Tercera entrega

Juan Manuel Ramírez, Juan Manuel Uribe y Juan Esteban Velandia

Fuente de los datos: https://www.datos.gov.co/Econom-a-y-Finanzas/Tarjetas-de-cr-dito-y-d-bito/h2jg-r3zg.

Link del repositorio GitHub: https://github.com/JuanManuel29/Proyecto_ID.git

Descripción:

La base de datos escogida es una compilación de datos referentes a las tarjetas de crédito y débito en Colombia. Para esto, se tienen múltiples datos referidos a elementos como tarjetas vigentes, abiertas, cerradas, tecnología y modos de uso, entre otros; lo cual es referido como subcuenta. Los datos son divididos entre personas naturales y jurídicas con tal de entender cuántas personas privadas y cuantas entidades corporativas hacen parte de los datos registrados dentro de cada subcuenta. Adicionalmente, se tienen datos para identificar la UCA o la entidad financiera que respalda el tipo de tarjeta (ej: MasterCard) y la entidad que realiza la oferta de estas tarjetas al público general junto a una clasificación de esta última bajo su actividad financiera. Teniendo esta base de datos identificada, se procedió a realizar la adaptación de los datos en Excel a postgreSQL para el procesamiento y la visualización de los mismos a través de Dash. Las visualizaciones se basaron en cuatro escenarios que serán explicados y analizados posteriormente.

Reglas de negocio:

1. A una subcuenta se le pueden asignar varios registros.

2. Se identifica una subcuenta por medio de un código, además posee una descripción.

3. Buscar identificar las características de las subcuentas, tipos de tarjetas e internacionalización por medio de la descripción.

4. Los registros cuentan con un número único, la fecha de registro, el número de datos por persona jurídica, persona natural y un total del número de datos que puede ser la suma de los tipos de personas o no.

5. Se debe poder identificar cuáles totales son las sumas y cuáles no, es decir, cuáles subcuentas son flujos y cuales stocks.

6. Poder identificar la cantidad de registros por subcuenta.

7. Los registros se ven agrupados por medio de las UCA. Una UCA puede tener varios registros.

8. Se debe percatar dentro de los registros, cuál fue la fecha donde más se registró cierta UCA.

9. Las UCA tienen un código único y su nombre.

10. Una clase posee un código y un nombre. Estas clasifican a las entidades por clases.

11. Las entidades requieren obligatoriamente de una clase para poder conocerse. Estas cuentan con un código que las identifica además de su nombre.

12. Las entidades ofrecen varias UCA Una entidad puede tener estar ofreciendo a varias UCA.

13. Se debe diseñar un ranking de UCA´s en base a cuáles tienen más subcuentas.

Diagrama entidad relación:

