Universidad Nacional de San Agustin de Arequipa

Año de la recuperación y consolidación de la economía peruana

Escuela Profesional de Ciencia de la Computación

Análisis Data Wrangling

Informe Data wrangling

A-2025

Docente: Dra. Ana Maria CuadrosValdivia

Javier Wilber Quispe Rojas

Índice

1. Contexto de mi DataSet

1.1. Boston_feature_df:

Este dataset contiene información sobre diferentes áreas de Boston, identificadas por códigos geográficos. Para cada área, muestra la proporción de diferentes grupos raciales como blancos, negros, asiáticos e hispanos, tanto de la población residente como de quienes llegan a esas zonas. También incluye datos sobre los ingresos de las personas en cada área, divididos en rangos económicos, y cómo estos ingresos se distribuyen entre quienes se trasladan hacia esas zonas. Además, registra la intensidad con la que la gente visita distintos tipos de lugares, como tiendas, restaurantes, hospitales o sitios religiosos. Cada área está geolocalizada con latitud y longitud, y se especifica la población total que vive ahí.

1.2. BostonMobility2021:

Este archivo muestra los movimientos de personas dentro de Boston durante 2021, indicando desde dónde salen y hacia dónde se dirigen. Cada fila representa un flujo de personas entre dos áreas geográficas, con las coordenadas de origen y destino, además del número de visitantes que se trasladaron en un período específico, generalmente semanal. También incluye estimaciones de la población en movimiento y, en algunos casos, datos sobre la cantidad de dispositivos móviles activos en el área durante el día. Este dataset es mucho más grande y permite analizar con detalle los patrones de movilidad dentro de la ciudad.

2. Descripción del data set

2.1. Cuantos Registros tiene cada DataSet

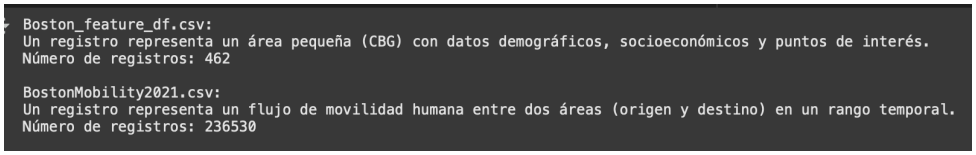

Figura 1: Cantidad de Registros por Data set

Boston_feature_df.csv:
Este dataset tiene 462 registros, lo que es bastante manejable para casi cualquier computadora actual. No es un volumen de datos grande, por lo que se puede procesar fácilmente en memoria sin necesidad de herramientas especiales o grandes recursos de CPU y RAM. Ideal para análisis rápidos y exploratorios.

BostonMobility2021.csv:
Con 236,530 registros, este dataset es considerablemente más grande, pero sigue siendo posible de manejar en la mayoría de las computadoras personales modernas, especialmente si cuentas con al menos 8 GB de RAM. Dependiendo del tipo de análisis , podrías experimentar un poco más de uso de CPU y memoria, pero en general no es un volumen masivo que requiera infraestructura de big data.

2.2. Descripcion DataSet Boston_feature_df.csv

¿Cuál es el objeto u entidad de estudio?
El dataset estudia diferentes áreas o sectores geográficos (probablemente census tracts o zonas censales) dentro de la ciudad de Boston. Cada fila representa una unidad espacial que contiene información demográfica, socioeconómica y de actividad social de esa zona.

Cuadro 1: Identificación y composición racial

Atributo	Significado / Descripción	Tipo de dato	Rango aproximado / Valores
GEOID	Identificador único del área geográfica (sector o tract)	Categórico (texto)	Código único por zona
white	Proporción de población blanca en el área	Numérico decimal	0 a 1
black	Proporción de población negra en el área	Numérico decimal	0 a 1
asian	Proporción de población asiática en el área	Numérico decimal	0 a 1
hispanic	Proporción de población hispana en el área	Numérico decimal	0 a 1
white_inflow	Proporción de población blanca que ingresa a esa área	Numérico decimal	0 a 1
black_inflow	Proporción de población negra que ingresa a esa área	Numérico decimal	0 a 1
hispanic_inflow	Proporción de población hispana que ingresa a esa área	Numérico decimal	0 a 1
asian_inflow	Proporción de población asiática que ingresa a esa área	Numérico decimal	0 a 1

