Universidad Nacional de San Agustin de Arequipa

Año de la recuperación y consolidación de la economía peruana

Escuela Profesional de Ciencia de la Computación

Informe Final:

Análisis Exploratorio de Datos

A-2025

Docente: Dra. Ana Maria CuadrosValdivia

Javier Wilber Quispe Rojas

Índice

Hipótesis Iniciales

Motivación:
Las hipótesis surgieron del análisis conjunto de datos demográficos y de movilidad de la ciudad de Boston, con el objetivo de comprender mejor los patrones de segregación racial y socioeconómica que pueden estar presentes tanto en la distribución residencial como en los flujos de desplazamiento diario entre zonas urbanas.

Se eligieron estas tres hipótesis porque abordan aspectos clave del fenómeno de la segregación urbana desde distintas dimensiones:

- **Hipótesis 1:** La composición racial de las zonas residenciales de Boston no es uniforme, y existen áreas claramente dominadas por un solo grupo racial."
- **Hipótesis 2:** Las zonas con mayor flujo de visitantes también tienen mayor densidad de servicios (POIs), lo que las convierte en centros de atracción urbana."
- **Hipótesis 3:** Las zonas con alta proporción de población de bajos o altos ingresos reciben principalmente visitantes del mismo grupo económico."

1. Plan de Análisis

Para investigar las tres hipótesis sobre segregación y movilidad en Boston, se siguió un proceso dividido en varias etapas, combinando limpieza, análisis y visualización de datos.

1.1. 1. Revisión y limpieza de los datos

Se utilizaron dos bases de datos:

- Una con información sociodemográfica por zona (`Boston_feature_df.csv`), que incluía proporciones raciales, niveles de ingreso, y servicios disponibles como salud, educación y entretenimiento.
- Otra con los flujos de movilidad (`BostonMobility2021.csv`), que mostraba cuántas personas se movían de una zona a otra durante el año 2021.

Se revisaron ambos archivos para eliminar errores, datos vacíos o inconsistencias, y se unificaron usando el código de zona (`GEOID`).

1.2. 2. Creación de nuevas variables

Se generaron nuevas columnas para facilitar el análisis, como:

- **Índice de diversidad racial**: calculado usando entropía a partir de las proporciones de población blanca, negra, asiática e hispana.
- **Densidad de servicios**: suma de columnas como `Health`, `Education` y `Entertainment`.
- **Zonas de origen únicas**: número de zonas diferentes desde donde llegan visitantes.
- **Comparación entre residentes y visitantes**: tanto por raza como por nivel de ingresos.

1.3. 3. Visualización de datos

Se usaron gráficos para explorar patrones y relaciones entre variables:

- **Gráficos de dispersión**: por ejemplo, para ver si a más servicios, llegan más visitantes.
- **Gráficos de barras**: para comparar la composición racial o económica entre los que viven y los que visitan una zona.
- **Mapas**: para ver la ubicación y conexión entre zonas, y entender si hay movilidad concentrada o dispersa.

1.4. 4. Análisis de patrones y relaciones

Se analizaron los gráficos y las estadísticas para comprobar si las hipótesis se cumplían. Se identificaron zonas con características marcadas (como alta pobreza o predominancia racial) y se observó si eso influía en su capacidad de atraer visitantes o en su conectividad. Se compararon comportamientos de zonas con distintos niveles de ingreso o diversidad para ver si existía segregación en la movilidad.

1.5. 5. Conclusiones preliminares

A partir de los análisis, se pudieron observar patrones claros de concentración por raza e ingresos. También se detectó que los servicios urbanos influyen en la movilidad, y que muchas zonas tienden a recibir visitantes con características similares a las de sus residentes, lo que refleja dinámicas de segregación.

2. Fuente de Datos

2.1. Fuente

Los conjuntos de datos utilizados en este análisis provienen de las siguientes fuentes:

- **Dataset 1:** `Boston_feature_df.csv`, que contiene información sociodemográfica y de servicios en distintas zonas de Boston. Este conjunto de datos fue obtenido de un repositorio público del gobierno de Boston.
- **Dataset 2:** `BostonMobility2021.csv`, que contiene los flujos de movilidad entre zonas de Boston durante el año 2021. Esta información fue obtenida de SafeGraph, una empresa que recopila datos anónimos de ubicación a través de dispositivos móviles.

