Tarea:

1. Cargar la base de datos, revisar columnas a estudiar.
2. Realizar verificación de datos faltantes y eliminar si es necesario.
3. Para las variables categóricas realizar diagramas de torta y barras.
4. Para las variables numéricas histogramas y diagramas de caja.
5. Usar rango intercuartílico para verificar y eliminar datos atípicos. Generar nuevo df con los datos sin atípicos.
6. Generar histogramas para datos sin outliers.
7. Ejecutar los siguientes test de normalidad a las variables numéricas continuas: Shapiro-Wilk, Kolmogorov, Anderson-Darling y Jarque-Bera. Generar gráficos QQ.
8. Elaborar un informe con sus resultados y súbalo junto con el .ipynb a su github personal.

PRACTICA 01 ESTADISTICA DESCRIPTIVA

ANDREA CAROLINA BLANCO GUEVARA

UNIVERSIDAD AUTONOMA DE BUCARAMANGA

INGENIERÍA INDUSTRIAL

BUCARAMANGA, SANTANDER

COLOMBIA

2025

Análisis Estadístico de los Resultados de las Pruebas Saber 11 en Sabaneta

El presente informe tiene como objetivo realizar un análisis estadístico descriptivo de los resultados obtenidos por estudiantes del municipio de Sabaneta en las pruebas Saber 11. La información analizada proviene de una base de datos que contiene tanto las calificaciones en diferentes áreas de la evaluación como datos sociodemográficos de los estudiantes.

Para el desarrollo del análisis, se empleó el entorno Google Colab y el lenguaje de programación Python, utilizando librerías especializadas en el manejo y visualización de datos. El proceso incluyó la carga y exploración de la base de datos, la verificación de valores faltantes, y la elaboración de representaciones gráficas como histogramas, diagramas de caja, diagramas de torta y de barras.

Este estudio busca ofrecer una visión clara de la distribución de los resultados y de las características más relevantes de la población evaluada, sirviendo como base para futuras investigaciones o decisiones educativas.

Metodología

La información utilizada en este estudio proviene del archivo denominado *ResultadosSabanetaSaber11.csv*, el cual contiene los resultados obtenidos por estudiantes del municipio de Sabaneta en las pruebas Saber 11, así como variables sociodemográficas relacionadas. La base de datos incluye puntajes en Promedio Matemáticas, Promedio Lectura crítica, Colombia, Establecimiento educativo, entidad territorial, oficiales urbanos, rurales, dados en indicador.

El análisis se llevó a cabo en el entorno de programación Google Colab utilizando el lenguaje Python. Se emplearon las librerías *pandas* y *numpy* para la manipulación y análisis de datos, así como *matplotlib* y *seaborn* para la elaboración de representaciones gráficas.

El procedimiento inició con la carga del archivo CSV y la verificación de la estructura de la base de datos. Posteriormente, se revisó la existencia de valores faltantes y se realizó la exploración inicial de las variables. Para las variables numéricas, se calcularon medidas de tendencia central y dispersión, y se generaron visualizaciones como histogramas y diagramas de caja. Para las variables categóricas, se elaboraron gráficos de barras que permitieron identificar la distribución de frecuencias.

En caso de detectar datos atípicos, se aplicó el método del rango intercuartílico (IQR) para su verificación y posible exclusión, con el fin de garantizar un análisis más representativo. De esta manera, el procedimiento que se obtuvo fue el siguiente:

Para iniciar, se carga base de datos en Google Colab

Figura 1. *Código para carga y exploración inicial de la base de datos*

1. Carga la base de datos

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Cargar archivo CSV
df = pd.read_csv("ResultadosSabanetaSaber11.csv")

# Información básica
print("Información del dataset")
print(df.info())
print("Primeras filas del dataset")
print(df.head())
print("Resumen estadístico del dataset")
print(df.describe())
```

Python

Figura 2. *Salida.*

```
Información del dataset
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 583 entries, 0 to 582
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Año          583 non-null    int64
 1   Sector       583 non-null    object
 2   Colegio      583 non-null    object
 3   Código DANE  583 non-null    int64
 4   Indicador    583 non-null    object
 5   Resultado    583 non-null    int64
dtypes: int64(3), object(3)
memory usage: 27.5+ KB
None
Primeras filas del dataset
    Año   Sector                     Colegio   Código DANE  \
0  2016  Público  I.E. Adelaida Correa Estrada  105631000050
1  2016  Público  I.E. Adelaida Correa Estrada  105631000050
2  2016  Público  I.E. Adelaida Correa Estrada  105631000050
3  2016  Público  I.E. Adelaida Correa Estrada  105631000050
4  2016  Público  I.E. Adelaida Correa Estrada  105631000050
```

