Informe Estadística Inferencial y Muestreo.

Tomas Santiago Bretón García – U00189852
Negocios Internacionales.

Universidad Autónoma de Bucaramanga.
Docente: Andrés Fabián Leal Archila.
Fecha: 08/08/2025

1. Resumen.

El análisis se llevó a cabo con resultados del ICFES de distintas instituciones educativas de Sabaneta entre 2016 y 2020. Primero se realizó una revisión completa de la información y, posteriormente, se organizó con el fin de eliminar datos incompletos y así trabajar únicamente con registros válidos. Luego se ordenaron las variables categóricas y se representaron mediante diagramas de barras y de torta; en el caso de las variables numéricas, se utilizaron histogramas y diagramas de caja. También se evaluó si los puntajes obtenidos por los estudiantes se encontraban muy alejados del promedio. Finalmente, se verificó si los puntajes seguían una distribución normal y se emplearon diagramas QQ para visualizarlos de manera más clara.

1.1 Abstract.

The analysis was conducted using ICFES results from various educational institutions in Sabaneta between 2016 and 2020. First, a thorough review of the information was conducted and, subsequently, the data was organized in order to eliminate incomplete entries, allowing work only with valid records. Then, the categorical variables were organized and represented using bar charts and pie charts; in the case of numerical variables, histograms and box plots were used. The scores obtained by students were also evaluated to determine whether they were significantly far from the average. Finally, it was verified whether the scores followed a normal distribution, and QQ plots were used to visualize them more clearly.

2. Introducción.

Este informe busca presentar un análisis claro de los resultados ICFES que se obtuvieron en Sabaneta durante el periodo de tiempo entre 2016 y 2020. Con el propósito de evaluar el desempeño de los estudiantes, permitiendo una mejor comprensión de estos a través de un análisis exhaustivo permitiendo identificar factores que influyeran en los resultados.

Para llevar esto a cabo, se realizó un análisis sobre la información y se buscó organizarla para tener certeza de que los datos estuvieran completos. Posterior a esto, se hizo una identificación de variables categóricas y numéricas, con las que se realizaron distintos diagramas como barras y torta para visualizar las categóricas e histograma y diagrama de caja para la variable numérica, esto permitió examinar los puntajes a detalle y gracias a esto se verifico si seguían una distribución normal.

Con este informe se busca ofrecer un análisis completo y concreto del rendimiento académico del municipio durante el respectivo periodo de tiempo, aportando datos útiles y realizando su respectivo análisis, dando una visión más clara de el porque se obtuvo esos resultados.

3. Metodología.

El análisis de los resultados del ICFES en instituciones educativas de Sabaneta entre 2016 y 2020 se desarrolló mediante un procedimiento estructurado, orientado a garantizar la integridad de los datos y la validez de las conclusiones. El proceso se organizó en las siguientes fases

3.1 Recolección de datos.

Para la recolección de datos del ICFES se utilizo una base de datos que contenía los resultados obtenidos en el periodo entre 2016 y 2020 en Sabaneta.

3.2 Depuración de la información.

Se hizo una revisión de los registros para poder realizar la identificación y eliminación de datos inconsistentes.

3.3 Organización de las variables.

Los datos obtenidos se organizaron en dos variables las cuales son categóricas y numéricas y para cada una respectivamente se realizaron diagramas de torta y barras para las categóricas y histograma y diagrama de caja para numéricas.

4. Objetivos.

4.1 General: Analizar los resultados del ICFES obtenidos por las instituciones educativas de Sabaneta durante el periodo 2016–2020, con el fin de identificar patrones, tendencias y factores que puedan influir en el rendimiento académico de los estudiantes.

4.2 Específicos.

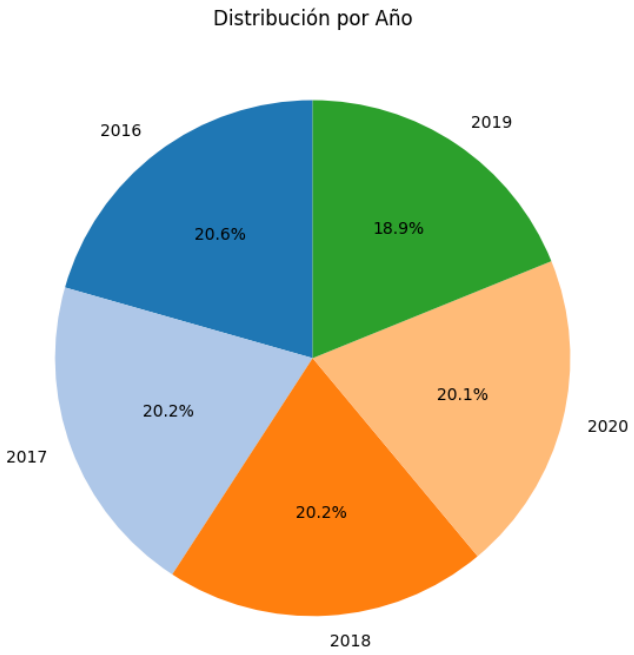
- Organizar y depurar la información para garantizar la integridad y validez de los datos utilizados en el análisis.

• Representar las variables categóricas y numéricas mediante gráficos adecuados que faciliten su interpretación.

