ENSAYO DATATON

Equipo Datalyticos

Descripción breve

Este documento corresponde al entregable solicitado en los requerimientos del reto Dataton realizado por Bancolombia. Describe los procedimientos, métodos, técnicas y estrategias utilizadas para hallar la solución al reto.

Carolina Flórez - Stiven López - Andrés González

Integrantes del Equipo

Tabla de contenido

Ensayo Datatón Bancolombia

Equipo: Datalyticos

Objetivo principal: Hallar la **clasificación** de las transacciones realizadas mediante la pasarela de pagos PSE (Pagos Seguros en Línea), desde una óptica del cliente, es decir si esta transacción es de tipo hogar, viaje, turismo, cuidado personal, etc.

Otros objetivos: Generar propuestas de valor a partir de la información suministrada, finalmente enfocadas al cliente.

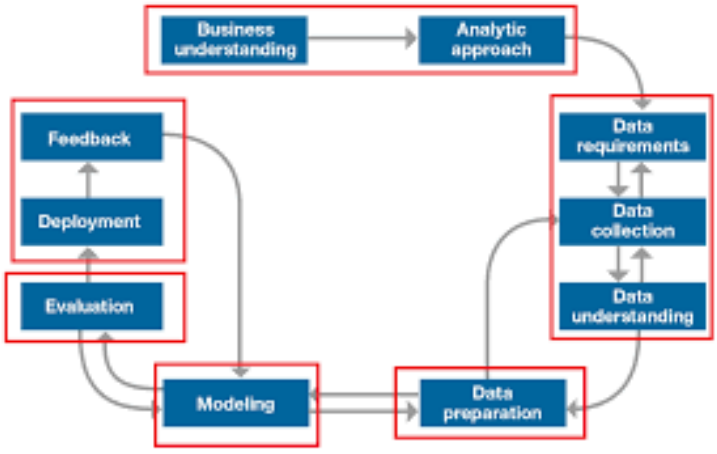
La realización del reto se llevará a cabo con **herramientas en su versión de uso libre**: DQ Analyzer, Python y Power BI.

Para trabajar con la totalidad de los datos, se levanta una instancia en **AWS (Amazon Web Services)**, con 32 GB de RAM, 8 Cores y 60 GB de Disco.

Los pasos a seguir para el desarrollo del objetivo estarán regidos bajo la metodología ASUM-DM, ya que tenemos una cantidad de datos considerable, y esta metodología extiende el modelo de CRISP-DM, considerando la infraestructura (las estaciones o PCs tradicionales no son las indicadas al momento de procesar y almacenar grandes cantidades de datos -> Big data, se deben considerar arquitecturas robustas que apoyen y solucionen esto), además aborda el elemento fundamental que es la gestión del proyecto, indispensable para alcanzar los objetivos de toda actividad (se tratara de emular la metodología de SCRUM, definiendo en un solo sprint todas las actividades a realizar para alcanzar la meta del proyecto, centrándonos en una comunicación personal y constante entre los integrantes del grupo).

Fases ASUM-DM

1. **Analizar**
2. **Diseñar**
3. **Configurar y construir**
4. **Desplegar**
5. **Optimizar**



El esfuerzo estará enfocado en las tres primeras etapas.

Desarrollo de la Metodología ASUM-DM

1. Entendimiento del negocio

Se analiza el tipo de usuario y negocio para el cual se está desarrollando el proyecto:

¿Qué hace un **banco**?

Sujeto A → Sujeto B

Sujeto A: Persona natural, Persona Jurídica, Empresa

Sujeto B: Persona natural, Persona Jurídica, Empresa

→ Se **relacionan** financieramente.

- **Estrategia:** Enfocada en el cliente
- **Foco del reto:** Transacciones PSE

Funcionamiento PSE

- **Usuario:**
- **Correo:**
- **Clave:**
- **Código de servicio:** viene predefinido ejemplo: 1001
- **Descripción del servicio:** Campo Libre, es el campo de la referencia en la fuente de datos
- **Cuenta recaudadora:** asociada al banco
- **Banco de la cuenta recaudadora:**

solo modifica la información el administrador de cuentas, además como información adicional de PSE, solo se permiten **transacciones mínimas de 1.600** $.

MCC: Es usado para clasificar el negocio por el tipo de bienes o servicios que provee.

2. Enfoque analítico

Surge la gran pregunta **¿qué tipo de aprendizaje de Machine learning elegir?** Para llevar a cabo el objetivo propuesto.

Por la descripción del reto se piensa en primera instancia en **aprendizaje supervisado**, entonces surge la inquietud: **¿qué variable de los datos hay disponible para hacer la clasificación deseada?**

No hay ninguna variable target para la clasificación de los gastos.

¿Qué hacer?

Por la definición del MCC (en POS) en este caso el sector del beneficiario, se podría identificar a que categoría de gastos personales está asociada la transacción, pero no se dispone en su totalidad de esta información, se podría usar la información que si conocemos y etiquetarla con base a la lista de la categoría de gastos sugerida, desarrollar un clasificador tomando el 70% de los datos para entrenamiento y el otro 30% para pruebas, y luego imputar el label a los datos restantes usando el clasificador, se debe tomar en cuenta que estos clasificadores se deben entrenar cada cierto tiempo para mejores resultados.

En la entrega del reto mencionaban que para el subsector bancos, las referencias eran muy variadas y no era posible clasificarlas solo por el subsector y/o descripción, por ejemplo, el pago de administración de edificios, tienen como subsector bancos, para atacar este tema, se realiza una clusterización (Clustering) que arroje una clasificación mas natural del campo referencia, se debe tener en cuenta que este campo que este campo es libre y se debe tratar como lenguaje natural.

3. Requisito de los datos

Para el tema de la clusterización y clasificación de las transacciones PSE, se requiere que la información de las variables relacionadas con la referencia, sector, subsector y descripción que se usa como fuente de información para los modelos, se requieren en tipo String y totalmente limpia, es decir se debe eliminar caracteres especiales, se deben remover las acentuaciones y dejar el campo en codificación UTF-8, se dejan cadenas con caracteres entre A-Z, a-z , dado que los números no aportan valor a los objetivos, se deben filtrar los stop words para el análisis, y toda la cadena debe estar en minúscula para evitar sensibilidades.

