Pontificia Universidad Javeriana

Proyecto de Procesamiento de Datos a Gran Escala

Entrega 1

Grupo n

Andres Felipe Galan Cardenas

Diego Alejandro Martínez Oviedo

Julian Andrey Mendez Rodriguez

Samuel Andres Lamilla Verjan

Steven Viscillinovick Robles Patiño

1. Entendimiento del negocio

El proyecto a realizar consiste en un análisis de datos a gran escala de múltiples conjuntos de datos, con la finalidad de identificar una relación entre los resultados de la prueba ICFES 11 y distintos factores como cobertura del servicio de internet, índices de pobreza en la región, niveles de educación, entre otros, para los municipios de Cundinamarca para el año 2023. El análisis se realiza con la finalidad de generar un plan de acción estratégico que optimice dichos indicadores y mejore la calidad educativa en Cundinamarca.

Para esta proyecto se toman en cuenta los treinta municipios con más población del departamento de Cundinamarca para el año 2023, los cuales corresponden a Soacha, Facatativá, Fusagasugá, Chía, Zipaquirá, Mosquera, Madrid, Girardot, Funza, Cajicá, Villa de San Diego de Ubaté, Tocancipá, Sibaté, Cota, La Mesa, Guaduas, La Calera, Villeta, Sopó, Pacho, El Colegio, Cogua, Tenjo, Tabio, Silvania, El Rosal, Chocontá, Suesca, Gachancipá y Villapinzón [1]; La muestra se restringe a los treinta municipios con más población con la finalidad de proporcionar un análisis más centrado y significativo en un periodo temporal cercano a la actualidad

1.1 Índice de Desempeño Fiscal (IDF)

Es una medición del desempeño de la gestión financiera de entidades territoriales, como municipios, departamentos u otros. El IDF es una herramienta que permite evaluar la sostenibilidad financiera de un territorio específico [2], y se puede clasificar dependiendo del puntaje que obtenga; las clasificaciones son las siguientes:

- **Sostenible:** Las entidades con más de 80 puntos, que poseen finanzas saludables, cumplen con límites legales de deuda y gasto, generan recursos propios y tienen una alta capacidad para proveer bienes y servicios a largo plazo se clasifican como “sostenibles” [3].

- **Solvente:** Las entidades que poseen puntajes entre 70 y 80 puntos, que tienen finanzas saludables pero tienen oportunidades de mejora se clasifican como “solventes” [3].

- **Vulnerable:** Las entidades que registran puntajes entre 60 y 70 puntos, que pueden cumplir con los límites legales de deuda y gasto, pero que dependen en gran medida de transferencias y presentan bajos niveles de inversión en formación bruta de capital son catalogados como “vulnerables” [3].

- **Riesgo:** Las entidades que presentan puntajes entre 40 y 60 puntos, que enfrentan dificultades significativas en su gestión fiscal y podrían estar en riesgo de insostenibilidad financiera se clasifican como “riesgosos” [3].

1.2 Situación general por municipio

- **Soacha:** Para finales de 2023 contaba con una población de 782647 habitantes; poseía un puntaje 49.89 para el IDF. Según el puntaje, Soacha se encuentra en riesgo.
- **Facatativá:** Para finales de 2023 contaba con una población de 166588 habitantes; poseía un puntaje de 41.89 para el IDF. Según el puntaje, Facatativá se encuentra en riesgo.
- **Fusagasugá:** Para finales de 2023 contaba con una población de 165143 habitantes; poseía un puntaje de 46.10 para el IDF. Según el puntaje, Fusagasugá se encuentra en riesgo.
- **Chía:** Para finales de 2023 contaba con una población de 158258 habitantes; poseía un puntaje de 51.04 para el IDF. Según el puntaje, Chía se encuentra en riesgo.
- **Zipaquirá:** Para finales de 2023 contaba con una población de 155618 habitantes; poseía un puntaje de 50.73 para el IDF. Según el puntaje, Zipaquirá se encuentra en riesgo.
- **Mosquera:** Para finales de 2023 contaba con una población de 151244 habitantes; poseía un puntaje de 42.95 para el IDF. Según el puntaje, Mosquera se encuentra en riesgo.
- **Madrid:** Para finales de 2023 contaba con una población de 134736 habitantes; poseía un puntaje de 47.03 para el IDF. Según el puntaje, Madrid se encuentra en riesgo.
- **Girardot:** Para finales de 2023 contaba con una población de 199446 habitantes; poseía un puntaje de 42.90 para el IDF. Según el puntaje, Girardot se encuentra en riesgo.
- **Funza:** Para finales de 2023 contaba con una población de 111675 habitantes; poseía un puntaje de 54.17 para el IDF. Según el puntaje, Funza se encuentra en riesgo.
- **Cajicá:** Para finales de 2023 contaba con una población de 98441 habitantes; poseía un puntaje de 56.73 para el IDF. Según el puntaje, Facatativá se encuentra en riesgo.
- **Villa de San Diego de Ubaté:** Para finales de 2023 contaba con una población de 50581 habitantes; poseía un puntaje de 50.97 para el IDF. Según el puntaje, Villa de San Diego de Ubaté se encuentra en riesgo.
- **Tocancipá:** Para finales de 2023 contaba con una población de 46918 habitantes; poseía un puntaje de 62.38 para el IDF. Según el puntaje, Tocancipá se encuentra vulnerable.
- **Sibaté:** Para finales de 2023 contaba con una población de 39761 habitantes; poseía un puntaje de 45.81 para el IDF. Según el puntaje, Sibaté se encuentra en riesgo.
- **Cota:** Para finales de 2023 contaba con una población de 39070 habitantes; poseía un puntaje de 46.00 para el IDF. Según el puntaje, Cota se encuentra en riesgo.

- **La Mesa:** Para finales de 2023 contaba con una población de 38759 habitantes; poseía un puntaje de 60.35 para el IDF. Según el puntaje, La Mesa se encuentra vulnerable.
- **Guaduas:** Para finales de 2023 contaba con una población de 35904 habitantes; poseía un puntaje de 45.14 para el IDF. Según el puntaje, Guaduas se encuentra en riesgo.
- **La Calera:** Para finales de 2023 contaba con una población de 35317 habitantes; poseía un puntaje de 45.13 para el IDF. Según el puntaje, La Calera se encuentra en riesgo.
- **Villeta:** Para finales de 2023 contaba con una población de 31632 habitantes; poseía un puntaje de 50.84 para el IDF. Según el puntaje, Villeta se encuentra en riesgo.
- **Sopó:** Para finales de 2023 contaba con una población de 30780 habitantes; poseía un puntaje de 58.53 para el IDF. Según el puntaje, Sopó se encuentra en riesgo.
- **Pacho:** Para finales de 2023 contaba con una población de 28412 habitantes; poseía un puntaje de 52.36 para el IDF. Según el puntaje, Pacho se encuentra en riesgo.
- **El Colegio:** Para finales de 2023 contaba con una población de 28185 habitantes
- **Cogua:** Para finales de 2023 contaba con una población de 26055 habitantes; poseía un puntaje de 52.39 para el IDF. Según el puntaje, Cogua se encuentra en riesgo.
- **Tenjo:** Para finales de 2023 contaba con una población de 26012 habitantes; poseía un puntaje de 52.47 para el IDF. Según el puntaje, Tenjo se encuentra en riesgo.
- **Tabio:** Para finales de 2023 contaba con una población de 25692 habitantes; poseía un puntaje de 41.35 para el IDF. Según el puntaje, Tabio se encuentra en riesgo.
- **Silvania:** Para finales de 2023 contaba con una población de 25347 habitantes; poseía un puntaje de 46.47 para el IDF. Según el puntaje, Silvania se encuentra en riesgo.
- **El Rosal:** Para finales de 2023 contaba con una población de 25330 habitantes; poseía un puntaje de 42.78 para el IDF. Según el puntaje, El Rosal se encuentra en riesgo.
- **Chocontá:** Para finales de 2023 contaba con una población de 24144 habitantes; poseía un puntaje de 44.01 para el IDF. Según el puntaje, Chocontá se encuentra en riesgo.
- **Gachancipá:** Para finales de 2023 contaba con una población de 20142 habitantes; poseía un puntaje de 47.82 para el IDF. Según el puntaje, Gachancipá se encuentra en riesgo.
- **Suesca:** Para finales de 2023 contaba con una población de 20120 habitantes; poseía un puntaje de 47.87 para el IDF. Según el puntaje, Suesca se encuentra en riesgo.