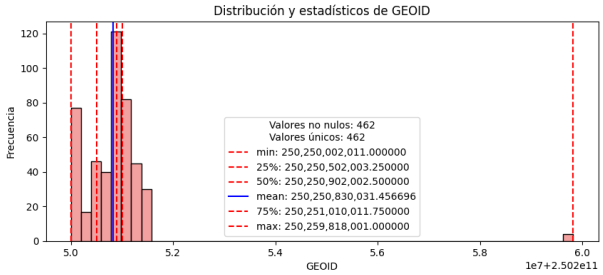

Figura 2: GEOID

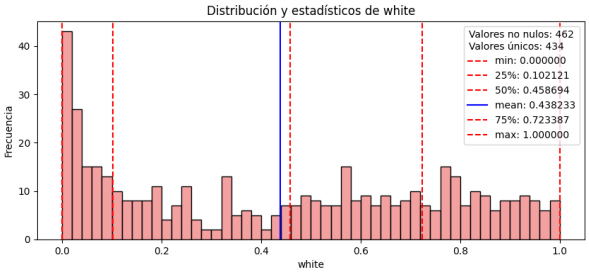

Figura 3: white

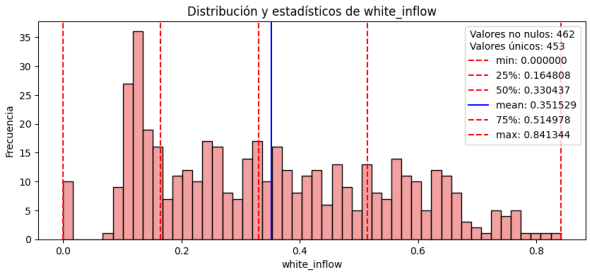

Figura 4: white_inflow

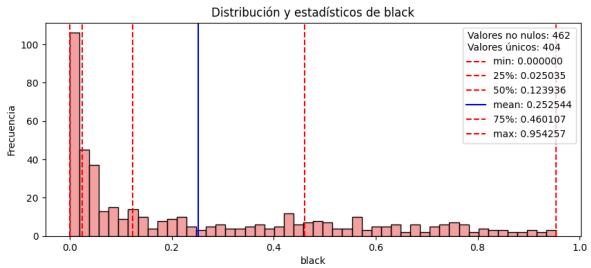

Figura 5: black

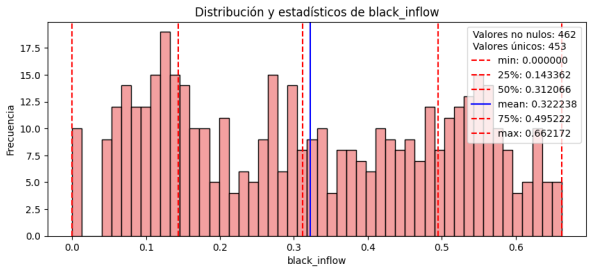

Figura 6: black_inflow

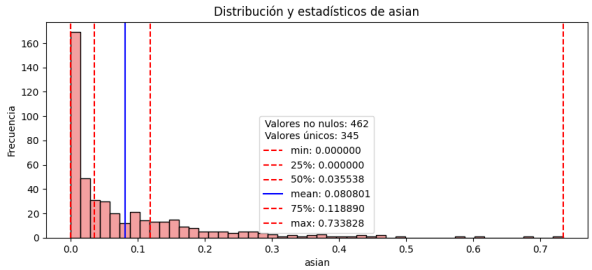

Figura 7: asian

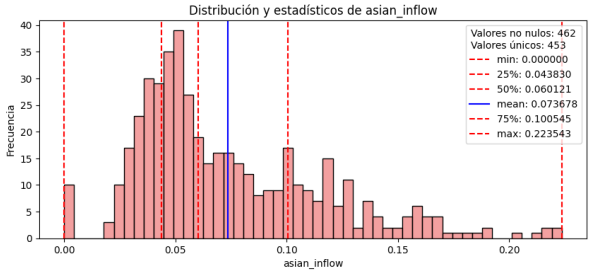

Figura 8: asian_inflow

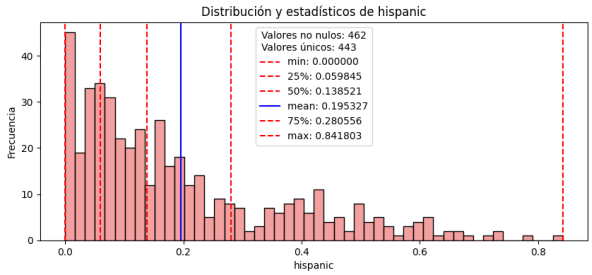

Figura 9: hispanic

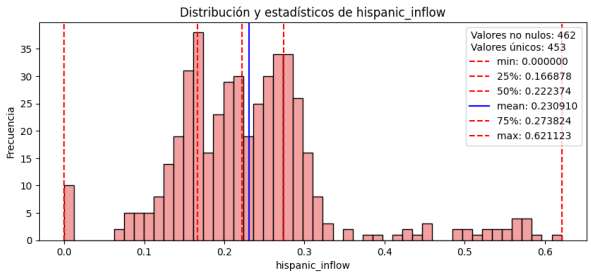

Figura 10: hispanic_inflow

Cuadro 2: Ingreso y flujo de ingreso

Atributo	Significado / Descripción	Tipo de dato	Rango aproximado / Valores
Under $50K	Proporción de población con ingresos menores a $50,000	Numérico decimal	0 a 1
$50K - $100K	Proporción con ingresos entre $50,000 y $100,000	Numérico decimal	0 a 1
$100K - $200K	Proporción con ingresos entre $100,000 y $200,000	Numérico decimal	0 a 1
Over $200K	Proporción con ingresos mayores a $200,000	Numérico decimal	0 a 1
Under $50K_inflow	Proporción con ingresos menores a $50,000 que ingresan a esa área	Numérico decimal	0 a 1
$50K - $100K_inflow	Proporción con ingresos entre $50,000 y $100,000 que ingresan a esa área	Numérico decimal	0 a 1
$100K - $200K_inflow	Proporción con ingresos entre $100,000 y $200,000 que ingresan a esa área	Numérico decimal	0 a 1
Over $200K_inflow	Proporción con ingresos mayores a $200,000 que ingresan a esa área	Numérico decimal	0 a 1