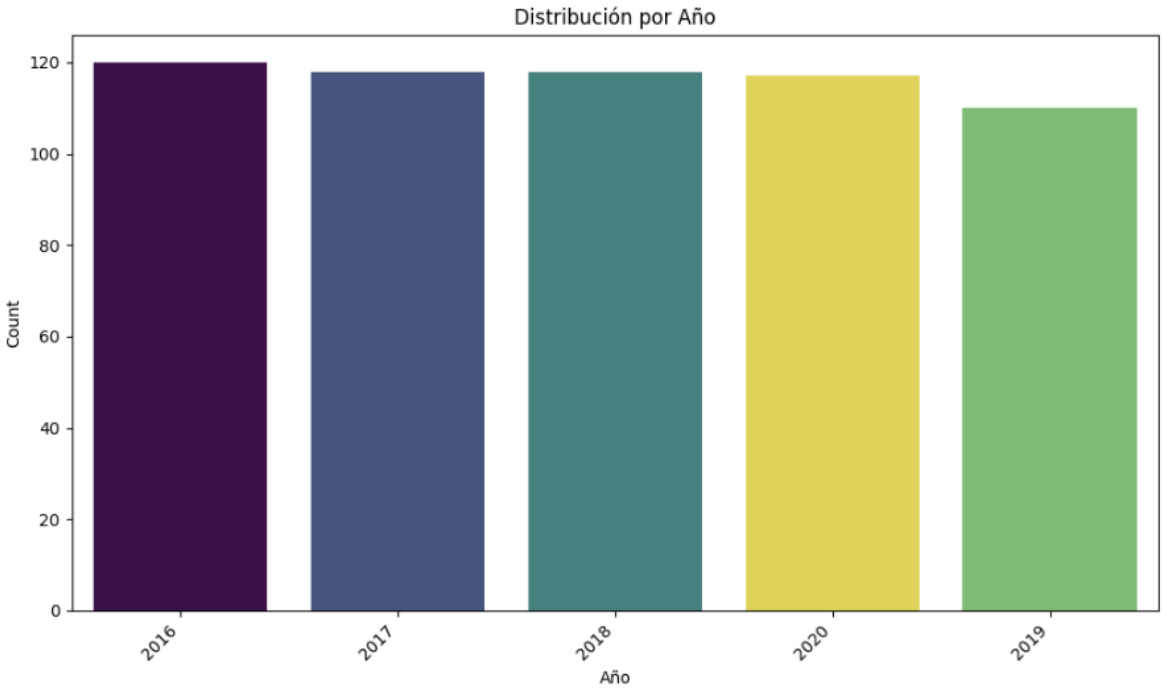
• Evaluar la distribución de los puntajes y detectar posibles valores atípicos que puedan influir en los resultados.

5. Análisis de Datos.

5.1. Análisis variables categóricas en diagrama de torta y barras.



Fuente: Propia.

