Trabajo Práctico 2: Análisis de Datos Móviles

1. Datos

Descargar datos del siguiente link:

`https://www.kaggle.com/code/yadrsv/eda-mobiles-dataset/input`

Bajar el archivo ZIP, extraer el CSV y colocarlo en el mismo directorio de la notebook de R.

2. Limpieza y Estandarización de Datos

Analizar y estandarizar el formato de columnas numéricas:

- Remover caracteres no numéricos.
- Renombrar columnas especificando sus unidades.

Ejemplo:

- De: `Launched Price (USA)` con valores como `USD 499`.
- A: `Launched Price (USA)_USD` con valores numéricos (499).

Aplicar lo anterior a las siguientes columnas: `Screen Size`, `Battery Capacity`, `Back Camera`, `Front Camera`, `RAM`, `Mobile Weight` y `Launched Price (USA)`.

Consejo: Crear una función general reutilizable.

3. Cálculo del Tamaño de Muestra

(**Ejercicio sin dataset**)
Se cree que el brillo promedio de celulares es 500 nits pero se quiere hacer un test estadístico como para confirmar lo que ya se cree.

a) Plantear las hipótesis H_0 y H_1.

b) ¿La hipótesis alternativa es unilateral o bilateral?

c) Antes de calcular el tamaño muestral, indicar cuáles son los errores tipo I (α) y tipo II (β) de nuestro experimento, a partir de las siguientes condiciones:

- Se desea detectar diferencias de ± 100 nits (o más) el 99 % de las veces que exista una diferencia real δ tan grande.
- Se quiere que, si el experimento se repitiese muchas veces, la proporción de veces que se rechaza la hipótesis nula bajo la hipótesis nula sea igual al 1 %.

d) Calcular el tamaño muestral usando:

$$\sigma = \frac{\delta \cdot \sqrt{n}}{z_{1-\alpha/2} + z_{1-\beta}}$$

sabiendo que el desvío muestral es $\sigma = 175$ (conocido exactamente). Indicar los valores de z y el resultado obtenido.

e) Recalcular el tamaño muestral utilizando R. Para ello, primero instalar y cargar el paquete `pwr`:

```
1 install.packages("pwr")
2 library(pwr)
```

Luego utilizar la función `pwr.norm.test()`, completando los parámetros adecuados:

```
1 pwr.norm.test(d = , sig.level = , power = , alternative =
      )
```

Donde:

- `d` es el efecto estandarizado, calculado como $d = \delta/\sigma$.
- `sig.level` es el nivel de significancia del test (α).
- `power` es la potencia del test ($1 - \beta$).
- `alternative` indica el tipo de hipótesis alternativa:
  - `"two.sided"` para una hipótesis bilateral ($H_1 : \mu \neq \mu_0$),
  - `"greater"` para una hipótesis unilateral derecha ($H_1 : \mu > \mu_0$),
  - `"less"` para una hipótesis unilateral izquierda ($H_1 : \mu < \mu_0$).

Indicar qué hacer si el resultado obtenido no es un número entero.

4. Estadísticas Descriptivas y Visualizaciones

Estadísticas descriptivas

Para las columnas limpias:

- Calcular: media, mediana, mínimo, máximo y desviación estándar.
- Identificar outliers con algun metodo no visual (Tests, Z-Scores, IQR, lo que quieran).
- Comentar si encuentran relaciones o correlaciones fuertes entre variables. (las van a usar más adelante)

Visualizaciones

- Realizar histogramas de al menos dos variables y describir sus formas (simétricas, sesgadas, multimodales, etc.).
- Generar boxplots para detectar outliers y, en caso de encontrar un outlier en la columna de precio USD, remover esa fila. Además, comparar agrupaciones por `Brand` o `Model` cuando sea posible.

Sugerencia: Utilizar las funciones `summary()`, `hist()`, `boxplot()` y la librería `ggplot2`.

5. Test de Normalidad

Seleccionar 2 columnas a eleccion y aplicar el test de Shapiro-Wilk (`shapiro.test`). Realizar un Q-Q plot para la visualización y, finalmente, interpretar los resultados (p-value, aceptación o rechazo de normalidad con $\alpha = 5\,\%$).

6. Correlación y Significancia

Elegir dos columnas relacionadas, por ejemplo:

- `Launched Price (USA)_USD` y `RAM_GB`, o
- `Battery Capacity_mAh` y `Launched Price (USA)_USD`.

- Realizar un diagrama de dispersión para observar la relación entre las variables.
- Calcular la correlación de Pearson utilizando `cor.test`.
- Reportar el valor de r, el p-value, la significancia ($\alpha = 5\,\%$) y la fuerza de la correlación.

7. Test t para una Muestra

Plantear una hipótesis sobre la media de una variable numérica, por ejemplo: `Mobile Weight_g`:

$$H_0 : \mu \leq 180 \text{ g} \quad ; \quad H_1 : \mu > 180 \text{ g}$$

Aplicar el test t:

```
1 t.test(variable, mu=180, alternative="greater")
```

Interpretar los resultados obtenidos (valor t, p-value y si se rechaza H_0 con $\alpha = 5\,\%$).

8. Test Chi-cuadrado de Independencia

- Crear una variable categórica denominada `High Battery`:
  - Asignar el valor 'Alta' si `Battery Capacity_mAh` es mayor o igual a la mediana.
  - Asignar "Baja"si es menor.
- Seleccionar una segunda variable categórica existente (por ejemplo, `Brand` o `Model`).
- Construir una tabla de contingencia y aplicar la función `chisq.test()`.
- Visualizar los resultados mediante un barplot apilado utilizando `geom_bar(position="fill")`.

9. Bootstrap para Evaluar la Incertidumbre del IQR

Supongamos que queremos estimar cuán dispersos son los precios de lanzamiento de los celulares en EEUU. Para ello, usaremos la columna `Launched Price (USA)`.

El objetivo es:

- Estimar el IQR (rango intercuartílico) de los precios de lanzamiento.
- Evaluar cuán seguros estamos de ese valor estimado usando bootstrap.

a) Calcular el IQR directamente sobre todos los datos disponibles en la columna `Launched Price (USA)`.

b) Realizar un bootstrap no paramétrico para obtener una distribución de valores posibles del IQR. Hacer minimo 10000 sampleos.
   - Tomar muchas muestras con reemplazo del mismo tamaño que el original.
   - Calcular el IQR de cada muestra.
   - Guardar todos los valores obtenidos.

(Pista: podés usar un for o la función `replicate()` lo que te sea mas comodo).

c) Calcular un intervalo de confianza del 95 % para el IQR utilizando los percentiles 2.5 y 97.5 de la distribución bootstrap obtenida.

d) Graficar un histograma de la distribución bootstrap del IQR e indicar en el gráfico los extremos del intervalo de confianza.

10. Distancias Promedio

Seleccionar estas 2 columnas del dataset: `Back Camera_MP` y `Front Camera_MP` y estandarizarlas en caso de que sigan en su escala original.

a) Calcular la **distancia promedio** entre todas las filas para esas columnas usando la Distancia Euclídea.

b) Repetir el mismo procedimiento para las combinaciones:

- `Back Camera_MP` y `Mobile Weight_g`.
- `Front Camera_MP` y `Mobile Weight_g`.