

Nombre del alumno: Gallardo Hernández Evelyn Aline

Matricula: 202118340

Profesor. Ebner Juárez Elías

Cuestionario Escrito 1er parcial.

Calificación:

Materia Análisis y Modelado de Datos

Valor total 30%

**Instrucciones:** contesta correctamente subrayando la respuesta correcta. Debes de entregar escrito a mano correctamente los códigos utilizados, así como compartir en GitHub un repositorio “cuestionario1\_nombrealumno” al usuario profebner.

**Problema 1:** Una empresa de retail ha recopilado datos de ventas de múltiples sucursales, pero presenta valores faltantes, datos duplicados y errores tipográficos. El equipo de análisis de datos necesita limpiar el dataset antes de realizar análisis.

**Tareas:**

1. Cargar un dataset en R
2. Identificar y manejar valores faltantes
3. Detectar y eliminar valores duplicados
4. Estandarizar formatos de nombres de productos

**Cuestionario de Evaluación**

1. ¿Qué función se usa para eliminar valores duplicados en un dataframe en R?
  - a) `remove_duplicates()`
  - b) `distinct()`
  - c) `filter_duplicates()`
2. ¿Cuál es la mejor manera de tratar valores faltantes en una columna numérica?
  - a) Eliminarlos directamente siempre
  - b) Imputarlos con la media o mediana
  - c) Dejar los valores faltantes sin cambios
3. ¿Qué paquete de R facilita la manipulación de datos de manera eficiente?
  - a) `ggplot2`
  - b) `tidyverse`
  - c)
  - d) `shiny`

**Problema 2 :** Un equipo de marketing necesita analizar datos de interacción en redes sociales, pero los datos están en diferentes formatos y escalas, lo que dificulta el análisis.

**Tareas:**

1. Convertir variables categóricas en factores
2. Normalizar valores numéricos
3. Crear nuevas variables derivadas
4. Convertir fechas en formato adecuado

### Cuestionario de Evaluación

1. ¿Qué función se usa para normalizar datos en R?
  - a) `normalize()`
  - b) `scale()`
  - c) `rescale()`
2. ¿Cuál es la ventaja de convertir variables categóricas en factores en R?
  - a) Permite realizar operaciones matemáticas en ellas
  - b) Mejora la eficiencia en el procesamiento y análisis
  - c) Hace que el dataset ocupe más memoria
3. ¿Qué función permite transformar una columna de texto en una fecha en R?
  - a) `to_date()`
  - b) `as.Date()`
  - c) `convert_date()`

**Problema 3:** Un analista de datos necesita fusionar dos datasets: uno con información de clientes y otro con sus compras. Es necesario unirlos de manera eficiente.

### Tareas:

1. Cargar y explorar los dos datasets en R.
2. Unir los datasets
3. Verificar si hay claves duplicadas o valores faltantes después de la fusión.
4. Realizar una consulta de resumen para verificar la correcta integración.

### Cuestionario de Evaluación

1. ¿Cuál de las siguientes funciones se usa para unir dos datasets en R por una clave común?
  - a) `merge()`
  - b) `left_join()`
  - c) `concat()`
2. ¿Qué función permite identificar si hay valores duplicados en una columna clave?
  - a) `table()`
  - b) `duplicated()`
  - c) `unique()`

3. ¿Qué ocurre si se usa `inner_join()` en lugar de `left_join()`?

- a) Se eliminan las filas sin coincidencias en ambas tablas
- b) Se mantienen todas las filas de la tabla izquierda
- c) Se duplican los valores de la clave

**Problema 4:** Un equipo financiero está analizando transacciones, pero ha detectado valores extremadamente altos o bajos en los datos. Es necesario identificar y manejar los outliers.

#### Tareas:

- 1. Identificar outliers mediante diagramas de caja
- 2. Usar el rango intercuartil para determinar límites de outliers.
- 3. Manejar los valores atípicos mediante eliminación o transformación
- 4. Comparar estadísticas antes y después del tratamiento.