- **Villapinzón:** Para finales de 2023 contaba con una población de 20062 habitantes; poseía un puntaje de 52.25 para el IDF. Según el puntaje, Villapinzón se encuentra en riesgo.

2. Selección de los datos a utilizar

2.1 Bases de datos

Se escogen cinco bases de datos para abordar el problema de negocio y poder responder a la problemática planteada:

- Internet Fijo Penetración Municipio
- Estadísticas en educación en preescolar, básica y media por municipio
- Pruebas ICFES
- Dirección de Descentralización y Fortalecimiento Fiscal
- Proyección de la población de cundinamarca 2023

2.2 Motivo de elección de las bases de datos

- **Internet Fijo Penetración Municipio:** Al momento de hablar de progreso académico el acceso a internet juega un papel fundamental dado que permite el acceso ilimitado a información de estudio; se quiere estudiar la relación de los resultados del ICFES respecto al municipio donde se registró el resultado, y como en Colombia no todos los municipios tienen acceso fijo a internet, se quiere observar si el acceso fijo a internet es un factor determinante al momento de evaluar los resultados de la prueba ICFES

- **Estadísticas en educación en preescolar, básica y media por municipio:** Esta base de datos permite examinar el sector escolar en los municipios elegidos; dado que se requiere hacer un análisis en el campo estudiantil para poder examinar el desempeño de los resultados de la prueba ICFES, la base de datos es ideal, ya que sus datos hacen referencia a la tasa de matriculación escolar, a la tasa de deserción escolar, a la tasa de reprobación escolar, entre otros.

- **Pruebas ICFES:** Esta base de datos es uno de los pilares de la investigación, ya que es necesario el conocer los resultados de la prueba ICFES para el año escogido en los municipios seleccionados, ya que sin la base de datos, no se podría realizar una relación entre los resultados y factores educativos/económicos de los municipios.

- **Dirección de Descentralización y Fortalecimiento Fiscal:** Para poder medir el desarrollo financiero; se toma esta base de datos para poder observar un factor económico en los municipios y poder observar si es un factor determinante al momento de los resultados de la prueba ICFES

- **Proyección de la población de cundinamarca 2023:** La base de datos posee la información de las población de los municipios en cundinamarca para el año 2023, es necesaria la proporción de población para poder comparar los resultados de la prueba ICFES entre los municipios evitando sesgos inesperados por la diferencia de población

3. Colección y descripción de datos

3.1 Carga y visualización de los datos en DataBricks

Primero se inicia el cluster (imagen 1), posterior a ello se empiezan a crear las tablas con las bases de datos (imagen 2) . Posteriormente se leen las tablas para corroborar que se cargaron correctamente (imagen 3) y se observan los tipos de variable de cada columna (imagen 4)