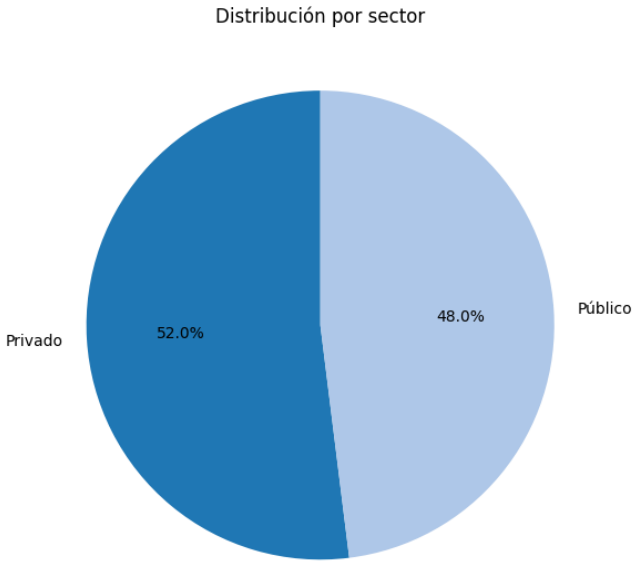


Fuente: Propia

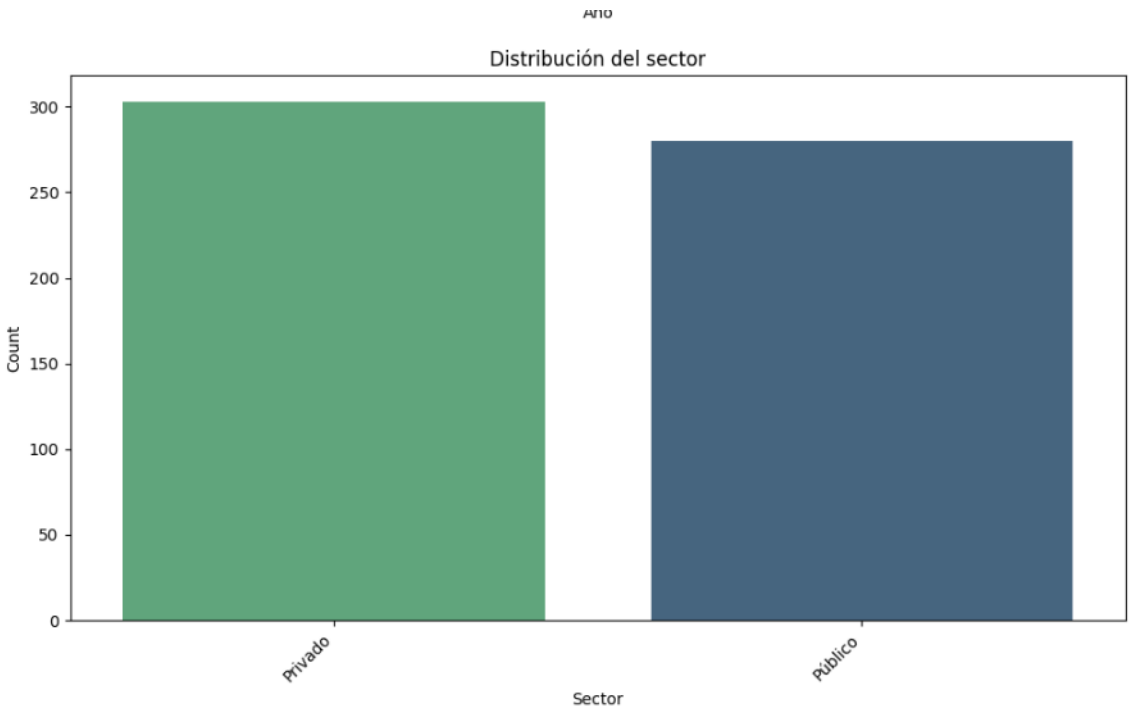
- En el diagrama de torta observamos la distribución por año de los estudiantes que presentaron el ICFES durante el periodo entre 2016y 2020.Podemos observar que los porcentajes de todos los años son similares pues 2016 tuvo 20.6% siendo el mayor porcentaje lo que indica que hubo un mayor número de estudiantes que presentaron el examen ese año, después tenemos los años 2017 y 2018 20.2% que tuvieron el mismo porcentaje por lo que en esos dos años hubo la misma cantidad de estudiantes que presentaron el examen, en 2019 el porcentaje disminuyo hasta 18.9% lo que indica que hubo menos participantes que en los 3 años anteriores y en el año 2020 tuvo un porcentaje del 20.1% lo que indica que hubo menos participantes que los años 2016, 2017 y 2018 pero hubo más participantes que en el año 2019. Esto indica que durante el 2019 hubo menos participantes que en el resto de los años y esto pudo haber sido debido a distintos factores que perjudicaron la participación los estudiantes factores que no se presentaron en los demás años debido a la similitud de los porcentajes o que si se presentaron lo hicieron en una menor frecuencia.

- En el diagrama de barras observamos la cantidad de estudiantes que presentaron el examen por cada año podemos observar que en el 2016 presentaron 120 estudiantes en el 2017 118-119 estudiantes aproximadamente en el 2018 la misma cantidad que en el 2017, en el 2020 115-116 estudiantes y en el 2019 podemos observar que este numero disminuyo podemos poner un aproximado de 110 estudiantes, este grafico nos permite observar y comparar la cantidad de estudiantes que participaron en el

examen en los diferentes años y permitiendo sacar distintas conclusiones al observar las barras lo que permite un análisis mas sencillo, aunque cabe recalcar que los valores no son exactos a simple vista puesto que la escala no va de 1 en 1 para poder decir con exactitud cuantos estudiantes fueron por esa razón se hace un aproximado a simple vista.



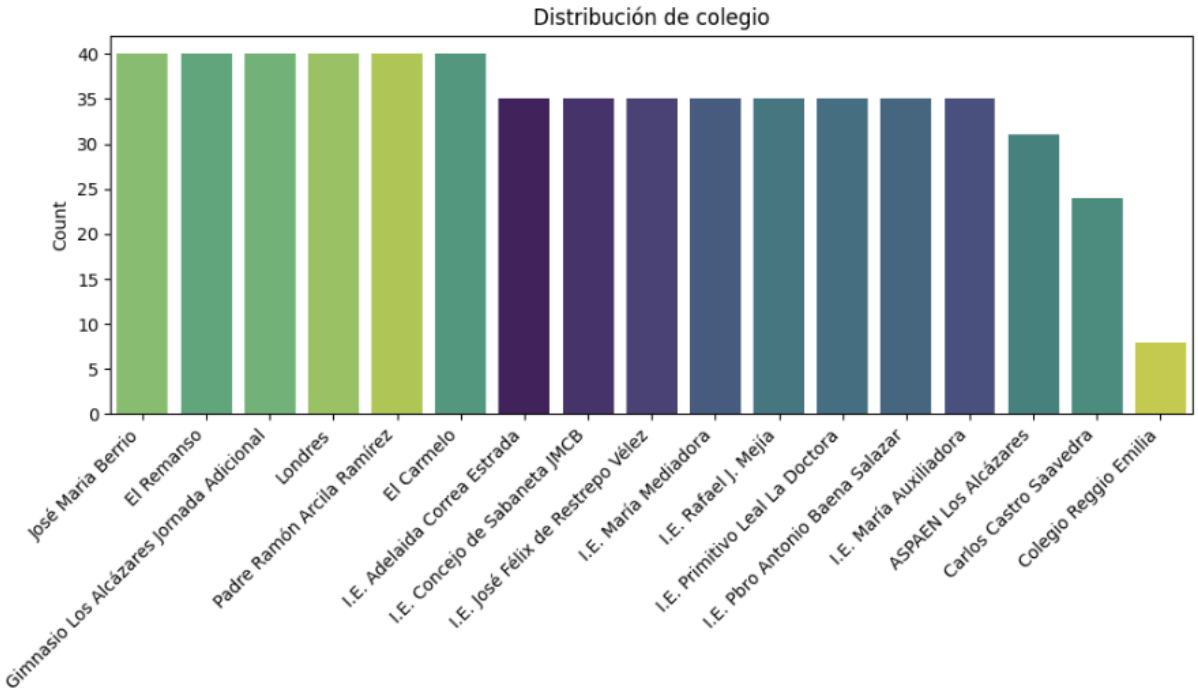
Fuente: Propia.



