

Práctica de Laboratorio: Análisis Exploratorio de Datos - Data Wrangling

Docente: Ana María Cuadros Valdivia

Alumna: Luciana Julissa Huaman Coaquira

Contexto:

Este estudio se centra en la utilización de datos de movilidad humana para comprender y monitorear la propagación de enfermedades infecciosas en Brasil. La movilidad de las personas es un factor crítico que influye en la velocidad y el alcance de los brotes epidémicos, y por ello, analizar patrones de desplazamiento es fundamental para la salud pública.

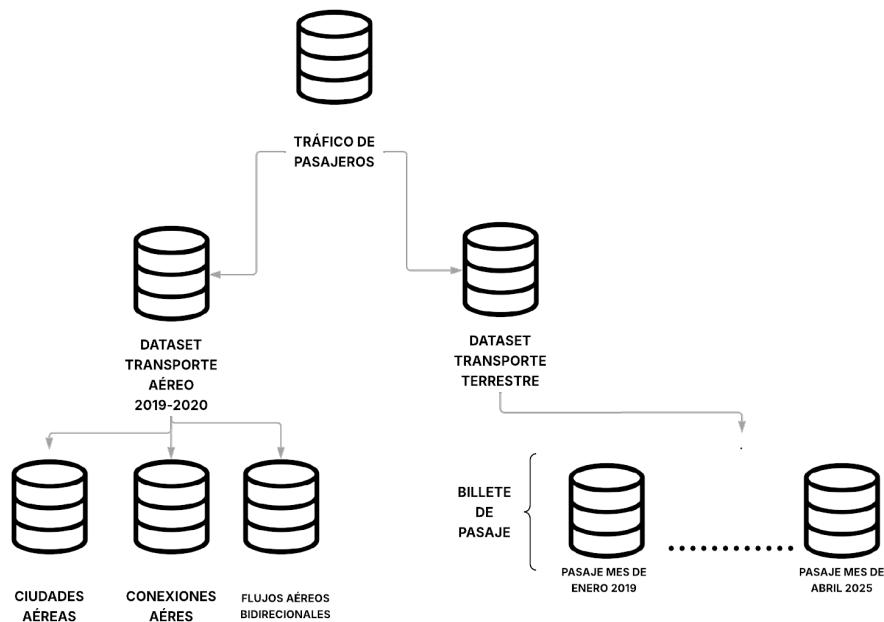
Los datos analizados provienen de fuentes oficiales que registran el transporte aéreo y terrestre, incluyendo conexiones entre ciudades, flujos de pasajeros y ventas de billetes. Estos datos, combinados con bases de datos demográficas y sanitarias, permiten construir modelos detallados de movilidad que ayudan a identificar rutas y nodos clave en la propagación de infecciones.

El análisis del comportamiento de estos datos revela la complejidad y heterogeneidad de los movimientos interurbanos, destacando la importancia de integrar distintas modalidades de transporte para obtener una visión completa. Este enfoque fue utilizado en la herramienta Epiflow[1].

Análisis del Comportamiento de Datos

El conjunto de datos está organizado en dos grandes grupos principales: transporte aéreo y transporte terrestre.

Dentro del transporte aéreo, se incluyen tres datasets: ciudades, conexiones y flujos origen-destino, que contienen información sobre la movilización de pasajeros y cargas entre las diferentes ciudades. Por otro lado, el transporte terrestre se compone de múltiples archivos con registros detallados de billetes de pasaje por usuario desde el 2019 hasta el 2025. Estos datasets se integran y procesan de forma estructurada para facilitar el análisis conjunto de la movilidad y sus impactos.



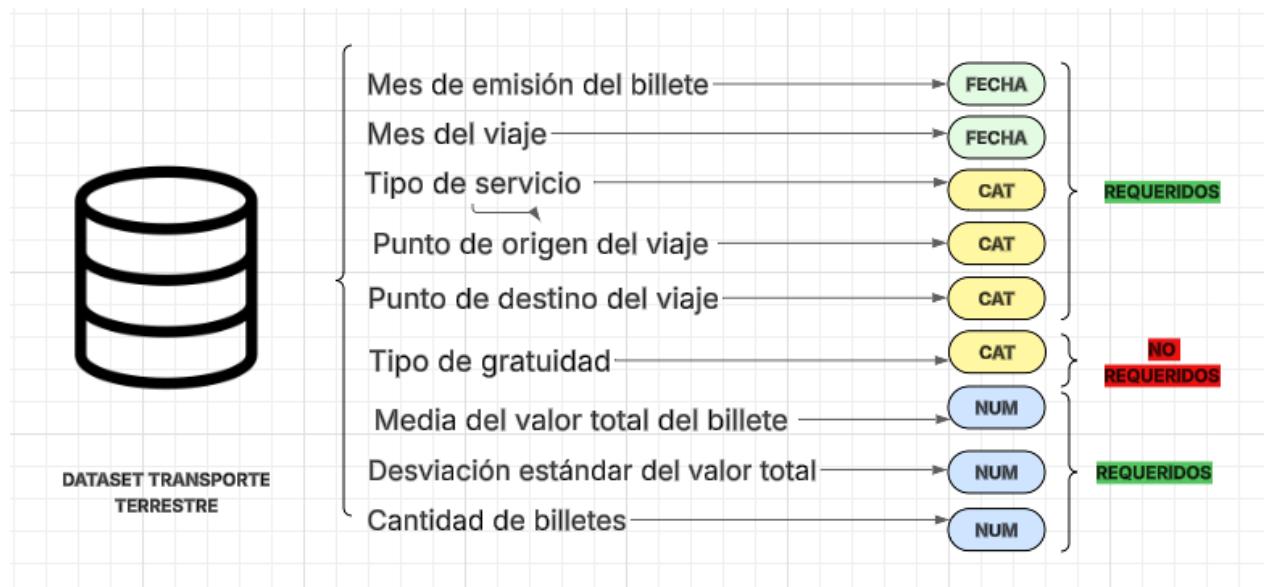
1.1 Descripción de los Datos

DATOS TERRESTRES

El dataset utilizado contiene registros detallados de billetes emitidos para el transporte terrestre interestadual e internacional de pasajeros en Brasil, correspondientes al período 2019-2025. Este conjunto incluye múltiples archivos CSV organizados por meses y años, donde cada archivo representa las ventas de billetes en distintas rutas y fechas.

Las variables principales de estos archivos incluyen:

- **mes_emissao_bilhete:** Mes y año en que se emitió el billete.
- **mes_viajem:** Mes y año en que se realiza el viaje.
- **ponto_origem_viajem:** Ciudad o terminal de origen del viaje.
- **ponto_destino_viajem:** Ciudad o terminal destino del viaje.
- **tipo_servico:** Modalidad del servicio de transporte (convencional, ejecutivo, leito, etc.).
- **tipo_gratuidade:** Categoría del descuento o gratuidad aplicada al billete (promocional, idoso, jovem, etc.).
- **media_valor_total:** Valor promedio pagado por el billete.
- **dp_valor_total:** Desviación estándar del valor del billete.
- **quantidade_bilhetes:** Cantidad de billetes vendidos en el registro.



tipo_gratuidade es solo útil si tu análisis incluye subsidios o gratuidades sociales; si no, no aporta al análisis general de viajes y volúmenes.