4. Recolección de los datos

- **dt_trxpse_personas_2016_2018_muestra_adjt**

Tabla con transacciones PSE durante 2016-09 a 2018-10 (muestra aleatoria de clientes persona -- 340 mil clientes --)

Campo	**Descripción**	**Tipo**
id_trn_ach	identificador único de transacción	string
id_cliente	id. único de cliente (pagador)	bigint
fecha	fecha de transacción	decimal(8,0)
hora	hora de transacción (HHMMSS)	decimal(6,0)
valor_trx	valor ($) transacción	double
ref1	texto libre referencia 1	string
ref2	texto libre referencia 2	string
ref3	texto libre referencia 3	string
sector	sector eco. receptor	varchar(24)
subsector	subsector eco. receptor	varchar(62)
descripcion	descripción subsector receptor	varchar(24)

- **dt_info_pagadores_muestra**

Tabla con (alguna) información demográfica de los pagadores (muestra aleatoria de 340 mil clientes)

Campo	**Descripción**	**Tipo**
id_cliente	id. único de cliente (pagador)	bigint
seg_str	segmento estructural	string
ocupacion	ocupación	string
tipo_vivienda	tipo de vivienda	string
nivel_academico	nivel académico	string
estado_civil	estado civil	string
genero	genero	string
edad	edad	int
ingreso_rango	rango de ingreso estimado	string

5. Entendimiento de datos:

Enfocados en las herramientas en su versión de uso libre, por practicidad se realiza una exploración inicial de datos en DQ Analyzer para el perfilamiento de todo el set de datos entregado por el banco relacionado con las transacciones, y otra en Python para transacciones y clientes, la cual nos entrega un análisis mas detallado de los datos.

Archivos análisis exploratorio en Python, análisis exploratorio en Python:

- **Exploracion_clientes. Ipynb (Esta en la ruta de GitHub)**
- **Exploracion_transacciones. ipynb (Esta en la ruta de GitHub)**

Se realizan dos procedimientos de entendimientos de los datos para el archivo de transacciones ya que por experiencia esta etapa es de suma importancia a la hora de proponer y/o implementar soluciones.

Al leer el **dt_trxpse_personas_2016_2018_muestra_adjt** se percibe que había filas que estaban haciendo saltos de fila donde no debería existir, campos \n, y campos de descripciones cuyo contenido había comas, estos inconvenientes fueron solucionados con Notepad ++.

Resumen General:

Input: **"datos_bc"**

Columns

Column Analyses

Quick filter: Advanced Filter

Expression	Type	Domain	Non-null	Null	Unique	Distinct	Min	Median	Max
id_trx	STRING	integer pat...	11.853.782	0	11.853.782	11.853.782	1000000017	296127621	373668411
id_cliente	LONG		11.853.782	0	40.966	338.606	1	144.906	338.606
fecha	LONG	day	11.853.782	0	0	761	20.160.901	20.171.122	20.181.001
hora	LONG		11.853.782	0	1.197	85.063	0	135.750	235.959
valor_trx	FLOAT		11.853.782	0	8.688.332	10.069.784	0,01	126.148,93	1.788.605.269,62
ref1	STRING		11.492.514	361.268	413.977	487.745	$ en fichas	Pago de fa...	ZZZZTA
ref2	STRING		6.801.921	5.051.861	141.210	171.900	´	CC	ZZY null
ref3	STRING		0	11.853.782	0	0			
sector	STRING	enum patt...	11.853.782	0	0	11	\N	\N	SERVICIOS NO FINANCIEROS
subsector	STRING	pattern	11.853.782	0	3	55	\N	\N	VALOR AGREGADO
descripcion	STRING		11.853.782	0	8	150	\N	\N	TURISMO Y AGENCIAS DE TURI...

Se tienen 11.853.782 transacciones, en un periodo de tiempo comprendido entre: el 01 de septiembre de 2016 y 01 octubre del 2018, con un total de clientes analizados de: 338.606, con un valor mínimo de transacción de 0,01, y un máximo de 1.788.605.269, se tiene en el campo estrella ref1 361.268 valores nulos, del mismo modo se percibe que el campo ref3 no genera valor dado que el 100 % de los datos es nulo.

Se tiene 11 categorías de sector, 55 categorías de subsector, 55 categorías de subsector y todos estos en su mínimo nivel de detalle que es la descripción hay 150.

Exploración campo a campo:

- **Id_trx:** Es de tipo String

Se puede observar que no hay registros duplicados en este campo, que la cantidad de registros en el set es de 11.852.782 millones y que no hay valores nulos.

Basic | Frequency | Domains | **Masks** | Quantiles | Groups

Mask Analysis

Value	Count	Percentage
DDDDDDDDD	11853778	100,00%
DDDDDDDDDD	4	0,00%

La mayoría de los registros tienen 9 dígitos.

- **Id_Cliente:** es tipo Bigint**:**

Counts

Type	Count
Non-null values:	11.853.782
Null values:	0
Distinct values:	338.606
Duplicate values:	11.515.176
Unique values:	40.966
Non-unique values:	297.640

En este campo no hay valores nulos, se presentan transacciones de 338.606 clientes distintos, hay 40.966 clientes que solo aparecen una única vez, y 297.640 clientes transaron mas de una vez en el periodo de tiempo.

Extremes

First Values:
1
2
3
4
5

Last Values:
338.602
338.603
338.604
338.605
338.606

Estos valores son consistentes ya que se percibe en la información que son campos autonuméricos en la maestra de clientes.

Frequency Analysis

Value	Count	Percentage
210.949	6187	0,05%
187.977	5776	0,05%
26.729	5727	0,05%
157.323	5261	0,04%
122.249	5139	0,04%
66.700	4357	0,04%
131.290	3950	0,03%
228.889	3471	0,03%
52.442	3379	0,03%
191.003	3238	0,03%
274.004	3156	0,03%
110.262	3049	0,03%
191.545	2994	0,03%
190.305	2915	0,02%

Estos son los id de clientes que mas transaron en ese periodo de tiempo.

Group Frequency Analysis

Group Size	Group Count	Percentage
1	40966	12,10%
2	23643	6,98%
3	16747	4,95%
4	12968	3,83%
5	10853	3,21%
6	9357	2,76%
7	8142	2,40%
8	7447	2,20%
9	6711	1,98%
10	6077	1,79%

Esta tabla muestra que cantidad de id_cliente transaron por PSE, 1 vez, 2 veces, 3 veces o más, es decir 23.643 id_cliente, solo aparecen 2 veces en el set de datos.

- **Fecha:** Es de tipo Bigint

Counts

Type	Count
Non-null values:	11.853.782
Null values:	0
Distinct values:	761
Duplicate values:	11.853.021
Unique values:	0
Non-unique values:	761

En este campo no hay valores nulos, hay 761 fechas diferentes, no hay valores que aparezcan una sola vez.

