Carolina Flórez - Stiven López - Andrés González

Integrantes del Equipo

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.

Ensayo dataton

Equipo Datalyticos

**Tabla de contenido**

* Objetivos
* Fases ASUM-DM
  + Entendimiento del negocio
  + Enfoque analítico
  + Requisito de los datos
  + Recolección de los datos
  + Entendimiento de datos
  + Preparación de datos
  + Modelamiento
  + Evaluación
  + Despliegue
  + Retroalimentación
* Solución Extendida
* Otras ideas a implementar
* Conclusiones

**Ensayo Dataton 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.

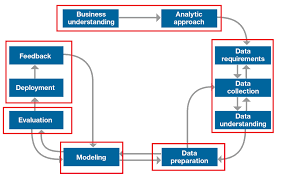
La realización del reto se llevará a cabo con **herramientas en su versión de uso libre**: DQ Analyzer, Python y Power BI.

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 maquinas se quedan cortas a la hora de procesar y almacenar, grandes cantidades de datos -> Big data, se deben considerar arquitecturas robustas que apoyen y solucionen esto), además aborda la parte de gestión de proyectos, indispensable para alcanzar cualquier proyecto (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.

1. **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.

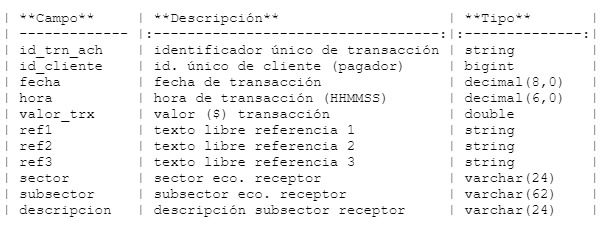
1. **Requisito de los datos**

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.

1. **Recolección de los datos**

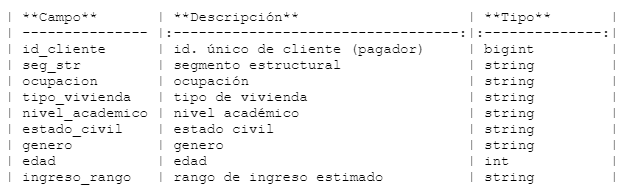
* **dt\_trxpse\_personas\_2016\_2018\_muestra\_adjt**

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



* **dt\_info\_pagadores\_muestra**

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



1. **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.

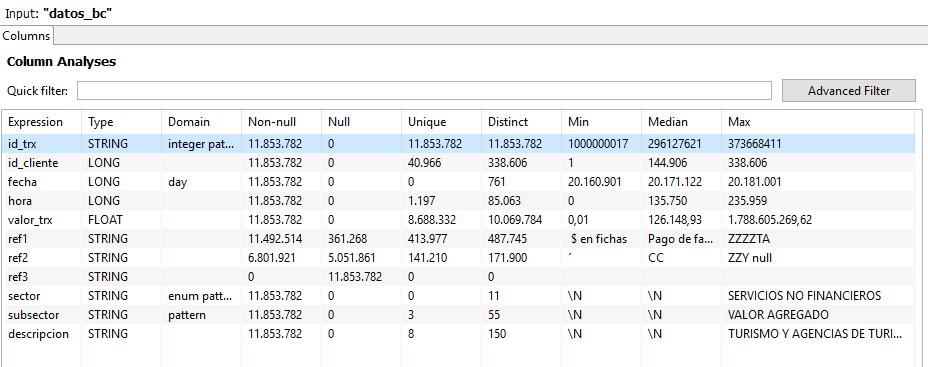
**Archivos análisis exploratorio en Python, análisis exploratorio en Python:**

* **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:**



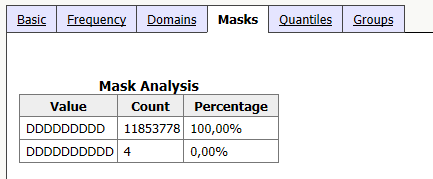
Se tienen 11.853.782 transacciones, en un periodo de tiempo comprendido entre: el 01 de septiembre de 2016 y 01 octubre del 2018, con un total de clientes analizados de: 338.606, con un valor mínimo de transacción de 0,01, y un máximo de 1.788.605.269, se tiene en el campo estrella ref1 361.268 valores 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:**

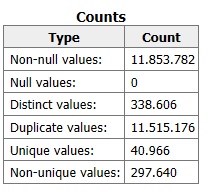
* **Id\_trx:** Es de tipo String

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.