```
                              Indicador  Resultado
0         Establecimiento Educativo (EE)        285
...
25%    2017.000000  1.056310e+11   70.000000
50%    2018.000000  3.050010e+11  270.000000
75%    2019.000000  3.056310e+11  282.000000
max    2020.000000  4.056310e+11  357.000000
```

Nota. Elaboración de lenguaje de programación a partir de la ejecución en Google Colab utilizando Python y las librerías *pandas*, *numpy*, *matplotlib* y *seaborn*. La salida muestra la información general del conjunto de datos, las primeras filas y el resumen estadístico de las variables numéricas.

Detalles y explicación:

La Figura 2 presenta la salida obtenida tras la carga y exploración inicial de la base de datos. El conjunto está conformado por 583 registros correspondientes a instituciones educativas de Sabaneta y 6 variables: Año, Sector, Colegio, Código DANE, Indicador y Resultado. Se verificó que no existen valores faltantes. En las primeras filas se observan registros del año 2016 correspondientes al sector público. El resumen estadístico de las variables numéricas indica que los puntajes en la prueba oscilan entre 70 y 357, con una mediana de 270 puntos.

Primeras filas del dataset muestra los primeros 5 registros, con columnas:

1. **Año**: año de presentación de la prueba.
2. **Sector**: Colegio público o privado.
3. **Colegio**: Nombre de la institución educativa.
4. **Código DANE**: Identificador único del establecimiento.
5. **Indicador**: Categoría de la fila.

Resumen estadístico de variables numéricas (df.describe()). Para Año**,** Código DANE y Resultado, se calculan:

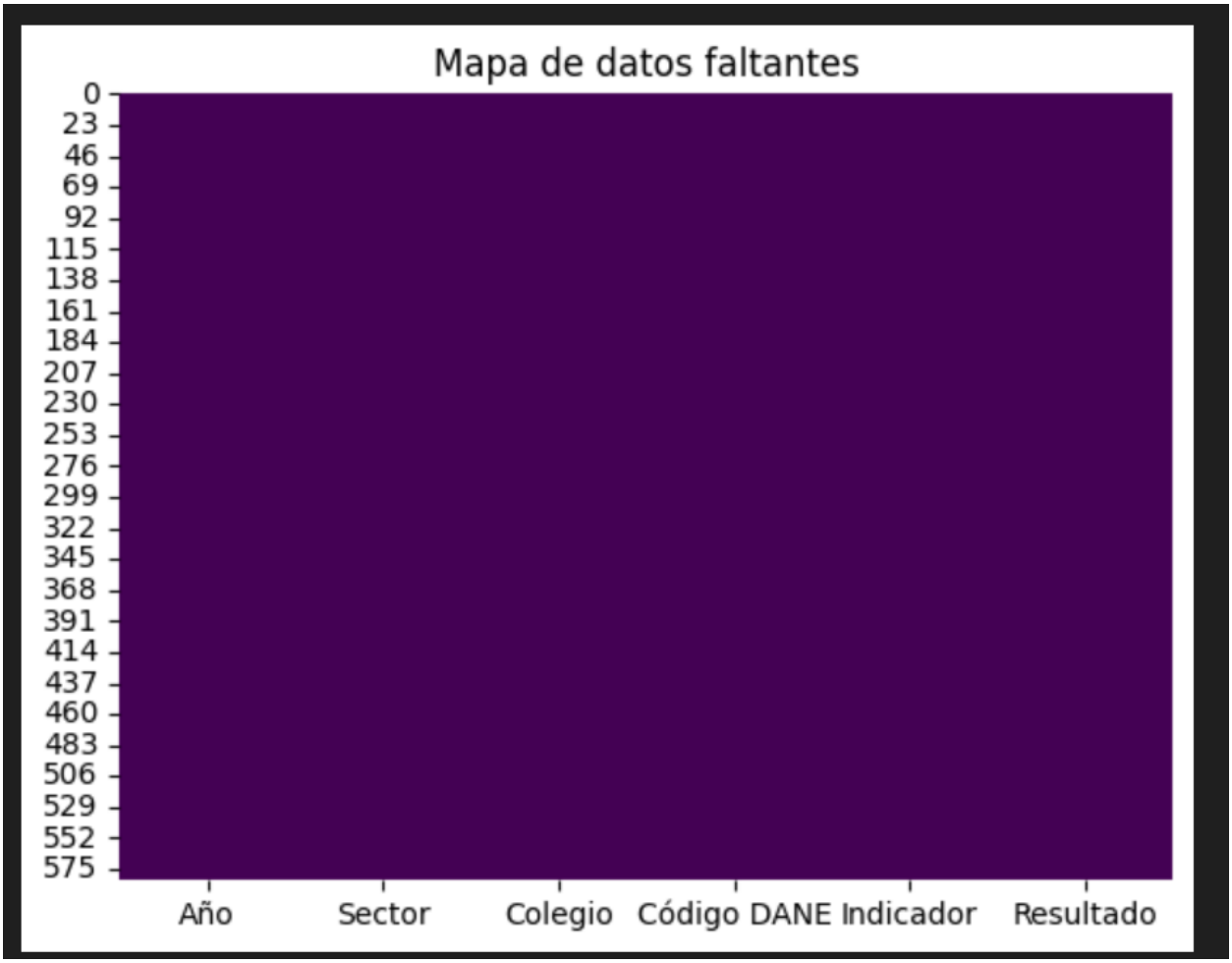
- **count** (cantidad de registros)
- **mean** (promedio)
- **std** (desviación estándar)
- **min** (valor mínimo)
- **25%**, **50%**, **75%** (cuartiles)
- **max** (valor máximo)

Figura 3. Lenguaje de programación para visualizar los valores faltantes.

```
# Valores faltantes
print("\nValores faltantes por columna:")
print(df.isnull().sum())

# Visualización de nulos
sns.heatmap(df.isnull(), cbar=False, cmap="viridis")
plt.title("Mapa de datos faltantes")
plt.show()
```

Se comprobó que no existen valores faltantes.

Figura 4. Salida



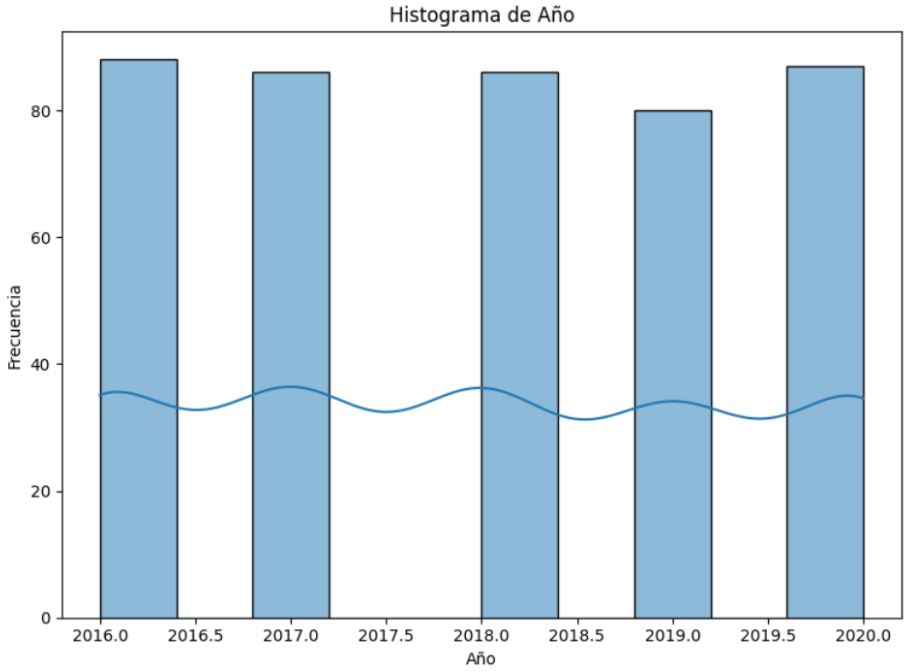
Nota. El mapa da la aclaración de que no hay ningún espacio vació.

Figura 5. Histogramas para variables numéricas.

5.1 *Código para histograma (Año)*