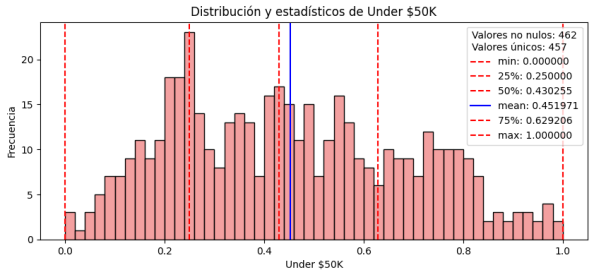

Figura 11: Under $50K

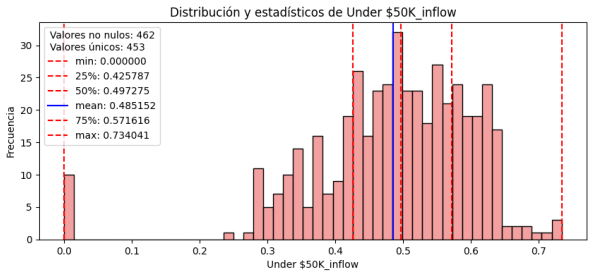

Figura 12: Under $50K_inflow

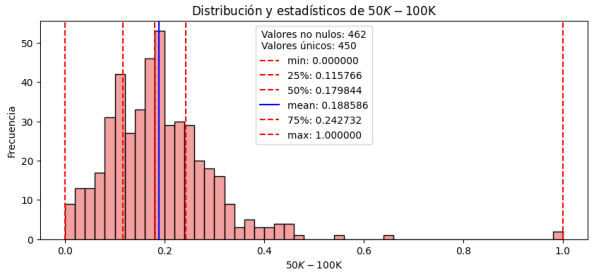

Figura 13: $50K - $100K

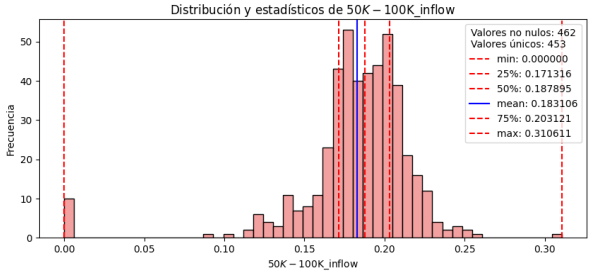

Figura 14: $50K - $100K_inflow

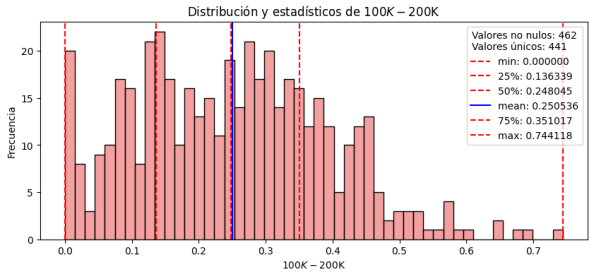

Figura 15: $100K - $200K

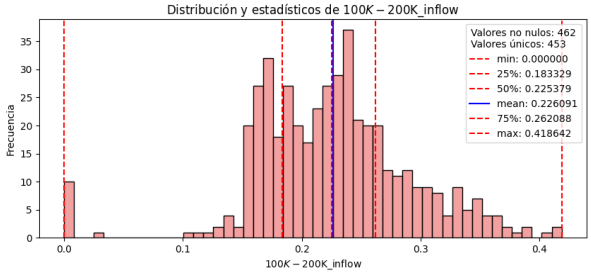

Figura 16: $100K - $200K_inflow

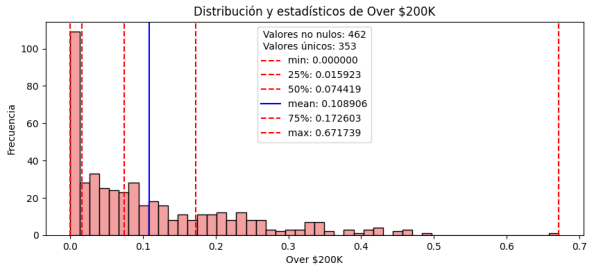

Figura 17: Over $200K

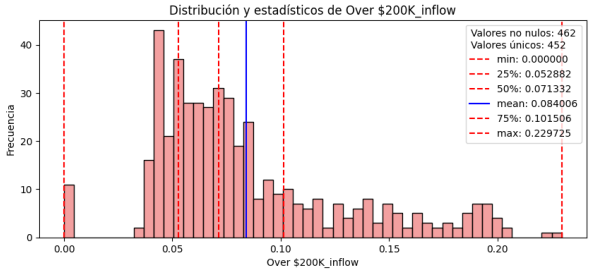

Figura 18: Over $200K_inflow

Cuadro 3: Actividades y geolocalización

Atributo	Significado / Descripción	Tipo de dato	Rango aproximado / Valores
Food	Nivel o cantidad de actividad relacionada con comida en el área	Numérico decimal	Valores positivos (escala no definida)
Shopping	Nivel o cantidad de actividad de compras en el área	Numérico decimal	Valores positivos (escala no definida)
Work	Nivel o cantidad de actividad laboral en el área	Numérico decimal	Valores positivos (escala no definida)
Health	Nivel o cantidad de actividad relacionada con salud en el área	Numérico decimal	Valores positivos (escala no definida)
Religious	Nivel o cantidad de actividad religiosa en el área	Numérico decimal	Valores positivos (escala no definida)
Service	Nivel o cantidad de actividad de servicios en el área	Numérico decimal	Valores positivos (escala no definida)
Entertainment	Nivel o cantidad de actividad de entretenimiento en el área	Numérico decimal	Valores positivos (escala no definida)
Grocery	Nivel o cantidad de actividad de supermercados/grocery en el área	Numérico decimal	Valores positivos (escala no definida)
Education	Nivel o cantidad de actividad educativa en el área	Numérico decimal	Valores positivos (escala no definida)
Arts/Museum	Nivel o cantidad de actividad cultural y museos en el área	Numérico decimal	Valores positivos (escala no definida)
Transportation	Nivel o cantidad de actividad de transporte en el área	Numérico decimal	Valores positivos (escala no definida)
Sports	Nivel o cantidad de actividad deportiva en el área	Numérico decimal	Valores positivos (escala no definida)
LATITUDE	Latitud geográfica del centro del área	Numérico decimal	Aproximadamente 42.x (Boston)
LONGITUDE	Longitud geográfica del centro del área	Numérico decimal	Aproximadamente -71.x (Boston)
total_population	Número total estimado de habitantes en el área	Numérico entero	Desde decenas hasta varios miles

