Taller 001 individual

Perfilación clientes servicios bancarios.

Laura Juliana Mora Páez[a,c] ,

Ana María Beltrán Cortés[b,c]

[a]*Estudiante de Maestría en Ingeniería de Sistemas y Computación*
[b]*Profesor, Departamento de Ingeniería Industrial*
[c]*Pontificia Universidad Javeriana, Bogotá, Colombia*

1. ENTENDIMIENTO DEL NEGOCIO

El sector financiero en Colombia está compuesto por un gran número de entidades bancarias, las cuales ofrecen diferentes alternativas para el manejo de los recursos monetarios de sus clientes. Para obtener un crédito se exigen una serie de requisitos que en gran parte de la población son difíciles de cumplir, sin embargo sale la opción del uso de las tarjetas de crédito (un ejemplo de créditos de consumo), que se otorgan con mayor facilidad y con cupos rotativos, que permiten al usuario la utilización de este crédito con la sola presentación de la tarjeta, y en la medida que su manejo sea responsable y con pagos oportunos, el cupo de la misma se va incrementando; adicionalmente, con la misma se puede obtener dinero en efectivo con la modalidad de "avances". Una de las cosas que puede representar desventaja con los créditos es sus tasas de interés, que se aproximan a los porcentajes de usura.[i]

Actualmente el banco Scotiabank Colpatria cuenta con tres franquicias para tarjetas de crédito, Visa, MasterCard y American Express, que se adaptan a las necesidades de cada cliente, con el fin de mejorar el flujo de efectivo del mismo a corto plazo, manejando cuotas de hasta 36 meses a nivel nacional, al igual que permite compras a nivel internacional.[ii]

Objetivo de negocio.

Diseñar nuevas estrategias para fidelizar a los clientes, a partir de nuevas promociones según sus hábitos de compras.

Objetivo de minería

Perfilar a los clientes del Scotiabank Colpatria, a partir de los datos sobre en sus hábitos de compras con sus tarjetas crédito, según la cantidad de transacciones realizadas durante el último mes.

2. ENTENDIMIENTO DE LOS DATOS

Se cuenta con una base de datos con 47.871 registros, pertenecientes a diversos clientes del banco Scotiabank Colpatria, de la cual se desconoce como fueron recolectados estos

datos, sin embargo, se sabe que pertenecen a las transacciones del último mes de los clientes que tienen tarjetas Visa, MasterCard o de otras franquicias, que han utilizado sus tarjetas, tanto a nivel nacional como internacional. La base de datos se presenta en un archivo de Excel “infoclientebanca.xlsx”.

A nivel de la granularidad, se observa que los datos, como se menciono anteriormente, están organizados por cliente; la temporalidad de la base de datos es correspondiente a un único mes de medición y no se cuenta con información geográfica y demográfica de cada cliente.

Explorar datos

La base de datos esta estructurada en forma de tabla. Cada fila representa un cliente y para cada uno se cuenta con 26 atributos. Con el fin de dar mayor claridad, a continuación se muestra la Tabla 1. Diccionario de datos para el proceso, en la cual se especifica el nombre del atributo, junto a una pequeña descripción de su significado, el tipo de dato y el conjunto posible de valores que puede tomar el atributo.

Tabla 1 Diccionario de variables

Diccionario de variables			
Variable	**Explicación**	**Tipo de dato**	**Valores posibles**
CLIENTE	Identificador del cliente (anonimizado)	Numérico	1-47.871
grupo_de_cliente	Clasificación del cliente en la segmentación del banco. Se desconocen los detalles	Char	A, B, C, D,E
Numero_de_transacciones	Numero de transacciones en el último mes	Numérico	0- 142
promedio_por_transaccion	Promedio por transacción en el último mes	Numérico	0- 6.262.025
transaccion_minima	Valor de transacción mínima en el último mes	Numérico	0- 6.148.920
transaccion_maxima	Valor de transacción máxima en el último mes	Numérico	0- 11.040.000
desviacion_estandar_por_transaccion	Desviación estándar del valor de las transacciones del último mes	Numérico	0- 5.419.241
porcentaje_visa_nacional	Porcentajes de uso de cada franquicia en el último mes por consumo nacional/internacional	Numérico	0-1
porcentaje_visa_internacional		Numérico	0-1
porcentaje_mastercard_nacional		Numérico	0-1
porcentaje_mastercard_internacional		Numérico	0-1
Porcentaje_otrafranquicia_nacional		Numérico	0-1
porcentaje_otrafranquicia_internacional		Numérico	0-1
porcentaje_nacional_total		Numérico	0-1

