Entendimiento del Negocio y Entendimiento de los Datos

Juan Pablo Arias Buitrago
Kevin Santiago Calderón Sánchez
Juan Andrés López Escalante
Paula Andrea Velásquez Romero
Juan José Álvarez Ortiz

Entrega 1: 1 Parcial

033702: Procesamiento de Datos a Gran Escala
Ing. John Jairo Corredor Franco, PhD

Pontificia Universidad Javeriana
Facultad de Ciencias
Bogotá D.C.

07 de marzo de 2025

Contenido

Justificación

En la actualidad, los avances tecnológicos han transformado diversos sectores, permitiendo una mayor toma de decisiones basada en datos. El procesamiento de grandes volúmenes de datos, también conocido como Big Data, ha cobrado una relevancia significativa, ya que facilita la obtención de Insights valiosos para la mejora de procesos y la optimización de recursos. Este proyecto se enmarca en un escenario de análisis de datos aplicado al ámbito educativo, específicamente en el ámbito territorial, donde se busca evaluar los resultados de los exámenes de ICFES en relación con la infraestructura de servicios de Internet y los niveles de pobreza de cada municipio. Este trabajo se enfoca en utilizar herramientas de Big Data como Apache Spark, con el fin de ofrecer una solución a preguntas de negocio específicas, además de generar un plan de acción para mejorar ciertos indicadores relacionados con los resultados de los exámenes. En particular, el proyecto aplica la metodología CRISP-DM, que proporciona una guía estructurada para el desarrollo de proyectos de análisis de datos.

Contexto del Proyecto

Este trabajo se en el contexto de una iniciativa educativa promovida por el Ministerio de Educación, cuyo objetivo es mejorar los resultados de los exámenes de ICFES a nivel municipal. A través del análisis de datos de diversas fuentes, se pretende identificar factores clave que puedan estar influyendo negativamente en estos resultados, y proponer soluciones basadas en evidencia. En este sentido, uno de los factores más relevantes es la infraestructura de servicios de Internet, que puede estar limitando el acceso a recursos educativos digitales, fundamental para el desarrollo académico de los estudiantes. Además, la pobreza a nivel municipal es otro factor que podría estar correlacionado con los bajos resultados en el ICFES. Este análisis se lleva a cabo por un equipo de consultoría formado por estudiantes de Ciencia de Datos, quienes aplican sus conocimientos para abordar este desafío y proponer un plan de acción que se ajuste a las necesidades del Ministerio de Educación.

Objetivo General

Realizar un análisis de los datos proporcionados por el Ministerio de Educación para identificar los factores que afectan los resultados de ICFES en los municipios, con el fin de diseñar un plan de acción basado en el procesamiento de grandes volúmenes de datos, que permita mejorar dichos resultados. Para ello, se utilizarán herramientas de Big Data como Apache Spark y se seguirá la metodología CRISP-DM para asegurar un enfoque estructurado y eficiente en el desarrollo del proyecto.

Entendimiento del Negocio y de los Datos

Problema Educativo

En Colombia, la calidad educativa presenta diferencias notables entre grandes ciudades como Bogotá y Medellín, y ciudades más pequeñas como Armenia y Neiva. Estas últimas, ubicadas en regiones con menores recursos y oferta académica, enfrentan desafíos significativos en comparación con las principales urbes del país. Factores como el acceso a recursos educativos, la conectividad a internet y el nivel socioeconómico de las familias influyen en los resultados de la prueba ICFES 11, evidenciando la desigualdad educativa en distintas zonas. Estos municipios fueron elegidos por su representatividad, calidad de datos y potencial para ofrecer un análisis contrastante. Además, se considerarán indicadores macroeconómicos como desempleo, pobreza e inversión pública para comprender mejor los desafíos educativos específicos y ofrecer una representación objetiva de la situación en otros municipios del país.

Objetivos Específicos

- Analizar la relación entre los resultados de ICFES y los indicadores macroeconómicos seleccionados.
- Desarrollar modelos predictivos para identificar factores clave que afectan los resultados de ICFES.
- Diseñar un plan de acción para mejorar los resultados de ICFES en los municipios con peores desempeños.

Conjunto de Datos Seleccionados

Considerando las bases de datos proporcionadas, se seleccionaron los siguientes conjuntos de datos para obtener información clave sobre la relación entre diversos factores y los resultados educativos. A continuación, se detallan las bases de datos y la razón de su selección en función de los objetivos de negocio.

Internet por Municipio

El acceso a Internet es uno de los factores clave en la calidad educativa, especialmente hoy donde las plataformas en línea y los materiales educativos digitales son fundamentales para el aprendizaje. Este conjunto de datos permitirá evaluar la relación entre la penetración de internet fijo en la población y los resultados de los exámenes ICFES.

Educación por Municipio

Este conjunto de datos es fundamental para evaluar la situación educativa en cada municipio, permitiendo identificar municipios con mayores dificultades en términos de cobertura y calidad educativa. A través de estas estadísticas, se puede explorar cómo las características del sistema educativo, como el número de estudiantes por institución o el tipo de cobertura, influyen en los resultados de la prueba ICFES.

Índice de Pobreza de Hogares por Persona

La pobreza es un factor determinante en el rendimiento académico, ya que limita el acceso a recursos educativos y afecta la capacidad de los estudiantes para concentrarse y estudiar de manera efectiva. En particular, la pobreza multidimensional tiene un impacto significativo en

la calidad de la educación, ya que abarca diversas privaciones que pueden influir en el desempeño escolar.

Uno de los factores clave es la **privación por logro educativo**, que mide el bajo nivel educativo en el hogar y puede afectar el apoyo académico que reciben los estudiantes. Asimismo, la **privación por analfabetismo** y la **privación por inasistencia escolar** reflejan las dificultades de acceso y permanencia en el sistema educativo, lo que puede llevar a tasas más altas de deserción y menor rendimiento en pruebas como el ICFES.

Otros factores estructurales incluyen la **privación por rezago escolar**, que evidencia las dificultades para avanzar en los niveles educativos adecuados según la edad, y la **privación por acceso a servicios para el cuidado de la primera infancia**, que puede impactar el desarrollo cognitivo y habilidades básicas en los primeros años de vida. Además, la **privación por desempleo de larga duración** y la **privación por empleo formal** afectan la estabilidad económica de los hogares, lo que a su vez influye en la posibilidad de costear materiales escolares, transporte y tecnología educativa.

La precariedad en las condiciones de vida también juega un rol importante. La **privación por falta de aseguramiento en salud** y las **barreras de acceso a salud** pueden generar problemas de salud no atendidos que afecten la asistencia y el rendimiento escolar.

Evaluar la relación entre estos indicadores y los resultados de los exámenes ICFES permitirá identificar los municipios más afectados por la pobreza y cómo estos factores influyen en la calidad educativa. Esto facilitará el diseño de estrategias específicas para mejorar las condiciones de aprendizaje y reducir las brechas en el acceso a la educación de calidad.

Resultados ICFES 11 por Municipio y Departamento

Los resultados de la prueba ICFES son el principal indicador de la calidad educativa en Colombia. Este conjunto de datos es esencial para analizar el rendimiento académico a nivel municipal y departamental, lo que permite correlacionar los puntajes obtenidos con otros factores como: nivel de estudio de los padres, número de personas con las cuales vives, acceso a internet, acceso a un computador, estudio en colegio bilingüe, estudio en colegio oficial, etc.

Ficha de Inversión Municipal PP

La inversión pública en educación es clave para mejorar la calidad de los servicios educativos. A través de estos datos, se puede evaluar la relación entre los recursos invertidos en educación y los resultados obtenidos en la prueba ICFES, permitiendo identificar si mayores inversiones en infraestructura educativa, formación docente u otros aspectos pueden tener un impacto positivo en el rendimiento académico.

Colección y Descripción de Datos

El análisis de la desigualdad educativa en Colombia se basa en la recopilación de datos provenientes de fuentes oficiales y abiertas, tales como datos abiertos de Colombia y el DANE (Departamento Administrativo Nacional de Estadística). Estos datos permiten realizar una evaluación detallada de la calidad educativa en diferentes regiones del país y su relación con diversos factores socioeconómicos.

Además, el estudio integra información de otras bases de datos gubernamentales y académicas que aportan un contexto más amplio sobre las condiciones educativas, económicas y sociales de los estudiantes. Esto posibilita la identificación de patrones, tendencias y desigualdades en el acceso a la educación.

Tipos de Datos Utilizados

El conjunto de datos empleado en el análisis incluye diversas variables que permiten examinar el desempeño académico y las condiciones socioeconómicas que influyen en él. Entre los tipos de datos más relevantes se encuentran:

- **Datos educativos:** Información sobre el rendimiento académico de los estudiantes en la prueba ICFES Saber 11, utilizada como referencia para medir la calidad de la educación en las distintas regiones del país.
- **Datos socioeconómicos:** Indicadores como el nivel de ingresos de las familias, el estrato socioeconómico y la disponibilidad de recursos educativos en los hogares.
- **Indicadores macroeconómicos:** Factores como el desempleo, la tasa de pobreza y la inversión pública en educación, los cuales influyen directamente en el acceso y la calidad de la enseñanza.

- **Infraestructura y conectividad:** Datos sobre el acceso a Internet, la disponibilidad de tecnología en los hogares y las condiciones de las instituciones educativas en cada región.

Estos datos han sido organizados y procesados para facilitar su análisis, asegurando su calidad mediante la limpieza y estructuración de las bases de datos.

Descripción General del Contenido de los Conjuntos de Datos

El conjunto de datos analizado incluye registros de estudiantes de diversas regiones de Colombia, permitiendo identificar disparidades en la calidad educativa según el contexto socioeconómico y las condiciones de infraestructura escolar. Para obtener información relevante, se ha llevado a cabo un Análisis Exploratorio de Datos (EDA), que permite

- **Detectar valores atípicos:** Identificar datos que se desvían significativamente de la norma y pueden afectar el análisis.
- **Analizar distribuciones:** Examinar cómo se distribuyen las variables clave, como el puntaje ICFES en distintas regiones y grupos socioeconómicos.
- **Explorar correlaciones:** Identificar relaciones entre variables, como el impacto del nivel socioeconómico en el rendimiento académico.