Extremes

First Values:
20.160.901
20.160.902
20.160.903
20.160.904
20.160.905

Last Values:
20.180.927
20.180.928
20.180.929
20.180.930
20.181.001

Las fechas están entre 20160901 y 20181001

Frequency Analysis

Value	Count	Percentage
20.180.502	37354	0,32%
20.180.903	35431	0,30%
20.180.402	35176	0,30%
20.180.515	34305	0,29%
20.180.801	34289	0,29%
20.180.615	34109	0,29%
20.180.815	33433	0,28%
20.180.605	33407	0,28%
20.180.301	33146	0,28%
20.180.703	32946	0,28%

Están son las fechas en las que más transacciones se hicieron, se pude deducir que las mayores transacciones se hacen a principio de mes y

- **Valor_trx:** esta variable es de tipo float

Statistics

Type	value
Minimum value:	0,01
Median value:	126.148,93
Maximum value:	1.788.605.269,62
Sum:	4.348.578.626.150,79
Average:	366.851,58
Variance:	2.545.179.426.271,83
Standard deviation:	1.595.361,85

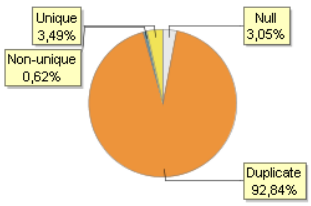
Extremes

First Values:
0,01
0,04
0,09
0,1
0,15

Last Values:
839.052.835,55
868.320.795,03
898.559.755,58
1.361.750.224,41

A simple vista se detectan valores atípicos tanto en los máximos como en los mínimos, en la definición de negocio se consideran transacciones atípicas de más de 50 millones.

- **Ref1**: Es de tipo String



Counts

Type	Count
Non-null values:	11.492.514
Null values:	361.268
Distinct values:	487.745
Duplicate values:	11.004.769
Unique values:	413.977
Non-unique values:	73.768

En el total de los datos hay 361.268 valores nulos, hay 487.745 referencias distintas sobre estas se debe hacer la segmentación, y 413.977 referencias que solo aparecen 1 sola vez en el set de datos, también se difiere que hay 73.768 referencias que aparecen más de 1 vez.

First Values:
$ en fichas
BEAUTY BRICK BLUSH; PRO CONCEALER; PRO CONCEALER; FINELINE EYELINER; EYE
BEAUTY BRICK BLUSH; VELVET CONTOUR STICK
BEAUTY BRICK EYESHADOW
BEAUTY BRICK EYESHADOW; BEAUTY BRICK EYESHADOW

Last Values:
Zzz
ZZZ
ZZZ|Booking|psepayment|Pago de tiquete flight FC date [Z
ZZZKZA
ZZZZTA

Se puede observar que hay datos inconsistentes o de poco valor para ser analizados.

Frequency Analysis

Value	Count	Percentage
	361268	3,05%
CC	1154094	9,74%
Pago de Saldo	513976	4,34%
Empresas Publicas de Medellin ESP	433147	3,65%
Transaccion_para_generacion_de_certificados_de_tradicion_y_libertad	282806	2,39%
CPV	276882	2,34%
Referencia: Contrato: Valor:	267165	2,25%
Pago de factura	251686	2,12%
NI	235812	1,99%
Ref pago express No |	230786	1,95%
Pago de factura Postpago	228912	1,93%
pagos	227021	1,92%
Recarga Nequi PSE	225938	1,91%
Pago a traves de PSE	175757	1,48%
Pago por payU Falabella	154521	1,30%
Pago facturas: CMR	142479	1,20%
NIT	138851	1,17%

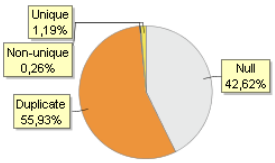
Estas son las referencias que mas aparecen en las transacciones, se puede observar que hay descripciones que no generan valor para el objetivo aproximadamente el 20%.

Mask Analysis

Value	Count	Percentage
	361268	3,05%
LL	1408778	11,88%
LLLL LL LLLLL	515094	4,35%
LLL	452259	3,82%
LLLLLLLL LLLLLLLL LL LLLLLLLL LLL	433147	3,65%
LLLL LL LLLLLLL LLLLLLLL	307528	2,59%
LLLLLLL	283588	2,39%
LLLLLLLLLLL_LLLL_LLLLLLLLLL_LL_LLLLLLLLLLLL_LL_LLLLLLLLL_L_LLLLLLLL	282806	2,39%
LLLL LL LLLLLLL	279500	2,36%
LLLLLLLLL: LLLLLLLL: LLLLL:	267165	2,25%
LLLLL	258330	2,18%
LLL LLLL LLLLLLL LL |	230786	1,95%
LLLLLLL LLLLL LLL	225957	1,91%
LLLL L LLLLLL LL LLL	175757	1,48%
LLLL LLL LLLL LLLLLLLLL	154523	1,30%
LLLL LLLLLLLL: LLL	142537	1,20%
LLLL LLLLLLL LLLLL LLLLLLLLL	125500	1,06%
LLLLLLLLL LL LLLL LLLLLLL LL |	120521	1,02%

Se puede observar referencias compuestas de dos y tres letras, se puede deducir que las referencias que tengan esta composición no aportan valor al negocio, aproximadamente el 20 % de la información.

- **Ref2**: Es un campo tipo String



Counts

Type	Count
Non-null values:	6.801.921
Null values:	5.051.861
Distinct values:	171.900
Duplicate values:	6.630.021
Unique values:	141.210
Non-unique values:	30.690

El 42% de la información es nula, hay 141.210 referencias que solo aparecen 1 vez, y en la información hay 171.900 referencias diferentes.

Extremes

First Values:
´
´ ETicket Avianca VXITIB
_ ADMIN@MAGRAVISCOM
_ ADMINISTRACION@SOYLATINLOVERCOM
_ ANAMARIALACOUTURE@GMAILCOM

Last Values:
ZZU null
ZZV null
ZZW null
ZZX null
ZZY null

Se percibe que en el campo hay información que no aporta valor.

Frequency Analysis

Value	Count	Percentage
	5051861	42,62%
CC	4681351	39,49%
IDC	803079	6,77%
TPNI	366847	3,09%
NO TIPIFICADO	207734	1,75%
NIT	127735	1,08%
BancoomevaPFA	61382	0,52%
cc	50714	0,43%
CE	37979	0,32%
Pago de obligacion:	18124	0,15%
* *	13749	0,12%
Cédula	13643	0,12%
TI	12728	0,11%
ID	11126	0,09%
CEDULA	10418	0,09%
Cedula	7522	0,06%
a CC	6565	0,06%
PP	4711	0,04%
BancoomevaPTC	4160	0,04%

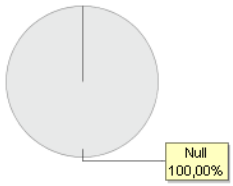
La mayoría de los campos son el tipo de identificación.

Quantiles

Percentage	Value
0,00%	´
10,00%	CC
20,00%	CC
30,00%	CC
40,00%	CC
50,00%	CC
60,00%	CC
70,00%	CC
80,00%	IDC
90,00%	NIT
100,00%	ZZY null

Se reafirma la hipótesis anterior, el campo no aporta gran valor, algunos casos como el de Matricula, podrían aportar valor, pero son una minoría (87).

