

Inferencia Estadística

Ejercicios

Resueltos

Grado en Ciencia de Datos e Ingeniería de Datos

AUTORES

- Víctor Aceña Gil
- Isaac Martín de Diego

2025-2026



Copyright © 2025 Víctor Aceña Gil, Isaac Martín de Diego. Esta obra está licenciada bajo CC BY-SA 4.0, Creative Commons Atribución-Compartir Igual 4.0 Internacional.

Índice de Soluciones

Prefacio	3
Filosofía pedagógica	3
¿Cómo usar este manual?	3
Metodología de las soluciones	3
Requisitos de software	4
Ejercicio 1	6
Ejercicio 2	7
Ejercicio 3	8
Ejercicio 4	9
Ejercicio 5	10
Ejercicio 6	12
Ejercicio 7	13
Ejercicio 8	15
Ejercicio 9	17
Ejercicio 10	18
Cargar el conjunto de datos	21
Ejercicio 1 (tidyverse y dplyr)	27

Prefacio

Este manual de soluciones complementa el libro **Inferencia Estadística** y está diseñado para proporcionar un apoyo integral al proceso de aprendizaje. Cada solución ha sido desarrollada con el mismo rigor teórico-práctico que caracteriza al curso, ofreciendo no solo la respuesta correcta, sino también el razonamiento estadístico y la interpretación práctica necesarios para una comprensión profunda.

Filosofía pedagógica

Al igual que el libro principal, este manual sigue un enfoque “**teórico-práctico**” sin concesiones. Las soluciones están diseñadas para:

- **Reforzar** la comprensión de los conceptos fundamentales mediante aplicaciones concretas
- **Desarrollar** la intuición estadística a través de interpretaciones razonadas
- **Conectar** la teoría con la práctica mediante código R completamente funcional
- **Fomentar** el pensamiento crítico sobre las limitaciones y supuestos de cada método

¿Cómo usar este manual?

Para maximizar el beneficio de este recurso:

1. **Intentar primero:** Resuelve cada ejercicio por tu cuenta antes de consultar la solución
2. **Estudiar el proceso:** No solo copies el código, entiende la lógica detrás de cada paso
3. **Experimentar:** Modifica los parámetros y observa cómo cambian los resultados
4. **Reflexionar:** Considera las implicaciones prácticas de cada resultado obtenido

Metodología de las soluciones

Cada solución incluye:

- **Código R completo:** Totalmente ejecutable y comentado

- **Explicaciones paso a paso:** Qué hace cada línea y por qué
- **Interpretación de resultados:** Qué significan los números obtenidos
- **Gráficos explicativos:** Visualización de conceptos clave
- **Consejos prácticos:** Cuándo y cómo usar cada técnica

! Uso responsable

Este manual es una herramienta de aprendizaje, no un sustituto del pensamiento propio. Utilízalo para verificar tu comprensión y mejorar tu técnica, pero siempre tras haber hecho un esfuerzo genuino por resolver los problemas de forma independiente.

Requisitos de software

Para ejecutar las soluciones necesitas tener instalados los siguientes paquetes de R:

```
# Paquetes principales
install.packages(c(
  "car",           # Diagnósticos avanzados
  "MASS",          # Datasets y funciones estadísticas
  "glmnet",        # Regularización
  "caret",         # Machine learning
  "pROC",          # Curvas ROC
  "fitdistrplus", # Ajuste de distribuciones
  "lmtest"         # Tests estadísticos
))
```

i Sobre los autores

Víctor Aceña Gil es graduado en Matemáticas por la UNED, máster en Tratamiento Estadístico y Computacional de la Información por la UCM y la UPM, doctor en Tecnologías de la Información y las Comunicaciones por la URJC y profesor del departamento de Informática y Estadística de la URJC. Miembro del grupo de investigación de alto rendimiento en Fundamentos y Aplicaciones de la Ciencia de Datos, DSLAB, de la URJC. Pertenece al grupo de innovación docente, DSLAB-TI.

Isaac Martín de Diego es diplomado en Estadística por la Universidad de Valladolid (UVA), licenciado en Ciencias y Técnicas Estadísticas por la Universidad Carlos III de Madrid (UC3M), doctor en Ingeniería Matemática por la UC3M, catedrático de Ciencias de la Computación e Inteligencia Artificial del departamento de Informática y Estadística de la URJC. Es fundador y coordinador del DSLAB y del DSLAB-TI.

Esta obra está bajo una licencia de Creative Commons Atribución-CompartirIgual 4.0 Interna-
cional.

Ejercicio 1

Pregunta: Describe en tus propias palabras qué es la Ciencia de Datos y su importancia en el análisis de grandes volúmenes de datos.

Solución: La Ciencia de Datos es una disciplina interdisciplinaria que se centra en la extracción de conocimiento significativo a partir de grandes conjuntos de datos. Es crucial en un mundo impulsado por los datos, ya que permite tomar decisiones informadas y hacer predicciones basadas en el análisis de datos. La Ciencia de Datos combina elementos de estadística, informática y conocimiento específico del dominio para interpretar datos y aplicar este conocimiento en diversas áreas como la medicina, las finanzas y la tecnología. Su importancia radica en su capacidad para transformar datos crudos en información valiosa que puede impulsar la innovación y la eficiencia en múltiples campos.

Ejercicio 2

Pregunta: Enumera las herramientas estadísticas que se utilizan en la inferencia estadística y explica brevemente su propósito.

Solución: En la inferencia estadística, se utilizan diversas herramientas para analizar datos y hacer generalizaciones sobre una población a partir de muestras:

- **Pruebas de Hipótesis:** Se utilizan para determinar si existe suficiente evidencia en una muestra de datos para inferir que una cierta condición es verdadera para toda la población.
- **Intervalos de Confianza:** Proporcionan un rango estimado que es probable que contenga el valor de un parámetro desconocido de la población, con un cierto nivel de confianza.
- **Análisis de Varianza (ANOVA):** Permite comparar tres o más medias de grupos para determinar si al menos una de las medias es diferente de las demás.
- **Chi-cuadrado (²):** Es una prueba que mide la discrepancia entre los datos observados y los datos que se esperarían según un modelo específico.
- **T-test:** Evalúa si las medias de dos grupos son estadísticamente diferentes entre sí.
- **Correlación:** Mide la relación entre dos variables y la fuerza de esta relación.
- **Estadística Bayesiana:** Utiliza la probabilidad para representar la incertidumbre sobre los parámetros del modelo y actualiza esta incertidumbre a medida que se obtienen más datos.
- **Métodos de Muestreo:** Incluyen técnicas para seleccionar muestras representativas de la población para realizar inferencias estadísticas.
- **Métodos No Paramétricos:** Son técnicas que no asumen una distribución específica de los datos y son útiles cuando no se cumplen los supuestos de los métodos paramétricos.

Ejercicio 3

Pregunta: Define los términos “población” y “muestra” y explica la diferencia entre ambos.