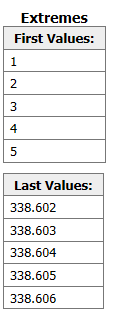


La mayoría de los registros tienen 9 dígitos.

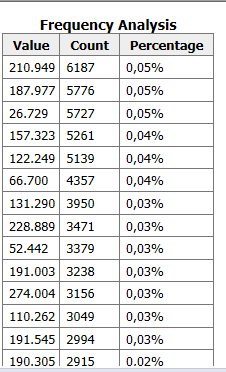
* **Id\_Cliente:** es tipo Bigint**:**



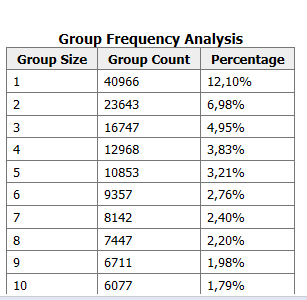
En este campo no hay valores nulos, se presentan transacciones de 338.606 clientes distintos, hay 40.966 clientes que solo aparecen una única vez, y 297.640 clientes transaron mas de una vez en el periodo de tiempo.



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

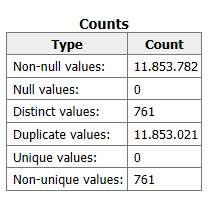


Estos son los id de clientes que mas transaron en ese periodo de tiempo.

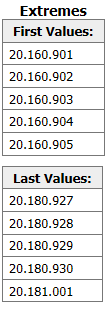


Esta tabla muestra que cantidad de id\_cliente transaron por PSE, 1 vez, 2 veces, 3 veces o más, es decir 23.643 id\_cliente, solo aparecen 2 veces en el set de datos.

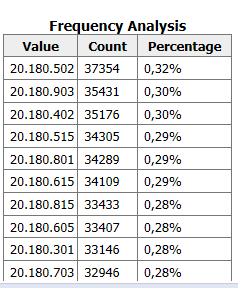
* **Fecha:** Es de tipo Bigint



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

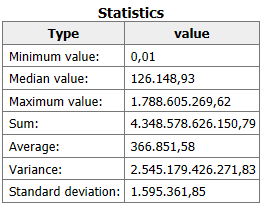


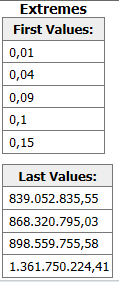
Las fechas están entre 20160901 y 20181001



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

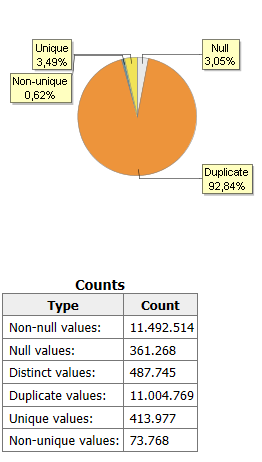
* **Valor\_trx:** esta variable es de tipo float



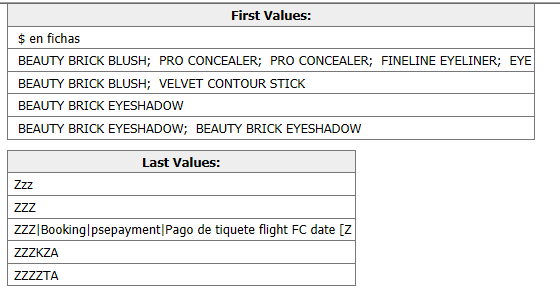


A simple vista se detectan valores atípicos tanto en los máximos como en los mínimos, en la definición de negocio se consideran transacciones atípicas de más de 50 millones.

