Trabajo Práctico 1: Análisis Inteligente de Datos

1. Temas a Evaluar

- Clasificación y exploración de datos.
- Técnicas de análisis univariado.
- Análisis descriptivo de datos estructurados (EDA).
- Visualización de la información y datos univariados.
- Análisis previo de los datos: detección de valores perdidos y casos atípicos.
- Muestreo.
- Introducción a técnicas de remuestreo: permutaciones y bootstrapping.
- División en entrenamiento y prueba.

2. Datos

El dataset a utilizar se encuentra en la siguiente dirección:

`https://www.kaggle.com/datasets/ahmedmohamed2003/cafe-sales-dirty-data-for-cleaning-training/data`

Se debe:

1. Descargar el archivo CSV comprimido (ZIP).

2. Descomprimir la carpeta.

3. Ubicar el archivo CSV en la misma carpeta en la que se encuentre su notebook.

2.1. Carga del Dataset en R

Utilice un notebook de R y ejecute el siguiente código para instalar el paquete `readr` y cargar el dataset:

```
install.packages("readr")  # Para leer archivos CSV
library(readr)
```

Luego, cargue el archivo CSV utilizando:

```
file_path <- "dirty_cafe_sales.csv"

cafe_sales_df <- read_csv(file_path)

head(cafe_sales_df)
```

3. Parte 1: Análisis Exploratorio Inicial

Realice el análisis de las columnas `Item`, `Quantity`, `Payment Method` y `Location` considerando lo siguiente:

a) Visualizar los valores presentes en cada columna, identificando la existencia de valores nulos y cadenas que indiquen errores (por ejemplo, `ERROR` o `UNKNOWN`).

b) Determinar los valores más frecuentes mediante representaciones gráficas (por ejemplo, diagramas de barras, diagramas circulares, etc.).

c) Contabilizar la cantidad de entradas que presentan valores nulos o cadenas `ERROR`/`UNKNOWN` para cada columna.

Adicionalmente, responda:

1. ¿Cuál es el `Item` (Coffee, Salad, Cake, etc.) más frecuente en la tabla?

2. ¿Cuál es la `Quantity` (cantidad de items pedidos) menos frecuente?

3. ¿Cuántas personas realizaron el pago con efectivo (`cash`)?

4. ¿Cuántos registros presentan `UNKNOWN`, `NA` o `ERROR` en la columna `Location`? Determine también la proporción que estos valores representan respecto al total de filas del dataset.

5. Eliminando todos los valores no numéricos de la columna `Price Per Unit`, determine el promedio, la mediana y la desviación estándar.

4. Parte 2: Análisis de Outliers y Patrones Temporales

Utilizando el dataset original sin modificaciones, realice lo siguiente:

a) Determine si existe algún outlier en alguna columna y justifique su respuesta.

b) Conociendo que el 1 de enero de 2023 fue domingo, responda:

   i. ¿Cuál es el día de la semana con menos transacciones?

   ii. ¿Cuál es el día de la semana con más transacciones?

   iii. ¿Cuántas transacciones se realizaron los viernes de ese año?

c) Identifique el mes con más transacciones y el mes con menos.

d) Genere diagramas de barras ordenados cronológicamente que representen estos resultados, y determine si se observa una tendencia evidente.

5. Parte 3: Completar y Corregir Datos

Con base en el dataset original sin cambios, realice las siguientes operaciones:

a) Se cumple la propiedad:

$$\texttt{Quantity} \times \texttt{Price Per Unit} = \texttt{Total Spent}.$$

   Dado que cada `Item` tiene un precio constante durante el año, se proporciona el siguiente *menú*:

Cake	3
Juice	3
Coffee	2
Cookie	1
Salad	5
Sandwich	4
Smoothie	4
Tea	2

Complete la columna `Price Per Unit` de tal manera que, en cada caso en que aparezca un `NA`, `ERROR` o `UNKNOWN`, se asigne el valor correspondiente al `Item` según el menú. Si el `Item` también es `NA`, `ERROR` o `UNKNOWN`, deje el valor como `NA`.

b) Complete las columnas `Quantity`, `Price Per Unit` y `Total Spent` en aquellos casos en que falte solamente una de estas tres variables, utilizando la relación entre ellas. En caso de imposibilidad de completar el valor, deje `NA`.

c) Complete la columna `Item` tomando como referencia el valor de `Price Per Unit`. En particular, si `Price Per Unit` es 3, se asume que se compró una `Cake` (y no un `Juice`); de igual forma, si es 4, se asume que se compró un `Sandwich` (y no un `Smoothie`). Si no es posible determinarlo, deje el valor como `NA`.

d) Como verificación ("sanity check"), tras realizar los tres procesos de completado secuencial, se debe obtener:

- 120 `NA` en la columna `Item`.
- 6 `NA` en la columna `Price Per Unit`.
- 23 `NA` en la columna `Quantity`.
- 23 `NA` en la columna `Total Spent`.

e) Rellene las siguientes columnas: `Item`, `Price Per Unit`, `Quantity`, `Total Spent`, `Payment Method`, `Location` y `Transaction Date` utilizando la moda (valor más frecuente, excluyendo `NA`, `ERROR` y `UNKNOWN`) de cada columna respectiva. Esto implica reemplazar no solo los `NA`, sino también los valores `ERROR` y `UNKNOWN`.

Al finalizar este proceso, su dataset no deberá contener ningún valor nulo, ni `ERROR` ni `UNKNOWN`.

6. Parte 4: Análisis con Bootstrapping y Visualización

Utilice el dataset corregido de la Parte 3 (sin valores faltantes) para realizar lo siguiente:

a) Establezca la semilla de aleatoriedad en 123.

b) Emplee técnicas de bootstrapping para generar un gráfico que muestre la distribución de la media de `Total Spent`. Realice 1000 resampleos con reemplazo de la totalidad de los datos.

c) Divida el dataset en tres bloques sin repetición, de tamaños 8000, 1000 y 1000, y compare estadísticos (a elección) de alguna columna.

d) Utilice `ggplot` para visualizar un aspecto del dataset que considere interesante.