

# DATATON.BC – 2018

Informe de resultados

---

Equipo ZenAI:

- Ana Isabel Rúa: [ana.rua@eia.edu.co](mailto:ana.rua@eia.edu.co)
  - Sebastián Uribe: [sebasuribe07@gmail.com](mailto:sebasuribe07@gmail.com)
  - Julián Andrés Aranzales Salgado: [aranzales@gmail.com](mailto:aranzales@gmail.com)
-

## Contenido

Descripción del reto .....	3
Metodología de la Solución Propuesta .....	4
1. Entendimiento del Negocio.....	5
2. Clasificación de las categorías de las transacciones.....	7
2.1 Entendimiento de los datos .....	7
2.2 Asignación de categoría para registros sin información .....	11
2.3 Asignación de categoría por subsector .....	12
2.4 Asignación de categoría partir de bolsa de palabras por Web Scraping.....	12
2.5 Asignación de categoría a partir de bolsa de palabras manual .....	14
2.6 Asignación de categorías mediante clusters .....	16
2.7 Descripción de las categorías identificadas.....	20
2.8 Conclusiones asignación de categorías .....	22
3. Modelo Clasificador de Categorías.....	23
3.1 Conclusiones Clasificador de Categorías .....	25
4. Análisis de Pagadores (Clientes).....	25
4.1 Entendimiento de los datos de los pagadores (clientes) .....	25
4.2 Preparación de los datos .....	28
4.3 Modelamiento - Características de cliente por categoría de trx .....	31
5. Aplicación .....	33

## Descripción del reto

Los datos entregados en este reto corresponden a transacciones realizadas por clientes persona del banco vía PSE. Estas transacciones, a diferencia de las transacciones realizadas vía POS, no cuentan con un código MCC atado a la transacción, que permite conocer la categoría de comercio a la que pertenece el establecimiento de comercio donde se realiza la transacción. Adicionalmente, muchas de estas transferencias por PSE corresponden a transferencias de pagos de servicios públicos, seguros, colegios, arrendamientos, y otros gastos que pueden ser denominados como gastos grandes. En el marco de un sistema de gestión de finanzas personales, poder categorizar adecuadamente estas transacciones que se realizan por PSE es de suma importancia para contar con una foto completa de la actividad de gastos de los clientes. Para este reto, los equipos participantes tendrán acceso a una muestra de transacciones PSE que corresponden a algo más de 300 mil clientes (persona), seleccionados de manera aleatoria. La tabla de transacciones cuenta con 11.8 millones de registros (uno para cada transacción), realizados entre septiembre de 2016 y octubre de 2018.

En el Banco ya se han llevado a cabo esfuerzos por categorizar transacciones provenientes del canal POS (con tarjetas débito y crédito), lo cual ha incluido, entre otras cosas, una depuración y limpieza de los códigos MCC. A continuación, mostramos a manera de referencia la categorización propuesta por el equipo:

1. Comida
2. Hogar
3. Cuidado personal
4. Entretenimiento
5. Educación
6. Transporte
7. Viajes
8. Ahorro
9. Pago de deudas
10. Ingresos
11. Retiros en efectivo
12. Mascotas
13. Moda
14. Tecnología y comunicaciones
15. Otros

## **Metodología de la Solución Propuesta**

Para la solución a este reto se utilizará la metodología CRISP-DM, mediante la cual se busca generar un resultado que haga un uso adecuado de los datos y permita resolver de forma efectiva el problema inicial planteado. De esta forma, se seguirán los siguientes pasos:

1. Entendimiento del Negocio
2. Entendimiento de los Datos
3. Preparación de los Datos
4. Modelamiento
5. Evaluación
6. Despliegue

Se debe tener presente también que a partir del entendimiento del negocio se plantearán varias soluciones diferentes, que puedan aportar desde diferentes perspectivas tanto a los clientes del banco (pagadores) como al banco por sí mismo.

## 1. Entendimiento del Negocio

En esta fase se buscará entender el negocio en lo referente al funcionamiento del botón de Pagos seguros en línea PSE. Este es un sistema centralizado y estandarizado que permite a las empresas ofrecer al Usuario la posibilidad de realizar pagos en línea, accediendo a sus recursos desde la Entidad Financiera donde los tiene.

PSE es un servicio de ACH COLOMBIA S.A. quien es miembro de la Asociación Nacional de Cámaras de Compensación Automatizadas de Estados Unidos conocida como entidad que rige los procedimientos, normas y formatos de los ACH en ese país, donde el sistema ACH existe hace más de 25 años. En Colombia, PSE cuenta con más de 6000 empresas suscritas para la realización de pagos mediante el portal, y además un listado de entidades financieras vinculadas, como:

- BANCO AGRARIO
- BANCO AV VILLAS
- BANCO CAJA SOCIAL
- BANCO COLPATRIA
- BANCO DAVIVIENDA
- BANCO DE BOGOTA
- BANCO DE OCCIDENTE
- BANCO GNB SUDAMERIS
- BANCO PICHINCHA S.A.
- BANCO POPULAR
- BANCO PROCREDIT
- BANCOLOMBIA
- BANCOOMEVA S.A.
- BBVA COLOMBIA S.A.
- CITIBANK
- ITAÚ
- BANCO FALABELLA

Algunas de las ventajas para las grandes, medianas y pequeñas empresas son:

- Confirma e identifica en línea y en tiempo real las transacciones.
- Concilia automáticamente la información.
- Ahorra gastos operativos, tiempo y recursos.
- Evita errores en pagos y/o recaudos.
- Acceso a 18 millones de cuentas corrientes/ahorros en 17 entidades financieras.
- Aumento en los niveles de recaudo.
- Descongestiona los puntos de atención

Y en cuanto a los usuarios, algunas de las ventajas son:

- Permite hacer transacciones sin moverse de su hogar u oficina.
- Brinda seguridad y agilidad al reducir el manejo de efectivo.
- Facilita y ofrece comodidad en sus pagos y/o compras.

- Disponible las 24 horas de día, 7 días a la semana y todos los días del año.
- Confirma en línea las transacciones

Información consultada en línea en: <https://bit.ly/2AuCbcY>

El reto que se tiene es lograr clasificar las transacciones en una serie de categorías definidas por el banco, para lo cual el proceso que se seguirá es:

- **Categorización de las transacciones**, para lo cual se utilizarán varias metodologías asociadas a minería de texto para asignar una categoría a cada transacción. Para esto se utilizarán metodologías como: Aplicación de relación campo categoría (específicamente para el campo "subsector"), Clasificación de transacciones que no tienen información, Web Scraping para obtener categorías de páginas que tengan información relevante, y uso de matrices TF para aplicarlas bolsas de palabras al set de datos, y finalmente clusterización.

Una vez clasificados los 12M de registros iniciales según las categorías definidas (se partió de las categorías propuestas por Bancolombia, y se agregaron dos adicionales), se procederá a generar un modelo que permita clasificar nuevas transacciones que entregue el banco, de tal forma que el proceso pueda llevarse a producción sobre nuevos registros, y se pueda generar valor tanto para el banco como para los clientes.

- **Categorización de los clientes**, para lo cual se hará uso de la información de los clientes para realizar un proceso de correlación que indique qué tipos de clientes son más propensos a realizar cada tipo de transacción, es decir, no analizar solo cuáles atributos son relevantes frente a cada clasificación, sino las categorías específicas de cada atributo.

Esta información se utilizará también para crear clusters que permitan agruparlos y a partir de ello poder implementar modelos que le permitan al banco generar valor mediante estrategias como cross-selling, y al cliente tener una base de comparación sobre la cual pueda recibir recomendaciones si su comportamiento de gasto es superior al de personas con características similares.

Finalmente, se proponen una serie de mockups que permitan llevar los modelos propuestos a la práctica. Aunque esto es algo que posiblemente requiera trabajos más profundos de co-creación, se hace esta aproximación para lograr que los modelos propuestos vayan más allá de solo el procesamiento de datos, y llegue a una propuesta de valor para el banco y sus clientes.

Cabe resaltar que todo el procesamiento de los datos estará acompañado de los respectivos procesos de exploración y preparación, para asegurar que los datos tengan una adecuada calidad y presentación, para que así puedan ser utilizados de la forma más precisa posible. Así mismo, se debe tener presente que se busca que el procesamiento sea eficiente, para que así se haga un óptimo uso de los recursos disponibles.

