de valores | 8 * (–2) = –16 |
| / | División de valores | –16/16 = –1 |
| Sqrt () | Raíz cuadrada de un número | Sqrt (9) = 3 |
| ^ | Potencia de un numero | 3^3 = 27 |
| registro () | Logaritmo natural | registro (120) = 4,79 |
| Exp () | Función exponencial | Exp (10) = 22.026,47 |

Fuente. Elaboración propia

Después de instalar R y tal vez mirar el sitio web y la consola, podemos comenzar a instalar RStudio. El software para los diferentes sistemas operativos se puede encontrar en el sitio web https://posit.co/download/rstudio-desktop/ (ver Figura 4).


Figura 4. Instalación de RStudio

Una vez más, continuamos siguiendo el enlace “Descargar” y accedemos a una página desde la cual es posible obtener diversas variantes de RStudio. Estas opciones incluyen versiones gratuitas para uso en escritorio y servidor, así como alternativas de pago tanto para escritorio como para servidor. Elegimos la versión gratuita para escritorio y, a continuación, navegamos hacia los archivos de instalación para seleccionar el archivo correspondiente a nuestro sistema operativo. Procedemos a seguir detenidamente las indicaciones de instalación, culminando con la instalación exitosa de RStudio en nuestra computadora. Si todos los pasos se han desarrollado de manera adecuada, estamos listos para iniciar RStudio sin contratiempos. Aunque, por lo general, no enfrentamos dificultades, en caso de que surjan, podemos recurrir a recursos en línea para buscar asistencia.

## ESTRUCTURA DE RSTUDIO

Al iniciar RStudio, se despliega una disposición en cuatro secciones. La ventana de script se sitúa en la parte superior izquierda, seguida por la Consola R en la parte inferior izquierda, la ventana de datos y objetos en la esquina superior derecha y, finalmente, la ventana del entorno en la esquina inferior derecha. Esta configuración resulta altamente beneficiosa para llevar a cabo análisis de datos. Si, tras abrir RStudio, la presentación no coincide con la que se muestra a continuación o si preferimos una organización diferente, tenemos la capacidad de ajustarla. Esto se logra dirigiéndonos a “Herramientas” en la barra de menú, seleccionando “Opciones globales” y realizando las configuraciones deseadas bajo la sección de “Disposición del panel”.

En la esquina superior izquierda se localiza la ventana del script (1). Este espacio se emplea para documentar con precisión todo el proceso de análisis de datos y almacenarlo de manera permanente. Aquí anotamos y reunimos los comandos completados, así como los complementamos con comentarios pertinentes, permitiendo su ejecución secuencial. El guion resultante se guarda para asegurar la disponibilidad del análisis incluso en años posteriores. En esta ventana, plasmamos la versión final de todo el análisis.

La esquina inferior izquierda alberga la Consola R (2), una herramienta con la que ya estamos familiarizados y que se instala junto a R. En esta consola, se presentan los

resultados de nuestro análisis conforme ejecutamos los comandos desde la ventana del script. Además, aquí encontramos un espacio para probar comandos, permitiéndonos introducirlos y evaluar su comportamiento. Esta función desempeña un papel crucial en la utilidad de la consola dentro de RStudio.

En la esquina superior derecha se ubica la ventana de datos y objetos (3), la cual nos proporciona una visualización de los conjuntos de datos disponibles en RStudio, es decir, aquellos que están cargados y en uso. No se limita solo a conjuntos de datos activos, sino que también muestra otros objetos generados durante el análisis, como vectores y funciones. Exploraremos más en profundidad este aspecto a lo largo del guion.

Finalmente, en la esquina inferior derecha se encuentra la ventana del entorno (4), que se subdivide en pestañas como Archivos, Gráficos, Paquetes y Ayuda, entre otras (ver Figura 5). La pestaña de Archivos muestra los archivos almacenados en el directorio activo. En la de Gráficos, visualizamos los gráficos generados; en Paquetes, se enumeran los paquetes instalados y existentes; y en Ayuda, encontramos recursos para solicitar asistencia en relación con R y las funciones disponibles. A lo largo del documento, profundizaremos más detalladamente en la función de ayuda.

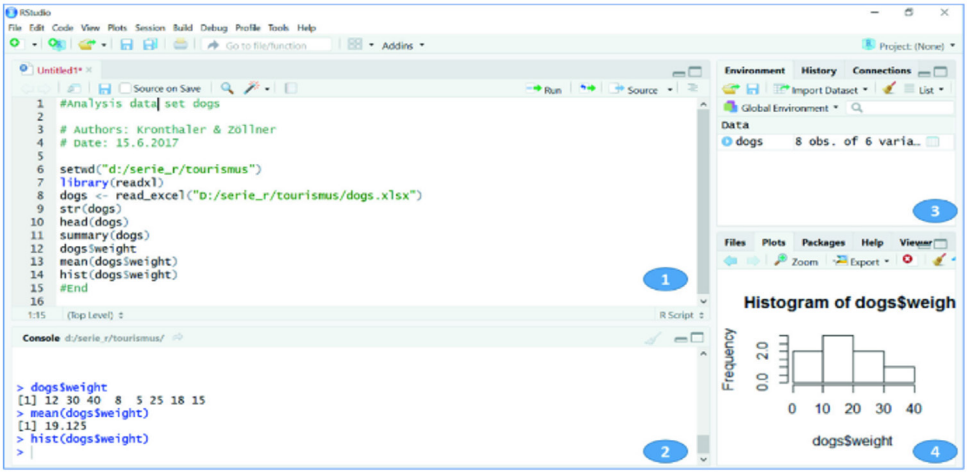



Figura 5. Partes de la interfaz de RStudio

# INSTALAR PAQUETES DE RSTUDIO

Ya hemos comentado varias veces que R o RStudio se basa en paquetes con los que podemos realizar análisis estadísticos. En RStudio hay dos formas de instalar y utilizar paquetes. Los paquetes se pueden instalar mediante menús o mediante el comando correspondiente. El comando es **install.packages()**.

Por ejemplo, se requiere el paquete readxl para leer archivos de Excel. Para recuperar e instalar el paquete correspondiente desde Internet podemos ingresar el siguiente comando:

> instalar.paquetes("readxl")

Es importante tener una conexión a Internet que funcione. **install.packages()** es el comando, entre paréntesis y comillas está el paquete a instalar. Un paquete importante para crear gráficos es, por ejemplo, ggplot2. Si queremos instalarlo, insertamos el nombre del paquete en el comando:

> instalar.paquetes("ggplot2").

No es necesario introducir el comando **install.packages()** a través de la ventana del script. Es mejor ejecutar el comando en la consola. Normalmente instalamos un paquete sólo una vez y, por lo tanto, el comando no forma parte de nuestro script.

Tenga en cuenta que un paquete instalado no se activa automáticamente. Para activar el paquete usamos el comando biblioteca () como se mencionó anteriormente. Sólo con este comando el paquete se activa y puede usarse para el análisis de datos. Para activar ggplot2 ingresamos el siguiente comando:

> biblioteca(ggplot2).

Este comando suele ser parte de nuestro script (lo veremos una y otra vez más adelante).

Si utilizamos la ruta basada en menús, vamos a través de la ventana de entorno, seleccionamos la pestaña Paquetes y hacemos clic en Instalar. Se abrirá una ventana en la que podremos introducir el paquete que quere-

mos instalar. RStudio hará sugerencias tan pronto como comencemos a escribir el nombre. La siguiente figura muestra la instalación basada en menús del paquete ggplot2. Por supuesto, finalizamos la instalación con el botón Instalar (ver Figura 6).

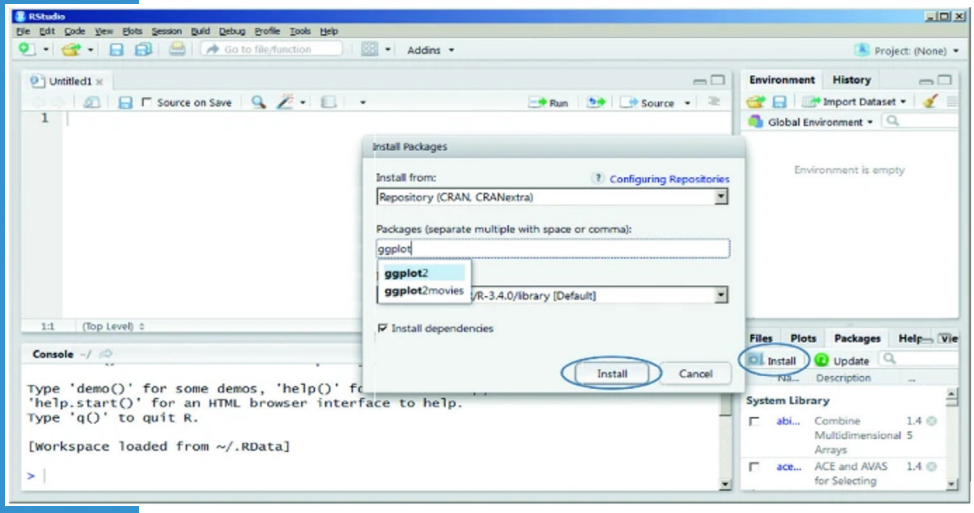



Figura 6. Instalación de paquetes en RStudio

# CÓMO LEER CON RSTUDIO

Antes de comenzar con el análisis de datos con RStudio, primero debemos ocuparnos más intensamente de la lectura de datos con RStudio (ver Figura 7). No hay análisis de datos sin datos. Hay dos formas de importar datos a RStudio. Podemos usar el botón Importar conjunto de datos en la ventana Datos y objetos (d Importar datos con RStudio) o podemos ingresar y ejecutar la sintaxis apropiada en la ventana Script. Se pueden importar diferentes formatos de datos a RStudio. Normalmente, los datos están disponibles como un archivo de texto, por ejemplo, con la extensión .csv, o como un archivo de Excel con la extensión .xlsx. Por lo tanto, analizaremos sólo estos dos formatos de datos.

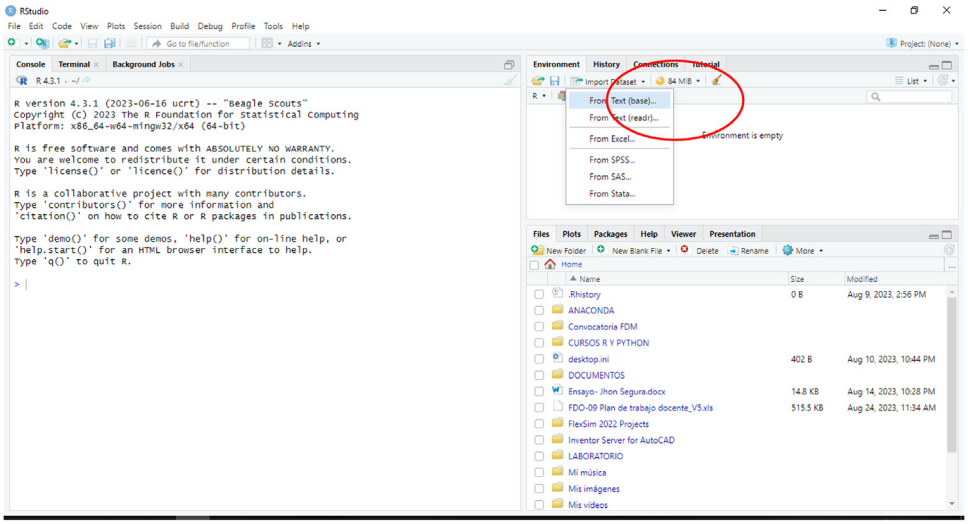



Figura 7. Importación de datos en RStudio

Si hacemos clic en el botón Importar conjunto de datos, podemos seleccionar el formato de datos apropiado, por ejemplo, desde Excel. Se abrirá una ventana donde podremos buscar e importar nuestro archivo de Excel. Dependiendo del formato, también podemos especificar aquí las opciones de importación. Para un archivo de Excel, por ejemplo, podemos especificar qué hoja del libro de Excel queremos importar o que la primera fila contenga los nombres de las variables. En esta ventana podemos ver la vista previa del conjunto de datos y comprobar visualmente si los datos se están leyendo correctamente.

La segunda opción es ingresar directamente el comando de importación en la ventana del script. Necesitamos saber el comando y dónde se encuentra el archivo de datos en nuestra computadora, es decir, la ruta. El comando para importar un archivo de Excel es (si el paquete readxl está instalado):

Como ya se mencionó, activamos el paquete readxl con el comando biblioteca (). El segundo paso es nombrar el conjunto de datos. Normalmente utilizamos el nombre del archivo de Excel para evitar confusiones. El archivo de Excel se llama perros. Entonces comenzamos el comando con perros seguido del operador de asignación “<-”. Luego sigue el comando read_excel (), con la ruta y el archivo de Excel entre comillas dentro de los corchetes.

Con la ruta debemos tener en cuenta que al contrario del procedimiento normal con el ordenador R no funciona con barras invertidas, sino con la barra diagonal "/".

Para leer los datos tenemos que escribir el comando y enviarlo. Podemos proceder paso a paso o enviar el comando de inmediato. Si procedemos paso a paso, colocamos el cursor en la línea respectiva (por supuesto, comenzamos con el comando biblioteca (()) y luego hacemos clic en el botón Ejecutar en la esquina superior derecha (ver Figura 8).

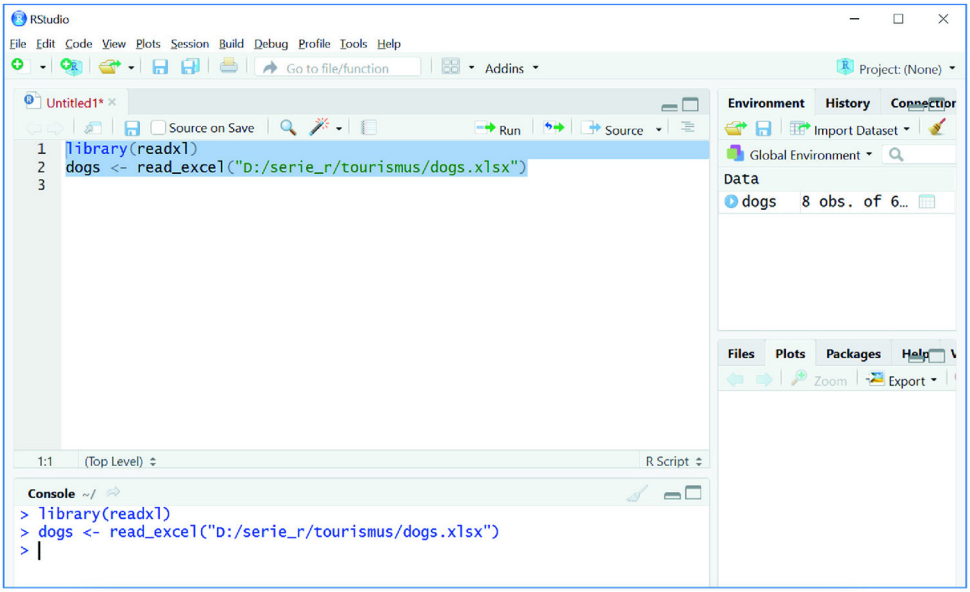



Figura 8. Código para leer la base de datos

Cuando la importación de datos se ha completado con éxito, el conjunto de datos (activo) aparece en la ventana superior derecha (Fig. 2.2: Importación de datos de Excel usando la sintaxis de comando. Si queremos importar un archivo CSV en lugar de un archivo Excel hacemos lo mismo. Si trabajamos mediante menús, seleccionamos Desde CSV en lugar de Desde Excel y especificamos los datos necesarios. Es importante que sepamos cómo se separan los personajes entre sí. Para el archivo CSV suele ser la coma "," o el punto y coma ";".

**mi primer programa**

Utilice la consola R como calculadora. Calcula los siguientes términos: (5 + 2) * (10 − 3), raíz cuadrada de 64, tercera raíz de 64, tercera raíz de 512, 4 elevado a la tercera, 8 elevado a la tercera, logaritmo de 500, logaritmo de 10, número de Euler exp (1), valor exponencial e x de x = 5.

Descargue el conjunto de datos dogs.xlsx de http://www.statistik-kronthaler.ch/ y guárdelo en una carpeta de su computadora portátil. Intente utilizar una ruta corta y una carpeta con un nombre sencillo. Recuerda la ruta y la carpeta; es el camino y el lugar para encontrar los datos.

Lea el conjunto de datos dogs.xlsx mediante el botón Importar conjunto de datos. Elimine el conjunto de datos con el comando rm() . Lea el conjunto de datos nuevamente con el comando de importación en la ventana del script. Elimina tu memoria de trabajo por completo con el comando correcto.

Lea el conjunto de datos dogs.xlsx e intente realizar las siguientes tareas:

Eche un vistazo al conjunto de datos con el comando correspondiente. Verifique si el conjunto de datos está completo mirando el principio y el final del conjunto de datos con los comandos relevantes. Eche un vistazo a la información general sobre el conjunto de datos, es decir, el tipo de datos de las variables, los valores característicos, etc. Produzca un conjunto de datos reducido con solo 6 observaciones y sin la variable raza. Nombra el conjunto de datos dogs_red. Trabaje con el conjunto de datos dogs_redproducido en la tarea 5. ¿Qué edad tiene el perro en la cuarta observación? Intente resolverlo mostrando solo este valor. Verifique el valor recibido con la ayuda del comando View (). Genere con humanage una nueva variable que se agregue al conjunto de datos dogs.xlsx. Has oído que un año de perro equivale sólo a 5 años humanos, no a 6. Además, convierte las variables sexo, tamaño y raza en factores.