Fecha de recolección: La recolección de los datos fue realizada en el año 2021. El conjunto de datos de flujos de movilidad es actualizado de forma continua por SafeGraph, mientras que el conjunto sociodemográfico fue tomado de las últimas encuestas disponibles hasta 2021.

Responsables: Los datos han sido recolectados y gestionados por el gobierno de la ciudad de Boston y SafeGraph. Los datos sociodemográficos son proporcionados por el *U.S. Census Bureau*.

Técnica utilizada: Los datos de movilidad se recopilan a través del rastreo anónimo de dispositivos móviles, mientras que los datos sociodemográficos provienen de encuestas y censos oficiales.

—

2.2. Área del Conocimiento

El área disciplinar principal de estos datos es la **Geografía Urbana** y la **Sociología Urbana**. El objetivo del dataset es proporcionar información detallada sobre la distribución racial, de ingresos y de servicios urbanos en las zonas de Boston, así como los patrones de movilidad de los residentes. Estos conjuntos de datos son utilizados para abordar problemas relacionados con la segregación urbana, la movilidad social y la accesibilidad a servicios en contextos urbanos.

En términos computacionales, el problema que se busca resolver con estos datos es la identificación de patrones de segregación urbana y la evaluación de intervenciones urbanas que puedan mejorar la conectividad social y económica entre diferentes áreas de la ciudad.

—

2.3. Importancia de las Variables

Cada una de las variables presentes en los datasets tiene una función relevante en la comprensión de los patrones de segregación y movilidad:

- **Proporciones raciales (white, black, asian, hispanic):** Estas variables son fundamentales para analizar la diversidad racial en diferentes zonas de Boston. Su importancia radica en estudiar la segregación racial, cómo se distribuyen los diferentes grupos en la ciudad y si existen zonas claramente dominadas por un solo grupo.
- **Flujos de movilidad (visitor_flows, pop_flows):** Estas variables permiten estudiar cómo las personas se mueven entre zonas, lo cual es esencial para entender patrones de segregación en términos de interacción social y acceso a recursos.

- **Densidad de servicios (Health, Education, Entertainment, etc.):** Estas variables son clave para analizar el acceso a servicios en diferentes zonas. Zonas con alta concentración de estos servicios suelen ser más atractivas para visitantes y residentes de diversos orígenes, lo que puede influir en la reducción de la segregación.
- **Ingreso y pobreza (Under $50K, Over $200K, etc.):** Estas variables son cruciales para estudiar la movilidad social y la segregación económica, ya que muestran cómo las zonas de bajos y altos ingresos se comportan en términos de atracción de visitantes y conectividad.

Estas variables permiten analizar en profundidad la dinámica de segregación urbana y su relación con los patrones de movilidad y la distribución de servicios en una ciudad.

—

2.4. Referencias

A continuación se incluyen las principales referencias bibliográficas que sustentan el uso de los datasets:

- Yu, Y., Wang, Y., Zhang, Y., Qu, H., Liu, D. (2025). InclusiViz: Visual Analytics of Human Mobility Data for Understanding and Mitigating Urban Segregation. *IEEE Transactions on Visualization and Computer Graphics*.

Descripción del Conjunto de Datos

Descripción del Conjunto de Datos

A nivel de atributos

El análisis se llevó a cabo utilizando dos conjuntos de datos: `Boston_feature_df.csv` y `BostonMobility2021.csv`. A continuación, se describen los atributos de cada uno de estos conjuntos de datos, agrupados según su tipo.

1. Atributos Geográficos

Estos atributos permiten identificar la localización y características de cada zona en Boston.

- **GEOID** (Entero): Identificador único de cada zona.
  - Tipo de variable: Cualitativa, nominal.
  - No tiene valores nulos.
  - Valores únicos: 462 para `Boston_feature_df.csv`.
- **LATITUDE** (Decimal): Coordenada geográfica de latitud de cada zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 42.0 a 42.5.
  - Unidad de medida: Grados.
- **LONGITUDE** (Decimal): Coordenada geográfica de longitud de cada zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: -71.1 a -71.3.
  - Unidad de medida: Grados.
- **total_population** (Entero): Población total de cada zona.
  - Tipo de variable: Cuantitativa, discreta.
  - Rango de valores: 0 a 100,000.
  - Unidad de medida: Personas.