Todo el proceso se llevará a cabo utilizando a Python como herramienta, y a Google Colab como plataforma. Los entregables incluyen este informe, junto con el Jupyter Notebook y los archivos de datos complementarios utilizados como parte del proceso (no los archivos entregados por el banco).

## 2. Clasificación de las categorías de las transacciones

Para este proceso se hará la carga de los archivos con la información y posteriormente todo el procesamiento mediante diferentes técnicas para lograr categorizar adecuadamente la mayor cantidad posible de transacciones, con la premisa de lograr que esta clasificación sea confiable, es decir, si no hay certeza de que una transacción sea de determinado tipo, no se clasificará como tal.

Cabe aclarar además que como parte del proceso se definió agregar 4 categorías adicionales para explicar mejor las transacciones:

- Gobierno e impuestos
- Seguros
- Almacenes de cadena
- Otros servicios financieros

### 2.1 Entendimiento de los datos

---

*Para ver en detalle el análisis descriptivo de los datos, remitirse al documento [descriptivo.ipynb](#)*

---

Al explorar el dataframe asociado a las transacciones se encuentra que para los campos de texto asociados a las referencias, y a la categorización del receptor hay múltiples campos con valores NaN o valores \N, por lo cual se procede cambiar estos valores por cadenas vacías. Así mismo, se procederá a realizar una limpieza general del texto mediante procesos como:

- Convertir todas las cadenas a minúsculas
- Cambiar caracteres especiales por espacios
- Volver a convertir las cadenas a tipo "category"
- Cambiar las letras con tildes y otros acentos, por la letra base

Para esto se crea una función, de tal forma que pueda ser luego utilizadas en todos los sets de palabras que se utilicen en el ejercicio.

Por otro lado, y teniendo presente que "ref1", "ref2", "ref3", "sector", "subsector" y "descripcion" tienen una cantidad muy baja de valores únicos comparados con la cantidad de registros, se procede a mantenerlos como tipo "category" para optimizar el uso de memoria.

	ref1	ref2	ref3	sector	subsector	descripcion
count	11866506	11866506	11866506	11866506	11866506	11866506
unique	476162	169384	1	11	55	3
top	cc					
freq	1154116	5078404	11866506	8559121	8559121	11866212

Se encuentra además que ref3 está vacía para todos los registros, pero se toma la decisión de mantener el campo para que el modelo siga aplicando a pesar de que a futuro lleguen transacciones que tengan valores en este campo.

Una vez realizado un ajuste inicial a los valores de los campos de texto, se procede a hacer una conversión de las columnas de fecha y hora para pasarlos de tipo "object" a tipo datetime. Durante este proceso se encuentra que hay 12724 filas que tienen valor de fecha nulo, aunque se identifica que también tienen inconvenientes en el campo valor\_trx pues este es NaN o tiene el valor de \N, lo cual no corresponde al tipo de dato que debe tener este campo. A partir de esto se toma la decisión de eliminar estos registros, con lo cual el dataset queda de 11.853.782 registros.

También se encuentra como parte del proceso que no todos los valores son de 6 dígitos como debería ser pues su formato se indica como HHMMSS. Esto suele pasar cuando este tipo de valores pasan a ser asignados a un campo numérico y los que comienzan por 0 pierden caracteres, por ejemplo, las 000102 que corresponde a las 00 de la mañana, con 1 minuto y 2 segundos se convierte en 102. Para mitigar esto se procede a agregar ceros a la izquierda para complementar la información y asegurar que todos los valores en este campo queden de 6 dígitos.

Finalmente, y para tener mayor flexibilidad con la información al momento de hacer los diferentes procesamientos, se crean cuatro nuevas columnas:

- Year
- Month
- DayOfMonth
- DayOfWeek
- Hour

De esta forma el dataframe queda:

id_trn_ach	id_cliente	valor_trx	ref1	ref2	ref3	sector	subsector	descripcion	DateTime	Year	Month	DayOfMonth
230435642	3	2122392.51	cc						2016-12-07 11:34:51	2016	12	7
222356110	10	148438.37	referencia contrato valor	cc					2016-10-16 00:34:24	2016	10	16
309137749	10	94025.19	cc						2018-01-20 19:50:42	2018	1	20
324614737	10	94430.070000000001	cc						2018-03-26 19:21:46	2018	3	26
235344690	18	670645.56999999999	medicina prepagada colsanitas	ce					2017-01-06 20:13:17	2017	1	6
295760261	338583	104534.5	eticket avianca tizi	eticket avianca tizi					2017-11-20 20:50:11	2017	11	20

Posteriormente, se procede a eliminar la columna "id\_trn\_ach" que no es útil para el ejercicio, y se crea una nueva columna de tipo categórico con una cadena vacía que indica la categoría asociada a la transacción. Ésta se irá poblando en la medida en que logre ser relacionada a través de los métodos que se desarrollan en las siguientes secciones.



Finalmente, se hace un procesamiento de la información asociada a las stopwords, que se utilizará para el proceso de minería de texto y que también involucra palabras, de tal forma que a estas se les aplique el mismo proceso de limpieza que se siguió con el dataframe de las transacciones.

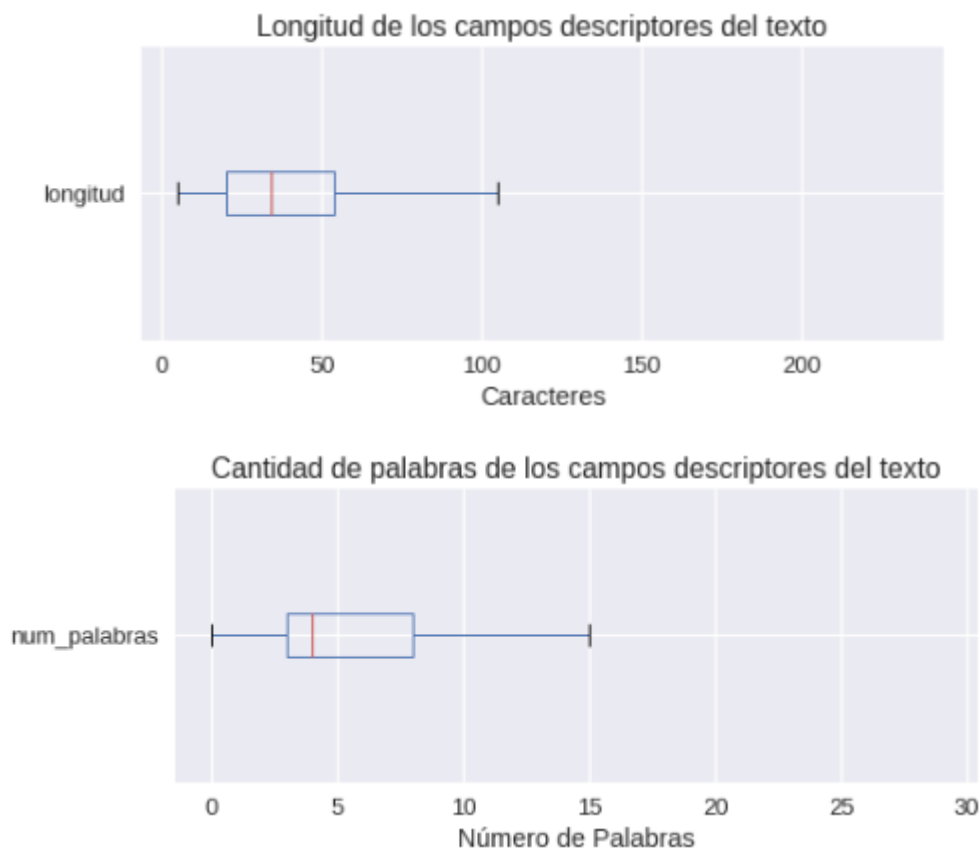
Muestra de algunas stopwords:

```
['a', 'a', 'aca', 'ademas', 'ademas', 'ahi', 'ajena', 'al', 'algo', 'algun']
```