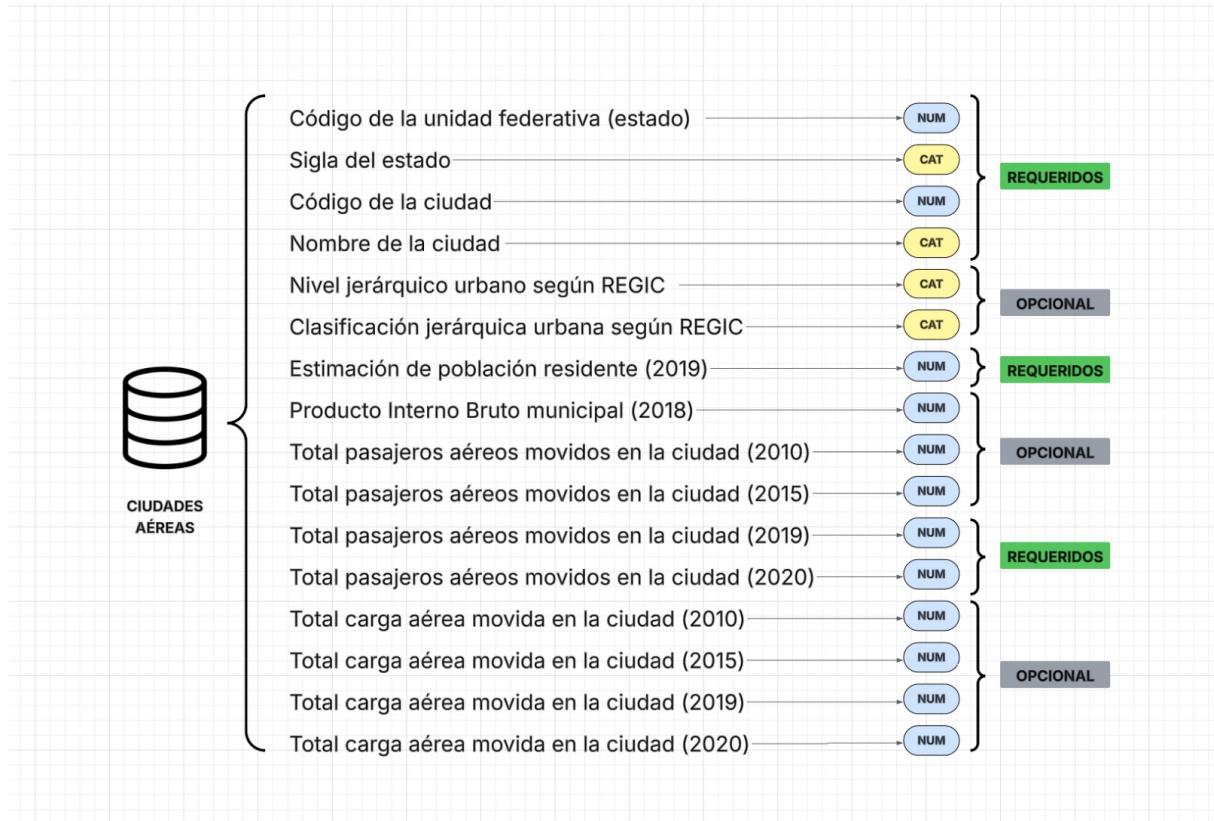
El resto de variables son clave para entender cuándo (temporal), dónde (origen/destino), qué tipo de viaje (servicio) y cuánto (cantidad y valor) viaja la gente.

Datos Aéreos

Los datos aéreos provienen de diferentes bases relacionadas con la movilidad aérea entre ciudades. Estas incluyen tres conjuntos de datos clave que reflejan la conexión entre ciudades, el movimiento de pasajeros, y los flujos de tráfico aéreo.

1. Ciudades Aéreas (LIG_AEREAS_2019-2020_cidades)

Este dataset contiene información sobre las ciudades en Brasil, incluyendo su jerarquía urbana, estimación de población, PIB y volumen de pasajeros y carga aérea movidos a lo largo de varios años.



Requeridas:

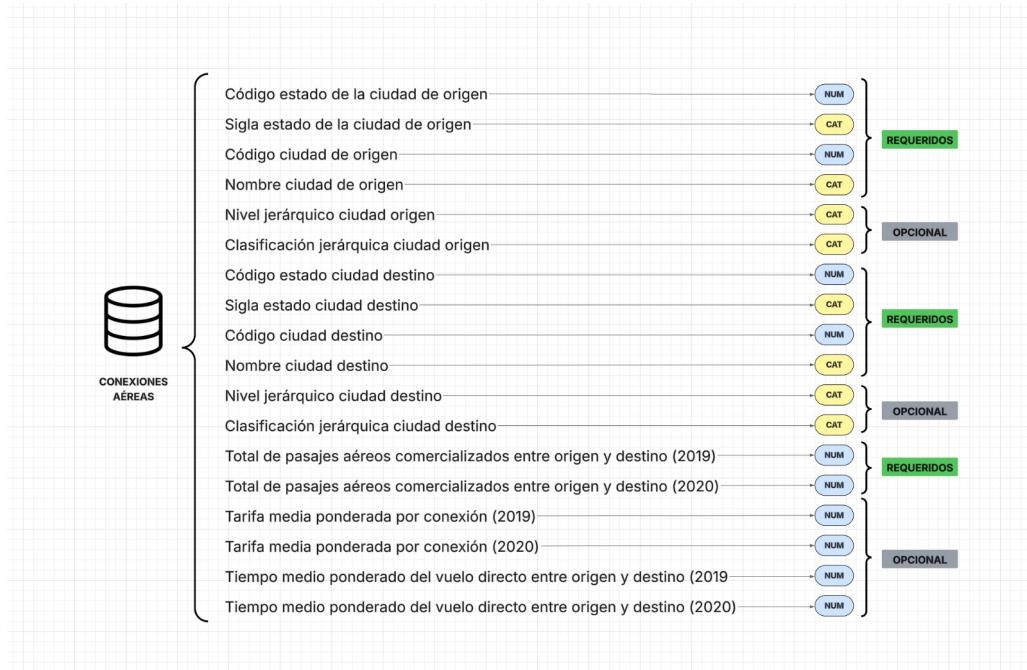
- **COD_UF** (Código de la unidad federativa): Es necesario para identificar el estado al que pertenece cada ciudad. Es crucial para cualquier análisis geográfico o de conectividad.
- **UF** (Sigla del estado): Similar al código de unidad federativa, pero más directo para asociar las ciudades con los estados en una representación más fácil de leer.
- **COD_CIDADE** (Código de la ciudad): Identificador único para cada ciudad, fundamental para diferenciar entre las ciudades en los análisis.
- **NOME_CIDADE** (Nombre de la ciudad): Permite referirse a las ciudades de forma legible. Es útil para visualizaciones o análisis descriptivos.

- **VAR01** (Estimación de población residente en 2019): La población de una ciudad es un factor clave para entender la demanda potencial de transporte aéreo.
- **VAR05** (Total de pasajeros aéreos movidos en la ciudad en 2019): Es la variable más importante para conocer la magnitud del tráfico aéreo de una ciudad en un periodo reciente.
- **VAR06** (Total de pasajeros aéreos movidos en la ciudad en 2020): Similar a la anterior, proporciona información actualizada y comparativa para medir tendencias recientes.