porcentaje_internacional_total		Numérico	0-1
porcentaje_manana	Porcentajes de uso de tarjeta en el último mes por bloque del día: mañana (6-12 a.m.) , tarde (12 a.m.- 6 p.m.) y noche (6 p.m-6a.m.)	Numérico	0-1
porcentaje_tarde		Numérico	0-1
porcentaje_noche		Numérico	0-1
porcDOMINGO	Porcentaje de uso en el último mes en cada uno de los días respectivos del mes	Numérico	0-1
porcLUNES		Numérico	0-1
porcMARTES		Numérico	0-1
porcMIERCOLES		Numérico	0-1
porcJUEVES		Numérico	0-1
porcVIERNES		Numérico	0-1
porcSABADO		Numérico	0-1
Sitio_consumo_masfrecuente	Clasificación MCC del grupo de sitios de consumo más frecuente	Char	109 lugares diferentes, como Clínica

Verificar calidad de datos

Con el fin de ver la calidad de los datos que se decide analizar, las medias de dispersión de los diferentes atributos, para esto se utiliza la función describe() en R teniendo como resultado la tabla presentada en la Ilustración 1 a continuación.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Numero_de_transacciones	1	47871	5.08	8.48	2.00	3.24	1.48	1.00	142	141	5.58	47.21	0.04
promedio_por_transaccion	2	47871	371602.69	579821.74	167173.75	238520.78	159616.89	1.00	6262025	6262024	3.79	18.97	2650.08
transaccion_minima	3	47871	253089.82	518155.91	80000.00	130138.47	80356.00	0.04	6148920	6148920	4.65	28.08	2368.23
transaccion_maxima	4	47871	580731.86	885419.14	257933.00	380961.72	266116.32	1.00	11040000	11039999	3.67	19.14	4046.81
desviacion_estandar_por_transaccion	5	47871	139883.50	318090.76	30440.94	65966.67	45131.74	0.00	5419241	5419241	5.29	41.23	1453.83
porcentaje_visa_nacional	6	47871	0.37	0.41	0.20	0.34	0.30	0.00	1	1	0.53	-1.39	0.00
porcentaje_visa_internacional	7	47871	0.04	0.16	0.00	0.00	0.00	0.00	1	1	4.92	24.24	0.00
porcentaje_mastercard_nacional	8	47871	0.54	0.43	0.57	0.55	0.64	0.00	1	1	-0.16	-1.67	0.00
porcentaje_mastercard_internacional	9	47871	0.03	0.14	0.00	0.00	0.00	0.00	1	1	5.85	35.32	0.00
Porcentaje_otrafranquicia_nacional	10	47871	0.01	0.09	0.00	0.00	0.00	0.00	1	1	8.59	81.24	0.00
porcentaje_otrafranquicia_internacional	11	47871	0.01	0.08	0.00	0.00	0.00	0.00	1	1	10.49	115.95	0.00
porcentaje_nacional_total	12	47871	0.93	0.23	1.00	1.00	0.00	0.00	1	1	-3.28	9.55	0.00
porcentaje_internacional_total	13	47871	0.07	0.23	0.00	0.00	0.00	0.00	1	1	3.28	9.55	0.00
porcentaje_manana	14	47871	0.29	0.36	0.08	0.24	0.12	0.00	1	1	0.97	-0.51	0.00
porcentaje_tarde	15	47871	0.58	0.39	0.64	0.60	0.54	0.00	1	1	-0.33	-1.40	0.00
porcentaje_noche	16	47871	0.13	0.26	0.00	0.06	0.00	0.00	1	1	2.26	4.19	0.00
porcDOMINGO	17	47871	0.14	0.28	0.00	0.06	0.00	0.00	1	1	2.18	3.64	0.00
porcLUNES	18	47871	0.13	0.27	0.00	0.06	0.00	0.00	1	1	2.30	4.42	0.00
porcMARTES	19	47871	0.13	0.27	0.00	0.06	0.00	0.00	1	1	2.28	4.33	0.00
porcMIERCOLES	20	47871	0.14	0.27	0.00	0.07	0.00	0.00	1	1	2.20	3.90	0.00
porcJUEVES	21	47871	0.14	0.27	0.00	0.07	0.00	0.00	1	1	2.24	4.12	0.00
porcVIERNES	22	47871	0.14	0.27	0.00	0.07	0.00	0.00	1	1	2.18	3.79	0.00
porcSABADO	23	47871	0.18	0.31	0.00	0.10	0.00	0.00	1	1	1.76	1.86	0.00