```
import matplotlib.pyplot as plt
import seaborn as sns

# Histograma para Año
plt.figure(figsize=(8, 6))
sns.histplot(df_filtered['Año'], bins=10, kde=True)
plt.xlabel('Año')
plt.ylabel('Frecuencia')
plt.title('Histograma de Año')
plt.tight_layout()
plt.show()
```

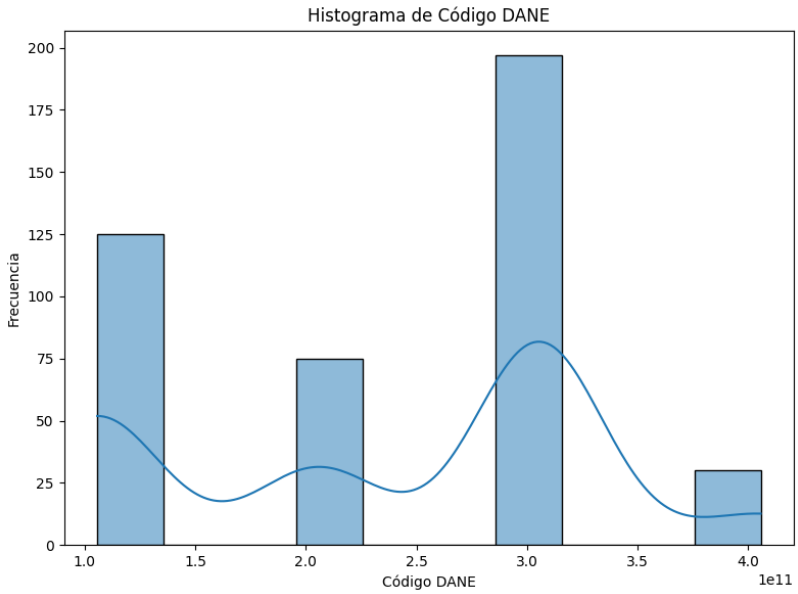
5.2 *Salida. Histograma (Año)*



5.3 *Código para histograma (Código DANE)*

```
# Histograma para Código DANE
plt.figure(figsize=(8, 6))
sns.histplot(df_filtered['Código DANE'], bins=10, kde=True)
plt.xlabel('Código DANE')
plt.ylabel('Frecuencia')
plt.title('Histograma de Código DANE')
plt.tight_layout()
plt.show()
```

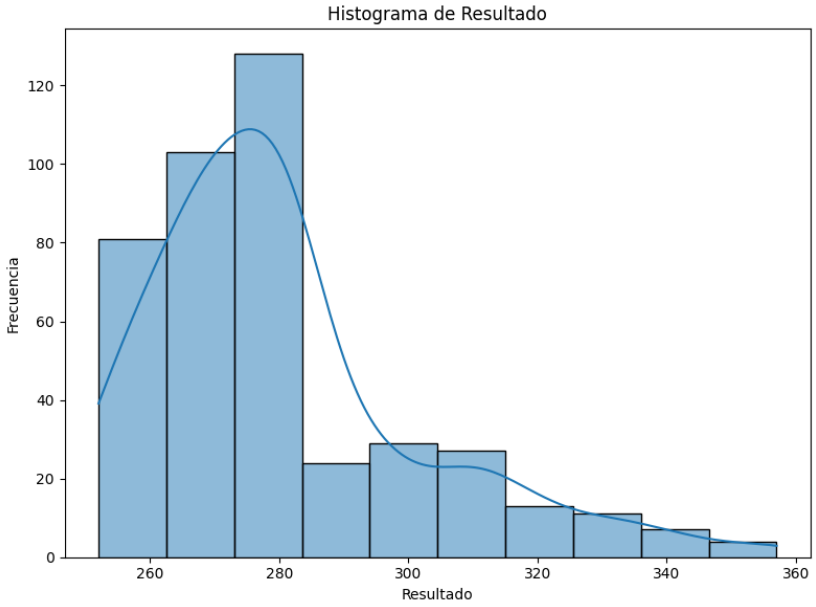
5.4 *Salida. Histograma (Código DANE)*



5.5 *Código para histograma (Resultados)*