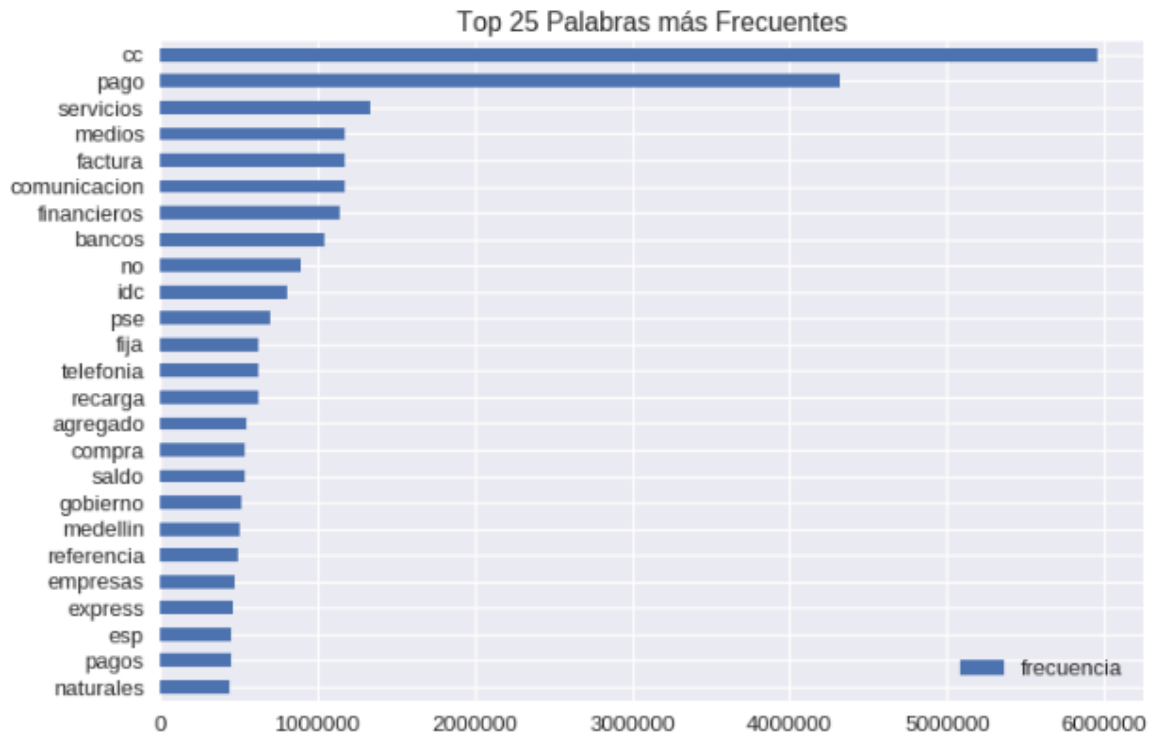
Ya con el set de datos de las transacciones preparado, se comienza a realizar el procesamiento de esta información para asignar la categoría mediante los siguientes pasos:

- Manualmente se asigna una categoría para aquellas transacciones en las que hay un valor de sector y subsector que está claramente definido y se puede asignar a una de las 15 categorías propuestas.
- Para el resto de los registros, se hará un proceso de minería de texto mediante el cual se concatenen los 6 campos de definición de la transacción, y luego se realice un proceso de preparación y posterior procesamiento que incluye aplicar bolsas de palabras obtenidas mediante Web Scraping y otras generadas por consenso del equipo, así como generar clusters sobre los cuales se haga una revisión de términos para clasificarlos en las categorías propuestas.

En primera instancia se encuentra que al concatenar los 6 campos con información de la transacción, y luego de la limpieza que se había hecho de los mismos, se obtiene la siguiente distribución de cantidad de caracteres y de cantidad de palabras:



Para la exploración inicial del texto se utilizará una matriz TF (Term Frequency) que permita identificar cuáles son los términos más y menos frecuentes, y a partir de ello hacer un proceso de preparación de los datos antes de proceder con los pasos siguientes. Se encuentra que los términos más frecuentes en todos el set de datos son:



AL revisar en detalle esta información se eucneutra que:

- El total de palabras identificadas en los textos eliminando las stopwords son 418.104
- De estas, solo aparecen hasta 100 veces un total de 413.742 palabras (únicas)

- Y aparecen máximo 11853 veces (0.1%) un total de 417825 palabras (poco frecuentes)
- Así mismo, hay un total de 653 que tienen hasta 2 caracteres

A partir de este análisis se identifican varios subsets de palabras asociados a:

- Palabras únicas: aparecen máximo 100 veces en los 12M de registros.
- Palabras poco frecuentes: aparecen máximo en el 0.1% de los 12M de registros.
- Palabras cortas: que son las que tienen máximo 2 caracteres

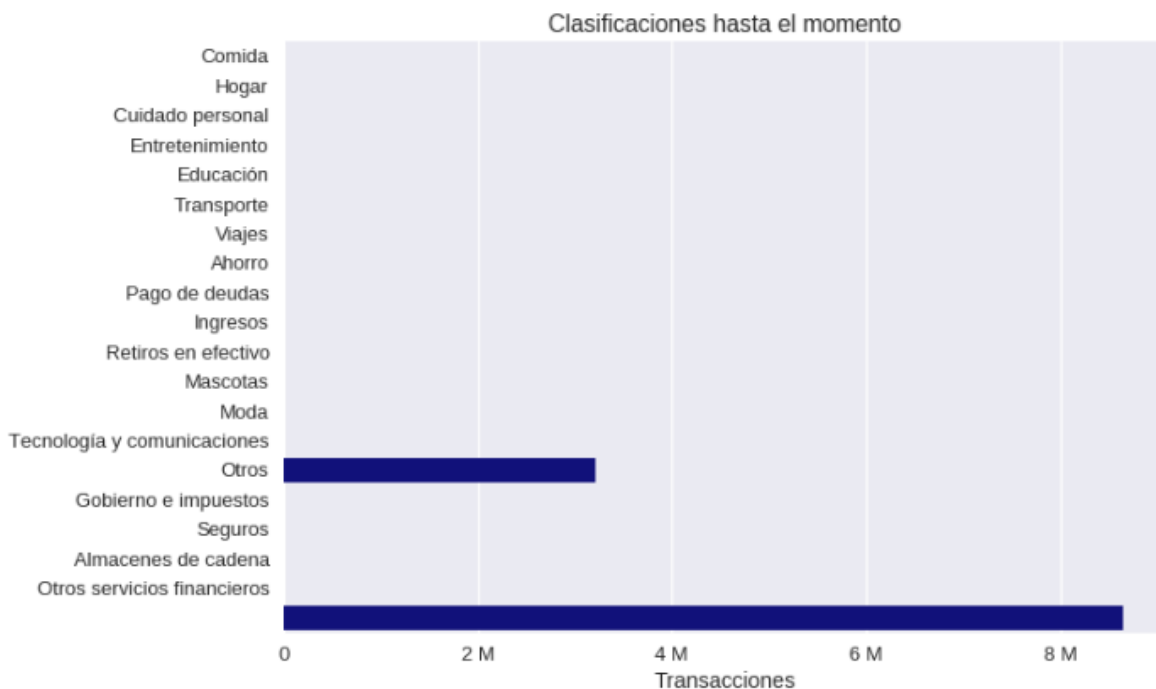
Se encuentra además que hay múltiples palabras que no agregan valor para el ejercicio, como por ejemplo: CC, pago, pagos, no, referencia, pse, web, cr, paymentid, entre otras. Todas estas palabras se incluyen en una lista llamada `no_relevant_words`, que se utilizará en los demás procesamientos.

## 2.2 Asignación de categoría para registros sin información

Para esta etapa del procesamiento se hará una identificación de aquellos registros que están vacíos o solo tienen palabras que se han identificado como no requeridas (`stopwords_spanish`, `no_relevant_cords`, `unique_words`, `two_char_words`), de tal forma que a todos ellos se les asigne la categoría objetivo "Otros", pues no se cuenta con información suficiente para hacer la clasificación.