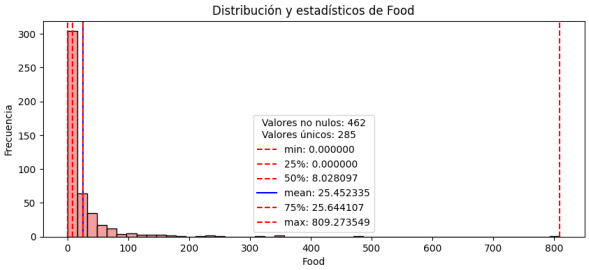

Figura 19: Food

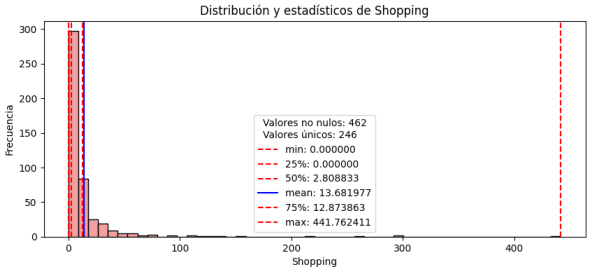

Figura 20: Shopping

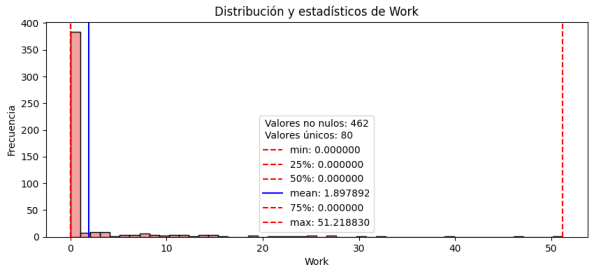

Figura 21: Work

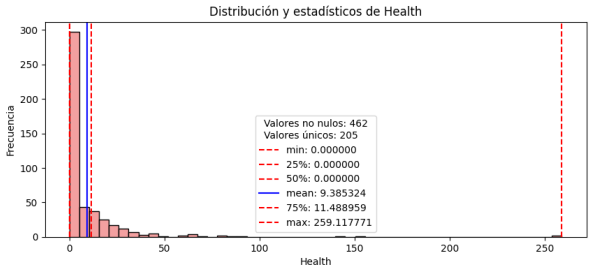

Figura 22: Health

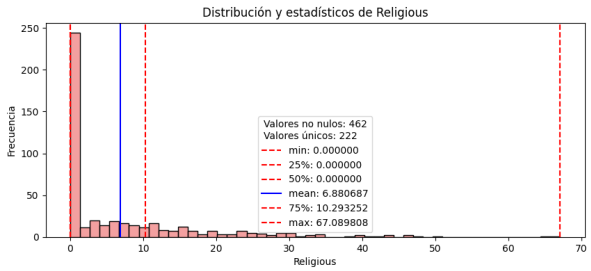

Figura 23: Religious

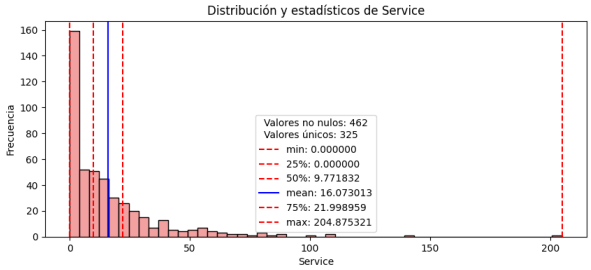

Figura 24: Service

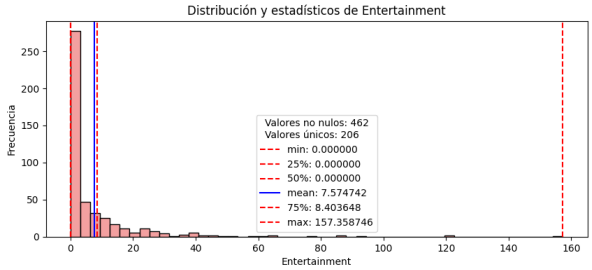

Figura 25: Entertainment

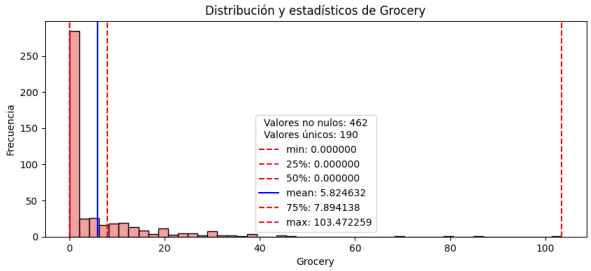

Figura 26: Grocery

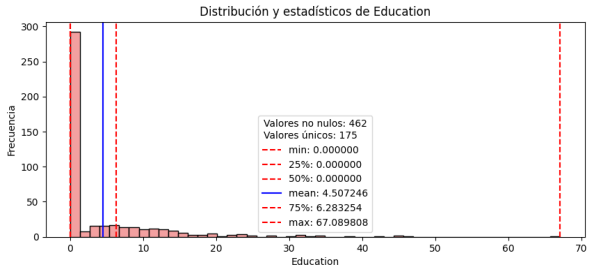

Figura 27: Education

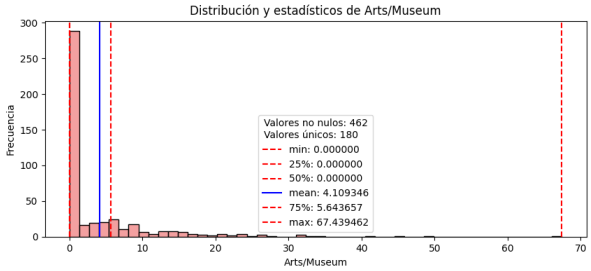

Figura 28: Arts_Museum

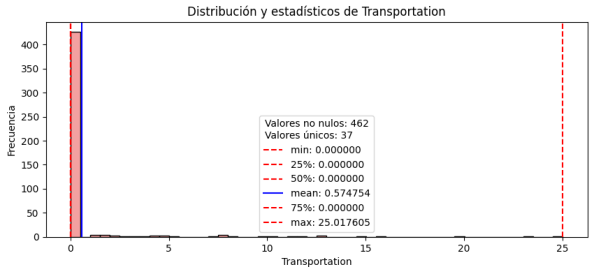

Figura 29: Transportation

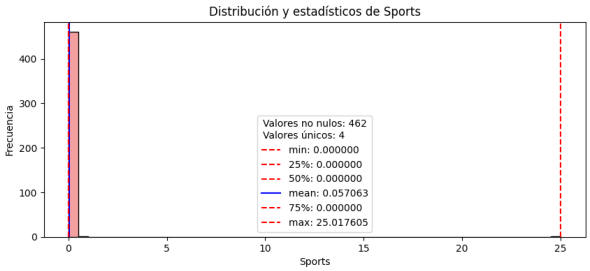

Figura 30: Sports

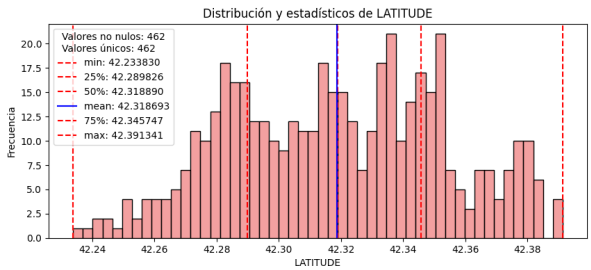

Figura 31: LATITUDE

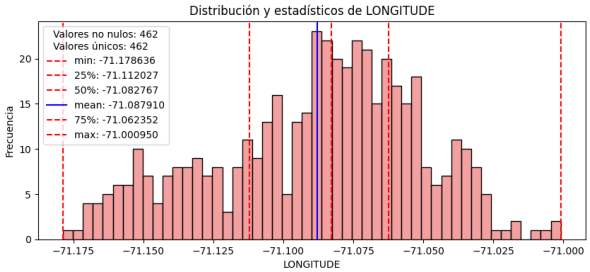

Figura 32: LONGITUDE

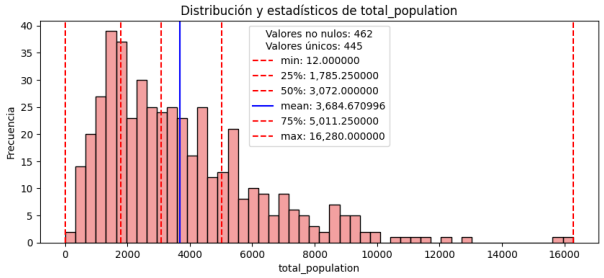

Figura 33: total_population

2.3. Descripción de las columnas del dataset BostonMobility2021.csv

¿Cuál es el objeto u entidad de estudio?

El objeto de estudio son los flujos de movilidad de personas entre diferentes áreas geográficas dentro de la ciudad de Boston. Es decir, cada registro representa la cantidad de personas que se desplazan desde una zona de origen hacia una zona de destino en un rango de fechas determinado.

Cuadro 4: Resumen del objeto de estudio y atributos del dataset BostonMobility2021

Atributo	Significado / Descripción	Tipo de dato	Rango aproximado / Valores
geoid_o	Identificador del área de origen del flujo	Categórico (texto)	Código único por zona
geoid_d	Identificador del área de destino del flujo	Categórico (texto)	Código único por zona
lng_o	Longitud del área de origen	Numérico decimal	Aproximadamente -71.x (Boston)
lat_o	Latitud del área de origen	Numérico decimal	Aproximadamente 42.x (Boston)
lng_d	Longitud del área de destino	Numérico decimal	Aproximadamente -71.x (Boston)
lat_d	Latitud del área de destino	Numérico decimal	Aproximadamente 42.x (Boston)
date_range	Rango de fechas del registro de movilidad	Categórico (texto)	Ejemplo: ”01/04/21 - 01/10/21”
visitor_flows	Número de visitantes entre áreas en el rango de fechas	Numérico entero	Desde 0 hasta valores variables
pop_flows	Población total estimada asociada al flujo o área	Numérico decimal	Valores positivos variados
census_block	Agrupación censal (datos mayormente ausentes)	Desconocido / vacío	Mayormente NaN
number_devices	Número estimado de dispositivos activos en área origen durante el día	Numérico entero o NaN	Valores numéricos o vacíos