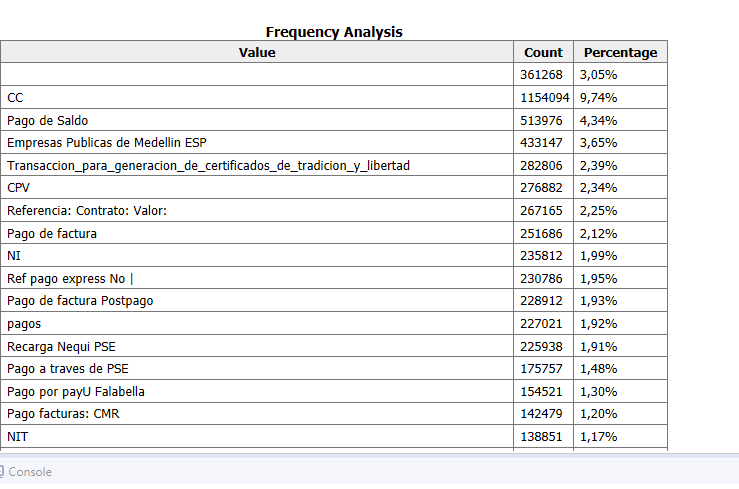
* **Ref1**: Es de tipo String



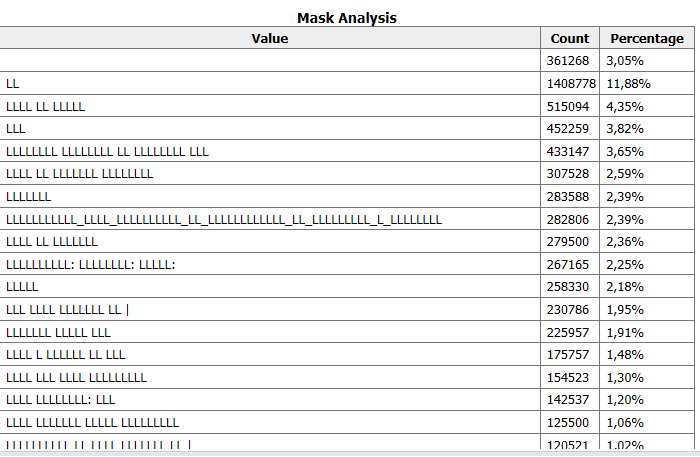
En el total de los datos hay 361.268 valores nulos, hay 487.745 referencias distintas sobre estas se debe hacer la segmentación, y 413.977 referencias que solo aparecen 1 sola vez en el set de datos, también se difiere que hay 73.768 referencias que aparecen más de 1 vez.



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

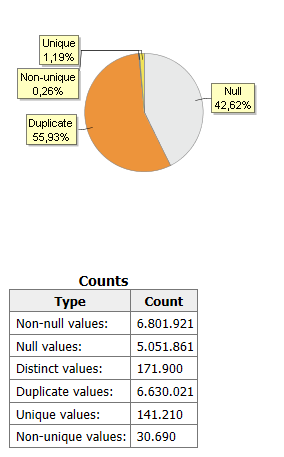


Estas son las referencias que mas aparecen en las transacciones, se puede observar que hay descripciones que no generan valor para el objetivo aproximadamente el 20%.

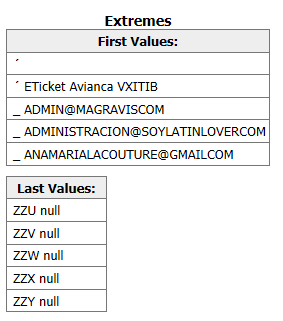


Se puede observar referencias compuestas de dos y tres letras, se puede deducir que las referencias que tengan esta composición no aportan valor al negocio, aproximadamente el 20 % de la información.