Solución:

- **Población:** Se refiere al conjunto completo de elementos o resultados que se están estudiando, del cual se desean obtener conclusiones. La población incluye a todos los individuos, mediciones, objetos o eventos que cumplen con un conjunto de especificaciones previamente definidas. Por ejemplo, si estamos estudiando la altura de los estudiantes de una universidad, la población sería la altura de **todos** los estudiantes de esa universidad.
- **Muestra:** Es un subconjunto de la población que se selecciona para representarla. La muestra debe ser representativa de la población para que las inferencias hechas a partir de ella sean válidas. Por ejemplo, si elegimos a 100 estudiantes al azar de la universidad mencionada anteriormente, esos 100 estudiantes constituirían una muestra de la población.

La **diferencia principal** entre ambos términos es el alcance. Mientras que la población es el grupo completo que se quiere estudiar, la muestra es solo una parte de ese grupo. Las muestras se utilizan porque a menudo es impráctico o imposible estudiar toda la población debido a limitaciones de tiempo, costo o logística. Por lo tanto, se selecciona una muestra para obtener estimaciones o pruebas sobre la población completa.

Ejercicio 4

Pregunta: ¿Qué es una distribución de probabilidad y cómo se relaciona con las variables cualitativas y cuantitativas?

Solución: Una **distribución de probabilidad** es una función matemática que describe la probabilidad de ocurrencia de los diferentes posibles resultados en un experimento. En otras palabras, asigna probabilidades a cada posible resultado de una variable aleatoria. Las distribuciones de probabilidad se relacionan con las variables cualitativas y cuantitativas de la siguiente manera:

- **Variables Cualitativas (o Categóricas):** Son aquellas que describen una calidad o categoría y no tienen un orden o medida numérica inherente. Las distribuciones de probabilidad para estas variables son conocidas como **distribuciones discretas** y asignan probabilidades a resultados específicos. Un ejemplo es la **distribución binomial**, que puede modelar eventos como el lanzamiento de una moneda, donde los resultados son categóricos (cara o cruz).
- **Variables Cuantitativas:** Son variables que se pueden medir en una escala numérica y tienen sentido hablar de valores mayores o menores. Las distribuciones de probabilidad asociadas a estas variables son **distribuciones continuas** y asignan probabilidades a intervalos de números. Por ejemplo, la **distribución normal** es una distribución continua que se utiliza comúnmente para modelar fenómenos naturales como la altura o el peso de individuos.

Ejercicio 5

Pregunta: Realiza un resumen descriptivo de un conjunto de datos utilizando medidas de tendencia central y dispersión.

Solución: Imaginemos un conjunto de datos que representa las calificaciones de un grupo de estudiantes en un examen:

```
datos=c(72, 85, 90, 68, 88, 76, 95, 89, 75, 80)
datos
```

```
[1] 72 85 90 68 88 76 95 89 75 80
```

Media:

$$\text{Media} = \frac{72 + 85 + 90 + 68 + 88 + 76 + 95 + 89 + 75 + 80}{10} = \frac{818}{10} = 81.8$$

Mediana: Ordenamos los datos: 68, 72, 75, 76, 80, 85, 88, 89, 90, 95. Como hay 10 valores, la mediana es el promedio de los dos valores centrales:

$$\text{Mediana} = \frac{80 + 85}{2} = \frac{165}{2} = 82.5$$

Moda: No hay un valor que se repita más de una vez, por lo que no hay moda en este conjunto de datos.

Rango:

$$\text{Rango} = 95 - 68 = 27$$

Varianza: Primero, calculamos la media de los cuadrados de las desviaciones respecto a la media:

$$\sigma^2 = \frac{(72 - 81.8)^2 + (85 - 81.8)^2 + \dots + (80 - 81.8)^2}{10} = 71.4$$

Desviación Estándar:

$$\sigma = \sqrt{71.4} \approx 8.45$$

Cuartiles: - **Primer Cuartil (Q1):** Valor en el percentil 25 (primera mitad inferior de los datos):

$$Q1 = 75$$

- **Tercer Cuartil (Q3):** Valor en el percentil 75 (primera mitad superior de los datos):

$$Q3 = 89$$

Resumen Descriptivo

- **Media:** 81.8
- **Mediana:** 82.5
- **Moda:** No hay moda
- **Rango:** 27
- **Varianza:** 71.4
- **Desviación Estándar:** 8.45
- **Primer Cuartil (Q1):** 75
- **Tercer Cuartil (Q3):** 89

Este resumen descriptivo proporciona una visión clara y concisa de las características principales del conjunto de datos, permitiendo una mejor comprensión de su distribución y variabilidad.

Ejercicio 6

Pregunta: Explica la diferencia entre estadística descriptiva e inferencial y proporciona un ejemplo de cómo se utiliza cada una.

Solución: La **estadística descriptiva** y la **estadística inferencial** son dos ramas principales de la estadística que tienen propósitos y métodos diferentes:

- **Estadística Descriptiva:** Se centra en resumir y describir las características de un conjunto de datos. Utiliza medidas como la media, mediana, moda, rango y desviación estándar para dar una visión general de los datos. Por ejemplo, si tenemos los resultados de una prueba de matemáticas de una clase, la estadística descriptiva podría incluir el cálculo de la media de las calificaciones, la calificación más alta, la más baja y la variabilidad de las calificaciones.
- **Estadística Inferencial:** Va más allá de la descripción de los datos y busca hacer predicciones o generalizaciones sobre una población basándose en una muestra de datos. Utiliza herramientas como pruebas de hipótesis, intervalos de confianza y regresión para inferir patrones y tomar decisiones. Por ejemplo, si queremos saber si un nuevo método de enseñanza es efectivo, podríamos aplicarlo a una muestra de estudiantes y usar la estadística inferencial para determinar si los resultados observados en la muestra pueden generalizarse a todos los estudiantes.

Solución:

Ejercicio 7

Pregunta: Diseña un experimento para ilustrar cómo el muestreo aleatorio simple puede ser utilizado para estimar una característica de una población.

Solución: Para ilustrar cómo el muestreo aleatorio simple puede ser utilizado para estimar una característica de una población, consideremos el siguiente experimento:

Objetivo del Experimento: Estimar la proporción de personas en una ciudad que prefieren el transporte público sobre otros medios de transporte.

Población: Todos los residentes de la ciudad que son mayores de edad y utilizan algún medio de transporte para desplazarse.

Característica de Interés: Preferencia por el transporte público.

Procedimiento:

1. **Definición de la Población:** Identificar a todos los residentes de la ciudad que son mayores de edad y utilizan algún medio de transporte.

2. **Selección de la Muestra:**

- Utilizar un registro actualizado de la población, como el padrón municipal, para obtener una lista de individuos.
- Seleccionar una muestra aleatoria de individuos utilizando un generador de números aleatorios.
- Determinar el tamaño de la muestra necesario para obtener resultados con un nivel de confianza y un margen de error deseado.

