Prueba Técnica Analítico/a de Ingeniería de Datos

Proceso de desarrollo:

- Inicialmente se extrajeron las bases de datos almacenadas en archivos CSV y se cargaron en un notebook para conocer qué tipo de data almacenaban, se observo la cantidad de registros, el tipo de dato de cada columna y si había presencia de valores faltantes para las tres bases, se eliminaron los valores duplicados para las bases que aplicaba.
- Luego se realizo un conteo de los diferentes valores de cada columna categórica de cada base, los numero de documento únicos y los numero de canales únicos, para las variables continuas se calcula la estadística descriptiva.
- Se sacaron las primeras conclusiones:

  Base de datos clientes:
  La dimensión de la base de datos de clientes discrepa con la cantidad de números de documentos únicos de los clientes, esto probablemente se da debido a que hay clientes que comparten el mismo número de documento pero tienen un tipo de documento diferente o porque hay un mismo cliente duplicado.
  Sin embargo, hay 1925 registros que se desconoce el tipo de documento ya que tiene el valor de "-", también hay 1532 registros dentro del campo tipo_persona con el valor "-" y 300 registros con el valor de "NATURAL". En el estado actual de la base no se detectan registros duplicados y no se ha identificado el número total de clientes únicos.
  Base de datos canales:
  La dimensión de la base de datos de canales coincide con la cantidad de códigos únicos de los canales (32974 canales únicos, 32974 registros), de momento no se encontraron, inconsistencia en los registros de la tabla.
  Base de datos transacciones:
  Es la base de datos con mayor número de registros, de momento no se encontraron errores en los registros de las columnas fecha_transaccion, cod_canal,num_doc, naturaleza y monto, solo se evidencio que en la columna tipo_doc hay 46138 registros con el valor de "-".
- Se subieron las bases a la lz utilizando una función de sparky , se realizaron las validaciones para garantizar que la data haya subido de la manera adecuada, luego se realizaron consultas para identificar el manejo adecuado de los registros con valores en "_".
  - Inicialmente en la tabla de clientes se contaron cuantos tipos de documentos tenían cada numero de documento, se identificaron que había varios numeros de documentos con hasta 4 clientes diferentes.
  - Se realizo el mismo ejercicio anterior pero para la base transaccional y en la revisión se encontró que los num_doc tenían hasta dos tipos de documento, pero había muchos de estos que el otro tipo de documento era el valor "-" .
  - Se sumaron el monto total transado por mes por diferente numero de documento y tipo de documento y en el resultado se encontró que se estaban duplicando transacciones a aquellos clientes que tenían dos tipo de

documento y que uno de esos tipos tenía como valor "-" (tabla con resultado: proceso_cumplimiento.validar_repetidos_trx_pa)

TIPO_DOC	NUM_DOC	FECHA	SUMA_MES	TIPO_DOC	NUM_DOC	FECHA	SUMA_MES
-	-8.1829E+18	1/01/2024	2000000	CEDULA DE CIUDADANIA	- 8,182,893,182,090,760,000	1/01/2024	2000000
-	-8.1829E+18	1/03/2024	800000	CEDULA DE CIUDADANIA	- 8,182,893,182,090,760,000	1/03/2024	800000
-	-8.1829E+18	1/05/2024	110000	CEDULA DE CIUDADANIA	- 8,182,893,182,090,760,000	1/05/2024	110000
-	-8.1829E+18	1/06/2024	1000000	CEDULA DE CIUDADANIA	- 8,182,893,182,090,760,000	1/06/2024	1000000
-	-8.1829E+18	1/07/2024	800000	CEDULA DE CIUDADANIA	- 8,182,893,182,090,760,000	1/07/2024	800000
-	-8.1829E+18	1/09/2024	2000000	CEDULA DE CIUDADANIA	- 8,182,893,182,090,760,000	1/09/2024	2000000
-	-8.1829E+18	1/10/2024	600000	CEDULA DE CIUDADANIA	- 8,182,893,182,090,760,000	1/10/2024	600000

Resultado - proceso_cumplimiento.validar_repetidos_trx_pa :

num_doc	id	mes	periodo	movimiento	mov	cod_canal	id_canal	naturaleza	e_s	tipo_doc	doc_tipo
-3.969E+18	-3.969E+18	1/09/2024	1/09/2024	4930000	4930000	2013946	2013946	ENTRADA	ENTRADA	CEDULA DE CIUDADANIA	-
-3.969E+18	-3.969E+18	1/09/2024	1/09/2024	4930000	4930000	2013946	2013946	ENTRADA	ENTRADA	CEDULA DE CIUDADANIA	-
-3.969E+18	-3.969E+18	1/09/2024	1/09/2024	4930000	4930000	2013946	2013946	ENTRADA	ENTRADA	CEDULA DE CIUDADANIA	-
-3.969E+18	-3.969E+18	1/09/2024	1/09/2024	4930000	4930000	2013946	2013946	ENTRADA	ENTRADA	CEDULA DE CIUDADANIA	-
6.474E+18	6.474E+18	1/07/2024	1/07/2024	600000	600000	3003148	3003148	SALIDA	SALIDA	CEDULA DE CIUDADANIA	-
6.474E+18	6.474E+18	1/07/2024	1/07/2024	600000	600000	3003148	3003148	SALIDA	SALIDA	CEDULA DE CIUDADANIA	-
6.474E+18	6.474E+18	1/07/2024	1/07/2024	600000	600000	3003148	3003148	SALIDA	SALIDA	CEDULA DE CIUDADANIA	-
6.474E+18	6.474E+18	1/07/2024	1/07/2024	600000	600000	3003148	3003148	SALIDA	SALIDA	CEDULA DE CIUDADANIA	-

VALIDACIÓN_PRUEBA.sql CONSULTAS_EXP_PRUEBA.sql

- Debido al hallazgo anterior se decidió solo trabajar con los registros transaccionales en donde el tipo de documento fuera diferente de "-", por lo tanto solo nos iba a cruzar con los clientes que también tuvieran un tipo de documento diferente a "-" en la tabla de clientes, debido a esto se crearon dos nuevas tablas en donde se almacenaba los registros tanto de clientes como transaccionales que cumplieran con esta condición, finalmente se realizo el cruce entre la tabla transaccional, de clientes y de canales, adicional se creo una tabla con la longitud y la latitud de todos los municipios con su respectivo código DANE y se creo la tabla que iba a ser consumida por el POWER BI en donde se realizo una agrupación par así reducir el numero de registros a cargar.

CONSULTA_DF.SQL

- La tabla con el cruce de las tres tablas se subió como insumo para realizar el análisis estadístico descriptivo pertinente en el notebook.
- Finalmente se dio paso a realizar las consultas necesarias para dar respuesta a:

¿Qué clientes han realizado transacciones en los últimos 6 meses por un monto total superior al 200% de sus ingresos mensuales y superiores al percentil 95 del total de la población por tipo de persona?

Reporte de clientes:

REPORTE_CLIENTES.xlsx pregunta_prueba.sql

- Finalmente se desarrolla el tablero de POWER BI con las tablas anteriormente mencionadas.