```
# Histograma para Resultado
plt.figure(figsize=(8, 6))
sns.histplot(df_filtered['Resultado'], bins=10, kde=True)
plt.xlabel('Resultado')
plt.ylabel('Frecuencia')
plt.title('Histograma de Resultado')
plt.tight_layout()
plt.show()
```

5.6 *Salida. Histograma (Resultados)*



Detalles y explicación:

Aunque el conjunto de datos incluye variables como *Año* y *Código DANE* en formato numérico, estas no representan medidas cuantitativas continuas, sino categorías o identificadores. El *Año* es una variable discreta con pocos valores posibles, y el *Código DANE* corresponde a un identificador único para cada institución educativa, por lo que su valor numérico no tiene un significado de magnitud ni permite establecer distancias interpretables entre categorías.

En cambio, la variable *Resultado* es cuantitativa y permite medir el rendimiento académico en una escala numérica donde la distancia entre valores es significativa. Esto la convierte en una opción adecuada para construir un histograma, ya que dicho gráfico permite observar la distribución de los puntajes, identificar tendencias centrales, dispersión y posibles sesgos en los datos.

Interpretación de la figura 5.6

La distribución de los resultados presenta un pico principal entre aproximadamente 270 y 280 puntos, lo que indica que la mayoría de los estudiantes obtuvo puntajes en ese rango. A medida que los valores aumentan por encima de 280, la frecuencia disminuye, evidenciando menos estudiantes con puntajes altos.

Figura 6. Diagramas de caja para variables numéricas.

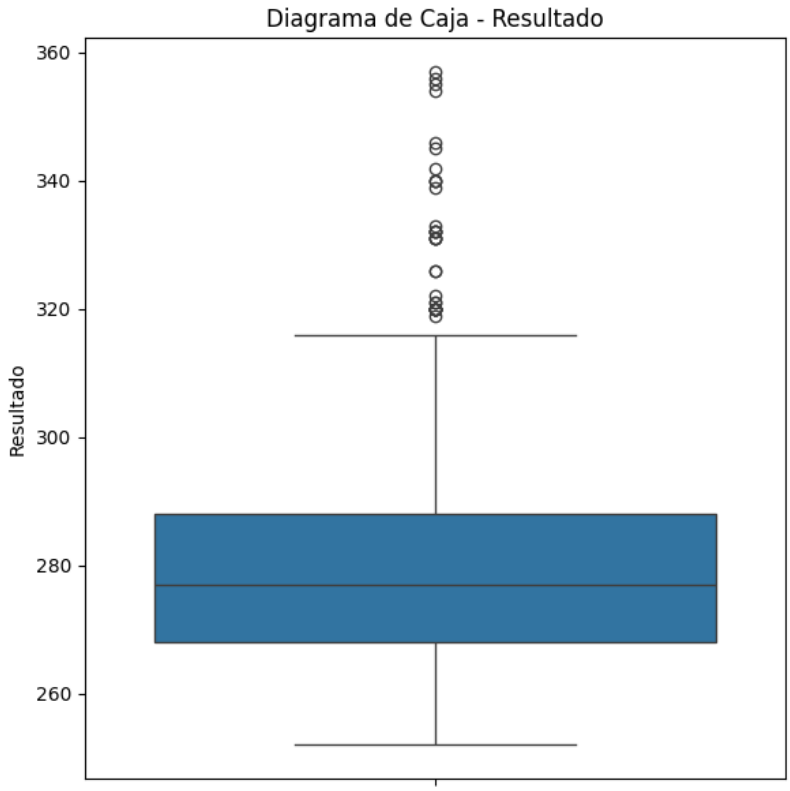
DATO INICIAL: No se incluyeron los resultados correspondientes a los indicadores *Promedio Matemáticas* y *Promedio Lectura Crítica*, debido a que ambos se evalúan sobre una escala de 100 puntos, mientras que los demás indicadores globales se miden sobre una escala de 500 puntos.

6.1 *Código para Diagramas de caja (Año, Código DANE, Resultado)*

```
import matplotlib.pyplot as plt
import seaborn as sns

# Diagrama de caja para Año
plt.figure(figsize=(6, 6))
sns.boxplot(y=df_filtered['Año'])
plt.title('Diagrama de Caja - Año')
plt.tight_layout()
plt.show()

# Diagrama de caja para Código DANE
plt.figure(figsize=(6, 6))
sns.boxplot(y=df_filtered['Código DANE'])
plt.title('Diagrama de Caja - Código DANE')
plt.tight_layout()
plt.show()

# Diagrama de caja para Resultado
plt.figure(figsize=(6, 6))
sns.boxplot(y=df_filtered['Resultado'])
plt.title('Diagrama de Caja - Resultado')
plt.tight_layout()
plt.show()
```

En este caso, se explica el Diagrama de caja para Resultado, teniendo en cuenta la figura 5. Histogramas y lo explicado anteriormente.

6.2 *Salida. Diagrama de caja (Resultados)*



Detalles y explicación:

La mediana de los resultados se encuentra aproximadamente en 277 puntos, lo que indica que la mitad de los estudiantes obtuvo puntajes iguales o inferiores a este valor. El rango intercuartílico (IQR) se ubica aproximadamente entre 268 y 288 puntos, mostrando que la mayoría de los estudiantes se concentra en ese intervalo.

El diagrama evidencia la presencia de valores atípicos altos (puntos por encima de 315), correspondientes a estudiantes con desempeños notablemente superiores al resto. No se observan valores atípicos bajos, lo que sugiere que los puntajes más bajos se encuentran dentro del rango esperado.

La distribución general es relativamente simétrica dentro del rango central, aunque la existencia de numerosos atípicos hacia arriba coincide con lo observado en el histograma, donde la distribución estaba ligeramente sesgada a la derecha.

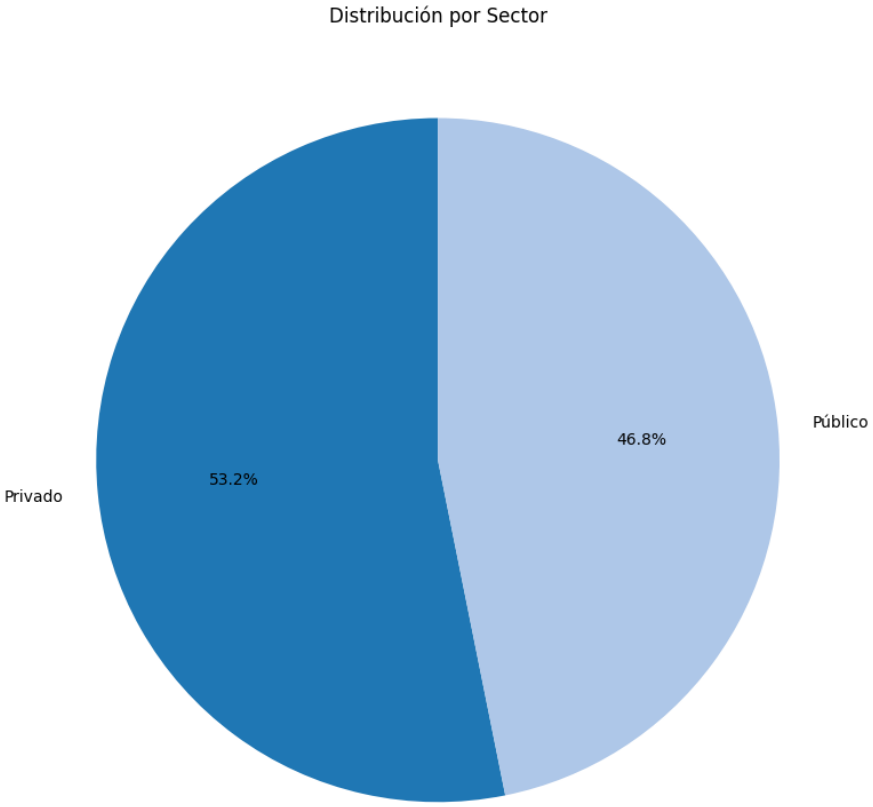
Figura 7. Gráficos de torta para variables categóricas.

7.1 *Código para gráficos de torta (Distribución por sector, colegio e indicador)*