- **Ref3:** Este campo es String.

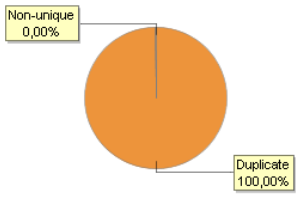


Counts

Type	Count
Non-null values:	0
Null values:	11.853.782
Distinct values:	0
Duplicate values:	0
Unique values:	0
Non-unique values:	0

No aporta valor, dado que todos los campos son nulos, esta columna se elimina del dataset.

- **Sector:** Este campo es de tipo String.



Counts

Type	Count
Non-null values:	11.853.782
Null values:	0
Distinct values:	11
Duplicate values:	11.853.771
Unique values:	0
Non-unique values:	11

Es un campo que no tiene valores nulos, hay 11 valores distintos.

Frequency Analysis

Value	Count	Percentage
\N	8546397	72,10%
MEDIOS DE COMUNICACION	1173421	9,90%
SERVICIOS FINANCIEROS	1061605	8,96%
GOBIERNO	520126	4,39%
RECURSOS NATURALES	442490	3,73%
SERVICIOS NO FINANCIEROS	69240	0,58%
CONSTRUCCION	25753	0,22%
PERSONAS	5238	0,04%
COMERCIO	4748	0,04%
MANUFACTURA INSUMOS	3655	0,03%
AGROINDUSTRIA	1109	0,01%

Este grafico nos muestra que el 72% de la información no aporta valor, a pesar de no existir valores nulos, se pude concluir que no se identifica el beneficiario para el 72% de la información en transacciones PSE.

- **Subsector:** Este campo es de tipo String.

Frequency Analysis

Value	Count	Percentage
\N	8546397	72,10%
BANCOS	1041893	8,79%
TELEFONIA FIJA	623771	5,26%
VALOR AGREGADO	548028	4,62%
ELECTRICIDAD	442489	3,73%
ADMINISTRACIÓN CENTRAL	331832	2,80%
MUNICIPIOS	95742	0,81%
SERVICIOS A EMPRESAS	45479	0,38%
CAJAS DE COMPENSACIÓN	39995	0,34%
ESTABLECIMIENTOS EDUCATIVOS	39611	0,33%
SERVICIOS A PERSONAS	22818	0,19%
OBRAS DE INFRAESTRUCTURA	18572	0,16%
EPS Y SALUD PREPAGADA (SALUD)	11759	0,10%
OTROS SERVICIOS FINANCIEROS	10886	0,09%
SEGUROS	8357	0,07%
TRANSPORTE TERRESTRE	5336	0,05%
PERSONAS	5238	0,04%
COMERCIO DE VARIEDADES Y VESTUARIO	3603	0,03%

Presenta el mismo comportamiento del campo de sector, aunque este es mas detallado. Si se pudiera identificar mas beneficiarios en el set de datos, este campo sería de gran valor, para la clasificación.

- **Descripción:** Este campo es String.

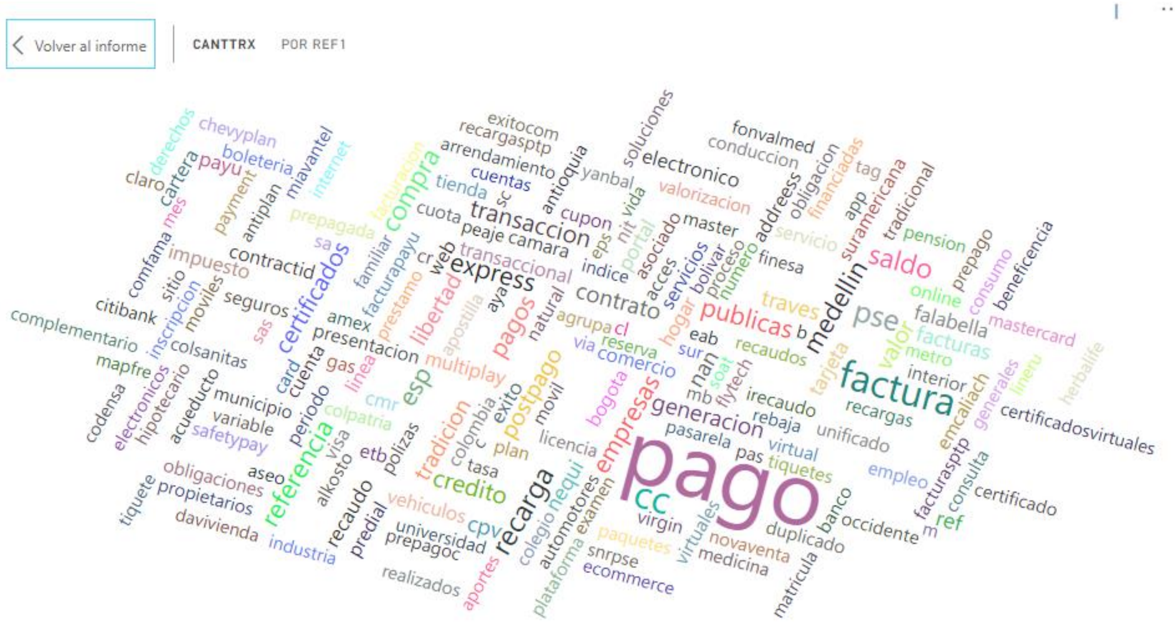
Frequency Analysis

Value	Count	Percentage
\N	8546397	72,10%
Bancos comerciales	1041893	8,79%
Servicios de telefonía fija	623771	5,26%
Actividades de telecomunicaciones inalámbricas	547971	4,62%
Generación de energía eléctrica	442489	3,73%
Actividades ejecutivas de la administración pública en gobierno central	324732	2,74%
Actividades ejecutivas de la administración pública en municipios	94690	0,80%
Actividades de seguridad social de afiliación obligatoria	39995	0,34%
Actividades de apoyo a la educación	21135	0,18%
Construcción de carreteras y vías de ferrocarril	18551	0,16%
Educación técnica profesional	13689	0,12%
Establecimientos que combinan diferentes niveles de educación	12774	0,11%
Actividades de seguridad social de afiliación obligatoria Entidades promotoras de salud	11759	0,10%
Procesamiento de datos	10584	0,09%
Actividades de asociaciones empresariales y de empleadores	8121	0,07%
Seguros generales	7261	0,06%

Presenta el mismo comportamiento de sector y subsector, 72% no se identifica, aunque es muy detallado este campo con el subsector se puede trabajar algunos aspectos de la calidad de la información.