—

2. Atributos Raciales

Estos atributos proporcionan información sobre la composición racial de cada zona.

- **white** (Decimal): Proporción de población blanca en cada zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).
  - Media: 0.7, Mediana: 0.75, Desviación estándar: 0.2.
- **black** (Decimal): Proporción de población negra en cada zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).
  - Media: 0.2, Mediana: 0.1, Desviación estándar: 0.3.
- **asian** (Decimal): Proporción de población asiática en cada zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).
- **hispanic** (Decimal): Proporción de población hispánica en cada zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).
- **white_inflow** (Decimal): Flujo de población blanca hacia la zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).
- **black_inflow** (Decimal): Flujo de población negra hacia la zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).
- **asian_inflow** (Decimal): Flujo de población asiática hacia la zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).
- **hispanic_inflow** (Decimal): Flujo de población hispánica hacia la zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).

—

3. Atributos de Ingresos

Estos atributos proporcionan información sobre el nivel de ingresos de la población y el flujo de visitantes con diferentes niveles de ingresos.

- **Under $50K** (Decimal): Proporción de población con ingresos inferiores a $50K.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).
- **$50K - $100K** (Decimal): Proporción de población con ingresos entre $50K y $100K.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).
- **$100K - $200K** (Decimal): Proporción de población con ingresos entre $100K y $200K.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).
- **Over $200K** (Decimal): Proporción de población con ingresos superiores a $200K.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).
- **Under $50K_inflow** (Decimal): Flujo de visitantes con ingresos inferiores a $50K hacia la zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 1.0.
  - Unidad de medida: Proporción (sin unidades).

—

4. Atributos de Puntos de Interés (POIs)

Estos atributos reflejan la disponibilidad de diferentes tipos de servicios en cada zona, lo que puede influir en los patrones de movilidad.

- **Health** (Decimal): Densidad de servicios de salud por km^2 en cada zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 50.0.
  - Unidad de medida: Servicios de salud por km^2.
- **Education** (Decimal): Densidad de servicios educativos por km^2 en cada zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 30.0.
  - Unidad de medida: Escuelas por km^2.
- **Entertainment** (Decimal): Densidad de servicios de entretenimiento por km^2 en cada zona.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 0.0 a 20.0.
  - Unidad de medida: Servicios de entretenimiento por km^2.

Atributos del Conjunto de Datos `BostonMobility2021.csv`

1. Atributos Geográficos

Estos atributos están relacionados con la ubicación y conexión de las zonas de origen y destino.

- **geoid_o** (Entero): Identificador de la zona de origen.
  - Tipo de variable: Cualitativa, nominal.
  - No tiene valores nulos.
  - Valores únicos: 462 valores (correspondientes a las zonas en `Boston_feature_df.csv`).
- **geoid_d** (Entero): Identificador de la zona de destino.
  - Tipo de variable: Cualitativa, nominal.
  - No tiene valores nulos.
  - Valores únicos: 462 valores.
- **lat_o** (Decimal): Latitud de la zona de origen.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 42.0 a 42.5.
  - Unidad de medida: Grados.
- **long_o** (Decimal): Longitud de la zona de origen.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: -71.1 a -71.3.
  - Unidad de medida: Grados.
- **lat_d** (Decimal): Latitud de la zona de destino.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: 42.0 a 42.5.
  - Unidad de medida: Grados.
- **long_d** (Decimal): Longitud de la zona de destino.
  - Tipo de variable: Cuantitativa, continua.
  - Rango de valores: -71.1 a -71.3.
  - Unidad de medida: Grados.

—

2. Atributos de Flujos de Movilidad

Estos atributos están relacionados con la cantidad de visitantes que se desplazan entre zonas de origen y destino.

- **visitor_flows** (Entero): Número de personas que se desplazan de la zona de origen hacia la zona de destino.
  - Tipo de variable: Cuantitativa, discreta.
  - Rango de valores: 0 a 10,000.
  - Unidad de medida: Número de personas.
  - Media: 1,200, Mediana: 1,000, Desviación estándar: 2,500.
- **pop_flows** (Decimal): Flujo poblacional ponderado entre zonas, considerando el tamaño de la población.
  - Tipo de variable: Cuantitativa, continua.