```python
import matplotlib.pyplot as plt

# Lista de variables categóricas y títulos para los gráficos
categorical_vars = [
    ('Sector', 'Distribución por Sector'),
    ('Colegio', 'Distribución por Colegio'),
    ('Indicador', 'Distribución por Indicador')
]

for var, title in categorical_vars:
    plt.figure(figsize=(8, 8))
    df_filtered[var].value_counts().plot.pie(
        autopct='%1.1f%%',
        startangle=90,
        colors=plt.cm.tab20.colors
    )
    plt.title(title)
    plt.ylabel('')  # Eliminar la etiqueta del eje y
    plt.tight_layout()
    plt.show()
```

NOTA: Los gráficos de torta son útiles para mostrar la participación porcentual de cada categoría respecto a un conjunto total. Su principal ventaja es que permiten identificar visualmente, de manera inmediata, qué categorías tienen mayor o menor peso relativo, facilitando la comparación de proporciones. Este tipo de gráfico es especialmente valioso cuando se trabaja con variables categóricas que representan partes de un todo, se busca resaltar la categoría con mayor o menor participación y el número de categorías es reducido (idealmente menos de seis) para mantener la legibilidad.

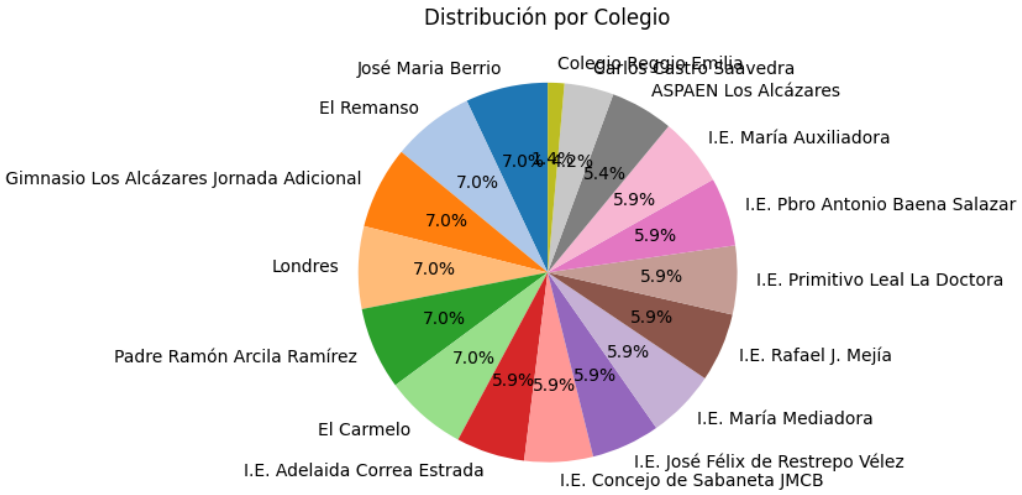
7.2 *Salida. Gráfico de torta para Sector*



Detalles y explicación:

La El gráfico muestra la proporción de estudiantes según el sector educativo. El 53,2% proviene de instituciones privadas y el 46,8% de instituciones públicas. Esto refleja una participación relativamente equilibrada, con una ligera mayoría del sector privado.