3. **Recolección de Datos:**

- Contactar a los individuos seleccionados y preguntarles si prefieren el transporte público sobre otros medios de transporte.
- Registrar las respuestas afirmativas y negativas.

4. **Análisis de Datos:**

- Calcular la proporción de respuestas afirmativas en la muestra.
- Utilizar esta proporción como una estimación puntual de la preferencia en la población total.

5. Estimación de la Población:

- Calcular un intervalo de confianza para la proporción estimada, lo que proporcionará un rango dentro del cual se espera que se encuentre la verdadera proporción de la población con un cierto nivel de confianza.

6. Conclusión:

- Presentar la proporción estimada y el intervalo de confianza como la estimación de la preferencia por el transporte público en la población.
- Discutir las limitaciones del estudio y la posibilidad de sesgo si la muestra no fue perfectamente aleatoria o si hubo una tasa de respuesta baja.

Ejercicio 8

Pregunta: Explica el Teorema Central del Límite y su relevancia en la inferencia estadística.

Solución: Teorema Central del Límite (TCL)

El Teorema Central del Límite es un principio fundamental en estadística que establece que, bajo ciertas condiciones, la distribución de la suma de un gran número de variables aleatorias independientes y idénticamente distribuidas (i.i.d.) tiende a aproximarse a una distribución normal (gaussiana), independientemente de la forma de la distribución original de las variables.

Enunciado del Teorema Central del Límite

Para una muestra de tamaño (n) tomada de una población con cualquier distribución de probabilidad con media (μ) y desviación estándar (σ), la distribución de la media muestral \bar{X} se aproximará a una distribución normal a medida que (n) se haga grande. Matemáticamente, si (X_1, X_2, \dots, X_n) son variables aleatorias independientes e idénticamente distribuidas con media (μ) y desviación estándar (σ), entonces la media muestral \bar{X} se distribuye aproximadamente como:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Importancia y Relevancia en la Inferencia Estadística

- Justificación del Uso de la Distribución Normal:** El TCL permite justificar el uso de la distribución normal en la inferencia estadística. Incluso si la población original no está distribuida normalmente, la distribución de las medias muestrales se aproximará a una distribución normal si el tamaño de la muestra es suficientemente grande.
- Construcción de Intervalos de Confianza:** Permite construir intervalos de confianza para estimar parámetros poblacionales. Por ejemplo, al estimar la media poblacional, podemos utilizar la distribución normal para determinar un rango donde probablemente se encuentra la media verdadera.
- Pruebas de Hipótesis:** Facilita la realización de pruebas de hipótesis. Dado que la distribución de la media muestral es aproximadamente normal, se pueden aplicar métodos estadísticos basados en la normalidad para decidir si rechazar o no una hipótesis nula.

4. **Simplificación de Cálculos:** El TCL simplifica el análisis de datos, ya que permite trabajar con la distribución normal, que tiene propiedades bien definidas y es ampliamente comprendida y tabulada, facilitando los cálculos y la interpretación de los resultados.
5. **Aplicación Universal:** Es aplicable en una amplia gama de situaciones, desde la economía y la biología hasta la ingeniería y las ciencias sociales, siempre que se cumplan las condiciones necesarias (independencia y tamaño de muestra grande).

Ejercicio 9

Pregunta: Compara y contrasta la estadística paramétrica y no paramétrica, dando ejemplos de cuándo se utilizaría cada una.

Solución: La **estadística paramétrica** y la **estadística no paramétrica** son dos enfoques de análisis estadístico que tienen diferentes supuestos y aplicaciones. La **estadística paramétrica** se utiliza cuando los datos se ajustan a ciertos supuestos y se conoce la distribución subyacente, lo que permite hacer inferencias más precisas si estos supuestos se cumplen. Por otro lado, la **estadística no paramétrica** es más flexible y se puede utilizar en una variedad más amplia de situaciones, especialmente cuando los datos no cumplen con los supuestos de los métodos paramétricos.

- **Estadística Paramétrica:**

- **Supuestos:** Asume que los datos de la muestra provienen de una población que sigue una distribución de probabilidad conocida, generalmente la distribución normal. También asume homogeneidad de varianzas y la independencia de las observaciones.
- **Uso:** Se utiliza cuando se conoce la forma de la distribución subyacente de los datos o cuando se tiene una muestra grande que, por el Teorema Central del Límite, tiende a una distribución normal.
- **Ejemplos de Herramientas:** T-test, ANOVA, regresión lineal.
- **Ejemplo de Uso:** Si queremos comparar las alturas promedio de dos grupos de personas y sabemos que las alturas siguen una distribución normal, podríamos usar un T-test paramétrico.

- **Estadística No Paramétrica:**

- **Supuestos:** No hace suposiciones sobre la forma de la distribución de la población. Es útil cuando no se cumplen los suposiciones de normalidad o cuando se trata con muestras pequeñas.
- **Uso:** Se aplica en situaciones donde no se conoce la distribución de los datos o cuando los datos son ordinales o nominales.
- **Ejemplos de Herramientas:** Test de Wilcoxon, Test de Kruskal-Wallis, Test de Chi-cuadrado.
- **Ejemplo de Uso:** Si queremos comparar las medianas de los tiempos de respuesta de dos grupos en una prueba y los datos son claramente no normales o son rangos en lugar de medidas, podríamos usar un test de Wilcoxon no paramétrico.

Ejercicio 10

Pregunta: Discute las diferencias entre los enfoques frecuentista y bayesiano en la inferencia estadística y da un ejemplo de aplicación para cada uno.

Solución: La inferencia estadística se basa en métodos que nos permiten hacer conclusiones sobre una población a partir de datos muestrales. Existen dos enfoques principales en la inferencia estadística: el enfoque frecuentista y el enfoque bayesiano. A continuación, se presentan las diferencias clave entre estos enfoques y ejemplos de aplicación para cada uno.

Enfoque Frecuentista

Características Principales:

1. **Interpretación de Probabilidad:** La probabilidad se interpreta como la frecuencia relativa de eventos en el largo plazo. Es decir, si un experimento se repite infinitas veces, la probabilidad de un evento es la proporción de veces que ocurre.
2. **Estimación de Parámetros:** Se basa en el concepto de estimación puntual y de intervalos de confianza. Los parámetros poblacionales se consideran fijos pero desconocidos, y los datos son aleatorios.
3. **Pruebas de Hipótesis:** Utiliza pruebas de hipótesis y valores p para decidir si rechazar la hipótesis nula. Las decisiones se basan en la frecuencia de observación de datos extremos bajo la suposición de que la hipótesis nula es verdadera.
4. **No usa Información Priori:** Los análisis frecuentistas no incorporan información previa sobre los parámetros; sólo se basan en los datos actuales.