Al aplicar este filtro se encuentra que quedan:

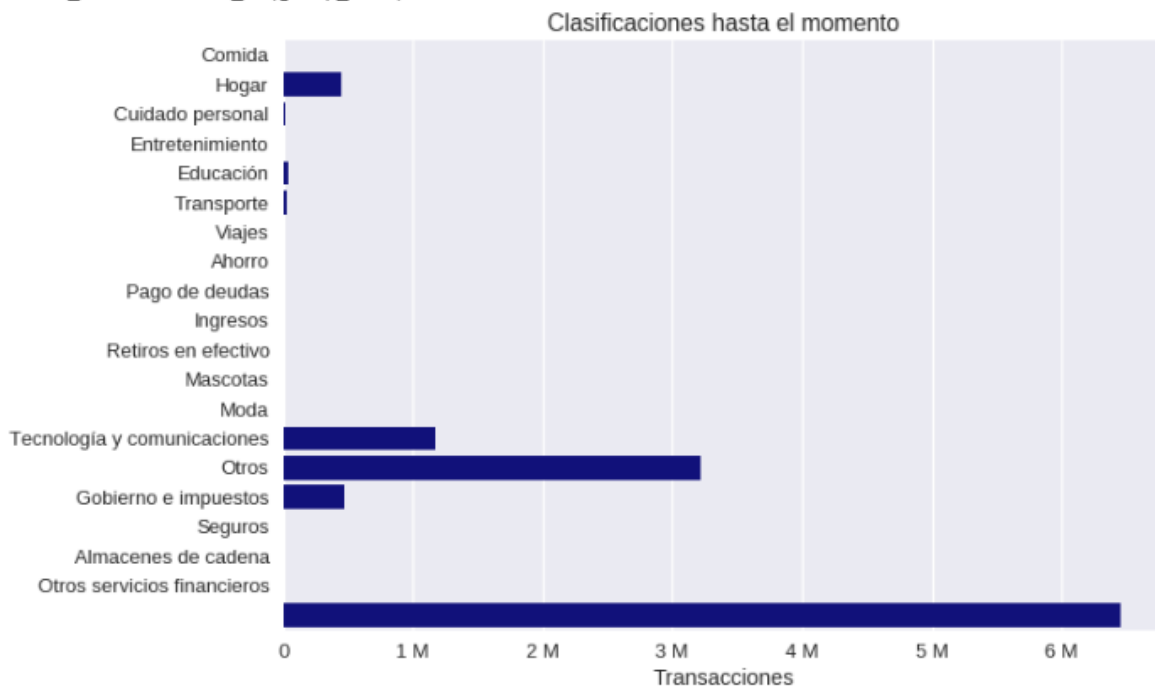
- 3.210.212 transacciones en la categoría "Otros" por falta de información.
- 8.643.570 transacciones aun pendiente por categorizar.



### 2.3 Asignación de categoría por subsector

Teniendo presente que muchas de las transacciones tienen ya asignado un subsector, pues el receptor del pago es cliente del banco, se hizo una clasificación manual de a qué categoría debería pertenecer ese subsector, y por ello en esta etapa del procesamiento se procede a asignar esa categoría a las transacciones asociadas.

Luego de aplicar estas clasificaciones se obtienen los siguientes resultados:



### 2.4 Asignación de categoría partir de bolsa de palabras por Web Scraping

Con la intención de tener una fuente para hacer la clasificación de las transacciones, se realizó Web Scraping a la página de PSE, en donde se encuentra el listado de empresas que utilizan la plataforma, relacionando una categoría para cada una:

[https://portal.psepagos.com.co/web/catalogo-pse?utm\\_source=url\\_pse&utm\\_medium=pse&utm\\_campaign=catalogopse](https://portal.psepagos.com.co/web/catalogo-pse?utm_source=url_pse&utm_medium=pse&utm_campaign=catalogopse)

De esta forma, se procedió a analizar el HTML de la página web de PSE, encontrando que existen un campo llamado empresa y otro llamado span en el que en algunas ocasiones se tiene el mismo nombre de la empresa, pero que en otras contiene una serie de palabras en relación al tipo de empresa. Se encuentran además que cada empresa está asociada a unas categorías propias de PSE, pero que son de utilidad para el ejercicio pues pueden asociarse directamente a las MCC que se tienen como objetivo.

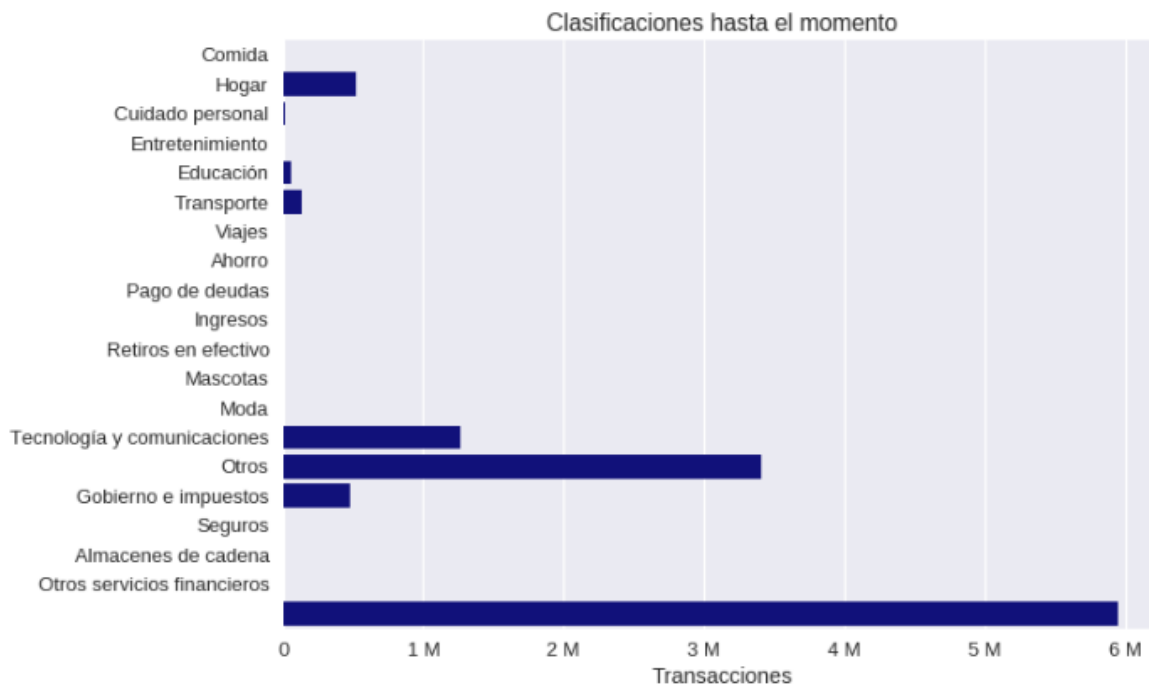
De esta forma, luego de cargar la información y procesarla, se logra obtener un dataframe con 3417 registros que agrupa listados de palabras asociados a categorías.

Para poder cruzar estos listados con el set de datos, se convierte la matriz TF que se había generado anteriormente a una matriz booleana, que indica si cada transacción tiene o no cada una de las diferentes palabras identificadas en todo el set de datos, excluyendo los listados mencionados anteriormente. Esto permite que el proceso sea mucho más rápido y genera salidas como:

```
Si: ['corporacion', 'mesa', 'yeguas', 'country', 'club'] - 1
Si: ['conjunto', 'residencial', 'retiro', 'santa', 'monica'] - 174
Si: ['asociaci', 'padres', 'familia', 'gimnasio', 'fontana'] - 1
Si: ['corporaci', 'mesa', 'yeguas', 'country', 'club'] - 79
Si: ['universidad', 'pontificia', 'bolivariana', 'seccional', 'bucaramanga'] - 144
Si: ['servicios', 'funebres', 'san', 'pedro', 'ltida'] - 146
Si: ['alarmas', 'multi', 'servicios', 'ltida'] - 121
Si: ['sociedad', 'portuaria', 'regional', 'buenaventura'] - 969
Si: ['sociedad', 'puerto', 'industrial', 'aguadulce'] - 283
Si: ['conjunto', 'residencial', 'santa', 'ana'] - 2
Si: ['sociedad', 'portuaria', 'regional', 'buenaventura'] - 969
Si: ['camara', 'comercio', 'aburra', 'sur'] - 5669
Si: ['jairo', 'alberto', 'arango', 'gomez'] - 229
Si: ['iglesia', 'centro', 'mundial', 'avivamiento'] - 78
Si: ['camara', 'comercio', 'santa', 'marta'] - 832
Si: ['colegio', 'santa', 'ana', 'fontibon'] - 218
Si: ['colegio', 'gimnasio', 'vermont', 'medellin'] - 58
Si: ['tour', 'vacation', 'hoteles', 'azul'] - 503
Si: ['tour', 'vacation', 'hoteles', 'azul'] - 503
Si: ['caceres', 'ferro', 'finca', 'raiz'] - 443
Si: ['camara', 'comercio', 'aburra', 'sur'] - 5669
Si: ['conjunto', 'residencial', 'saint', 'moritz'] - 110
Si: ['tour', 'vacation', 'hoteles', 'azul'] - 503
Si: ['universidad', 'san', 'buenaventura', 'cali'] - 2
Si: ['global', 'mercado', 'turismo'] - 2
```

Una vez finalizado el proceso, se encuentra que este método permitió identificar más de 1.5M de transacciones, lo cual es un resultado interesante teniendo presente que se partió de la categorización que ya tiene el mismo PSE en su página Web. La única limitante de este método es que se depende de que esta categorización esté bien hecha, y que la información en la página se actualice permanentemente para incluir las nuevas empresas que firmen convenios con PSE.

