Nombre: Yan Carlos Rodríguez Ospinos

ID: U00106308

Correo: yrodriguez143@unab.edu.co

INFORME PRÁCTICA 1 – ESTADÍSTICA DESCRIPTIVA

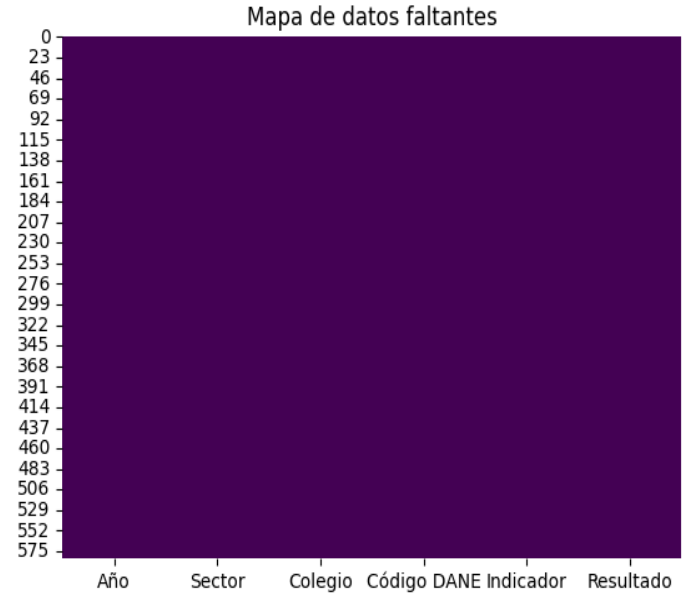
Base de datos: ResultadosSabanetaSaber11.csv

1. Se realiza el cargue de la base de datos y se revisa a las columnas a estudiar y se identifica las siguientes variables y tipos de variables:

Variable	Tipo de Variable	Dtype
Año	Categórica	object
Sector	Categórica	object
Colegio	Categórica	object
Código DANE	Categórica	object
Indicador	Categórica	object
Resultado	Numéricas	int64

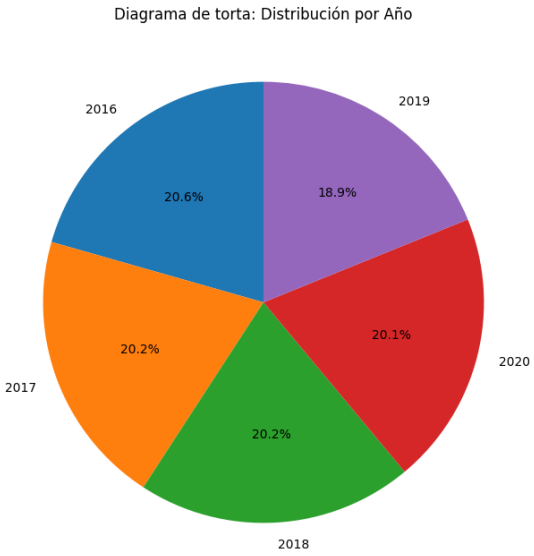
2. Se verifica que no haya datos faltantes

```
Valores faltantes por columna:
Año            0
Sector         0
Colegio        0
Código DANE    0
Indicador      0
Resultado      0
dtype: int64
```