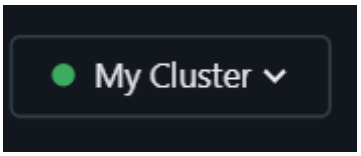


Imagen 1: Inicio de cluster

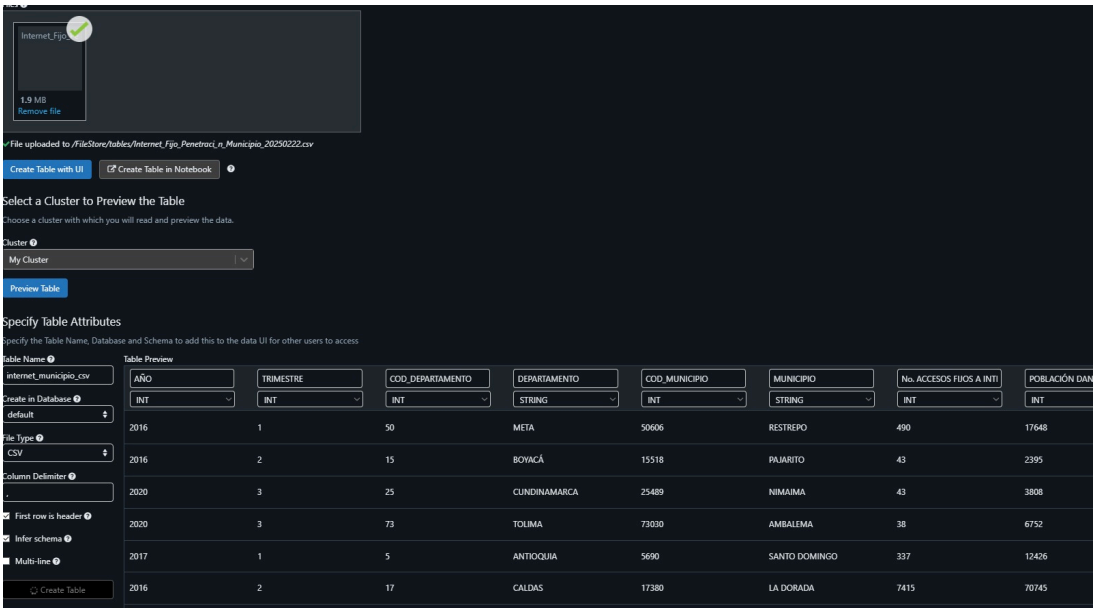


Imagen 2: Creación de tablas a partir de los csv

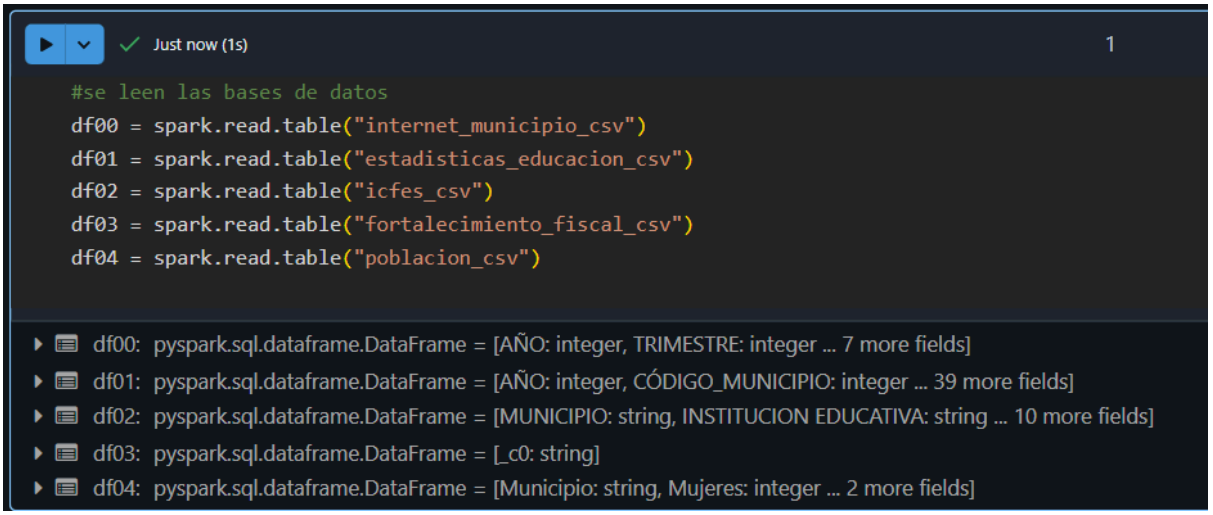


Imagen 3: Carga de tablas en el notebook

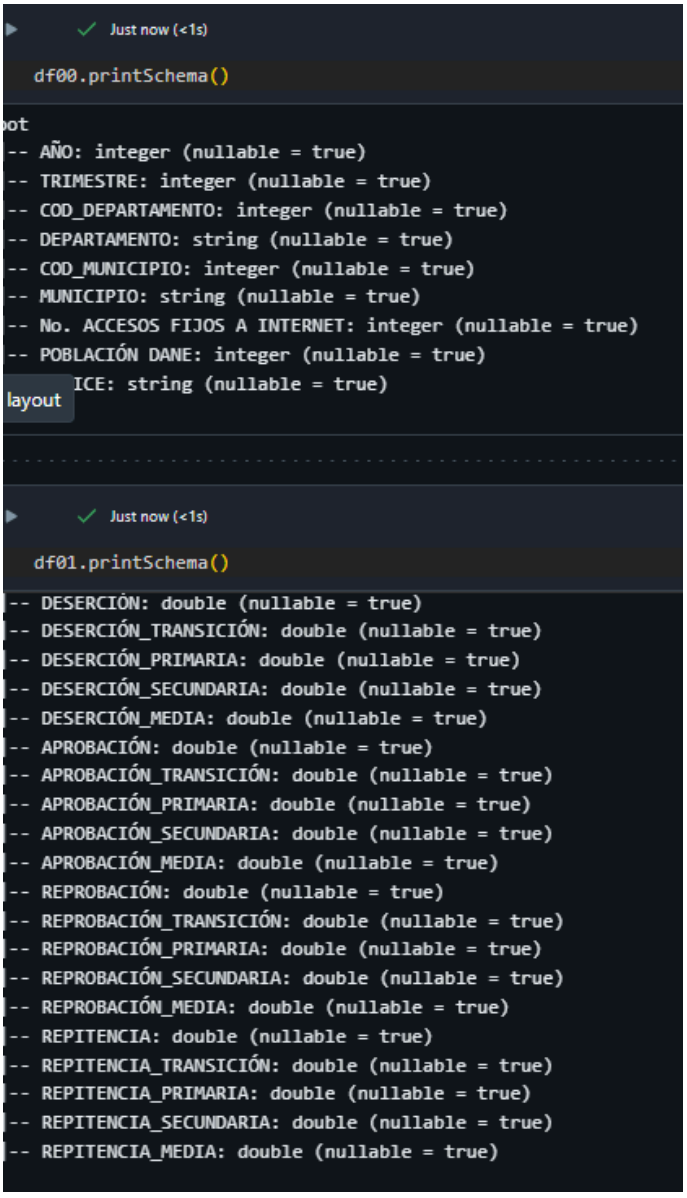


Imagen 4: Visualización de tipo de variables

3.2 Descripción de los conjuntos de datos

Para este análisis usamos las siguientes bases de datos proporcionadas por datos abiertos del gobierno nacional.

- **Internet Fijo Penetración Municipio:** Esta base de datos, nos muestra el número de suscriptores con acceso dedicado a Internet para cada uno de los departamentos y municipios de Colombia, según los datos reportados por los proveedores al último día de cada trimestre.
  Tipos de datos:
  - **AÑO**: es un dato tipo texto que representa el año del cual se investigó el dato. TRIMESTRE, es un dato tipo texto que representa el trimestre del año del cual se investigó el dato. }
  - **COD_DEPARTAMENTO**, es una variable tipo texto, que representa el código del departamento.
  - **DEPARTAMENTO**, es una variable tipo texto, que representa el departamento donde se tomó el dato.
  - **COD_MUNICIPIO**, es una variable tipo texto, que representa el código del municipio de donde se tomó el dato.
  - **MUNICIPIO**, es una variable tipo texto, que representa el código del municipio de donde se tomó el dato. MUNICIPIO, es una variable tipo texto. que representa el municipio de donde se tomó el dato.
  - **No. ACCESOS FIJOS A INTERNET**, es una variable tipo texto. que representa el número de accesos fijos a internet del municipio/departamento.
  - POBLACIÓN DANE, es una variable tipo texto, que representa la población registrada por el Dane del municipio/departamento.
    INDICE, es una variable tipo texto del cual no se tiene información

.

- **Estadísticas en educación en preescolar, básica y media por municipio**:
  Esta base de datos, contiene información estadística de los niveles preescolar, básica y media relacionada con indicadores sectoriales por Municipio sin atípicos, desde el 2011 hasta 2023.
  Tipos de datos: AÑO: es una variable tipo texto que representa la vigencia del indicador.
  - **CÓDIGO_MUNICIPIO:** es una variable tipo texto que indica el código del municipio.
  - **MUNICIPIO:** es una variable tipo texto que señala el nombre del municipio.
  - **CÓDIGO_DEPARTAMENTO:** es una variable tipo texto que representa el código DANE del departamento.

- **DEPARTAMENTO:** es una variable tipo texto que indica el nombre del departamento.
- **CÓDIGO_ETC:** es una variable tipo texto que señala el código DANE de la Entidad Territorial Certificada (ETC).
- **ETC:** es una variable tipo texto que no se tiene información sobre su uso
- **POBLACIÓN_5_16:** es una variable tipo texto que muestra la población en edad teórica de estudiar (5 a 16 años) según proyecciones del DANE.
- **TASA_MATRICULACIÓN_5_16:** es una variable tipo numérica que indica la proporción de la población entre 5 y 16 años que asiste al sistema educativo.
- **COBERTURA_NETA:** es una variable tipo numérica que refleja la relación entre estudiantes matriculados en transición, primaria, secundaria y media con la edad teórica (5 a 16 años) y la población correspondiente.
- **COBERTURA_NETA_SECUNDARIA:** es una variable tipo Número. Relación de estudiantes matriculados en secundaria con la edad teórica (11 a 14 años) y la población correspondiente.
- **COBERTURA_NETA_MEDIA:** es una variable tipo Número. Relación de estudiantes matriculados en media con la edad teórica (15 a 16 años) y la población correspondiente.
- **COBERTURA_BRUTA:** es una variable tipo Número. Relación de estudiantes matriculados en transición, primaria, secundaria y media respecto a la población en edad teórica.
- **COBERTURA_BRUTA_TRANSICIÓN:** es una variable tipo Número. Relación de estudiantes matriculados en transición respecto a la población en edad teórica (5 años).
- **COBERTURA_BRUTA_PRIMARIA:** es una variable tipo Número. Relación de estudiantes matriculados en primaria respecto a la población en edad teórica (6 a 10 años).
- **COBERTURA_BRUTA_SECUNDARIA:** es una variable tipo Número. Relación de estudiantes matriculados en secundaria respecto a la población en edad teórica (11 a 14 años).
- **COBERTURA_BRUTA_MEDIA:** es una variable tipo Número. Relación de estudiantes matriculados en media respecto a la población en edad teórica (15 a 16 años).
- **TAMAÑO_PROMEDIO_DE_GRUPO:** es una variable tipo Número. Número promedio de estudiantes por grupo.
- **SEDES_CONECTADAS_A_INTERNET:** es una variable tipo Número. Porcentaje de sedes oficiales conectadas a internet.
- **DESERCIÓN:** es una variable tipo Número. Tasa de deserción intra-anual del sector oficial.

- **DESERCIÓN_TRANSICIÓN:** es una variable tipo Número. Tasa de deserción en transición.
- **DESERCIÓN_PRIMARIA:** es una variable tipo Número. Tasa de deserción en primaria.
- **DESERCIÓN_SECUNDARIA:** es una variable tipo Número. Tasa de deserción en secundaria.
- **DESERCIÓN_MEDIA**: es una variable tipo Número. Tasa de deserción en media.
- **APROBACIÓN**: es una variable tipo Número. Tasa de aprobación de estudiantes del sector oficial.
- **APROBACIÓN_TRANSICIÓN**: es una variable tipo Número. Tasa de aprobación en transición.
- **APROBACIÓN_PRIMARIA:** es una variable tipo Número. Tasa de aprobación en primaria.
- **APROBACIÓN_SECUNDARIA:** es una variable tipo Número. Tasa de aprobación en secundaria.
- **APROBACIÓN_MEDIA:** es una variable tipo Número. Media de tasa de aprobación.
- REPROBACIÓN: es una variable tipo Número. Tasa de reprobación de estudiantes del sector oficial.
- **REPROBACIÓN_TRANSICIÓN:** es una variable tipo Número. Tasa de reprobación en transición.
  REPROBACIÓN_PRIMARIA: es una variable tipo Número. Tasa de reprobación en primaria.
- **REPROBACIÓN_SECUNDARIA:** es una variable tipo Número. Tasa de reprobación en secundaria.
- **REPROBACIÓN_MEDIA:** es una variable tipo Número. Media de tasa de reprobación.
- **REPITENCIA:** es una variable tipo Número. Tasa de repitencia del sector oficial.
- **REPITENCIA_TRANSICIÓN**: es una variable tipo Número. Tasa de repitencia en transición.
- **REPITENCIA_PRIMARIA**: es una variable tipo Número. Tasa de repitencia en primaria.
- **REPITENCIA_SECUNDARIA**: es una variable tipo Número. Tasa de repitencia en secundaria.
- **REPITENCIA_MEDIA:** es una variable tipo Número. Media de la tasa de repitencia.