Los resultados de este proceso permitieron obtener los siguientes totales de transacciones por categoría:



Se debe tener presente también que a pesar de este avance importante en la categorización, aun quedan casi 6M de transacciones por clasificar, por lo cual se continuará el proceso mediante un nuevo acercamiento diferente detallado en la siguiente sección.

## 2.5 Asignación de categoría a partir de bolsa de palabras manual

Para este proceso se partirá de la misma matriz TF, convertida a booleana, que se utilizó en la metodología de clasificación anterior. La diferencia es que acá se realizará la categorización a partir de una bolsa de palabras creada manualmente por los miembros del equipo, donde se relacionaron palabras "comunes" que consideramos podían estar asociadas a cada una de las categorías.

A continuación, se muestra un ejemplo de esta matriz:

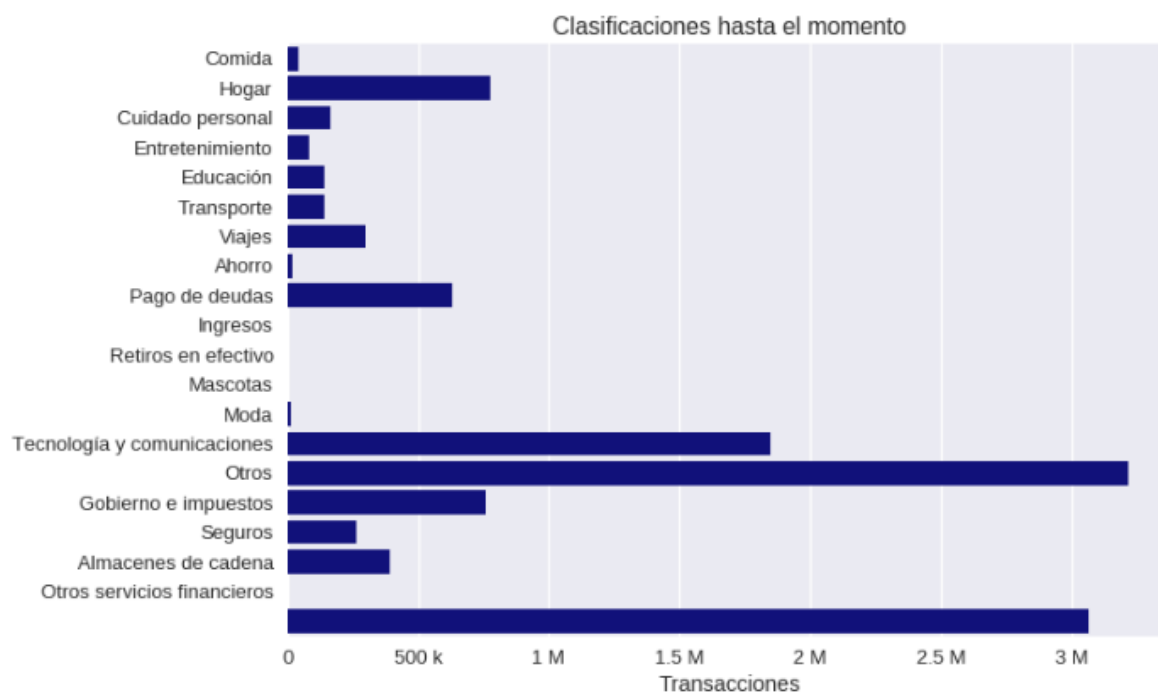
Comida	Hogar	Cuidado personal	Entretenimiento	Educación	Transporte	Viajes	Ahorro	Pago de deudas
novaventa	gas	yanbal	boleteria	jardin	flytech	tiquetes	ahorro voluntario	credito
restaurante	acueducto	eps	boletas	colegio	flypass	avianca	aporte voluntario	amex
restaurantes	eab	colsanitas	boleta	universidad	licencia conduccion	flight	pensiones voluntarias	visa
alimento	empresas publicas de medellin	salud sura	deporte	pregado	transporte	booking	ahorro	mastercard
alimentos	alcantarillado	medicina	procinal	matricula	accenorte	ada	old mutual	master card

Para la aplicación de esta metodología se sigue un proceso similar al del paso anterior, aplicando cada grupo de palabras sobre la matriz TF convertida a booleana y asignando la clasificación respectiva a los registros que cumplan con las condiciones.

Cabe resaltar además, que a pesar de que no se muestra el detalle en este informe, este procesamiento incluye todas las transformaciones requerida de los datos para poder hacer comparaciones adecuadas. A continuación, una muestra de parte de la salida del proceso:

```
Si: ['empresas', 'publicas', 'medellin'] - 433147
Si: ['flash', 'mobile'] - 2152
Si: ['impuesto', 'predial'] - 72935
Si: ['recarga', 'tag'] - 24215
Si: ['club', 'campestre'] - 544
Si: ['cruz', 'roja'] - 17
Si: ['camara', 'comercio'] - 90879
Si: ['centro', 'salud'] - 1
Si: ['licencia', 'conduccion'] - 26758
Si: ['impuesto', 'vehicular'] - 4810
Si: ['impuesto', 'vehicular'] - 4810
Si: ['industria', 'comercio'] - 15572
Si: ['canchas', 'futbol'] - 487
Si: ['cancha', 'futbol'] - 178
Si: ['servicios', 'educativos'] - 6988
Si: ['aportes', 'obligaciones'] - 10833
Si: ['tarjeta', 'profesional'] - 1355
Si: ['servicios', 'metro'] - 13749
Si: ['cine', 'colombia'] - 143
```

Se encuentra que con esta herramienta, cuya complejidad es baja pues solo requiere que un grupo de personas acuerden términos comunes asociados a cada categoría, se lograrían cubrir casi 7 millones de transacciones, aunque acá se suman algunas que ya tenían categoría asignada a partir de uno de los métodos anteriores, por lo cual, luego de aplicar esta metodología, aun quedan aproximadamente 3M de transacciones pendientes.

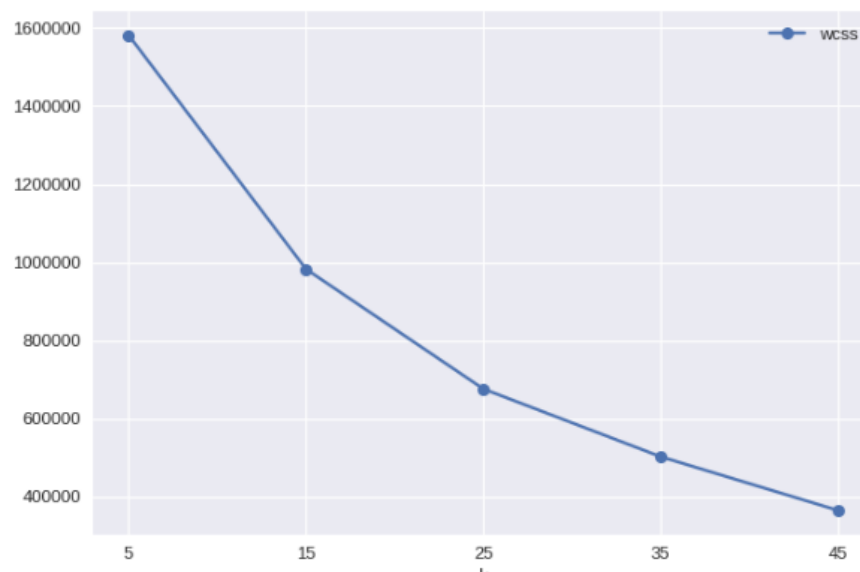


En la siguiente sección se hará una exploración adicional que busca terminar de clasificar las transacciones aun pendientes.

## 2.6 Asignación de categorías mediante clusters

En esta etapa se hace una exploración de los datos que aun no tienen categoría mediante un ejercicio de clusterización, bajo el cual se busca identificar mediante el método del codo cuál es la cantidad ideal de clusters para identificar los datos, y posteriormente analizar la información que quede en cada cluster para validar si se identifican nuevas transacciones a agrupar en las diferentes categorías. Cabe resaltar que este procesamiento se hace solamente sobre los registros a los que no se les ha asignado aun una categoría.