Ejemplo de Aplicación Frecuentista:

Imaginemos que una empresa desea saber si un nuevo medicamento es efectivo para reducir la presión arterial. Realizan un ensayo clínico donde:

- **Hipótesis Nula (H_0):** El medicamento no tiene efecto en la presión arterial (la media de la reducción de la presión arterial es 0).
- **Hipótesis Alternativa (H_1):** El medicamento reduce la presión arterial (la media de la reducción de la presión arterial es mayor que 0).

El análisis frecuentista implicaría:

1. Recoger datos muestrales de pacientes.
2. Calcular la media y la desviación estándar de la reducción en la presión arterial.

3. Realizar una prueba t para comparar la media muestral con 0.
4. Calcular el valor p para determinar la significancia estadística.
5. Rechazar o no la hipótesis nula en función del valor p y el nivel de significancia establecido (por ejemplo, 0.05).

Enfoque Bayesiano

Características Principales:

1. **Interpretación de Probabilidad:** La probabilidad se interpreta como un grado de creencia o confianza sobre la ocurrencia de un evento, dado el conocimiento disponible.
2. **Estimación de Parámetros:** Los parámetros poblacionales se tratan como variables aleatorias con distribuciones de probabilidad. Utiliza la distribución a priori (información previa) y los datos observados para obtener la distribución a posteriori.
3. **Actualización de Conocimientos:** Aplica el Teorema de Bayes para actualizar la probabilidad a medida que se dispone de nueva información.
4. **Incorporación de Información Priori:** Utiliza información previa sobre los parámetros en forma de distribuciones a priori, que se combinan con la información de los datos para obtener las distribuciones a posteriori.

Ejemplo de Aplicación Bayesiana:

Supongamos que un médico quiere estimar la probabilidad de que un paciente tenga una enfermedad dada, basándose en un resultado positivo de una prueba diagnóstica y en el conocimiento previo sobre la prevalencia de la enfermedad y la precisión de la prueba.

1. **Información a Priori:** El médico tiene una estimación previa (a priori) de la prevalencia de la enfermedad (por ejemplo, 1% de la población tiene la enfermedad).
2. **Datos Observados:** La sensibilidad (probabilidad de un resultado positivo dado que el paciente tiene la enfermedad) es 90% y la especificidad (probabilidad de un resultado negativo dado que el paciente no tiene la enfermedad) es 95%.
3. **Aplicación del Teorema de Bayes:**

- ($P(E|+)$): Probabilidad de tener la enfermedad dado un resultado positivo.
- ($P(+|E)$): Sensibilidad.
- ($P(+|\neg E)$): Probabilidad de un falso positivo (1 - especificidad).
- ($P(E)$): Prevalencia de la enfermedad.
- ($P(\neg E)$): Probabilidad de no tener la enfermedad (1 - prevalencia).

$$P(E|+) = \frac{P(+|E) \cdot P(E)}{P(+|E) \cdot P(E) + P(+|\neg E) \cdot P(\neg E)}$$

Sustituyendo los valores:

$$P(E|+) = \frac{0.90 \cdot 0.01}{0.90 \cdot 0.01 + 0.05 \cdot 0.99} = \frac{0.009}{0.009 + 0.0495} = \frac{0.009}{0.0585} \approx 0.154$$

El resultado indica que, dado un resultado positivo de la prueba, la probabilidad a posteriori de tener la enfermedad es aproximadamente 15.4%.

Conclusión

Frecuentista:

- No utiliza información previa.
- Se basa en la frecuencia de los eventos.
- Adecuado para análisis donde no se dispone de información previa o se quiere evitar la subjetividad.

Bayesiano:

- Utiliza información previa (a priori).
- Actualiza las probabilidades a medida que se dispone de nueva información.
- Adecuado para situaciones donde la información previa es relevante y valiosa.

Ambos enfoques son útiles y válidos, y la elección entre ellos depende del contexto del problema, la disponibilidad de información previa y las preferencias del investigador.

Cargar el conjunto de datos

```
library(ggplot2)
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

```
library(readr)
```

```
url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank.zip"
download.file(url, "bank.zip")
unzip("bank.zip", "bank-full.csv")
bank_data <- read.csv("bank-full.csv", sep = ";")
```

```
# Pregunta 1: Estructura del DataFrame
head(bank_data)
```

	age	job	marital	education	default	balance	housing	loan	contact	day
1	58	management	married	tertiary	no	2143	yes	no	unknown	5
2	44	technician	single	secondary	no	29	yes	no	unknown	5
3	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5
4	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5
5	33	unknown	single	unknown	no	1	no	no	unknown	5
6	35	management	married	tertiary	no	231	yes	no	unknown	5

month duration campaign pdays previous poutcome y

```

1   may      261      1     -1      0  unknown no
2   may      151      1     -1      0  unknown no
3   may       76      1     -1      0  unknown no
4   may       92      1     -1      0  unknown no
5   may      198      1     -1      0  unknown no
6   may      139      1     -1      0  unknown no

```

```
dim(bank_data)
```

```
[1] 45211    17
```

```
# Pregunta 2: Resumen del DataFrame
summary(bank_data)
```

	age	job	marital	education
Min.	:18.00	Length:45211	Length:45211	Length:45211
1st Qu.	:33.00	Class :character	Class :character	Class :character
Median	:39.00	Mode :character	Mode :character	Mode :character
Mean	:40.94			
3rd Qu.	:48.00			
Max.	:95.00			
	default	balance	housing	loan
Length:	45211	Min. : -8019	Length:45211	Length:45211
Class :	character	1st Qu.: 72	Class :character	Class :character
Mode :	character	Median : 448	Mode :character	Mode :character
		Mean : 1362		
		3rd Qu.: 1428		
		Max. :102127		
	contact	day	month	duration
Length:	45211	Min. : 1.00	Length:45211	Min. : 0.0
Class :	character	1st Qu.: 8.00	Class :character	1st Qu.: 103.0
Mode :	character	Median :16.00	Mode :character	Median : 180.0
		Mean :15.81		Mean : 258.2
		3rd Qu.:21.00		3rd Qu.: 319.0
		Max. :31.00		Max. :4918.0
	campaign	pdays	previous	poutcome
Min. :	1.000	Min. : -1.0	Min. : 0.0000	Length:45211
1st Qu.:	1.000	1st Qu.: -1.0	1st Qu.: 0.0000	Class :character
Median :	2.000	Median : -1.0	Median : 0.0000	Mode :character
Mean :	2.764	Mean : 40.2	Mean : 0.5803	
3rd Qu.:	3.000	3rd Qu.: -1.0	3rd Qu.: 0.0000	

```

Max.    :63.000   Max.    :871.0   Max.    :275.0000
      y
Length:45211
Class :character
Mode  :character