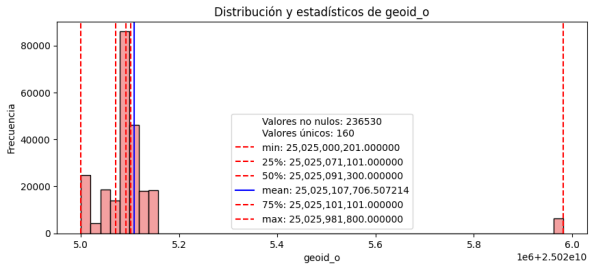

Figura 34: geoid origen

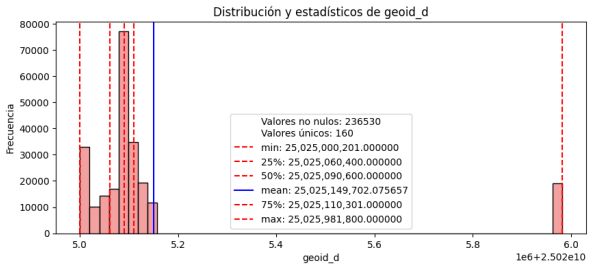

Figura 35: geoid destino

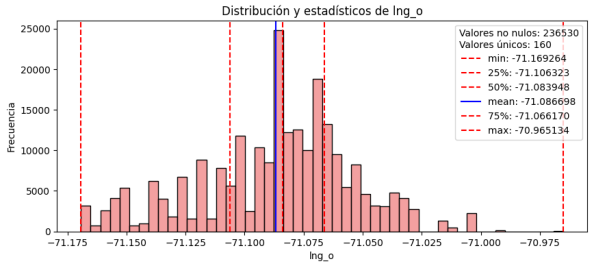

Figura 36: LONGITUDE Origen

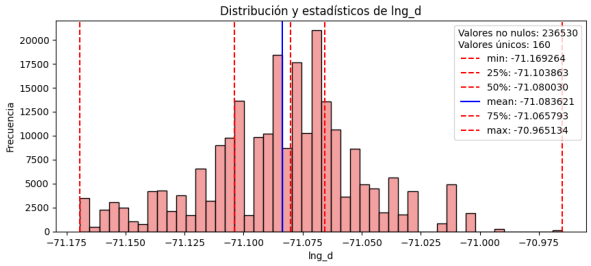

Figura 37: LONGITUDE Destino

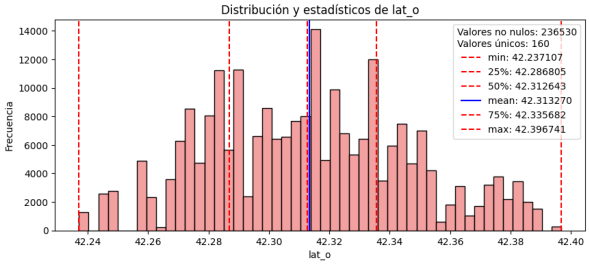

Figura 38: LATITUDE Origen

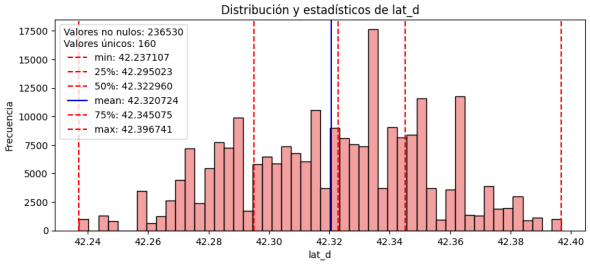

Figura 39: LATITUDE Destino

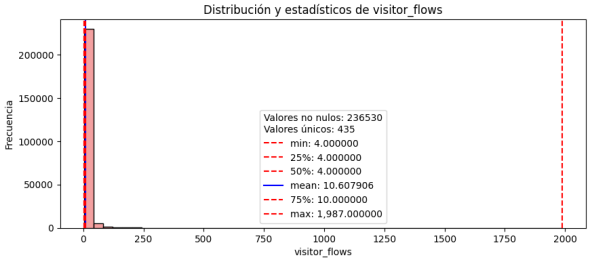

Figura 40: Flujo de persona

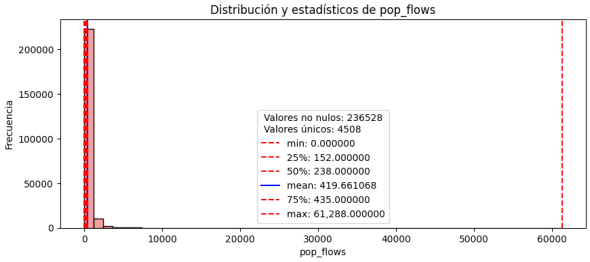

Figura 41: Población total

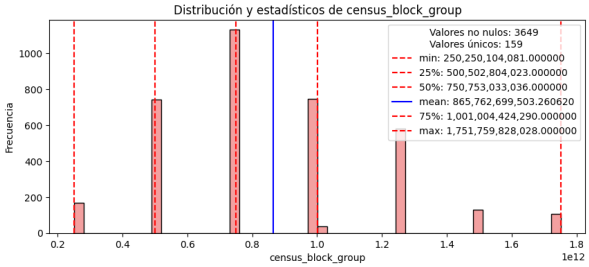

Figura 42: Agrupación censal

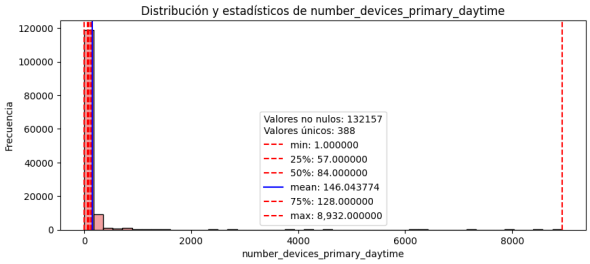

Figura 43: dispositivos activos

2.4. ¿Que contiene cada Registro

2.4.1. Descripción de un Registro del dataset Boston_feature_df

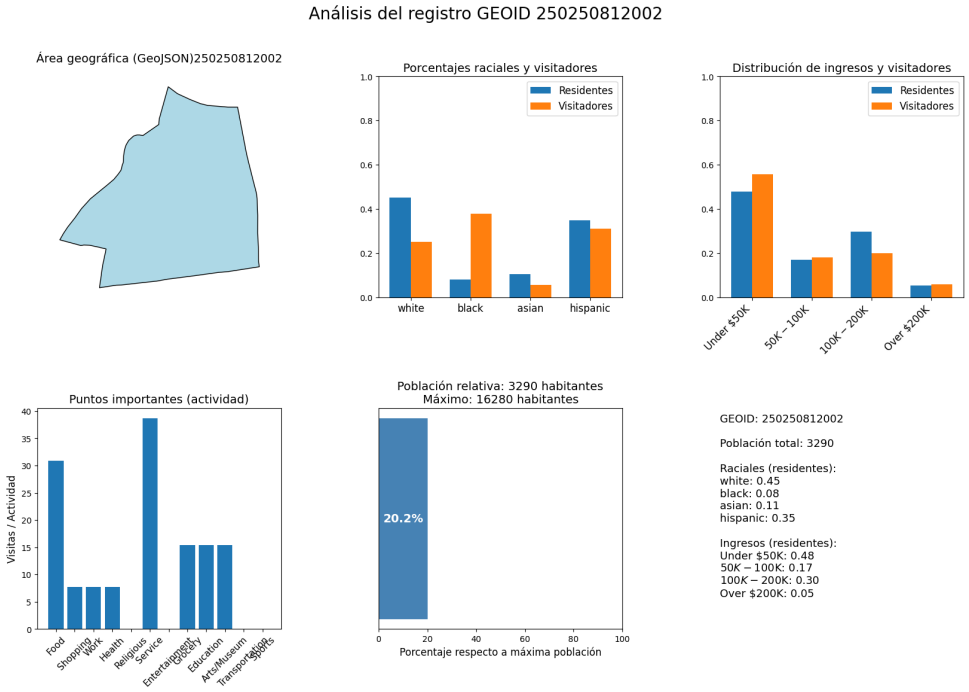

Figura 44: Registro Geoid:250250812002

Cada fila en el dataset representa un área geográfica pequeña dentro de Boston, identificada por un código único llamado `GEOID`. Este código corresponde a un bloque censal o grupo de bloques con características propias.

Tomando como ejemplo el registro con `GEOID = 250250812002`, podemos describir la fila de la siguiente manera:

- **Ubicación geográfica:** La fila corresponde a un polígono delimitado en el mapa que define el área específica de ese bloque censal.
- **Composición racial y visitantes:** Se presentan las proporciones de la población residente dividida en grupos raciales: *white*, *black*, *asian* y *hispanic*, así como la composición de los visitantes que llegan a esa área. Esto refleja la diversidad y dinámica demográfica de la zona.
- **Distribución de ingresos y visitantes:** La información incluye la distribución de ingresos de los residentes en cuatro rangos económicos y la composición similar de los visitantes, brindando un panorama socioeconómico detallado.
- **Puntos importantes y actividad:** Se muestran las intensidades de visitas o actividades en diferentes tipos de lugares, como comida, tiendas, trabajo, salud, servicios y entretenimiento, que reflejan el uso y atractivo de la zona.
- **Población relativa:** Se presenta como la poblacion total de esta área.