Lo primero que se hace, como ya se mencionó es hacer un ejercicio de clusterización bajo diferentes cantidades de clusters, para validar el error relativo de cada uno, y a partir de ello identificar la cantidad ideal. Para este caso se hizo la validación para  $K = 5, 15, 25, 35$  y  $45$ , obteniendo los siguientes resultados:

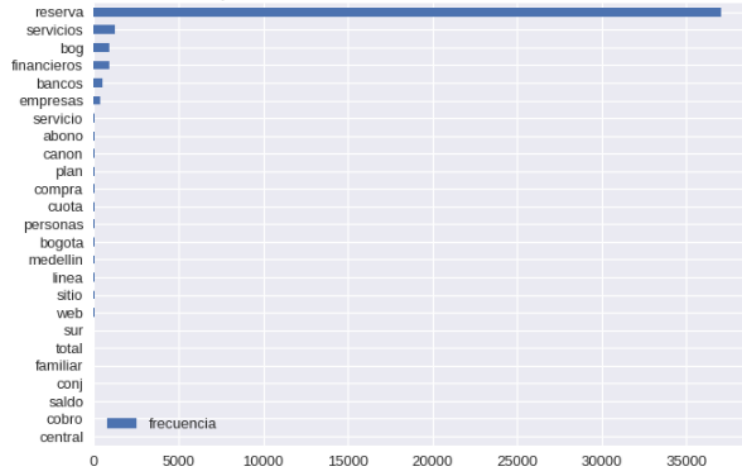


Se encuentra hasta este punto que hay un codo con  $k=15$ , por lo cual se procede a tomar esta cantidad de clusters y a hacer un análisis de las palabras asociadas a cada uno de ellos, para a partir de esto definir cuál es la categoría respectiva e ir avanzando en la asignación de una clasificación por grupos de palabras.

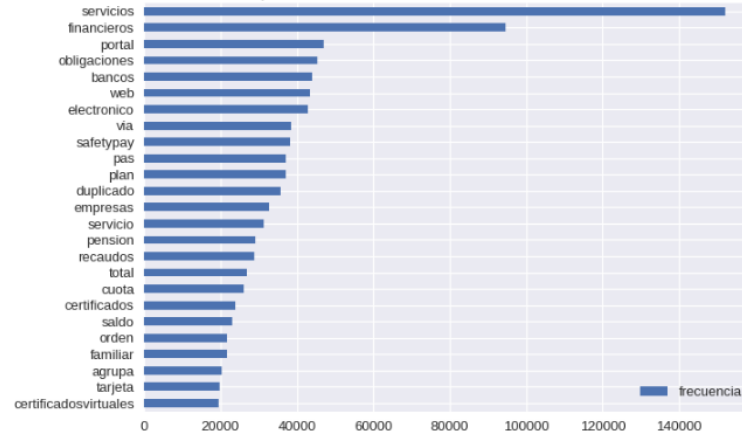
A continuación, algunas muestras de las palabras más frecuentes por cada cluster:



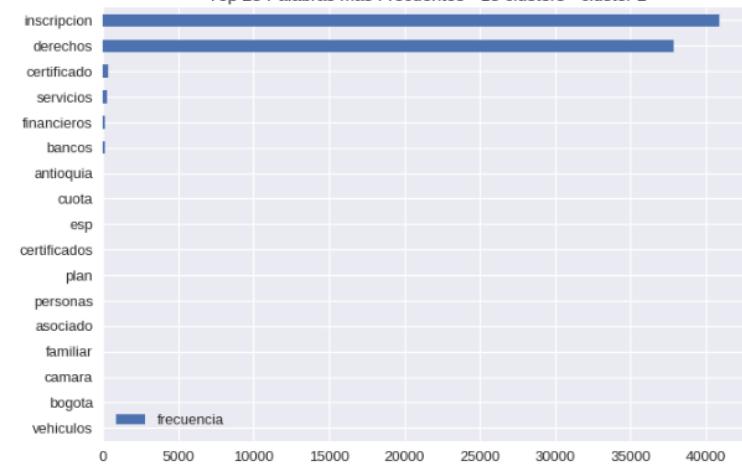
Top 25 Palabras más Frecuentes - 15 clusters - cluster 0



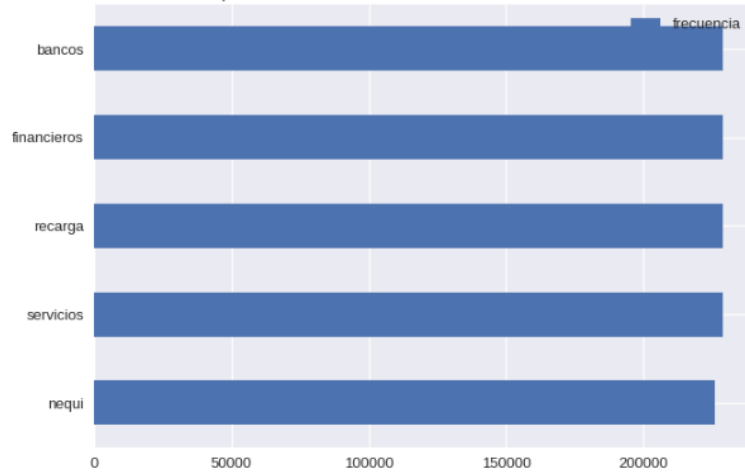
Top 25 Palabras más Frecuentes - 15 clusters - cluster 1



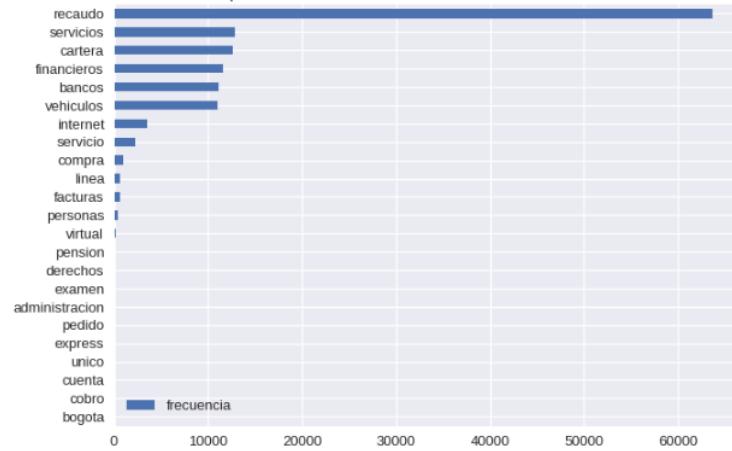
Top 25 Palabras más Frecuentes - 15 clusters - cluster 2



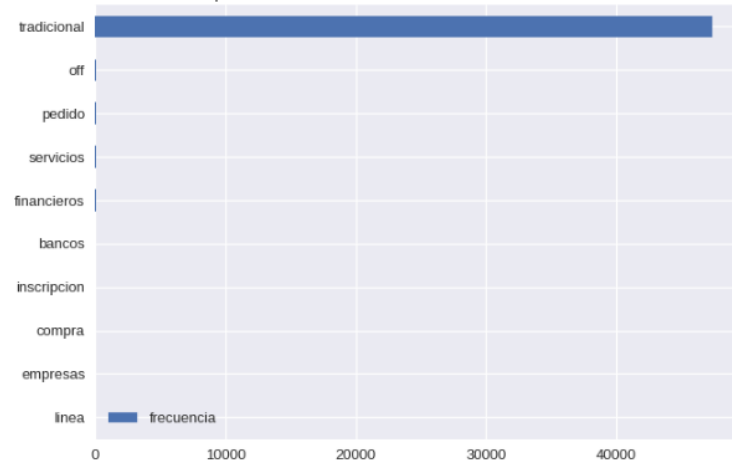
Top 25 Palabras más Frecuentes - 15 clusters - cluster 6