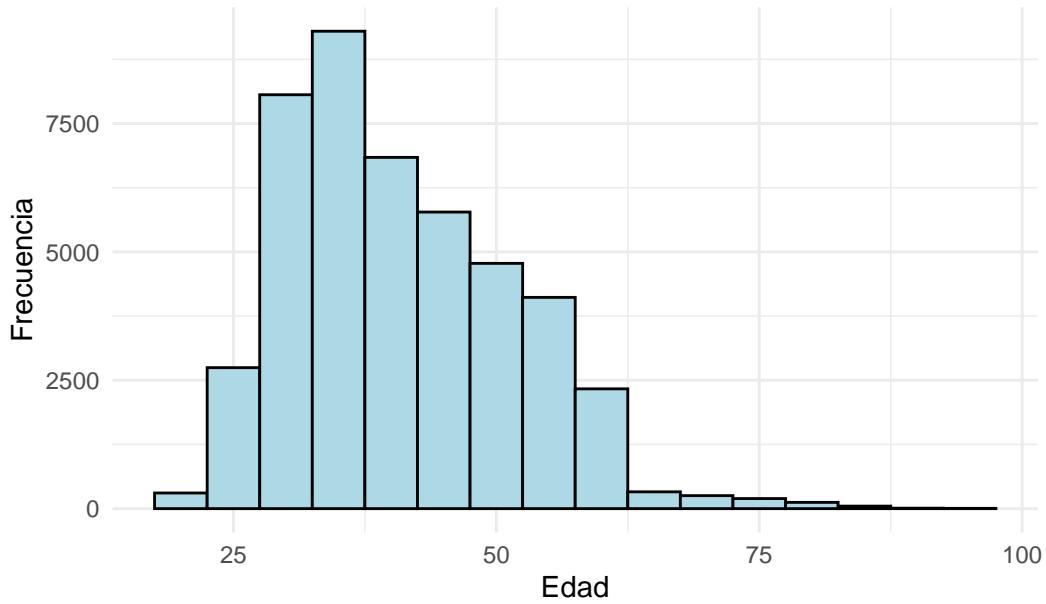
```

```
# Pregunta 3: Distribución de la Edad
summary(bank_data$age)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.00	33.00	39.00	40.94	48.00	95.00

```
ggplot(bank_data, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  labs(title = "Distribución de la Edad de los Clientes",
       x = "Edad",
       y = "Frecuencia") +
  theme_minimal()
```

Distribución de la Edad de los Clientes



```
# Pregunta 4: Balance Promedio  
mean(bank_data$balance, na.rm = TRUE)
```

```
[1] 1362.272
```

```
mean(bank_data$balance[bank_data$y == "yes"], na.rm = TRUE)
```

```
[1] 1804.268
```

```
# Pregunta 5: Frecuencia de Contacto  
summary(bank_data$campaign)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	1.000	2.000	2.764	3.000	63.000

```
# Pregunta 6: Análisis de Duración  
summary(bank_data$duration)
```

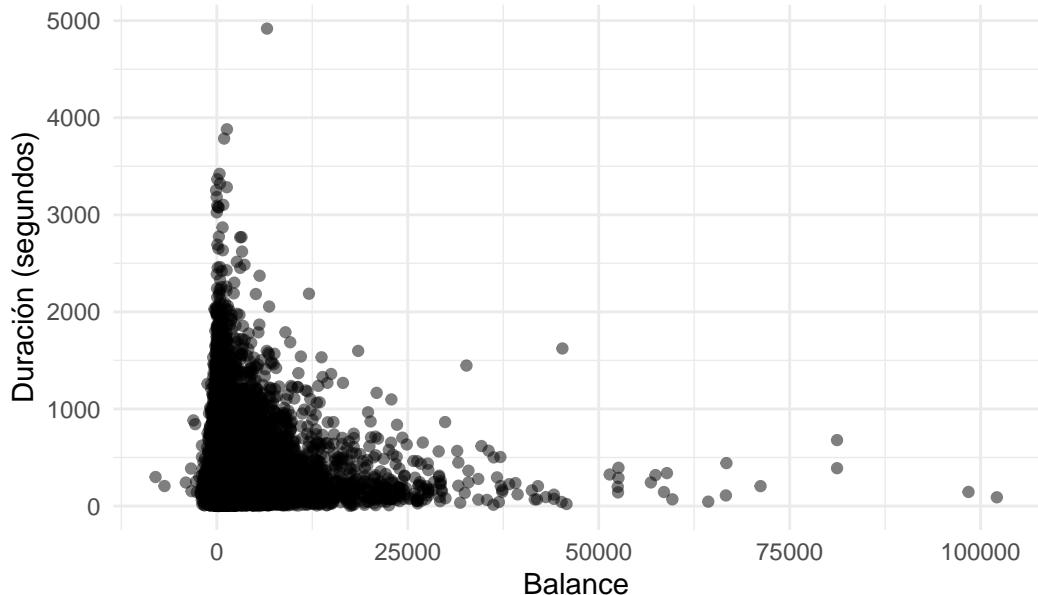
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	103.0	180.0	258.2	319.0	4918.0

```
mean(bank_data$duration[bank_data$y == "yes"], na.rm = TRUE)
```

```
[1] 537.2946
```

```
# Pregunta 7: Relación entre Balance y Duración  
ggplot(bank_data, aes(x = balance, y = duration)) +  
  geom_point(alpha = 0.5) +  
  labs(title = "Relación entre Balance y Duración del Último Contacto",  
       x = "Balance",  
       y = "Duración (segundos)") +  
  theme_minimal()
```

Relación entre Balance y Duración del Último Contacto



```
# Pregunta 8: Segmentación por Trabajo
bank_data %>% group_by(job) %>%
  summarise(media_balance = mean(balance, na.rm = TRUE),
            mediana_balance = median(balance, na.rm = TRUE))
```

```
# A tibble: 12 x 3
  job          media_balance mediana_balance
  <chr>           <dbl>             <dbl>
1 admin.        1136.            396
2 blue-collar   1079.            388
3 entrepreneur  1521.            352
4 housemaid    1392.            406
5 management    1764.            572
6 retired       1984.            787
7 self-employed 1648.            526
8 services      997.             340.
9 student        1388.            502
10 technician   1253.            421
11 unemployed   1522.            529
12 unknown       1772.            677
```

```
# Pregunta 9: Análisis de Contactos Anteriores  
summary(bank_data$pdays)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.0	-1.0	-1.0	40.2	-1.0	871.0