- **Pruebas ICFES:** Esta base de datos, contiene información del comparativo clasificación de planteles y las pruebas saber 11o
  Tipos de datos:

- **MUNICIPIO:** Variable tipo texto. Municipio donde fue tomado el ICFES por el participante.
- **INSTITUCIÓN EDUCATIVA:** Variable tipo texto. Colegio donde del cual es el participante que tomó el ICFES.
- **AÑO 2014:** Variable tipo texto. Registro de los ICFES realizados en 2014.
- **AÑO 2015:** Variable tipo texto. Registro de los ICFES realizados en 2015.
- **AÑO 2016:** Variable tipo texto. Registro de los ICFES realizados en 2016.
- **AÑO 2017:** Variable tipo texto. Registro de los ICFES realizados en 2017.
- **AÑO 2018:** Variable tipo texto. Registro de los ICFES realizados en 2018.
- **AÑO 2019:** Variable tipo texto. Registro de los ICFES realizados en 2019.
- **AÑO 2020:** Variable tipo texto. Registro de los ICFES realizados en 2020.
- **AÑO 2021:** Variable tipo texto. Registro de los ICFES realizados en 2021.
- **AÑO 2022:** Variable tipo texto. Registro de los ICFES realizados en 2022.
- **AÑO 2023:** Variable tipo texto. Registro de los ICFES realizados en 2023.

- **Dirección de Descentralización y Fortalecimiento Fiscal:** A continuación en esta base de datos, se presentan los resultados del desempeño fiscal de las entidades territoriales vigencia 2023, de conformidad con lo establecido en la Ley 617 de 2000 y según los criterios de evaluación establecidos por la Dirección de Descentralización y Fortalecimiento Fiscal del Departamento Nacional de Planeación para la Metodología de cálculo del indicador.
  Tipos de datos:
  - **Código:** es una variable tipo texto. Indica el código del municipio según el DANE.
  - **Departamento:** es una variable tipo texto. Representa el nombre del departamento al que pertenece el municipio.
  - **Municipio:** es una variable tipo texto. Muestra el nombre del municipio evaluado.
  - **Categorías:** es una variable tipo texto. Define la categoría administrativa del municipio (e.g., capital, categoría especial, 1ra, 2da...).
  - **Dotaciones Iniciales:** es una variable tipo texto. Refleja las dotaciones o recursos iniciales con los que cuenta el municipio.

- **Ciudad capital:** es una variable tipo texto. Indica si el municipio es una ciudad capital (Sí/No).
- **Dependencia de las Transferencias:** es una variable tipo numérico. Mide el grado de dependencia del municipio respecto a las transferencias del gobierno.
- **Calificación Dependencia de las Transferencias:** es una variable tipo numérico. Asigna una calificación basada en la dependencia de transferencias.
- **Relevancia FBK fijo:** es una variable tipo numérico. Representa la relevancia de la formación bruta de capital fijo en la economía del municipio.
- **Calificación Relevancia FBK fijo:** es una variable tipo numérico. Muestra la calificación otorgada a la relevancia del FBK fijo.
- **Endeudamiento (Total):** es una variable tipo numérico. Refleja el nivel de endeudamiento total del municipio.
- **Calificación Endeudamiento Total:** es una variable tipo numérico. Asigna una calificación al nivel de endeudamiento total del municipio.
- **Ahorro Corriente:** es una variable tipo numérico. Indica la capacidad del municipio para generar ahorro corriente.
- **Calificación Ahorro Corriente:** es una variable tipo numérico. Evalúa el desempeño del municipio respecto al ahorro corriente.
- **Balance Primario:** es una variable tipo numérico. Mide el balance entre ingresos y gastos antes del pago de la deuda.
- **Calificación Balance Primario:** es una variable tipo numérico. Refleja la calificación otorgada al balance primario del municipio.
- **Resultados:** es una variable tipo numérico. Indica los resultados financieros o de gestión obtenidos por el municipio.
- **Calificación Resultados:** es una variable tipo numérico. Asigna una calificación a los resultados obtenidos por el municipio.
- **Holgura:** es una variable tipo numérico. Representa la capacidad del municipio para manejar sus finanzas sin comprometer su estabilidad.
- **Calificación Holgura:** es una variable tipo numérico. Muestra la calificación otorgada a la holgura financiera del municipio.
- **Capacidad de programación y recaudo de ingresos:** es una variable tipo numérico. Indica la eficiencia del municipio en programar y recaudar ingresos propios.
- **Calificación capacidad de programación y recaudo de Ingresos:** es una variable tipo numérico. Evalúa la eficiencia en la programación y el recaudo de ingresos.
- **Capacidad de Ejecución de Inversión:** es una variable tipo numérico. Mide la capacidad del municipio para ejecutar proyectos de inversión.
- **Calificación Capacidad de Ejecución de Inversión:** es una variable tipo numérico. Refleja la calificación otorgada a la capacidad de ejecución de inversiones.

- **Bonificación Esfuerzo Propio:** es una variable tipo numérico. Muestra el incentivo recibido por el municipio por su esfuerzo propio en la generación de ingresos.
- **Bono Catastro:** es una variable tipo numérico. Indica un bono otorgado al municipio relacionado con la gestión catastral.
- **Resultados Gestión:** es una variable tipo numérico. Evalúa los resultados obtenidos por la gestión municipal.
- **Gestión +Bonos:** es una variable tipo numérico. Representa la evaluación de la gestión del municipio considerando los bonos recibidos.
- **Calificación Resultados Gestión:** es una variable tipo numérico. Asigna una calificación a los resultados de la gestión del municipio.
- **Nuevo IDF (sin bonos):** es una variable tipo numérico. Muestra el Índice de Desempeño Fiscal sin considerar los bonos.
- **Nuevo IDF:** es una variable tipo numérico. Indica el Índice de Desempeño Fiscal del municipio, incluyendo los bonos.
- **Rango:** es una variable tipo texto. Clasifica al municipio dentro de un rango o categoría según su desempeño.

- **Proyección de la población de cundinamarca 2023**: Este dataset nos indica una proyección estimada de cundinamarca para 2023
  Tipos de datos:
  - **Municipio**: es una variable tipo texto que indica el nombre del municipio de Cundinamarca.
  - **Mujeres**: es una variable tipo numérica que representa la población femenina del municipio.
  - **Hombres**: es una variable tipo numérico que representa la población masculina del municipio.
  - **Total**: es una variable tipo numérica que indica la población total del municipio, sumando hombres y mujeres.

4. Exploración de los datos

Para esta sección el objetivo es entender cómo están estructurados los datos, cómo se comportan y cuáles pueden llegar a ser los datos relevantes para el estudio.

Los datos fueron cargados y se realizaron las siguientes verificaciones:

- **Visualización de las hojas disponibles en los archivos Excel** para seleccionar las relevantes.
- **Extracción de las primeras filas de cada conjunto de datos** con ***head()*** para inspeccionar su estructura y formato.

- **Diccionario de datos** donde se especifican los nombres de los data frames en donde se guardaron los datos.
- **Identificación de tipos de datos** con ***info()*** para verificar si las variables tienen el formato adecuado.
- **Resumen estadístico** con ***describe()*** para obtener medidas de tendencia central, dispersión y distribución de los datos.

Lo siguiente que se realizó fueron un total de 8 formas diferentes de análisis a los datasets, donde se usaron estadística descriptiva, gráficos de barras, regresión, entre otros.