#### Cuestionario de Evaluación

- 1. ¿Cuál es una forma común de identificar outliers en un dataset?
  - a) Usar un histograma
  - b) Aplicar la técnica del rango intercuartil (IQR)
  - c) Convertir los valores en  $c' \{$
  - d) eros
- 2. ¿Qué gráfico es más adecuado para visualizar outliers?
  - a) Diagrama de caja
  - b) Gráfico de dispersión
  - c) Gráfico de barras
- 3. ¿Cuál es una estrategia válida para manejar outliers en un dataset?
  - a) Eliminarlos sin análisis previo
  - b) Sustituirlos por la media o mediana
  - c) Ignorarlos completamente

**Problema 5:** Se ha recopilado información de una encuesta con respuestas en formato de texto, pero se necesita transformar las variables categóricas en valores numéricos para análisis estadístico.

#### Tareas:

- 1. Convertir variables cualitativas en numéricas
- 2. Aplicar codificación
- 3. Comparar cómo los modelos de machine learning reaccionan a diferentes codificaciones.

#### Cuestionario de Evaluación

- 1. ¿Por qué es importante codificar variables categóricas en modelos predictivos?
  - a) Porque los modelos solo aceptan datos numéricos

- b) Porque mejora la visualización de datos
- c) No es importante codificarlas
- 2. ¿Qué técnica de codificación de variables categóricas crea múltiples columnas binarias?
  - a) One-hot encoding
  - b) Label encoding
  - c) Scaling
- 3. ¿Qué función en R se usa para transformar variables categóricas en factores numéricos?
  - a) factorize()
  - b) as.factor()
  - c) convert()

**Problema 6:** Un hospital ha recolectado datos de pacientes, pero algunas variables como presión arterial y nivel de glucosa tienen valores faltantes. El equipo de análisis necesita decidir cómo tratarlos antes de realizar estudios estadísticos.

#### Tareas

1. Cargar el dataset en R usando read.csv().
2. Identificar los valores faltantes con is.na() y summary().
3. Aplicar distintas estrategias para manejarlos: eliminación (na.omit()), imputación con la media (tidyverse::replace\_na()), o interpolación.
4. Comparar los efectos de cada estrategia en el dataset final.

#### Cuestionario de Evaluación

1. ¿Qué función en R permite identificar valores faltantes en un dataframe?
  - a) missing\_values()
  - b) is.na()
  - c) find\_NA()
2. ¿Cuál es una estrategia válida para manejar valores faltantes en una columna numérica?
  - a) Eliminarlos sin analizar su impacto
  - b) Imputarlos con la media o la mediana
  - c) Dejar los valores sin cambios y proceder con el análisis
3. ¿Cuál es una posible desventaja de eliminar todas las filas con valores faltantes?
  - a) Puede reducir la cantidad de datos y afectar la representatividad
  - b) No hay ninguna desventaja
  - c) Mejora la calidad de los datos siempre

**Problema 7:** Una empresa de inversiones necesita comparar el desempeño financiero de diversas empresas, pero los datos están en distintas escalas. Se requiere normalizar y estandarizar los datos para hacer comparaciones justas.

#### Tareas

1. Cargar el dataset de indicadores financieros.
2. Aplicar estandarización utilizando `scale()`.
3. Aplicar normalización con la fórmula  $(x - \min(x)) / (\max(x) - \min(x))$ .
4. Evaluar las diferencias entre ambas transformaciones y decidir cuál es más adecuada.

### Cuestionario de Evaluación

1. ¿Cuál es la diferencia entre estandarización y normalización?
  - a) La estandarización ajusta los valores a una media de 0 y desviación estándar de 1, mientras que la normalización los escala entre 0 y 1
  - b) No hay diferencia entre ambas técnicas
  - c) La normalización siempre da mejores resultados
2. ¿Qué función de R permite estandarizar datos?
  - a) `normalize()`
  - b) `scale()`
  - c) `standardize()`
3. ¿En qué caso es más útil la normalización en lugar de la estandarización?
  - a) Cuando los datos tienen distribuciones con valores extremos
  - b) Cuando se requiere comparar datos en diferentes escalas
  - c) Cuando se trabaja con variables categóricas

**Problema 8:** Una empresa de comercio electrónico tiene un dataset con información de clientes y otro con el historial de compras. Se necesita fusionar ambas bases para

**Tareas**

1. Cargar los dos datasets en R.
2. Fusionar los datos usando `left_join()` de dplyr.
3. Detectar y manejar duplicados con `distinct()`.
4. Verificar si hay inconsistencias después de la integración.