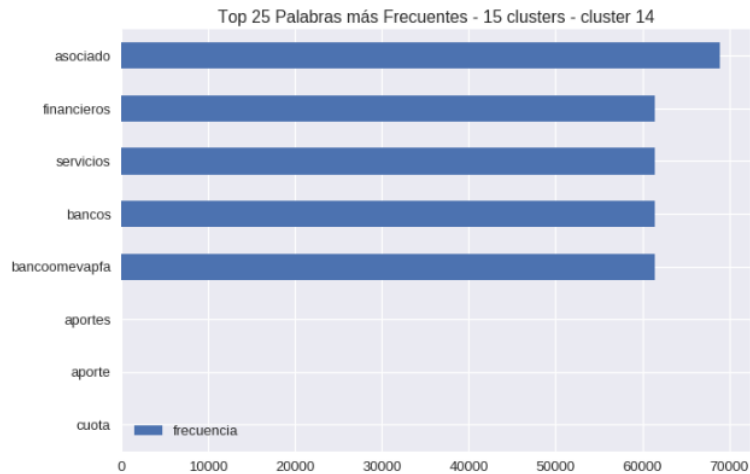


Top 25 Palabras más Frecuentes - 15 clusters - cluster 12



Top 25 Palabras más Frecuentes - 15 clusters - cluster 13



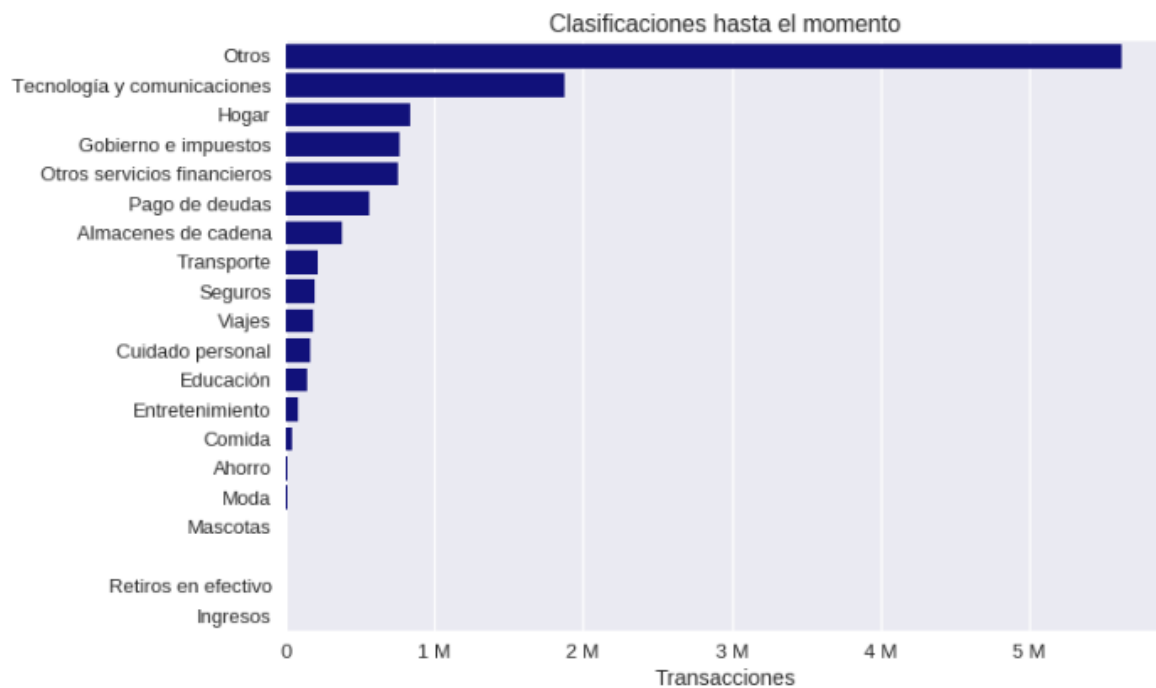


Se encuentra que hay un cluster grande con más de 1M de registros y que tiene una gran dispersión de términos diferentes, que no son claramente relacionables frente a una categoría.

Por otro lado, se identifica que varios de los clusters tienen como factor común las palabras "servicios" y "financieros", lo cual está atado a pagos de clientes de Bancolombia que se hacen a otros bancos, pero para los cuales no es fácil identificar cuál es su destino final. Debido a esto se procede a crear una nueva categoría denominada "Otros servicios bancarios" donde se agruparán todas estas transacciones.

Así mismo, se identifica que en general los demás clusters están asociados a términos en los que es difícil identificar cuál es su categoría, por lo cual estos se llevan a la categoría "Otros"

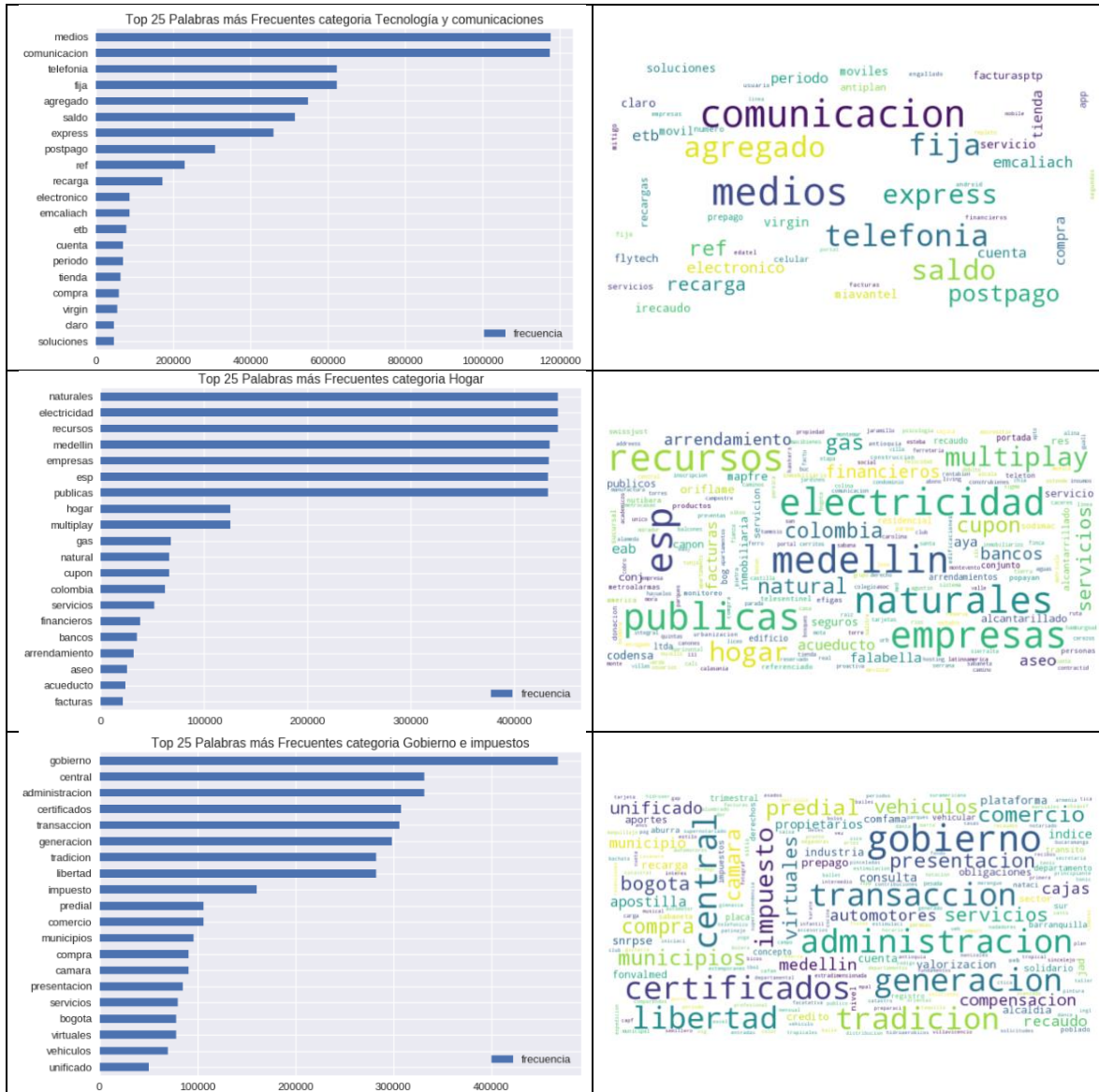
Luego de este procesamiento, se obtienen los resultados definitivos del proceso de categorización de las transacciones, que genera los siguientes resultados ordenados según la cantidad asociada.