Fuente: Propia

- El diagrama muestra la distribución de los estudiantes que presentaron el icfes basado en la variable categórica del sector educativo, este grafico revela que un 52% pertenece a instituciones privadas por lo que el 48% restante pertenece a instituciones públicas. Esta diferencia de porcentajes no es amplia, pero si nos indica que entre 2016 y 2020 los estudiantes del sector privado tuvieron predominancia a la hora de presentar el icfes, esto puede ser provocado por la preferencia de la educación privada sobre la publica por ejemplo sin embargo puede haber otros factores que hacen que la balanza se incline hacia el sector privado. Este resultado permite observar posibles diferencias en lo que se refiere al rendimiento académico y que están asociadas al tipo de institución a la que pertenecen.

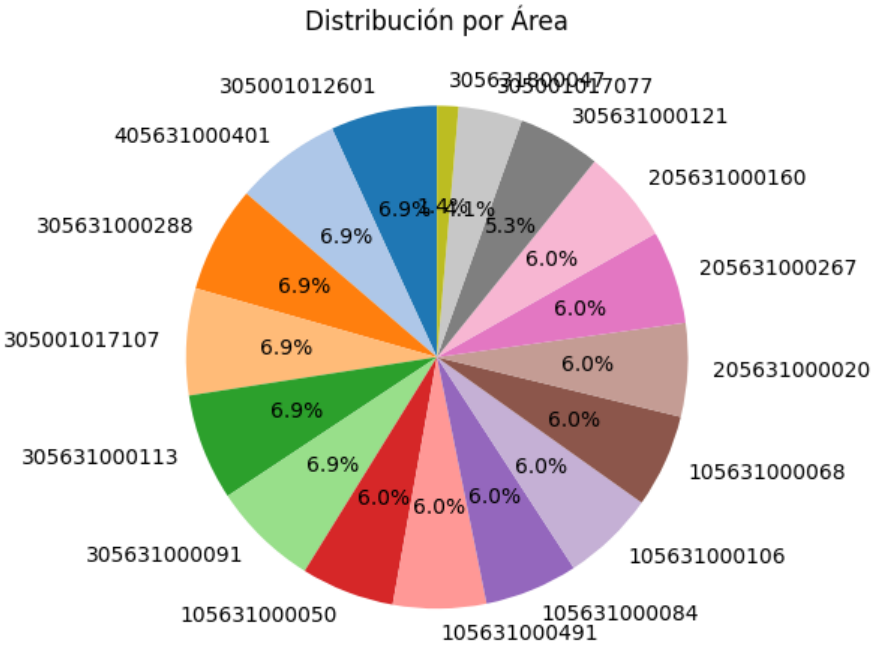
- En el diagrama de barras se observa que del sector privado participaron 300 estudiantes mientras que en el sector publico se puede aproximar a que participaron uno 270 estudiantes que es una cantidad menor a comparación del sector privado y gracias al uso de este diagrama es más fácil saber la diferencia de estudiantes a diferencia de si utilizamos el de torta pues este se da por números y el de torta es representado por porcentajes por lo que la interpretación de datos es de cierta forma más sencilla.