Produzca su primer script pequeño con la ayuda del conjunto de datos dogs.xlsx (puede hacer lo mismo con el conjunto de datos reducido producido en la tarea 5), guárdelo y guarde el informe también como un archivo de Word. Para ello incluya todas las cuestiones relevantes y utilice para el análisis de datos los siguientes comandos:

resumen(perros);

hist(perros$edad);

numSummary (perros [,"edad", drop=FALSE], estadísticas=c("media", "sd", "IQR", "cuantiles"), cuantiles=c(0,.25,.5,.75, 1));

boxplot (perros$edad~perros$f_sexo);

numSummary (perros [,"edad", drop=FALSE], grupos=perros$f_sex, estadísticas=c("media", "sd", "IQR", "cuantiles"), cuantiles=c(0,.25, .5,.75,1)).

**Una primera aplicación de análisis de datos con RStudio**

En la Figura 9, se ilustra un análisis de datos inicial de pequeña escala. Ahora, tomemos un momento para explorar en detalle este análisis y comprender la lógica subyacente en RStudio.

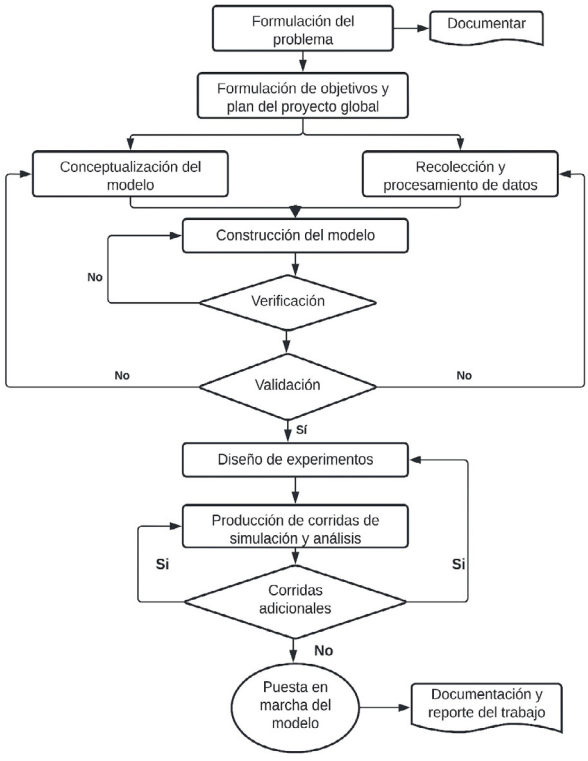



Figura 9. Estructura metodológica de un modelo de simulación

# VENTAJAS DE UTILIZAR RSTUDIO

RStudio es un entorno de desarrollo integrado (IDE) especialmente diseñado para trabajar con el lenguaje de programación R, que se utiliza principalmente en estadísticas y análisis de datos.

**Interfaz Amigable:** RStudio proporciona una interfaz de usuario intuitiva y amigable que facilita la escritura, prueba y depuración de código en R. Esto lo convierte en una excelente opción tanto para principiantes como para usuarios experimentados.

**Editor de Código:** RStudio incluye un editor de código R enriquecido con características como resaltado de sintaxis, auto-completado y atajos de teclado que agilizan la escritura de código y mejoran la productividad.

**Exploración de Datos Interactiva:** RStudio permite la exploración de datos de forma interactiva. Puedes ver y modificar objetos, tablas y gráficos en tiempo real, lo que facilita la comprensión de tus datos.

**Administración de Proyectos:** RStudio incluye herramientas para gestionar proyectos de manera eficiente, lo que te permite organizar tus archivos y mantener un entorno de trabajo limpio y estructurado.

**Gráficos Integrados:** RStudio ofrece una variedad de paquetes y herramientas para crear visualizaciones de datos de alta calidad de manera sencilla. Los gráficos se pueden ver y ajustar en tiempo real.

**Depuración de Código:** el IDE incluye herramientas para depurar el código de manera efectiva, lo que facilita la identificación y corrección de errores en tu programa.

**Integración con Git:** RStudio se integra perfectamente con sistemas de control de versiones como Git, lo que facilita el seguimiento de cambios en tu código y la colaboración en proyectos.

**Extensibilidad:** RStudio es altamente extensible a través de paquetes y complementos. Esto significa que puedes

personalizar tu entorno de trabajo agregando funcionalidades adicionales según tus necesidades.

**Documentación y Ayuda Integradas:** RStudio ofrece acceso a la documentación de R y paquetes directamente desde el IDE, lo que facilita la búsqueda de ayuda y la resolución de problemas.

**Comunidad Activa:** RStudio cuenta con una comunidad activa de usuarios y desarrolladores, lo que significa que es fácil encontrar recursos, tutoriales y soluciones a problemas comunes en línea.

**Multiplataforma:** RStudio es compatible con Windows, macOS y Linux, lo que permite a los usuarios trabajar en sus plataformas preferidas.

**Versión Gratuita:** RStudio ofrece una versión gratuita y de código abierto, lo que lo hace accesible para usuarios con presupuestos limitados.

## DESVENTAJAS DE RSTUDIO

**Curva de Aprendizaje:** a pesar de ser amigable para principiantes, RStudio y R en general pueden tener una curva de aprendizaje pronunciada, especialmente para personas nuevas en la programación o el análisis de datos. Comprender la sintaxis de R y las mejores prácticas puede llevar tiempo.

**Uso Específico para R:** RStudio está diseñado principalmente para trabajar con R. Si necesitas trabajar con otros lenguajes de programación, como Python, tendrás que buscar soluciones alternativas o usar IDEs **más versátiles.**

**Requiere Conocimiento de R:** para aprovechar al máximo RStudio, es necesario tener un conocimiento sólido de R. Esto puede ser una desventaja si necesitas un entorno más amigable para usuarios que no son programadores.

**Rendimiento Limitado para Grandes Conjuntos de Datos:** R puede tener dificultades para manejar grandes

conjuntos de datos debido a su naturaleza de lenguaje interpretado. RStudio, como IDE, no soluciona por completo este problema y puede experimentar ralentizaciones en la manipulación de datos masivos.

**Interfaz Basada en Ventanas:** algunos usuarios pueden preferir un IDE con una interfaz basada en pestañas en lugar de ventanas, lo que puede hacer que la organización de múltiples scripts y archivos sea un poco menos eficiente.

**Recursos del Sistema:** RStudio puede ser más exigente en términos de recursos del sistema en comparación con otros IDEs **más ligeros. Si trabajas en una computadora con recursos limitados, es posible que experimentes problemas de rendimiento.**

**No es una Solución Todo en Uno para Ciencia de Datos:** aunque RStudio es excelente para la programación en R y el análisis de datos, no proporciona todas las herramientas necesarias para todo el flujo de trabajo de ciencia de datos. Es posible que necesites utilizar otros programas y herramientas en combinación con RStudio.

**Dependencia de Paquetes de Terceros:** la funcionalidad adicional en RStudio, como la generación de informes con R Markdown, depende de paquetes de terceros, lo que puede generar problemas de compatibilidad y actualizaciones.

**Problemas de Compatibilidad entre Versiones:** a veces los scripts y paquetes escritos en una versión de R, no son completamente compatibles con versiones posteriores, lo que puede requerir ajustes en tu código y causar problemas de migración.

**Personalización Limitada para Interfaz:** aunque RStudio es extensible a través de complementos y paquetes, la personalización de la interfaz del usuario es limitada en comparación con algunos otros IDEs.

# CAPÍTULO II

# EL ROL DE LA ESTADÍSTICA EN EL ANÁLISIS DE DATOS

## ¿QUÉ ES LA ESTADÍSTICA?

Es una ciencia que se encarga de proporcionar diferentes métodos y procedimientos para recoger, clasificar, resumir, encontrar regularidades y analizar datos, así como para hacer inferencias a partir de ellos, con el fin de ayudar en la toma de decisiones y, en su caso, formular predicciones. La estadística surge como la herramienta ideal para cercar los efectos de la incertidumbre inherentes a la gran mayoría de los procesos biológicos, químicos, industriales y de comportamiento humano, donde predominan los efectos del azar y la incertidumbre. Esto nos lleva a preguntarnos cuál es el papel de la estadística en la simulación. Además, cual es el impacto en los datos finales (Correoso Espinosa et al., 2011).

## LA ESTADÍSTICA EN LA SIMULACIÓN

En el contexto estadístico, entendemos por simulación, la técnica de muestreo estadístico controlado, que se utiliza conjuntamente con un modelo, para obtener respuestas aproximadas a preguntas que surgen en problemas complejos de tipo probabilístico (Diharce, 2008).

La simulación puede entenderse como el arte de construir modelos para estudiar el comportamiento de un sistema real. Un modelo es una representación de un sistema de interés a través de elementos concretos o abstractos. Diversas simulaciones se desarrollan con base en modelos matemáticos de carácter probabilístico o de corte determinístico que no incluye lo aleatorio; la naturaleza del sistema a estudiar generalmente proporciona indicios acerca de cuál ha de ser el modelo más apropiado para su análisis. Para entender cómo se aplica la estadística en la simulación, se debe conocer los tipos de estadísticas que se encuentran (Ángel & Pantoja, 2014).

# DIVISIÓN DE LA ESTADÍSTICA

La estadística se divide en dos grandes áreas:

### Estadística descriptiva

La estadística descriptiva comprende métodos para organizar, resumir y presentar los datos de forma informativa. Su propósito es únicamente exploratorio y se limita a describir lo que se observa en una población o muestra. Su objetivo es la exploración sin restricciones de los datos en busca de regularidades interesantes. Las conclusiones se aplican únicamente a los individuos y circunstancias de los que se obtuvieron los datos, y son informales y se basan en lo que se observa en los datos (Ángel & Pantoja, 2014).

Ejemplo: Descripción de la producción mensual de café durante el año 2016 a través de una tabla o gráfico lineal o de barras, además se pueden comparar las variaciones porcentuales del año 2016 con respecto al 2015.

### Estadística inferencial

La estadística inferencial consiste en el proceso inductivo que permite inferir o generalizar a toda la población las características observadas en una muestra. Su objetivo es responder a preguntas concretas que se plantearon antes de obtener los datos. Las conclusiones se aplican a un grupo más amplio de individuos o situaciones, son

formales y se hace explícito el grado de confianza en ellas (Ángel & Pantoja, 2014).

Ejemplo: con base en una muestra aleatoria regional de 550 estudiantes en el Valle del Cauca, se encontró que el 45% de ellos están en primaria; esta proporción es generalizada a toda la población del departamento.

# TÉRMINOS BÁSICOS DE LA ESTADÍSTICA

En general, la estadística hace uso del método científico para realizar diferentes procesos de análisis de observaciones o datos que luego se convierten en información valiosa para la toma de decisiones en diferentes ámbitos. Sin embargo, la metodología estadística es a menudo mal interpretada debido a un proceso inadecuado que parte de los datos, ignorando su origen y el proceso de obtención de estos. Este proceso ignora elementos importantes en el análisis estadístico como los instrumentos de medición utilizados, la población encuestada, las preguntas formuladas o las variables, las escalas de medición, etc. Por eso es importante aclarar algunos conceptos que no deben faltar en el análisis estadístico, como la validez de un estudio (Villarreal Larrinaga & Landeta Rodríguez, 2010).

### Validez

En un estudio o investigación, la validez es el grado de valor que se confiere a la información. Una medida es válida cuando mide lo que pretende medir. La validez puede ser de dos tipos: interna y externa (Martínez-García & Martínez-Caro, 2009).

### Validez interna

Es el grado en que la medición refleja la situación que pretende medir, a través de esta validez es posible responder a las siguientes preguntas ¿El instrumento es adecuado, la medición está bien realizada, la medición puede repetirse y obtener los mismos resultados, la medición puede repetirse y obtener los mismos resultados?

Ejemplo: si se quisiera medir la distancia entre dos objetos A y B (ver Figura 10), ¿Qué instrumento podríamos utilizar que proporcione datos confiables?

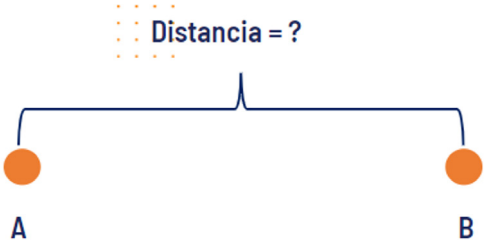



Figura 10. Distancia entre dos objetos

**Validez externa**

Es el grado en que la medición puede generalizarse a otras situaciones no medidas, depende en gran proporción de la conformación adecuada de la muestra a través de métodos estadísticos.

Ejemplo**:** un estudio desea concluir sobre el ingreso promedio por vivienda de las familias de la ciudad Cali. Si en el estudio solo se encuestan familias que pertenecen al estrato 1 ¿Se podría concluir que el ingreso promedio calculado sobre ese grupo representa el promedio de ingresos para toda la ciudad?

La respuesta es NO, la información obtenida a partir de ese estudio no podrá ser generalizada a toda la población dado que sólo se tuvo en cuenta cierto sector de la ciudad y se ignoraron las familias del estrato 2 al 6 (ver Figura 11).


Figura 11. Selección de muestra

**Individuo.** Son los elementos u objetos que tienen información sobre el fenómeno estudiado, es decir, aquellos objetos descritos por un conjunto de datos. Los individuos pueden ser personas, pero también pueden ser animales o cosas (Kelmansky, 2010).
**Ejemplo:** Si se estudia el peso promedio de los estudiantes en una carrera de la universidad, cada uno de los estudiantes son los individuos.

**Variable.** Es cualquier característica de interés de un individuo. Una variable puede tomar distintos valores para distintos individuos (Kelmansky, 2010).
**Ejemplo:** la edad de los estudiantes de primer semestre o el número de habitantes por vivienda.

**Población.** Es el conjunto de todos los elementos de interés en un estudio, sobre los cuales se desea información y hacia los cuales se extenderán las conclusiones (Kelmansky, 2010).
**Ejemplo:** El total de estudiantes matriculados en la universidad o en total d la población en la ciudad de Cali. 3.5.

**Muestra.** Es cualquier subconjunto representativo de la población, sobre el que se realizan los estudios para obtener conclusiones acerca de las características de la población (Kelmansky, 2010).
**Ejemplo:** 100 estudiantes seleccionados del total de matriculados en la universidad o 500 personas seleccionadas de la población de Cali (ver Figura 12).

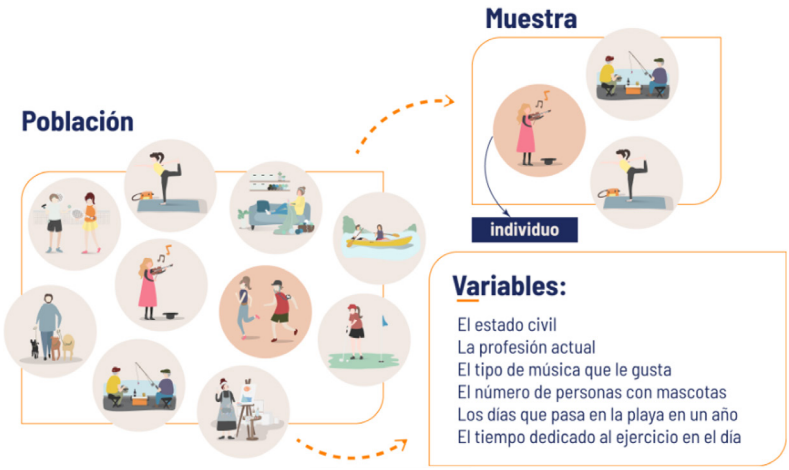



Figura 12. Selección de muestra

Parámetro. Indicador estadístico calculado teniendo en cuenta los elementos de la población (Kelmansky, 2010).

Ejemplo: Edad promedio de todos los estudiantes matriculados en la universidad o los ingresos promedio de toda la población de la ciudad de Cali.

Estimador. Indicador estadístico calculado teniendo en cuenta los elementos de la muestra (Kelmansky, 2010).

Ejemplo: edad promedio de una muestra de 100 estudiantes matriculados en la universidad o los ingresos promedio de una muestra de 500 personas de la ciudad de Cali (ver Figura 13).

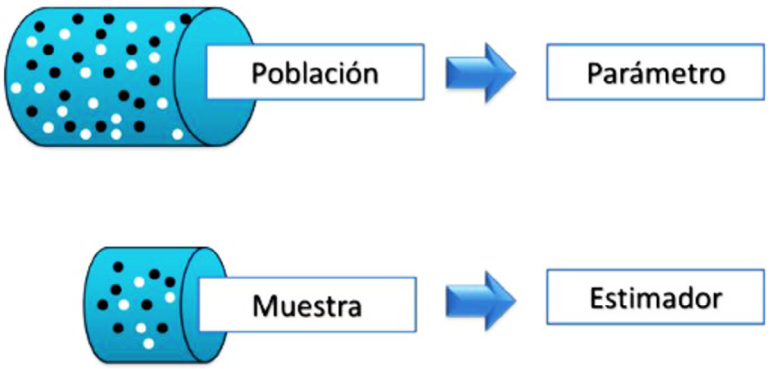



Figura 13. Concepto de estimador

# TIPOS DE VARIABLES Y ESCALAS DE MEDICIÓN

**Variables cualitativas**

Sus valores (categorías o modalidades) no se pueden asociar naturalmente a un número, ni se pueden hacer operaciones matemáticas con ellos. Estas variables presentan dos escalas de medición: nominal y ordinal (Hidalgo Troya, 2019).