Ilustración 1 Medidas de atributos base de datos "Banco"

Como se puede observar, la kurtosis en los diferentes atributos es diferente a 0, indicándonos que puede existir una alta presencia de datos atípicos, por los cuales, datos como la media, se pueden ver arrastrados, adicionalmente en algunos datos se nos presentan kurtosis negativas, indicándonos que los datos pueden estar divididos en dos grupos; por otro lado se observan asimetrías pequeñas, con excepción del atributo “porcentaje_otrafranquicia_internacional” ya que presenta un skew en 10.49, finalmente la desviación estándar que se presenta en la mayoría de casos es 0, exceptuando los atributos ligados a las transacciones (Numero_de_transacciones, promedio_por_transaccion,…)

Posteriormente se pasa a observar los datos graficados, por comodidad del documento solo se resaltan algunas gráficas; en la Ilustración 2 que se muestra a continuación se presentan los atributos ligados a las transacciones en forma de Boxplot, como se puede

observar los 5 atributos presentan una alta presencia de datos atípicos, confirmando así lo que nos indicaba la kurtosis.

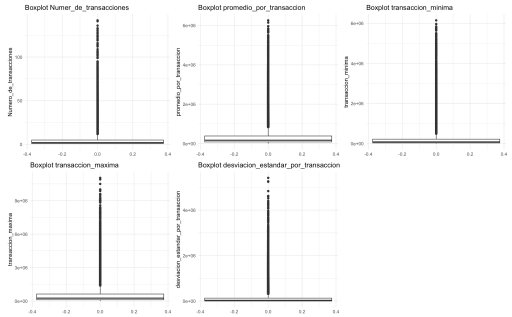

Ilustración 2 Boxplot atributos de transacción

En la ilustración 3 se presentan los Boxplot de los atributos relacionados al horario, en el cual los clientes realizan sus transacciones, como se puede observar el atributo de la noche presenta varios datos atípicos, mientras que para el caso de la tarde, se observa que la mayoría de clientes realizan transacciones en este horario, por otro lado en la mañana no se observan datos atípicos, sin embargo no se presentan tantos clientes que realicen transacciones en este horario.

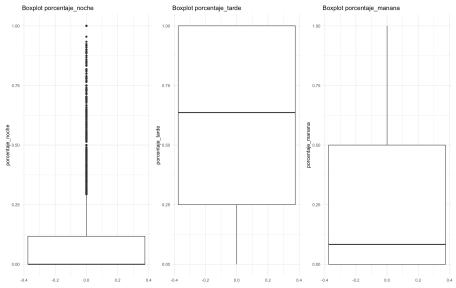

Ilustración 3 Boxplot atributos de horarios

En la Ilustración 4 se presentan los Boxplot de los atributos relacionados al porcentaje de transacciones realizadas según el día de la semana, como se puede observar los 7 atributos tienen una alta presencia de datos atípicos, como menciono la kurtosis, adicionalmente se observa que el día donde más transacciones ocurren es el sábado.

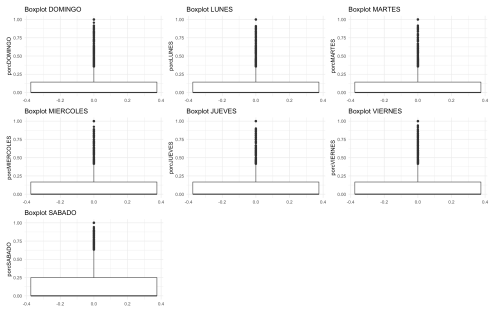

Ilustración 4 Boxplot atributos de días

A continuación, en la Ilustración 5, se presentan los histogramas de consumo según la franquicia de la tarjeta del cliente y si fueron a nivel nacional o internacional; como se puede observar en todos los atributos se presenta una alta cantidad de usuarios que no usaron sus tarjetas, sin embargo, se aprecia que el porcentaje de uso de visa a nivel nacional, si presenta varios usuarios realizando todas sus transacciones con esta franquicia.