### Cuestionario de Evaluación

1. ¿Qué función en R se usa para unir datasets por una columna común?
  - a) `merge()`
  - b) `left_join()`
  - c) `combine()`
2. ¿Qué ocurre si se usa `inner_join()` en lugar de `left_join()`?
  - a) Se eliminan las filas sin coincidencias en ambas tablas
  - b) Se mantienen todas las filas de la tabla izquierda
  - c) Se duplican las filas sin coincidencias
3. ¿Cómo se identifican valores duplicados en R?
  - a) `duplicated()`
  - b) `unique()`
  - c) `filter_duplicates()`

**Problema 9:** Un equipo de calidad de una fábrica detectó que ciertos valores de producción están fuera de lo esperado. Se necesita identificar y decidir qué hacer con estos valores atípicos.

### Tareas

1. Visualizar los datos con un diagrama de caja usando `ggplot2::geom_boxplot()`.
2. Determinar outliers utilizando el rango intercuartil (IQR).
3. Aplicar estrategias para manejarlos: eliminación, transformación o imputación.
4. Analizar el impacto de cada estrategia en el dataset.

### Cuestionario de Evaluación

1. ¿Cómo se detectan valores atípicos en un conjunto de datos?
  - a) Usando diagramas de caja y la técnica del rango intercuartil

- b) Eliminando cualquier dato que parezca extraño
  - c) Usando solo la media y la desviación estándar
2. ¿Cuál de los siguientes métodos es adecuado para visualizar outliers?
- a) Gráfico de barras
  - b) Diagrama de caja
  - c) Histograma
3. ¿Cuál es una estrategia válida para manejar valores atípicos?
- a) Siempre eliminarlos
  - b) Analizar su impacto y considerar imputaciones o transformaciones
  - c) Ignorarlos y proceder con el análisis

**Problema 10:** Se han recopilado respuestas de una encuesta donde las variables son de tipo categórico (por ejemplo, satisfacción del cliente: "baja", "media", "alta"). Se requiere convertir estos datos en formato numérico para análisis estadístico.

### Tareas

1. Convertir variables categóricas en factores con `as.factor()`.
2. Aplicar codificación one-hot con `model.matrix()`.
3. Evaluar cómo estas transformaciones impactan en modelos de regresión.

### Cuestionario de Evaluación

1. ¿Por qué es importante codificar variables categóricas en modelos predictivos?
  - a) Porque los modelos estadísticos requieren datos numéricos
  - b) Porque es obligatorio para todas las variables
  - c) No es necesario codificarlas
2. ¿Qué técnica de codificación crea múltiples columnas binarias?
  - a) One-hot encoding
  - b) Label encoding
  - c) Scaling
3. ¿Qué función permite convertir una variable categórica en un factor en R?
  - a) `as.factor()`
  - b) `convert()`
  - c) `factorize()`



## Problema 1

Cargar el dataset

```
data <- read.csv("data.csv")
```

Manejo de valores faltantes.

```
print(is.na(data))
```

```
print(summary(data))
```

#Media.

```
for (col in colnames(data)) {  
  if (any(is.na(data[[col]]))) {  
    if (is.numeric(data[[col]])) {  
      media[[col]] <- ifelse(is.na(data[[col]]), median(col), data[[col]])  
    } else {
```

Moda

```
tabla_freq <- table(data[[col]])  
moda_col <- names(tabla_freq[which.max(tabla_freq)])  
data[[col]] <- ifelse(is.na(data[[col]]), moda_col, data[[col]])
```