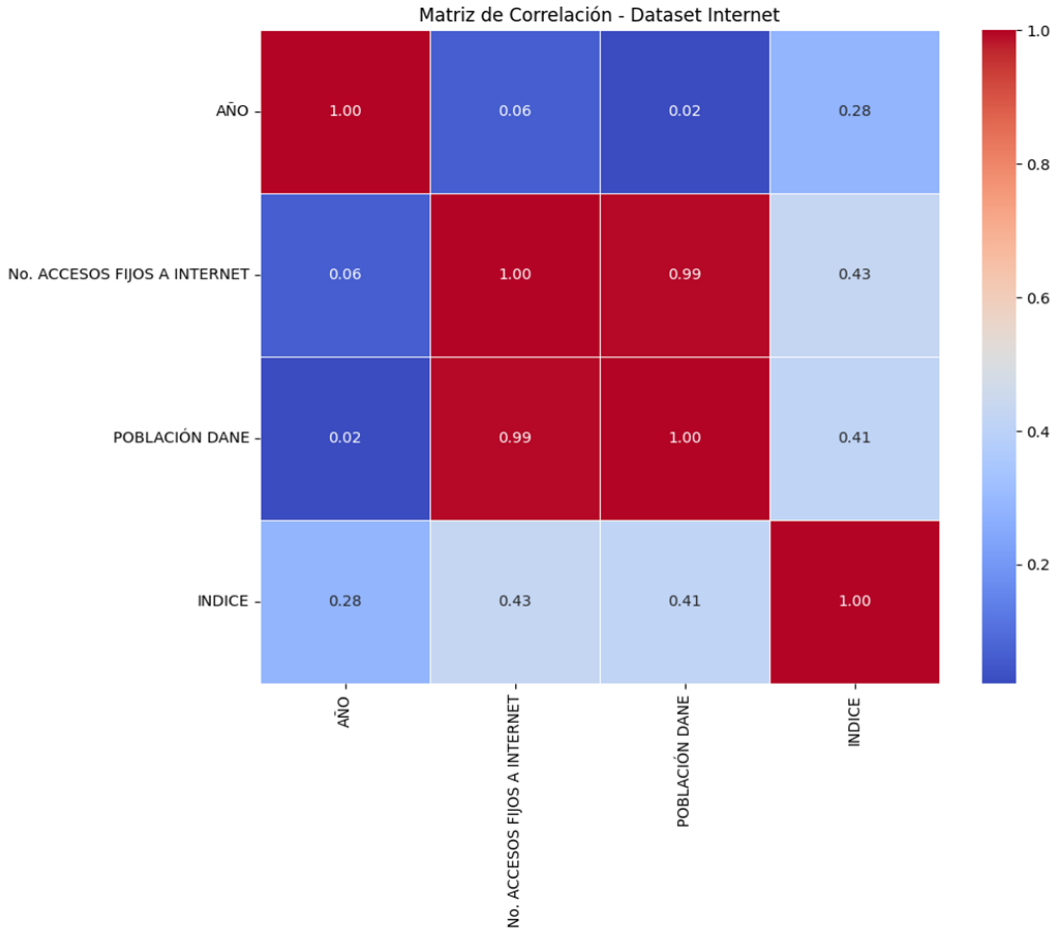
El análisis de estos datos no solo permite comprender las diferencias en la calidad educativa en Colombia, sino también generar estrategias y políticas públicas orientadas a reducir la brecha de desigualdad en el acceso a la educación de calidad.

Exploración de los datos

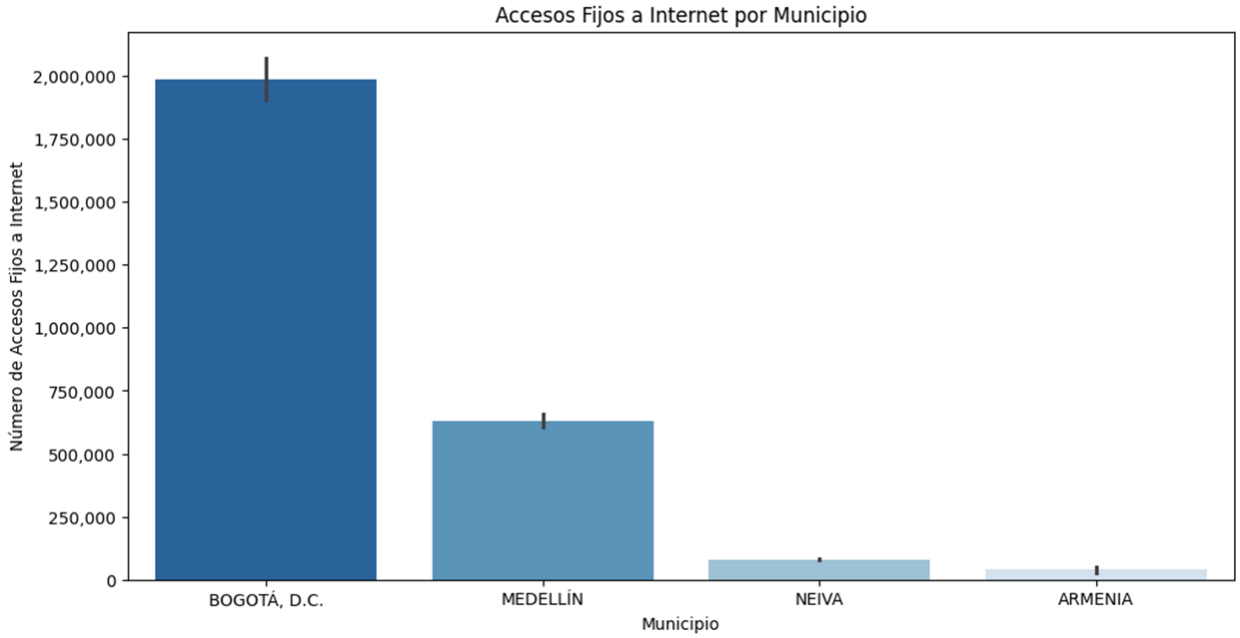
Internet por Municipio

El análisis de estos datos revela una cobertura temporal equilibrada desde 2016 hasta 2022, con una distribución de trimestres homogénea. Los accesos fijos a internet presentan una media de aproximadamente 555,370, pero con una alta desviación estándar, lo que sugiere una gran variabilidad entre municipios. La población también muestra una amplia dispersión, con valores que oscilan entre 5,085 y 7.87 millones de habitantes. En cuanto a la unicidad de los datos, hay 7 valores únicos para el año, 4 para el trimestre, 4 para el código de departamento, 5 para el código de municipio, 139 para los accesos fijos a internet, 35 para la población y 133 para el índice. No hay valores faltantes en ninguna columna, lo que indica que el conjunto de datos está completo. Además, no se encontraron filas duplicadas, lo que confirma la consistencia en la estructura del dataset.

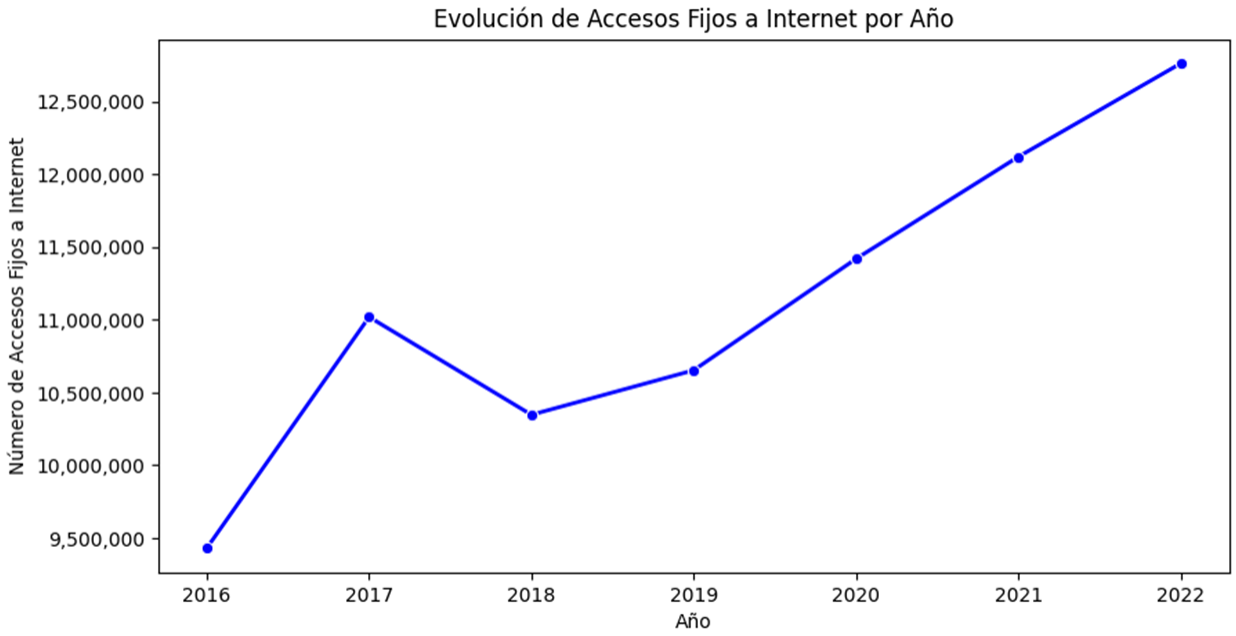
El análisis de los gráficos de cajas y bigotes muestra que tanto los accesos fijos a internet como la población municipal tienen distribuciones sesgadas con valores atípicos altos, indicando que unos pocos municipios concentran gran parte de la conectividad y la población. La mayoría de los municipios tienen bajos accesos a internet en relación con su población. En contraste, la variable **AÑO** está uniformemente distribuida entre 2016 y 2022, sin valores extremos. Esto sugiere una cobertura de datos consistente en el tiempo, mientras que la conectividad muestra desigualdades significativas entre municipios.