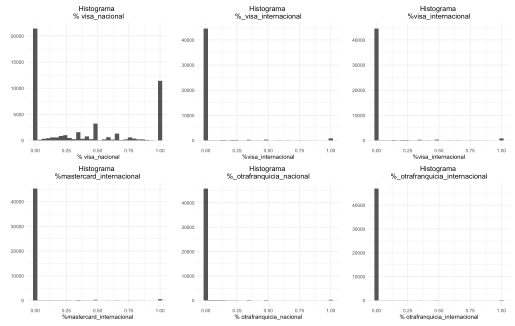

Ilustración 5 Histograma atributos de franquicia

Finalmente, con el fin de observar los atributos categóricos de “grupo_de_cliente” y “Sitio_consumo_masfrecuente”, se realizaron sus respectivos histogramas, sin embargo por espacio en el documento solo se muestra el del “grupo_de_cliente”, donde se puede ver que la gran mayoría de clientes pertenecen al grupo “A”, no obstante se desconocen los detalles de esta segmentación ya realizada por el banco.

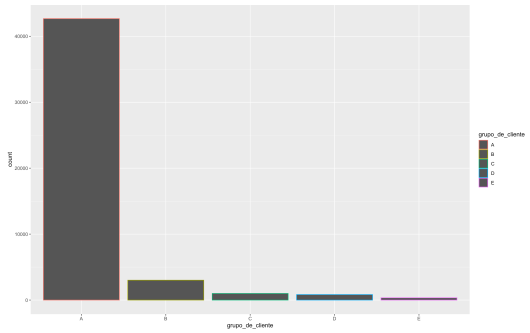

Ilustración 6 Diagrama de barras grupos

3. PREPARACIÓN DE LOS DATOS

Con el fin de preparar los datos se tuvo en cuenta la naturaleza de cada variable y a su vez, la relevancia que tiene para la segmentación teniendo en cuenta las características bajo las cuales se quiere hacer la misma. Cabe recalcar que se realizaron diferentes versiones y aquí se presentan solo los resultados de una de ellas. Para la primera se tomaron todos los atributos de la base de datos, se pasa a aplicar el logaritmo de 1+x con el fin de atacar asimetrías y posteriormente se deja todo en una misma escala; para el segundo caso, se toman los atributos de transacciones y franquicias, se realiza el mismo procedimiento que para todos los atributos; el siguiente caso se toman los atributos de transacciones y horarios, aplicando el mismo procedimiento mencionado anteriormente; con el penúltimo caso se toman los atributos de transacciones y días aplicando el mismo

procedimiento antes mencionando. Los casos mencionados anteriormente, no son presentados, porque la mayoría están ligados solo a un par de grupos de atributos y la primera al tomar todos los datos es muy sensible ya que se tiene el 100% de todo.

Finalmente en el ultimo caso, en el cual el documento esta enfocado, se depuraron y retiraron de la base los atributos relacionadas con el grupo de cliente, las transacciones mínima y máxima, ya que serian redundantes con el promedio de transacción, los atributos ligados a las franquicias, también fueron retirados debido a que como se pudo observar en las secciones anteriores, la gran mayoría de datos eran 0; se crean dos variables derivadas, por un lado porcFinde donde se suman los porcentajes de uso de los días viernes y sábado, que son los días con mayor consumo, por otro lado se crea porcHorario, donde se sumaron los porcentajes de uso en los horarios de mañana y noche, debido a que no abarcaban demasiados consumidores. Con todos estos atributos se crea un data set, donde se agregan los atributos de: el número de transacciones, el promedio de transacción y el porcentaje total de transacciones a nivel nacional.