Eliminar Valores

```
data_sin_duplicados <- distinct(data) #Fusion distinct
```

Estandarizar formatos

```
if ("nombre_producto" %in% colnames(data)) {  
  data$nombre_producto <- tolower(trimws(data$nombre_producto))
```





☐ LUN ☐ MAR ☐ MIE ☐ JUE ☐ VIE ☐ SAB ☐ DOM

FECHA:

## Problema 2.

Conversión de variables

```
red$typeInteracción <- as.factor(red$typeInteracción)
red$plataforma <- as.factor(red$plataforma)
red$Nombre de usuario <- as.factor(red$NombreUsuario)
red$fecha <- as.Date(red$fecha, format = "%d/%m/%Y")
red <- red[, !name(red) %in% "Unnamed: 7"]

normalizar_minmax <- function(x) {
  return((x - min(x, na.rm = TRUE)) / (max(x, na.rm = TRUE) -
    min(x, na.rm = TRUE)))
}
red$numeroInteracciones <- normalizar_minmax(red$numeroInteracciones)
summary(red$numeroInteracciones)
```

Fechas derivadas

```
red$díaSemana <- weekdays(red$fecha)
red$mes <- format(red$fecha, "%Y-%m")
red$anio <- format(red$fecha, "%Y")
red$díaDesdePrimera <- as.numeric(red$fecha, na.rm = TRUE))

if (sum(!is.na(red$numeroInteracciones)) > 0) {
  red$numeroInteraccionesNom <- normalizar_minmax(red$numInt)
}
```

Convertir fechas

```
if (!inherits(red$fecha, "Date")) {
  red$fecha <- as.Date(red$fecha, tryFormats = c("%d/%m/%Y",
    "%Y-%m-%d", "%m/%d/%Y"))
}
```

### Problema 3

Librería

library(dplyr)

Unir datasets

```
fusionado <- merge(clientes, compras, by = "id_cliente", all = TRUE)
```

Verificar datos duplicados

```
duplicados <- fusionado %>% group_by(id_cliente) %>% filter(n()  
> 1)
```

```
print(duplicados)
```

Consultar resumen

```
resumen <- fusionado %>%
```

```
group_by(id_cliente) %>%
```

```
summarise(total_compras = n(),
```

```
monto_total = sum(monto, na.rm = TRUE))
```

### Problema 4

```
transacciones <- data.frame()
```

```
Monto = c(rnorm(50, mean = 1000, sd = 200);
```

```
5000, 5500, 6000
```

```
200, 150)
```

```
ggplot(transacciones, aes(x = Monto)) +
```

```
geom_boxplot(fill = "skyblue", color = "black") +
```

```
labs(title = "Diagrama de caja",
```

```
x = "Monto de transacción") +
```

```
theme_minimal()
```

Rango Inter cuartil

```
Q1 <- quantile(transacciones$Monto, 0.25)
```

```
Q3 <- quantile(transacciones$Monto, 0.75)
```

```
IQR <- Q3 - Q1
```





☐ DOM ☐ MAR ☐ ABR ☐ MAY ☐ JUN ☐ JUL ☐ AGO ☐ SEPT ☐ OCT ☐ NOV ☐ DIC

FECHA:

limite inferior  $L = Q1 - 1.5 * IQR$   
limite superior  $= Q3 + 1.5 * IQR$

Outliers

Outliers  $\leftarrow$  transacciones \$ monto [transacciones \$ monto < limite inferior | transacciones \$ monto > limite superior]

print(outliers)