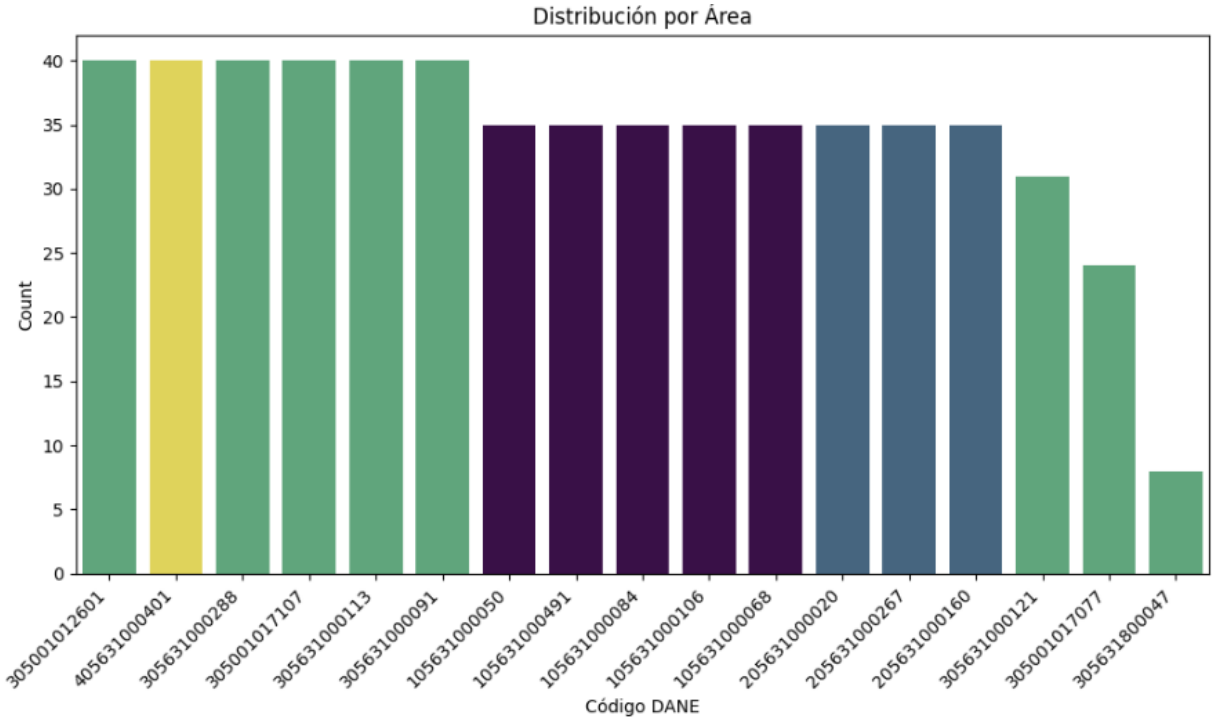
Fuente: Propia

- En este diagrama de barras podemos observar la distribución de los estudiantes por colegios, podemos observar que a la izquierda los primeros 6 colegios poseen una cantidad igual de estudiantes participantes a estos les siguen los colegios desde la

7ma casilla hasta la numero 14 que también poseen la misma cantidad de estudiantes y los últimos 3 colegios poseen cantidades distintas de estudiantes participantes, a simple vista podemos observar que hay una diferencia grande entre los colegios que mas participaron estudiantes y el que menos participo esto puede deberse a en el colegio Reggio Emilia que obtuvo la menor cantidad de participantes no existieran suficientes candidatos para presentar el ICFES durante ese periodo de tiempo, gracias al grafico podemos observar que los datos están organizados de mayor a menor en dirección de izquierda a derecha que permite una clara visualización de la diferencia de participantes por colegio y que gracias a esto se facilita el análisis de esta distribución, la diferencia como mencione puede ser debido a la diferencia de alumnos en cada institución y esto puede derivar gracias a las otras variables categóricas presentes como el sector, el área ya que estos son factores que pueden provocar que un estudiante se matricule en un colegio o en otro.

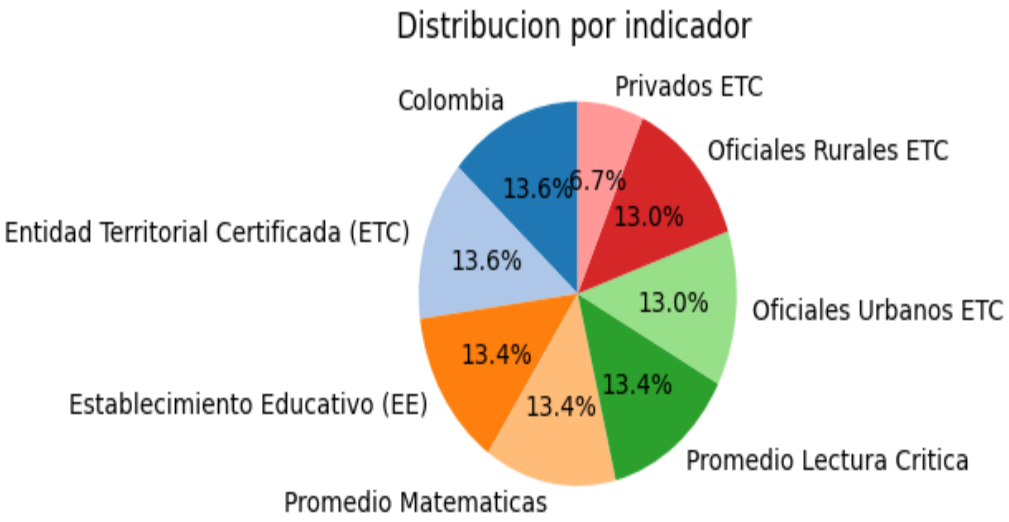


Fuente: Propia.