La matriz de correlación muestra la relación entre las variables del dataset de acceso a internet. Se observa una correlación casi perfecta (0.99) entre el número de accesos fijos a internet y la población, lo que indica que en municipios con mayor población, el número de accesos a internet tiende a ser mayor. También hay una correlación moderada entre el índice y los accesos fijos a internet (0.43) y entre el índice y la población (0.41), sugiriendo que estos factores influyen en la penetración de internet. En contraste, la variable "AÑO" muestra correlaciones débiles con el resto de las variables, con valores de 0.06 para accesos fijos, 0.02 para población y 0.28 para el índice, lo que indica que el tiempo tiene una relación baja con la variabilidad de estos datos. En general, la gráfica sugiere que la disponibilidad de internet está fuertemente ligada al tamaño de la población, mientras que otros factores tienen menor impacto en la variabilidad de los accesos.



La gráfica muestra el número de accesos fijos a internet en distintos municipios, destacando una fuerte concentración en Bogotá, D.C., seguido por Medellín, mientras que Neiva y Armenia tienen valores significativamente más bajos. Esta distribución es esperada debido a la relación positiva entre la población y el acceso a internet, como se observó en la matriz de correlación. Bogotá, al ser la ciudad más grande del país, lidera en cantidad de accesos, mientras que Medellín, otro centro urbano importante, ocupa el segundo lugar. En contraste, Neiva y Armenia, con poblaciones más reducidas, presentan menor cantidad de accesos. Esto sugiere que el acceso a internet fijo está fuertemente influenciado por el tamaño de la población y la infraestructura tecnológica disponible en cada municipio.

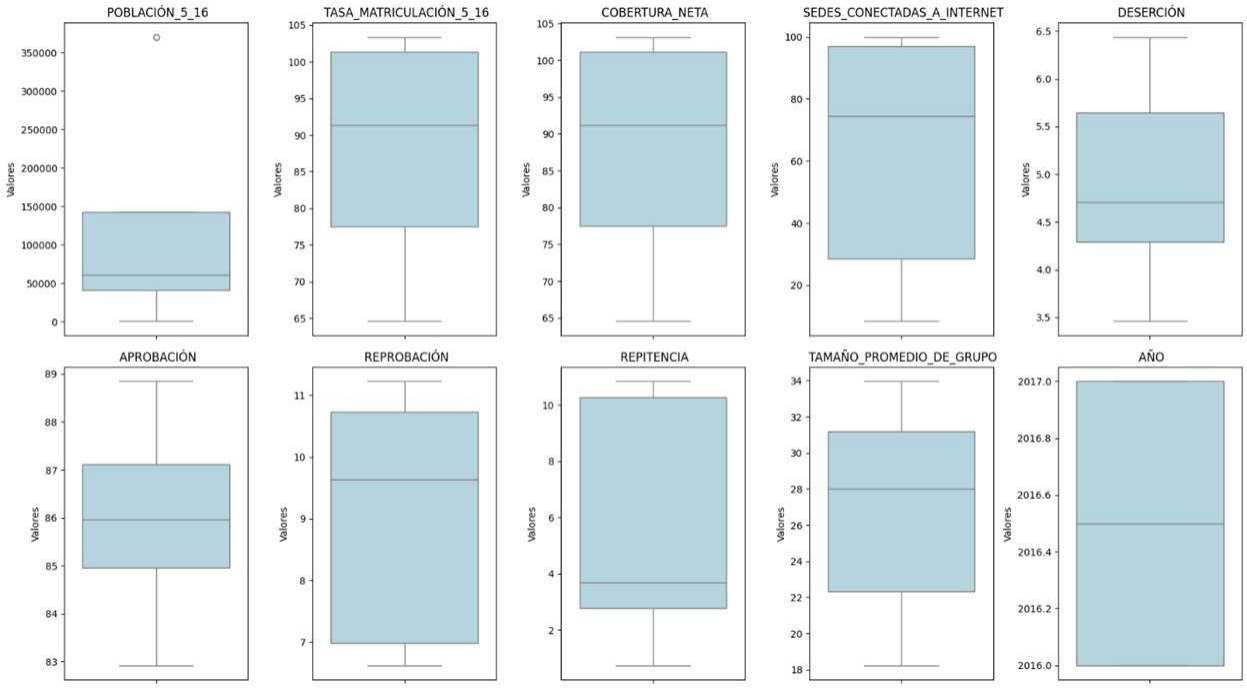


La gráfica muestra la evolución del número de accesos fijos a internet entre los años 2016 y 2022. Se observa una tendencia general al alza, con un crecimiento significativo desde 2016 hasta 2017, seguido de una leve disminución en 2018. A partir de 2019, la cantidad de accesos vuelve a aumentar de manera sostenida, alcanzando su punto más alto en 2022.

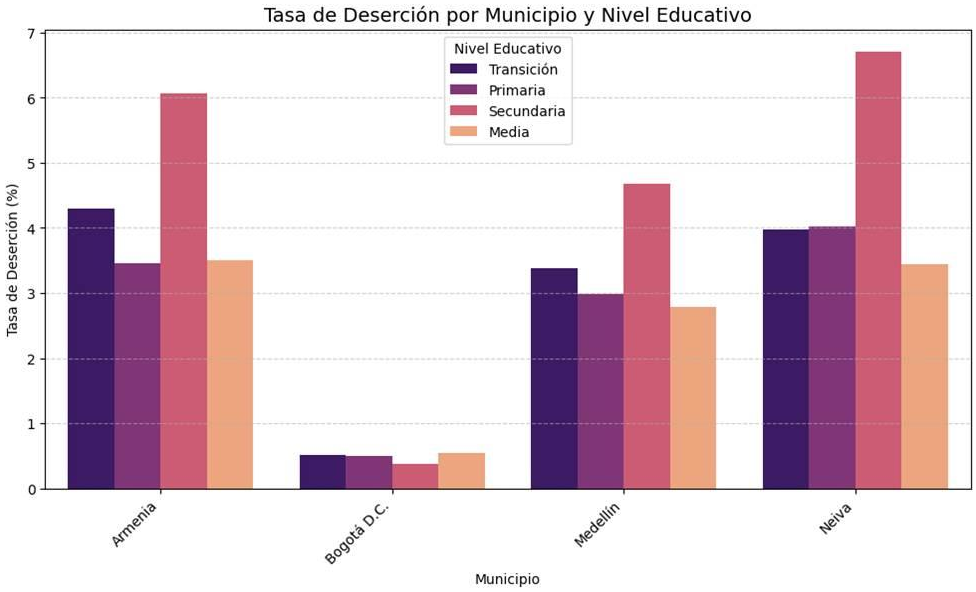
Educación por Municipio

El análisis de los indicadores educativos muestra una tendencia positiva en la cobertura y matriculación de estudiantes entre 5 y 16 años, con un aumento sostenido en los últimos años.

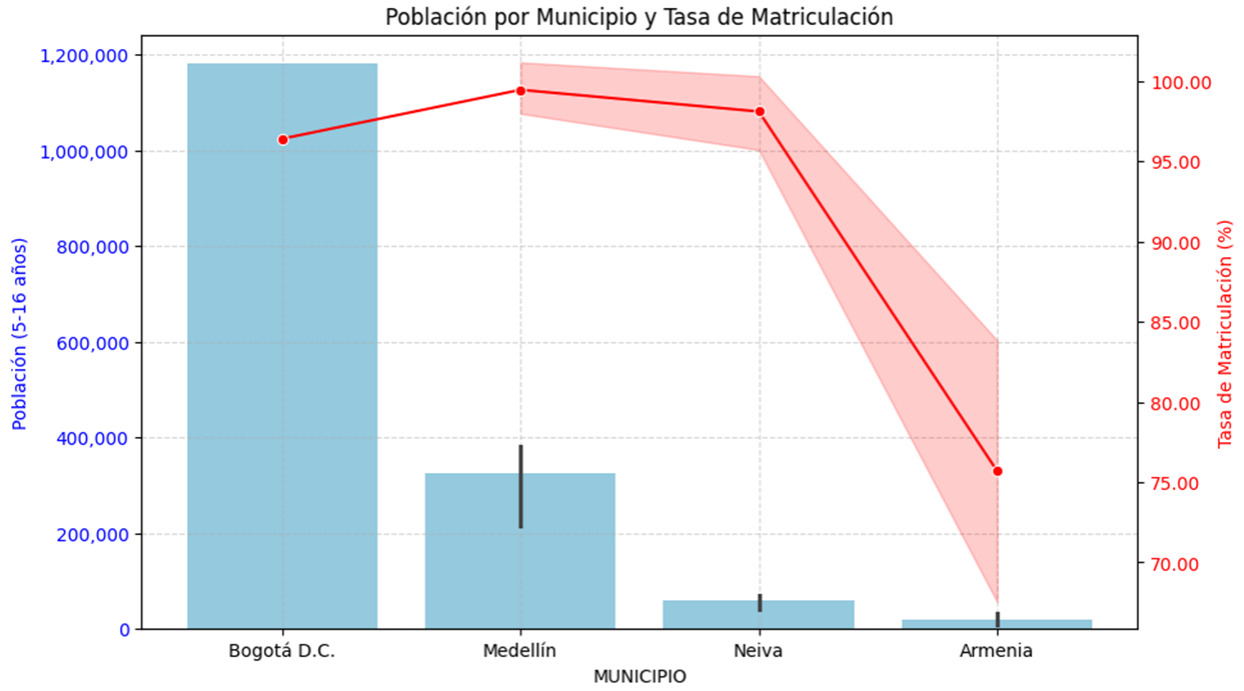
La cobertura neta en transición, primaria y secundaria ha mejorado, aunque en la educación media sigue siendo más baja en comparación con los otros niveles. A pesar de estos avances, aún hay desafíos importantes, como la deserción y la repitencia, que afectan principalmente a los niveles de secundaria y media. La tasa de reprobación también es un factor preocupante, ya que es más alta en secundaria, lo que indica dificultades en la permanencia y éxito académico de los estudiantes en esta etapa. Además, el acceso a internet en las sedes educativas es limitado, lo que puede afectar la calidad de la educación y el acceso a recursos digitales. Estos datos reflejan la necesidad de implementar estrategias para mejorar la retención escolar, fortalecer la calidad educativa y garantizar una mayor equidad en el acceso a herramientas tecnológicas.