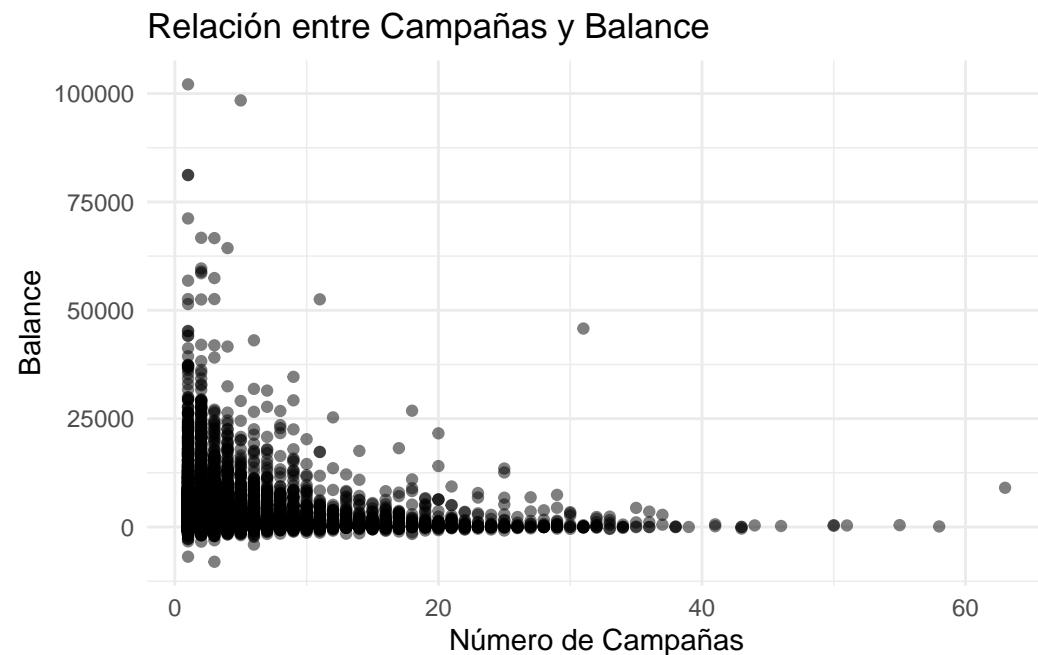
```
mean(bank_data$pdays[bank_data$y == "yes"], na.rm = TRUE)
```

```
[1] 68.70297
```

```
# Pregunta 10: Estudio General  
# Puedes utilizar técnicas de análisis más avanzadas como correlación, regresión, etc.  
cor(bank_data$balance, bank_data$duration, use = "complete.obs")
```

```
[1] 0.02156038
```

```
ggplot(bank_data, aes(x = campaign, y = balance)) +  
  geom_point(alpha = 0.5) +  
  labs(title = "Relación entre Campañas y Balance",  
       x = "Número de Campañas",  
       y = "Balance") +  
  theme_minimal()
```



Ejercicio 1 (tidyverse)

```
library(tidyr)
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

```
# A partir del siguiente dataframe realizar las siguientes operaciones de limpieza de datos:
set.seed(1)
stocks <- data.frame(
  time = as.Date('2009-01-01') + 0:9,
  Walmart = rnorm(10, 20, 1),
  Target = rnorm(10, 20, 2),
  Walgreens = rnorm(10, 20, 4)
)
#      time   Walmart   Target Walgreens
# 1 2009-01-01 19.37355 23.02356 23.67591
# 2 2009-01-02 20.18364 20.77969 23.12855
# 3 2009-01-03 19.16437 18.75752 20.29826
# 4 2009-01-04 21.59528 15.57060 12.04259
# 5 2009-01-05 20.32951 22.24986 22.47930
# 6 2009-01-06 19.17953 19.91013 19.77549
# 7 2009-01-07 20.48743 19.96762 19.37682
# 8 2009-01-08 20.73832 21.88767 14.11699
# 9 2009-01-09 20.57578 21.64244 18.08740
# 10 2009-01-10 19.69461 21.18780 21.67177
```

```

# Como se puede observar hay un problema de clave-valor en las compañías con sus observaciones
# Transformar los datos para que tengan una clave stock y el valor sea el precio.
# Por lo tanto se requiere la función "gather".

# Opción 1:
new_stocks <- gather(data = stocks, key = stock, value = price, Walmart, Target, Walgreens)

# Opción 2:
new_stocks <- gather(data = stocks, key = stock, value = price, Walmart:Walgreens)

# Opción 3:
new_stocks <- gather(data = stocks, key = stock, value = price, -time)
# El último argumento, -time, significa que todas las columnas excepto el tiempo
# contienen los pares clave-valor.

# Devolver el dataframe al estado original utilizando la función "spread".
original_stocks <- spread(data = new_stocks, key = stock, value = price)

# Utilizando el operador tubería %>% se desea realizar las siguientes operaciones anidadas.
# 1) Transformar los datos para que tengan una clave stock y el valor sea
# el precio mediante la función "gather".
# 2) Agrupar los datos por la clave stock mediante la función "group_by".
# 3) Obtener el precio mínimo y máximo utilizando la función "summarise".

stocks %>%
  gather(key = stock, value = price, Walmart:Walgreens)%>%
  group_by(stock) %>%
  summarise(min = min(price), max = max(price))

```

stock	min	max
Target	15.6	23.0
Walgreens	12.0	23.7
Walmart	19.2	21.6

#####

```

# Ejercicio 2 (dplyr)

library(dplyr)
library(nycflights13)

# COMPROBACION.
# Observamos los distintos dataframes que nos proporcionan.
# Utilizamos el nombre del paquete y doblemente dos puntos (::) para comprobarlo.
# Tambien se puede utilizar el nombre del dataframe si previamente estamos familiarizados.

# PRIMERA OBSERVACION.
# Comprobamos las variables de cada uno de los datasets que nos proporcionan
# mediante la instrucción "head".
print(head(flights))

# A tibble: 6 x 19
  year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
  <int> <int> <int>     <int>           <int>      <dbl>    <int>           <int>
1 2013     1     1      517            515        2       830            819
2 2013     1     1      533            529        4       850            830
3 2013     1     1      542            540        2       923            850
4 2013     1     1      544            545       -1      1004           1022
5 2013     1     1      554            600       -6      812            837
6 2013     1     1      554            558       -4      740            728
# i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
#   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dttm>

print(head(airports))

# A tibble: 6 x 8
  faa      name          lat   lon   alt   tz dst tzone
  <chr> <chr>        <dbl> <dbl> <dbl> <dbl> <chr> <chr>
1 04G    Lansdowne Airport 41.1 -80.6 1044  -5 A America/Ne-
2 06A    Moton Field Municipal Airport 32.5 -85.7  264  -6 A America/Ch-
3 06C    Schaumburg Regional 42.0 -88.1  801  -6 A America/Ch-
4 06N    Randall Airport    41.4 -74.4  523  -5 A America/Ne-
5 09J    Jekyll Island Airport 31.1 -81.4   11  -5 A America/Ne-
6 0A9    Elizabethton Municipal Airport 36.4 -82.2 1593  -5 A America/Ne-

```

```

print(head(weather))