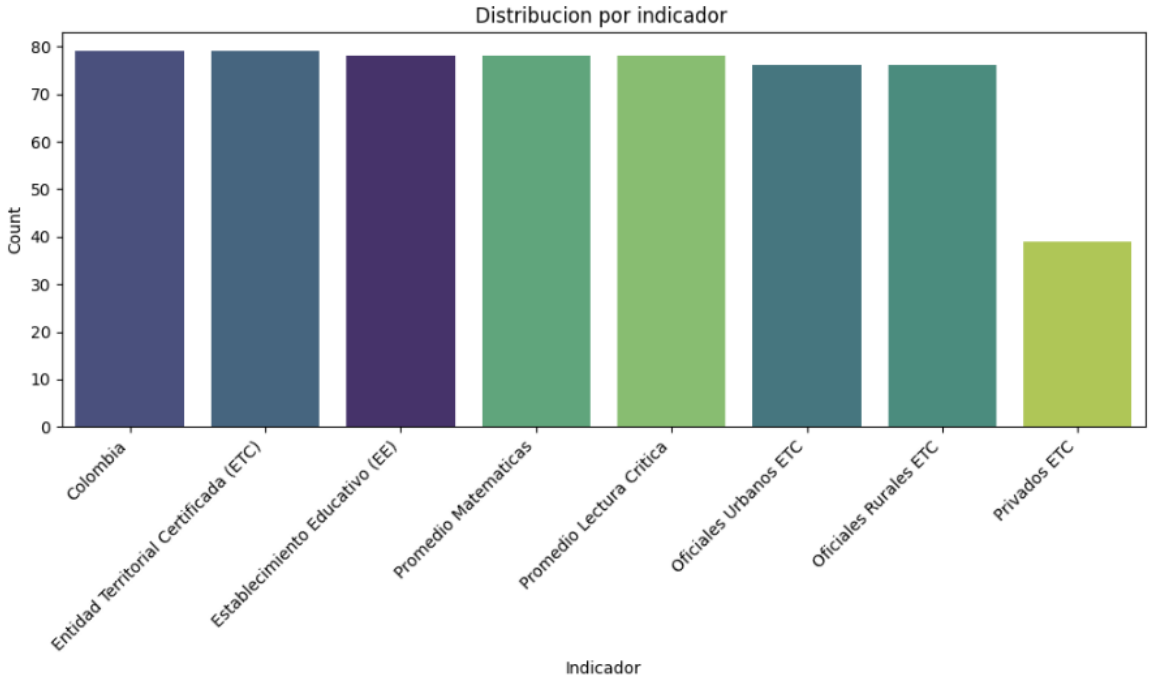


Fuente: Propia

- Podemos observar dos diagramas uno de torta que se mide por porcentajes y uno de barras que se da una cantidad que dependiendo de la escala puede ser exacta o será necesario aproximar, observando los dos gráficos podemos ver a simple vista que hay unas áreas que comparten un mismo numero de participantes o porcentajes similares de estudiantes pero también podemos observar que hay 3 que son totalmente diferentes a los demás datos proporcionados, gracias a esto podemos inferir que al ser distribuciones por área podemos hablar sobre zonas de Sabaneta que es de donde se recolectaron los datos por lo que podemos inferir que son distintos estratos socioeconómicos y que este puede ser un factor de peso para el hecho que de unas zonas hayan tantos participantes pero de otras esta cifra disminuya puesto que la educación al fin y al cabo también depende de lo socioeconómico y es un factor a tener en cuenta, con el grafico de barra podríamos decir que las zonas que comparten los 40 estudiantes son las zonas de mayor estrato socioeconómico las que comparten los 34 estudiantes aproximadamente son un estrato menor que las zonas de 40 estudiantes pero mayor que las otras zonas que no comparten misma cantidad de estudiantes con ninguna otra, si observamos detenidamente los gráficos son prácticamente iguales a los gráficos de colegios puesto que colegios están en ciertas áreas por lo que compartirán los mismos datos.



Fuente: Propia.