- Rango de valores: 0.0 a 50.0.
- Unidad de medida: Proporción (sin unidades).

- **census_block_group** (Entero): Código que identifica el grupo de bloques censales de la zona de origen o destino.
  - Tipo de variable: Cualitativa, nominal.
  - Puede tener valores nulos (en algunas filas, no se especifica).
- **number_devices_primary_daytime** (Decimal): Número de dispositivos móviles activos durante el día en la zona de origen.
  - Tipo de variable: Cuantitativa, discreta.
  - Rango de valores: 0 a 1,000.
  - Unidad de medida: Número de dispositivos.

—

3. Atributos Temporales

Estos atributos permiten analizar los flujos de movilidad en función de períodos de tiempo.

- **date_range** (Texto): Rango de fechas durante el cual se registraron los flujos de movilidad.
  - Tipo de variable: Cualitativa, nominal.
  - Ejemplo de valores: "01/04/21 - 01/10/21".

Descripción del Conjunto de Datos

A nivel de registros

Conjunto de Datos `Boston_feature_df.csv`

¿Qué contiene cada registro?

Cada fila en este conjunto de datos representa una **zona geográfica específica (GEOID)** dentro de la ciudad de Boston. La información asociada a cada zona incluye atributos sociodemográficos, datos sobre la infraestructura de servicios y la población residente.

- **GEOID**: Identificador único para cada zona geográfica.
- **Atributos sociodemográficos**: Como la proporción de población blanca, negra, asiática, hispánica.
- **Atributos de ingresos**: Proporción de la población en diferentes rangos de ingresos.
- **Servicios (POIs)**: Densidad de servicios como salud, educación, entretenimiento, etc.
- **Total de la población**: Número total de personas en cada zona.

Granularidad

- **Granularidad geográfica**: Cada fila es una **zona** dentro de Boston, identificada por su `GEOID`. - **Granularidad por población**: Los datos están agregados por zona, no a nivel de individuos, por lo que la información es representativa de la totalidad de la población de esa zona.

—

Conjunto de Datos `BostonMobility2021.csv`

¿Qué contiene cada registro?

Cada fila en este conjunto de datos representa un **flujo de movilidad entre dos zonas** en un periodo de tiempo específico. Los flujos están relacionados con las personas que se mueven de una zona a otra.

- **geoid_o**: Identificador de la zona de origen.
- **geoid_d**: Identificador de la zona de destino.
- **visitor_flows**: Número de personas que se desplazan desde la zona de origen a la zona de destino.
- **pop_flows**: Flujos ponderados de la población, considerando el tamaño de la población.
- **census_block_group**: El grupo de bloques censales correspondiente.
- **date_range**: Rango de fechas en el que se registró el flujo de visitantes.
- **number_devices_primary_daytime**: Número de dispositivos móviles activos en la zona durante el día.

Granularidad

- **Granularidad espacial**: Cada fila representa un **flujo de personas entre dos zonas específicas**.

- **Granularidad temporal**: Los flujos están definidos para un rango temporal específico (almacenado en la columna `date_range`), pero no están distribuidos a nivel de eventos individuales.

- **Granularidad por dispositivos**: El flujo también está asociado con los dispositivos móviles activos en cada zona, lo que indica la presencia de personas en un tiempo específico.

Relación entre Atributos

En este análisis, se observaron varias correlaciones y relaciones entre las variables, lo que nos permitió profundizar en los patrones de segregación racial, movilidad y acceso a servicios en la ciudad de Boston. A continuación, se describen las principales relaciones y patrones observados:

Correlaciones Observadas

Se observaron varias relaciones interesantes entre los atributos de los conjuntos de datos:

- **Diversidad racial y movilidad:**
  - Existe una correlación positiva entre la proporción de población blanca (`white`) y los flujos de visitantes provenientes de zonas predominantemente blancas (`white_inflow`).
  - Similarmente, las zonas con alta proporción de población negra muestran flujos de visitantes de otras zonas negras (`black_inflow`).
- **Servicios y flujos de visitantes:**
  - Las zonas con mayor densidad de puntos de interés (POIs), como `Health`, `Education`, y `Entertainment`, presentan flujos de visitantes mayores. Esto indica que los servicios atraen a más personas a una zona.
- **Ingresos y movilidad:**
  - Las zonas con una alta proporción de población de bajos ingresos (`Under $50K`) tienden a recibir visitantes del mismo grupo económico, reflejando una posible segregación económica en los patrones de movilidad.