- **Escala de medición nominal.** En las variables con escala de medición nominal, sus valores no se pueden ordenar, es decir que no existe una forma particular para ordenar sus valores (Hidalgo Troya, 2019).

**Ejemplos:** la religión (católico, cristiano, evangélico, ateo), el género (mujer, hombre), la nacionalidad (colombiano, venezolano, ecuatoriano), etc.

- **Escala de medición ordinal.** En las variables con escala de medición ordinal, sus valores tienen un orden natural o se puede identificar un orden jerárquico entre sus valores o categorías (Hidalgo Troya, 2019).
  **Ejemplos:** el nivel educativo (primaria, secundaria, universitario, especialización, maestría, doctorado, postdoctorado), el estrato socioeconómico (1, 2, 3, 4, 5, 6), los meses del año (enero, febrero, marzo), etc.

### Variables cuantitativas

Toman valores numéricos y se pueden hacer operaciones matemáticas con ellos. Estas se clasifican en dos grupos: continuas y discretas (Hidalgo Troya, 2019).

- **Variables continuas.** Las variables continuas pueden tomar cualquier valor real dentro de un intervalo (Hidalgo Troya, 2019).
  **Ejemplos:** la estatura de los estudiantes de este curso, las temperaturas registradas al medio día durante el mes pasado, el peso de los profesores de estadística, etc.
- **Variables discretas.** Las variables discretas sólo toman valores enteros (Valenzuela et al., 2021) .
  **Ejemplos:** el número de hijos, el número de carros, el número de materias matriculas este semestre, etc. Tanto las variables continuas como las discretas presentan dos escalas de medición: intervalo y razón.
- **Escala de medición de intervalo.** En estas variables el cero (0) es un valor arbitrario, no implica la ausencia de una característica (Valenzuela et al., 2021).
  **Ejemplo:** una temperatura de cero grados no indica que no hay temperatura, sólo es un valor que toma la variable para esa escala.
- **Escala de medición de razón.** El cero (0) refleja ausencia de la característica, implica la ausencia de una característica.
  **Ejemplo:** un valor cero en el número de hijos quiere decir que esa persona no tiene hijos

# VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD

Uno de los conceptos más importantes en estadística de probabilidad es el de variables aleatorias que es una característica medible que toma diferentes valores con probabilidades determinadas. Las variables aleatorias de un sistema se deben modelar a ciertas ecuaciones matemáticas que sean capaces de reducir su variabilidad. En la mayoría de los casos dicha variabilidad puede clasificarse dentro de alguna distribución de probabilidad, por lo que uno de los pasos más importantes en todo proceso de modelado de simulación es encontrar el tipo de distribución de probabilidad que modela la variabilidad de los datos de entrada del sistema. Por lo tanto, una distribución de probabilidad se define como una función que asigna a cada suceso definido sobre la variable aleatoria la probabilidad que dicho suceso ocurra (Beltran-Pellicer et al., 2018).

# DISTRIBUCIONES DE PROBABILIDAD

A continuación, se exponen las distribuciones de probabilidad de mayor uso en el modelado de sistemas, sus aplicaciones prácticas y su relación con el software R:

**Distribución de probabilidad discreta**

Estas describen la probabilidad de que ocurran valores específicos. Dentro de las principales y más utilizadas (Teresa et al., 2018).

- Distribución uniforme discreta
- Binomial
- Poisson
- Geométrica
- Bernoulli

- Distribución uniforme discreta. Esta distribución describe el comportamiento de una variable discreta que puede tomar n valores distintos con la misma probabilidad cada uno de ellos, esto ocurre cuando los valores son enteros consecutivos. Esta distribución asigna igual probabilidad a todos los valores enteros entre el límite inferior y superior que define el recorrido de la variable. Por ejemplo, cuando se observa el número obtenido mediante el lanzamiento de un dado, los valores posibles siguen una distribución uniforme discreta en (1,2,3,4,5,6) y la probabilidad de cada cara es 1/6.

Valores:
x: a, a+1, a+2,.....,b número entero
Parámetro**:**
a: mínimo, donde a es un número entero
b: máximo, donde b es un número entero a < b

- Distribución binomial. La distribución binomial es una distribución discreta de mucha importancia que surge en muchas aplicaciones estadísticas. Esta distribución aparece al realizar repeticiones independientes de un experimento que tenga respuesta binaria, generalmente clasificado como éxito o fracaso. Por ejemplo, esa respuesta la podemos encontrar en los resultados de los exámenes parciales, aprobado o no aprobado. La variable discreta que cuenta el número de éxitos en n pruebas independientes de ese experimento cada una con la misma probabilidad de éxito igual a p, sigue una distribución binomial de parámetros n y p.

Valores:
x. 0,1,2,.....n
Parámetros:
n: Número de pruebas en donde n > 0, entero
p: probabilidad de éxito 0 < p < 1

- Distribución de poisson. Esta distribución surge cuando un evento o seceso ocurre aleatoriamente en el espacio o tiempo. La variable asociada es el número de ocurrencias del evento en un intervalo o espacio continuo, por lo tanto, es una variable aleatoria discreta que toma valores mayores a 0.

Por Ejemplo, un número de pacientes que llegan a un consultorio en un tiempo estipulado, el número de llamadas que recibe un servicio de atención a urgencias durante una hora, estos ejemplos siguen una distribución de Poisson.

Valores:
x: 0,1,2,3,,,,,,,,n
Parámetros:
Lambda (λ): media de la distribución, λ > 0

Distribución geométrica**.** Esta distribución permite calcular la probabilidad de los datos, a partir de un número k de repeticiones hasta obtener un éxito por primera vez. La distribución geométrica se utiliza en la distribución de tiempo de espera. Esta distribución presenta la denominada propiedad de Markov.

Valores:
x: 0,1,2,,,,,,n
Parámetros:
p: probabilidad de éxito 0 < p < 1

Distribución de Bernoulli. Es una distribución de probabilidad discreta que toma valor de 1 para la probabilidad de éxito y valor 0 para la probabilidad de fracaso. Consiste en realizar un experimento aleatorio una sola vez y observar si el suceso ocurre o no, siendo p la probabilidad que esto sea exitoso y q = 1 – p la probabilidad que no lo sea fracaso. Existen muchas ocasiones en las que se presenta una situación de esta naturaleza. Cada uno de los experimentos es independiente de los restantes o sea que la probabilidad del resultado de un experimento no depende del resultado del resto.

Valores:
x: 0,1,2,.....n
Parámetros:
p: probabilidad
Valor de éxito: valor para representar el éxito del experimento
Valor de fracaso: Valor para representar el fracaso del experimento

**Distribución de probabilidad continuas**

Son aquellas que describen una variable aleatoria que no sean valores específicos.

Distribución uniforme. Esta distribución es útil para describir una variable aleatoria con probabilidad constante sobre el intervalo [a,b] en el que se encuentra definida. Esta distribución presenta una singularidad importante: la probabilidad de un suceso dependerá exclusivamente de la amplitud del intervalo considerado y no de su posición en el campo de variación de la variable.

Campo de variación: $a \leq x \leq b$

Parámetros: a: mínimo del recorrido; b: máximo del recorrido

Distribución normal. Es una probabilidad más importante del cálculo de probabilidades y de la estadística en general. La importancia de esta distribución queda evidenciada por ser el límite de numerosas variables aleatorias, discretas y continas.

Campo de variación

$-\infty < x < \infty$

Parámetros:

μ: (Mu)**,** media de la distribución, $-\infty < x < \infty$

λ: (Sigma), desviación estándar de la distribución, $\lambda > 0$

Se expone un ejercicio con la finalidad de determinar la distribución de probabilidad a la cual se ajustan los datos.

# CAPÍTULO 3

# DATOS DE ENTRADA A UN MODELO DE SIMULACIÓN

El análisis de datos se alza como un poderoso recurso para abordar, evaluar y visualizar tanto los procesos ya existentes como los recién concebidos, sin incurrir en los riesgos asociados a experimentar directamente en sistemas reales. En esencia, esta disciplina otorga a las organizaciones una valiosa oportunidad de examinar sus operaciones desde una perspectiva sistemática, permitiéndoles obtener una comprensión más profunda de las relaciones de causa y efecto entre diversos elementos, además de habilitar una mayor capacidad de anticipación y predicción de situaciones futuras.

El análisis de datos se convierte en una herramienta estratégica al facilitar la evaluación, reevaluación y medición de una serie de factores críticos. Por ejemplo, es posible utilizar esta disciplina para determinar la satisfacción del cliente con respecto a un nuevo proceso implementado, o para medir con precisión la utilización de recursos en dicho proceso, lo que a su vez permite identificar oportunidades de optimización. Además, el análisis de datos puede ayudar a reducir el tiempo necesario para alcanzar la eficiencia óptima, lo que es esencial en un entorno competitivo.

Sin embargo, para que todas estas ventajas se materialicen, es necesario llevar a cabo una preparación rigurosa de los datos. Esto implica la identificación y corrección de valores atípicos, que pueden introducir ruido y sesgar la

interpretación de los resultados finales. Dicha preparación es una fase crítica en el proceso de análisis de datos, ya que garantiza la calidad y la confiabilidad de los resultados, lo que a su vez permite a las organizaciones tomar decisiones más informadas y efectivas para mejorar sus operaciones y alcanzar sus objetivos estratégicos. (Vargas-Rojas, 2021).

## PREPARACIÓN DE LOS DATOS

La preparación de datos en un proyecto de simulación desempeña un papel fundamental en la optimización de la utilidad y aplicabilidad del conjunto de datos. Estas técnicas tienen como objetivo principal poner en condiciones los datos para que puedan ser procesados de manera eficaz por algoritmos de clasificación, segmentación o regresión, lo que a su vez facilita la obtención de información valiosa y resultados confiables. En este proceso de preparación, se pueden identificar diversas tareas agrupadas en los siguientes bloques temáticos, como se ha destacado en investigaciones previas (Ruiz Manosalva, 2019):

- **Tareas de limpieza de datos:** estas tareas se centran en la corrección o eliminación de datos ruidosos o inválidos que podrían afectar la calidad y precisión del análisis. Esto implica identificar y tratar valores atípicos, datos faltantes o inconsistencias en el conjunto de datos. La limpieza de datos garantiza que la información utilizada sea confiable y precisa.
- **Tareas de normalización de datos:** la normalización se ocupa de estandarizar la escala y la distribución de los datos, lo que facilita su comparación y análisis. Al poner los datos en el mismo rango, se elimina el sesgo potencial que podría surgir de unidades de medida diferentes y permite una mejor interpretación de las relaciones entre las variables.
- **Tareas de discretización de datos:** en ocasiones, es necesario convertir variables continuas en categóricas para satisfacer los requisitos específicos de ciertos algoritmos o para abordar situaciones particulares. Este proceso, conocido como desratización, implica la creación de intervalos o cate-

gorías para las variables continuas, lo que puede simplificar el análisis y mejorar la precisión de ciertos modelos.

- **Tareas de reducción de la dimensionalidad:** la reducción de la dimensionalidad se enfoca en la simplificación de conjuntos de datos que contienen numerosas variables, lo que puede ayudar a desarrollar modelos con un menor número de características, reduciendo así la complejidad computacional y permitiendo un análisis más eficiente. Esto es especialmente útil cuando se trabaja con grandes volúmenes de datos, ya que puede mejorar significativamente la velocidad de procesamiento y la interpretación de resultados.
- La aplicación cuidadosa de estas técnicas de preparación de datos es esencial para garantizar que los resultados de un proyecto de simulación sean confiables y proporcionen información valiosa para la toma de decisiones. La calidad de los datos y su adecuación para su uso posterior son elementos críticos en la generación de insights precisos y efectivos en el ámbito de la simulación.

## LIMPIEZA DE DATOS

En el proceso de limpieza de datos (en inglés, data cleansing o data scrubbing) se llevan a cabo actividades de detección, eliminación o corrección de instancias corrompidas o inapropiadas en los juegos de datos.

A nivel de valores de atributos se gestionan los valores ausentes, los erróneos y los inconsistentes. Un ejemplo podrían ser los valores fuera de rango (outliers).

El proceso de integración de datos puede ser una de las principales fuentes de incoherencias en los datos. Fruto de la fusión de juegos de datos distintos, se pueden generar inconsistencias que deben ser detectadas y subsanadas (Morales, 2021).

# GRÁFICO DE DISPERSIÓN

Es un gráfico en el cual se representan las parejas (λ, λ) de las variables observadas. La forma que toma ilustra acerca de la posible asociación existente (ver Figura 14).

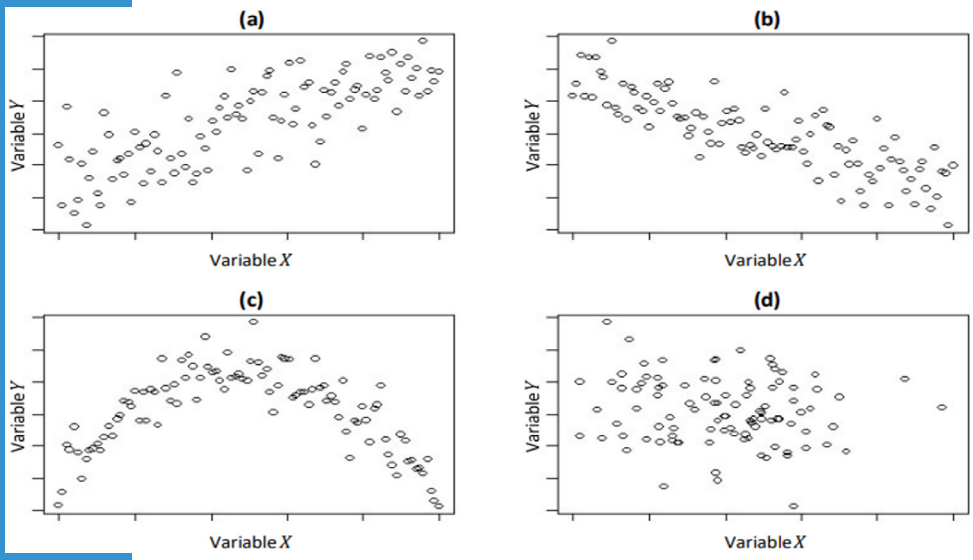



Figura 14. a) Relación lineal positiva (directa o creciente) b) Relación lineal negativa (indirecta o decreciente) c) Relación no lineal d) Falta de relación

En el proceso de elaboración de un modelo de análisis de datos, se torna fundamental llevar a cabo un análisis exhaustivo de los datos que serán empleados en la modelación. Este análisis se divide en dos componentes esenciales: el análisis descriptivo y el análisis inferencial. Ambos desempeñan un papel esencial en la obtención de información valiosa y en la generación de simulaciones precisas.

Tomemos como ejemplo una empresa dedicada a la producción de botellas de agua, la cual opera con múltiples estaciones de trabajo interconectadas. Para simular eficazmente su sistema de producción, es imperativo adentrarse en los detalles de sus procesos. Esto implica la recopilación de datos relevantes, como la frecuencia de llegada de materias primas, los tiempos necesarios para el alistamiento de las máquinas y los tiempos de procesamiento por botella en cada máquina.

Un paso crucial en este proceso es la acumulación de un conjunto de datos lo suficientemente amplio y representativo. En general, se requieren al menos 100 observaciones para que los datos sean estadísticamente significativos y se pueda proceder con la generación de distribuciones de probabilidad. Estos datos sirven como base para crear un modelo que refleje de manera precisa el comportamiento del sistema de producción.

En este contexto, el uso de herramientas como el software R se convierte en una herramienta poderosa. Una vez que se han recolectado los datos suficientes, R permite generar distribuciones de probabilidad que pueden utilizarse para simular el rendimiento del sistema en condiciones diversas. Esto incluye la capacidad de generar datos aleatorios que se ajusten a las distribuciones basadas en el histórico de datos recolectados, lo que proporciona una base sólida para la simulación y el análisis de diferentes escenarios y condiciones de operación.

En resumen, el análisis de datos descriptivo e inferencial es un proceso fundamental en la modelación y simulación de sistemas. La recopilación y el análisis de datos precisos y representativos sientan las bases para generar modelos efectivos y tomar decisiones informadas en la optimización de operaciones, como en el caso de la producción de botellas de agua. La utilización de herramientas como R agiliza el proceso al permitir la generación de datos aleatorios que se ajusten a distribuciones probables, lo que contribuye a la creación de simulaciones realistas y útiles (Morales, 2021).

Para comprender estos conceptos a continuación se presenta un histórico de datos (ver Tabla 3).