subsector	descripcion	ref1	canttrx	valor_trx	%
\n	n	cc	1.076.633	100.573.603.479.985.700.000,00	
telefonia fija	servicios telefonia fija	pago saldo	511.630	30.126.934.538.072.760.000,00	
electricidad	generacion energia electrica	empresas publicas medellin esp	427.106	17.237.352.151.040.190.000,00	
\n	n	pago factura postpago	306.362	12.515.555.110.277.850.000,00	
\n	n	pago factura	291.949	15.719.289.468.978.920.000,00	
administraciÃ³n central	actividades ejecutivas administracion publica gobierno central	transaccion generacion certificados tradicion libertad	281.970	792.007.170.503.258.400,00	
\n	n	referencia contrato valor	277.596	10.152.034.810.189.800.000,00	
\n	n	nan	266.037	19.110.979.895.565.650.000,00	
bancos	bancos comerciales	cpv	254.409	17.501.731.731.988.080.000,00	
valor agregado	actividades telecomunicaciones inalambricas	ref pago express	229.869	3.203.038.778.004.179.000,00	
bancos	bancos comerciales	recarga nequi pse	217.983	8.408.967.178.048.729.000,00	
\n	n		208.626	21.336.612.130.941.270.000,00	
valor agregado	actividades telecomunicaciones inalambricas	referencia pago express	184.345	9.723.193.188.489.050.000,00	
\n	n	pagos	180.481	24.009.856.180.749.380.000,00	
\n	n	pago traves pse	173.223	5.848.203.732.647.762.000,00	
\n	n	pago payu falabella	134.555	11.342.169.378.242.000.000,00	
\n	n	pago facturas cmr	125.524	10.712.887.825.911.650.000,00	
\n	n	pago factura hogar multiplay	125.079	6.610.661.536.018.252.000,00	
\n	n	pago pse portal transaccional exito	105.104	8.091.739.565.853.213.000,00	
\n	n	nit	104.594	11.396.756.815.873.170.000,00	
\n	n	factura	101.235	3.615.780.047.872.017.000,00	
\n	n	cartera	91.905	6.029.429.483.129.838.000,00	
\n	n	pago suramericana	90.609	5.084.400.586.248.893.000,00	
			10.906.328	**658.043.984.825.170.000.000,00**	

Se observa que en el campo ref1 donde la descripción es CC, no se puede categorizar ya que no aporta mucho valor, las referencias por las que más transan son telefonía y pagos de servicios públicos.



La palabra “pago” es la que mas se repite, esta palabra puede entrar en una lista de personalizada de STOPWORDS, al igual que los nombres de los meses, los días de la semana, los nombres de los municipios.

También se repite “públicas”, “servicios”, “Medellín” y “recargas”, se puede concluir que servicios públicos y recargas de celular, son los conceptos por los que más transan en PSE.

6. Preparación de los datos

En esta etapa se requiere el tratamiento de lenguaje natural, por lo tanto, se realiza un proceso de ETL (Extracción, transformación y carga) a los datos entregados, centrado principalmente en el campo de ref1, se limpia y normaliza, se eliminan vacíos, se eliminan caracteres especiales, se eliminan stop words, se pasan a minúscula los datos, se filtran outliers y aquella información que no aporte valor.

7. Modelamiento

En esta etapa se tienen dos modelos, **uno de aprendizaje supervisado y otro de aprendizaje no supervisado.**

Por el lado del aprendizaje supervisado se lleva a cabo el siguiente proceso:

- Se extrae un dataset, con los campos de sector, subsector y descripción, se imputa el label con la categoría de gasto, según el criterio del equipo, con base a la lista dada por el banco.
- En este dataset se excluye los registros donde:
- **Sector = ‘personas’**
- **Subsector in (‘bancos’, ‘cajas de compensación’)**
- **Sector = ‘servicios no financieros’ and subsector = ‘servicios a personas’ and descripción in (‘actividades asociaciones ncp’, ‘actividades servicios personales ncp’),** ya que estos pueden generar ruido en la clasificación de gastos personales, por ejemplo, en el subsector bancos, hay referencias que son pagos de administración de edificio (este set de datos será tratado con **clusterización**)
- Se implementa un clasificador con el método **Random Forrest (mejores resultados) y Multinomial Näives Bayes(**genera buenos resultados, pero es mejor Random Forrest**)**,

- Se toma 70% de los datos para entrenamiento y 30% para validación.
- Para evaluar el modelo se usa la matriz de confusión.
- Se usa TF-IDF, que permite identificar el nivel de similitud de las frases comparándola con el cluster.

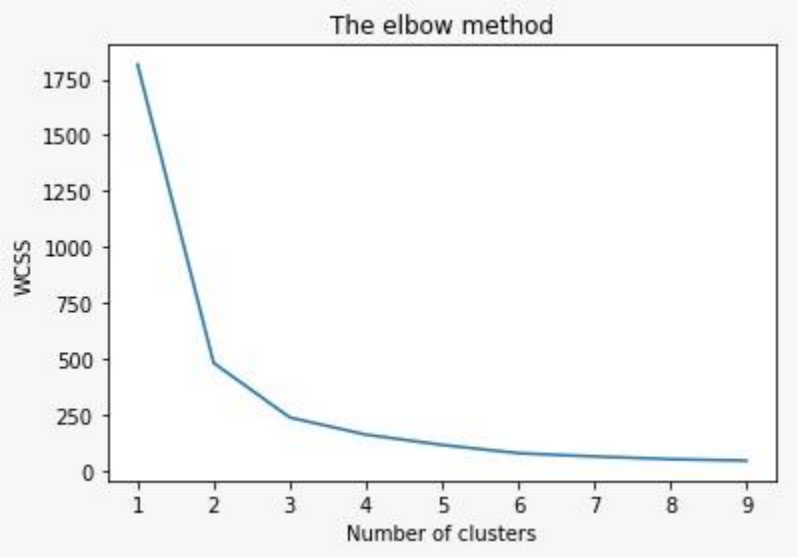
Se sugiere balancear los datos con muestras sintéticas, para corregir el sesgo que se presenta en el clasificador, por ejemplo, con metodologías como SMOTE.

Por el lado del aprendizaje no supervisado se lleva a cabo el siguiente proceso:

- Se extrae un dataset con el campo de referencia1, con los siguientes filtros:
- **Sector = 'personas'**
- **Subsector in ('bancos', 'cajas de compensación')**
- **Sector = 'servicios no financieros' and subsector = 'servicios a personas' and descripción in ('actividades asociaciones ncp', 'actividades servicios personales ncp'),**

Ya que estos pueden generar ruido en la clasificación de gastos personales, por ejemplo, en el subsector bancos, hay referencias que son pagos de administración de edificio.