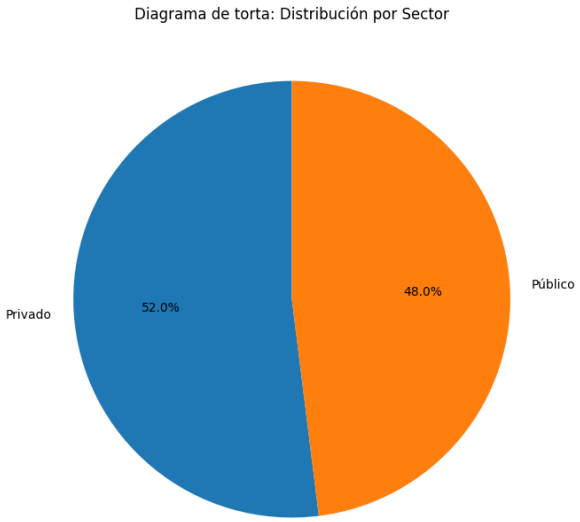


3. Se realizan diagramas de barra y torta para los datos tipo categórico, y se eligen los siguientes diagramas:

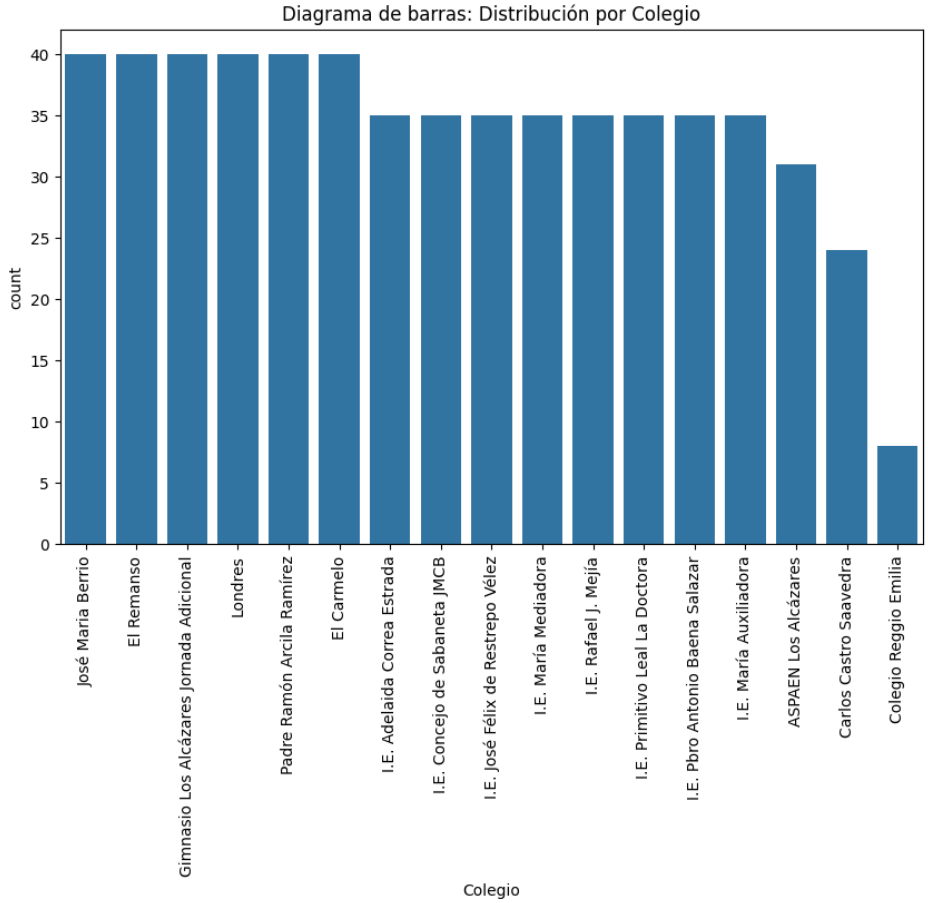
Año:



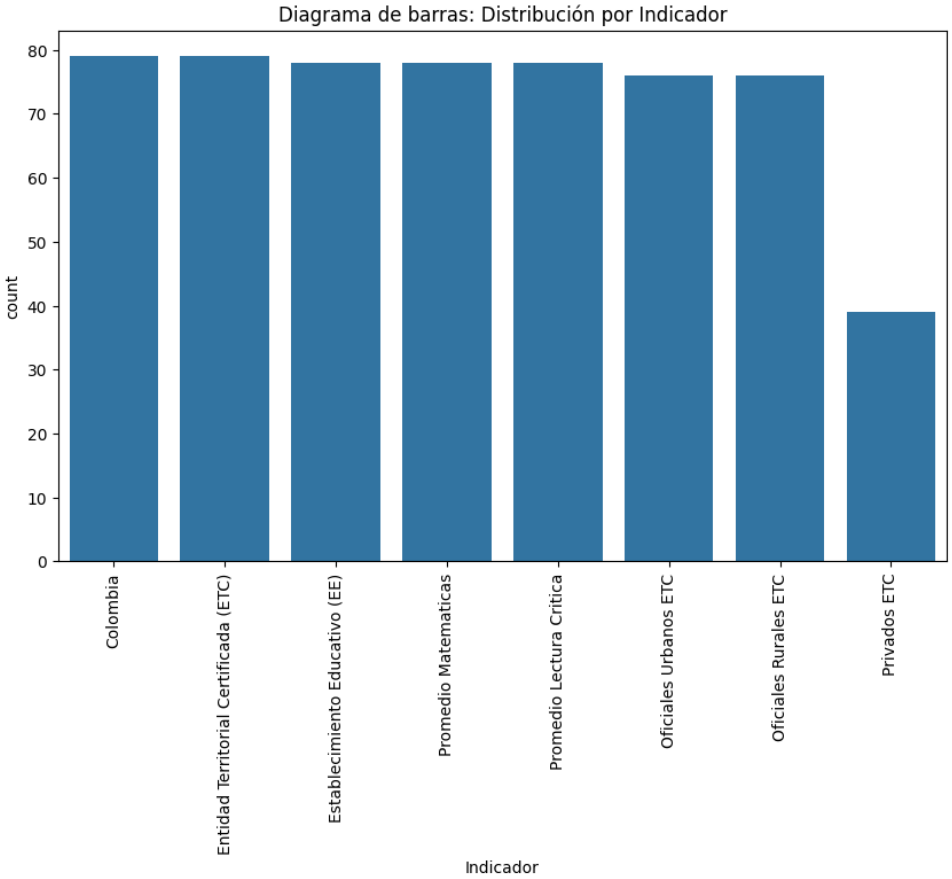
Sector:



Colegio:

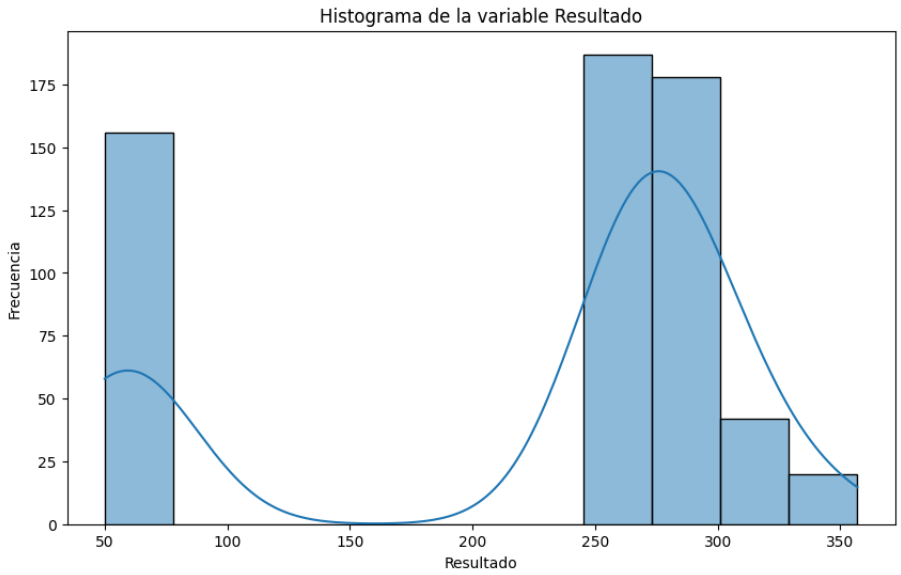


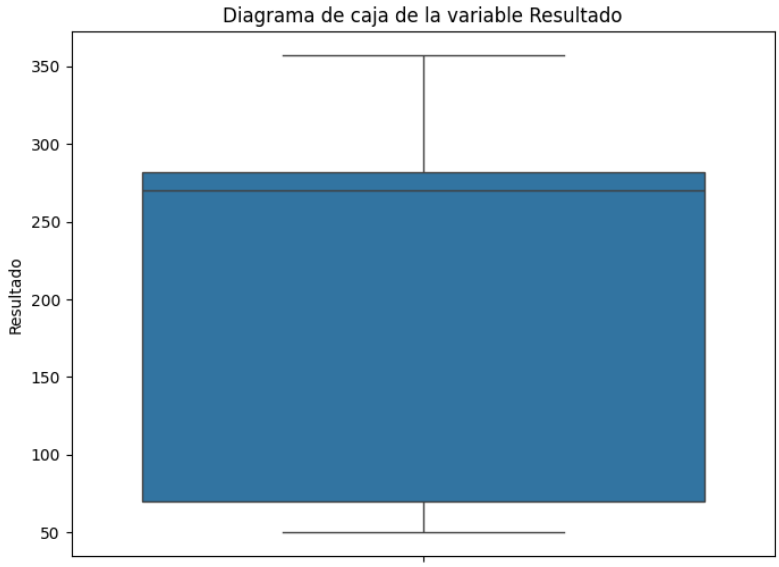
Indicador:



Para la variable Código Dane ya que corresponde a un dato relacionado directamente al Colegio donde se realizó la prueba.