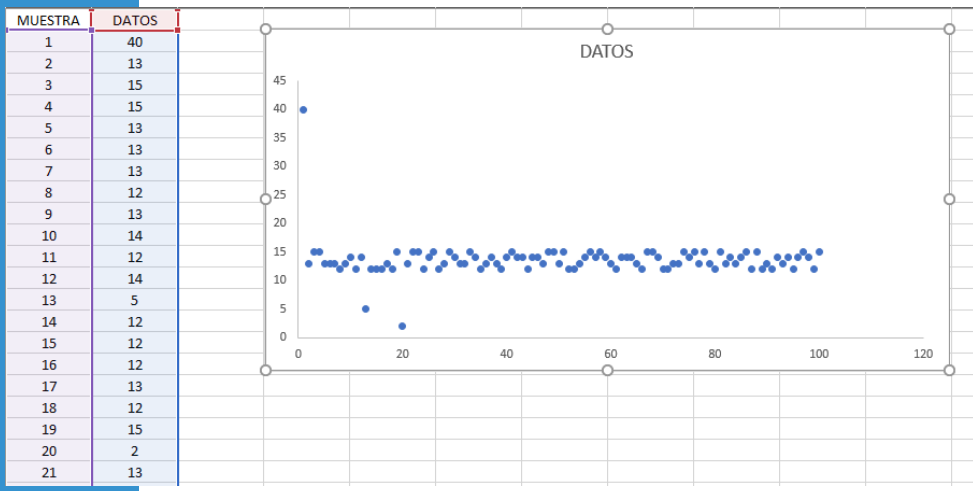
**Tabla 2. Datos históricos**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 14 | 12 | 15 | 13 | 14 | 12 | 14 | 13 | 13 | 13 |
| 13 | 13 | 13 | 14 | 15 | 13 | 15 | 12 | 40 | 14 |
| 15 | 15 | 14 | 15 | 14 | 14 | 15 | 13 | 13 | 12 |
| 12 | 12 | 40 | 14 | 12 | 12 | 15 | 14 | 13 | 13 |
| 14 | 12 | 14 | 12 | 5 | 14 | 15 | 13 | 13 | 14 |
| 15 | 2 | 15 | 14 | 13 | 15 | 14 | 14 | 13 | 12 |
| 15 | 13 | 15 | 12 | 14 | 15 | 12 | 13 | 13 | 14 |
| 15 | 14 | 15 | 12 | 12 | 14 | 13 | 14 | 15 | 14 |
| 15 | 13 | 15 | 15 | 12 | 13 | 13 | 13 | 12 | 15 |
| 15 | 12 | 12 | 14 | 13 | 13 | 14 | 13 | 15 | 15 |

Fuente. Elaboración propia

Para realizar el diagrama de dispersión se procede de la siguiente forma:

Se seleccionan ambos conjuntos de datos sin considerar los títulos. En la pestaña insertar se da clic en el botón Insertar gráfico de dispersión (X, Y) o de burbujas (ubicado en el centro de la parte superior de la pantalla) como se puede ver en la Figura 15.




| MUESTRA | DATOS |
|---|---|
| 1 | 40 |
| 2 | 13 |
| 3 | 15 |
| 4 | 15 |
| 5 | 13 |
| 6 | 13 |
| 7 | 13 |
| 8 | 12 |
| 9 | 13 |
| 10 | 14 |
| 11 | 12 |
| 12 | 14 |
| 13 | 5 |
| 14 | 12 |
| 15 | 12 |
| 16 | 12 |
| 17 | 13 |
| 18 | 12 |
| 19 | 15 |
| 20 | 2 |
| 21 | 13 |

Figura 15. Gráfico de dispersión

El gráfico de dispersión anterior muestra que la mayoría de los datos fueron tomados correctamente, pero se pueden ver tres datos atípico.

## ¿QUÉ SON DATOS ATÍPICOS?

Un valor atípico es una observación que numéricamente es muy distinta al resto de elementos de una muestra (ver Figura 16). Estos datos nos pueden causar problemas en la interpretación de lo que ocurre en un proceso o en una población. Por ejemplo, en el cálculo de la resistencia media a compresión simple de unas probetas de hormigón, la mayoría se encuentran entre 25 y 30 MPa. ¿Qué ocurriría si, de repente, medimos una probeta con una resistencia de 60 MPa? La mediana de los datos puede ser 27 MPa, pero la resistencia media podría llegar a 45 MPa. En este caso, la mediana refleja mejor el valor central de la muestra que la media (Orellana Cordero & Cedillo, 2020).

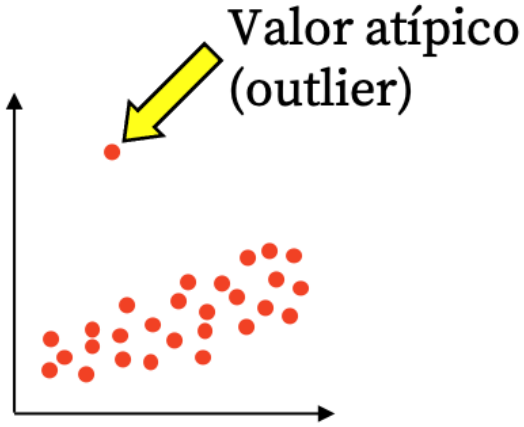



Figura 16. Datos atípicos

**¿Qué hacemos con esos valores atípicos?**

La opción de ignorarlos a veces no es la mejor de las soluciones posibles si pretendemos conocer qué ha pasado con estos valores. Lo cierto es que distorsionan los resultados del análisis, por lo que hay que identificarlos y tratarlos de forma adecuada. A veces se excluyen si son resultado de un error, pero otras veces son datos potencialmente interesantes en la detección de anomalías.

Los valores atípicos pueden deberse a errores en la recolección de datos válidos que muestran un comportamiento diferente, pero reflejan la aleatoriedad de la variable en estudio. Es decir, valores que pueden haber aparecido como parte del proceso, aunque parezcan extraños. Si los valores atípicos son parte del proceso, deben conservarse. En cambio, si ocurren por algún tipo de error, lo adecuado es su eliminación (Orellana Cordero & Cedillo, 2020).

# CAPÍTULO 4

# INTRODUCCIÓN AL ANÁLISIS DE DATOS CON R

En este capítulo, nos embarcaremos en un emocionante viaje para explorar el mundo del análisis de datos utilizando R, un lenguaje de programación y un entorno de software para el análisis estadístico y gráficos. Nuestro objetivo principal será aplicar un modelo de regresión lineal simple, una herramienta poderosa y ampliamente utilizada en estadística que nos permite entender y predecir la relación entre dos variables cuantitativas.

Para ilustrar la aplicación de la regresión lineal simple, nos centraremos en un caso de estudio concreto: las ventas de artículos. Este caso nos permitirá entender cómo ciertas características o factores pueden influir en las ventas de un artículo. A lo largo del capítulo, aprenderemos cómo cargar y manipular datos en R, cómo visualizarlos y, finalmente, cómo ajustar e interpretar nuestro modelo de regresión lineal. Al final del capítulo, seremos capaces de utilizar R para analizar y predecir las ventas de artículos basándonos en los datos disponibles.

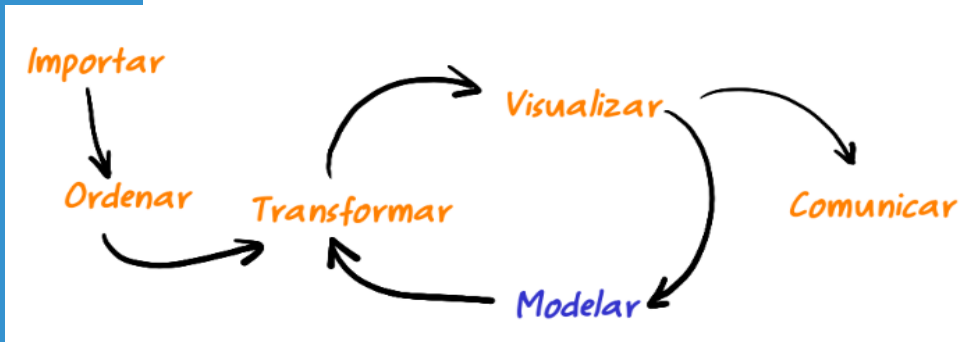



Figura 17. Pasos para realizar un análisis de datos.

Dependiendo el problema a ser abordado, se utiliza una herramienta estadística diferente para el proceso de modelación (ver Figura 17). Dentro de los modelos más utilizados se destacan: el modelo de Regresión Lineal, la Regresión Logística, el Análisis de Componente Principales, el Análisis Discriminante, el Modelo Lineal Generalizado, el Modelo Lineal Mixto Generalizao el modelo de Regresión de Cox y el modelo de series de tiempo.

Una de las aplicaciones de más uso ciencia de datos es el modelo de regresión lineal. Esta técnica permite medir la relación que pueda existir entre una variable llamada variable respuesta y un conjunto de variables independientes.

## CASOS COMO LA ESTIMACIÓN

Estimación de las ventas a partir de los gastos en publicidad. Si en una empresa gasta 10 millones de pesos en publicidad ¿cuánto podrán ser sus ingresos?

Consumo de un producto a partir de su precio. Si se presupuesta vender 500 unidades de un producto, ¿Qué valor debería tener el producto?

Precio de una casa a partir de sus características. ¿Qué valor comercial tiene una casa en particular de acuerdo con sus características?

Relación entre el peso de un bebe y su edad. ¿Qué peso esperado debe tener un bebe con un mes de edad?

El modelo se basa en una relación lineal entre las variables (línea recta) (ver Figura 18)

$$Y = a + bX$$

Donde:
Y= variable dependiente
X= variable independiente
a =intercepto
b = pendiente de la recta

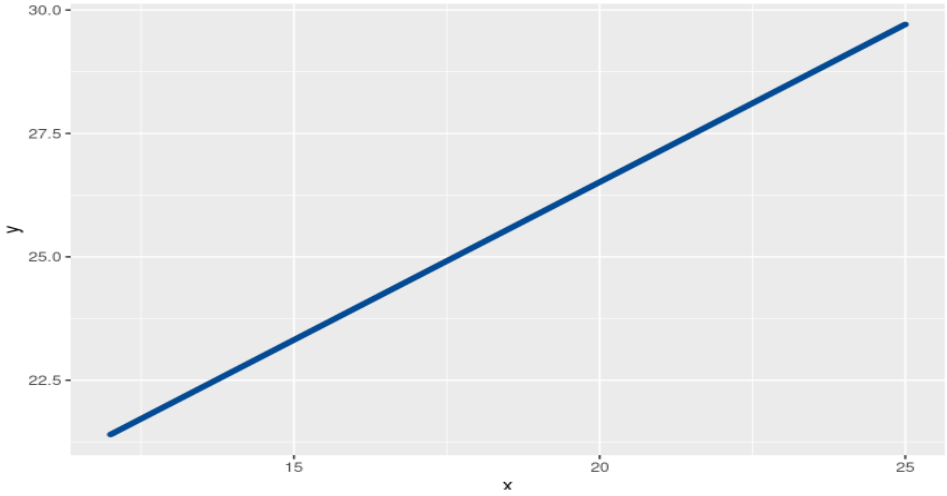



Figura 18. Recta y=13.7+0.64x=13.7+0.64

Cuando la relación no se ajusta de manera perfecta a una línea recta (λ≠0), el problema se centra en la estimación de los coeficientes de la línea recta que mejor se ajuste a datos procedentes de una muestra de las variables X y Y en n parejas (x,y).

Con el propósito de desarrollar el caso de estudio en R de descarga la base de ventas que contiene el *paqueteMOD*

Vendedor: id del vendedor
Ventas: valor de las ventas
Clientes: número de clientes

## Carga de los datos

```
library(paqueteMOD)
library(tidyverse)
data(ventas)
```

## Descripción de las variables

```
summarytools::descr(ventas[,2:3])
```

```
Descriptive Statistics

                   clientes   ventas
----------------- ---------- --------
             Mean      96.00    45.00
          Std.Dev      42.76    12.89
              Min      36.00    28.00
               Q1      72.00    31.00
           Median      84.00    43.00
               Q3     128.00    56.00
              Max     180.00    70.00
              MAD      53.37    17.79
              IQR      50.00    18.50
               CV       0.45     0.29
         Skewness       0.37     0.29
      SE.Skewness       0.58     0.58
         Kurtosis      -0.91    -1.16
          N.Valid      15.00    15.00
        Pct.Valid     100.00   100.00
```

```
library(ggplot2)
library(patchwork)
p1 = ggplot(ventas, aes(x=clientes))+geom_boxplot()
p2 = ggplot(ventas, aes(x=ventas))+ geom_boxplot()
p3 = ggplot(ventas, aes(x=clientes))+geom_density()
p4 = ggplot(ventas, aes(x=ventas))+ geom_density()

(p1 + p2)/(p3 + p4)
```

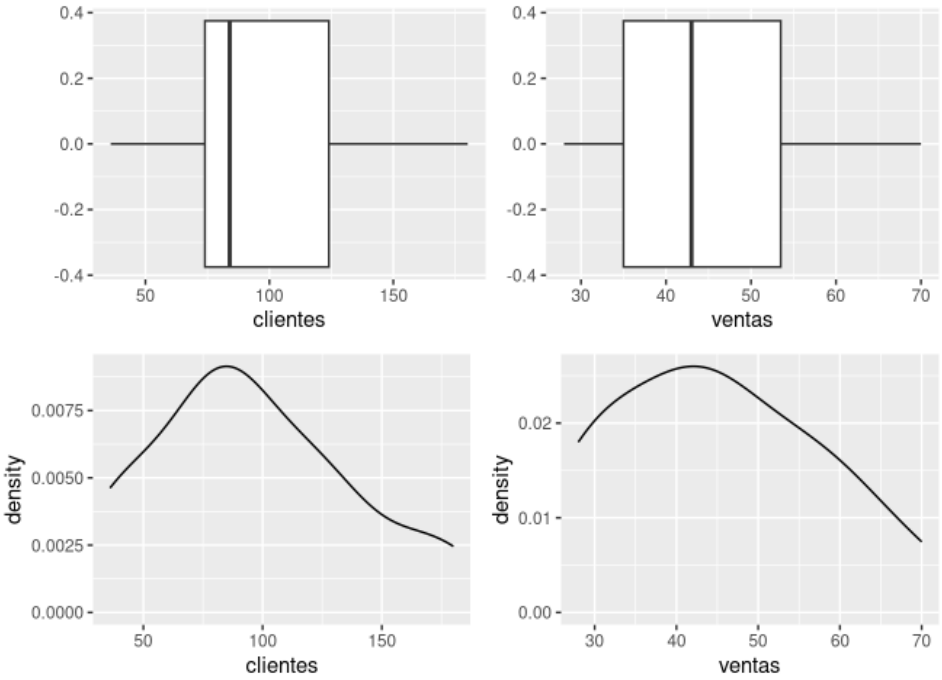



**Diagrama de dispersión**

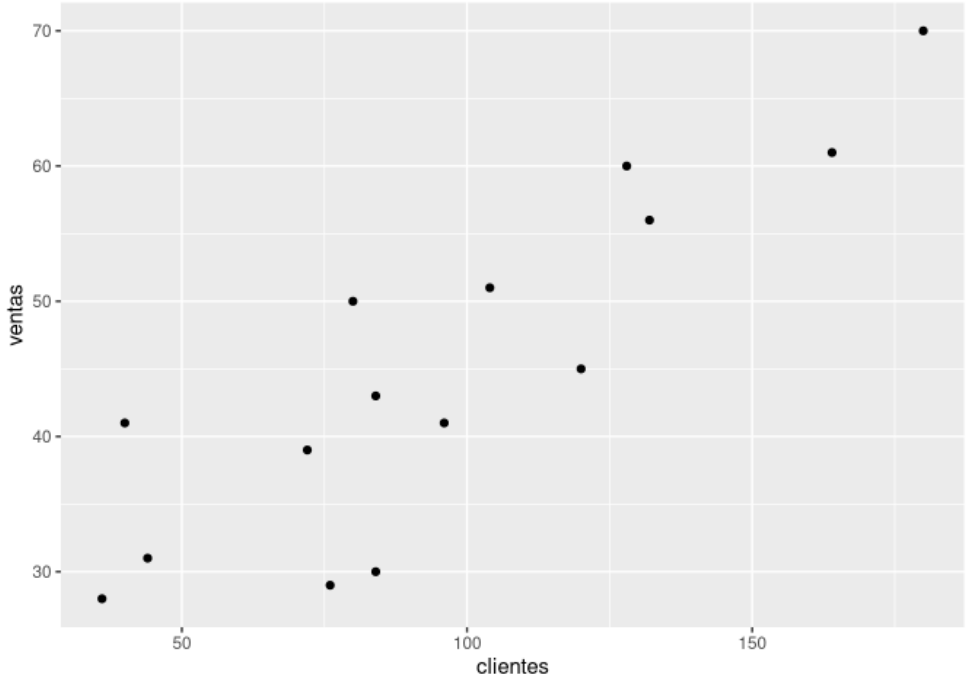
En este diagrama se pretende visualizar la relación existente entre las variables, si se puede ajustar a una línea recta o si por el contrario se debe realizar una transformación.

