DESCRIPCIÓN ANÁLISIS

Samuel Huertas*
* Autor correspondiente: samhuro3@gmial.com

Exploración de Datos:

A continuación se muestra la exploración realizada a las tres bases de datos. Vale aclarar que las bases de datos son archivos .csv los cuales van a ser procesados con python y con la ayuda de la librería de pandas se va a realizar la respectiva exploración. Las bases de datos van a constar de:

1. ***exa_barrios_cali*:** esta base de datos tiene almacenada una lista de los barrios de Cali con su respectivo código o id. Tiene un total de 337 registros, sin valores nulos y duplicados, presenta dos columnas:
    a. id_barrio: está columna almacena el número de identificación del barrio, tipo de dato (int64). Esta columna inicialmente presenta el nombre de 'codigo' pero se cambio a 'id_barrio' para realizar una posterior unión con otras bases de datos.
    b. nombre: está columna almacena el nombre que tiene el barrio, tipo de dato (object, string)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 337 entries, 0 to 336
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   id_barrio  337 non-null    int64 
 1   nombre     337 non-null    object
dtypes: int64(1), object(1)
memory usage: 5.4+ KB
```

Figura 1. *Resumen de las columnas de la base de datos* ***exa_barrios_cali.***

2. ***exa_dispositivos_cali*:** está base de datos cuenta con información sobre los dispositivos que cuenta el banco para realizar transacciones. Está base de datos cuenta con un total de 37284 registros, no tiene valores nulos ni duplicados. Presenta las siguientes columnas:

    a. tipo: almacena los canales físicos presentes en la ciudad de Cali (POS: establecimiento-POS, PAC: PAC, CB: corresponsales bancarias, SAI: sucursales, DISPENSADOR: cajero dispensador y MF: cajero multifuncional)

    b. cod_dispositivo: se almacena el número de identificación de los dispositivos presentes en la ciudad de Cali, tipo de dato (int64). Esta columna inicialmente se llamaba 'codigo' pero se reemplazó por 'cod_dispositivo' para realizar una posterior unión con otras bases de datos. Los números de identificación de los dispositivos se encuentran en el rango de [18, 6921701174]

    c. latitud: almacena las coordenadas de latitud del dispositivo

    d. longitud: almacena las coordenadas de longitud del dispositivo

e. id_barrio: está columna almacena el número de identificación del barrio, tipo de dato (int64).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37284 entries, 0 to 37283
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   tipo             37284 non-null  object
 1   cod_dispositivo  37284 non-null  object
 2   latitud          37284 non-null  float64
 3   longitud         37284 non-null  float64
 4   id_barrio        37284 non-null  object
dtypes: float64(2), object(3)
memory usage: 1.4+ MB
```

Figura 2. *Resumen de las columnas de la base de datos* ***exa_dispositivos_cali.***