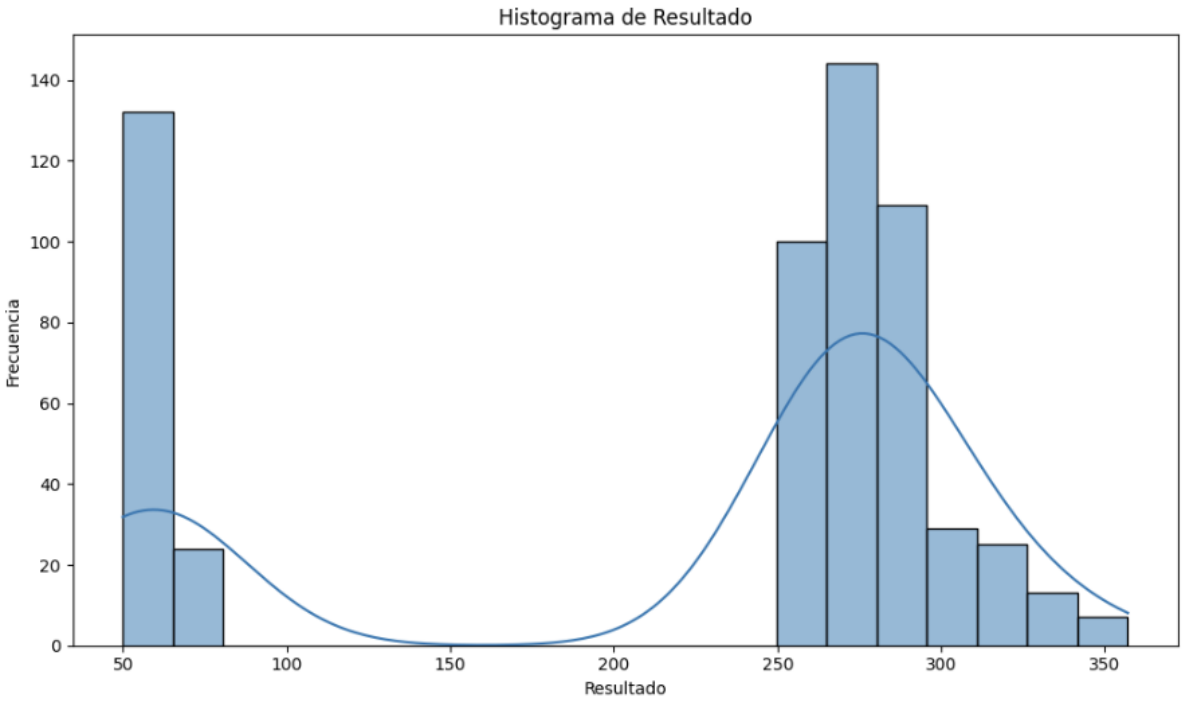
Fuente: Propia

- En el grafico de torta podemos observar porcentajes similares entre los indicadores Colombia y ETC que comparten un 13.6%, esto también se observa entre EE, promedio de matemáticas y promedio de lectura critica que comparten un porcentaje de 13.4%, los indicadores de Oficiales Rurales ETC y Oficiales Urbanos ETC también comparten un porcentaje que seria 13% por lo que se puede decir que los porcentajes son similares entre sí, sin embargo si observamos el indicador Privados ETC observamos que solo tiene un 6.7% que marca una diferencia grande

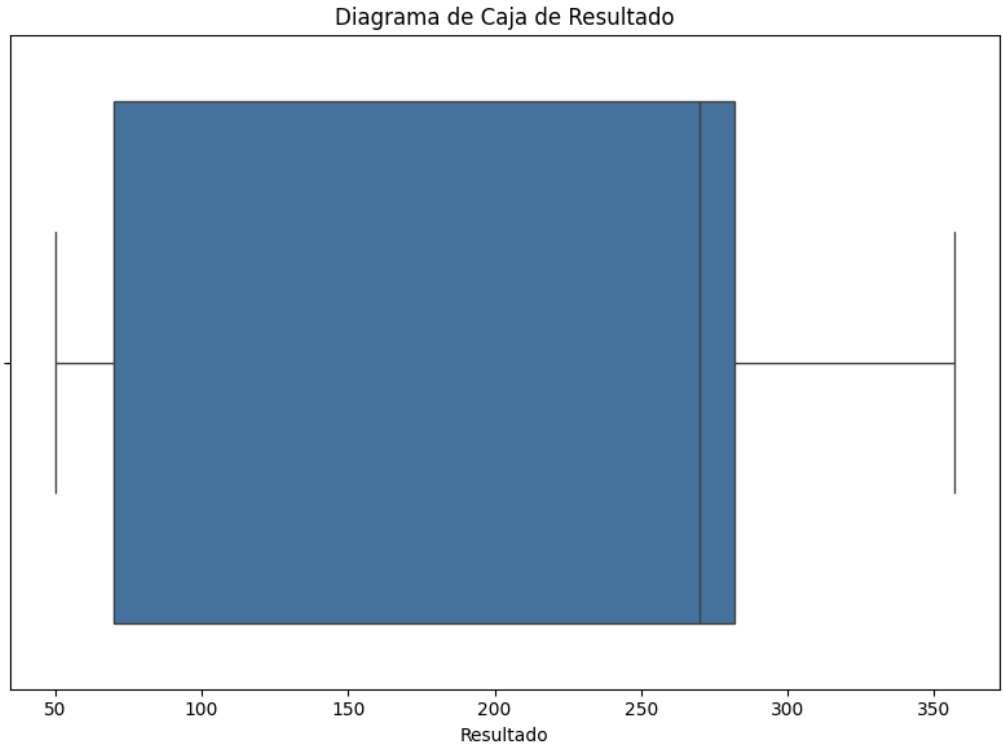
frente a los demás valores dados en si se puede decir que esta distribución es equilibrada a excepción de este último porcentaje mencionado pues su valor es muy distinto a los demás.

- En el grafico de barras podemos observar en vez de porcentajes la cantidad exacta se podría decir en cada indicador por lo que también se es mas notable como es equilibrado en casi todos los indicadores, pero hay uno que genera un desequilibrio grande en los datos que se nos proporcionaron.

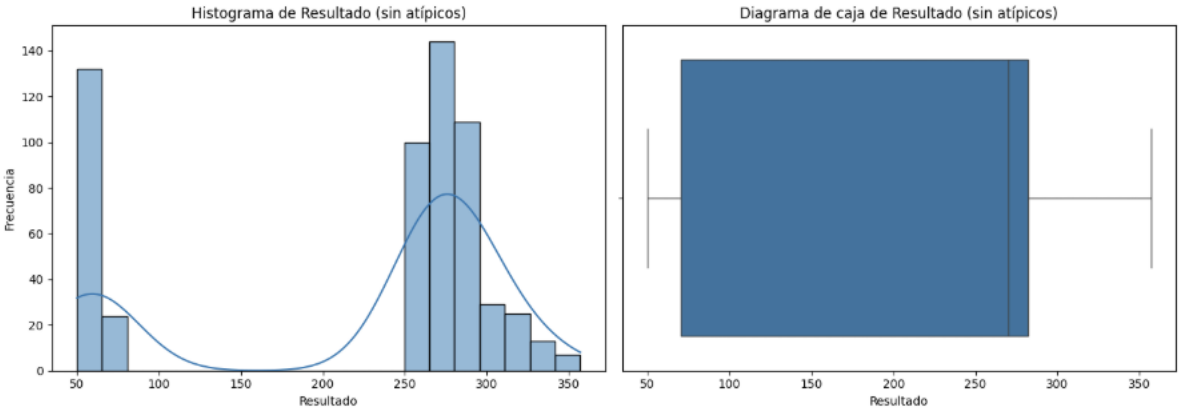
5.2. Análisis variables numéricas en histograma y diagramas de caja.