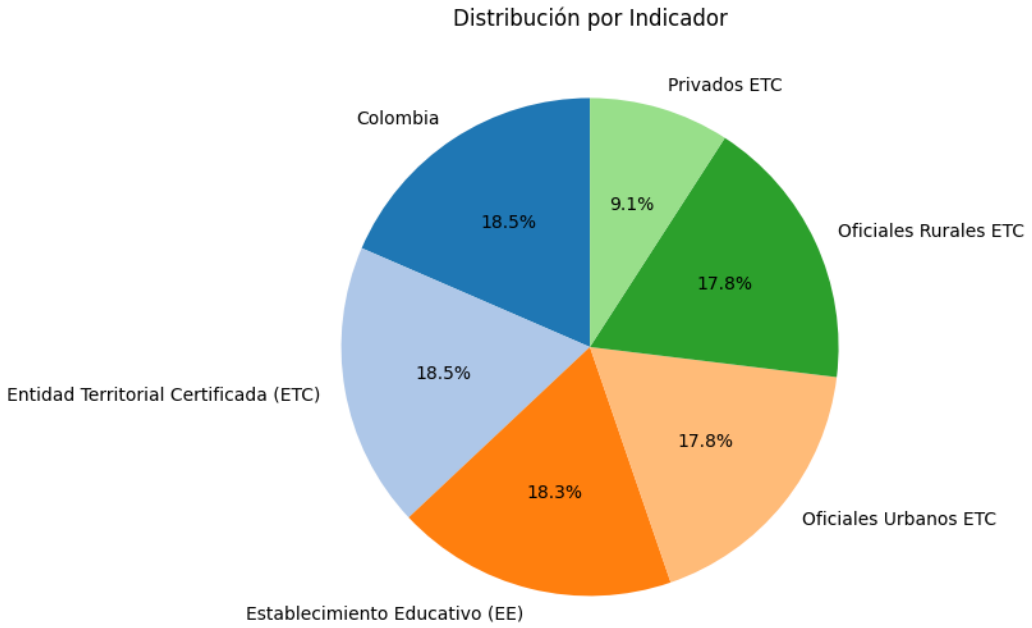
7.3 *Salida. Gráfico de torta para Colegio*



Detalles y explicación:

Este gráfico de torta desglosa la participación de cada colegio en el conjunto de datos. La mayoría de las instituciones tienen una participación cercana al 7% o 5,9%, lo que sugiere que no hay un colegio que concentre de forma desproporcionada la muestra. Sin embargo, hay instituciones con porcentajes más bajos (3,4% y 4,2%), lo que indica menor representación en el análisis. Es necesario aclarar que visualmente el Diagrama de tortas para visualizar la distribución por colegio no es el más eficaz, debido a la cantidad de colegios que se obtiene, para ello, se podría recurrir en un diagrama de barras.

7.4 *Salida. Gráfico de torta para Indicador*



Detalles y explicación:

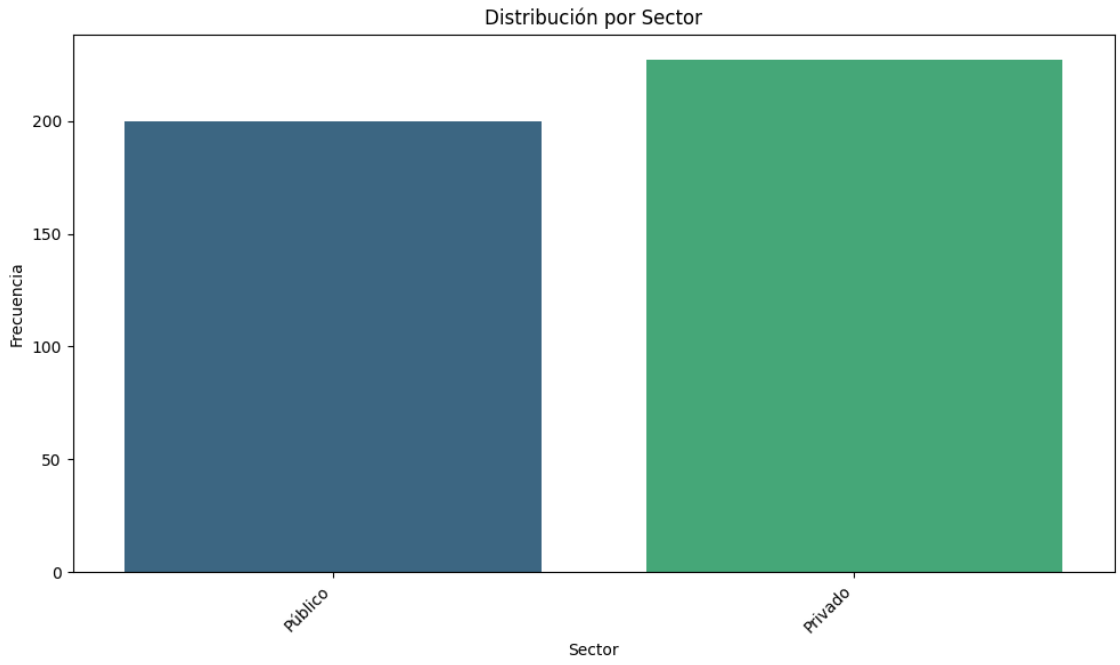
Acá se muestra la proporción de distintos indicadores educativos a nivel institucional o territorial. Las categorías *Colombia*, *Entidad Territorial Certificada (ETC)*, y *Establecimiento Educativo (EE)* representan alrededor del 18% cada una, mientras que *Oficiales Urbanos ETC* y *Oficiales Rurales ETC* se sitúan cerca del 17,8%. La categoría *Privados ETC* tiene la menor participación, con 9,1%, lo que podría indicar menor presencia de este grupo en los datos evaluados.

Figura 8. Gráficos de barras para variables categóricas.

8.1 *Código para gráfico de barras*