1. Estadísticos Descriptivos

Se calcularon estadísticas descriptivas para cada conjunto de datos, como el promedio, la desviación estándar, el valor mínimo y máximo. Esto permite conocer la distribución de los valores y detectar posibles anomalías o tendencias.

2. Dimensiones de los DataFrames

Se analizó el número de filas y columnas en cada conjunto de datos para entender su estructura y la cantidad de información disponible.

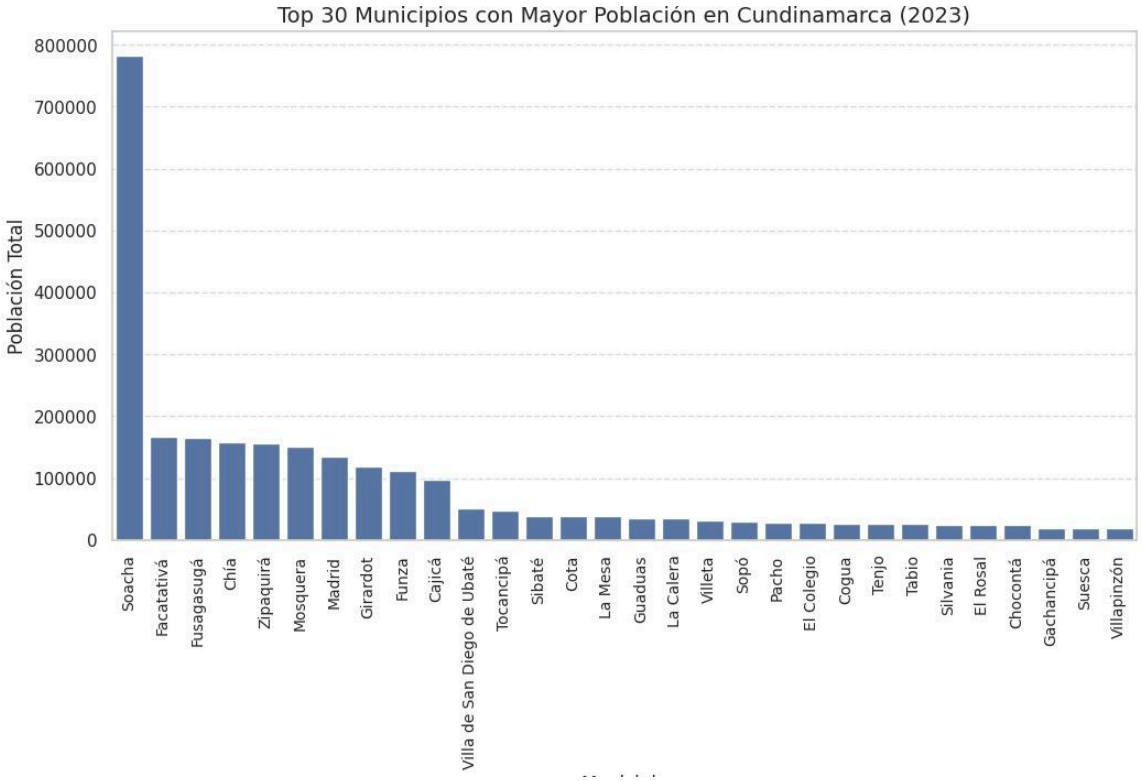
3. Tipos de Datos en los DataFrames

Se examinó el tipo de datos en cada columna para asegurar la coherencia y verificar que se puedan realizar operaciones correctas sobre ellos, como conversiones numéricas o categóricas.

4. Gráfica de Distribución de Población

Se seleccionaron los 30 municipios con mayor población en Cundinamarca, ya que son nuestros municipios de estudio. Se realizó un gráfico de barras para visualizar la distribución de la población en los municipios más grandes.

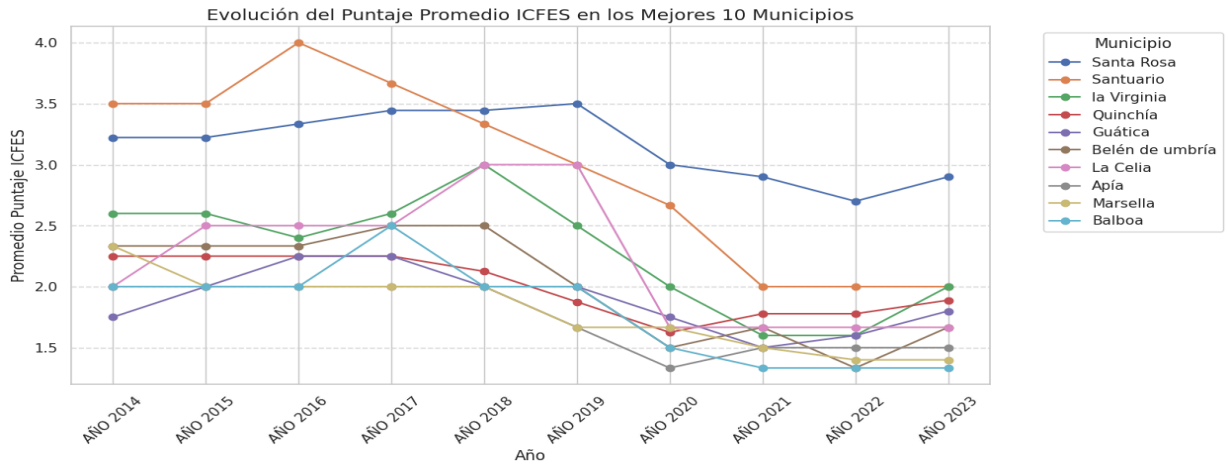
Figura 1: Top 30 Municipios con Mayor Población en Cundinamarca (2023)



5. Comparación del Puntaje Promedio ICFES

Se analizaron los puntajes promedio del ICFES en distintos municipios a lo largo de los años. Se hizo una conversión de las categorías de puntaje a valores numéricos y se graficó la evolución del puntaje en los 10 municipios desde el año 2014 hasta el año 2023.

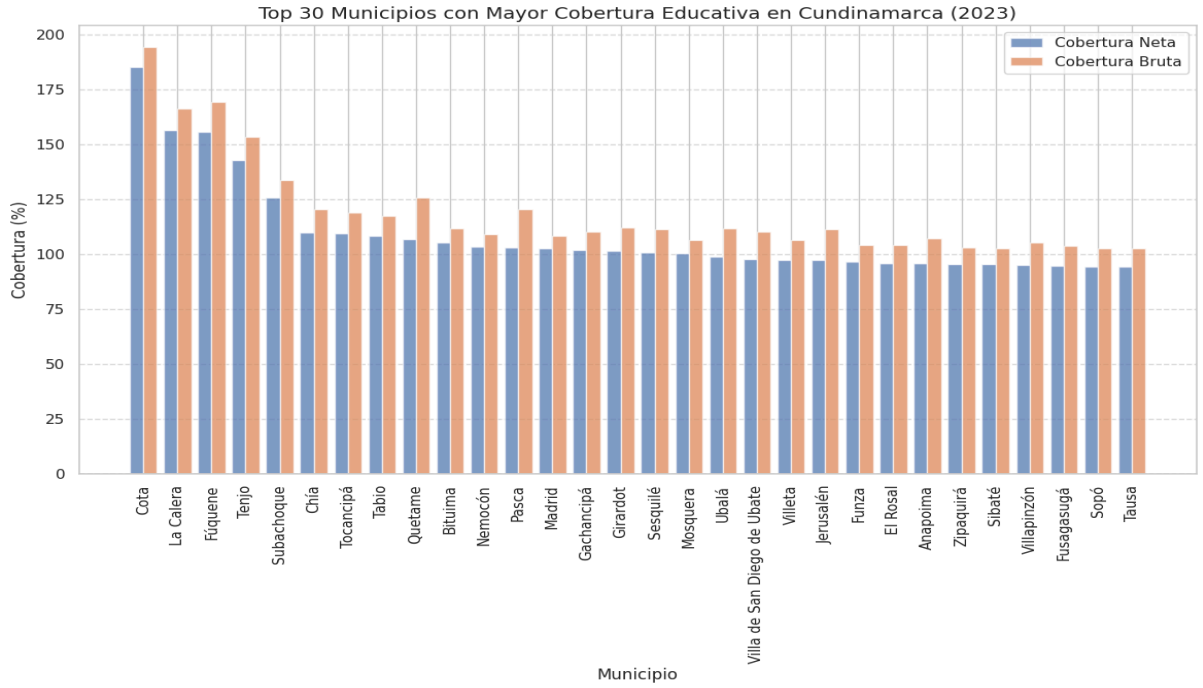
Figura 2: Evolución del Puntaje Promedio ICFES en los Mejores 10 Municipios



6. Comparación de Municipios con Mayor Cobertura Educativa

Se identificaron los 30 municipios con mayor cobertura educativa (neta y bruta) en el último año disponible. Se realizó un gráfico comparativo para visualizar las diferencias en acceso a la educación entre los municipios.