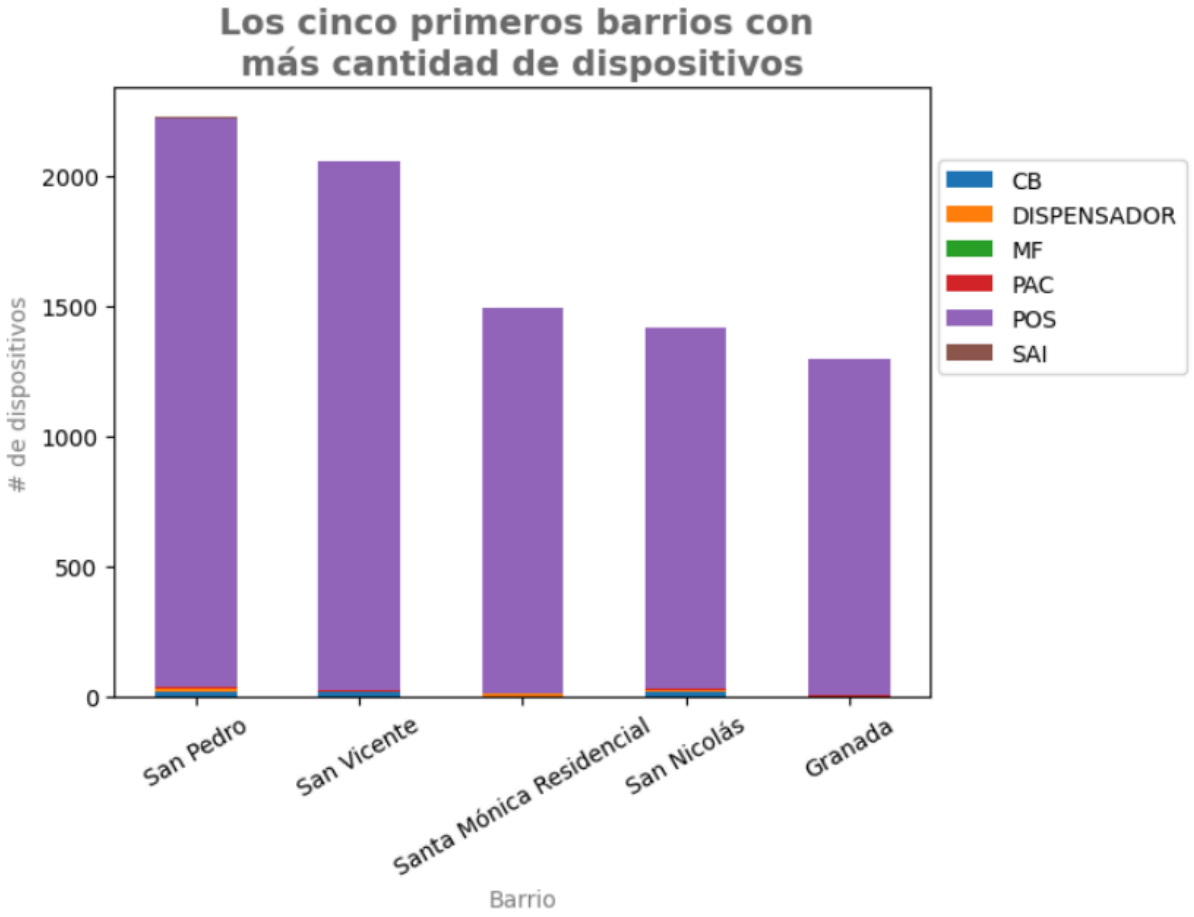

Figura 3. *Muestra los cinco barrios con más dispositivos y el tipo de dispositivo, vemos que la mayoría de los tipos de dispositivos en los barrios con más dispositivos son 'POS'.*