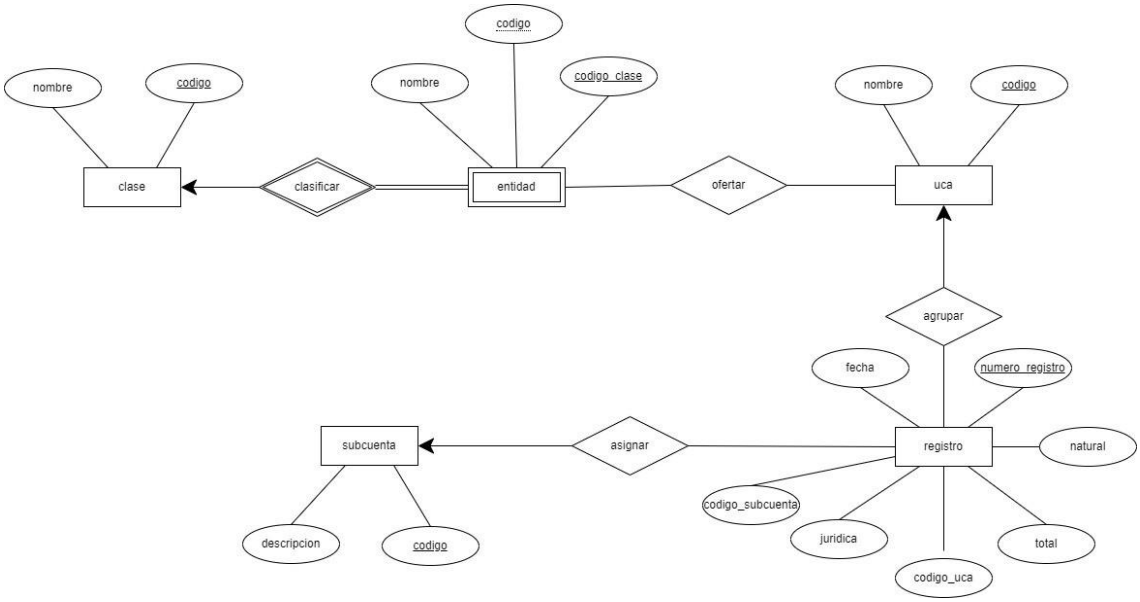
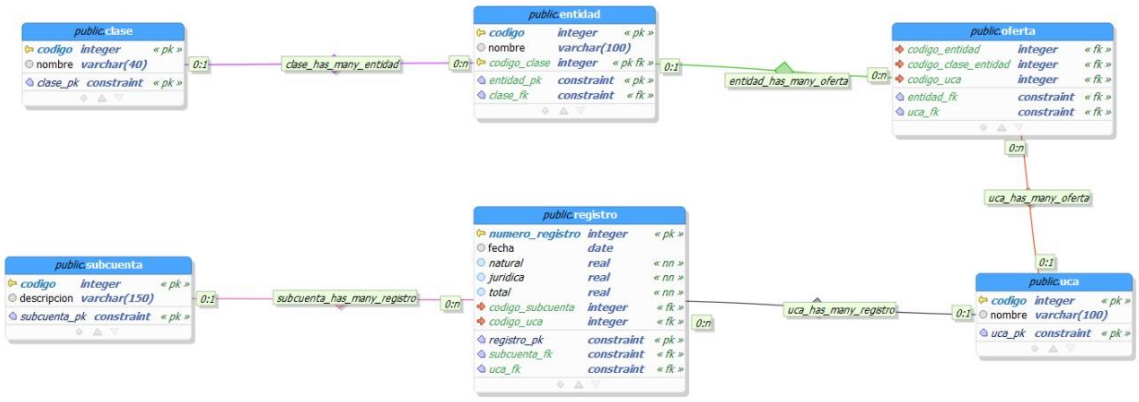
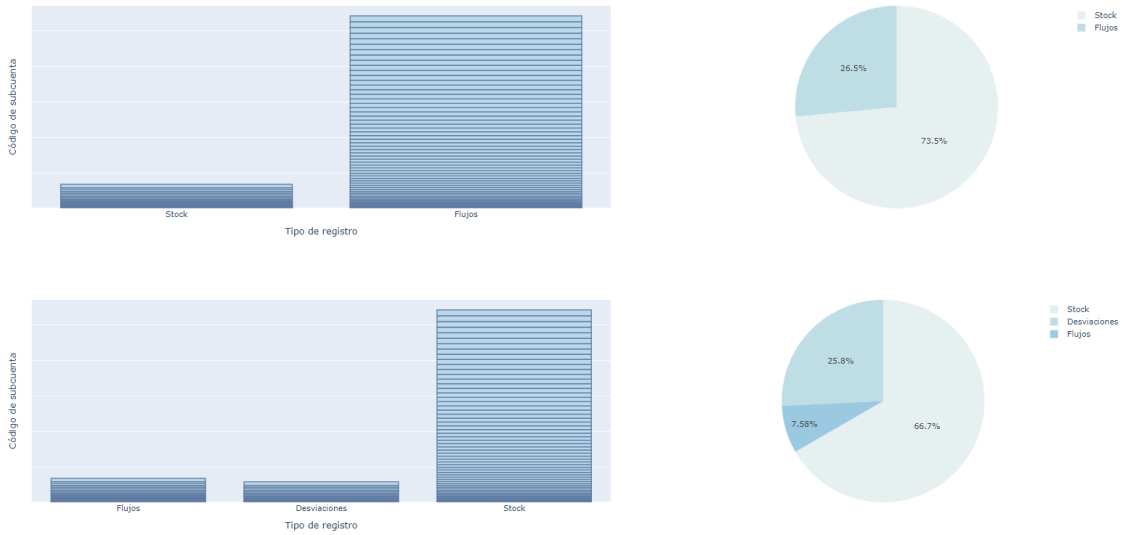