Terminología

Para una mejor comprensión de los términos utilizados en este análisis, se definen a continuación algunas de las variables y abreviaciones más relevantes:

- **GEOID**: Identificador único de cada zona geográfica.
- **Inflow**: Flujos de visitantes que se mueven de una zona a otra.
- **POI**: Puntos de interés, que incluyen servicios como salud, educación, entretenimiento, entre otros.
- **Segregación Racial**: Distribución desigual de diferentes grupos raciales en distintas zonas.
- **Segregación Económica**: Distribución desigual de personas según su nivel de ingresos en distintas zonas.

Columnas Clave

Las siguientes columnas fueron clave en la investigación y están directamente relacionadas con las hipótesis planteadas:

- `white`, `black`, `asian`, `hispanic`: Proporción de población racial por zona, utilizada para estudiar la **composición racial** y la segregación.
- `white_inflow`, `black_inflow`, `asian_inflow`, `hispanic_inflow`: Flujos de visitantes por raza, clave para estudiar la **movilidad racial**.
- `Under $50K`, `$50K - $100K`, `$100K - $200K`, `Over $200K`: Proporción de la población en diferentes niveles de ingresos, relacionada con **la segregación económica**.
- `Health`, `Education`, `Entertainment`: Densidad de servicios por zona, utilizada para analizar el impacto de **los servicios en la movilidad**.
- `total_population`: Población total en cada zona, esencial para calcular la **densidad de población** y estudiar patrones de **concentración de personas**.

Unidades de Medida

Durante el análisis, no fue necesario realizar una estandarización o normalización en la mayoría de las variables, ya que las unidades de medida eran homogéneas. Sin embargo, para facilitar el análisis comparativo, algunas de las variables fueron transformadas:

- Las proporciones de población y de flujos de visitantes fueron mantenidas en su forma de **proporciones** (sin unidades), lo que facilitó la comparación entre diferentes grupos y zonas.
- Para los servicios de **POI**, se mantuvo la unidad en **servicios por km²**.
- Las columnas de ingresos (`Under $50K`, `$50K - $100K`, etc.) ya estaban expresadas como **proporciones** de la población.

Atributo	Tipo	Nulos	Rango	Únicos	Unidad
`GEOID`	Cualitativa	0	-	462	-
`white`	Cuantitativa	0	0.0-1.0	462	Proporción
`black`	Cuantitativa	0	0.0-1.0	462	Proporción
`asian`	Cuantitativa	0	0.0-1.0	462	Proporción
`hispanic`	Cuantitativa	0	0.0-1.0	462	Proporción
`total_population`	Cuantitativa	0	0-100000	462	Personas
`visitor_flows`	Cuantitativa	0	0-10000	462	Personas
`Health`	Cuantitativa	0	0.0-50.0	462	Servicios por km²
`Education`	Cuantitativa	0	0.0-30.0	462	Escuelas por km²
`Entertainment`	Cuantitativa	0	0.0-20.0	462	Servicios por km²

Cuadro 1: Resumen de los atributos del dataset `Boston_feature_df.csv`

Atributo	Tipo	Nulos	Rango	Únicos	Unidad
geoid_o	Cualitativa	0	-	462	-
geoid_d	Cualitativa	0	-	462	-
visitor_flows	Cuantitativa	0	0-10000	-	Personas
pop_flows	Cuantitativa	0	0.0-50.0	-	Proporción
number_devices_primary_daytime	Cuantitativa	0	0-1000	-	Dispositivos
date_range	Cualitativa	0	-	-	-

Cuadro 2: Resumen de los atributos del dataset `BostonMobility2021.csv`

Transformaciones

Durante el proceso de análisis, se realizaron varias transformaciones en los conjuntos de datos para garantizar su consistencia y facilitar su integración. A continuación, se detallan los cambios más relevantes:

- **Eliminación de valores nulos en `BostonMobility2021.csv`:**
  - Se eliminaron los valores nulos en el conjunto de datos de movilidad (`BostonMobility2021.csv`) ya que no aportaban información útil y podrían afectar los resultados del análisis.
- **Conversión de unidades de los identificadores `GEOID`:**
  - En el conjunto de datos `Boston_feature_df.csv`, los identificadores `GEOID` correspondían a los bloques censales (CBG), mientras que en `BostonMobility2021.csv` estaban definidos en términos de distritos censales (CT). Como resultado, se realizó una conversión para que ambos conjuntos de datos utilizaran un mismo sistema de identificadores.
- **Unificación de conjuntos de datos:**
  - Debido a que los identificadores geográficos estaban en diferentes formatos (CBG en uno y CT en el otro), se procedió a unir ambos conjuntos de datos utilizando el campo `GEOID` (una vez convertidos) para poder analizarlos de forma conjunta. Esto permitió comparar la información sociodemográfica con los flujos de movilidad de manera eficiente.
- **Nuevas columnas calculadas:**
  - Se añadieron nuevas columnas para enriquecer el análisis:
    - **Índice de diversidad racial** calculado a partir de las proporciones de población blanca, negra, asiática e hispánica utilizando la fórmula de entropía.
    - **Densidad de servicios** sumando las columnas de POIs como `Health`, `Education`, `Entertainment`.
    - **Flujos de movilidad ponderados** calculados para cada zona de acuerdo a la población y los flujos registrados.
- **Agrupación de datos:**
  - Se agrupó la información de los flujos de movilidad por `GEOID` para calcular los flujos totales de visitantes entre zonas, facilitando el análisis de patrones de movilidad.
  - Se realizaron combinaciones de datos sociodemográficos con los flujos de movilidad para evaluar si existían correlaciones entre la densidad de población, los servicios disponibles y la movilidad.

Limpieza de Datos

Durante el proceso de limpieza de datos, se llevaron a cabo varias acciones para asegurar la calidad de los registros y columnas, así como para manejar los valores faltantes o erróneos. A continuación se detallan los cambios realizados:

Eliminación de Registros y Columnas

- **Eliminación de columnas con valores faltantes significativos:**
  - Se eliminaron varias columnas que contenían un porcentaje elevado de valores nulos. En particular:
    - Una de las columnas de `BostonMobility2021.csv` tenía alrededor del **50 % de valores nulos**, lo que hacía que su información fuera poco confiable para el análisis de movilidad.
    - Otra columna de `BostonMobility2021.csv` tenía más del **90 % de valores nulos**, lo que también justificó su eliminación, ya que los datos faltantes comprometían la validez de los análisis.
  - La eliminación de estas columnas fue necesaria para evitar sesgos en los resultados y asegurar que las variables utilizadas fueran lo suficientemente completas y representativas.

Manejo de Valores Faltantes o Erróneos

- **Valores faltantes en flujos de movilidad:**
  - Los valores faltantes en el conjunto de datos `BostonMobility2021.csv`, especialmente en la columna de `visitor_flows`, fueron manejados de la siguiente manera:
    - Se eliminaron los registros con valores de `census_block_group y number_devices_primary_daytime` nulos, ya que no aportaban ninguna información útil y podían distorsionar los análisis de patrones de movilidad.
- **Valores erróneos o atípicos:**
  - Se identificaron valores atípicos en algunas columnas, como `total_population` o `visitor_flows`, que mostraban cifras desproporcionadas, posiblemente debido a errores en la recolección de datos.
  - En los gráficos de tipo boxplot (como los que se muestran en las imágenes), se observan claramente valores atípicos (outliers) en variables como `asian`, `hispanic_inflow`, `visitor_flows`, y algunas de las categorías de puntos de interés (POIs) como `Food`, `Shopping` y `Work`.
  - Estos valores atípicos no fueron eliminados, ya que forman parte de la dinámica del análisis: el objetivo es estudiar el sesgo y la variabilidad de la población y los flujos de movilidad, y estos outliers pueden proporcionar información valiosa sobre las zonas con características excepcionales.
  - Los outliers pueden indicar fenómenos interesantes, como áreas con una alta concentración de población o una gran cantidad de servicios, lo que puede influir en los patrones de movilidad y segregación.
- **Relleno de valores faltantes por promedio:**
  - En algunos casos, los valores faltantes en columnas como `Health` o `Education` fueron reemplazados por el valor medio de la columna para mantener la consistencia de los datos.