Opcionales:

- **NIVEL_CID** (Nivel jerárquico urbano según REGIC) y **CLASS_CID** (Clasificación jerárquica urbana según REGIC): Aunque son útiles para clasificar y entender la importancia de las ciudades en términos de conectividad, no son estrictamente necesarias para realizar análisis de tráfico aéreo básico.
- **VAR02** (PIB municipal 2018), **VAR03** (Pasajeros movidos en 2010), **VAR04** (Pasajeros movidos en 2015), **VAR07** (Carga aérea movida en 2010), **VAR08** (Carga aérea movida en 2015), **VAR09** (Carga aérea movida en 2019), **VAR10** (Carga aérea movida en 2020): Aunque estos datos son útiles para análisis longitudinales y comparativos, no son necesarios si el enfoque está únicamente en el tráfico de pasajeros más reciente (2019 y 2020).
-

2. Conexiones Aéreas (LIG_AEREAS_2019-2020_ligacoes)



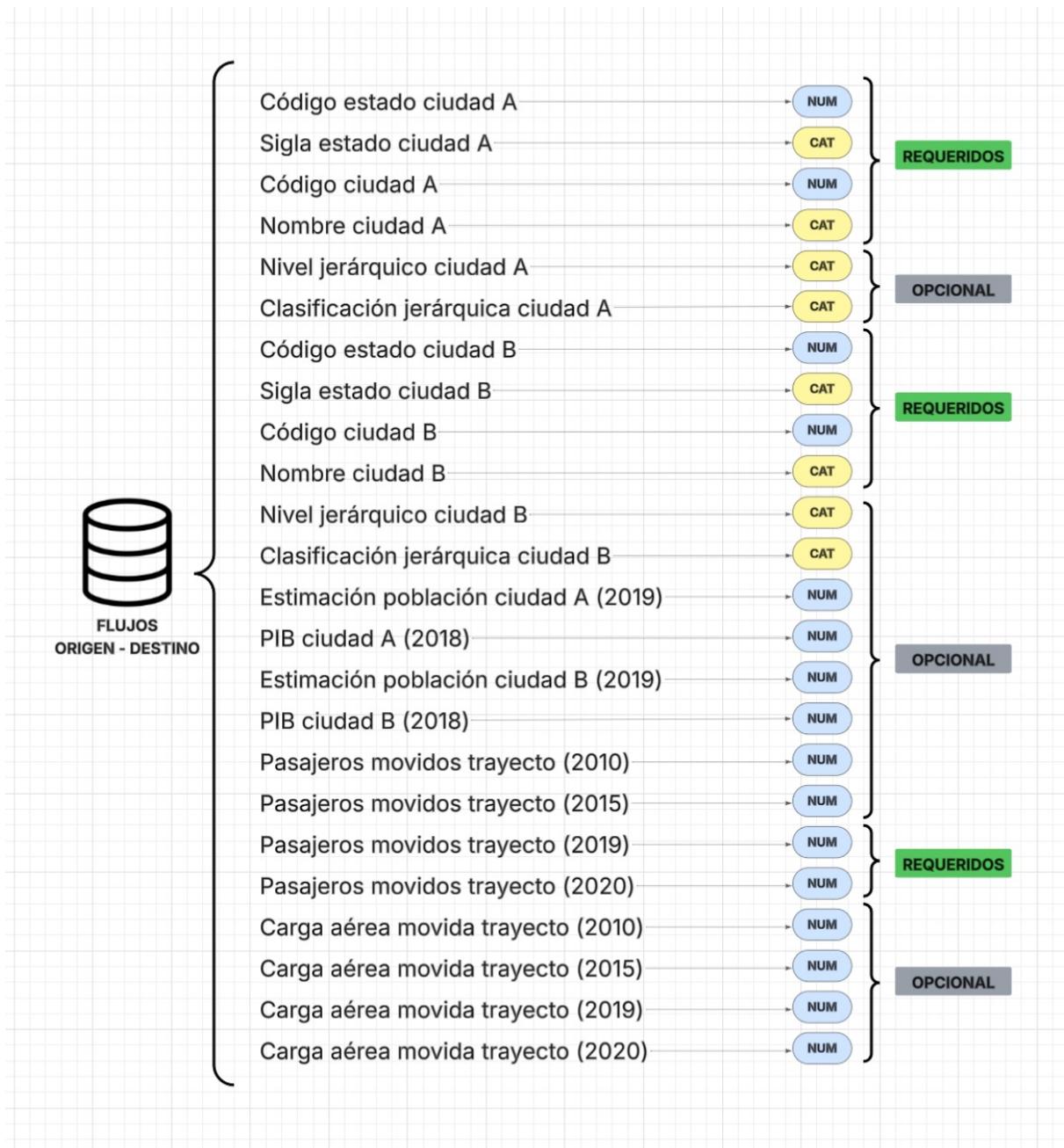
Requeridas:

- **COD_UF_O** y **UF_O** (Código y sigla del estado de la ciudad de origen): Necesarios para identificar el origen del vuelo y asociarlo con un estado específico.
- **COD_CID_O** y **NOME_CID_O** (Código y nombre de la ciudad de origen): Los códigos de las ciudades de origen son claves para diferenciar entre vuelos y destinos.
- **COD_UF_D** y **UF_D** (Código y sigla del estado de la ciudad de destino): Iguales a las variables de origen, pero para el destino del vuelo.
- **COD_CID_D** y **NOME_CID_D** (Código y nombre de la ciudad de destino): Permiten identificar la ciudad de destino, lo cual es necesario para analizar la conectividad entre ciudades.
- **VAR01** y **VAR02** (Total de pasajes comercializados en 2019 y 2020): Los datos de pasajes vendidos son esenciales para medir la demanda de vuelos entre ciudades.

OPcionales:

- **NIVEL_O**, **CLASS_O** (Nivel y clasificación de la ciudad de origen) y **NIVEL_D**, **CLASS_D** (Nivel y clasificación de la ciudad de destino): Aunque pueden proporcionar información valiosa sobre la jerarquía de las ciudades en la red aérea, no son estrictamente necesarios para la análisis de la cantidad de pasajes comercializados o la conectividad básica entre ciudades.
- **VAR03**, **VAR04** (Tarifa media ponderada por conexión en 2019 y 2020) y **VAR05**, **VAR06** (Tiempo medio ponderado del vuelo directo entre origen y destino en 2019 y 2020): Estas variables son útiles si el análisis busca comparar tarifas o tiempos de vuelo, pero no son esenciales para estudiar la conectividad y el volumen de pasajes

3. Flujos Origen-Destino(LIG_AEREAS_2019-2020_fluxos_od)



Requeridas:

- **COD_UF_A, UF_A, COD_CID_A, NOME_CID_A** (Código y nombre del estado y ciudad A): Son necesarios para identificar el origen de los flujos de pasajeros y carga.
- **COD_UF_B, UF_B, COD_CID_B, NOME_CID_B** (Código y nombre del estado y ciudad B): Iguales a las variables de origen, pero para la ciudad destino. Son claves para cualquier análisis de flujos entre ciudades.
- **VAR07** (Pasajeros movidos entre las ciudades en 2019) y **VAR09** (Pasajeros movidos entre las ciudades en 2020): Estas variables permiten medir la cantidad de pasajeros que viajan entre las ciudades, lo cual es esencial para el análisis de movilidad.

Opcionales:

- **NIVE_A, CLASS_A** (Nivel y clasificación de la ciudad A) y **NIVEL_B, CLASS_B** (Nivel y clasificación de la ciudad B): Estos datos son útiles para el análisis de la jerarquía y la importancia relativa de las ciudades, pero no son esenciales para el análisis de flujos.
- **VAR01, VAR02, VAR03, VAR04** (Estimaciones de población y PIB de las ciudades A y B) y **VAR05, VAR06, VAR10, VAR11, VAR12, VAR13** (Datos de pasajeros y carga aérea movidos en los años anteriores): Proporcionan contexto adicional útil en análisis más profundos, pero no son necesarios si se busca un análisis centrado en los flujos más recientes de pasajeros y carga.