Basados en la kurtosis se presentan datos atípicos, y con el fin de dejar todos los atributos en una misma escala, para evitar que algunos pesen más que otros, se pasa a realizar el proceso anteriormente mencionado, aplicando el logaritmo de 1+x y modificando la escala de los datos. Es así como se obtiene como resultado, las medidas de tendencia central y de dispersión presentadas en la Ilustración 7 a continuación.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Numero_de_transacciones	1	47871	0	1	-0.41	-0.15	0.78	-0.93	4.61	5.54	1.16	1.04	0
promedio_por_transaccion	2	47871	0	1	-0.07	-0.03	0.97	-9.69	3.01	12.70	0.08	1.17	0
porcentaje_nacional_total	3	47871	0	1	0.30	0.30	0.00	-4.21	0.30	4.51	-3.47	10.92	0
porcentaje_tarde	4	47871	0	1	0.25	0.07	1.10	-1.58	0.99	2.57	-0.53	-1.23	0
porcFinde	5	47871	0	1	-0.23	-0.10	1.03	-0.92	1.72	2.64	0.60	-1.11	0
porcHorario	6	47871	0	1	0.00	-0.03	1.66	-1.12	1.39	2.51	0.14	-1.53	0

Ilustración 7 Medidas de atributos tras procesos para Bancomix

Adicionalmente, y con el fin de facilitar la perfilación de los futuros grupos resultados del proceso de clustering, se crea un nuevo atributo “sitiosNuevos”, donde se toman los cinco lugares donde más transacciones se realizan, y al resto se les asigna el nombre “Otro”, con el fin de reducir las 109 opciones de lugares de transacción que se presentaban en un inicio.

4. MODELACIÓN

Con el propósito de determinar la cantidad de clusters para realizar utilizando K-means, se usó para cada uno de los casos mencionados anteriormente, el diagrama de codo debido a la alta cantidad de datos, alternativas como silueta y gap no eran posibles de correr en el computador. Para el primer caso el diagrama de codo, indicó que se utilizaran alrededor de 9 clusters, para el caso donde se utilizan los atributos de transacciones y franquicias, el diagrama indicó que son alrededor de 3 clusters, en el caso de atributos de transacciones y horarios, en el diagrama se observa que son alrededor de 4 clusters, el caso de atributos de transacciones y días, se observa que la mejor opción para el número de clusters son 5; finalmente, para el caso presentado en el diagrama de codo, como se puede observar en la Ilustración 8, la opción de utilizar 4 clusters.

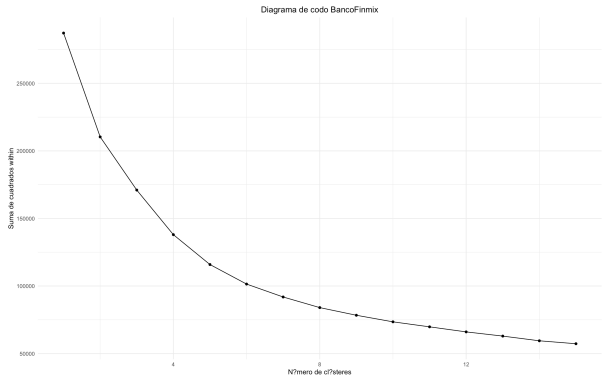

Ilustración 8 Diagrama de codo Bancomix

Tras aplicar k-means con el data set de datos preparados y teniendo en cuenta la cantidad de clusters sugeridos por el diagrama de codo, a continuación se presentan los centroides obtenidos en un mapa de calor.

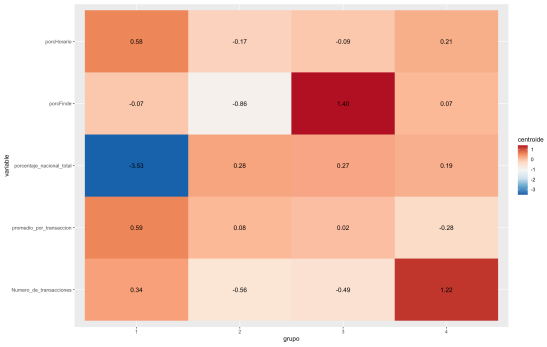

Ilustración 9 Diagrama de calor clusters Bancomix

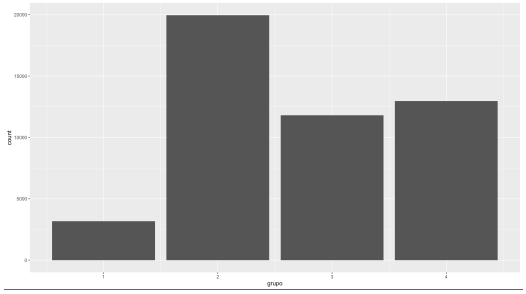

Ilustración 10 Diagrama de barras clusters Bancomix

5. EVALUACIÓN

A continuación se presenta en la Ilustración 11, la validación de los resultados obtenidos, donde se observan que para los grupos 2 y 3, los resultados son muy buenos, para el 4 los resultados fueron regulares, sin embargo para el grupo 1 fueron malos, esto debido a que es el grupo donde menos usuarios quedaron catalogados.

```
> #la validaci?n del resultado. >0.75 o .85 muy bueno; <.6 malo
> kclusters$bootmean
[1] 0.5659517 0.8051661 0.7950751 0.6761262
```

Ilustración 11 Validación clustering (k-means) con 5 atributos.