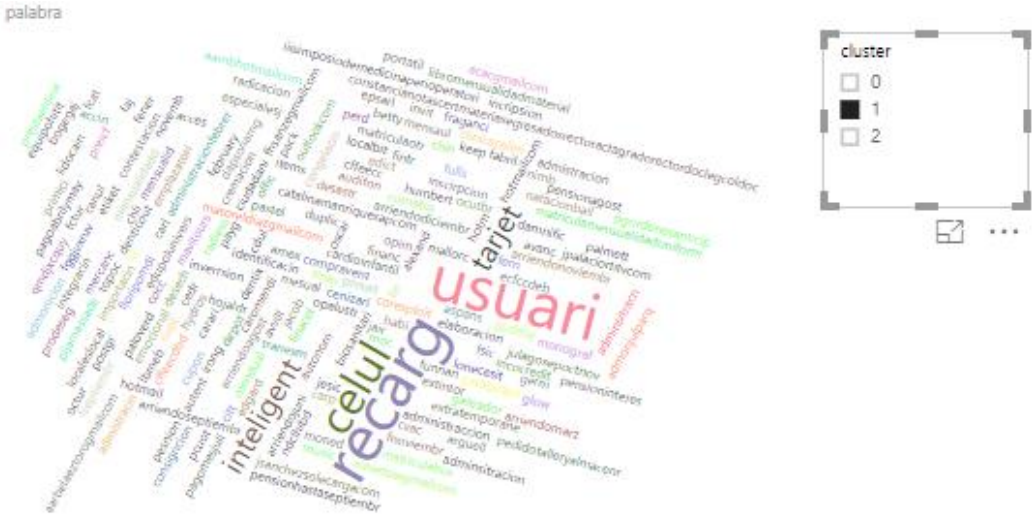
- Se implementa el algoritmo de **K-means**
- Se hacen pruebas reduciendo la dimensión con **TSNE**, dado que los datos eran no lineales, pero los resultados arrojados no fueron los esperados.
- Se vectoriza con **word2vec** de la librería gensim.
- Se selecciona el número óptimo de clúster, con **Elbow.**



Según la gráfica anterior generada por el método Elbow, el número adecuado de clusters es de 3

- Se evalúa el modelo con la silueta.
- Con **Power BI** se realizan nubes de palabra para darle una etiqueta a cada clúster.

Las palabras que más sobresalen son arriendo, administración, apto, este tipo de gastos se imputo con la etiqueta sugerida de **pago de deudas**.



La palabra que más sobre sale es el pago del celular, tarjet, inteligent, este tipo de descripciones se etiqueto en la categoría de **tecnología y telecomunicaciones**.

Este no se visualiza un patrón común de palabras se etiqueta como **otros**.

8. Evaluación

- **Para evaluar la clusterización se usa:**

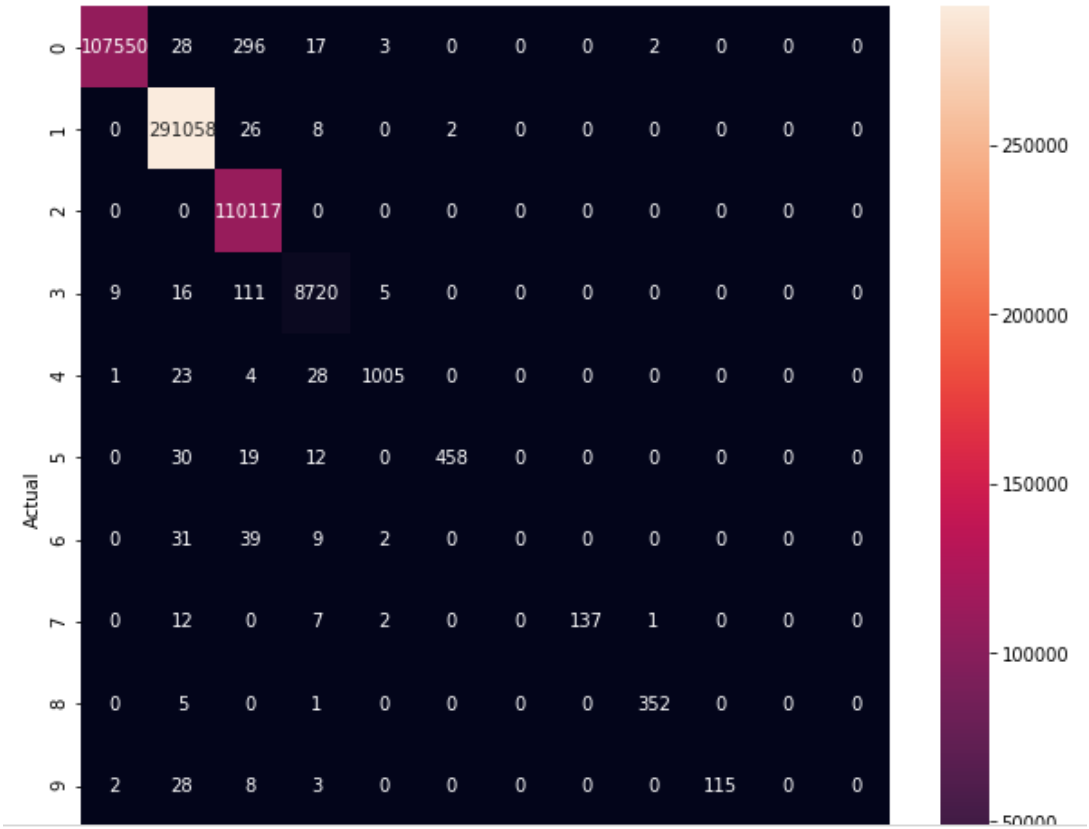
Silhouette_score:

0.58932287

Lo cual indica que el número de clusters es óptimo.

- **Para evaluar el clasificador se usa:** Matriz de confusión.

Por ambos métodos (Random Forrest y Multinomial Näives Bayes) genera una matriz similar.



La matriz arroja los resultados esperados y es que en la diagonal queden los datos más grandes, además los colores indican mayor concentración de datos para esas etiquetas, por tanto, esto puede generar sesgos de clasificación hacia esas etiquetas.

9. Despliegue

Esta etapa no aplica para el reto ya que solo existe un ambiente.