Se pueden utilizar las funciones:

plot(x,y)

gplot(data, aes (x,y)) + geom_point ()

```
library(ggplot2)
ggplot(ventas, aes(x=clientes, y=ventas))+
  geom_point()
```




## Normalidad de las variables

```
shapiro.test(ventas$clientes)
```

```
    Shapiro-Wilk normality test

data:  ventas$clientes
W = 0.95329, p-value = 0.5777
```

```
shapiro.test(ventas$ventas)
```

```
    Shapiro-Wilk normality test

data:  ventas$ventas
W = 0.94876, p-value = 0.5052
```

### Inferencia sobre la correlación

```
cor.test(ventas$ventas, ventas$clientes)
```

```
    Pearson's product-moment correlation

data:  ventas$ventas and ventas$clientes
t = 6.2051, df = 13, p-value = 3.193e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6325270 0.9542427
sample estimates:
      cor
0.8646318
```

### Estimación del modelo por MCO

Para la estimación del modelo requerimos las funciones:

modelo = lm (fórmula, data, weights): estima el modelo MCO
summary(modelo): visualiza el objeto

```
modelo1=lm(ventas ~ clientes, ventas)
summary(modelo1)
```

```
Call:
lm(formula = ventas ~ clientes, data = ventas)

Residuals:
    Min      1Q  Median      3Q     Max
-11.873  -2.861   0.255   3.511  10.595

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.9800     4.3897   4.552 0.000544 ***
clientes      0.2606     0.0420   6.205 3.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.72 on 13 degrees of freedom
Multiple R-squared:  0.7476,    Adjusted R-squared:  0.7282
F-statistic:  38.5 on 1 and 13 DF,  p-value: 3.193e-05
```

La anterior salida se divide en cuatro partes:

Forma funcional del modelo

```
Call:
lm(formula = ventas ~ clientes, data = ventas)
```

Resumen de los residuales

```
Residuals:
    Min      1Q  Median      3Q     Max
-11.873  -2.861   0.255   3.511  10.595
```

Estimación MCO de los coeficientes λ0 y λ1

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.9800     4.3897   4.552 0.000544 ***
clientes      0.2606     0.0420   6.205 3.19e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

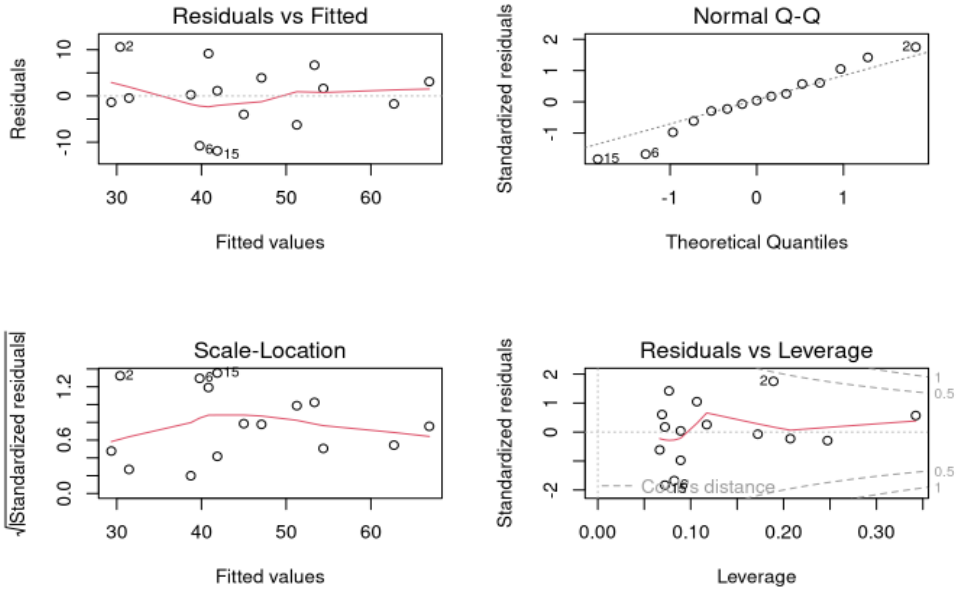
Indicadores de significancia del modelo

```
Residual standard error: 6.72 on 13 degrees of freedom
Multiple R-squared:  0.7476,    Adjusted R-squared:  0.7282
F-statistic:  38.5 on 1 and 13 DF,  p-value: 3.193e-05
```

## Validación de los supuestos sobre los errores

Inicialmente se revisar los gráficos generados con el objeto modelo1

```
par(mfrow = c(2, 2))
plot(modelo1)
```




Ahora utilizando pruebas de hipótesis

Supuesto de normalidad de los errores

```
e= modelo1$residuals

# Ho: los errores tienen distribución normal
shapiro.test(e)
```

```
    Shapiro-Wilk normality test

data:  e
W = 0.96985, p-value = 0.8559
```

## Supuesto de independencia de errores

```
library(lmtest)
# Ho: los errores sin independientes ( no estan relacionados)
dwtest(modelo1)
```

```
    Durbin-Watson test

data:  modelo1
DW = 1.6497, p-value = 0.2226
alternative hypothesis: true autocorrelation is greater than 0
```

## Supuesto de varianza constante

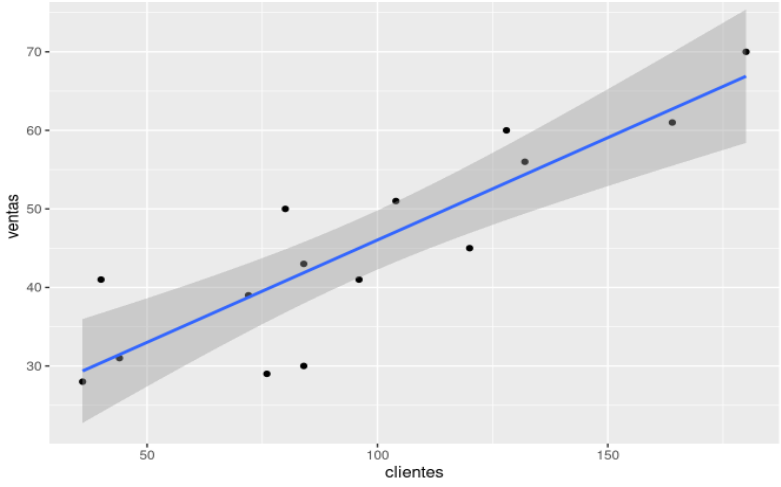
```
# Ho: la varianza de los erroes es constante
gqtest(modelo1)
```

```
    Goldfeld-Quandt test

data:  modelo1
GQ = 0.76554, df1 = 6, df2 = 5, p-value = 0.6276
alternative hypothesis: variance increases from segment 1 to 2
```

## Supuesto de linealidad del modelo

### Representación gráfica del modelo

```
ggplot(ventas, aes(clientes, ventas)) +
            geom_point() +
            geom_smooth(method = "lm",  level = 0.95, formula = y ~ x)
```



## Transformaciones

En caso de requerir transformaciones a las variables del modelo planteado con el fin de mejorar sus indicadores y el cumplimiento loa supuestos.

Modelo log-lin

```
modelo2=lm(log(ventas) ~ clientes, ventas)
summary(modelo2)
```

```
Call:
lm(formula = log(ventas) ~ clientes, data = ventas)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29863 -0.07507  0.00534  0.09009  0.26277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.224418   0.109256  29.512 2.68e-13 ***
clientes    0.005660   0.001045   5.414 0.000118 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1673 on 13 degrees of freedom
Multiple R-squared:  0.6927,    Adjusted R-squared:  0.6691
F-statistic: 29.31 on 1 and 13 DF,  p-value: 0.0001183
```

Modelo lin-log

```
modelo3=lm(ventas ~ log(clientes), ventas)
summary(modelo3)
```

```
Call:
lm(formula = ventas ~ log(clientes), data = ventas)

Residuals:
    Min      1Q  Median      3Q     Max
-14.333  -4.114   1.720   4.551  12.475

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -50.075     19.635  -2.550 0.024178 *
log(clientes)   21.307      4.376   4.869 0.000307 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.96 on 13 degrees of freedom
Multiple R-squared:  0.6458,    Adjusted R-squared:  0.6186
F-statistic: 23.71 on 1 and 13 DF,  p-value: 0.0003066
```

Modelo log-log

```
modelo4=lm(log(ventas) ~ log(clientes), ventas)
summary(modelo4)
```

```
Call:
lm(formula = log(ventas) ~ log(clientes), data = ventas)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35174 -0.06117  0.00826  0.10636  0.31173

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.6562     0.4552   3.638 0.003003 **
log(clientes)   0.4732     0.1015   4.664 0.000443 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1845 on 13 degrees of freedom
Multiple R-squared:  0.626, Adjusted R-squared:  0.5972
F-statistic: 21.76 on 1 and 13 DF,  p-value: 0.0004428
```

## Comparación de modelos

```
library(stargazer)
#stargazer(modelo1, modelo2, modelo3, modelo4,type="html", title="Comparación de modelos")
stargazer(modelo1, modelo2, modelo3, modelo4, type="text", df=FALSE)
```

```
================================================================
                                Dependent variable:
                    ---------------------------------------------
                     ventas   log(ventas)  ventas   log(ventas)
                       (1)        (2)        (3)        (4)
----------------------------------------------------------------
clientes            0.261***   0.006***
                    (0.042)    (0.001)

log(clientes)                             21.307***  0.473***
                                           (4.376)    (0.101)

Constant            19.980***  3.224***   -50.075**  1.656***
                    (4.390)    (0.109)    (19.635)   (0.455)

----------------------------------------------------------------
Observations           15         15         15         15
R2                   0.748      0.693      0.646      0.626
Adjusted R2          0.728      0.669      0.619      0.597
Residual Std. Error  6.720      0.167      7.960      0.185
F Statistic        38.503***  29.310***  23.706***  21.756***
================================================================
Note:                                *p<0.1; **p<0.05; ***p<0.01
```

Al comparar los valores de R2, el modelo1 presenta el mayor porcentaje de explicación de la variabilidad de Y

### Predicción de Y0

```
predict(modelo1, data.frame(clientes=20), interval = "prediction", level = 0.95)
```

```
      fit      lwr      upr
1 25.1925 8.688234 41.69677
```

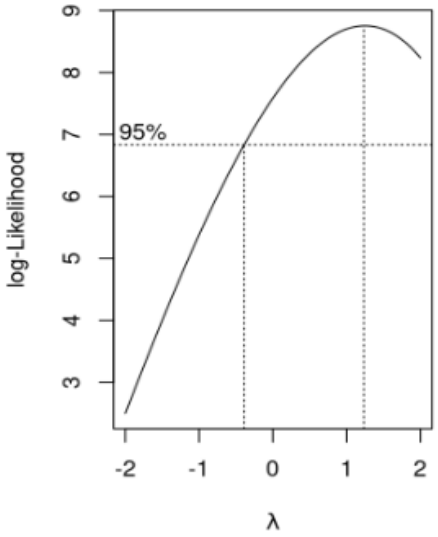
### Evaluación de la linealidad del modelo

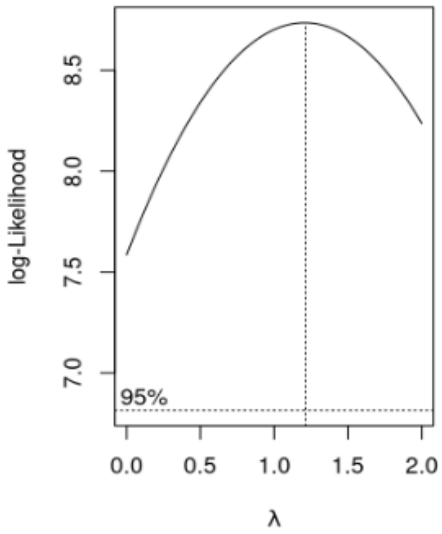
Para ello utilizamos la función boxcox() del paquete MASS

```
library(paqueteMOD)
data(ventas)
modelo1=lm(ventas ~ clientes, ventas)
summary(modelo1)
```

```

Call:
lm(formula = ventas ~ clientes, data = ventas)

Residuals:
    Min      1Q  Median      3Q     Max
-11.873  -2.861   0.255   3.511  10.595

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.9800     4.3897   4.552 0.000544 ***
clientes      0.2606     0.0420   6.205 3.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.72 on 13 degrees of freedom
Multiple R-squared:  0.7476,    Adjusted R-squared:  0.7282
F-statistic:  38.5 on 1 and 13 DF,  p-value: 3.193e-05
```

```
par(mfrow = c(1,2))
boxcox(lm(ventas$ventas ~ ventas$clientes), lambda = -2:2)
#Se repite el proceso pero esta vez estrechando el rango de valores de lamb
bc<-boxcox(lm(ventas$ventas ~ ventas$clientes), lambda = 0:2)
```




```
(lambda <- bc$x[which.max(bc$y)])
```

```
[1] 1.212121
```

El valor de $= 1.212121 $ indica que la variable dependiente debe continuar en la forma lineal Y1

El siguiente código en R te permitirá acceder a un tutorial que te permitirá revisar lo aprendido:

```
install.packages("learnr")                    # solo una vez
install.packages("devtools")                  # solo una vez
devtools::install_github("dgonxalex80/paqueteMOD")   #descarga paquete
learnr::run_tutorial("Tutorial101", "paqueteMOD")  # carga Tutorial101
```

Después de correr estos códigos te debe aparecer:

# CAPÍTULO 5

# CASO DE APLICACIÓN CON R

## ANÁLISIS DE LOS PRECIOS DE LAS VIVIENDAS

En un entorno en constante cambio, el mercado inmobiliario es un sector crítico que influye en la vida de muchas personas. La adquisición o venta de una vivienda es una de las decisiones financieras más significativas que una persona puede tomar en su vida. Por lo tanto, es esencial comprender y evaluar adecuadamente los factores que influyen en los precios de las viviendas. En este contexto, se llevará a cabo un análisis exhaustivo de los precios de las viviendas con el objetivo de proporcionar información valiosa y perspicacia sobre este mercado en constante evolución.

Exploración de la base de datos[1]

```
library(readxl)
datos <- read_excel("C:/Users/Hp/Downloads/datos_vivienda1 (1).xlsx")
View(datos)
```

1 La información aquí suministrada se encuentra encuentra en repositorio del autor, en https://rpubs.com/jhonsegura25/950142

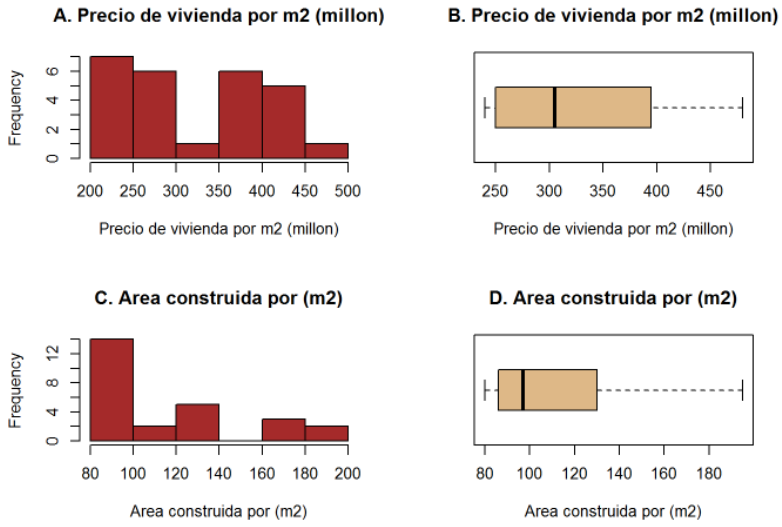
Se realiza un análisis exploratorio de las variables precio de vivienda (millones de pesos COP) y área de la vivienda (metros cuadrados). Incluir gráficos e indicadores apropiados interpretado.

```
library(ggplot2)
library(ggpubr)
```

```
attach(datos)
par(mfrow=c(2,2))
summary(datos)
```

```
##  Area_contruida  precio_millon
##  Min.   : 80.0   Min.   :240.0
##  1st Qu.: 86.0   1st Qu.:251.2
##  Median : 97.0   Median :305.0
##  Mean   :115.7   Mean   :332.1
##  3rd Qu.:130.0   3rd Qu.:395.0
##  Max.   :195.0   Max.   :480.0
```

```
hist(precio_millon, breaks = 6, main =  "A. Precio de vivienda por m2 (millon)", xlab =  "Precio de vivienda por m2 (millo
n)", col = "#A52A2A")

boxplot(precio_millon, main =  "B. Precio de vivienda por m2 (millon)", horizontal = TRUE, xlab =  "Precio de vivienda por m
2 (millon)", col="#DEB887")


hist(Area_contruida, breaks = 6, main = "C. Area construida por (m2)", xlab =  "Area construida por (m2)",col = "#A52A2A" )

boxplot(Area_contruida, main = "D. Area construida por (m2)",  horizontal = TRUE, xlab =  "Area construida por (m2)",col =
"#DEB887")
```



Los resultados arrojados en las gráficas, el Precio de vivienda por $m^2$ (millón) y área construida por $m^2$ (A y B) se identifica que el precio tiene una concentración entre 250 a 400 millones de pesos colombianos, Además en el gráfico A se evidencia una asimetría positiva, concluyendo que son muy pocos los apartamentos que cuestan 400 millones de pesos o más. Por otro lado, el precio de las viviendas en promedio es de 332.1 millones de pesos y la media corresponde a 305 millones de pesos, esto nos permite observar una afectación en el costo promedio de las viviendas, ya que se tienen valores en los extremos del precio de vivienda por $m^2$ (millón), esto se puede validar con el valor de la desviación estándar que es de $ 82 millones de pesos. Finalmente, en la variable área construida, el promedio de $m^2$ es de 115.7$m^2$, además se identifica una concentración entre 86 $m^2$ a 130 $m^2$. La distribución del área de las viviendas muestra una asimetría positiva bastante pronunciada. En el gráfico de caja, el 50% de las viviendas tienen 86 a 97 $m^2$, el porcentaje restante se encuentran en 97$m^2$ y 130 $m^2$.