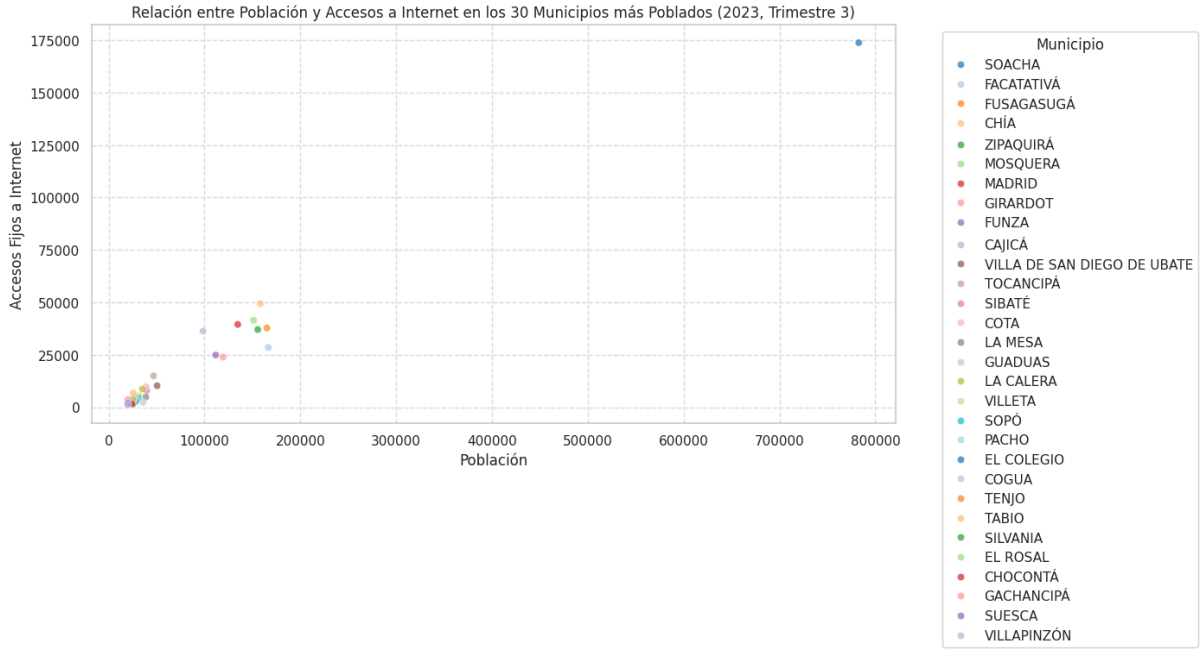
Figura 3: Top 30 Municipios con Mayor Cobertura Educativa en Cundinamarca (2023)



7. Relación entre Población y Accesos a Internet

Se analizó la relación entre la cantidad de habitantes y el número de accesos fijos a internet en los 30 municipios más poblados. Se utilizó un diagrama de dispersión y se calculó el coeficiente de correlación para determinar el grado de asociación entre ambas variables. Dicho coeficiente nos arrojó como resultado 0.99, indicando una fuerte relación que entre mayor cantidad de personas en un municipio, hay mayor puntos de acceso de internet.

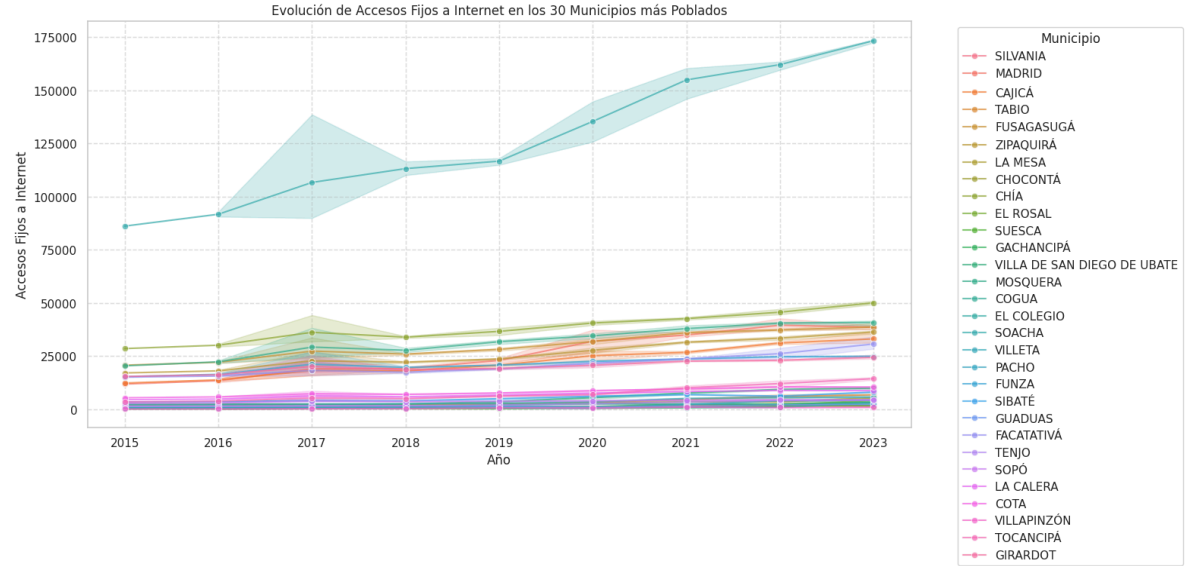
Figura 4: Relación entre Población y Accesos a Internet en los 30 Municipios más Poblados (2023, Trimestre 3)



8. Evolución de los Accesos de Internet en los Municipios

Se visualizó la evolución de los accesos fijos a internet en los 30 municipios más poblados a lo largo de los años. Esto permite identificar tendencias de crecimiento o estancamiento en la conectividad de la región.

Figura 5: Evolución de Accesos Fijos a Internet en los 30 Municipios más Poblados



5. Reporte de calidad de los datos:

Para esta sección se hizo el conteo de los datos faltantes, nulos o vacíos con el fin de poder plantear soluciones a realizar para poder filtrar, limpiar y transformar los datos.

Como resultado obtuvimos que:

Bases de Datos Completas (Sin Datos Faltantes):

- ***IDF Municipios y IDF Departamentos***: No presentan valores faltantes, lo que indica que los datos financieros y de gestión fiscal de municipios y departamentos están completos.
- ***Población***: Todos los municipios tienen datos de población total, desglosada por género.
- ***Internet****:* No hay datos faltantes en el acceso a internet en municipios, lo que facilita su uso para evaluar la conectividad.

Bases de Datos con Valores Faltantes:

- ***Educación (3.13% de datos faltantes):***
    - Faltan datos en variables clave como **Tasa de Matrícula (115 valores faltantes), Cobertura Neta (111), Cobertura Bruta (68), Tamaño Promedio de Grupo (7013), y Sedes Conectadas a Internet (6817).**
    - La falta de datos en estas variables puede dificultar el análisis sobre acceso y calidad educativa en algunos municipios.
    - Es relevante notar que la mayor cantidad de valores faltantes está en **Tamaño Promedio de Grupo y Conectividad de Sedes Escolares**, lo que sugiere que estos datos no se recopilan sistemáticamente en todas las regiones.
- ***ICFES (0.40% de datos faltantes):***
    - Existen **valores faltantes en los años 2014, 2015 y 2016**, pero la información de 2017 en adelante está completa.
    - Es posible que los registros de estos años anteriores no fueran capturados o que haya inconsistencias en la fuente de datos.