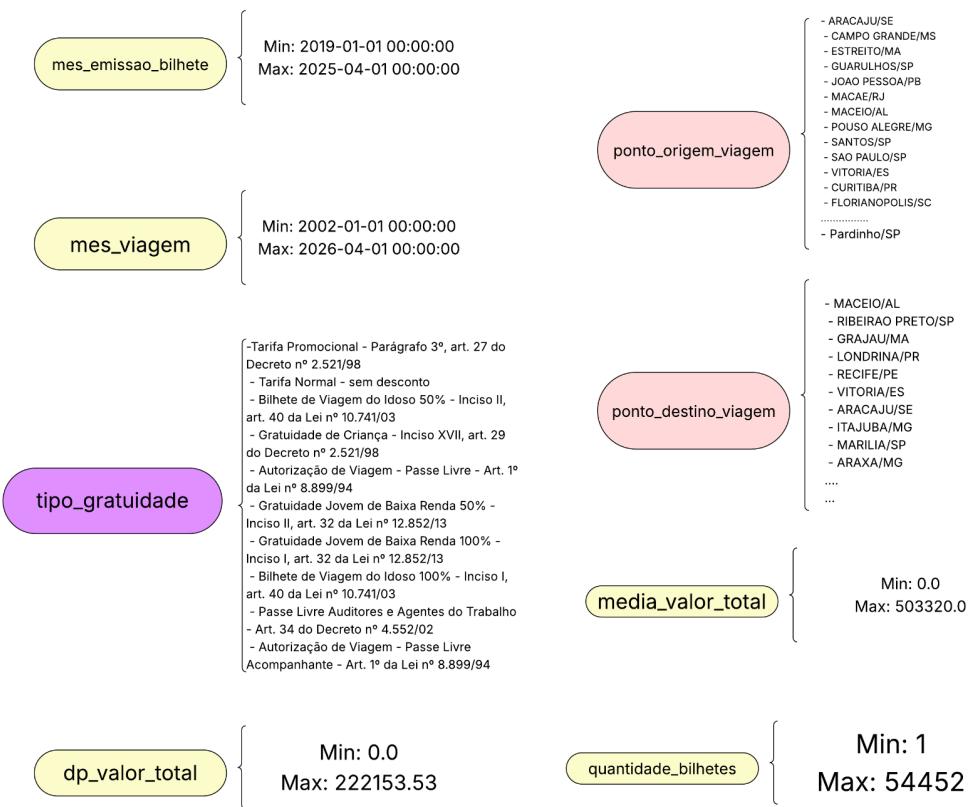
Al analizar los datos es importante entender que datos con discretos o continuos

Pregunta	Ligações	Cidades	Fluxos_OD
¿Cuántos registros hay?	El número total de registros es 1233. Esto significa que el dataset tiene un tamaño moderado.	El número total de registros es 213. Es un dataset relativamente pequeño, lo que facilita su manejo.	El número total de registros es 6749. Este dataset tiene un tamaño considerable, lo que sugiere que podría ser más exigente en términos de procesamiento.
¿Son demasiados pocos?	El número de registros en ligacoes es 1233, lo que podría requerir optimización si el análisis es más complejo o si los datos crecen.	El número de registros en ciudades es pequeño, por lo que no hay problemas de capacidad ni necesidad de optimizaciones.	El número de registros en fluxos_od es 6749, lo que podría requerir optimización, especialmente si se realizan cálculos complejos o el volumen de datos aumenta.
¿Son muchos y no tenemos suficiente capacidad (CPU + RAM)?	El dataset 'ligacoes' ocupa 0.85 MB en memoria. Esto es manejable para el procesamiento sin problemas de capacidad.	El dataset 'ciudades' ocupa 0.08 MB en memoria, lo que lo hace totalmente manejable y no presenta problemas de rendimiento.	El dataset 'fluxos_od' ocupa 5.09 MB, lo cual sigue siendo manejable, aunque más grande que los otros, y aún debería ser procesable sin problemas significativos.
¿Hay registros duplicados?	No hay registros duplicados en 'ligacoes', lo que asegura que los datos sean únicos y los análisis serán precisos.	No hay registros duplicados en 'ciudades', lo que garantiza que no haya sesgo en los resultados debido a la repetición de datos.	No hay registros duplicados en 'fluxos_od', lo que asegura que los resultados del análisis no estén distorsionados por datos repetidos.

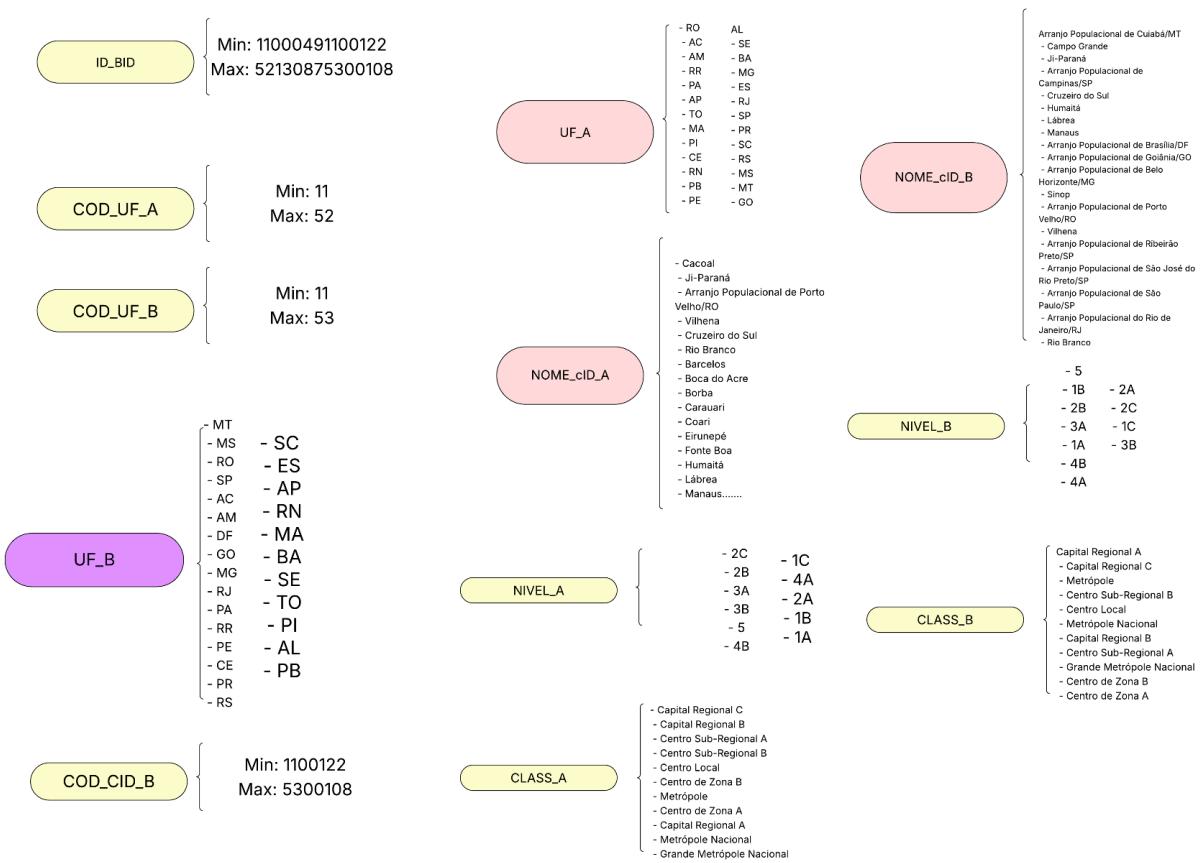
Al categorizar los valores obtenemos, podemos precisar datos únicos , observamos el caso de dataset terrestre y aéreo