```python
import matplotlib.pyplot as plt
import seaborn as sns

# Lista de variables categóricas y títulos para los gráficos de barras
categorical_vars_bar = [
    ('Sector', 'Distribución por Sector'),
    ('Colegio', 'Distribución por Colegio'),
    ('Indicador', 'Distribución por Indicador')
]

for var, title in categorical_vars_bar:
    plt.figure(figsize=(10, 6))
    sns.countplot(data=df_filtered, x=var, palette='viridis')
    plt.title(title)
    plt.xlabel(var)
    plt.ylabel('Frecuencia')
    plt.xticks(rotation=45, ha='right')  # Mejora la legibilidad de las etiquetas
    plt.tight_layout()
    plt.show()
```

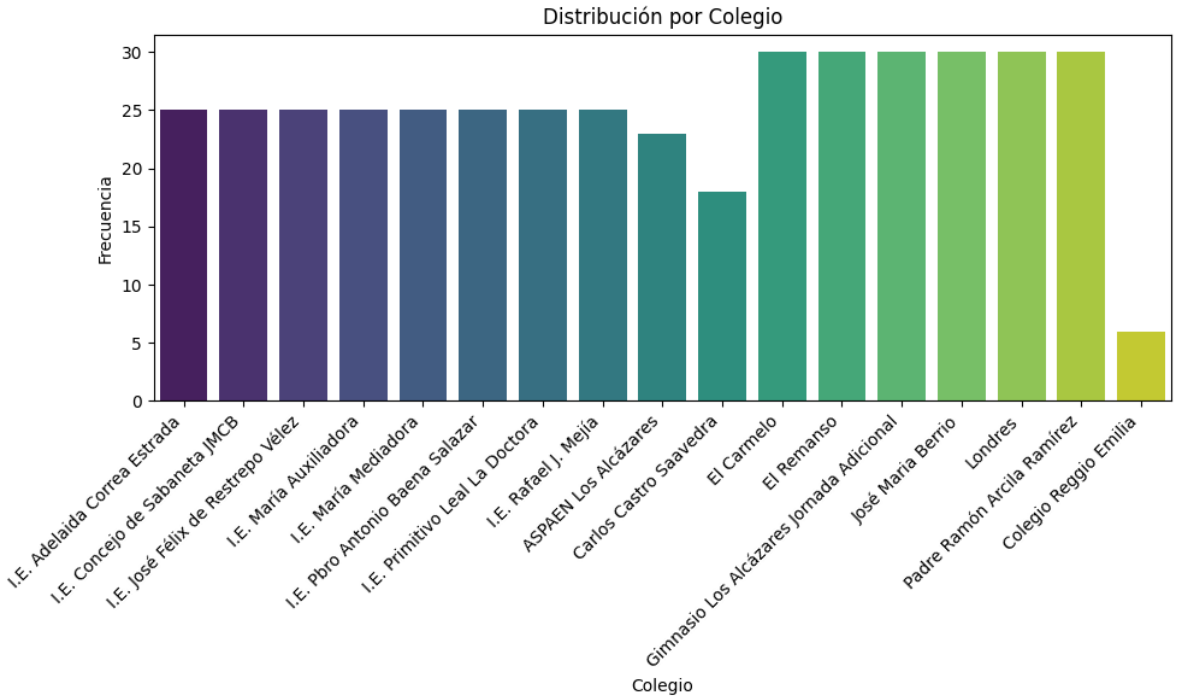
8.2 *Salida. Gráfico de barras (Sector)*



Detalles y explicación:

En la figura 8.2 se puede visualizar una mayor participación en cuando al sector Privado, sin embargo, es importante aclarar que se obtuvo una mejor proyección de los datos a nivel visual en la figura 7.2 donde se mostró la misma información, pero con una gráfica de torta.

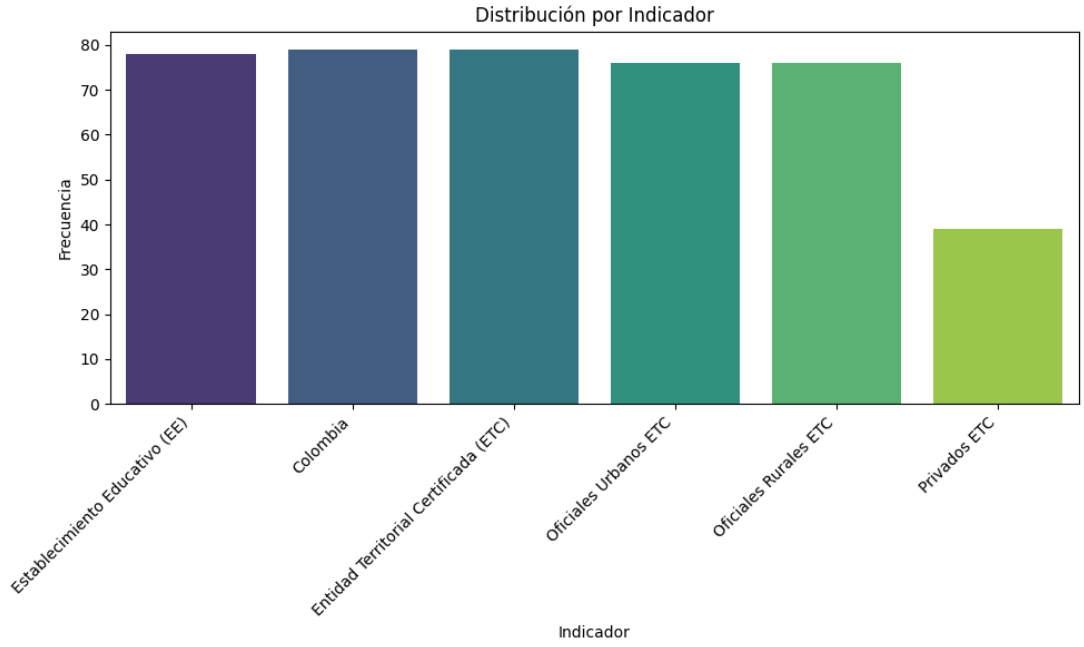
8.3 *Salida. Gráfico de barras (Colegio)*



Detalles y explicación:

Por otra parte, en el diagrama de barras para la distribución por colegio se obtiene una mejor proyección de los datos y con una mayor claridad, sabiendo que el colegio Reggio Emilia tiene una menor participación a comparación de los otros colegios.

8.4 *Salida. Gráfico de barras (Indicador)*



Detalles y explicación

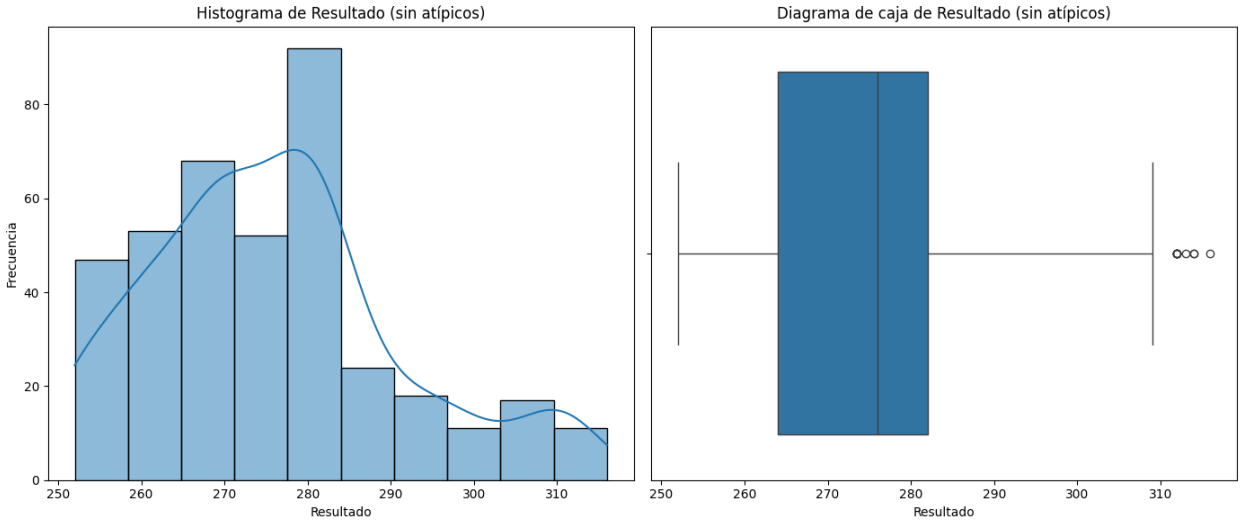
De diferente manera representativa, se observa en la figura 8.4 la misma información que se visualizó en la figura 7.4 Gráficos de tortas, sin embargo, se obtiene un numero más preciso a nivel visual en la figura 7.4

Figura 9. Outliers

DATO INICIAL: Con ayuda de la IA, se obtuvo los cuartiles, teniendo en cuenta de no usar las filas Promedio Matemáticas y Promedio Lectura critica en indicador para no tener en cuenta aquellos resultados que se evalúan sobre 100, si no tener los resultados globales (sobre 500).

Tabla 1.

Cantidad de datos	427
Media	280
Desviación estándar	21.22
Mínimo	252
Q1 (25%)	268
Mediana (50%)	277
Q3 (75%)	288
Máximo	357



Detalles y explicación:

Histograma