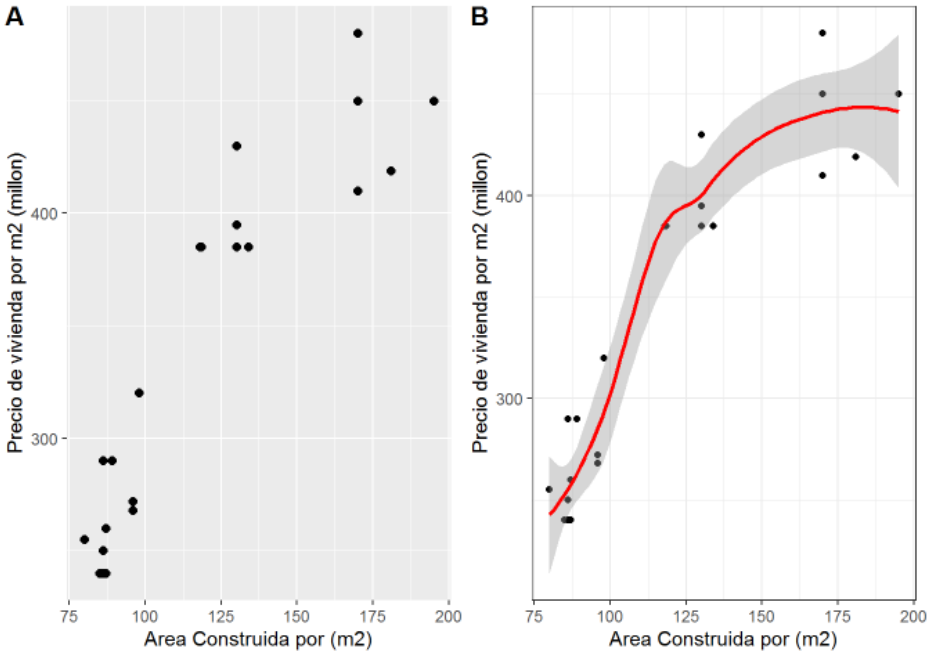
Realice un análisis exploratorio bivariado de datos enfocado en la relación entre la variable respuesta (y = precio) en función de la variable predictora (x = área) - incluir gráficos e indicadores apropiados interpretados.

```
attach(datos)
```

```
## The following objects are masked from datos (pos = 3):
##
##     Area_contruida, precio_millon
```

```
g1= ggplot(datos, aes(Area_contruida, precio_millon)) +geom_point(size =2, shape = 16) + ylab("Precio de vivienda por m2 (millon)") + xlab("Area Construida por (m2)")

g2=ggplot(datos,aes(y=precio_millon,x=Area_contruida))+geom_point()+theme_bw()+geom_smooth(col="red")+ ylab("Precio de vivienda por m2 (millon)") + xlab("Area Construida por (m2)")

ggarrange(g1, g2, labels = c("A", "B"))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
cor.test(Area_contruida, precio_millon)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Area_contruida and precio_millon
## t = 11.422, df = 24, p-value = 3.45e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8255707 0.9634168
## sample estimates:
##       cor
## 0.9190295
```

Como conclusión del análisis bivariado, se identifica que existe una relación directa entre las variables área construida y precio de las viviendas, es decir que si aumenta una de las variables la otra aumenta. Además, se aplica una prueba de correlación, la cual nos arroja un coeficiente r = 0.9190295, afirmando que hay una correlación fuerte y positiva entre las dos variables.

Estime el modelo de regresión lineal simple entre λλλλλλ = λ(λλλλ) + λ. Los coeficientes del modelo λ0, λ1 en caso de ser correcto.

```
mod=lm(precio_millon ~ Area_contruida)
summary(mod)
```

```
##
## Call:
## lm(formula = precio_millon ~ Area_contruida)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.673 -25.612  -6.085  24.875  67.650
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      86.234     22.479   3.836 0.000796 ***
## Area_contruida    2.124      0.186  11.422 3.45e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.05 on 24 degrees of freedom
## Multiple R-squared:  0.8446, Adjusted R-squared:  0.8381
## F-statistic: 130.5 on 1 and 24 DF,  p-value: 3.45e-11
```

El modelo queda de la siguiente manera: Y = 86.234 + 2.124 (Área_construida ($m^2$))

Como resultado del$\lambda 0$ = 86.23. Nos indica que el precio base de una vivienda es de 86.234 millones de pesos, es decir cuando el área de construcción en $m^2$ es 0.

Como resultado del$\lambda 1$ = 2.124. Esto nos indica que por cada metro adicional se espera un aumento en el precio de la vivienda de 2.124 millones.

Construir un intervalo de confianza (95%) para el coeficiente $\lambda 1$, interpretar y concluir si el coeficiente es igual a cero o no. Compare este resultado con una prueba de hipótesis t.

```
confint(mod)
```

```
##                   2.5 %     97.5 %
## (Intercept)    39.83983 132.627917
## Area_contruida  1.74017   2.507771
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = precio_millon ~ Area_contruida)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.673 -25.612  -6.085  24.875  67.650
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      86.234     22.479   3.836 0.000796 ***
## Area_contruida    2.124      0.186  11.422 3.45e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.05 on 24 degrees of freedom
## Multiple R-squared:  0.8446, Adjusted R-squared:  0.8381
## F-statistic: 130.5 on 1 and 24 DF,  p-value: 3.45e-11
```

Como resultado se tiene un intervalo de confianza para el valor λ1 que es de 1.74017 millones a 2.507771 millones, concluyendo que no contiene el valor de 0. Por lo tanto, Con el 95% de confianza se rechazar la hipótesis nula, estableciendo que la variable área construida es significativa y el precio depende de esta variable. El resultado de la prueba T como se muestra en el resumen es menor al 5%, rechazando la hipótesis nula, la cual es que el parámetro λ1 es igual a 0.

Calcule e interprete el indicador de bondad y ajuste R2

```
summary(mod)$adj.r.squared
```

```
## [1] 0.8381408
```

Como resultado del indicador de bondad y ajuste R2, se concluye que el modelo tiene un asertividad del 83.81% para determinar con exactitud el precio de las viviendas, este es un modelo confiable y tiene un buen desempeño para realizar predicciones, pero se debe tener en cuenta el valor del error que es del 16.19%.

¿Cuál sería el precio promedio estimado para un apartamento de 110 metros cuadrados? ¿Considera entonces con este resultado que un apartamento en la misma zona

con 110 metros cuadrados en un precio de 200 millones sería una buena oferta? ¿Qué consideraciones adicionales se deben tener?

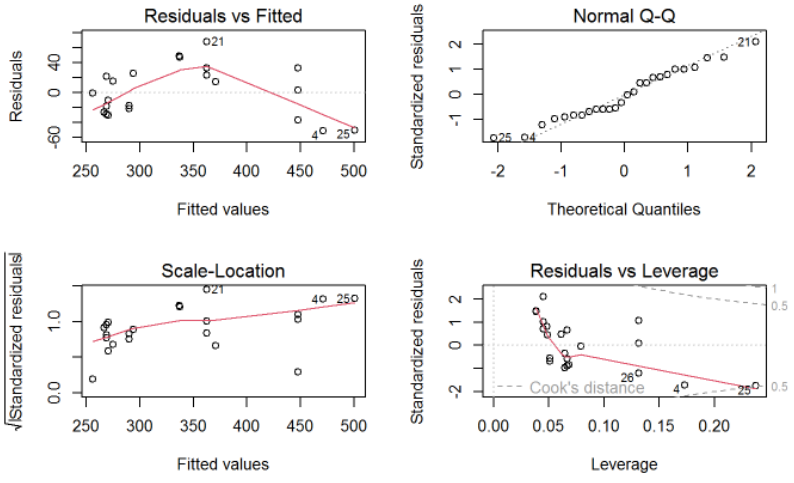
```
predict(mod,newdata = list(Area_contruida=110))
```

```
##        1
## 319.8706
```

Con el modelo determinado, se estima que el precio de un apartamento con un área de 110 $m^2$ tendría un costo de 319 millones de pesos colombianos. Con un nivel de confianza del 95%. Con este resultado se puede decir que es una buena oferta un apartamento de 110 metros cuadrados con un precio de 200 millones de pesos. Pero se deben tener en cuenta que este precio está sujeto a variables diferentes variables como estrato, número de habitaciones, entre otras. Por otro lado, se debe considerar el porcentaje de asertividad del modelo el cual es de 83.81%, lo cual puede afectar la predicción del precio de los apartamentos. Es decir, que el precio arrojado por el modelo para un apartamento de 100 $m^2$ puede estar sobreestimando.

Realice la validación de supuestos del modelo por medio de gráficos apropiados, interpretarlos y sugerir posibles soluciones si se violan algunos de ellos.

```
par(mfrow=c(2,2))
plot(mod)
```



- Errores aleatorios con media 0: se observa que, aunque los errores están centrados en 0, reflejan que no son aleatorios, ya que se aprecia claramente una forma parabólica con concavidad hacia abajo en la distribución de los errores. Por tanto, no se cumple la hipótesis de aleatoriedad.
- varianza constante: en consecuencia, la variabilidad de los errores no es constante; por otra parte, para los valores bajos de los precios tienen menos variabilidad que los valores altos. Por lo tanto, no se cumple la hipótesis de la varianza constante.
- Independencia: en este caso no es necesario revisar, ya que los datos son transversales, es decir, los datos no varían con el tiempo.
- Normalidad: según el gráfico Q-Q normal, los puntos están bastante alineados con la línea recta, sin embargo, se observa que algunos datos de los extremos se alejan de la línea de normalidad.

De acuerdo con el análisis anterior, se refleja un incumplimiento en algunos supuestos, por lo tanto, se decide realizar una transformación a las variables, esto se apoya en el valor de R2 que puede ser mejorado, esto permitirá optimizar el modelo de regresión lineal y así tener un mejor desempeño en la validación de los supuestos de los errores.

De ser necesario realice una transformación apropiada para mejorar el ajuste y supuestos del modelo.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

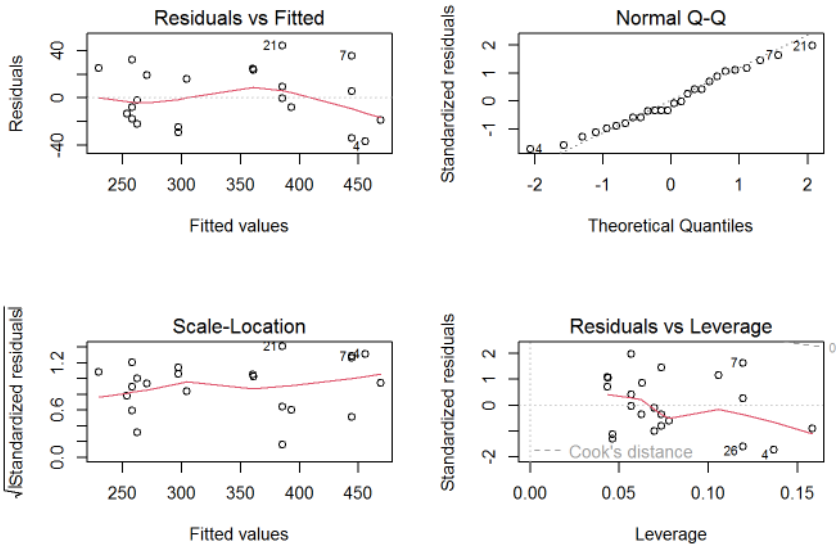
```
datos_hip <- mutate(datos, Area_Hip= 1/Area_contruida)
head(datos_hip)
```

```
## # A tibble: 6 x 3
##   Area_contruida precio_millon Area_Hip
##            <dbl>         <dbl>    <dbl>
## 1             86           250  0.0116
## 2            118           385  0.00847
## 3            130           395  0.00769
## 4            181           419  0.00552
## 5             86           240  0.0116
## 6             98           320  0.0102
```

```
md2 <- lm(datos_hip$precio_millon ~ datos_hip$Area_Hip)
summary(md2)
```

```
##
## Call:
## lm(formula = datos_hip$precio_millon ~ datos_hip$Area_Hip)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.987 -16.743  -5.023  18.547  44.379
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           635.35      18.27   34.77  < 2e-16 ***
## datos_hip$Area_Hip -32464.72    1895.32  -17.13 5.84e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.05 on 24 degrees of freedom
## Multiple R-squared:  0.9244, Adjusted R-squared:  0.9212
## F-statistic: 293.4 on 1 and 24 DF,  p-value: 5.839e-15
```

```
par(mfrow=c(2,2))
plot(md2)
```




**Planteamiento del nuevo modelo**

Y = 635.35 - 32464.72(Área construida)

Utilizando la transformación logarítmica, volvimos a validar los datos. Como resultado obtuvimos una mejora significativa, en el gráfico de los residuos frente a los predichos, los errores están centrados en 0 y no presentan ningún patrón, esto significa que hay aleatoriedad. Por lo tanto, podemos concluir que con la transformación logarítmica se cumplen los supuestos de aleatoriedad en los residuos, media 0, varianza constante, normalidad e independencia.

De ser necesario compare el ajuste y supuestos del modelo inicial y el transformado.

Con la transformación logarítmica que se aplica al modelo, se obtuvo como resultado que el R2 pasó de 0.8446 a 0.9244 por lo que se obtuvo una mejora del 9%

# ANÁLISIS DE ROTACIÓN DE LOS EMPLEADOS DE UNA EMPRESA

El análisis que a continuación se detalla se refiere a una empresa dentro del sector productivo que enfrenta un de-

safío significativo en términos de alta rotación de su personal. La constante entrada y salida de empleados ha generado preocupación y, en última instancia, ha llevado a la empresa a tomar medidas para abordar este problema.

Para abordar la problemática de la alta rotación, la empresa ha decidido llevar a cabo un análisis en profundidad utilizando una base de datos que ha acumulado a lo largo del tiempo. Esta base de datos contiene información valiosa sobre sus empleados, incluyendo detalles relacionados con la rotación, como salarios, carga de trabajo, edades y otros factores relevantes.

El primer paso en este análisis involucra la exploración minuciosa de la base de datos para identificar patrones y tendencias. Se lleva a cabo un análisis estadístico exhaustivo que permite comprender mejor la dinámica detrás de la rotación de personal. Se buscan correlaciones y relaciones entre las variables para discernir qué factores pueden estar contribuyendo de manera significativa a la alta rotación

Una vez identificados los factores clave, se procede a realizar una evaluación más detallada de su influencia en la rotación. Esto implica examinar cómo el salario, la carga de trabajo, la edad y otros factores interaccionan y contribuyen al problema de la rotación. Esta etapa del análisis se lleva a cabo con el propósito de adquirir una comprensión más profunda y precisa de cómo estos elementos se relacionan con la retención del personal.

Con base en los hallazgos y conclusiones extraídos de este análisis, la empresa está en condiciones de proponer y desarrollar estrategias efectivas para reducir la rotación de personal. Estas estrategias pueden abarcar desde ajustes en la estructura salarial, la redistribución de la carga de trabajo, programas de desarrollo profesional para empleados de diferentes edades, y otras iniciativas que aborden los factores identificados como contribuyentes a la rotación.

En resumen, el análisis completo de la base de datos y la evaluación de los factores relacionados con la rotación son pasos esenciales para abordar un problema crítico en la empresa. A través de este enfoque, se busca mejorar la retención del personal y, en última instancia, fortalecer la estabilidad y el desempeño general de la organización en el competitivo entorno del sector productivo. Este análisis representa un compromiso por parte de la empresa para mantener y cuidar a su valioso recurso humano.