El objetivo de este paso es identificar y reportar todos los valores únicos presentes en las columnas categóricas de los datasets, eliminando cualquier repetición para facilitar la interpretación de los datos y su uso en análisis posteriores

DATSET TERRESTRE

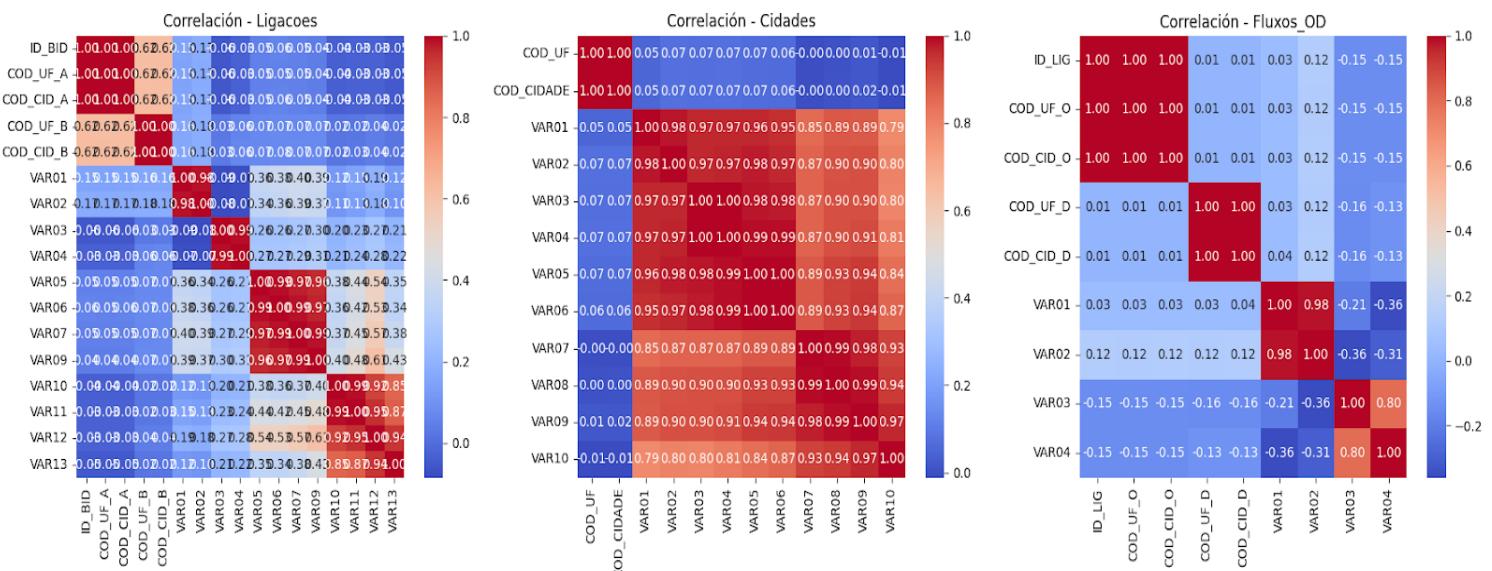


DATASET AÉREO



Sí, los datos en los tres datasets presentan diferentes unidades de medida. Por ejemplo, las columnas relacionadas con la población y el PIB utilizan personas y reales (R\$), respectivamente. Las columnas que representan carga aérea utilizan kilogramos (kg), mientras que las relacionadas con el tiempo de vuelo están en minutos. Además, las tarifas aéreas están en reales (R\$). Si se van a realizar comparaciones o análisis conjuntos, es importante normalizar estas unidades para que los datos sean coherentes.

Los datos categóricos en los tres datasets incluyen nombres de ciudades, unidades federativas y niveles jerárquicos. Ejemplos son UF_A, NOME_CID_A en Ligações y UF, NOME_CIDADE en Cidades. Para usar estos datos en modelos de machine learning, es necesario convertirlos a valores numéricos. Esto se puede hacer mediante técnicas como One-Hot Encoding o Label Encoding, que permiten que los modelos de análisis procesen estos datos de manera eficiente.

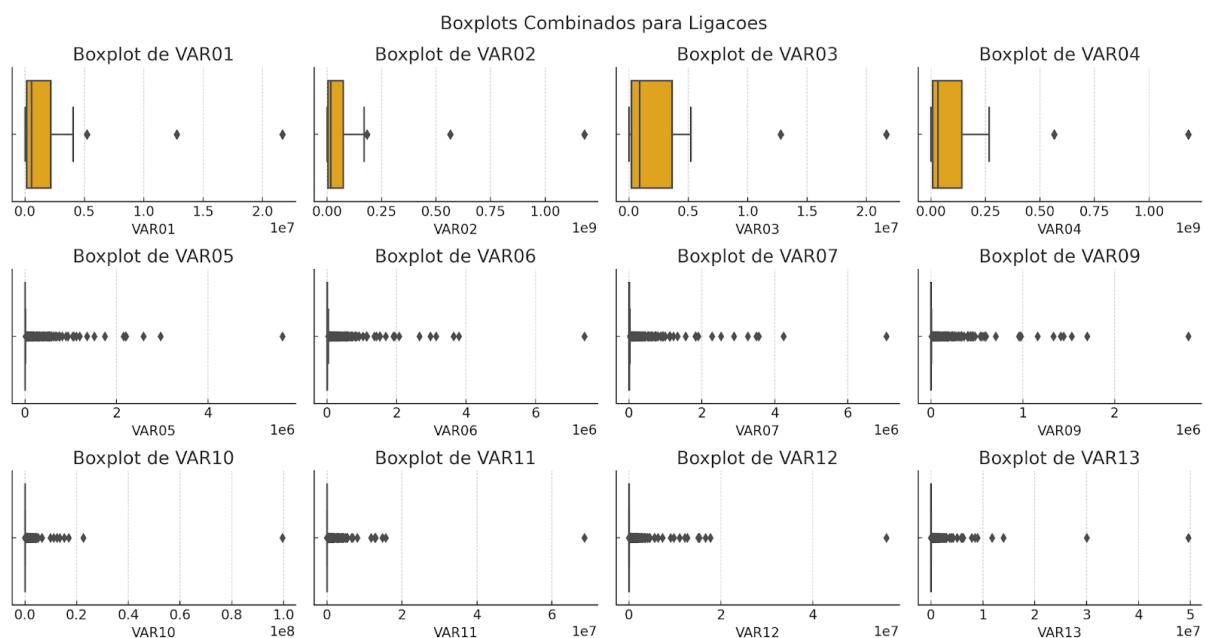


Se realizó la investigación sobre outliers en las columnas de los distintos datasets

LIG_AEREAS_2019-2020_ligacoex.xls

Se detectaron varios outliers en columnas como:

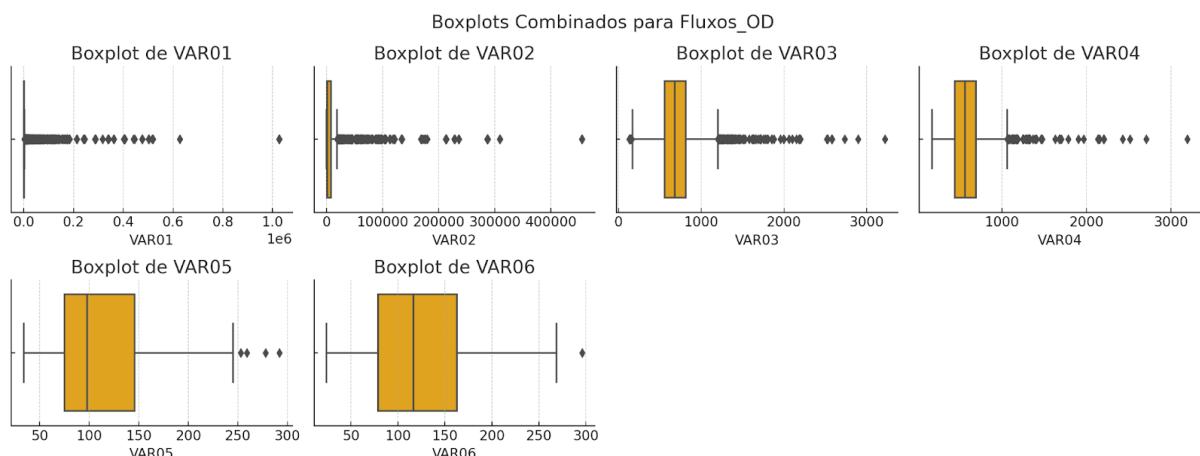
- Total de pasajeros aéreos movimentados entre las Ciudades de origen y destino (2019)
- Producto Interno Bruto (PIB) de la Ciudad A (2018)
- Estimación de la población residente de la Ciudad A (2019),
- Producto Interno Bruto (PIB) de la Ciudad B (2018), entre otras.



LIG_AEREAS_2019-2020_cidades.xlsx

En el dataset Fluxos_OD, los outliers también fueron identificados en varias columnas como:

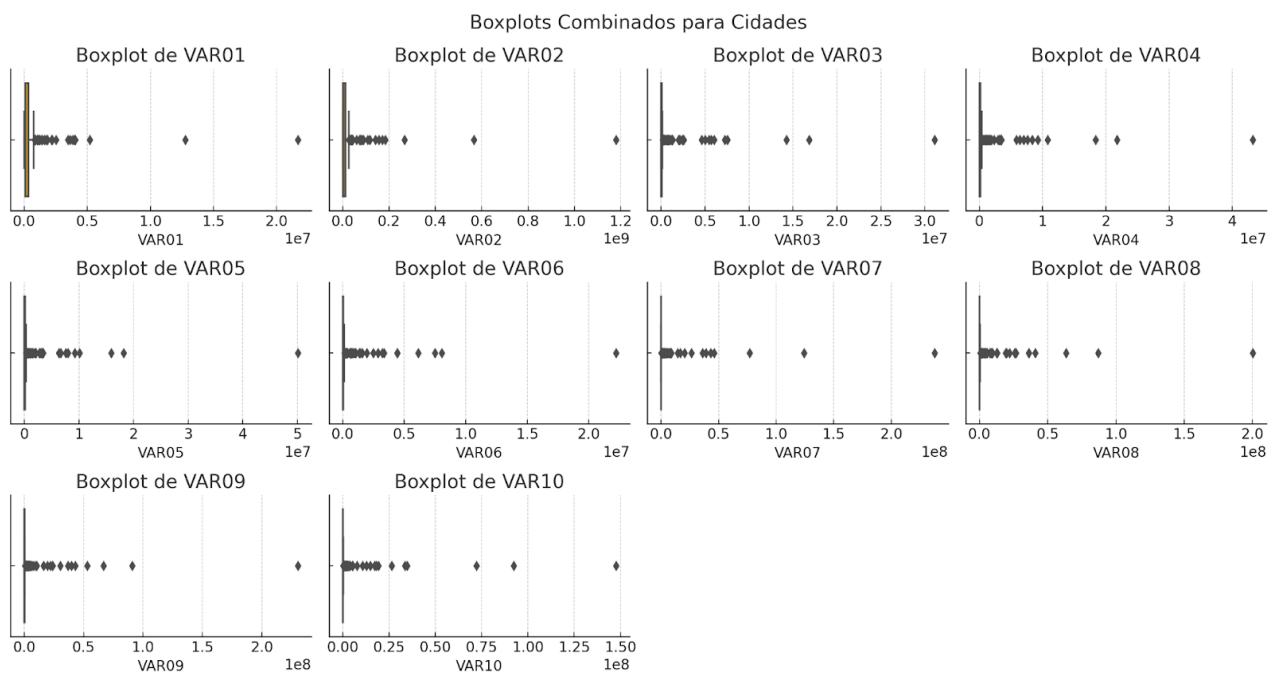
- Total de pasajes aéreos comercializados entre las Ciudades de origen y destino (2019)
- Producto Interno Bruto (PIB) de la Ciudad A (2018)
- Estimación de la población residente de la Ciudad A (2019), etc.



LIG_AEREAS_2019-2020_fluxos_od.xlsx

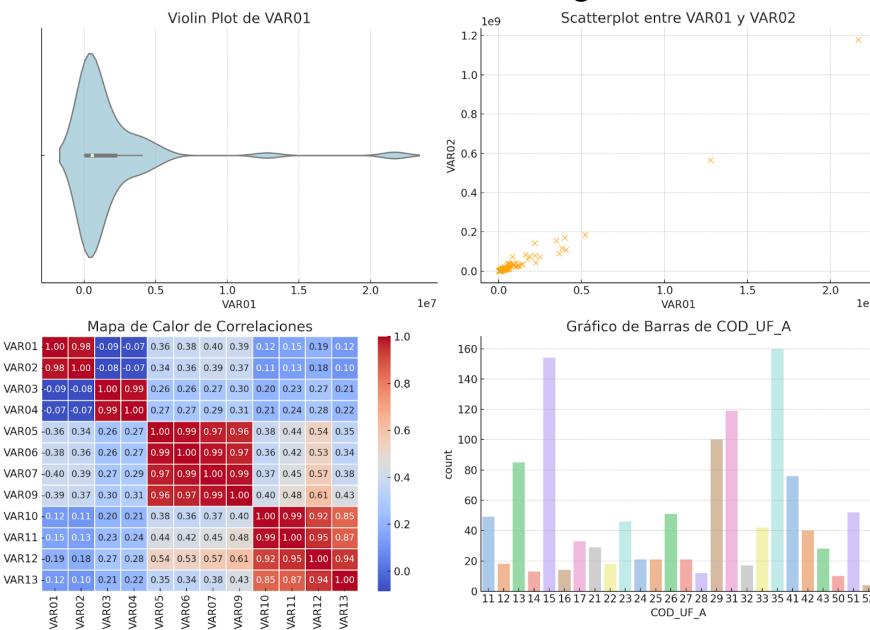
Finalmente, en el dataset Cidades, se observaron outliers significativos en varias columnas, incluyendo:

- Estimación de la población residente de la Ciudad A (2019)
- Producto Interno Bruto (PIB) de la Ciudad A (2018)
- Estimación de la población residente de la Ciudad B (2019), entre otras.