En la Ilustración 10 se puede observar como quedaron divididos los clientes en los diferentes grupos, donde resaltan el grupo 1 con menos clientes y el grupo 2 donde más clientes hay, mientras que los grupos 3 y 4 están bastante parejos en la cantidad de clientes presentes.

Con el fin de dar una mayor claridad a los grupos obtenidos, en la Tabla 2 se presenta una descripción de los mismos, junto con sus datos más representativos a nivel de transacciones, al igual que en esta se mencionan las relaciones de los grupos con los

lugares de mayor consumo obtenidos a través de un mosaico (el cual se puede observar en el script).

Tabla 2 Descripción grupos obtenidos con k-means

Grupo	Descripción	Promedio cantidad de transacciones	Promedio transacción
1	En este grupo se encuentran las personas que más tienden a comprar en horarios de la mañana o noche, adicionalmente no tienden a comprar los días viernes y sábados, como tampoco realizan muchas transacciones. Finalmente se observa que las características por las que resaltan, es que son clientes que no realizan transacciones a nivel nacional y son quienes más gastan por transacción.	7,27	689.036
2	En este grupo es donde más clientes se encuentran y son los clientes que no compran mucho en las mañanas y noches, no tienden a comprar durante los viernes y sábados, la cantidad de transacciones que realizan no son muy altas, de los cuatro grupos son quienes más realizan transacciones a nivel nacional y finalmente no gastan mucho por transacción. Adicionalmente este grupo no realiza muchas transacciones en supermercados, almacenes y otros lugares.	1,85	426.635
3	En este grupo se presentan clientes que no realizan muchas transacciones, no compran en horarios de mañana ni noche, tampoco tienden a gastar mucho por transacción, sin embargo resaltan por ser quienes más compran durante los viernes y sábados y tienen a realizar varias transacciones a nivel nacional. Adicionalmente las compras de estos clientes se concentran principalmente en almacenes y supermercados.	2,02	375.836
4	En este grupo se presentan las personas que menos gastan por transacción, sin embargo realizan transacciones en horarios de mañana y noche, los días viernes y sábados, a nivel nacional y la característica por la que más resaltan son los usuarios que más transacciones realizan. Adicionalmente este grupo realiza muchas transacciones en supermercados, almacenes y otros lugares.	12,3	205.537

6. DESPLIEGUE (RECOMENDACIONES DE NEGOCIO)

- Implementar una campaña nacional, incentivando el consumo de las tarjetas de crédito a través de convenios con los lugares donde más transacciones se realizan.
- Para el grupo 1 teniendo en cuenta que sus transacciones no son a nivel nacional, sus transacciones son de un costo muchísimo más alto se propone disminuir los intereses en estas transacciones.
- Para el grupo 3 se proponen realizar alianzas con supermercados y almacenes para facilitar las compras de estos clientes.
- Para todos los grupos se sugiere analizar si cuentan con cuentas de ahorros con el banco, con el fin de que no tengan que cargar con efectivo, permitirles el pago de sus tarjetas de crédito a través de sus cuentas de ahorro.
- Para el grupo 4 teniendo en cuenta la cantidad de transacciones que se realizan con el fin de incentivar aún más estas transacciones, se propone disminuir el monto de la cuota de manejo según la cantidad de transacciones realizadas en el mes, al igual que disminuir el costo de los intereses según estos.
- Se sugiere para futuros análisis, agregar variables demográficas como la edad de los clientes y su genero, con el fin de conocer más de ellos y poder entender más su comportamiento a la hora de realizar transacciones.

BIBLIOGRAFÍA

[i] Entrevista a Contador público

[ii] "Tarjeta de Crédito - Solicita la tuya en línea." *Scotiabank Colpatria*, https://www.scotiabankcolpatria.com/personas/tarjetas-de-credito. Accessed 25 Feb. 2022.