2.4.2. Granularidad del dataset Boston Model Feature

El dataset Boston Model Feature contiene información dividida y organizada por áreas geográficas pequeñas, identificadas mediante la columna `GEOID`, que representa zonas como bloques censales o grupos de bloques. Esto significa que la granularidad espacial es fina o media, ya que los datos están detallados para cada área específica dentro de la ciudad de Boston.

Esta organización permite realizar análisis precisos sobre características demográficas, flujos de población y patrones de uso del espacio a nivel local. En lugar de tener solo un resumen general de toda la ciudad, los datos muestran diferencias y variaciones entre las distintas zonas.

En resumen, el dataset tiene una granularidad espacial que permite realizar análisis detallados y focalizados en áreas específicas de Boston.

2.4.3. Descripción de un registro del dataset BostonMovility.csv

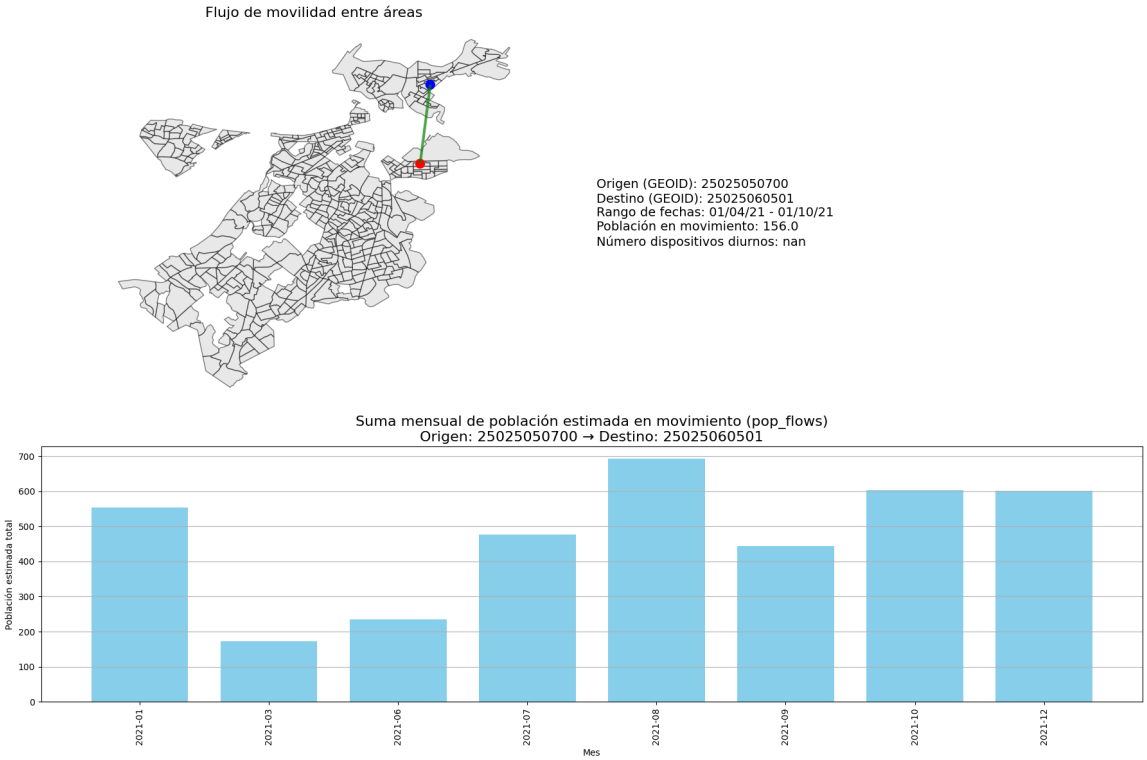

Figura 45: Registro de Boston Movility

Cada registro representa un flujo de movilidad entre dos áreas geográficas específicas durante un rango de fechas determinado. En detalle, cada fila contiene:

- **Área de origen (GEOID):** Un código único que identifica la zona desde donde salen las personas (en el gráfico, marcado con un punto rojo).
- **Área destino (GEOID):** Un código único que identifica la zona a donde llegan las personas (en el gráfico, marcado con un punto azul).
- **Coordenadas (latitud y longitud):** Ubicación espacial de las áreas de origen y destino en el mapa.
- **Rango de fechas:** Periodo durante el cual se mide el flujo de movilidad (por ejemplo, del 01/04/21 al 01/10/21).
- **Población en movimiento (`pop_flows`):** Estimación del total de personas que se movilizan entre el área de origen y destino en ese periodo.
- **Número de dispositivos móviles diurnos:** Cantidad de dispositivos detectados durante el día que soportan la estimación de movilidad (en algunos casos puede no estar disponible).
- **Gráfica de barras mensual:** Muestra la evolución mensual estimada del flujo de población en movimiento entre las dos áreas durante el año 2021.