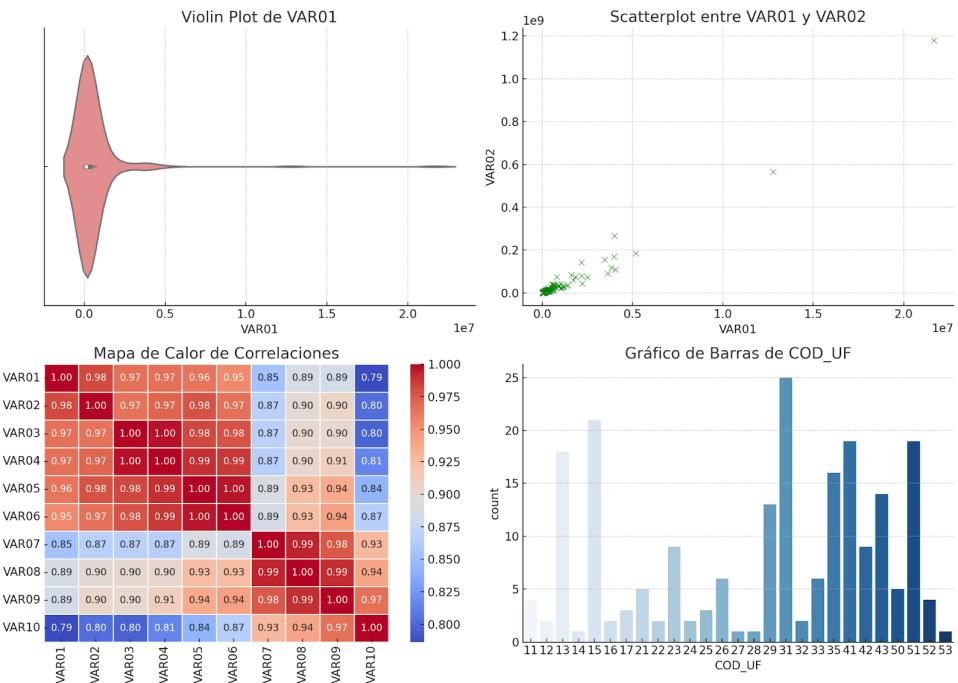


VISUALIZACIÓN GENERAL DE DATOS

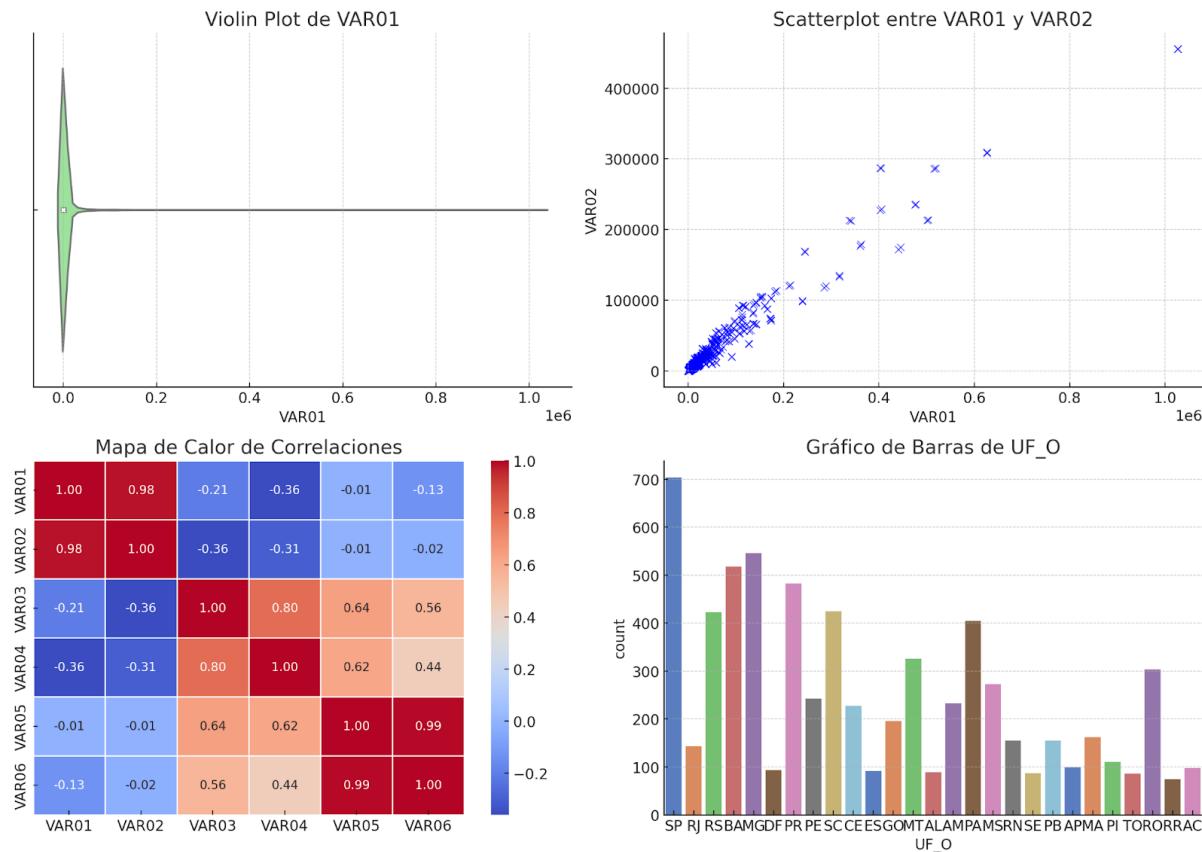
LIG_AEREAS_2019-2020_ligacoes.xls



LIG_AEREAS_2019-2020_cidades.xlsx



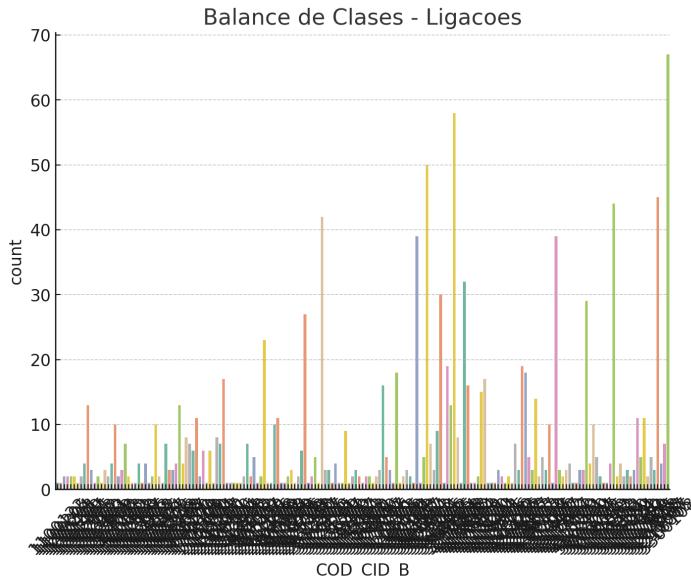
LIG_AEREAS_2019-2020_fluxos_od.xlsx



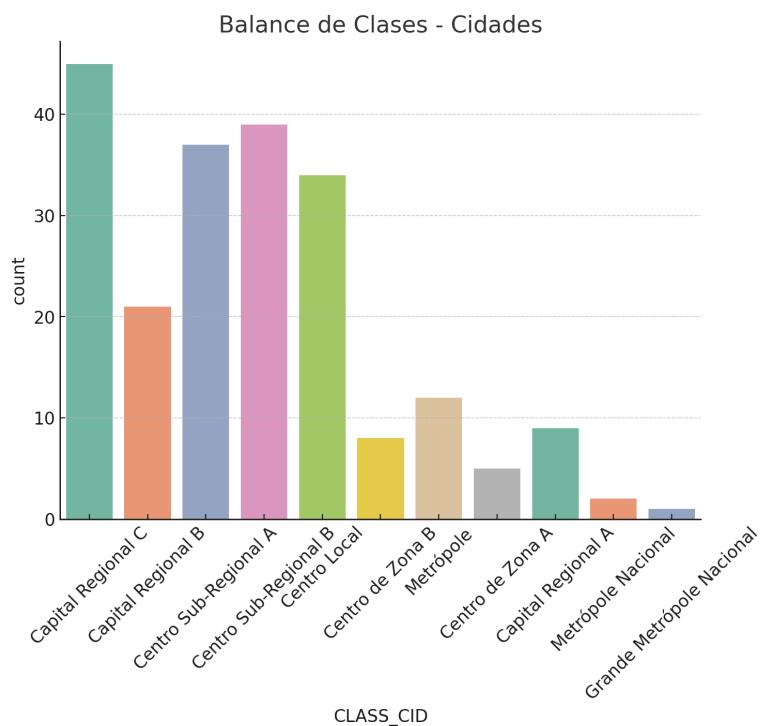
Paso 4. Encuentra un problema potencial en tus datos.