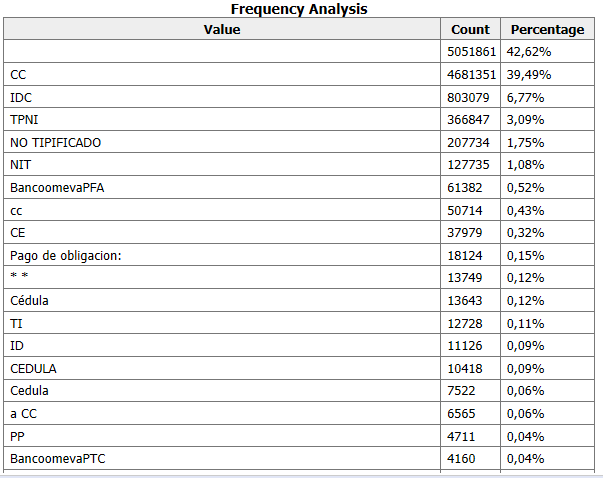
* **Ref2**: Es un campo tipo String



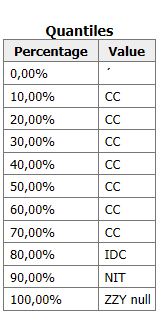
El 42% de la información es nula, hay 141.210 referencias que solo aparecen 1 vez, y en la información hay 171.900 referencias diferentes.



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

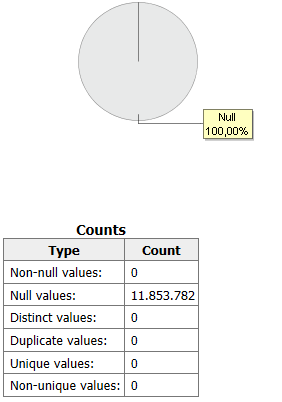


La mayoría de los campos son el tipo de identificación.



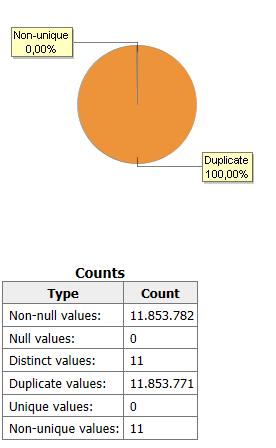
Se reafirma la hipótesis anterior, el campo no aporta gran valor, algunos casos como el de Matricula, podrían aportar valor, pero son una minoría (87).

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

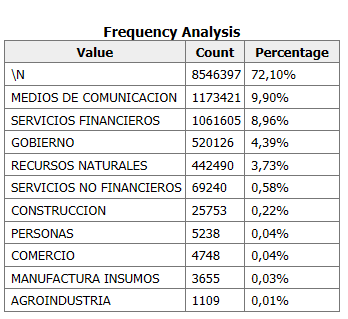


No aporta valor, dado que todos los campos son nulos, esta columna se elimina del dataset.

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

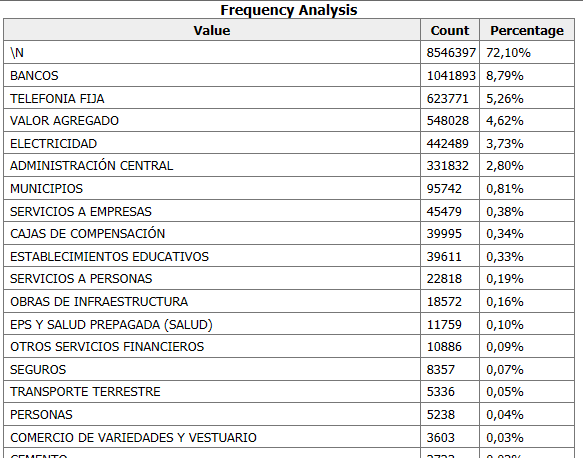


Es un campo que no tiene valores nulos, hay 11 valores distintos.



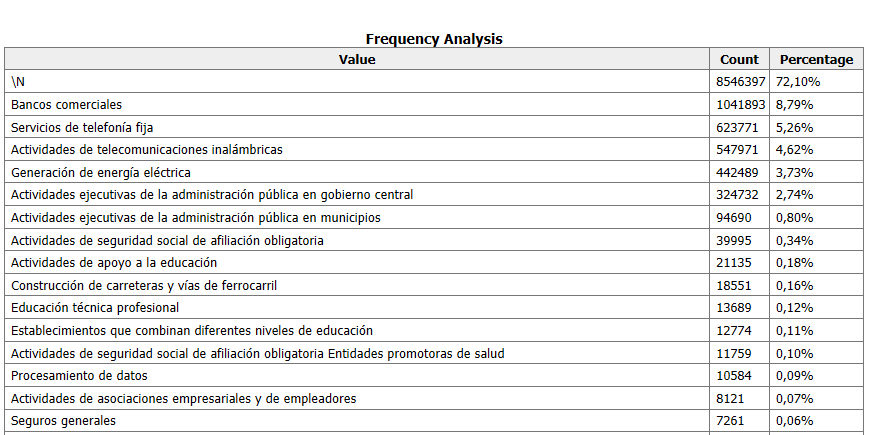
Este grafico nos muestra que el 72% de la información no aporta valor, a pesar de no existir valores nulos, se pude concluir que no se identifica el beneficiario para el 72% de la información en transacciones PSE.