2.4.4. Granularidad del dataset Boston Mobility

La granularidad se refiere al nivel de detalle con el que se representan los datos en el conjunto. En el caso del dataset Boston Mobility, podemos distinguir dos tipos principales de granularidad: espacial y temporal.

Granularidad espacial:

- Los datos están representados por **GEOIDs**, que corresponden a áreas pequeñas como bloques censales o grupos de bloques, lo que permite un análisis detallado de la movilidad a nivel local.
- Además, se incluyen las coordenadas precisas de latitud y longitud para los puntos de origen y destino, facilitando la visualización geográfica exacta de los flujos de movilidad.
- Esta granularidad fina o media es especialmente útil para estudios urbanos, planificación de transporte y análisis de patrones de movimiento a pequeña escala.

Granularidad temporal:

- El campo `date_range` contiene rangos semanales (por ejemplo, "01/04/21 - 01/10/21"), lo que indica que los datos están agregados a nivel de semana.
- No se cuenta con datos diarios ni horarios, por lo que la granularidad temporal es media, adecuada para análisis de tendencias y patrones semanales, pero no para estudios con alta resolución temporal.
- Esta resolución semanal es suficiente para identificar cambios significativos en la movilidad debido a eventos o políticas que afectan períodos de tiempo de varios días.

2.5. Manejo de columnas con valores nulos en el dataset Boston Mobility

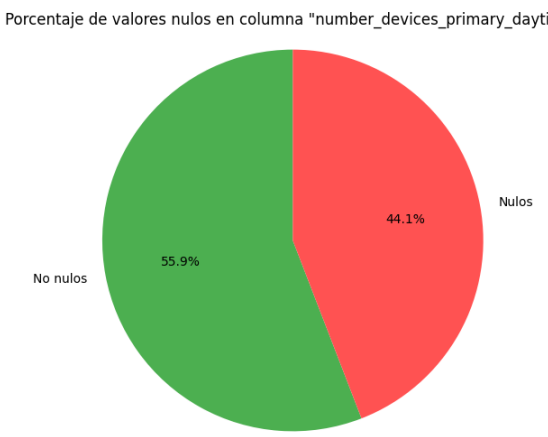

Figura 46: Valores Nulos

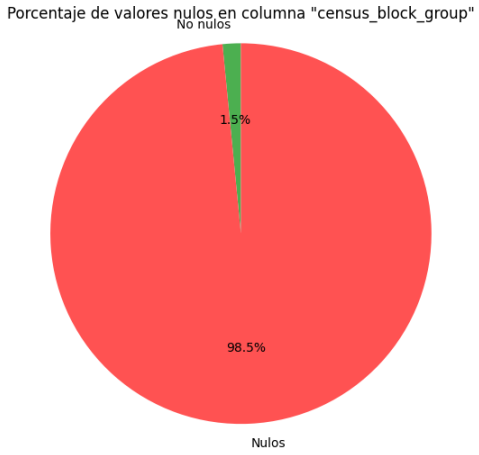

Figura 47: Valores NUlos

En el análisis del dataset Boston Mobility se identificaron dos columnas con un alto porcentaje de valores nulos:

- **census_block_group** con aproximadamente un 98.5 % de valores nulos. Esta columna presenta una ausencia casi total de datos, lo que dificulta cualquier intento confiable de imputación. Además, es redundante con las columnas `geoid_o` y `geoid_d`, que ya proporcionan una identificación geográfica precisa.
- **number_devices_primary_daytime** con cerca del 44 % de valores nulos. Aunque esta variable es útil para validar la representatividad de los datos de movilidad, la elevada proporción de datos faltantes implica que su imputación podría introducir sesgos significativos sin una estrategia sólida y datos adicionales de soporte.

Por lo tanto, se recomienda eliminar ambas columnas del análisis debido a:

- La redundancia y falta de datos en `census_block_group`, que limita su utilidad.
- El riesgo de sesgo y ruido que la imputación de `number_devices_primary_daytime` podría generar, dado el alto porcentaje de datos faltantes.

En consecuencia, es preferible concentrarse en las columnas limpias y completas que ofrecen información clave para el análisis de movilidad, garantizando así la calidad y confiabilidad de los resultados.