- Si es un problema de tipo supervisado:
 - ¿Cuál es la columna de “salida”? ¿binaria, multiclasificación?
 - ¿Está balanceado el conjunto salida?

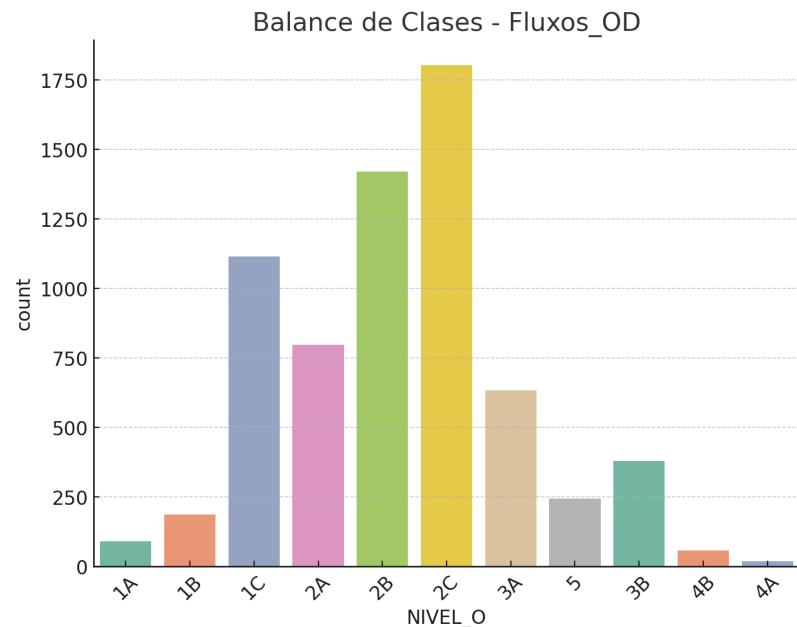
LIG_AEREAS_2019-2020_ligacoes.xls



LIG_AEREAS_2019-2020_cidades.xlsx

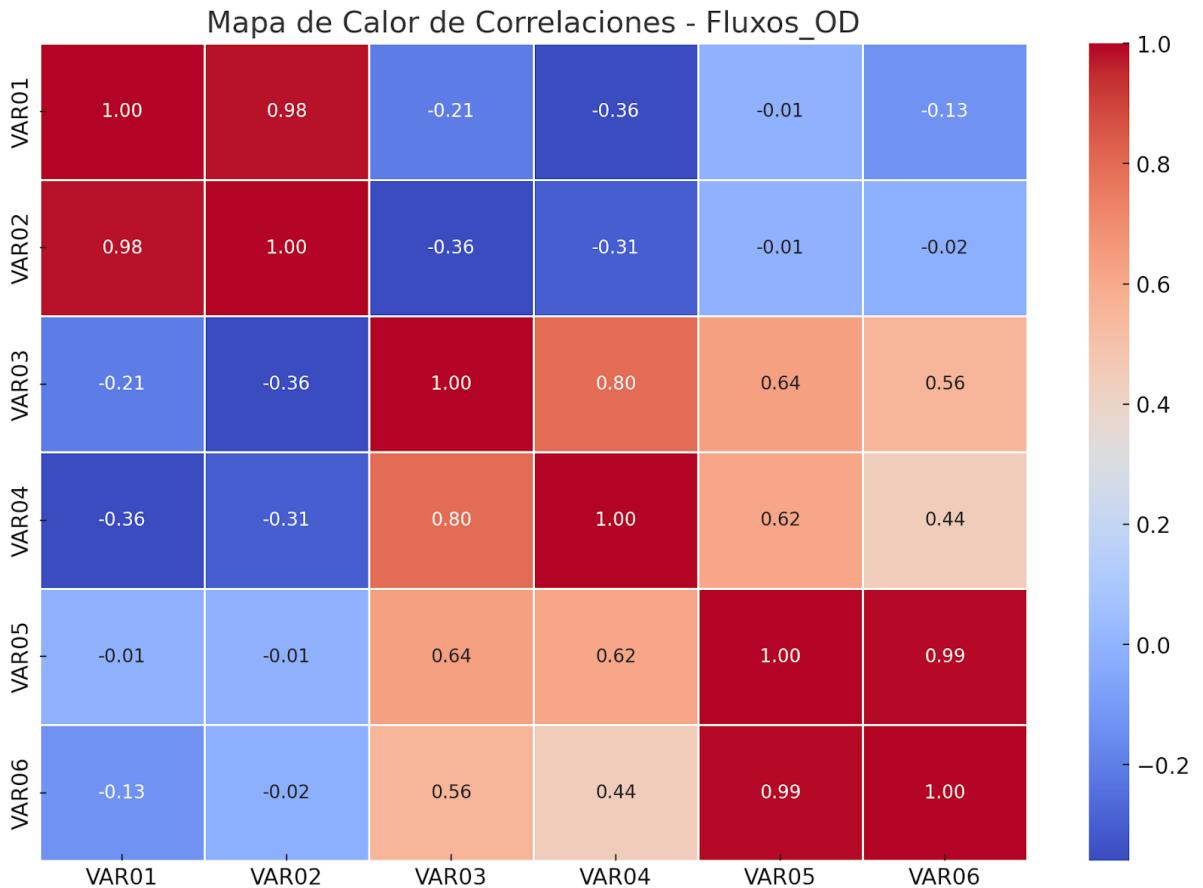


LIG_AEREAS_2019-2020_fluxos_od.xlsx



Basándonos en el análisis del balance de clases que realizamos previamente, podemos concluir que hay un desbalance en las clases. Algunas clases tienen muchas muestras, mientras que otras solo tienen una. Esto puede limitar la capacidad de generalización del modelo, especialmente para las clases menos representadas.

LIG_AEREAS_2019-2020_fluxos.xlsx



Conclusión

- El análisis ha mostrado que los datasets son problemas supervisados multiclase, con desbalance significativo en las clases de salida. Esto implica que las clases con menos muestras podrían afectar el rendimiento del modelo, por lo que se recomienda aplicar técnicas de sobremuestreo o submuestreo.
- He identificado que algunas variables están altamente correlacionadas entre sí, lo que sugiere redundancia. Esto puede generar problemas de multicolinealidad, por lo que se recomienda eliminar variables redundantes o utilizar técnicas de selección de características.
- Aunque no hay valores faltantes en los datasets, se identificaron outliers en las distribuciones de las variables. Estos pueden afectar las predicciones, por lo que se

deben investigar para determinar si son errores de carga o eventos excepcionales que deben ser mantenidos.

- La ausencia de una columna de fecha en los datasets indica que no estamos ante un problema de series temporales, lo que simplifica el modelo al eliminar la necesidad de manejar la temporalidad.
- Las correlaciones sorprendentes entre algunas variables, como las que presentan relaciones altas, sugieren que algunas de ellas podrían estar midiendo el mismo fenómeno, lo que debe considerarse al seleccionar características.
- A pesar de tener suficiente variedad en las clases, el desbalance de clases podría limitar la capacidad de generalización del modelo. Se recomienda aplicar estrategias para mejorar el manejo de este desbalance y la precisión en las clases menos representadas.

ANEXOS

https://colab.research.google.com/drive/1CaleuaUfSu06McJ_omxJ9NxhIehKyPmJ?usp=sharing