## Descripción de la base de datos

Se cargan los datos para ser visualizados, identificando la cantidad de variables que fueron abordadas en el estudio.

```
library(readxl)
Datos = read_excel("E:/MAESTRIA JAVERIANA/SEMESTRE 1/Métodos y simulación/Actividad 1/Datos_Rotacion.xlsx")
names(Datos)
```

```
##  [1] "Rotación"                  "Edad"
##  [3] "Viaje de Negocios"         "Departamento"
##  [5] "Distancia_Casa"            "Educación"
##  [7] "Campo_Educación"           "Satisfacción_Ambiental"
##  [9] "Genero"                    "Cargo"
## [11] "Satisfación_Laboral"       "Estado_Civil"
## [13] "Ingreso_Mensual"           "Trabajos_Anteriores"
## [15] "Horas_Extra"               "Porcentaje_aumento_salarial"
## [17] "Rendimiento_Laboral"       "Años_Experiencia"
## [19] "Capacitaciones"            "Equilibrio_Trabajo_Vida"
## [21] "Antigüedad"                "Antigüedad_Cargo"
## [23] "Años_ultima_promoción"     "Años_acargo_con_mismo_jefe"
```

## Identificación de variables categóricas y variables cuantitativas

En la base de datos se visualizó un total de 8 variables categóricas y 16 variables cuantitativas, esto se realiza para analizar la relación entre las variables y que tipo de relación se espera.

```
str(Datos)
```

```
## tibble [1,470 x 24] (S3: tbl_df/tbl/data.frame)
##  $ Rotación                  : chr [1:1470] "Si" "No" "Si" "No" ...
##  $ Edad                      : num [1:1470] 41 49 37 33 27 32 59 30 38 36 ...
##  $ Viaje de Negocios         : chr [1:1470] "Raramente" "Frecuentemente" "Raramente" "Frecuentemente" ...
##  $ Departamento              : chr [1:1470] "Ventas" "IyD" "IyD" "IyD" ...
##  $ Distancia_Casa            : num [1:1470] 1 8 2 3 2 2 3 24 23 27 ...
##  $ Educación                 : num [1:1470] 2 1 2 4 1 2 3 1 3 3 ...
##  $ Campo_Educación           : chr [1:1470] "Ciencias" "Ciencias" "Otra" "Ciencias" ...
##  $ Satisfacción_Ambiental    : num [1:1470] 2 3 4 4 1 4 3 4 4 3 ...
##  $ Genero                    : chr [1:1470] "F" "M" "M" "F" ...
##  $ Cargo                     : chr [1:1470] "Ejecutivo_Ventas" "Investigador_Cientifico" "Tecnico_Laboratorio" "Investig
ador_Cientifico" ...
##  $ Satisfación_Laboral       : num [1:1470] 4 2 3 3 2 4 1 3 3 3 ...
##  $ Estado_Civil              : chr [1:1470] "Soltero" "Casado" "Soltero" "Casado" ...
##  $ Ingreso_Mensual           : num [1:1470] 5993 5130 2090 2909 3468 ...
##  $ Trabajos_Anteriores       : num [1:1470] 8 1 6 1 9 0 4 1 0 6 ...
##  $ Horas_Extra               : chr [1:1470] "Si" "No" "Si" "Si" ...
##  $ Porcentaje_aumento_salarial: num [1:1470] 11 23 15 11 12 13 20 22 21 13 ...
##  $ Rendimiento_Laboral       : num [1:1470] 3 4 3 3 3 3 4 4 4 3 ...
##  $ Años_Experiencia          : num [1:1470] 8 10 7 8 6 8 12 1 10 17 ...
##  $ Capacitaciones            : num [1:1470] 0 3 3 3 3 2 3 2 2 3 ...
##  $ Equilibrio_Trabajo_Vida   : num [1:1470] 1 3 3 3 3 2 2 3 3 2 ...
##  $ Antigüedad                : num [1:1470] 6 10 0 8 2 7 1 1 9 7 ...
##  $ Antigüedad_Cargo          : num [1:1470] 4 7 0 7 2 7 0 0 7 7 ...
##  $ Años_ultima_promoción     : num [1:1470] 0 1 0 3 2 3 0 0 1 7 ...
##  $ Años_acargo_con_mismo_jefe : num [1:1470] 5 7 0 0 2 6 0 0 8 7 ...
```

Seleccionar 3 variables categóricas (distintas de rotación) y 3 variables cuantitativas, que se consideran que están relacionadas con la rotación, estas son:

**Variables categóricas:**
Horas extras Viaje de Negocios Genero

**Variables cuantitativas:**
Edad Años de experiencia Satisfacción laboral

Con las variables expuestas anteriormente se plantean las siguientes hipótesis

**Hipótesis planteadas para el estudio:**

Se espera que las horas extra se relacionen con la rotación, ya que las personas podrían tener una sobrecarga laboral y degradación de la salud. La hipótesis es que las personas que trabajan horas extra tienen mayor posibilidad de rotar que las que no trabajan extra.

Se espera que la frecuencia por viaje de negocios se relacione con la rotación, esto a que descuidan aspectos familiares y perjudica gravemente la salud. La hipótesis es que las personas que viajan por negocios con mayor frecuencia tienen una alta posibilidad de rotar que las que no viajan con frecuencia.

Se espera que el tipo de género se relacione con la alta rotación, esto a que la carga de trabajo no es asignada equitativamente, dejando a la mujer trabajos de fuerza. La hipótesis es que las mujeres tienen mayor frecuencia de rotar en la empresa que las que los hombres.

Se espera que la edad se relacione con la rotación en la empresa, ya que los de mayor edad tienen un mejor balance entre vida y trabajo. La hipótesis es que las personas de mayor edad tienen una mayor posibilidad de rotar que los más jóvenes.

Se espera que los años de experiencia se relacionen con la rotación, ya que las personas pueden acceder a mejores salarios y una mejor calidad de vida. La hipótesis es que las personas que tiene mayores años de experiencia tienen mayor posibilidad de rotar que las que están apenas iniciando en el campo laboral.

Se espera que la satisfacción laboral de las personas se relacione con la rotación en la empresa, ya que no se

cuenta con un sistema de incentivos y se presentan malas condiciones laborales. La hipótesis es que las personas que tienen una baja satisfacción laboral tienen mayor posibilidad de rotar que las que tiene un nivel alto.

### Análisis univariado

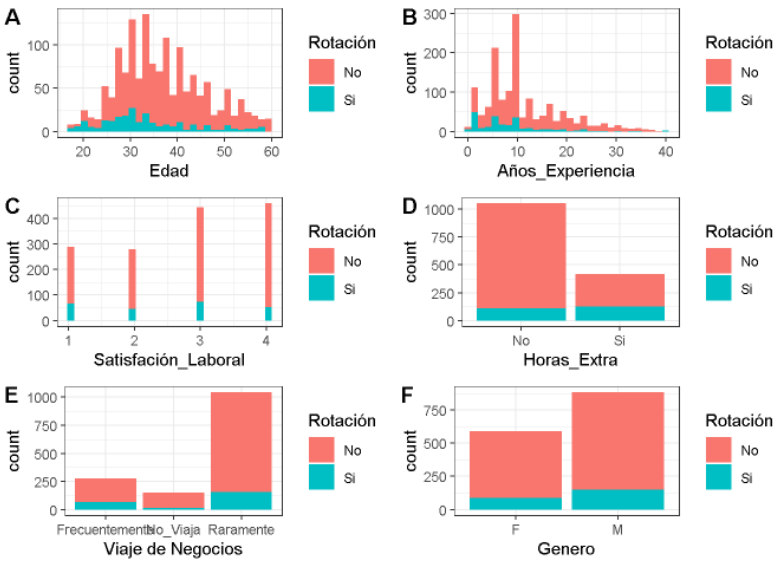
Se presentan los resultados que se obtienen al hacer el análisis estadístico univariado, mediante el cual se determinan las principales estadísticas descriptivas de cada una de las variables.

```
library(ggplot2)
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.1.3
```

```
g1=ggplot(Datos,aes(x=Edad, fill=Rotación))+geom_histogram()+theme_bw()
g2=ggplot(Datos,aes(x=Años_Experiencia, fill=Rotación))+geom_histogram()+theme_bw()
g3=ggplot(Datos,aes(x=Satisfación_Laboral, fill=Rotación))+geom_histogram()+theme_bw()
g4=ggplot(Datos,aes(x=Horas_Extra, fill=Rotación))+geom_bar()+theme_bw()
g5=ggplot(Datos,aes(x=`Viaje de Negocios`, fill=Rotación))+geom_bar()+theme_bw()
g6=ggplot(Datos,aes(x=Genero, fill=Rotación))+geom_bar()+theme_bw()
ggarrange(g1,g2,g3,g4,g5,g6,labels = c("A", "B","C", "D","E", "F"),ncol = 2, nrow = 3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
Datos$Edad_grupo=cut(Datos$Edad,breaks = c(0,30,40,50,60))
require(table1)
```

```
## Loading required package: table1
```

```
## Warning: package 'table1' was built under R version 4.1.3
```

```
##
## Attaching package: 'table1'
```

```
## The following objects are masked from 'package:base':
##
##     units, units<-
```

```
y <- table1::table1(~ Edad+Años_Experiencia+Satisfación_Laboral+
Horas_Extra+`Viaje de Negocios`+Genero | Rotación, data = Datos)
y
```

| | No (N=1233) | Si (N=237) | Overall (N=1470) |
|---|---|---|---|
| **Edad** | | | |
| Mean (SD) | 37.6 (8.89) | 33.6 (9.69) | 36.9 (9.14) |
| Median [Min, Max] | 36.0 [18.0, 60.0] | 32.0 [18.0, 58.0] | 36.0 [18.0, 60.0] |
| **Años_Experiencia** | | | |
| Mean (SD) | 11.9 (7.76) | 8.24 (7.17) | 11.3 (7.78) |
| Median [Min, Max] | 10.0 [0, 38.0] | 7.00 [0, 40.0] | 10.0 [0, 40.0] |
| **Satisfación_Laboral** | | | |
| Mean (SD) | 2.78 (1.09) | 2.47 (1.12) | 2.73 (1.10) |
| Median [Min, Max] | 3.00 [1.00, 4.00] | 3.00 [1.00, 4.00] | 3.00 [1.00, 4.00] |
| **Horas_Extra** | | | |
| No | 944 (76.6%) | 110 (46.4%) | 1054 (71.7%) |
| Si | 289 (23.4%) | 127 (53.6%) | 416 (28.3%) |
| **Viaje de Negocios** | | | |
| Frecuentemente | 208 (16.9%) | 69 (29.1%) | 277 (18.8%) |
| No_Viaja | 138 (11.2%) | 12 (5.1%) | 150 (10.2%) |
| Raramente | 887 (71.9%) | 156 (65.8%) | 1043 (71.0%) |
| **Genero** | | | |
| F | 501 (40.6%) | 87 (36.7%) | 588 (40.0%) |
| M | 732 (59.4%) | 150 (63.3%) | 882 (60.0%) |

Se identifica que el 40% del personal de la empresa son mujeres y el 60% son hombres para un total de 1.470, donde el 18.8% del personal que trabaja en la empresa tiene viajes frecuentes, un 10.2% no viajan y el 71% viaja raramente. Por otro lado, el 71.7% del personal no trabajan horas extras, solamente el 28.3% si trabaja horas extras. El resultado de la satisfacción laboral en general se tiene una media de 3 y una dispersión media de los datos de 2.73, así mismo los años de experiencia se tiene una media de 10 y una dispersión media de los datos 11.3, finamente en la edad se tiene una media de 36 años, además se presenta una concentración en el rango de 25 a 40 años.

Con las gráficas y los datos de la tabla, podemos analizar que en el análisis univariado las variables horas extras y viaje de negocios inciden en la decisión de los empleados para presentar su renuncia.

## Análisis bivariado

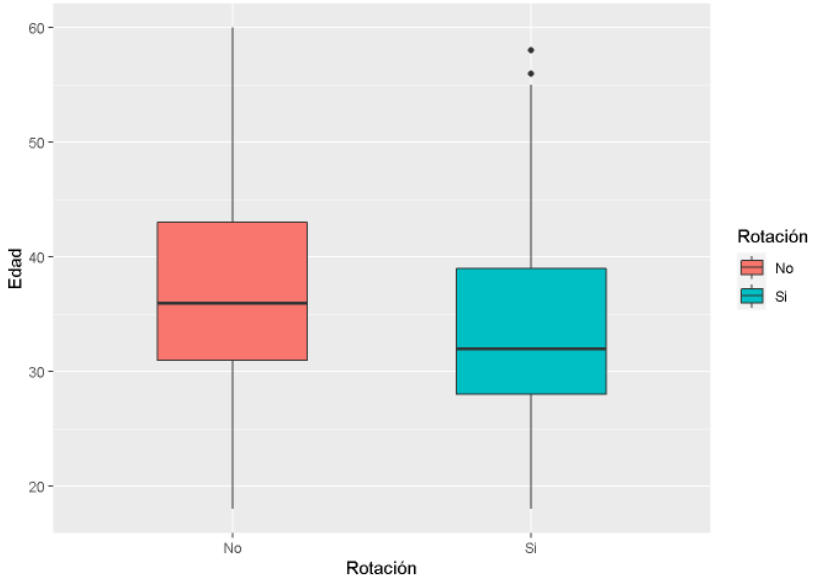
### Hipótesis 1:

Se espera que la edad se relacione con la rotación en la empresa, ya que los de mayor edad tienen un mejor balance entre vida y trabajo. La hipótesis es que las personas de mayor edad tienen una mayor posibilidad de rotar que los más jóvenes.

```
t.test(Datos$Edad~Datos$Rotación)
```

```
##
##  Welch Two Sample t-test
##
## data:  Datos$Edad by Datos$Rotación
## t = 5.828, df = 316.93, p-value = 1.38e-08
## alternative hypothesis: true difference in means between group No and group Si is not equal to 0
## 95 percent confidence interval:
##  2.618930 5.288346
## sample estimates:
## mean in group No mean in group Si
##         37.56123         33.60759
```

```
ggplot(Datos,aes(Rotación, Edad,fill= Rotación))+geom_boxplot(width=0.5)
```




Dado que este valor p es inferior a 0,05, podemos rechazar la hipótesis nula y concluir que existe una diferencia estadísticamente significativa entre la edad y la rotación de los empleados, pero hay una mayor posibilidad que una persona joven rote de la empresa. Además, en el boxplot se puede ratificar la diferencia significativa entre las variables edad y rotación, donde las personas jóvenes tienen más posibilidades de rotar. Esto se ratifica con la prueba de hipótesis T student, con un nivel de confianza del 95%, donde se rechaza la hipótesis nula, confirmando que las personas más jóvenes tienen mayores posibilidades de rotar.

**Hipótesis 2:**

Se espera que la satisfacción laboral de las personas se relacione con la rotación en la empresa, ya que no se cuenta con un sistema de incentivos y se presentan malas condiciones laborales. La hipótesis es que las personas que tienen una baja satisfacción laboral tienen mayor posibilidad de rotar que las que tiene un nivel alto.

```
t.test(Datos$Satisfación_Laboral~Datos$Rotación)
```

```
##
##  Welch Two Sample t-test
##
## data:  Datos$Satisfación_Laboral by Datos$Rotación
## t = 3.9261, df = 328.59, p-value = 0.0001052
## alternative hypothesis: true difference in means between group No and group Si is not equal to 0
## 95 percent confidence interval:
##  0.1547890 0.4656797
## sample estimates:
## mean in group No mean in group Si
##         2.778589         2.468354
```

```
ggplot(Datos,aes(Rotación, Satisfación_Laboral,fill= Rotación))+geom_boxplot(width=0.5)
```

Dado que este valor p es inferior a 0,05, podemos rechazar la hipótesis nula y concluir que existe una diferencia estadísticamente significativa entre la satisfacción laboral y la rotación de los empleados, concluyendo que no es una variable significativa para el estudio. Además, en el boxplot se puede ratificar la diferencia significativa entre las variables Satisfacción laboral y rotación.

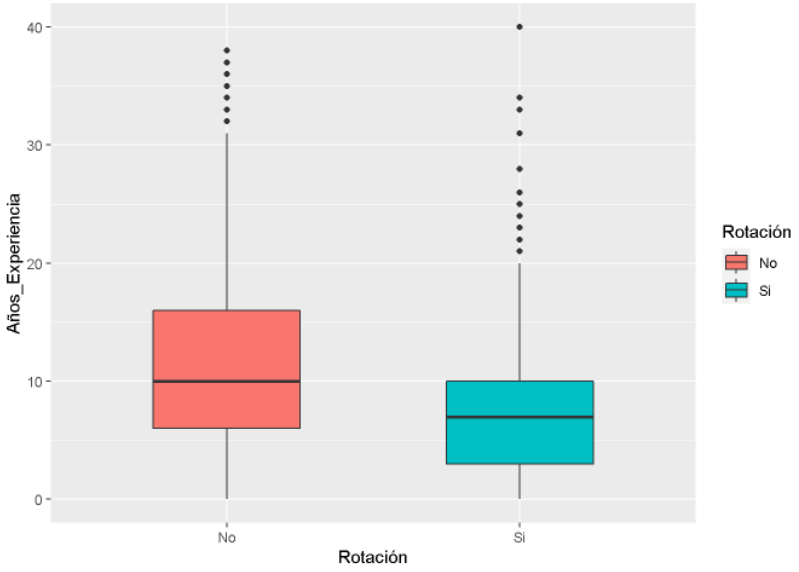
**Hipótesis 3:**

Se espera que los años de experiencia se relacionen con la rotación, ya que las personas pueden acceder a mejores salarios y una mejor calidad de vida. La hipótesis es que las personas que tiene mayores años de experiencia tienen mayor posibilidad de rotar que las que están apenas iniciando en el campo laboral.

```
t.test(Datos$Años_Experiencia~Datos$Rotación)
```

```
##
##  Welch Two Sample t-test
##
## data:  Datos$Años_Experiencia by Datos$Rotación
## t = 7.0192, df = 350.88, p-value = 1.16e-11
## alternative hypothesis: true difference in means between group No and group Si is not equal to 0
## 95 percent confidence interval:
##  2.604401 4.632019
## sample estimates:
## mean in group No mean in group Si
##        11.862936         8.244726
```

```
ggplot(Datos,aes(Rotación, Años_Experiencia,fill= Rotación))+geom_boxplot(width=0.5)
```




Dado que este valor p es inferior a 0,05, podemos rechazar la hipótesis nula y concluir que existe una diferencia entre los años de experiencia y la rotación de los empleados. Además, en el boxplot se puede ratificar la diferencia significativa entre las variables Años experiencia y rotación, donde las personas con menos de 10 años de experiencia tienen mayor posibilidad de rotar. Esto se ratifica con la prueba de hipótesis T student, con un nivel de confianza del 95%, donde se rechaza la hipótesis nula.