Después de analizar estos resultados, planteamos posibles y diferentes soluciones:

- Eliminación de filas o columnas con demasiados valores nulos
- Imputación con la media o mediana para datos numéricos
- Uso de modelos predictivos para estimar valores perdidos
- Completar valores con base en información externa o reglas de negocio.

6. Preguntas sobre el negocio

- ¿Existe una relación significativa entre la penetración de internet fijo en los municipios y los puntajes promedio del ICFES?
- ¿Cómo afecta el Índice de Desempeño Fiscal (IDF) de los municipios a los resultados del ICFES 11?

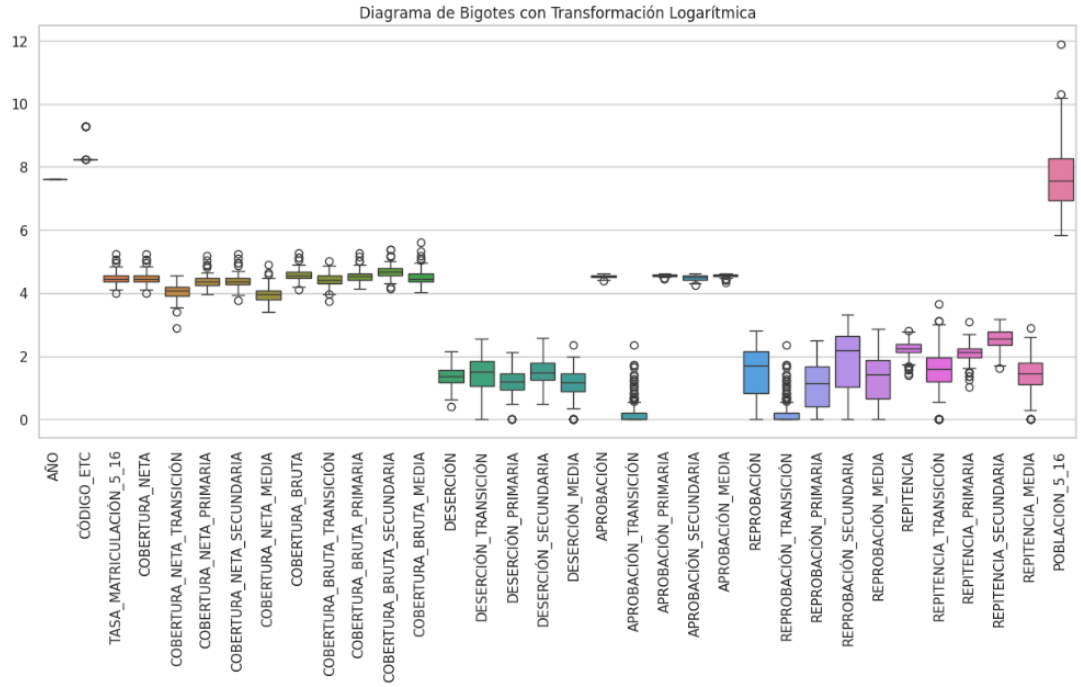
- ¿Cuál es la tendencia en la evolución de los accesos a internet fijo en los municipios más poblados de Cundinamarca en los últimos años?
- ¿Cuál es el impacto de la tasa de matrícula y la cobertura educativa en los resultados de la prueba ICFES 11?
- ¿Existe correlación entre la tasa de deserción escolar en los municipios y los puntajes promedio de la prueba ICFES?
- ¿Cuáles municipios tienen el mejor y peor desempeño en la prueba ICFES 11 y qué factores pueden influir en estas diferencias?
- ¿Cómo ha variado la relación entre el acceso a internet y el desempeño académico en los últimos años en Cundinamarca?
- ¿En qué medida los municipios con mayor inversión en educación tienen mejores resultados en la prueba ICFES?

Filtros, limpieza y transformación inicial

DataFrame: Educaion

1. **Limpieza de Datos:**
    - Se analizaron las columnas con valores nulos y se eliminaron aquellas con una cantidad significativa de datos faltantes.
    - Se revisó la consistencia de los valores numéricos en cada columna.
2. **Análisis de Valores Atípicos:**
    - Se generaron diagramas de bigotes para identificar valores extremos en diferentes variables.
    - Se aplicó una escala logarítmica para mejorar la visualización de datos con alta dispersión.
    - Se realizó un análisis específico de la variable APROBACIÓN_TRANSICIÓN, encontrando múltiples valores atípicos y una

distribución inusual.



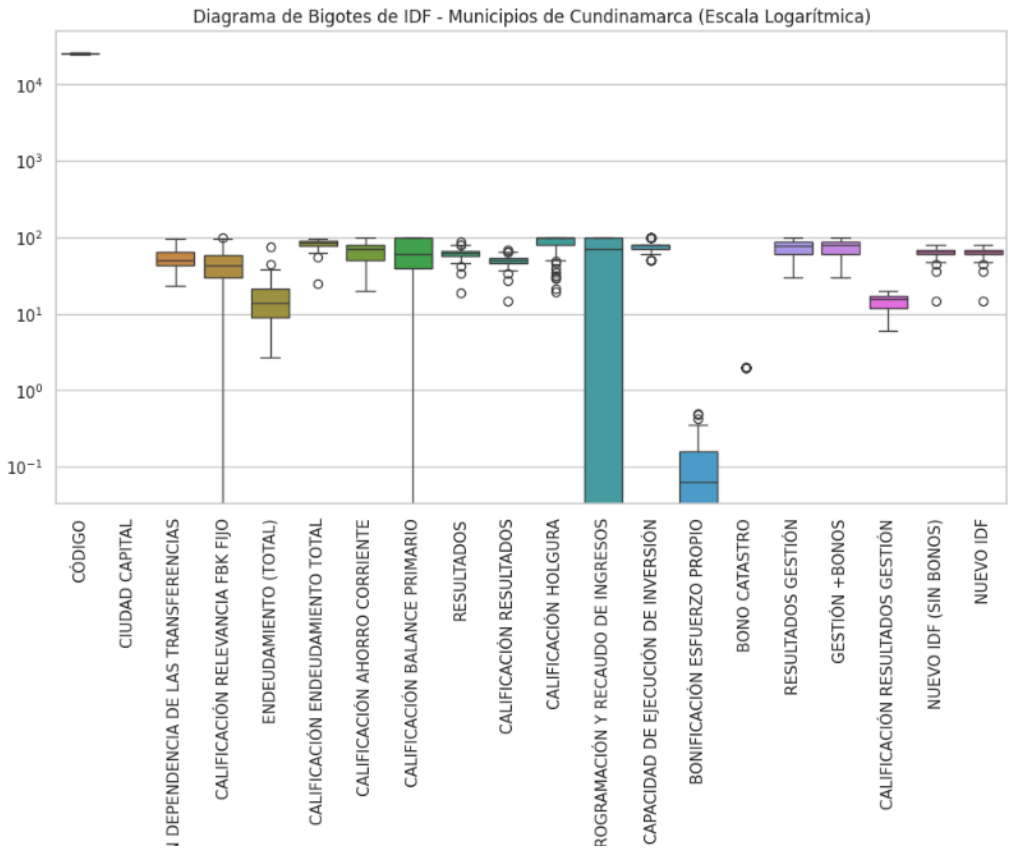
3. **Transformación y Evaluación de la Variable APROBACIÓN_TRANSICIÓN:**
    - Se aplicó una transformación logarítmica para evaluar su distribución.
    - Se identificó que esta variable contenía datos inconsistentes en casi todas las filas.
    - Debido a la falta de confiabilidad de esta variable, se optó por eliminarla del conjunto de datos para evitar sesgos en futuros análisis.

DataFrame: IDF Municipios

1. **Filtrado de Datos:** Se realizó la selección de los municipios pertenecientes al departamento de Cundinamarca dentro del dataframe, asegurando que el análisis se centrará exclusivamente en esta región.