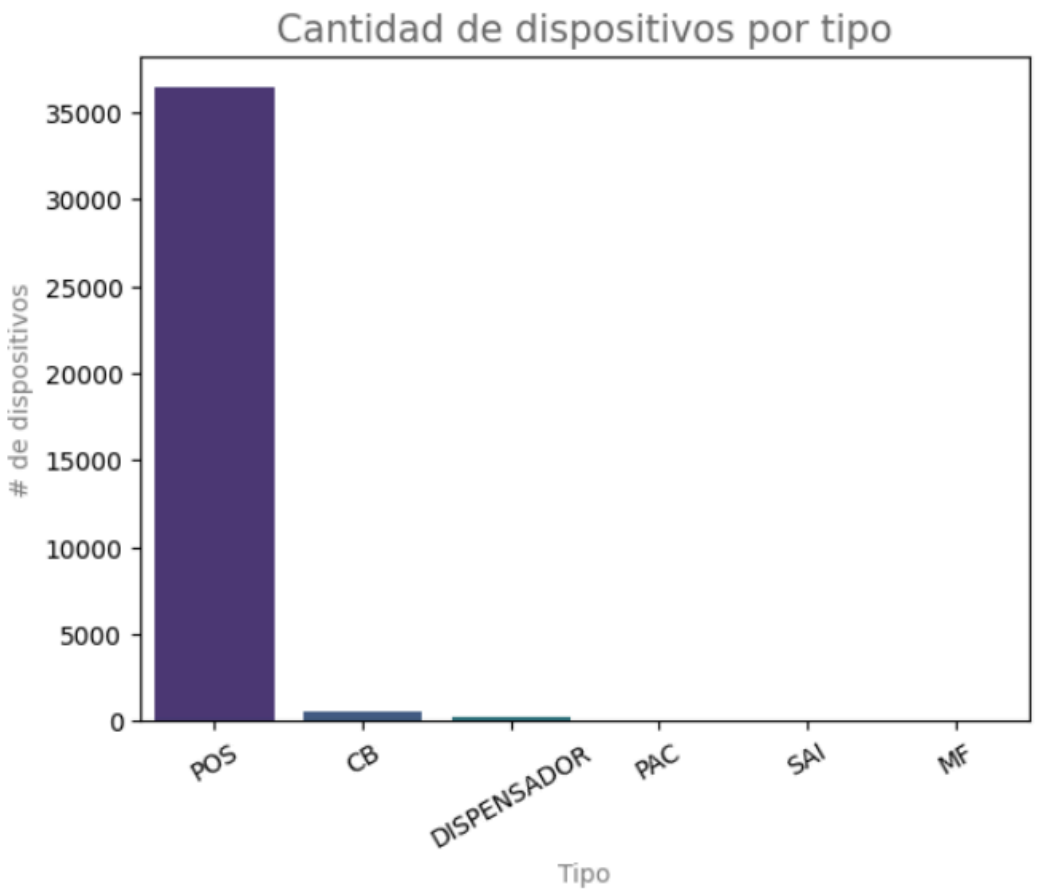

Figura 4. *Muestra la cantidad de tipos de dispositivos en la ciudad de Cali, se puede observar que la mayoría de los tipos de dispositivos son 'POS'.*

3. ***exa_trx_clientes*:** contiene la información de las transacciones realizadas por cada uno de los clientes de la ciudad de Cali, cuenta con un total de 93446 registros, no tiene valores nulos ni duplicados. Presenta las siguientes columnas:

    a. num_doc: en está columna se encuentra almacenado el número de documento del cliente que realizó la transacción.

    b. tipo_doc: cual es el tipo de documento del cliente (1: ciudadanos nacionales, 2: extranjeros, 3: empresas, 4 y 9: entre otros)

    c. canal: almacena los canales físicos presentes en la ciudad de Cali (POS: establecimiento-POS, PAC: PAC, CB: corresponsales bancarias, SAI: sucursales, DISPENSADOR: cajero dispensador y MF: cajero multifuncional)

    d. cod_dispositivo: se almacena el número de identificación de los dispositivos presentes en la ciudad de Cali, tipo de dato (int64).

    e. num_trx: el número de transacciones realizadas por el cliente

    f. mnt_total_trx: el monto total de las transacciones.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 93446 entries, 0 to 93445
Data columns (total 6 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   num_doc          93446 non-null  int64
 1   tipo_doc         93446 non-null  int64
 2   canal            93446 non-null  object
 3   cod_dispositivo  93446 non-null  int64
 4   num_trx          93446 non-null  int64
 5   mnt_total_trx    93446 non-null  float64
dtypes: float64(1), int64(4), object(1)
memory usage: 4.3+ MB
```

Figura 5. *Resumen de las columnas de la base de datos* ***exa_trx_clientes.***

Llaves de cruce:

Como se mencionó anteriormente se realizaron unos cambios a algunas de las columnas de las base de datos, esto se efectuó para cruzar entre ellas las bases de datos, estas llaves se mencionan a continuación:

1. id_barrio: está llave nos permite realizar el cruce entre ***exa_dispositivos_cali*** y ***exa_barrios_cali,*** lo cual nos da a conocer el nombre del barrio en el cual se encuentra el dispositivo.
2. cod_dispositivo: nos permite relacionar la base de datos ***exa_trx_clientes*** con ***exa_dispositivos_cali*** permitiendo conocer en qué dispositivo se realizó la transacción.

Transformación de columnas:

Las columnas a las cuales se les realizó una transformación se nombran a continuación:

- num_doc
- tipo_doc
- cod_dispositivo
- id_barrio

Está transformación que se efectuó en las columnas se realizó debido a que estas se utilizan para identificar una persona, un dispositivo o un barrio, por lo cual se necesita que estas sean de tipo string y no int64.

Otra transformación realizada es a las columnas de longitud y latitud de las coordenadas de cada uno de los dispositivos, ya que para poder graficar los puntos en un plano y que sea fiel a la distribución que tienen en un mapa se debe de realizar dicha transformación. Para ejecutar dicha transformación se utilizó la libraria pyproj y el módulo Transformer. Con esta transformación se logró obtener la siguiente gráfica.

Figura 6. *Distribución de los dispositivos en la ciudad de Cali.*

Análisis Ejercicios:

a) Primera Pregunta

Para dar respuesta a esta pregunta se siguieron los siguientes pasos:

1. Unir las bases de datos para identificar en qué barrio se realizaron las transacciones. Se procedió a eliminar las columnas que no aportan información (las columnas de coordenadas, tipo de canal y tipo de documento)
2. Agrupar los datos por cliente y barrio para calcular el dinero total transado por cada cliente en cada barrio y para cada grupo de cliente y barrio, sumar el dinero total transado por el cliente.
3. Calcular el porcentaje del dinero total transado por el cliente en cada barrio con respecto al total transado por el cliente en todos los barrios.
4. Identificar los barrios donde al menos el 51% del dinero total transado por cada cliente está concentrado.

