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

Se cree que el brillo promedio de celulares es 500 nits ($\sigma = 175$, conocida exactamente) y se desea detectar diferencias de ± 100 nits. Se plantean las siguientes condiciones:

- Error tipo I ($\alpha < 5\,\%$).
- Error tipo II ($\beta < 1\,\%$).

a) Plantear las hipótesis H_0 y H_1.

b) ¿La hipótesis alternativa es unilateral o bilateral?

c) Calcular el tamaño muestral usando:

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

Indicar los valores de z y el resultado obtenido.

d) Recalcular utilizando la función R:

```
power.t.test(delta = , sd = , sig.level = , power = ,
   type = "one.sample", alternative = )
```

Indicar qué hacer si el resultado obtenido no es entero.

4. Estadísticas Descriptivas y Visualizaciones

Estadísticas descriptivas

Para las columnas limpias:

- Calcular: media, mediana, mínimo, máximo y desviación estándar.
- Identificar outliers visualmente (por ejemplo, analizando rangos grandes).
- Comentar hallazgos relacionados con la homogeneidad y la variabilidad.

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 una o dos columnas (por ejemplo, `Launched Price (USA)_USD` y `Battery Capacity_mAh`) 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:

```
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. Más Ejercicios de Tamaño de Muestra

I. **Diferencia detectada:** Utilizar:

```
power.t.test(delta=0.5, sd=1, power=0.9, sig.level=0.05,
   alternative="two.sided", type="one.sample")
```

¿Qué tamaño muestral se requiere?

II. **Variar α:** Discutir qué ocurre con el tamaño muestral si se varía el valor de α.

III. **Discusión:** Analizar el efecto en β cuando α cambia y n se mantiene fijo; además, discutir cómo influye un aumento en n.

10. Bootstrap para Intervalos de Confianza

Para una columna numérica positiva (por ejemplo, `Launched Price (USA)_USD`):

a) Emplear el método Bootstrap (no paramétrico) para generar la distribución Bootstrap de la media.

b) Calcular el intervalo de confianza al 95 % (utilizando percentiles).

c) Graficar un histograma que incluya los límites del intervalo de confianza.