Manejar Outliers

transacciones - filtrados  $\leftarrow$  transacciones [transacciones \$ monto > limite inferior & transacciones \$ monto <= limite superior]

transformaciones

## Practica 5

Codificación One Hot

```
encuesta$id <- 1:nrow(encuesta)
formula_one_hot <- as.formula(paste("n", paste(colnames(encuesta)
[apply(encuesta, 1, function(x) collapse = "f", "-1"))])
one_hot_encoded <- model.matrix(formula_one_hot, data = encuesta)
one_hot_encoded_df <- as.data.frame(one_hot_encoded)
encuesta_codificada_one_hot <- merge(encuesta, one_hot_encoded_df,
by.x = "id", by.y = "row.names")
encuesta_codificada_one_hot$hotid <- NULL
```

```
if(!is.null(encuesta_codificada_one_hot$variable_numerica)) {
  formula_label <- as.formula(paste("variable_numerica", paste(colnames(
encuesta_codificada_one_hot) [apply(encuesta_codificada_one_hot, 1,
function(x) collapse = "f", "-1"))])
  modelo_label <- lm(formula_label, data = encuesta_codificada_one_hot)
  print("Resumen de modelo")
```



## Practica 6

```
hospital <- read.csv("hospital.csv")
```

```
sum(is.na(hospital))
```

```
colSums(is.na(hospital))
```

```
hospital_limpio_nacait <- na.omit(hospital)
```

```
hospital_imputado_medica <- hospital[is.na(hospital) == FALSE, ]
```

```
Medica_Presion_arterial = replace_na(Presion_arterial, na.rm = TRUE)
```

```
Glucosa = replace_na(Glucosa, mean(Glucosa, na.rm = TRUE))
```

## Practica 7

```
empresa <- read.csv("empresa.csv")
```

```
columnas_numericas <- datos[, c("Ingresos", "Utilidades Neto", "ROE",  
"Liquidez", "Endeudamiento", "premio")]
```

```
datos_estandarizados <- as.data.frame(scale(columnas_numericas))
```

Normalización

$$\frac{(x - \min(x))}{\max(x) - \min(x)}$$
,  $normalize \leftarrow función(x)$

$$\frac{(x - \min(x, na.rm = TRUE))}{(\max(x, na.rm = TRUE) - \min(x, na.rm = TRUE))}$$

$$\min(x, na.rm = TRUE)$$

```
datos_normalizados <- data.frame(apply
```

Evaluación

```
resumen_estandarizado <- summary(datos_estandarizados)
```





☐ LUN ☐ MAR ☐ MIÉ ☐ JUE ☐ VIE ☐ SAB ☐ DOM

FECHA:

## Practica 8

```
Clientes <- read.csv("Clientes.csv")
```

```
fusionado <- left_join(Clientes, compres, by = "id_cliente")
```

```
fusionado <- fusionado %>% distinct()
```

```
duplicados <- fusionado %>% group_by(id_cliente) %>% filter(n() > 1)
```

```
resumen <- fusionado %>%
```

```
group_by(id_cliente) %>%
```

```
summarize(monto_tot = sum(monto, na.rm = TRUE))
```

## Practica 9

```
set.seed(123)
```

```
Produccion <- data.frame(
```

```
  Unidades = rnorm(50, mean = 500, sd = 50)
```

```
  800, 850, 900
```

```
  200, 150)
```

```
ggplot(Produccion, aes(x = Unidades)) +
```

```
geom_boxplot(aes(fill = "lightblue", color = "black")) +
```

```
labs(title = "Diagrama de caja de produccion")
```

```
Q1 <- quantile(Produccion$Unidades, 0.25)
```

```
Q2 <- quantile(Produccion$Unidades, 0.75)
```

```
IQR <- Q3 - Q1
```

```
Produccion = sin_outliers <- Produccion[Produccion$Unidades >=
```

```
limite_inferior & Produccion$Unidades <= limite_superior, ]
```



☐ LUN ☐ MAR ☐ MIE ☐ JUE ☐ VIE ☐ SAB ☐ DOM

FECHA:

## Practica 10

`Data$ Satisfaccion <- as.factor (data$ Satisfaccion)`

`Data$ Servicio <- as.factor (data$ Servicio)`

`One_hot_encoded <- model.matrix (v Satisfaccion + Servicio ~ 1,  
data = data)`

`Data$ Score <- rnorm (10, mean = 50, sd = 10)`

`Modelo <- lm (score ~., data = as.data.frame (One_hot_  
encoded)) summary (Modelo)`