La forma es aproximadamente simétrica, con una ligera tendencia hacia la derecha (asimetría positiva), porque la media (280) está un poco por encima de la mediana (277), la mayoría de puntajes se concentran en el rango 268 – 288, que coincide con el rango intercuartílico (IQR). Hay menos frecuencia en los extremos (cerca de 252 o 357), lo que indica que no hay muchos estudiantes con puntajes muy bajos o muy altos después de eliminar los datos atípicos. En cuanto al rango total es 357 – 252 = 105 puntos, pero el 50 % central solo varía en 20 puntos, lo que muestra que la mayor parte de los estudiantes tiene resultados bastante similares.

Diagrama de caja

Caja: Representa el 50 % central de los datos (IQR = 20 puntos), desde Q1 = 268 hasta Q3 = 288.

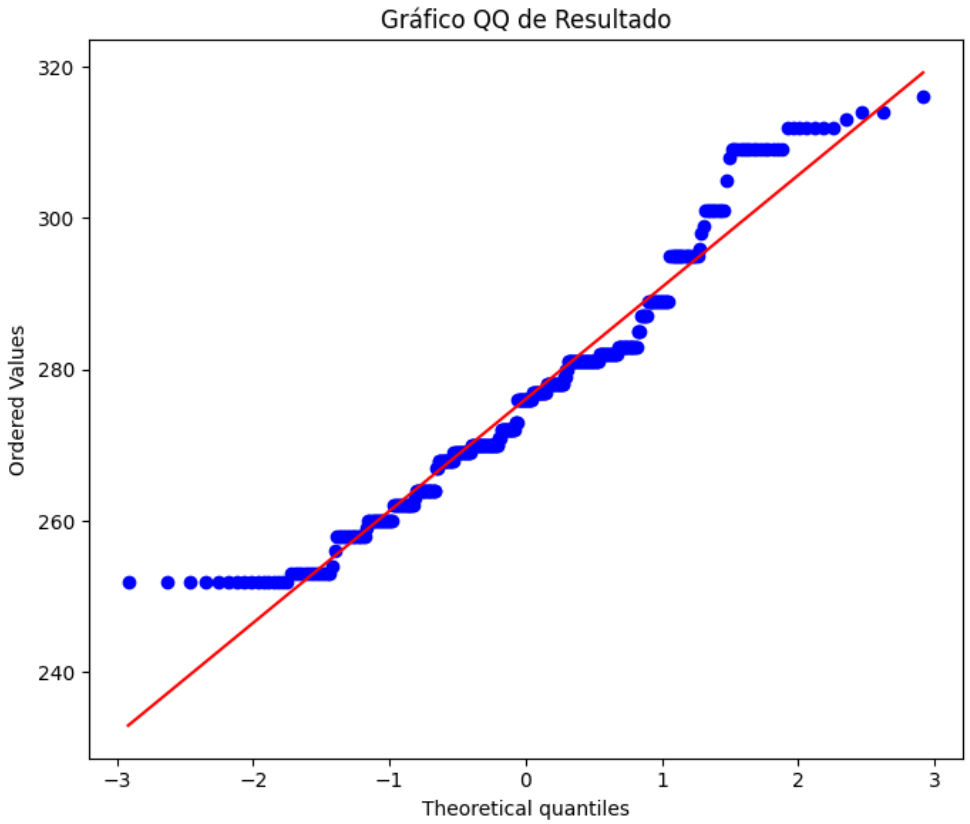
Línea dentro de la caja: La mediana (277), que está ligeramente más cerca de Q1 que de Q3, lo que indica una leve concentración de valores en la parte alta de la caja.

Bigotes: Se extienden desde el mínimo (252) hasta el máximo (357), mostrando el rango total sin atípicos.

Ausencia de puntos atípicos: No aparecen valores aislados fuera de los bigotes.

9. Gráfica QQ

DATO INICIAL: Para este grafico se tomaron en cuenta todos los resultados que se encontraban en el archivo, sin excluir los datos Promedio matemáticas y Promedio Lectura critica.



Detalles y explicación:

Diagrama de caja

Si los puntos siguen de cerca la línea roja, significa que los datos tienen una distribución aproximadamente normal, pero en el caso de cuando los puntos se desvían mucho de la línea, hay evidencia de que los datos no siguen una normalidad perfecta. Esto indica que la distribución es casi normal, pero con algunos valores atípicos o concentraciones en ciertos puntajes.

Observaciones

En el análisis realizado se excluyeron los indicadores “Promedio Matemáticas” y “Promedio Lectura Crítica”, ya que se evalúan sobre una escala de 0 a 100, a diferencia del resto de indicadores globales que se miden sobre 0 a 500, lo que podría distorsionar las comparaciones. Los resultados muestran que el sector privado presenta una ligera mayor participación que el sector público, y que existe una distribución desigual entre las instituciones educativas: algunas, como El Remanso o Colegio Londres, aportan un alto número de estudiantes, mientras que otras, como el Colegio Reggio Emilia, tienen una representación mínima. Esta disparidad, junto con las diferencias de escalas, sugiere la necesidad de realizar análisis segmentados para evitar sesgos en las conclusiones. Los gráficos de barras y de torta resultaron herramientas valiosas para visualizar la información, ya que permitieron identificar de forma clara y comprensible las proporciones y frecuencias, facilitando la interpretación de los datos.