4. Se realiza el histograma y diagrama de caja para los datos “Resultados”

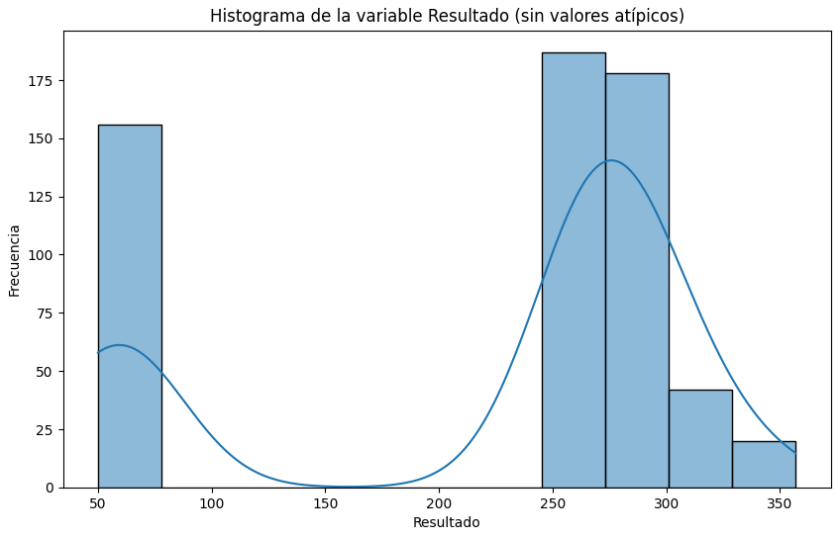


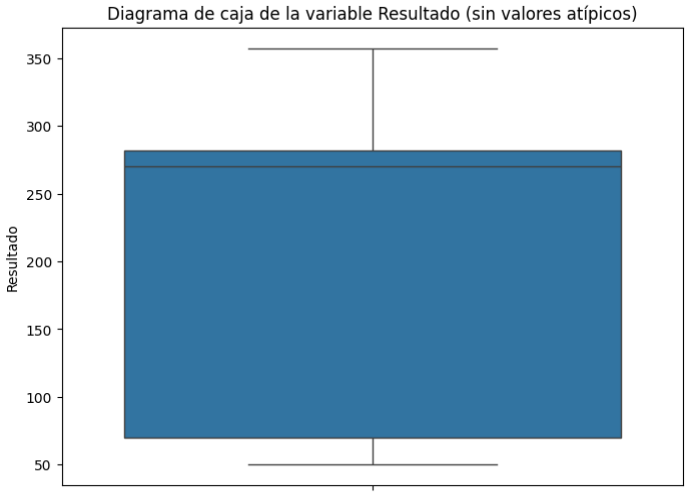


5. Se aplica rango intercuartílico para verificar y eliminar los datos atípicos para las variables numéricas. Se genera un nuevo df con los datos ya verificados (sin atípicos)
   Se genera nuevamente el histograma y diagrama de caja, y se evidencia que no se encontraron datos atípicos, ya que no hubo cambios en la visualización de los datos.

```
[15] # Histograma para la variable 'Resultado' sin valores atípicos
     plt.figure(figsize=(10, 6))
     sns.histplot(data=df_no_outliers, x='Resultado', kde=True)
     plt.title('Histograma de la variable Resultado (sin valores atípicos)')
     plt.xlabel('Resultado')
     plt.ylabel('Frecuencia')
     plt.show()

     # Diagrama de caja para la variable 'Resultado' sin valores atípicos
     plt.figure(figsize=(8, 6))
     sns.boxplot(data=df_no_outliers, y='Resultado')
     plt.title('Diagrama de caja de la variable Resultado (sin valores atípicos)')
     plt.ylabel('Resultado')
     plt.show()
```



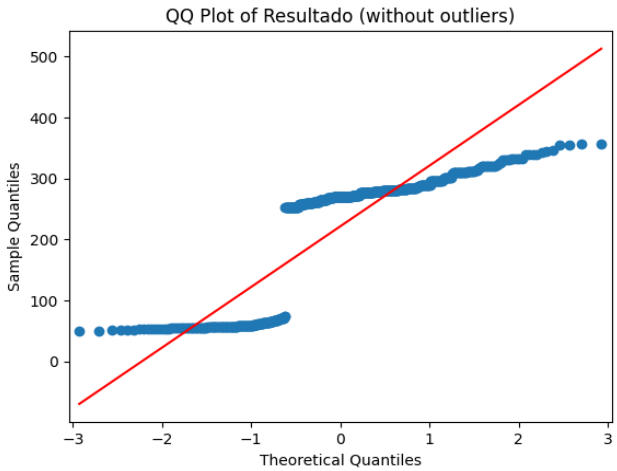


6. Se ejecutan los test de normalidad a las variables numéricas, es decir, se genera los test para la variable "resultado" ya que esta se considera como una variable numérica continua, estos test se realizan con los valores ya definidos en df_no_outliers. Los test son los siguientes:

    - Shapiro-Wilk
    - Kolmogorov-Smirnov
    - Anderson-Darling
    - Jarque-Bera

```
Shapiro-Wilk Test: Statistic=0.7180, p-value=0.0000
Kolmogorov-Smirnov Test: Statistic=1.0000, p-value=0.0000
Anderson-Darling Test: Statistic=80.5745
  At 15.0% significance level, the data does not look normal (critical value 0.5720)
  At 10.0% significance level, the data does not look normal (critical value 0.6520)
  At 5.0% significance level, the data does not look normal (critical value 0.7820)
  At 2.5% significance level, the data does not look normal (critical value 0.9120)
  At 1.0% significance level, the data does not look normal (critical value 1.0850)
Jarque-Bera Test: Statistic=105.4944, p-value=0.0000
```

Se generan los gráficos QQ.



CONCLUSIONES

- En la verificación inicial no visualizaron valores faltantes en los datos, lo cual facilitó el análisis de los datos.
- Se realizaron gráficos de barras y de torta a las variables categóricas (Año, Sector, Colegio, Código DANE e Indicador) donde muestran la distribución de las entradas dentro de cada categoría. Esto nos da una idea de la composición de los datos.
- El histograma y el diagrama de caja de la variable Resultado nos muestra la distribución de esta variable numérica y nos ayuda a visualizar la presencia de posibles valores atípicos.
- Se identificó y mostró el DataFrame con y sin valores atípicos en la columna numérica “Resultado”. Se realizaron los test de normalidad: Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling y Jarque-Bera, el cual nos indican que la variable "Resultado" (incluso sin valores atípicos) no se distribuye normalmente, ya que los valores p son inferiores al nivel de significancia típico de 0,05 y el estadístico de Anderson-Darling es superior a los valores críticos.
- Si los datos se distribuyeran normalmente, los puntos del gráfico cuantitativo se encontrarán cerca de la línea recta roja. En nuestro gráfico cuantitativo, los puntos se desvían significativamente de la línea, especialmente en los extremos. Esta desviación visual respalda firmemente la conclusión de los test de estadísticas: la variable "Resultado" no sigue una distribución normal.