2.6. Analisis outliers en los Datasets

2.6.1. outliers Dataser Boston feature

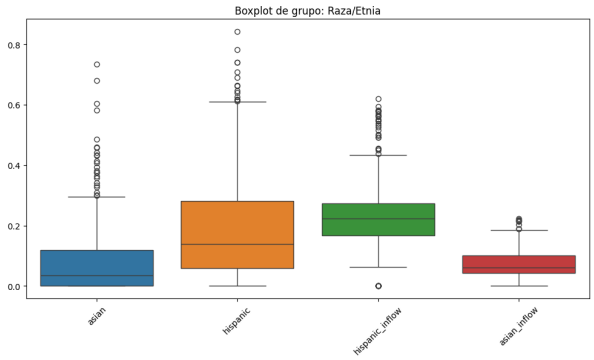

Figura 48: outliers Race

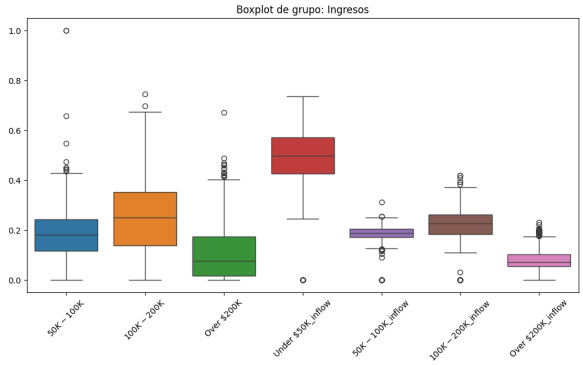

Figura 49: outliers income

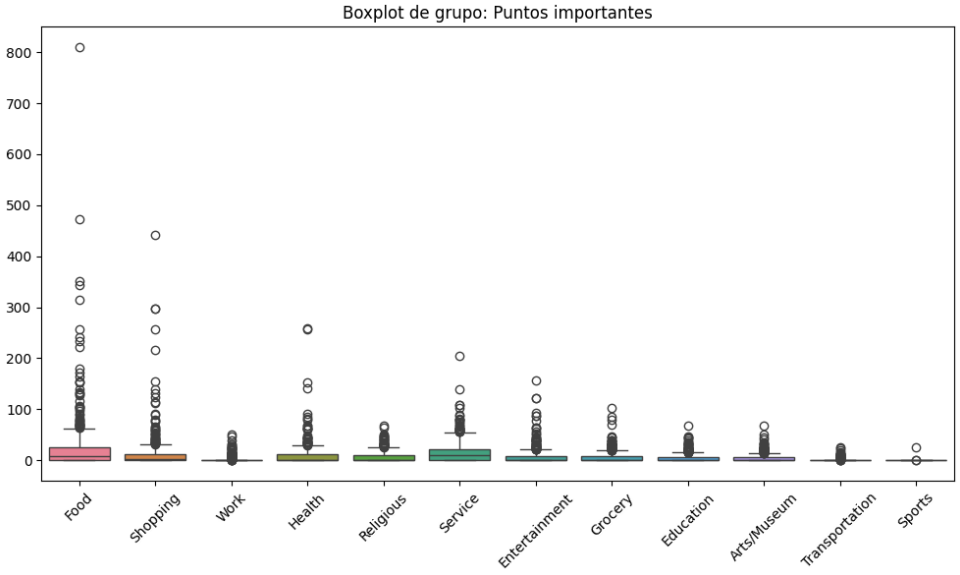

Figura 50: outliers Puntos importantes

En el análisis de segregación por grupos sociales, ingresos y la relación con puntos importantes, es fundamental reflexionar cuidadosamente sobre el tratamiento de los outliers. Estos valores extremos no siempre representan errores o datos atípicos sin sentido; en muchos casos, reflejan condiciones reales y relevantes del entorno social y económico. Por ejemplo, algunos barrios pueden presentar una concentración muy alta de un determinado grupo étnico o niveles de ingresos que difieren marcadamente del promedio, lo cual puede ser justamente el fenómeno que se busca estudiar.

Eliminar automáticamente los outliers podría implicar perder información crucial sobre disparidades y concentraciones que son esenciales para entender la segregación. Sin embargo, también es cierto que si los outliers se deben a errores de medición o registros incorrectos, su inclusión podría sesgar los resultados. Por ello, es recomendable primero investigar el origen y la plausibilidad de estos valores extremos.

No se aconseja eliminar outliers de forma indiscriminada. En su lugar, es preferible mantenerlos para captar con mayor precisión las desigualdades y patrones sociales presentes en los datos. En caso de decidir su exclusión, esta acción debe estar claramente justificada y respaldada por un análisis riguroso. También es útil realizar análisis comparativos, con y sin outliers, para evaluar el impacto de estos valores en los resultados.

En cuanto a las variables relacionadas con puntos importantes, los outliers pueden indicar zonas con una actividad particular o elevada, como barrios con una gran cantidad de comercios o centros culturales. Esta información es valiosa para comprender mejor los patrones de movilidad y segregación urbana.

En resumen, el manejo de los outliers debe hacerse con cautela, priorizando la interpretación social y económica de los datos y evitando la eliminación sistemática que pueda afectar la validez del análisis. Mantener estos valores extremos, siempre que sean plausibles, contribuye a obtener una visión más completa y realista de la segregación y sus determinantes.

Análisis de los outliers en las columnas `pop_flows` y `visitor_flows`

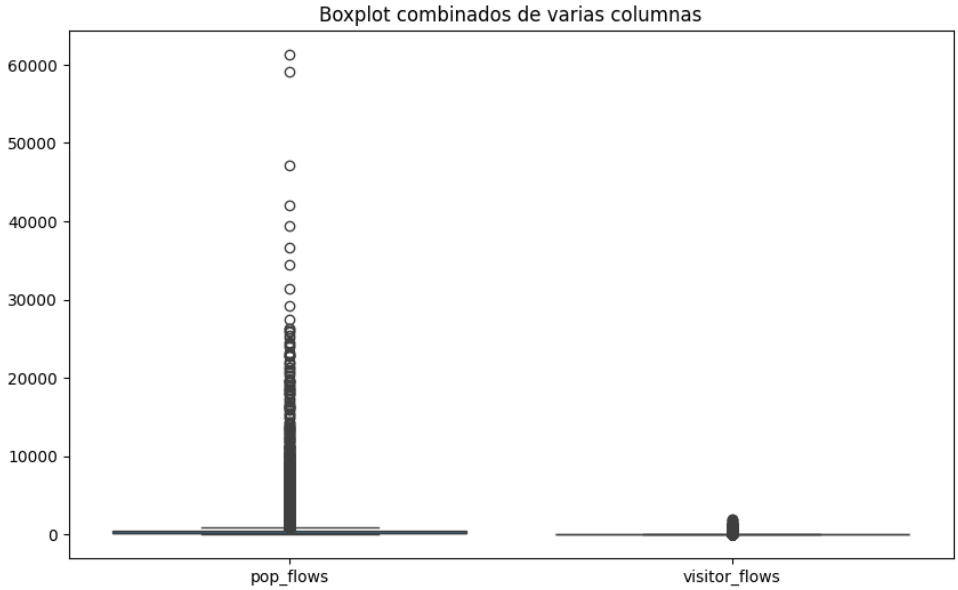

Figura 51: Caption

En el análisis de las columnas `pop_flows` y `visitor_flows` del dataset, se identificaron valores atípicos (outliers) mediante la visualización de boxplots. Los outliers en la columna `pop_flows` muestran valores extremadamente altos en comparación con la mayoría de los datos, lo que podría indicar concentraciones inusuales de personas en determinadas áreas. Estos valores podrían ser representaciones válidas de eventos excepcionales, como grandes concentraciones de población debido a actividades o eventos especiales. Por lo tanto, eliminarlos podría distorsionar la comprensión de estos patrones extremos y perder información clave sobre fenómenos significativos.

Por otro lado, la columna `visitor_flows` también presenta outliers, pero estos son de menor magnitud que los de `pop_flows`. A pesar de su menor impacto en la escala de los datos, se debe considerar si estos outliers corresponden a visitas recurrentes a zonas específicas o si son datos erróneos.

En general, no es necesario eliminar estos outliers de inmediato. Si bien pueden ser datos extremos, es importante evaluar si representan fenómenos legítimos. En caso de que estos valores atípicos sean producto de errores de medición o no representen eventos significativos, su eliminación podría ser considerada. Sin embargo, si los outliers están asociados a eventos válidos y específicos, eliminarlos podría resultar en una pérdida importante de información.

Se recomienda proceder con un análisis adicional que considere el impacto de los outliers sobre los resultados, comparando los análisis con y sin estos valores extremos. Esto permitirá comprender mejor cómo los outliers afectan las conclusiones del estudio y determinar si es adecuado su tratamiento o eliminación.