Prueba Técnica Ingeniero de Datos

Estephania Arenas Marulanda

Diagnóstico de la información

En primer lugar, se analiza la información que contiene cada una de las tablas compartidas, para encontrar la relación entre ellas y la mejor manera de abordar el ejercicio.

Se encuentra que la relación entre las tablas es la siguiente:

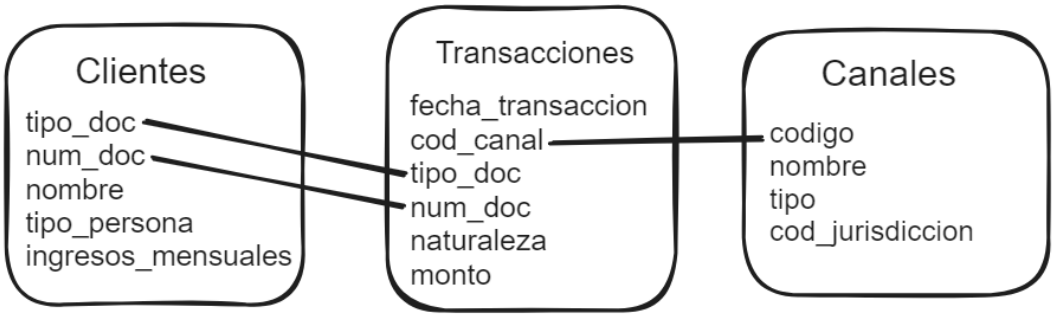

Figura 1. Llaves de cruce de la información inicial

Para lograr que se crucen entre ellas los campos cod_canal de la tabla transacciones y código de la tabla canales se deben llevar al mismo formato. Además, se encuentra en la tabla clientes, la columna num_doc se encuentra con algunos signos negativos, por lo tanto, se asumirá que el numero se encuentra correcto y lo erróneo es el símbolo, el cual se quitará ya que no hay números de documento negativos.

Para la tabla de clientes se eliminan registros duplicados por tipo_doc y num_doc y se eliminan datos que no se encuentren con la totalidad de las columnas llenas

Luego de conocer esta información, se llevan cada una de las tablas compartidas a Python para comenzar con el tratamiento de la información.

La prueba se trabajará en jupyter notebook para tener una ejecución paso a paso de la información, allí se detallará lo realizado en el código.

Análisis exploratorio de datos

En el archivo de Python ***Solucion.ipynb*** en primer lugar, se realiza una exploración de los datos con la información anteriormente mencionada. Se realiza una unión de tablas por medio de la función merge con las llaves de cruce encontradas en la Figura 1; Todo esto con el fin de generar una tabla total que se usará para la solución de la prueba.

En la limpieza de datos se eliminan los clientes que tengan información incompleta como el tipo_doc con '-', de igual modo en la tabla de transacciones.

En la segunda etapa de la solución, se separa la pregunta en dos partes:

1. Clientes que han realizado transacciones en los últimos 6 meses por un monto total superior al 200% de sus ingresos mensuales.
2. Clientes superiores al percentil 95 del total de la población por tipo de persona.

Se genera una tabla con cada una de las poblaciones anteriores, para luego cruzarlas y obtener los clientes que cumplen con las dos condiciones.

Construcción de la metadata

La metadata de las tablas con mas relevancia dentro del proyecto se encuentra alojada en el archivo ***metadata.md*** estas fueron las tablas que se incluyeron

1. **tabla_total:** Tabla que contiene el cruce entre los 3 archivos csv entregados
2. **trx_ult_6_meses:** Tabla que contiene las transacciones de salida de los últimos 6 meses
3. **cliente_exceden:** Contiene clientes que exceden por 200% o más sus ingresos mensuales en el total de transacciones durante los últimos 6 meses
4. **clientes_superan_percentil**: Contiene clientes que superan el percentil 95 de las transacciones realizadas del total de la población por tipo de persona
5. **trx_porcentil**: Contiene clientes que superan el percentil 95 y los clientes que superan por el 200% sus ingresos en transacciones de salida. Esta es la audiencia solicitada

Visualización Power Bi

A continuación, se describirá lo realizado para la visualización del proyecto en power bi. En este caso se usó como insumo la tabla total, para dar un reporte detallado de los movimientos de los clientes por tipo de persona (Natural, Jurídica), tener en cuenta las entradas y salidas de dinero y los movimientos de los clientes a través del tiempo. Además, con la información del código de jurisdicción y *usando información obtenida en bases de datos públicas del DANE*, se pudo visualizar la información georreferenciada por cada cliente, mostrando los lugares en donde se realizaron movimientos y el canal en el que se realizaron.

Además, se implementó un filtro con el que se puede dividir la población visualizada en 2 grandes grupos: los afiliados que cumplen la condición del 200% y el percentil 95 y los afiliados que no cumplen, este filtro permite entender a mayor profundidad el comportamiento que tienen estas dos poblaciones.

Todo esto queda en el archivo ***visualizacion.pbix***

Tratamiento de datos SQL

Se realiza el mismo procedimiento de Python, pero en sql para la generación de la información solicitada. Se trabaja con la librería sparky, usada a diario y extensivamente en mis actividades como analista, para la carga de las tablas a la lz y realizar el tratamiento de estas en entorno de impala, así:

```
# uso de la librería sparky para cargue de los archivos a impala, para el tratamiento de los datos en sql
```

```
from sparky_bc import Sparky
#inicializar sparky
sp = Sparky(username = "esarenas",dsn = 'impala-virtual-prd',hostname =
"sbmdeblze004.bancolombia.corp")
#cargar df a la lz
sp.subir_df(clientes_df, 'proceso.clientes_df_eam')
sp.subir_df(transacciones_df, 'proceso.transacciones_df_eam')
sp.subir_df(canales_df, 'proceso.canales_df_eam')
```

Los querys usados para el tratamiento y la extracción de la información por medio de sql están consignados en el archivo ***solucion.sql***

La salida generada por el sql se encuentra en el archivo *trx_porcentil_eam.csv*

Conclusiones

La extracción y análisis de información a partir de las tablas proporcionadas permitió identificar comportamientos de clientes con alto gasto, como aquellos que superan el 200% de sus ingresos en transacciones, lo cual puede ser un indicativo de riesgo financiero. Además, al separar los datos por tipo de persona, se observó que existen diferencias significativas en los patrones de gasto entre una persona natural y jurídica, lo que puede influir en decisiones estratégicas de marketing, servicio al cliente y oferta en general.

El análisis también facilitó la detección de tendencias de gasto a lo largo del tiempo y el uso de diferentes canales de transacción. Esta información es crucial para optimizar recursos y mejorar la experiencia del cliente. A su vez, se puede utilizar para identificar irregularidades que podrían sugerir fraude, permitiendo a la empresa gestionar mejor el riesgo.

La información obtenida brinda una base sólida para tomar decisiones informadas que mejoren la estrategia comercial, la atención al cliente y la gestión del riesgo.