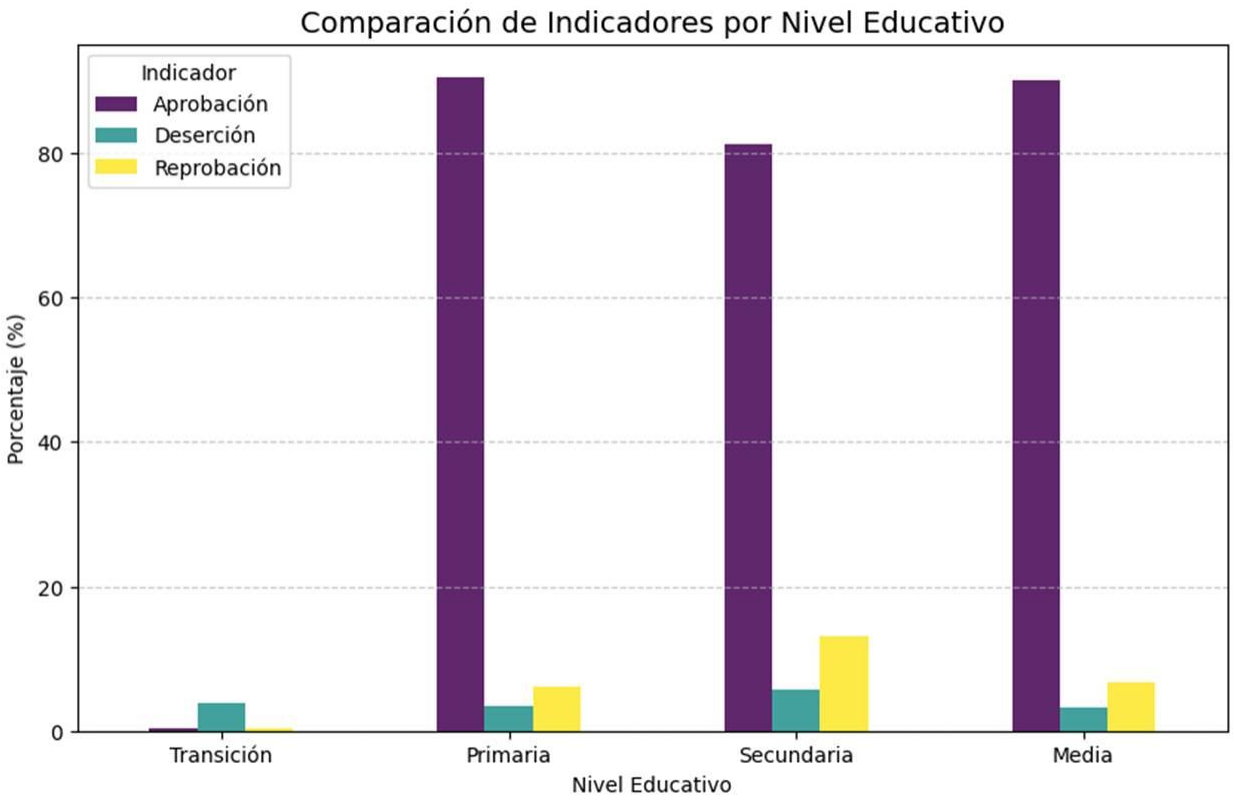
El gráfico muestra la distribución de diversas variables educativas mediante diagramas de caja. Se observa que la población entre 5 y 16 años varía significativamente entre regiones, con algunos valores atípicos elevados. La tasa de matriculación y cobertura neta son altas, pero presentan variabilidad en ciertas áreas. La conectividad a internet en sedes educativas es desigual, con una mediana en torno al 50%. La deserción se mantiene en niveles bajos (3.5%-6.5%), mientras que la aprobación ronda el 86%, aunque con tasas de reprobación y repitencia moderadas en algunas regiones. El tamaño promedio de los grupos es de 28-30 estudiantes, reflejando diferencias en infraestructura escolar. Los datos corresponden a los años 2016-2017, evidenciando desafíos en equidad educativa, conectividad y retención estudiantil.



El gráfico muestra la tasa de deserción escolar por municipio y nivel educativo. Se observa que Neiva y Armenia tienen los valores más altos, especialmente en secundaria, con tasas que superan el 6%. Medellín presenta tasas intermedias, con un pico en secundaria, mientras que Bogotá D.C. tiene la deserción más baja en todos los niveles, con valores cercanos al 1%. En general, la secundaria es el nivel con mayor deserción en la mayoría de los municipios, mientras que la transición y la educación media presentan tasas más moderadas. Esto sugiere que la permanencia en el sistema educativo es más crítica en la educación secundaria, lo que podría estar relacionado con factores socioeconómicos o barreras de acceso en estas etapas.



El gráfico muestra la población de niños y adolescentes entre 5 y 16 años por municipio (barras azules, eje izquierdo) y la tasa de matriculación correspondiente (línea roja, eje derecho). Bogotá D.C. tiene la población más alta, superando 1.2 millones, y una tasa de matriculación cercana al 100%. Medellín, con una población significativamente menor, también mantiene una tasa alta. Neiva, a pesar de tener una población mucho menor que Bogotá, presenta una tasa de matriculación ligeramente superior, lo que sugiere una mayor cobertura proporcional en el sistema educativo. En contraste, Armenia tiene la población más baja y la menor tasa de matriculación, cercana al 75%. La caída en la tasa de matriculación a medida que la población disminuye sugiere que los municipios más pequeños pueden enfrentar mayores dificultades en el acceso y permanencia en el sistema educativo.

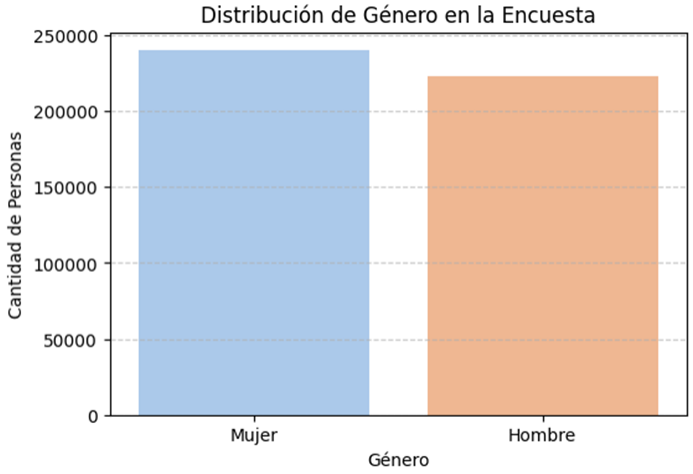


El gráfico muestra que la aprobación es el indicador predominante en todos los niveles educativos, con valores superiores al 80%, aunque disminuye ligeramente en secundaria. La deserción se mantiene baja y sin variaciones significativas. La reprobación, en cambio, es más baja en transición y primaria, aumenta notablemente en secundaria y disminuye nuevamente en media. Esto reafirma que la secundaria es la etapa con mayores desafíos académicos para los estudiantes.

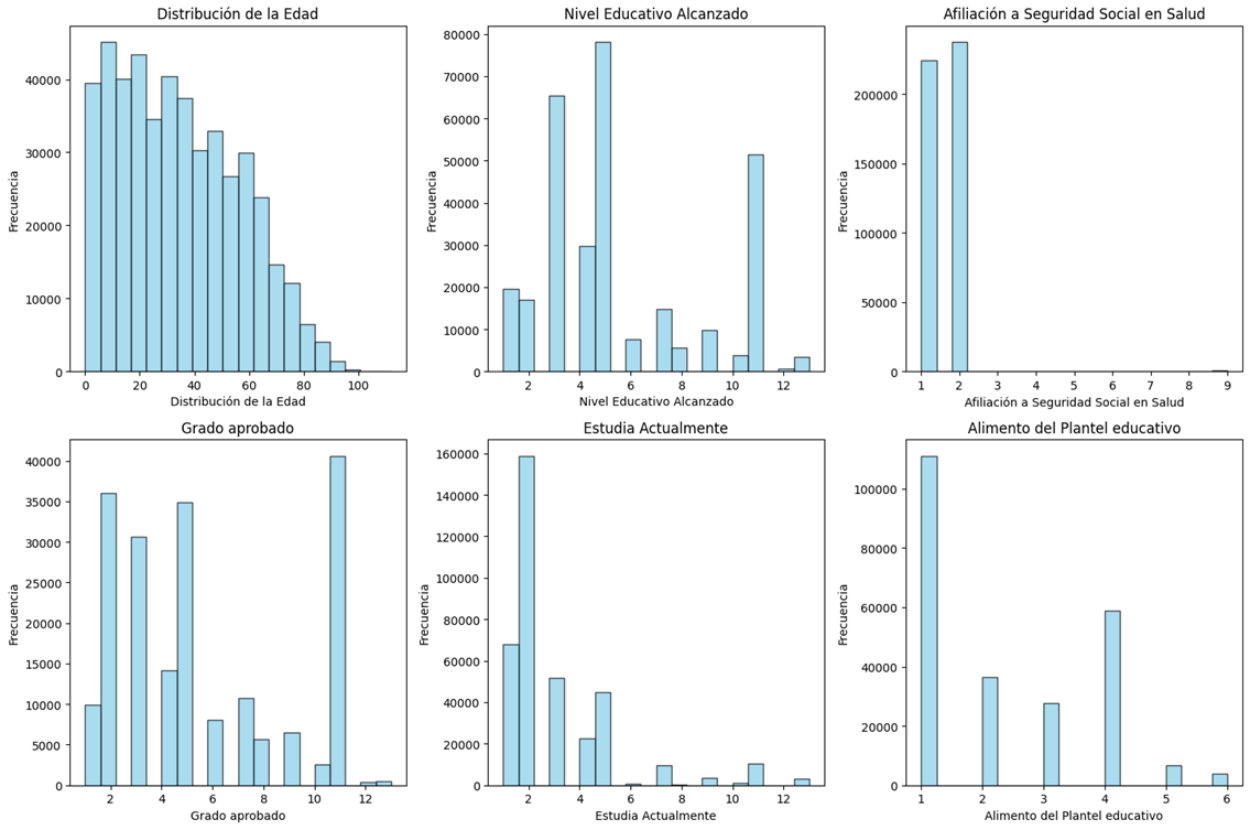
Índice de Pobreza de Hogares por Persona