Diagrama relacional (normalizado en tercera forma normal):



Primer caso: Clasificación de subcuentas según características de registros

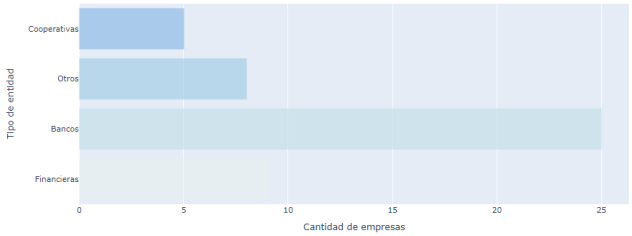


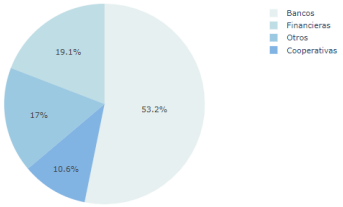
El primer caso de análisis se basó en buscar diferenciar aquellos registros considerados como flujos, donde la suma de personas naturales y jurídicas es igual al total, y stocks, donde la suma mencionada no fuese igual al total. Adicionalmente, se encontró que algunas subcuentas pertenecían a ambas clasificaciones, por lo cual se consideró un tercer conjunto para visualizar, donde se agregaba el caso de las desviaciones.

Para esto se buscó realizar dos aproximaciones; una donde se tuviesen los valores agrupados, pero identificables por su código y otro donde se toman solo las proporciones sin desagregación por cada subcuenta. La primera visualización planteada fue usar un gráfico de barras: al pasar el cursor por encima de las barras se pueden identificar los códigos y nombres de cada subcuenta. Esta gráfica es eficiente para poder extraer, no solo la separación, sino también identificar cuáles son las subcuentas clasificadas; sin embargo, esta metodología usa la suma de los códigos como altura dando una representación sesgada de la proporción. El segundo tipo de visualización, diagrama de "pie", es más simple dando una mejor idea de la proporción entre los datos, pero no nos dice nada sobre cada subcuenta en particular.

Los resultados muestran que la mayoría de los datos son considerados stocks. Esto implica que la mayoría de los datos son actualizaciones a un total histórico. Sin embargo, agregar el caso de desviaciones muestra que la proporción inicialmente obtenida está sesgada, debido a que estos datos desviados pueden implicar uno de dos casos: los datos de flujo pueden parecer stocks debido a la generación de nuevas empresas donde su primer registro de un flujo se comporta como un stock o los datos de stocks son reportados con errores por parte de algunas entidades.