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



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.



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.



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.

1. **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.

1. **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.

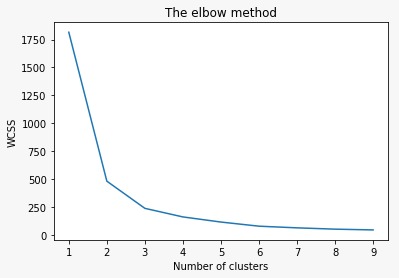
Se sugiere balancear los datos con muestras sintéticas, para corregir el sesgo que se presenta en el clasificador, por ejemplo, con metodologías como SMOTE.

**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’),**

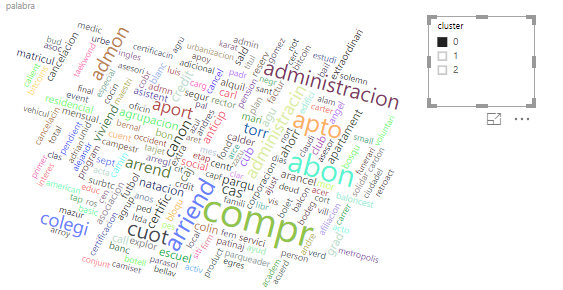
Ya que estos pueden generar ruido en la clasificación de gastos personales, por ejemplo, en el subsector bancos, hay referencias que son pagos de administración de edificio.

* 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.**

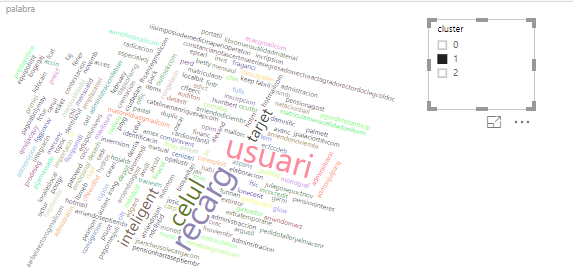


Según la gráfica anterior generada por el método Elbow, el número adecuado de clusters es de 3

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



Las palabras que más sobresalen es 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**.

1. **Evaluación**

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

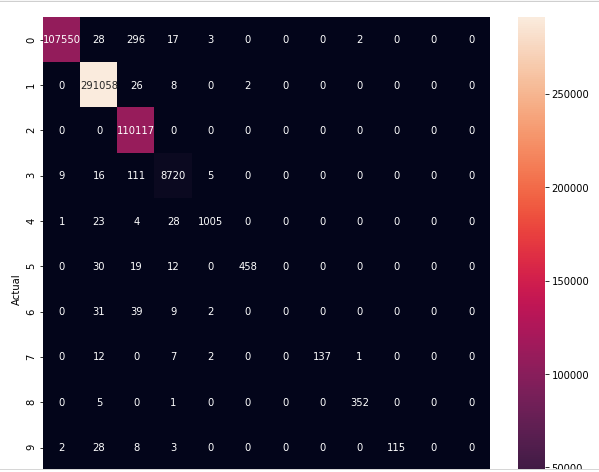
Silhouette\_score:

0.58932287

Lo cual indica que el número de clusters fue correctamente asignado.

* **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.