El análisis descriptivo de los datos revela información sobre la estructura de la base de datos, la distribución de las variables y la calidad de los datos. La base de datos contiene 462,884 registros en 28 columnas, con variables relacionadas con demografía, salud, educación, empleo y condiciones de vida. La media de edad (P6040) es de 34 años, con una desviación estándar de 22 años, mostrando una distribución amplia. En cuanto a género (P6020), se observa un equilibrio con una media cercana a 1.5 (posiblemente indicando codificación binaria). Variables de salud y empleo presentan alta variabilidad y valores atípicos, como P6090 (afiliación a seguridad social), con un rango amplio. La variable FEX_C (factor de expansión) muestra una alta dispersión,

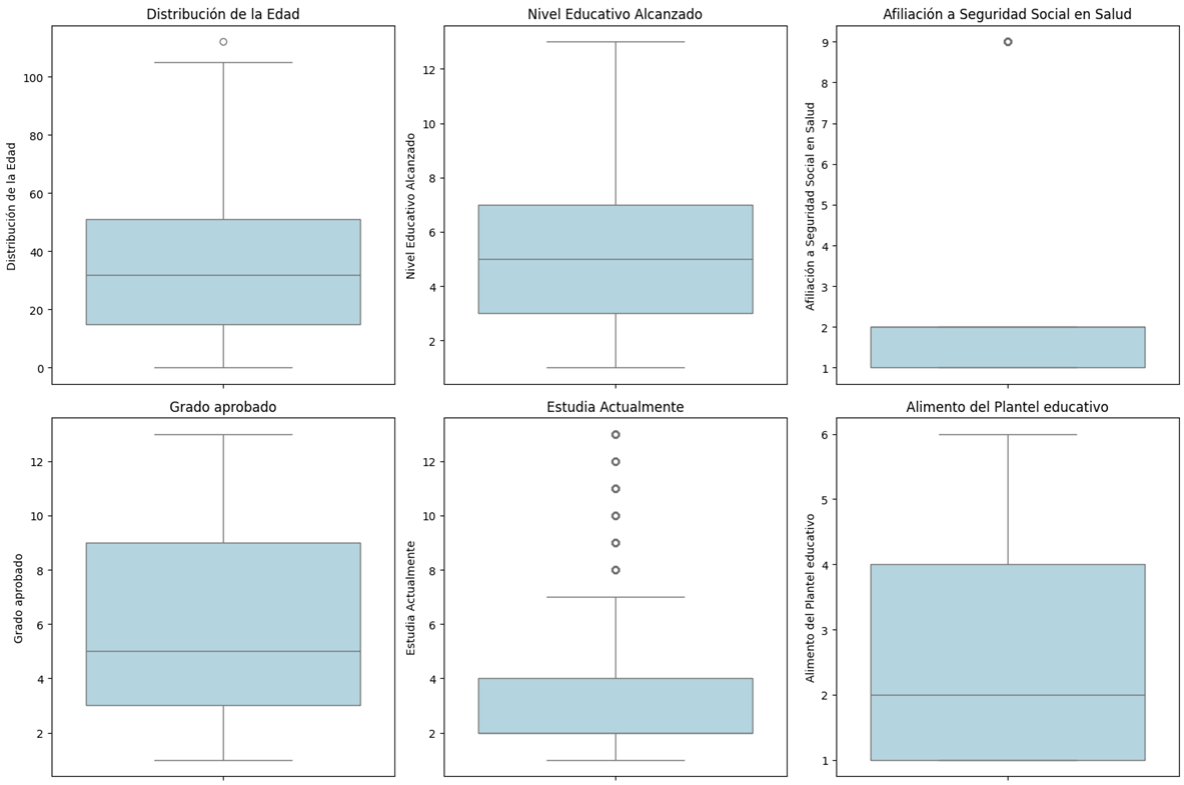
indicando diferencias significativas en la ponderación de la muestra. Existen múltiples valores nulos en varias variables, como P8563 (problemas de salud) y P6180 (acceso a alimentación en instituciones educativas), lo que puede afectar el análisis de tendencias y correlaciones. Además, variables categóricas como P6240 (actividad principal) tienen un número limitado de categorías únicas, lo que sugiere que los datos están estructurados en opciones predefinidas. La presencia de valores extremos en algunas variables (como P7250, semanas buscando empleo, con un máximo de 520) podría influir en el análisis de empleo. En general, la base de datos proporciona una visión detallada de las condiciones socioeconómicas de la población encuestada, aunque la cantidad de datos faltantes y la dispersión de ciertos valores podrían requerir limpieza y normalización antes de realizar inferencias más precisas.



El gráfico muestra la distribución de género en una encuesta, donde la cantidad de mujeres supera ligeramente a la de hombres. La muestra total supera las 400,000 personas, con una proporción aproximada de 53-55% mujeres y 45-47% hombres. La diferencia, aunque no muy grande, es significativa y podría reflejar una mayor participación femenina en el estudio o la composición demográfica del área encuestada.



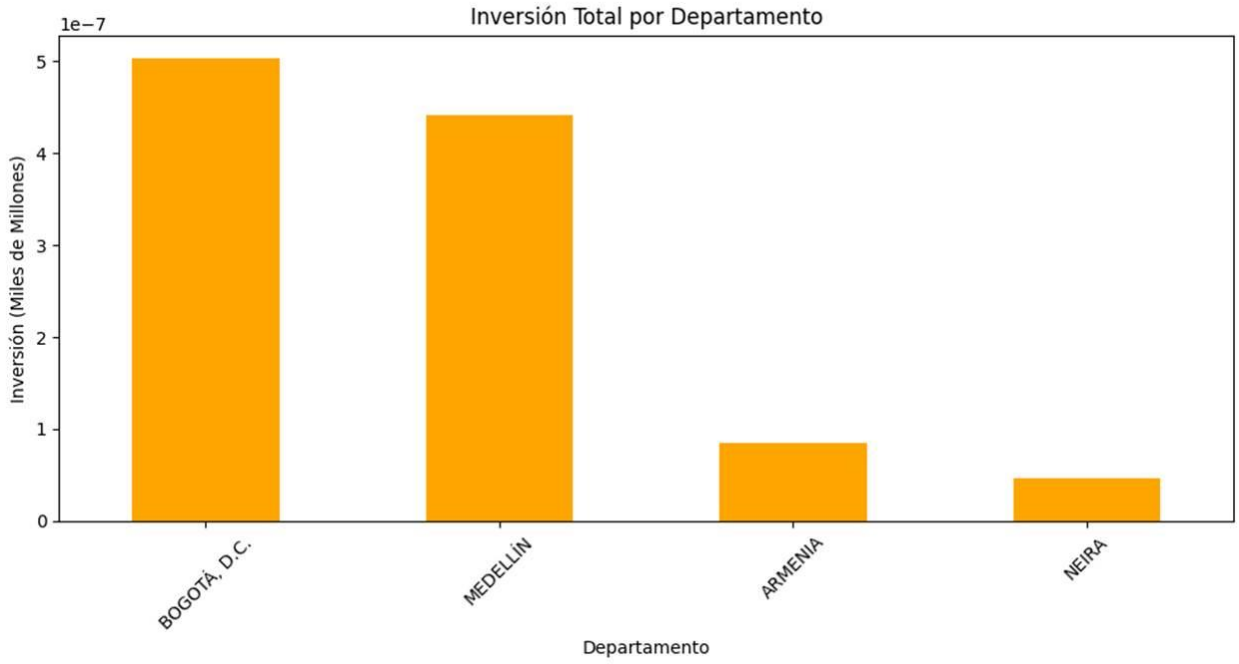
Los gráficos muestran una población mayoritariamente joven, con alta afiliación a la seguridad social en salud y predominancia en los niveles educativos básicos y secundarios, aunque con menor presencia en educación superior. La mayoría no estudia actualmente, pero muchos han recibido alimentación en sus planteles educativos. Estos datos reflejan una estructura poblacional con buen acceso a educación y salud, aunque con desafíos en la continuidad educativa en niveles superiores.



Los diagramas de caja muestran la distribución y dispersión de las variables. La edad tiene una mediana cercana a los 30 años, con valores atípicos en edades avanzadas. El nivel educativo alcanzado y el grado aprobado tienen distribuciones centradas en la educación básica y media, con pocos casos en educación superior. La afiliación a seguridad social está altamente concentrada en una categoría, con algunos valores atípicos. La variable "Estudia actualmente" presenta varios valores extremos, indicando que algunos siguen estudiando más allá de la media. Finalmente, la distribución del acceso a alimentos en el plantel educativo es amplia, con una mediana baja, sugiriendo que no todos los estudiantes reciben este beneficio.