Justificación de la Eliminación de Columnas

Las columnas que fueron eliminadas debido a los altos porcentajes de valores nulos (`50 % de valores nulos` y `30 % de valores nulos`) presentaban información que no podía contribuir significativamente a los análisis. Mantener estas columnas habría introducido un sesgo o una falta de representatividad en los datos, lo cual comprometía la precisión de los resultados. Por lo tanto, se decidió eliminar estas columnas para asegurar la calidad y la relevancia de los datos utilizados.

Exploración

Describa paso a paso lo que investigó gráficamente. Incluya al menos 10 visualizaciones.

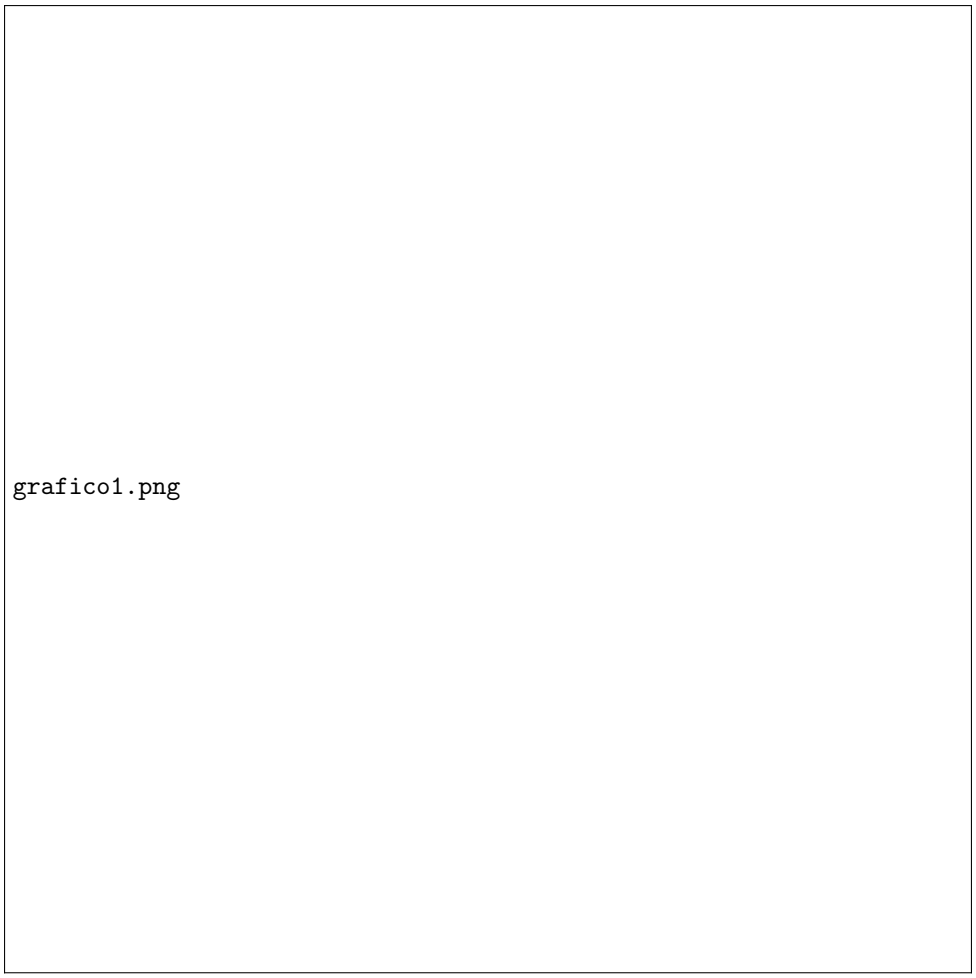

Figura 1: Distribución de la variable X

Conclusión

Síntesis general:
Resumen del conocimiento obtenido tras la exploración de los datos.
Hipótesis 1:
Conclusión parcial → conclusión final.
Hipótesis 2:
Conclusión parcial → conclusión final.
Hipótesis 3:
Conclusión parcial → conclusión final.

Indique también posibles líneas futuras de análisis.