Segundo caso: Análisis de distribución de empresas según tipos de entidades



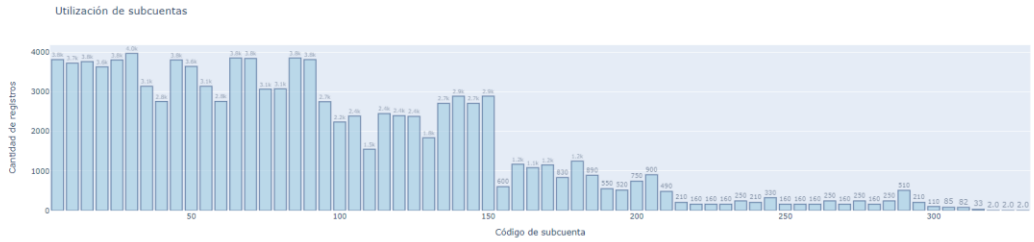


Aprovechando la existencia de la tabla “clase” que diferencia las empresas en tipos de entidades según sus actividades financieras, se procedió a buscar como se daba esta distribución. En este sentido se optó por usar dos métodos de visualización. Primero, se optó por el uso de un gráfico de barras simples horizontales; esto debido a que no es necesario identificar el contenido de las clasificaciones (pertenencia individual de las empresas) y la dirección horizontal permitía una mejor visualización de las magnitudes de cada agrupación.

En términos de desventajas principalmente se tiene que la comparación entre los grupos puede dar ideas inexactas sobre la distribución y el medio matemático requeriría extraer manualmente los valores de cada grupo. Al igual que en el primer caso, se usó el grafico de “pie” teniendo la ventaja de dar una idea más exacta de la proporción de cada clasificación al usar porcentajes, pero se pierde la identificación de la cantidad de empresas.

Los datos muestran una mayor proporción de bancos, siendo estos más de la mitad de las empresas. El resultado es bastante esperado debido a que los bancos comerciales tienden a ser la principal conexión entre el mercado financiero y el público general.

Tercer caso: Utilización de subcuentas

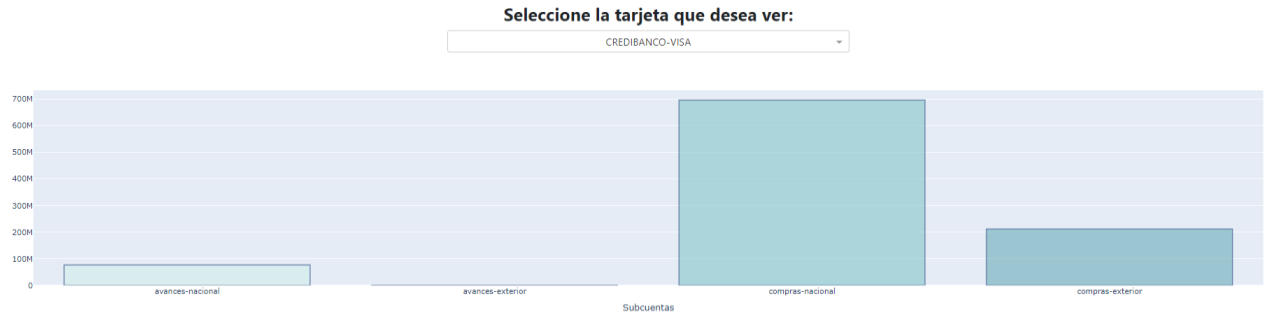


En el tercer caso se buscó hallar la utilización de cada subcuenta (tomando utilización como cantidad de registros pertenecientes a esta). Se consideró que un único gráfico de barras era apropiado debido a que permite identificar el código de cada subcuenta junto a su cantidad de registros. Esta visualización tiene los beneficios de presentar la cantidad encima de cada barra, siendo una aproximación para los casos con más de mil registros; y el de también reportar los valores al pasar el cursor encima de las barras, siendo este preciso para cualquier cantidad de datos. En términos de desventajas, el grafico base cumple al objetivo deseado, sin embargo, en términos de análisis a mayor profundidad, como organización por

ranking, se tiene que la visualización es estática y requeriría ordenación manual de los resultados.

La utilización de subcuentas, al estar organizada por el código, muestra una tendencia decreciente de la cantidad de registros, esto se da por que las subcuentas tienden a estar más asociadas a algún tipo de entidad, por lo cual las primeras subcuentas son utilizadas por la mayoría de las entidades que análogamente poseerían la mayoría del mercado, mientras las últimas pueden ser subcuentas muy específicas a la naturaleza de grupos pequeños de empresas.