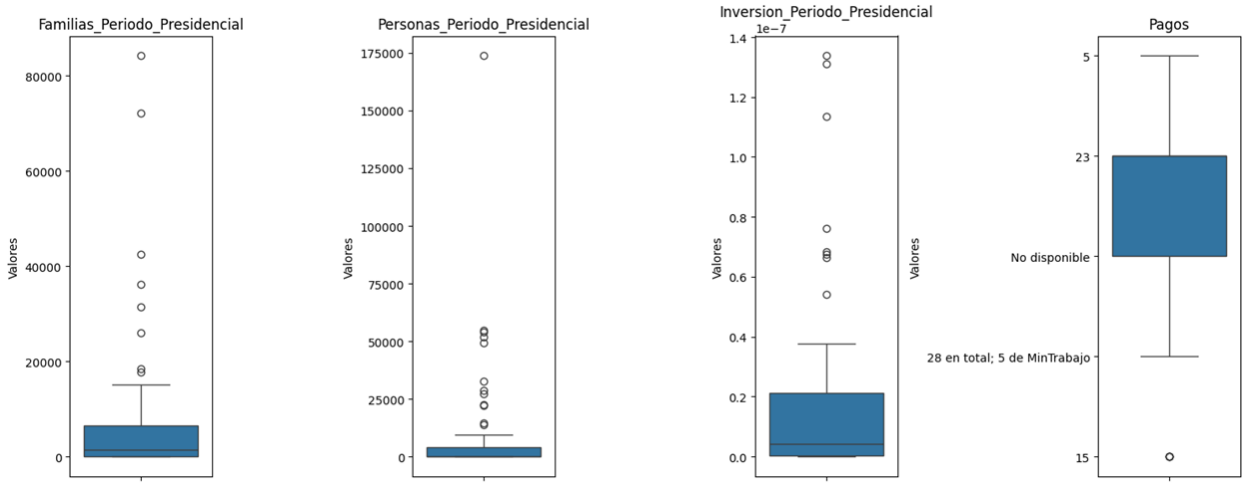
Ficha de Inversión Municipal PP

El conjunto de datos contiene 56 registros sin valores nulos ni filas duplicadas. La inversión durante el periodo presidencial varía ampliamente, con una media de aproximadamente 19.2 mil millones y un máximo que supera los 133.7 mil millones, reflejando una gran dispersión en los recursos asignados. La cantidad de familias y personas beneficiadas también presenta una distribución sesgada, con valores medianos relativamente bajos en comparación con sus respectivos máximos (84,185 familias y 173,698 personas). Además, hay una notable concentración en algunos valores, dado que solo existen 42 valores únicos en la columna de familias beneficiadas y 33 en la de personas atendidas. La variabilidad en los códigos de departamento y municipios es baja, lo que sugiere que los datos provienen de un número limitado de regiones. En cuanto a los programas y entidades, hay una diversidad de iniciativas (16 programas y 12 descripciones únicas), aunque con una sola entidad reportada, lo que indica centralización en la gestión.

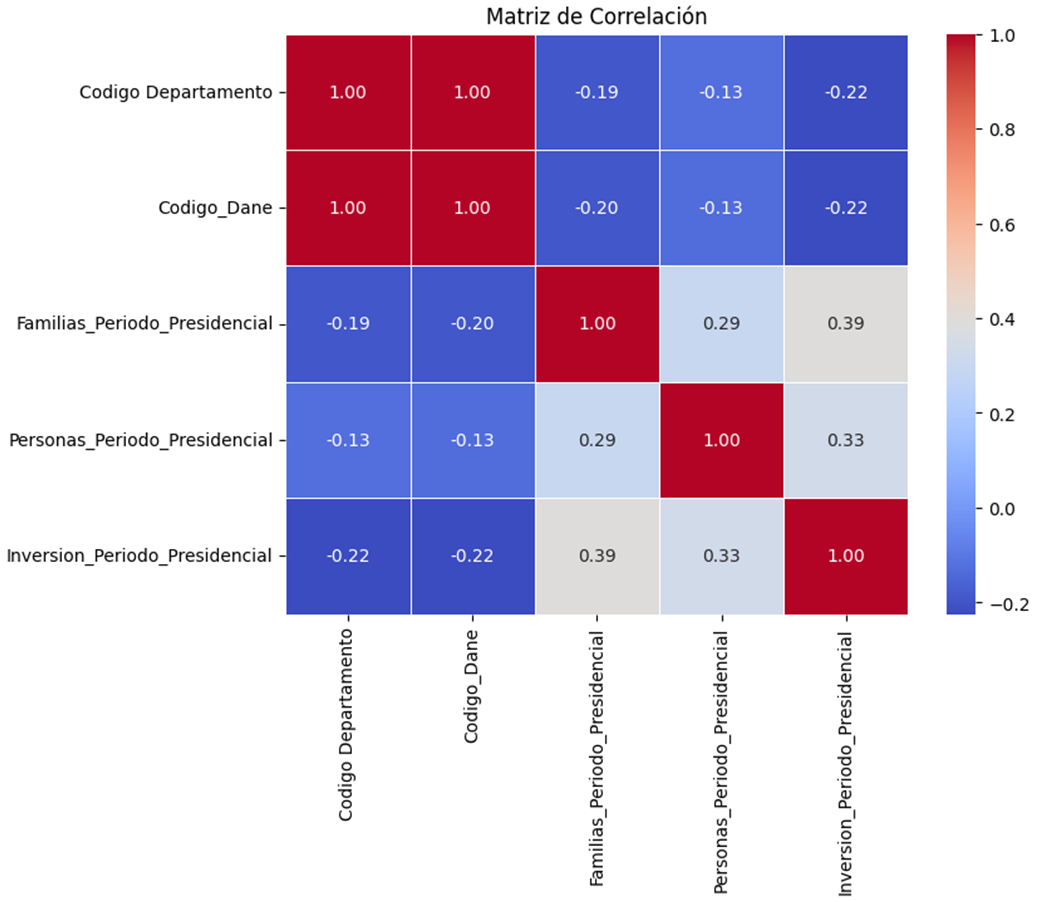


El gráfico muestra la inversión total por departamento en miles de millones, con una distribución desigual de los recursos. Bogotá, D.C. y Medellín concentran la mayor parte de la inversión, con Bogotá liderando con la cifra más alta. Armenia y Neira presentan inversiones significativamente menores, con una diferencia notable en comparación con las dos primeras

ciudades. Esta distribución sugiere una fuerte centralización de los recursos en las principales ciudades del país, mientras que otras regiones reciben una inversión comparativamente baja. Además, la diferencia entre Bogotá y Medellín respecto a los otros dos departamentos es considerable, lo que podría reflejar factores como densidad poblacional, necesidades específicas o prioridades gubernamentales en la asignación de recursos.



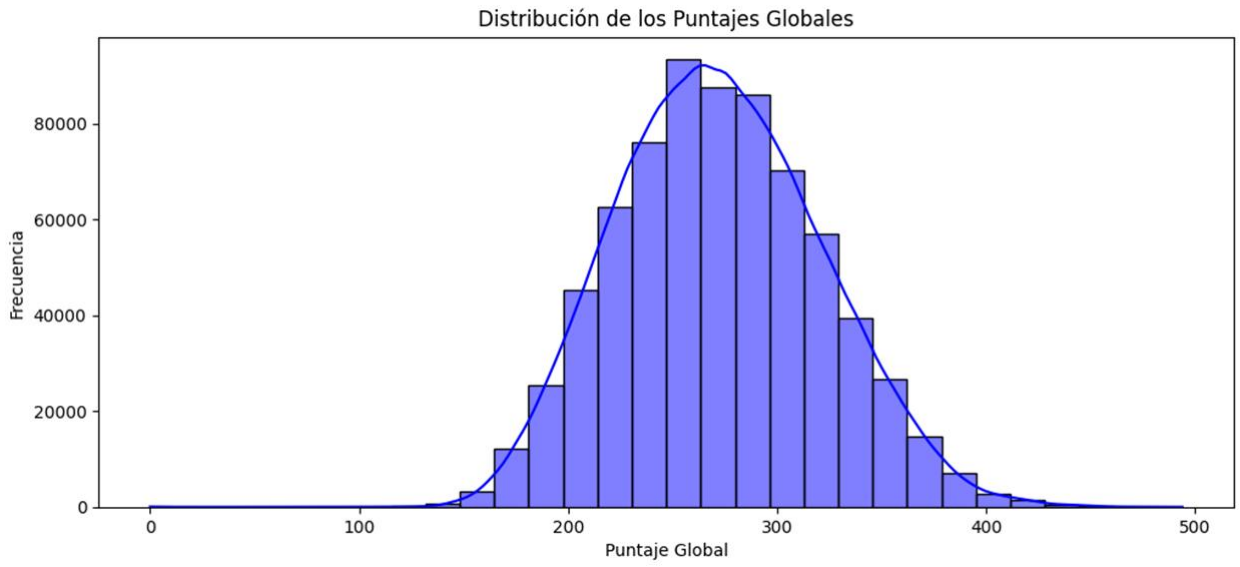
Los diagramas de caja muestran distribuciones sesgadas con valores atípicos altos en todas las variables analizadas (familias, personas e inversión en el período presidencial). La mayoría de los valores son bajos, pero algunos casos extremos elevan significativamente el promedio, lo que indica una distribución desigual del impacto. En el caso de los pagos, la dispersión es notable y parece incluir datos categóricos. En general, los datos sugieren que la inversión y el beneficio no son homogéneos, sino que están concentrados en ciertos grupos o regiones.



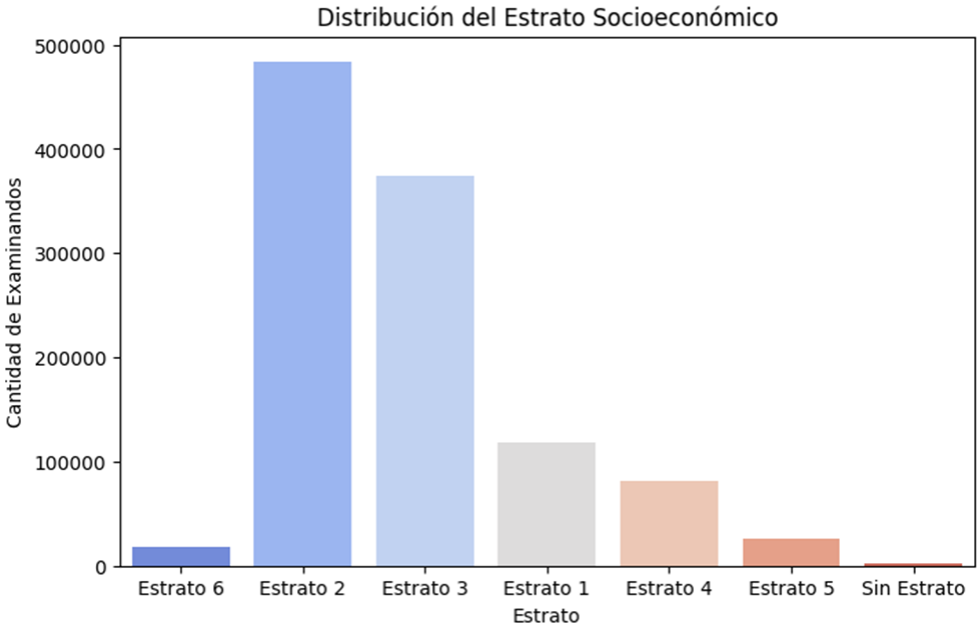
La matriz de correlación muestra una relación positiva entre el número de familias y personas beneficiadas con la inversión en el período presidencial, aunque con una intensidad moderada (0.39 y 0.33, respectivamente). Sin embargo, los códigos de departamento y DANE tienen correlaciones negativas débiles con todas las variables, lo que sugiere que la asignación de inversión y beneficiarios no está directamente vinculada a la ubicación geográfica. En general, la relación entre inversión y beneficiarios indica cierta coherencia, pero la dispersión sugiere otros factores que influyen en la distribución de los recursos.