```
San Vicente                    575
San Pedro                      368
Unicentro Cali                 359
Urbanización San Juaquín       288
El Sena                        283
Santa Mónica Residencial       235
El Cedro                       232
Lili                           230
Urbanización Ciudad Jardín     220
Chipichape                     208
Name: nombre, dtype: int64
```

Figura 7. *Los 10 primeros barrios en los cuales al menos el 51% del dinero total es transado por cada cliente está concentrado.*

b) Segunda Pregunta

Se comenzó a hacer una filtración a la base de datos ***exa_trx_clientes*** para solo obtener las transacciones realizadas en dispositivos tipo POS. Luego se realiza una unión entre ***exa_barrios_cali*** y ***exa_dispositivos_cali*** para saber a cual barrio pertenecía cada dispositivo, a esta unión se procedió a unirla con la filtración realizada a ***exa_trx_clientes*** para saber el barrio en donde se realizó la transacción con dispositivos tipo POS. Se crea una nueva columna para identificar los clientes que solo compran en un solo lugar, para esto se utilizaron las columnas que tienen los documentos de los clientes y el barrio donde se realizó la transacción, posteriormente se realizó una lista en donde se almacenaron estos clientes únicos y se realizó una filtración teniendo en cuenta esta lista. Seguido a esto se realizó la agrupación teniendo en cuenta el barrio y el documento del cliente. Se encontraron los cinco barrios con más transacciones en Cali y se filtró la agrupación anteriormente mencionada con los nombres de estos cinco barrios.

	nombre_barrio	num_transacciones	num_clientes
0	Parcelaciones Pance	1027.0	235
1	San Vicente	847.0	299
2	Santa Mónica Residencial	1200.0	321
3	Urbanización Ciudad Jardín	1175.0	254
4	Urbanización San Juaquín	1027.0	373

Figura 8. *Los cinco barrios con mayor cantidad de clientes únicos que realizan transacciones en dispositivos tipo POS.*

c) Tercera Pregunta

Para dar solución al tercer punto de la prueba técnica se realizó una unión entre la base de datos ***exa_barrios_cali*** y ***exa_dispositivos_cali*** para saber a cual barrio pertenecía cada dispositivo, posteriormente se realizó un filtrado para buscar sólo los dispositivos localizados en el barrio ‘Caney’, se siguió con el mapeo de la longitud y la latitud a coordenadas cartesianas, esto se logró haciendo uso de la proyección de Mercator. Una vez finalizado el mapeo se realizó un ciclo for para encontrar la distancia euclidiana de un dispositivo a cada uno de los otros dispositivos, y para finalizar se organizó

el Data Frame resultante de manera descendente para que los datos con la distancia euclidiana más grande quedaran de primeros.

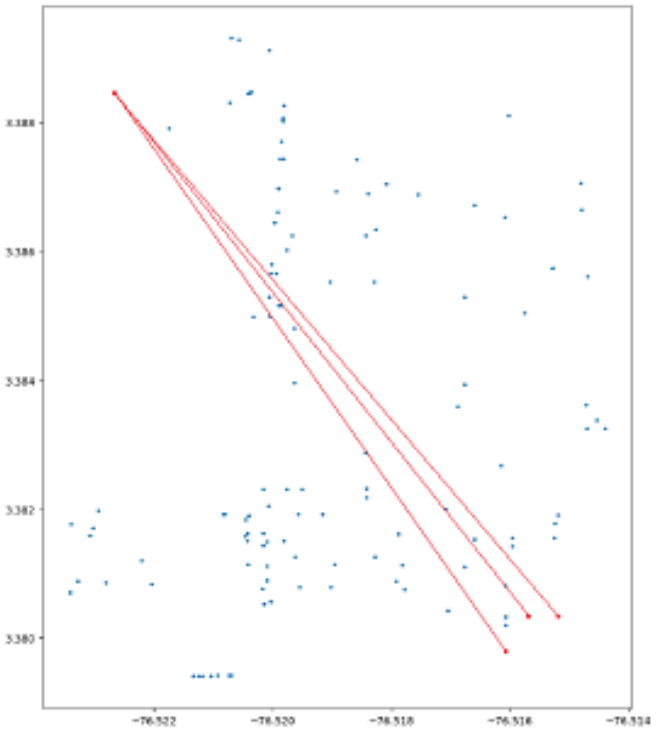

Figura 9. *Los dispositivos más alejados entre sí, que pertenecen al barrio Caney.*