# A tibble: 6 x 15
  origin year month   day hour  temp dewp humid wind_dir wind_speed wind_gust
  <chr>  <int> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 EWR    2013     1     1     1  39.0  26.1  59.4    270    10.4    NA
2 EWR    2013     1     1     2  39.0  27.0  61.6    250     8.06   NA
3 EWR    2013     1     1     3  39.0  28.0  64.4    240    11.5    NA
4 EWR    2013     1     1     4  39.9  28.0  62.2    250    12.7    NA
5 EWR    2013     1     1     5  39.0  28.0  64.4    260    12.7    NA
6 EWR    2013     1     1     6  37.9  28.0  67.2    240    11.5    NA
# i 4 more variables: precip <dbl>, pressure <dbl>, visib <dbl>,
#   time_hour <dttm>

print(head(airlines))

# A tibble: 6 x 2
  carrier name
  <chr>   <chr>
1 9E      Endeavor Air Inc.
2 AA      American Airlines Inc.
3 AS      Alaska Airlines Inc.
4 B6      JetBlue Airways
5 DL      Delta Air Lines Inc.
6 EV      ExpressJet Airlines Inc.

print(head(planes))

# A tibble: 6 x 9
  tailnum year type          manufacturer model engines seats speed engine
  <chr>  <int> <chr>        <chr>       <chr>  <int> <int> <int> <chr>
1 N10156  2004 Fixed wing ~ EMBRAER   EMB~~    2     55    NA Turbo~
2 N102UW   1998 Fixed wing ~ AIRBUS INDU~ A320~    2    182    NA Turbo~
3 N103US   1999 Fixed wing ~ AIRBUS INDU~ A320~    2    182    NA Turbo~
4 N104UW   1999 Fixed wing ~ AIRBUS INDU~ A320~    2    182    NA Turbo~
5 N10575   2002 Fixed wing ~ EMBRAER   EMB~~    2     55    NA Turbo~
6 N105UW   1999 Fixed wing ~ AIRBUS INDU~ A320~    2    182    NA Turbo~

```

```
# Comprobamos las variables de cada uno de los datasets que nos proporcionan
# mediante la instrucción "summary".
print(summary(flights))
```

year	month	day	dep_time	sched_dep_time
Min. :2013	Min. : 1.000	Min. : 1.00	Min. : 1	Min. : 106
1st Qu.:2013	1st Qu.: 4.000	1st Qu.: 8.00	1st Qu.: 907	1st Qu.: 906
Median :2013	Median : 7.000	Median :16.00	Median :1401	Median :1359
Mean :2013	Mean : 6.549	Mean :15.71	Mean :1349	Mean :1344
3rd Qu.:2013	3rd Qu.:10.000	3rd Qu.:23.00	3rd Qu.:1744	3rd Qu.:1729
Max. :2013	Max. :12.000	Max. :31.00	Max. :2400	Max. :2359
			NA's :8255	
dep_delay	arr_time	sched_arr_time	arr_delay	
Min. : -43.00	Min. : 1	Min. : 1	Min. : -86.000	
1st Qu.: -5.00	1st Qu.:1104	1st Qu.:1124	1st Qu.: -17.000	
Median : -2.00	Median :1535	Median :1556	Median : -5.000	
Mean : 12.64	Mean :1502	Mean :1536	Mean : 6.895	
3rd Qu.: 11.00	3rd Qu.:1940	3rd Qu.:1945	3rd Qu.: 14.000	
Max. :1301.00	Max. :2400	Max. :2359	Max. :1272.000	
NA's :8255	NA's :8713		NA's :9430	
carrier	flight	tailnum	origin	
Length:336776	Min. : 1	Length:336776	Length:336776	
Class :character	1st Qu.: 553	Class :character	Class :character	
Mode :character	Median :1496	Mode :character	Mode :character	
	Mean :1972			
	3rd Qu.:3465			
	Max. :8500			
dest	air_time	distance	hour	
Length:336776	Min. : 20.0	Min. : 17	Min. : 1.00	
Class :character	1st Qu.: 82.0	1st Qu.: 502	1st Qu.: 9.00	
Mode :character	Median :129.0	Median : 872	Median :13.00	
	Mean :150.7	Mean :1040	Mean :13.18	
	3rd Qu.:192.0	3rd Qu.:1389	3rd Qu.:17.00	
	Max. :695.0	Max. :4983	Max. :23.00	
	NA's :9430			
minute	time_hour			
Min. : 0.00	Min. :2013-01-01 05:00:00.00			
1st Qu.: 8.00	1st Qu.:2013-04-04 13:00:00.00			
Median :29.00	Median :2013-07-03 10:00:00.00			
Mean :26.23	Mean :2013-07-03 05:22:54.64			
3rd Qu.:44.00	3rd Qu.:2013-10-01 07:00:00.00			

```
Max. :59.00  Max. :2013-12-31 23:00:00.00
```

```
print(summary(airports))
```

faa	name	lat	lon
Length:1458	Length:1458	Min. :19.72	Min. :-176.65
Class :character	Class :character	1st Qu.:34.26	1st Qu.:-119.19
Mode :character	Mode :character	Median :40.09	Median :-94.66
		Mean :41.65	Mean :-103.39
		3rd Qu.:45.07	3rd Qu.:-82.52
		Max. :72.27	Max. : 174.11
alt	tz	dst	tzone
Min. :-54.00	Min. :-10.000	Length:1458	Length:1458
1st Qu.: 70.25	1st Qu.: -8.000	Class :character	Class :character
Median : 473.00	Median : -6.000	Mode :character	Mode :character
Mean :1001.42	Mean : -6.519		
3rd Qu.:1062.50	3rd Qu.: -5.000		
Max. :9078.00	Max. : 8.000		

```
print(summary(weather))
```

origin	year	month	day
Length:26115	Min. :2013	Min. : 1.000	Min. : 1.00
Class :character	1st Qu.:2013	1st Qu.: 4.000	1st Qu.: 8.00
Mode :character	Median :2013	Median : 7.000	Median :16.00
	Mean :2013	Mean : 6.504	Mean :15.68
	3rd Qu.:2013	3rd Qu.: 9.000	3rd Qu.:23.00
	Max. :2013	Max. :12.000	Max. :31.00
hour	temp	dewp	humid
Min. : 0.00	Min. : 10.94	Min. :-9.94	Min. : 12.74
1st Qu.: 6.00	1st Qu.: 39.92	1st Qu.:26.06	1st Qu.: 47.05
Median :11.00	Median : 55.40	Median :42.08	Median : 61.79
Mean :11.49	Mean : 55.26	Mean :41.44	Mean : 62.53
3rd Qu.:17.00	3rd Qu.: 69.98	3rd Qu.:57.92	3rd Qu.: 78.79
Max. :23.00	Max. :100.04	Max. :78.08	Max. :100.00
	NA's :1	NA's :1	NA's :1
wind_dir	wind_speed	wind_gust	precip
Min. : 0.0	Min. : 0.000	Min. :16.11	Min. :0.000000
1st Qu.:120.0	1st Qu.: 6.905	1st Qu.:20.71	1st Qu.:0.000000