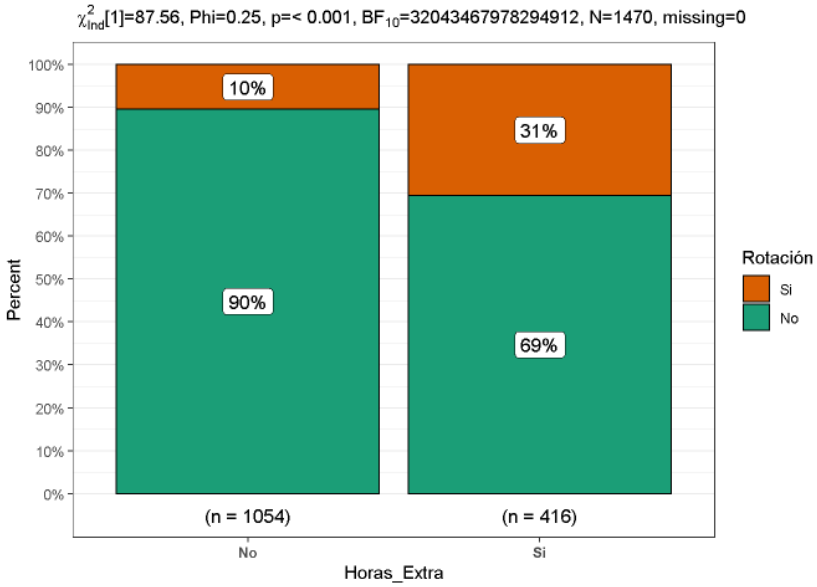
**Hipótesis 4:**

```
require(CGPfunctions)
```

```
## Loading required package: CGPfunctions
```

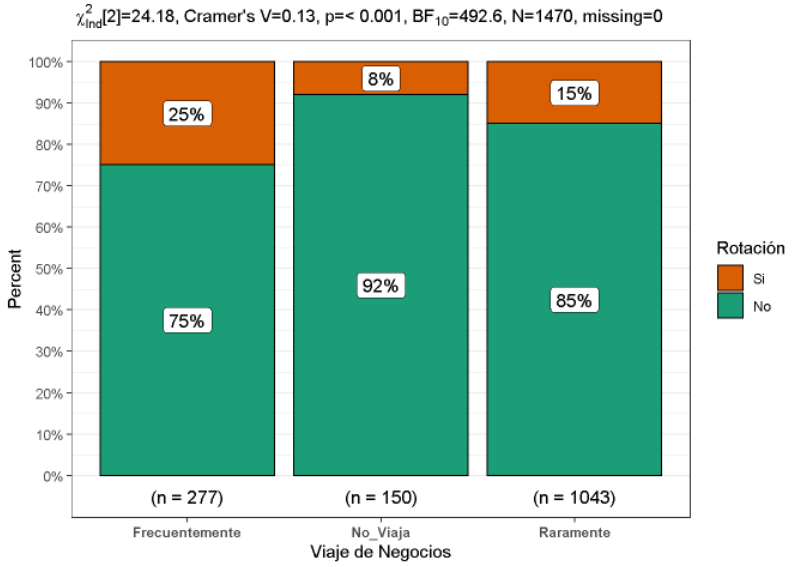
```
## Warning: package 'CGPfunctions' was built under R version 4.1.3
```

```
PlotXTabs2(data = Datos,x = Horas_Extra,y = Rotación)
```




En la Gráfica se observa que del total de personas que tiene horas extras, es decir 416, el 31% rotan de la empresa, el cual es un porcentaje importante, y el 69% de las personas a pesar de que tiene horas extras no rotan. Por otro lado, las personas que no viajan solamente el 10% rotan, el cual es un porcentaje bajo. Para el estudio la variable horas extras permite tomar decisiones para minimizar la rotación de los empleados.

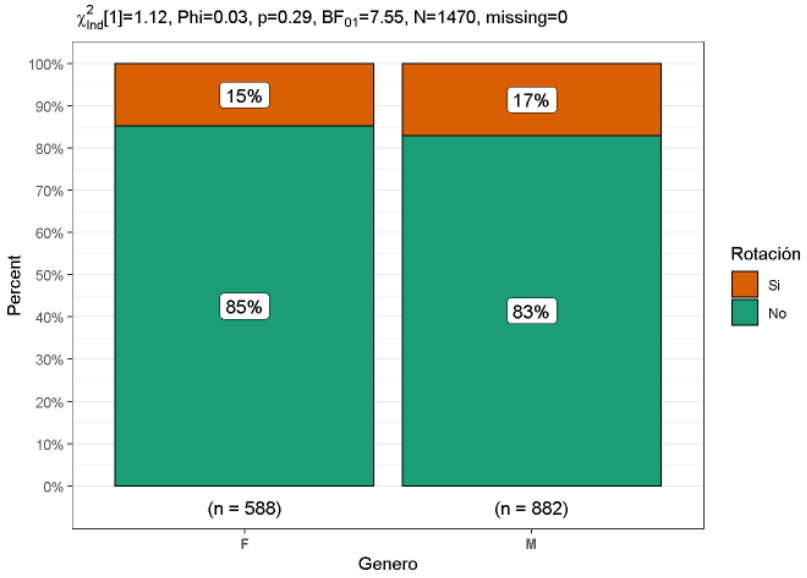
```
require(CGPfunctions)
PlotXTabs2(data = Datos,x = `Viaje de Negocios`, y = Rotación)
```



En la Gráfica se observa que del total de personas que viajan con frecuencia, es decir 277, el 25% rotan de la empresa, y el 75% de las personas a pesar viajar a cada momento no rotan. Así mismo las personas que rara vez viajan solo el 15% rotan de la empresa, y el 85% no rotan. Concluyendo que las personas que viajan con frecuencia son las que más posibilidades tienen de rotar dentro de la empresa, esto permitirá establecer mejoras para minimizar este impacto.

**Hipótesis 5:**

```
require(CGPfunctions)
PlotXTabs2(data = Datos,x = Genero, y = Rotación)
```




En el Gráfica de comparación entre rotación del personal y el género se puede ver un equilibrio en el porcentaje de rotación, para los hombres solo 15% rotan de 588 y de las mujeres solo el 17% rotan de 882. Concluyendo que la variable género no es un determinante para aumentar la rotación.

# CONCLUSIONES

Se puede concluir que, entre las variables cuantitativas escogidas, la edad y los años de experiencia pueden explicar la rotación de los empleados de esta empresa, pero además en las variables cualitativas las horas extras y la frecuencia de viaje también aportan a que se den una alta rotación. Con estas variables, se proponen las siguientes estrategias para disminuir la rotación:

- Como estrategia para disminuir la rotación del personal de la compañía, se propone establecer una política dentro de la empresa para disminuir las horas extras de los trabajadores. Se propone contratar personal capacitado adicional o realizar un equilibrio de carga laboral de los empleados, lo cual no necesita que la empresa realice una inversión. Estas opciones permitirán disminuir el índice de rotación y mejorar la calidad de vida de los empleados.
- Se propone establecer más incentivos tanto a nivel profesional como personal a las personas que viajan con frecuencia, esto les permitirá mantener una mejor calidad de vida. Además de ofrecerles a los empleados estabilidad laboral y generar una mayor confianza entre empresa y empleado.
- Para captar el talento joven en la empresa se propone implementar el teletrabajo, programas de salud y bienestar, oportunidades de crecimiento que son factores relevantes a la hora de decidir si un trabajador joven se queda. Además, a los jóvenes hoy les interesa seguir estudiando y capacitándose para estar a la vanguardia. Una compañía que les apoye con flexibilidad de tiempos y con incentivos económicos para estos fines, genera un factor decisivo para ellos. Por otro lado, los entornos menos rígidos y con políticas más abiertas permitan a los jóvenes interactuar entre todos los niveles.
- Se propone establecer un programa para los trabajadores que van ganando experiencia en la empresa para ir escalando posiciones, asignándoles progresivamente mayor responsabilidad, con un cargo adecuado al nuevo nivel. Por supuesto, esto también se debe reflejar en un aumento de retribución.

# GLOSARIO

**Análisis de Datos:** El proceso de examinar, limpiar, transformar y modelar datos con el objetivo de descubrir patrones, tendencias y obtener información relevante.

**R:** Un lenguaje de programación y entorno de software utilizado para el análisis estadístico y visualización de datos.

**DataFrame:** una estructura de datos en R que organiza los datos en filas y columnas, similar a una tabla de una base de datos.

**Variable:** una característica o atributo que se está midiendo o registrando, como la edad, el ingreso o la temperatura.

**Vector:** una estructura de datos unidimensional que almacena valores del mismo tipo, como un vector numérico o un vector de caracteres.

**Factor:** una variable categórica en R que representa categorías o niveles, como “Bajo”, “Medio” y “Alto”.

**Gráfico de Barras:** un tipo de gráfico que representa datos categóricos usando barras verticales u horizontales de diferentes longitudes.

**Histograma:** un gráfico que muestra la distribución de una variable numérica en intervalos o “bins”.

**Regresión:** un análisis estadístico que examina la relación entre una variable dependiente y una o más variables independientes para predecir valores futuros.

**Clustering:** un método de análisis de datos que agrupa observaciones similares en conjuntos o “clusters” basados en características comunes.

**Estadística Descriptiva:** el uso de estadísticas para resumir y describir características de un conjunto de datos, como la media, la mediana y la desviación estándar.

**Test de Hipótesis:** un procedimiento estadístico para evaluar si hay evidencia suficiente para aceptar o rechazar una afirmación sobre una población.

**Bootstrap:** una técnica de remuestreo que se utiliza para estimar la incertidumbre de una estadística al construir múltiples muestras a partir del conjunto de datos original.

**Correlación:** una medida estadística que evalúa la relación entre dos variables, indicando si cambian juntas y en qué dirección.

**Box Plot:** un gráfico que muestra la distribución de una variable en forma de caja, con los cuartiles y los valores atípicos.

**Machine Learning:** un campo de la inteligencia artificial que se enfoca en desarrollar algoritmos que permiten a las computadoras aprender y mejorar a partir de datos.

**Validación Cruzada:** una técnica que evalúa el rendimiento de un modelo dividiendo los datos en conjuntos de entrenamiento y prueba múltiples para reducir el sesgo de la evaluación.

**Modelo Predictivo:** un modelo estadístico o de aprendizaje automático que se utiliza para hacer predicciones sobre datos no vistos.

**API de R:** un conjunto de funciones y herramientas que permiten la comunicación y la interacción de R con otros programas y sistemas.

**Paquete de R:** una extensión o librería de R que proporciona funcionalidades adicionales, como ggplot2 para visualización o caret para modelado.

**DataFrame Tidy:** un formato de datos estructurado de manera ordenada que facilita el análisis y manipulación de datos utilizando paquetes como dplyr y tidyr.

**Regresión Lineal:** un método de análisis que modela la relación lineal entre una variable dependiente y una o más variables independientes.

**Visualización de Datos:** la representación gráfica de datos que ayuda a comprender patrones, relaciones y tendencias en los datos.

**Efecto de Interacción:** una relación entre variables que afecta la respuesta de manera conjunta, en lugar de forma independiente.

**Reducción de Dimensión:** la técnica para reducir el número de variables en un conjunto de datos, como PCA (Análisis de Componentes Principales).

**Overfitting:** un problema en el modelado que ocurre cuando un modelo se ajusta en exceso a los datos de entrenamiento y tiene un mal rendimiento en datos no vistos.

**Curva ROC:** una representación gráfica de la sensibilidad frente a la especificidad de un modelo de clasificación.

**P-valor:** una medida que ayuda a evaluar la evidencia en contra de una hipótesis nula en un test de hipótesis.

# REFERENCIAS


Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5

Ángel, M., & Pantoja, B. (2014). *La simulación en el contexto de una didáctica de la estadística y la probabilidad INVESTIGACIÓN*.

Beltran-Pellicer, P., Godino, J. D., & Giacomone, B. (2018). Elaboración de Indicadores Específicos de Idoneidad Didáctica en Probabilidad: Aplicación para la Reflexión sobre la Práctica Docente. *Bolema: Boletim de Educação Matemática*, *32*(61), 526-548. https://doi.org/10.1590/1980-4415v32n61a11

Chong, J., Wishart, D. S., & Xia, J. (2019). Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Current Protocols in Bioinformatics*, *68*(1), e86. https://doi.org/10.1002/cpbi.86

Chong, J., & Xia, J. (2020). Using MetaboAnalyst 4.0 for Metabolomics Data Analysis, Interpretation, and Integration with Other Omics Data. En S. Li (Ed.), *Computational Methods and Data Analysis for Metabolomics* (Vol. 2104, pp. 337-360). Springer US. https://doi.org/10.1007/978-1-0716-0239-3_17

Correoso Espinosa, Y., Chávez Jiménez, M., & Puig Vázquez, L. (2011). Alternativas Didácticas de la Estadística Inferencial en el Pregrado de las Ciencias de la Salud. *Revista Información Científica*, *72*(4), 1-9.

Diharce, E. V. (2008). Técnicas De Simulación Para El Análisis Estadístico De Datos De Medición. *Centro de investigación en Matemáticas*.

Hidalgo Troya, A. (2019). Técnicas estadísticas en el análisis cuantitativo de datos. *Revista Sigma*, *15*(1), 28-44.

Kelmansky, D. (2010). *Estadística para todos*.

Kronthaler, F., & Zöllner, S. (2021). *Data Analysis with RStudio: An Easygoing Introduction*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-62518-7

Liu, C., Liu, W., Zhang, G., Wang, Y., Jiang, J., Yang, Z., & Wu, W. (2022). Conjunctional Relationship between Se-

rum Uric Acid and Serum Nickel with Non-Alcoholic Fatty Liver Disease in Men: A Cross-Sectional Study. *International Journal of Environmental Research and Public Health*, *19*(11), 6424. https://doi.org/10.3390/ijerph19116424

Martínez-García, J. A., & Martínez-Caro, L. (2009). La validez discriminante como criterio de evaluación de escalas: ¿Tteoría o estadística? *Universitas Psychologica*, *8*(1), 27-36.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. *F1000Research*, *10*, 33. https://doi.org/10.12688/f1000research.29032.2

Morales, N. (2021). *Implementación de motor de limpieza de datos para la estandarización de los datos insumo del análisis de riesgos en una entidad del Estado*.

Ochoa, G. S., Billingsley, M. C., & Synovec, R. E. (2023). Using solid-phase extraction to facilitate a focused tile-based Fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data: Comparative analysis of aerospace fuel composition. *Analytical and Bioanalytical Chemistry*, *415*(13), 2411-2423. https://doi.org/10.1007/s00216-022-04348-1

Orellana Cordero, M. P., & Cedillo, P. (2020). Detección de valores atípicos con técnicas de minería de datos y métodos estadísticos. *Enfoque UTE*, *11*(1), 56-67. https://doi.org/10.29019/enfoque.v11n1.584

Ruiz Manosalva, G. (2019). Modelo de análisis de datos utilizando técnicas de aprendizaje supervisado y no supervisado, para identificar patrones en la información generada por los pacientes, sometidos a juegos diseñados como un instrumento de apoyo terapéutico. *Universidad de Bogotá*, 18-19.

Schöneich, S., Ochoa, G. S., Monzón, C. M., & Synovec, R. E. (2022). Minimum variance optimized Fisher ratio analysis of comprehensive two-dimensional gas chromatography / mass spectrometry data: Study of the pacu fish metabolome. *Journal of Chromatography A*, *1667*, 462868. https://doi.org/10.1016/j.chroma.2022.462868

Teresa, M., Aceituno, C., Civil, I., & De, C. (2018). *Enseñanza de función variable aleatoria , función de probabi-*

*lidad y función de distribución de probabilidad por medio de la Teoría de Registros de Representaciones Semióticas y la Teoría de Situaciones Didácticas Enseñanza de función variable aleatoria , .*

Trinklein, T. J., & Synovec, R. E. (2022). Simulating comprehensive two-dimensional gas chromatography mass spectrometry data with realistic run-to-run shifting to evaluate the robustness of tile-based Fisher ratio analysis. *Journal of Chromatography A*, *1677*, 463321. https://doi.org/10.1016/j.chroma.2022.463321

Valenzuela, S., Batanero, C., & Begué, N. (2021). Recursos en Internet para el estudio de los estadísticos de orden. *Revista Sergipana de Matemática e Educação Matemática*, *6*(1), 1-22. https://doi.org/10.34179/revisem.v6i1.14419

Vargas-Rojas, J. C. (2021). Uniformity trials simulation to determine the statistical power in yield rice trials. *Agronomia Mesoamericana*, *32*(1), 196-208. https://doi.org/10.15517/am.v32i1.40870

Villanueva, R. A. M., & Chen, Z. J. (2019). Elegant Graphics for Data Analysis. *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 160-167. https://doi.org/10.1080/15366367.2019.1565254

Villarreal Larrinaga, O., & Landeta Rodríguez, J. (2010). El estudio de casos como metodología de investigación científica en dirección y economía de la empresa. Una aplicación a la internacionalización. *Investigaciones Europeas de Direccion y Economia de la Empresa*, *16*(3), 31-52. https://doi.org/10.1016/S1135-2523(12)60033-1

Wang, Z., Liu, C., Zhao, J., & Shao, L. (2022). Extending the Fisher Information Matrix in Gravitational-wave Data Analysis. *The Astrophysical Journal*, *932*(2), 102. https://doi.org/10.3847/1538-4357/ac6b99

Wu, P., Lou, S., Zhang, X., He, J., Liu, Y., & Gao, J. (2021). Data-Driven Fault Diagnosis Using Deep Canonical Variate Analysis and Fisher Discriminant Analysis. *IEEE Transactions on Industrial Informatics*, *17*(5), 3324-3334. https://doi.org/10.1109/TII.2020.3030179

Corporación Universitaria Comfacauca

Unicomfacauca

ACREDITACIÓN INSTITUCIONAL EN ALTA CALIDAD

Vigilada MinEducación