Fuente: Propia



Fuente: Propia



Fuente: Propia

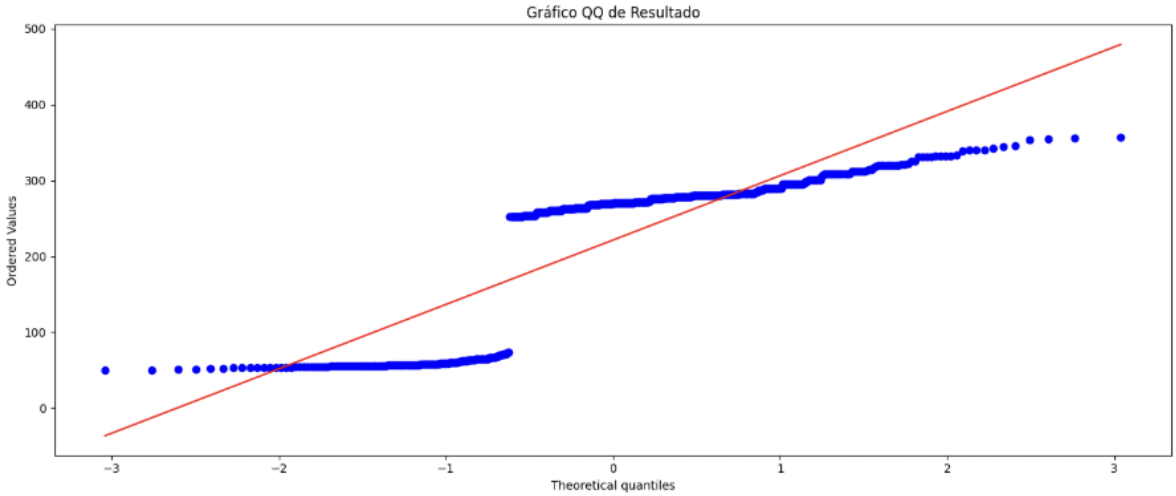
- En el histograma podemos observar dos grupos, el primero seria entre 50 y 80 aproximadamente que sería un grupo pequeño a comparación del segundo que se encuentra entre 250 y 300 algunos incluso llegando a 350 lo que indica que unos resultados son bajos y otros son altos, si se observa detenidamente observamos que la variable numérica de resultados no sigue una distribución normal puesto que existen dos picos y esto se conoce como bimodalidad que se traduce en dos picos claros en lugar de uno solo (Dela Cruz & Jackson, 2023). Además de esto para que sea una distribución normal el pico debería encontrarse en el medio ósea los datos deberían encontrarse en el medio sin embargo esto no pasa en el histograma y por ultimo se observa que existe un hueco entre los rangos de resultados entre el grupo

1 y 2 y esto rompe la forma de campana que es otro indicador que no es distribución normal.

- En el diagrama de caja observamos que la mayoría de los datos o de resultados se agrupan en una zona alta de la escala, pero como sabemos también existen valores bajos lo que indica variabilidad de datos como mencione anteriormente y al presentarse valores así de distantes afectan la distribución de los datos.

- En los diagramas sin datos atípicos se observa que es lo mismo antes y después de hacer esa comprobación por lo que podemos intuir que no existen valores atípicos en los datos recolectados.

5.3. Análisis diagramas QQ



Fuente: Propia

- En el diagrama QQ los puntos azules son valores y como podemos observar estos se alejan bastante de la línea roja, en un gráfico QQ la línea azul debería ir a la par de la línea roja o es lo que se espera ya que la línea roja indica la distribución normal pero como en este caso no la siguen, en otras palabras, "un gráfico Q-Q permite comparar visualmente una distribución de datos con una distribución teórica como la normal" (Waples, 2024). En este caso observamos que en vez de una sola línea son 2, como se ha mencionado anteriormente esto indica que existen dos grupos de datos distintos que generan una discontinuidad que se refleja en el diagrama, también notamos unas desviaciones que vendrían a ser los puntos que se

separan de la línea roja y por ultimo observamos como los grupos de datos se encuentran de forma horizontal y no de forma vertical que muestra que se acumulan valores en esas zonas, en conclusión el diagrama no muestra que los datos siguen una distribución normal.

6. Bibliografía

- Dela Cruz, A., & Jackson, C. (2023, noviembre 21). *Unimodal & bimodal distributions: Definition & examples*. Study.com. Recuperado de Study.com
- Waples, J. (2024, 18 de noviembre). *Q-Q plot*. DataCamp. Recuperado de DataCamp