Resultados ICFES 11

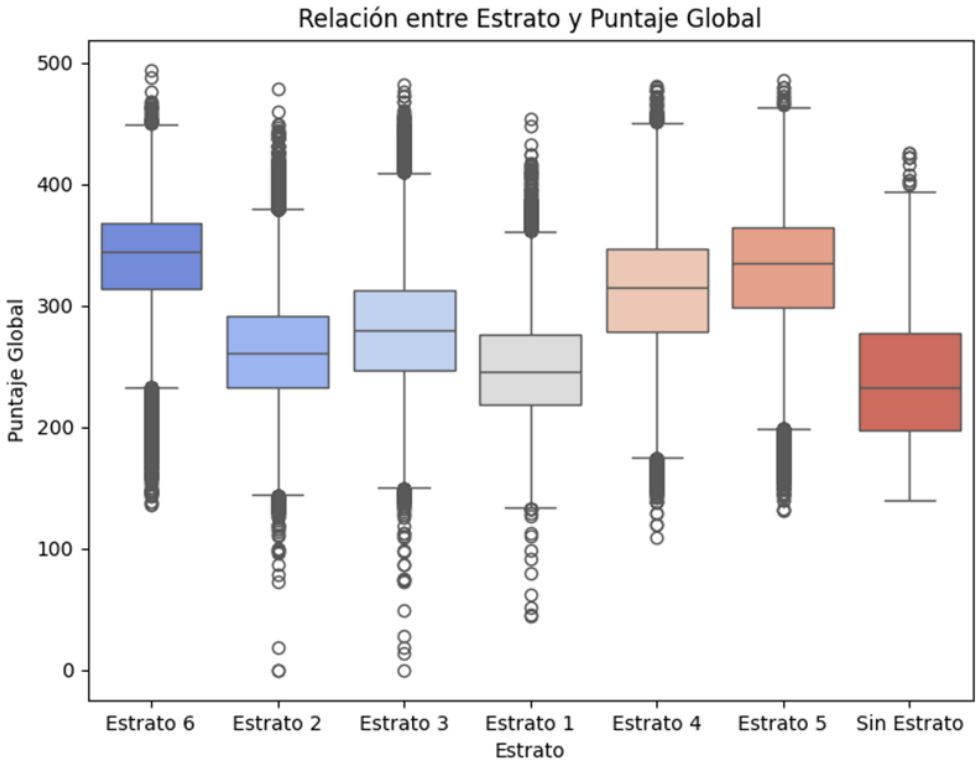
Los datos corresponden a un gran conjunto de registros (más de un millón) relacionados con el desempeño académico en pruebas estandarizadas, donde se incluyen códigos de instituciones, municipios y departamentos, así como puntajes en diversas áreas. La media de los puntajes oscila entre 52 y 55, con una desviación estándar de aproximadamente 10-13 puntos, indicando una dispersión moderada en los resultados. Se observan valores atípicos con puntajes negativos o en cero, lo que sugiere posibles errores en la captura de datos. Además, algunas variables presentan un alto número de valores nulos, lo que podría afectar ciertos análisis. No se encontraron filas duplicadas, lo que sugiere un adecuado manejo de la base de datos en términos de unicidad de registros.



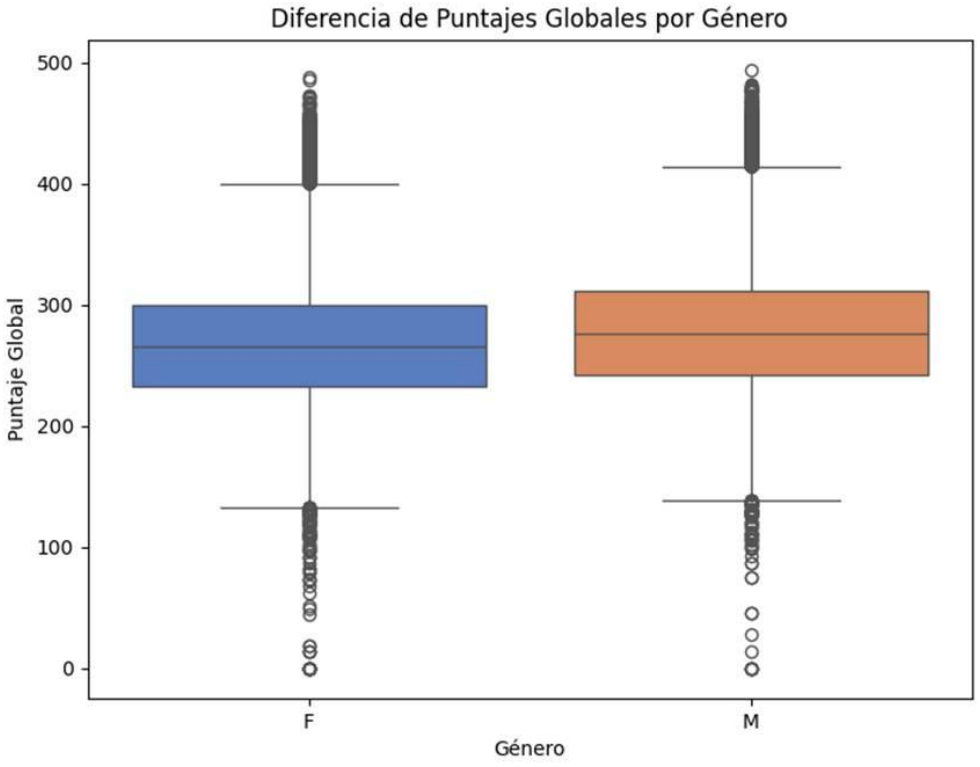
La distribución de los puntajes globales es aproximadamente normal, con mayor concentración entre 200 y 350 puntos y un pico alrededor de 275-300. La curva muestra una ligera asimetría hacia la derecha, indicando la presencia de algunos puntajes excepcionalmente altos. La mayoría de los estudiantes obtienen puntajes en el rango medio, mientras que los valores extremos son menos frecuentes. Esto sugiere un desempeño generalmente equilibrado, aunque con algunos casos sobresalientes o con dificultades.



La gráfica muestra la distribución de los examinados según su estrato socioeconómico, evidenciando una marcada concentración en los estratos 2 y 3, que en conjunto representan la mayoría de los participantes. El estrato 2 es el más numeroso, con una cantidad de examinados cercana a los 500,000, seguido por el estrato 3, con una cifra también elevada, pero menor en comparación. En contraste, los estratos 1, 4 y 5 presentan una cantidad significativamente menor de examinados, con el estrato 1 siendo el más representativo dentro de este grupo. Por otro lado, el estrato 6 y la categoría "Sin Estrato" tienen una presencia casi marginal, lo que indica que muy pocas personas de los sectores más altos o no clasificados participaron en la evaluación. Estos resultados sugieren que el acceso a este tipo de exámenes está fuertemente influenciado por la distribución socioeconómica, con una mayor representación de personas de ingresos bajos y medios, posiblemente debido a factores como acceso a educación, oportunidades y expectativas académicas.



La gráfica de cajas muestra la relación entre el estrato socioeconómico y el puntaje global obtenido en el examen. Se observa una tendencia en la que los puntajes tienden a ser más altos en los estratos más altos (estrato 6, 5 y 4), mientras que los estratos más bajos (1, 2 y 3) presentan medianas menores y una mayor dispersión en los puntajes. En particular, el estrato 6 tiene la mediana más alta y una menor variabilidad en comparación con los demás, lo que sugiere que los estudiantes de este grupo obtienen resultados más consistentes y elevados. Por otro lado, el estrato 1 muestra una mediana significativamente inferior, con una gran cantidad de valores atípicos en los extremos inferiores, lo que indica que muchos estudiantes de este grupo obtienen puntajes bajos. El estrato "Sin Estrato" presenta una distribución amplia, pero su mediana es más baja en comparación con los demás. En general, la gráfica sugiere una posible correlación entre el nivel socioeconómico y el rendimiento académico, donde los estudiantes de estratos más altos tienden a obtener mejores puntajes, lo que podría estar relacionado con el acceso a mejores oportunidades educativas y recursos de aprendizaje.



La gráfica de cajas compara la distribución de los puntajes globales en función del género (F: femenino, M: masculino). A simple vista, se observa que la mediana de los puntajes es similar en ambos grupos, lo que indica que no hay una diferencia sustancial en el rendimiento central entre hombres y mujeres. Sin embargo, la dispersión de los datos muestra algunas diferencias notables.

Ambos grupos presentan una distribución relativamente simétrica, con un rango intercuartílico (IQR) parecido, lo que sugiere que la mayoría de los estudiantes, independientemente de su género, obtienen puntajes dentro de un rango similar. No obstante, en la parte superior de la distribución, hay una leve tendencia a que los hombres tengan valores ligeramente más altos en el puntaje global, aunque la diferencia no parece ser muy marcada.

En la parte inferior de la distribución, se observa que ambos grupos presentan valores atípicos (outliers), lo que indica la existencia de estudiantes con puntajes extremadamente bajos.

En conclusión, la gráfica sugiere que no hay una diferencia significativa entre los puntajes globales obtenidos por hombres y mujeres. Aunque puede haber ligeras variaciones en la dispersión de los datos y en los valores más altos alcanzados, la tendencia general indica que el rendimiento académico, en términos de puntaje global, es comparable entre ambos grupos.