```

Median :220.0   Median : 10.357   Median :24.17    Median :0.000000
Mean   :199.8   Mean   : 10.518   Mean   :25.49    Mean   :0.004469
3rd Qu.:290.0   3rd Qu.: 13.809   3rd Qu.:28.77    3rd Qu.:0.000000
Max.   :360.0   Max.   :1048.361   Max.   :66.75    Max.   :1.210000
NA's   :460     NA's   :4         NA's   :20778
pressure      visib       time_hour
Min.   : 983.8   Min.   : 0.000   Min.   :2013-01-01 01:00:00.0
1st Qu.:1012.9   1st Qu.:10.000   1st Qu.:2013-04-01 21:30:00.0
Median :1017.6   Median :10.000   Median :2013-07-01 14:00:00.0
Mean   :1017.9   Mean   : 9.255   Mean   :2013-07-01 18:26:37.7
3rd Qu.:1023.0   3rd Qu.:10.000   3rd Qu.:2013-09-30 13:00:00.0
Max.   :1042.1   Max.   :10.000   Max.   :2013-12-30 18:00:00.0
NA's   :2729

```

```
print(summary(airlines))
```

```

carrier          name
Length:16        Length:16
Class :character Class :character
Mode  :character Mode  :character

```

```
print(summary(planes))
```

```

tailnum          year        type        manufacturer
Length:3322      Min.   :1956   Length:3322      Length:3322
Class :character  1st Qu.:1997   Class :character  Class :character
Mode  :character  Median :2001    Mode  :character  Mode  :character
                           Mean   :2000
                           3rd Qu.:2005
                           Max.   :2013
                           NA's   :70
model            engines      seats       speed
Length:3322      Min.   :1.000   Min.   : 2.0   Min.   : 90.0
Class :character  1st Qu.:2.000   1st Qu.:140.0  1st Qu.:107.5
Mode  :character  Median :2.000   Median :149.0  Median :162.0
                           Mean   :1.995   Mean   :154.3  Mean   :236.8
                           3rd Qu.:2.000   3rd Qu.:182.0  3rd Qu.:432.0
                           Max.   :4.000   Max.   :450.0  Max.   :432.0
                           NA's   :3299
engine
Length:3322

```

```
Class :character  
Mode  :character
```

```
# Simplificar los dataframes originales a 100 observaciones. Renombrarlos  
# introduciendo la coletilla "_simple".  
  
flights_simple <- head(flights,100)  
airports_simple <- head(airports,100)  
weather_simple <- head(weather,100)  
airlines_simple <- head(airlines,100)  
planes_simple <- head(planes,100)  
  
# Selecciona los tipos de aerolinea ("carrier") mediante la instruccion "select"  
# y el operador "unique" concatenados con el operador tuberia %>%.  
airlines_simple %>% unique %>% select(carrier)  
  
# A tibble: 16 x 1  
  carrier  
  <chr>  
1 9E  
2 AA  
3 AS  
4 B6  
5 DL  
6 EV  
7 F9  
8 FL  
9 HA  
10 MQ  
11 OO  
12 UA  
13 US  
14 VX  
15 WN  
16 YV
```

```

# Obtener la media y el maximo de asientos ("seats") que tienen los aviones.
# Utilizar el operador tuberia %>%.
planes_simple %>% summarise(mean = mean(seats),max_engines = max(seats))

# A tibble: 1 x 2
  mean max_engines
  <dbl>      <int>
1   105.        330

# Ordenar los aviones por numero de motores ("engines") y numero de asientos ("seats").
result1 <- arrange(planes_simple,engines,seats)
print(result1)

# A tibble: 100 x 9
  tailnum year type          manufacturer model engines seats speed engine
  <chr>   <int> <chr>        <chr>       <chr>    <int> <int> <int> <chr>
1 N10156  2004 Fixed wing multi~ EMBRAER   EMB~~     2     55   NA Turbo~
2 N10575  2002 Fixed wing multi~ EMBRAER   EMB~~     2     55   NA Turbo~
3 N11106  2002 Fixed wing multi~ EMBRAER   EMB~~     2     55   NA Turbo~
4 N11107  2002 Fixed wing multi~ EMBRAER   EMB~~     2     55   NA Turbo~
5 N11109  2002 Fixed wing multi~ EMBRAER   EMB~~     2     55   NA Turbo~
6 N11113  2002 Fixed wing multi~ EMBRAER   EMB~~     2     55   NA Turbo~
7 N11119  2002 Fixed wing multi~ EMBRAER   EMB~~     2     55   NA Turbo~
8 N11121  2003 Fixed wing multi~ EMBRAER   EMB~~     2     55   NA Turbo~
9 N11127  2003 Fixed wing multi~ EMBRAER   EMB~~     2     55   NA Turbo~
10 N11137 2003 Fixed wing multi~ EMBRAER   EMB~~     2     55   NA Turbo~

# i 90 more rows

# Averigua que numero de cola comparten los dataframes "flights_simple"
# y "planes_simple" que has creado anteriormente.
# Obten su aerolinea ("carrier")
shared <- inner_join(flights_simple,planes_simple,by="tailnum") # -> N14228
shared_carrier <- shared$carrier
print(shared_carrier)

[1] "EV"

```

```

# Cruzar los datos de vuelos ("flights") con los aviones ("planes")
# por el numero de cola ("tailnum") que no coincidan.
# De esos obtener aquellos con 2 o mas motores.
# Finalmente obtener los distintos modelos de avión que satisfacen las premisas anteriores.
fp <- anti_join(planes_simple, flights_simple, by="tailnum")
engines_fp <- filter(fp, engines >= 2)
result2 <- unique(engines_fp$model) # No queremos los repetidos. Por lo tanto usamos "unique"
print(result2)

[1] "EMB-145XR" "A320-214"  "EMB-145LR" "737-824"   "767-332"   "757-224"

# Crea una nueva variable que calcule el retraso total sumando los
# delays acumulados ("dep_delay") y ("arr_delay").
# Almacena el dataframe resultante en "flights_total".
flights_total <- mutate(flights_simple, total_delay=dep_delay+arr_delay)

# En base a la variable anteriormente obtenida, devuelve los aviones que
# han llegado con antelación a su destino.
filter(flights_total, total_delay < 0)

# A tibble: 57 x 20
  year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
  <int> <int> <int>    <int>        <int>     <dbl>    <int>        <int>
1 2013     1     1      544           545      -1       1004        1022
2 2013     1     1      554           600      -6       812         837
3 2013     1     1      557           600      -3       709         723
4 2013     1     1      557           600      -3       838         846
5 2013     1     1      558           600      -2       849         851
6 2013     1     1      558           600      -2       853         856
7 2013     1     1      558           600      -2       923         937
8 2013     1     1      559           559       0       702         706
9 2013     1     1      559           600      -1       854         902
10 2013    1     1      600           600       0       851         858
# i 47 more rows
# i 12 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
# tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
# hour <dbl>, minute <dbl>, time_hour <dttm>, total_delay <dbl>

```