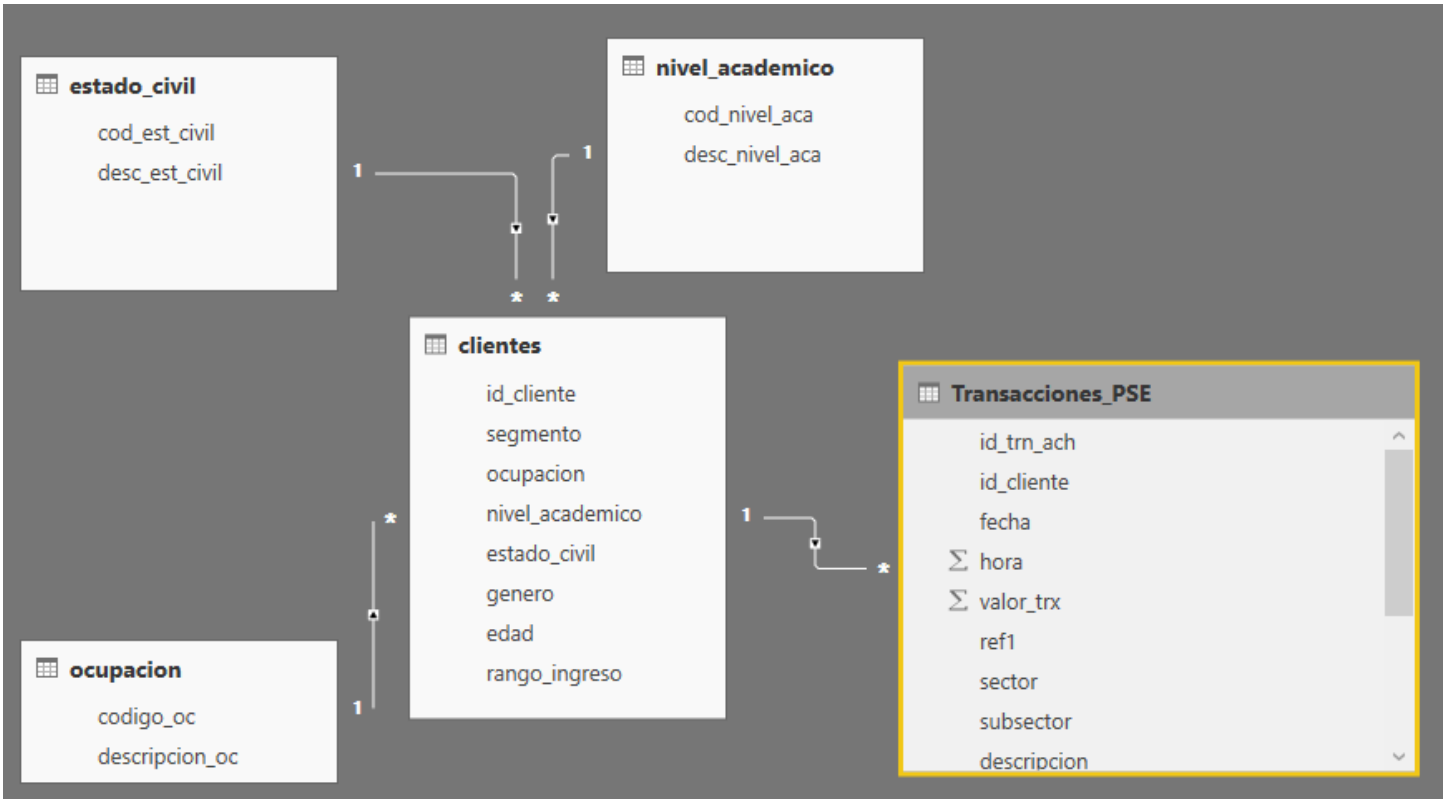
10. Retroalimentación

Al recolectar los resultados del modelo implementado, el banco obtiene retroalimentación sobre el rendimiento del modelo y observa como afecta su entorno de despliegue.

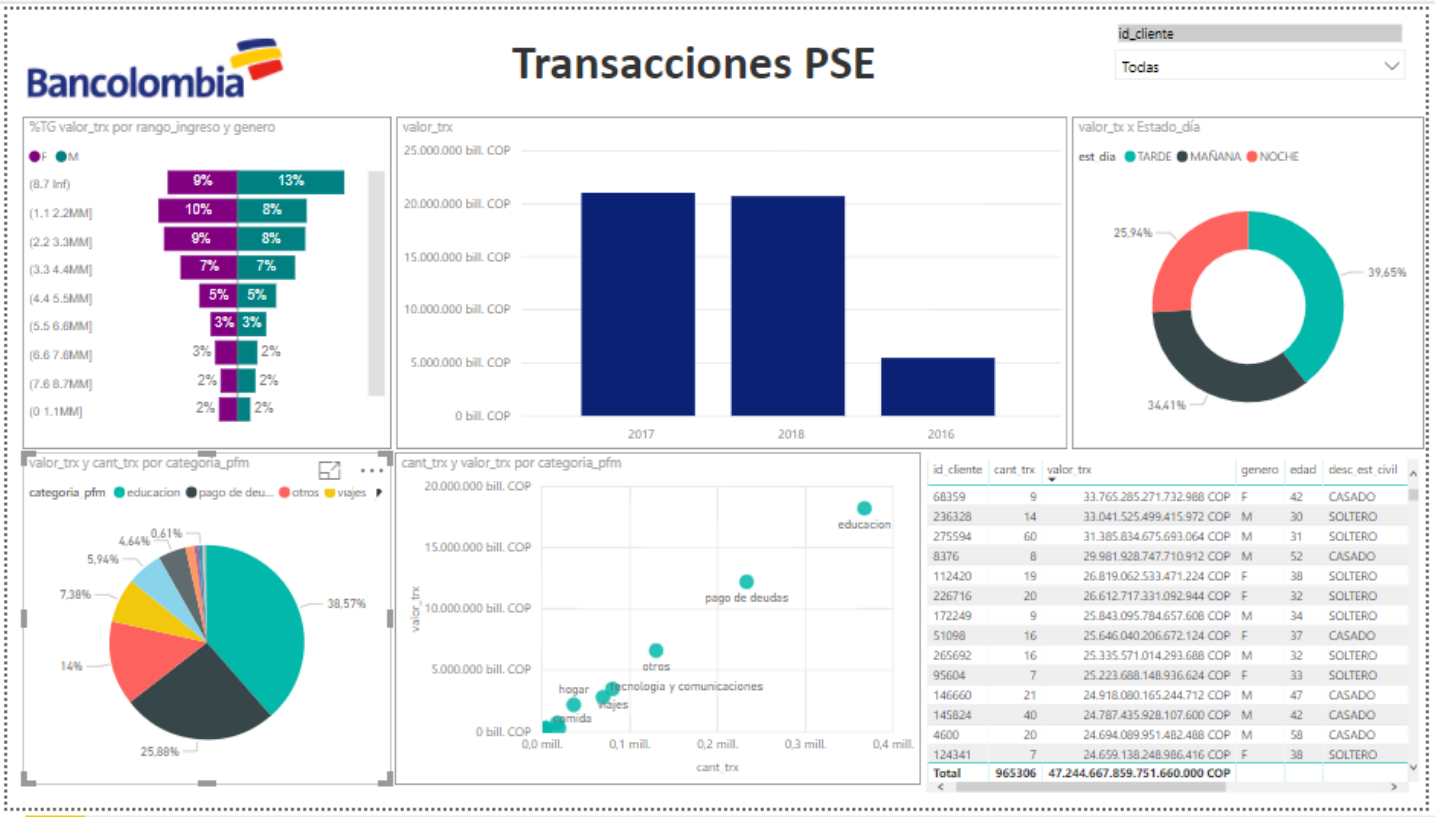
SOLUCIÓN EXTENDIDA

- Luego de exponer e implementar una solución al reto principal, se lleva a cabo la analítica descriptiva con la herramienta Power BI, de la cual se puede extraer gran conocimiento y valor, para conocer a mayor profundidad **el cliente** y sus **diferentes interacciones** con las diferentes variables suministradas.