Cuarto caso: Comparación de uso de tarjetas de crédito según UCA



El último caso buscaba realizar comparaciones del uso de las tarjetas de crédito para realizar avances o compras y simultáneamente, para comparar que cada una sí fuera realizada nacionalmente o desde el exterior. Esta comparación se realizó por cada UCA y se puede navegar entre estas mediante el menú desplegable. Esta visualización tiene la ventaja usual de los gráficos de barras, al dar una representación de la magnitud o cantidad de la subcuenta, y ofrecer la cantidad exacta al pasar el cursor en el Dash. Además, separar la ventana mediante un desplegable permite mantener una mayor facilidad de visualización al no tener todas las cuentas en una sola gráfica; esto también permite identificar que UCAs no tienen registros en alguna de las subcuentas con mayor facilidad. La principal desventaja de la visualización es que no permite comparación directa entre las UCA debido a que se necesitaría manualmente extraer ambas graficas.

La mayoría de UCAs muestran tener un mayor uso nacional que en el exterior, lo cual es esperado, debido a que se trata de un mercado nacional. Entre avances y compras, se evidencia una fuerte preferencia por realizar directamente compras con las tarjetas que solicitar avances de dinero. Esto último puede explicarse por altas tasas de interés, facilidad de uso para compras directas o desconocimiento del mecanismo de avances.

Conclusiones:

Durante la realización de este proyecto, se pasaron por varios inconvenientes para seleccionar y adecuar la base de datos. Inicialmente para la selección se tuvieron problemas con la temporalidad de los datos debido a que muchas bases ideales fueron publicadas en el semestre anterior. Adicionalmente, se tuvieron problemas con el tamaño de la base de datos dado que la mayoría no tenía suficientes variables(columnas) o entradas(tuplas) para realizar análisis detallados y complejos.

Una vez se decidió usar las tarjetas de crédito y débito se encontró problemas en la interdependencia de los datos una vez se separaron en tablas, esto se dio en parte por los códigos, debido a que llegaban a depender de factores como el tipo de entidad y la UCA generando así tablas débiles que, modificando los códigos, funcionaban perfectamente como tablas fuertes. Por otro lado, esta interdependencia generaba problemas adicionales como manejar el registro como tabla o como una relación con atributos, lo cual generaba dificultades adicionales en el diseño de las relaciones.

Al momento de cargar los datos se tuvieron problemas con los códigos debido a que siendo llaves primarias generaban repeticiones y por tanto entidades débiles innecesarias, como se mencionó previamente, se optó por las modificaciones de los mismos para subcuentas. Similarmente, el formato de las fechas y los textos que, al estar en español, usaba tildes las cuales requirieron un procesamiento previo. Por parte de las conexiones de bases de datos, el proceso de creación de sentencias SQL y el paso a Python fue sencillo y sin impedimento gracias a las soluciones realizadas.

Por otra parte, aparecieron problemas en la creación de los gráficos, en particular para el primer caso de análisis, puesto que se requería de hacer un gráfico con dos variables cualitativas, algo que no es usual en los gráficos de barras. Por otro lado, la separación del cuarto caso por UCAs causó la necesidad de separar la sentencia de Python para SQL por casos.

Finalmente, en lo que refiere netamente a la programación de la aplicación en Dash, el reto más grande estuvo en organizar el 'layout' de la aplicación y en programar la actualización del gráfico para el último caso. Las dificultades para organizar todos los títulos y textos dentro de la aplicación se debieron a que nuestros conocimientos del uso del html en Dash no son los más avanzados, y por esto, casi siempre se obtenían errores de falta de agrupamiento de varios elementos. Para el caso de la actualización del gráfico, la dificultad estuvo en la utilización de una de las funciones 'callback' propias de Dash para poder actualizar el gráfico en tiempo real según la información que el usuario deseara ver. Cabe aclarar que para hacer esto posible, también se tuvieron que reorganizar las consultas SQL para este caso.