2. **Verificación de Valores Nulos:** Se evaluó la existencia de valores nulos en las columnas del dataframe. No se identificaron valores faltantes, lo que indica que el conjunto de datos está completo y no requiere limpieza en este aspecto.

3. **Detección de Valores Atípicos:** Se generó un diagrama de bigotes (boxplot) con escala logarítmica para detectar posibles valores atípicos en las variables del IDF. La inspección visual del gráfico no mostró problemas significativos con los datos, lo que sugiere que la distribución es adecuada y no se requiere intervención adicional.



DataFrame: IDF Departamento:

1. **Filtrado de Datos:** Se realizó la selección del departamento de Cundinamarca dentro del dataframe, asegurando que el análisis se centrará exclusivamente en esta región.

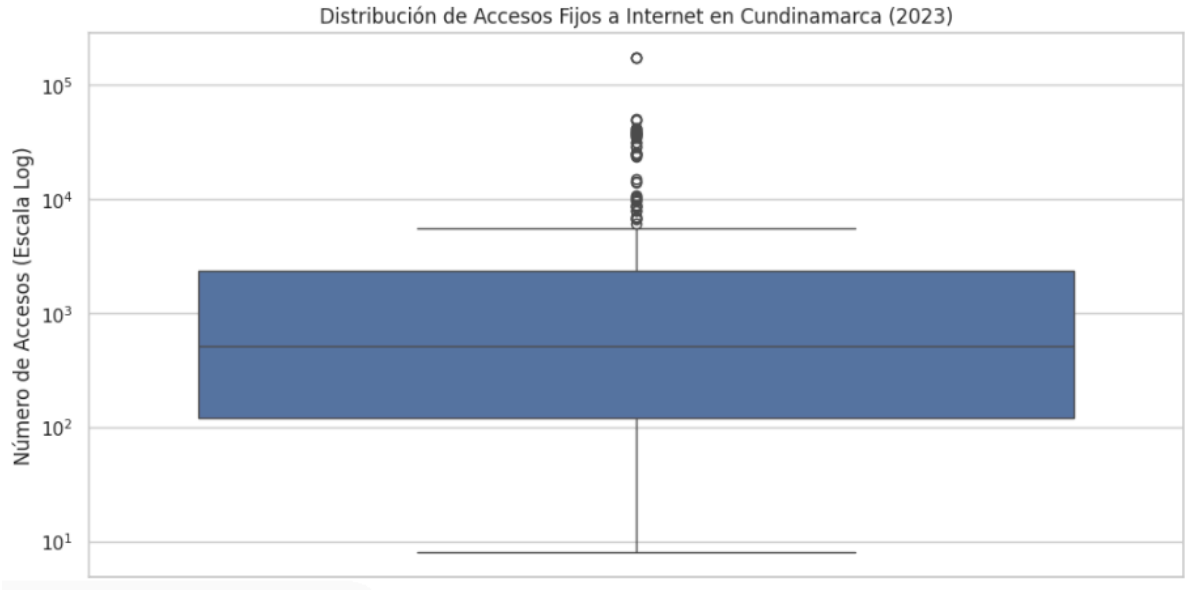
2. **Verificación de Valores Nulos:** Se evaluó la existencia de valores nulos en las columnas del dataframe. No se identificaron valores faltantes, lo que indica que el conjunto de datos está completo y no requiere limpieza en este aspecto.

DataFrame: IDF Poblacion:

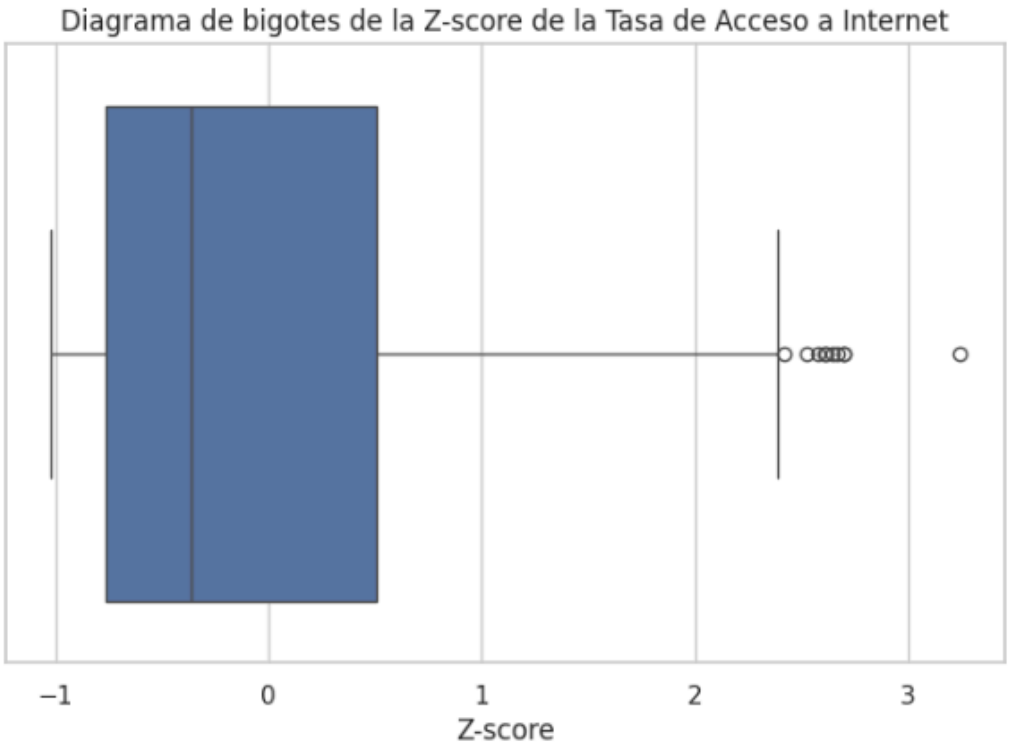
1. **Verificación de Valores Nulos:** Se evaluó la existencia de valores nulos en las columnas del dataframe. No se identificaron valores faltantes, lo que indica que el conjunto de datos está completo y no requiere limpieza en este aspecto.
2. **Detección de Valores Atípicos:** Se generó un diagrama de bigotes (boxplot) con escala logarítmica para detectar posibles valores atípicos en las variables del dataframe poblacion. La inspección visual del gráfico no mostró problemas significativos con los datos, lo que sugiere que la distribución es adecuada y no se requiere intervención adicional.

DataFrame: IDF Internet:

1. **Filtrado de Datos:** Inicialmente, se filtraron los datos del Data Frame internet para considerar exclusivamente la información correspondiente al año 2023 y al departamento de Cundinamarca. Este proceso permitió reducir el conjunto de datos y centrarnos en el análisis de la región de interés.

2. **Identificación de Valores Atípicos:** Se realizó un diagrama de bigotes (boxplot) para examinar la distribución del número de accesos fijos a internet en los municipios de Cundinamarca. A partir del gráfico, se observó que varios municipios presentaban valores significativamente por encima de la media. Esto se debía a diferencias en la población y la cobertura del servicio, lo que sugería la necesidad de una estandarización para una mejor comparación entre municipios.
3. 

4.
5. **Estandarización de la Tasa de Acceso a Internet:** Para hacer una comparación más justa entre municipios con diferentes tamaños de población, se aplicó la normalización Z-score para identificar municipios con tasas de acceso significativamente diferentes de la media.



6.

Referencias

[1] https://telencuestas.com/censos-de-poblacion/colombia/2023/cundinamarca

[2]https://colaboracion.dnp.gov.co/CDT/Desarrollo%20Territorial/Resultados_Piloto_Nuevo_IDF_presentacion%20_lanzamiento_final.pdf

[3]https://www.dnp.gov.co/Prensa_/Noticias/Paginas/dnp-publica-indice-de-desempeno-fiscal-idf-de-departamentos-y-municipios.aspx?utm_source=chatgpt.com

[4]https://www.datos.gov.co/Ciencia-Tecnolog-a-e-Innovaci-n/Internet-Fijo-Penetraci-n-Municipio/fut2-keu8/about_data