Reporte de Calidad de Datos

En este apartado se presenta un análisis de la calidad de los datos utilizados en el proyecto, con el fin de identificar los data sets con valores faltantes, inconsistencias o posibles problemas en las bases de datos seleccionadas. Además, se proponen estrategias para el tratamiento de estos problemas, asegurando la integridad y confiabilidad del análisis.

Conteo de Valores Faltantes

Durante la exploración de los conjuntos de datos, se identificaron los siguientes valores faltantes:

- ***Internet por Municipio:*** Se encontraron valores nulos en el porcentaje de acceso a Internet en algunos municipios, lo que puede afectar el análisis de correlación con los resultados de las pruebas ICFES.
- ***Educación por Municipio***: Algunas variables relacionadas con la cobertura educativa presentan registros incompletos, especialmente en municipios con menor densidad poblacional.
- ***Índice de Pobreza de Hogares por Persona***: Existen datos ausentes en ciertos indicadores de pobreza multidimensional, lo que puede influir en la precisión de los modelos predictivos.

Estrategias para el Tratamiento de Valores Faltantes

Para mitigar el impacto de los valores faltantes, se proponen las siguientes estrategias:

Para valores numéricos:

- ***Media*:** Se utilizará cuando la distribución de la variable sea aproximadamente normal. Por ejemplo, si el puntaje promedio de una prueba en un municipio falta para algunos registros, se reemplazará con la media de los valores conocidos.
- ***Mediana*:** Si los datos presentan una distribución sesgada o con valores extremos, se utilizará la mediana, ya que es menos sensible a los outliers. Por ejemplo, si hay

valores faltantes en el índice de pobreza y la distribución es asimétrica, se empleará la mediana del conjunto de datos.

Para valores categóricos:

- **Moda (categoría más frecuente):** Para variables como el tipo de institución (pública/privada) o el acceso a Internet (sí/no), se llenarán los valores faltantes con la categoría más frecuente en el conjunto de datos.

- **Asignación basada en patrones:** Si ciertos municipios tienen tendencias similares en ciertas variables (por ejemplo, municipios rurales suelen tener menor acceso a Internet), se imputarán los valores según los patrones observados en datos similares

Eliminación de Registros Incompletos

Si un registro tiene una cantidad significativa de valores faltantes (más del 50% de sus variables), su eliminación puede ser la mejor opción para evitar sesgos en el análisis. Se aplicarán los siguientes criterios:

- **Eliminación de filas con más del 50% de datos ausentes:** Esto es útil cuando la falta de información impide un análisis confiable. Por ejemplo, si un municipio tiene valores faltantes en la mayoría de las variables educativas y socioeconómicas, se eliminará del conjunto de datos.

- **Eliminación de columnas con demasiados valores faltantes:** Si una variable presenta más del 60-70% de datos ausentes, se evaluará su eliminación, ya que su inclusión podría introducir ruido en el análisis.

Interpolación de Datos:

Para variables continuas con valores faltantes en función de una progresión lógica o geográfica, se aplicará interpolación. Este método estima los valores en función de los datos disponibles en registros cercanos.

- Interpolación Lineal: Se empleará para variables como el acceso a Internet, donde los datos pueden variar gradualmente entre municipios cercanos. Si en un municipio faltan datos sobre conectividad, pero los municipios adyacentes tienen información, se estimará el valor en función de una progresión lineal.
- Interpolación Polinómica o Spline: Para series de datos donde la evolución en el tiempo o en el espacio no es lineal, se podrá utilizar interpolación polinómica de segundo o tercer grado.
- Estimación basada en datos geográficos: En casos donde la interpolación matemática no sea adecuada, se podrá utilizar métodos basados en la distancia entre municipios o regiones con características similares.

Detección de Inconsistencias

Se identificaron algunos posibles errores en los datos, como diferencias en la denominación de municipios y valores extremos en ciertas variables. Para corregir esto, se debe:

- Estandarización de nombres de municipios para evitar registros duplicados.
- Revisión y eliminación de valores atípicos mediante técnicas estadísticas como el uso de percentiles o la desviación estándar.

Conclusiones y Próximos Pasos

La calidad de los datos es un factor crítico para la validez del análisis. El siguiente paso es aplicar las estrategias mencionadas y realizar una validación posterior para verificar la efectividad del tratamiento de los valores faltantes y de las inconsistencias detectadas.

Filtros, Limpieza y Transformación Inicial

En esta sección se describen los procesos de limpieza, filtrado y transformación inicial aplicados a cada conjunto de datos, con el fin de mejorar su calidad y asegurar su utilidad en el análisis posterior.

ICFES Bogotá

Se eliminaron las columnas con más del 35% de valores nulos para evitar sesgos en el análisis. Los valores faltantes en variables numéricas fueron imputados con la media, mientras que en variables categóricas se reemplazaron con la moda. Se aplicó el método del rango Inter cuartil (IQR) para eliminar valores atípicos en las variables numéricas.

ICFES Medellín

En general, el conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto.

ICFES Neiva

En general, el conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto. Además, se identificaron datos fuera de los límites de puntuación establecidos, como valores superiores a 100 en ciertas áreas, a pesar de que este era el máximo permitido.

ICFES Armenia

En general, el conjunto de datos se encontraba en buenas condiciones; sin embargo, presentaba una cantidad significativa de valores nulos en variables esenciales para el desarrollo del proyecto. Además, se identificaron datos fuera de los límites de puntuación establecidos, como valores superiores a 100 en ciertas áreas, a pesar de que este era el máximo permitido.

Inversión

No presentó valores nulos, pero se realizó una verificación de valores atípicos en la variable de inversión por periodo presidencial. Se evaluó la inversión per cápita como una métrica adicional para futuros análisis.

Pobreza

Presentó un alto porcentaje de valores nulos en varias columnas, algunas superando el 95%. Se eliminaron aquellas con más del 35% de datos faltantes, mientras que las variables restantes fueron imputadas con la media (para valores numéricos) y la moda (para valores categóricos). Se aplicó el método IQR para detectar y eliminar valores atípicos en las variables numéricas.

Internet

No contenía valores nulos, pero se transformó la variable de accesos a internet, calculando una métrica de accesos por cada 1,000 habitantes para permitir comparaciones

equitativas entre municipios. Se eliminaron valores atípicos en esta métrica utilizando el método IQR.

Estos procesos aseguran que los datos utilizados en el análisis sean más representativos y confiables, minimizando el impacto de datos inconsistentes o extremos en los resultados del estudio.

Educación

En general, el conjunto de datos contenía múltiples columnas irrelevantes para el proyecto, incluyendo algunas que presentaban información sobre educación en niveles de transición y primaria, los cuales no eran pertinentes para el análisis. Además, se identificaron incongruencias en ciertas columnas, como discrepancias en los totales de estudiantes, donde el número total reportado era menor que la cantidad de estudiantes en bachillerato. Asimismo, se encontraron columnas con un alto porcentaje de datos nulos. Debido a estos factores, se decidió eliminar dichas columnas para garantizar la coherencia y calidad del análisis.

Planteamiento de Preguntas sobre los Datos

Condiciones Sociales y Pobreza

1. ¿Cuál es el porcentaje de hogares en condiciones de hacinamiento crítico en cada municipio y cómo se compara con la media nacional?
2. ¿Cuáles son las diferencias en la tasa de empleo formal entre municipios con altos y bajos niveles de pobreza multidimensional?
3. ¿Cuál es el porcentaje de población sin aseguramiento en salud en cada municipio y qué relación tiene con la tasa de mortalidad infantil?

Educación y Acceso a la Educación

4. ¿En qué municipios se presenta la mayor privación por rezago escolar y cómo afecta esto el acceso a educación superior en estos territorios?
5. ¿Cuál es el porcentaje de jóvenes en edad escolar secundaria que no están asistiendo a los centros educativos en los municipios?
6. ¿Cuáles son los municipios con menor acceso a tecnología educativa y qué impacto tiene en los resultados académicos?
7. ¿Cómo afecta la falta de conectividad a internet la tasa de escolaridad en los municipios?

Infraestructura y Desarrollo Municipal

8. ¿Existe una relación directa entre el nivel de inversión municipal en infraestructura y la mejora en las condiciones de vida de la población vulnerable?

Referencias

- Rodríguez, J. D. & Vidal, G. E. (2022). *Factores asociados con la brecha digital en los resultados de las Pruebas Saber 11 en el departamento del Quindío para el periodo 2021-2*. Recuperado de, http://hdl.handle.net/10554/63005.
- Alba Lorena Ballesteros-Alfonso1 Ph. D. Nubia Yaneth Gómez-Velasco. (s/f). *Desigualdad de resultados pruebas Saber-11 antes y durante la pandemia covid-19*. Scielo.org. Recuperado el 5 de marzo de 2025, de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1692-715X2022000300046
- *¿En qué consiste la limpieza de los datos?* (s/f). Amazon.com. Recuperado el 5 de marzo de 2025, de https://aws.amazon.com/es/what-is/data-cleansing/
- Haider, K. (2024, enero 2). ¿Qué es la calidad de los datos y por qué es importante? *Astera*. https://www.astera.com/es/type/blog/data-quality/
- ¿Qué es la transformación de datos? (2024, octubre 10). *Ibm.com*. https://www.ibm.com/es-es/think/topics/data-transformation
- *R Para Ciencia de Datos - 7 Análisis exploratorio de datos (EDA)*. (s/f). Hadley.Nz. Recuperado el 7 de marzo de 2025, de https://es.r4ds.hadley.nz/07-eda.html
- Reinoso, B., & Paula, M. (2015). *Efecto del nivel de pobreza colombiana en los resultados del ICFES, PRUEBA SABER 11*. Recuperado el 4 de marzo de 2025, de https://intellectum.unisabana.edu.co/handle/10818/21522 Universidad de la Sabana.
- *RPubs - La brecha académica en las pruebas ICFES: ¿Un reflejo de la desigualdad en Colombia?* (s/f). Rpubs.com. Recuperado el 7 de marzo de 2025, de https://rpubs.com/Foodweb/La-brecha-academica-en-las-pruebas-ICFES