Modelo de datos creado en Power BI

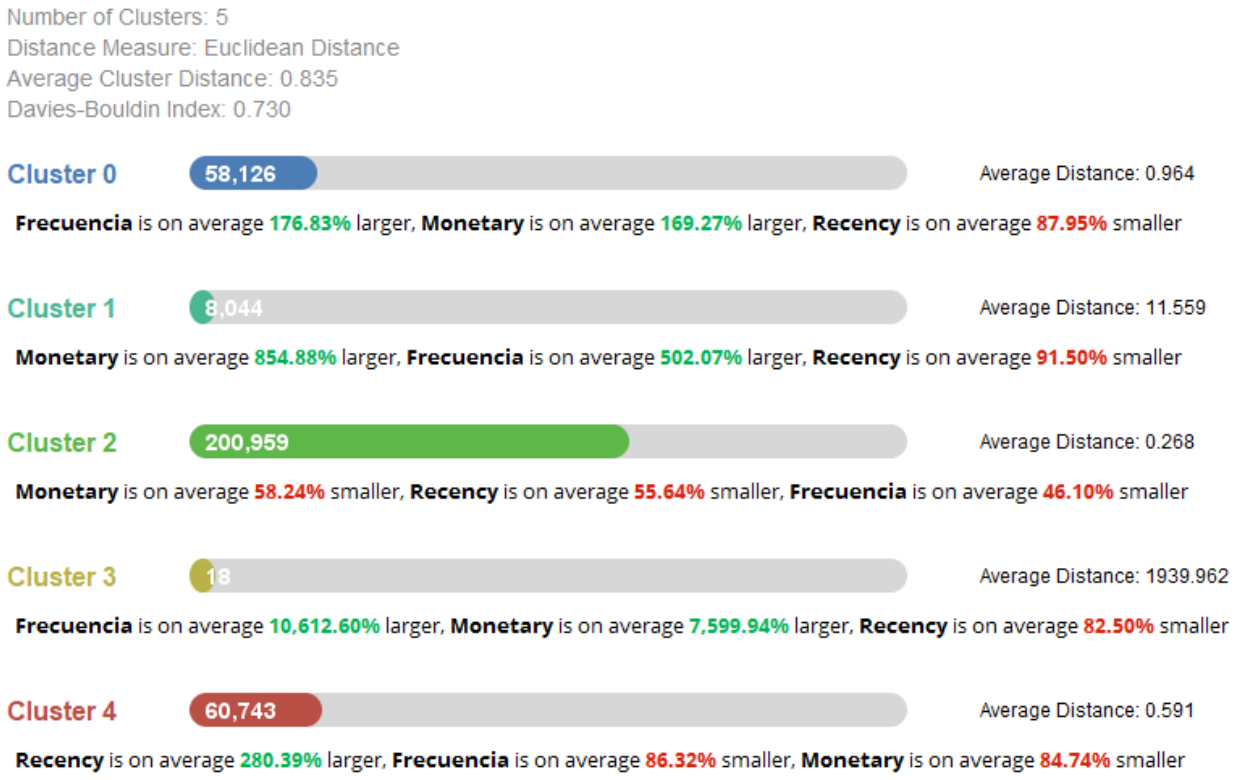


Dashboard desarrollado con Power BI



Nota: El tablero está disponible para ser consultado, no se almacena en el Github, por políticas del reto.

- Con la información disponible de las transacciones es posible identificar la frecuencia, la recencia y el monto total transado por los usuarios a través de la plataforma de PSE. Utilizando un modelo RFM y aplicando una clusterización por KMeans, se pueden agrupar los clientes según sus comportamientos de compra, por ejemplo, en el clúster 0 observamos todos los usuarios que suelen utilizar esta aplicación de forma frecuente, con montos muy altos y lo han hecho de forma reciente. Con este tipo de estudios junto con la etiqueta obtenida por medio los resultados anteriores, se le puede otorgar mayor valor agregado a las aplicaciones de PFM, según los hábitos de compra y los canales de transacciones de los clientes.



OTRAS IDEAS A IMPLEMENTAR

Se presentan otras ideas que se pueden desarrollar con los datos suministrados, pero que requieren de más tiempo para ser ejecutadas.

- Se pueden segmentar los clientes con base a las categorías en las que gastan (con el algoritmo de K-means), y con esto hacer comparaciones anónimas entre clientes, se pueden comparar el valor de los gastos de los clientes de un clúster, como los categorías de gastos, son similares ver que hace el que menos gasta, para hacer que todo ese clúster tienda a ese comportamiento y promover el ahorro de dinero a la hora de pagar o comprar, bienes y/o servicios.
- Con un algoritmo de regresión, se puede predecir cual sería el gasto de cada cliente, esto se podría ver por categoría.
- Se podría implementar reglas de asociación en las transacciones (algoritmo a priori), agrupando la información por mes, para encontrar patrones ocultos y generar nuevas estrategias o ideas, enfocadas en el cliente.
- Conociendo cuales son aquellas categorías en las cuales gasta mas el cliente, se puede implementar un sistema de recomendación (con KNN), el cual permita al cliente adquirir los mismos bienes o servicios, pero ahorrando dinero, del mismo modo se pueden sugerir establecimientos que sean clientes de Bancolombia y así mantener el flujo de dinero en el banco.
- Para enriquecer los datos se puede contar con fuentes externas como lo son las fuentes del DANE, para tener identificados en su totalidad los nit del beneficiario y así hacer que el clasificador sea mas exacto, ya que se clasificaría por sector, subsector, descripción y referencia.
- También se puede enriquecer la data de los clientes con la información de las redes sociales, para generar un mejor perfil de los clientes, sus interacciones y sus sueños.
- Se puede adicionar información de todas las transacciones (gastos e ingresos del cliente) que pasan por el banco, y **categorizar mediante analítica**, adicional se pude estudiar la viabilidad de tomar la foto de la factura de los gastos en efectivo y que se incluyan en la aplicación de PFM, ya que una de las debilidades de estas herramientas es tener que ingresar manualmente estos valores.