1. **Despliegue**

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

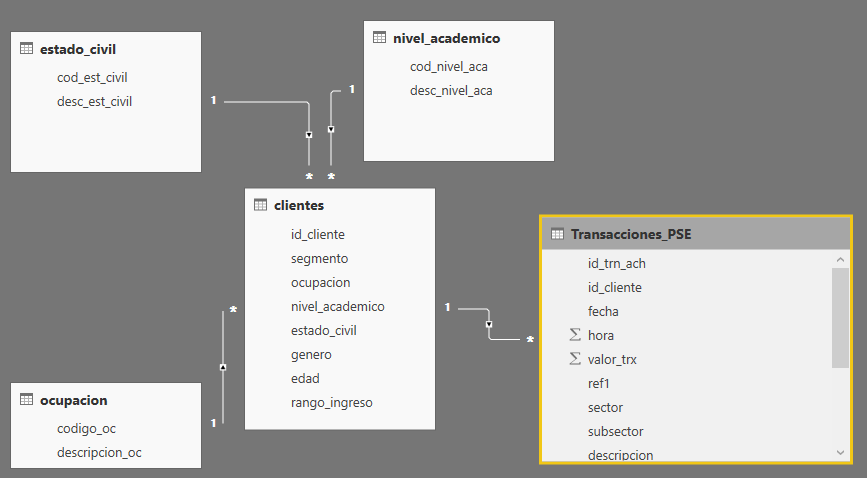
1. **Retroalimentación**

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

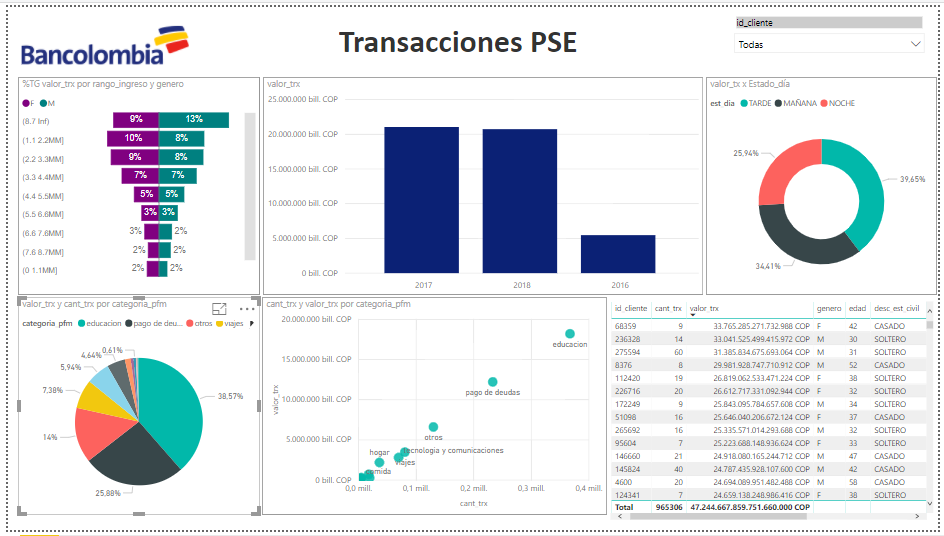
**SOLUCIÓN EXTENDIDA**

* 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**

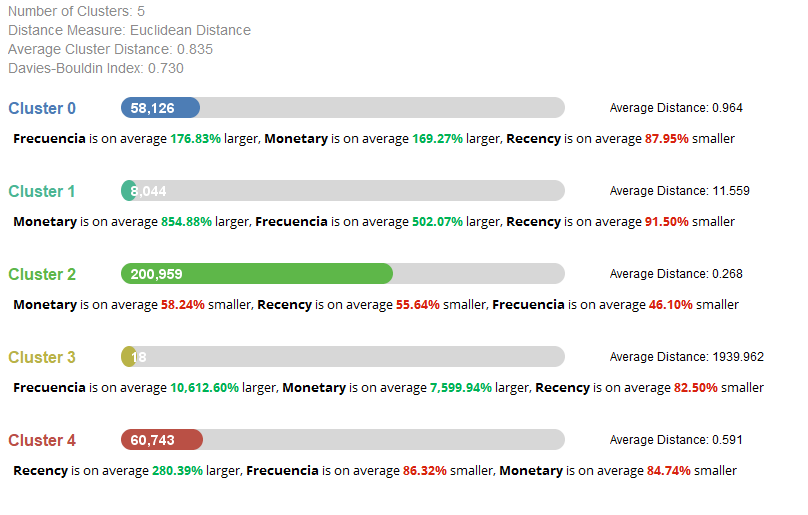


**Dashboard desarrollado con Power BI**



**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.