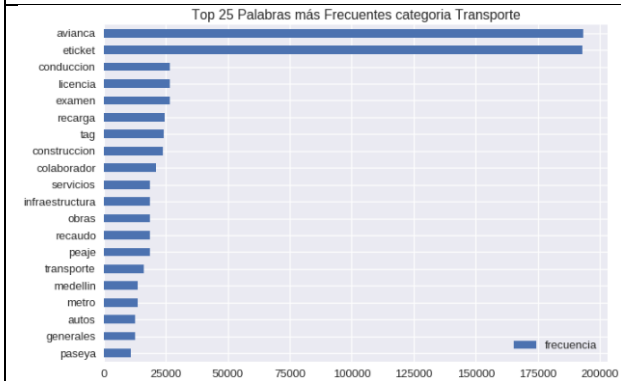
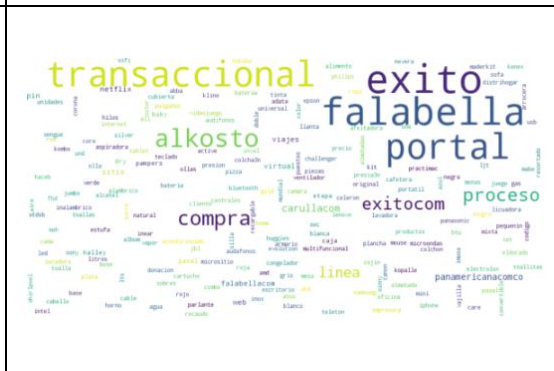
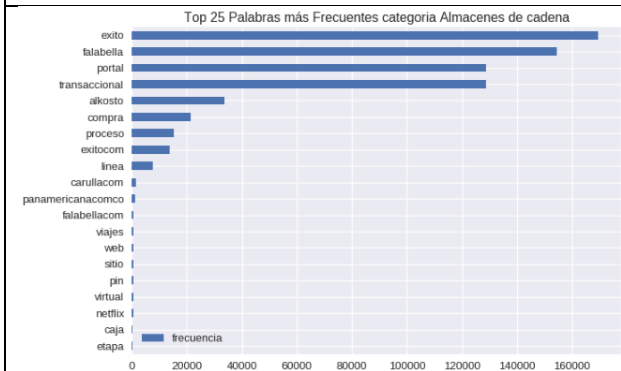
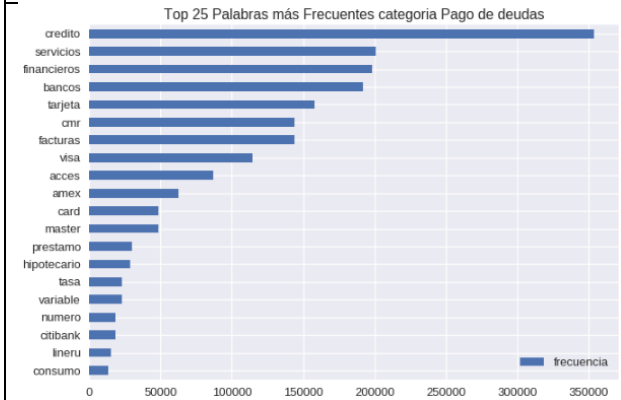
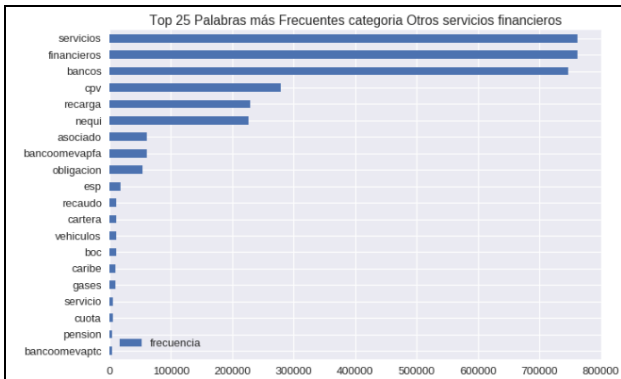


## 2.7 Descripción de las categorías identificadas

Finalmente, y para cerrar el proceso de clasificación, se hace una descripción de las diferentes categorías identificadas, presentando para cada una las palabras más frecuentes countplot y un WordCloud que presentan las palabras más comunes para cada categoría.

A continuación, algunos ejemplos de las clasificaciones realizadas:







### 3. Modelo Clasificador de Categorías

Teniendo presente que una de las aplicaciones de este ejercicio está asociada a generar valor desde los sistemas PFM, en esta sección se buscará generar un modelo de clasificación que pueda ser usado en conjunto con un PFM para facilitar a sus usuarios llevar sus cuentas personales, al lograr clasificar las transacciones PSE que realice en las categorías definidas.

Así mismo, y teniendo presente que los modelos que generalmente tienen mejor desempeño en la clasificación de texto son Naive Bayes y la Máquina de Soporte Vectorial lineal, el procesamiento de texto y los clasificadores a construir se basarán en estos.

El procesamiento incluye los siguientes pasos:

1. Cargar el dataframe con la información.
2. Aplicarle los filtros de texto definidos desde la sección 2.
3. Crear una sola columna de texto, que agrupa los 6 campos de texto de categorización.
4. Carga los listados de palabras a excluir, definidas en la sección 2.

Una vez se cuenta con esta información ya procesada, se procede con el entrenamiento del modelo, lo cual incluye:

1. Hacer un sub-samplio de todos los registros, pues son demasiados y muchos contienen información redundante, por lo cual se busca disminuir el procesamiento requerido sin afectar los resultados.
2. Sobre la cantidad de registros resultantes, hacer un Split 50-50 (entrenamiento – evaluación).
3. Aplicar un modelo Multinomial Naives Bayes, bajo el cual se obtuvieron los siguientes resultados:

	precision	recall	f1-score	support
Ahorro	0.01	0.97	0.01	18656
Almacenes de cadena	0.94	1.00	0.97	390398
Comida	0.94	0.73	0.82	43512
Cuidado personal	0.97	0.95	0.96	163117
Educación	0.88	0.90	0.89	140626
Entretenimiento	0.76	0.79	0.78	83129
Gobierno e impuestos	1.00	0.99	0.99	762644
Hogar	1.00	0.89	0.94	800103
Mascotas	0.08	0.60	0.14	1080
Moda	0.43	0.58	0.49	10785
Otros	0.48	0.01	0.03	3271961
Pago de deudas	0.99	0.99	0.99	615335
Seguros	0.94	0.99	0.97	261874
Tecnología y comunicaciones	1.00	0.98	0.99	1547198
Transporte	0.98	0.95	0.97	99156
Viajes	1.00	0.96	0.98	298910
micro avg	0.60	0.60	0.60	8508484
macro avg	0.77	0.83	0.74	8508484
weighted avg	0.79	0.60	0.60	8508484

Se encuentra que los resultados son mixtos, y algunas categorías tuvieron una muy baja precisión en la clasificación, por lo cual se procede a evaluar un clasificador bajo SVM.

4. Aplicar un modelo de Linear Support Vector Classification, que se hizo también bajo Split 50-50 del dataset ajustado mediante sub-sample, pero cuyo desempeño final se evaluó frente al 100% de los datos de las transacciones para lograr una evaluación muy estricta. Los resultados obtenidos fueron los siguientes:

	precision	recall	f1-score	support
Ahorro	0.92	0.99	0.95	18656
Almacenes de cadena	1.00	1.00	1.00	390398
Comida	0.53	0.84	0.65	43512
Cuidado personal	0.97	0.95	0.96	163117
Educación	0.88	0.96	0.92	140626
Entretenimiento	0.98	0.80	0.88	83129
Gobierno e impuestos	1.00	1.00	1.00	762644
Hogar	1.00	0.95	0.97	800103
Mascotas	0.67	0.46	0.55	1080
Moda	0.93	0.58	0.72	10785
Otros	0.98	0.99	0.99	3271961
Pago de deudas	1.00	1.00	1.00	615335
Seguros	1.00	0.99	1.00	261874
Tecnología y comunicaciones	1.00	0.99	0.99	1547198
Transporte	0.91	1.00	0.95	99156
Viajes	1.00	0.98	0.99	298910
avg / total	0.99	0.98	0.99	8508484

Se encuentra que el desempeño general es muy bueno, pues la mayor parte de las categorías tuvo una precisión mayor a 0.9, muchas tuvieron 1, y solo hubo 2 categorías con desempeño bajo, Hogar y Mascotas, que justamente son algunas de las que menos registros tienen.



El nivel promedio de precision fue de 0.99 al evaluar el modelo sobre todo el dataset.

### **3.1 Conclusiones Clasificador de Categorías**

Luego de aplicar estos modelos se encuentra que se tienen buenos resultados en cuanto a desempeño para el clasificador basado en Máquinas de Soporte Vectorial. Esto permitiría efectivamente hacer una adecuada clasificación de las transacciones vía PSE del cliente, facilitando que este pueda verlas directamente agrupadas en la aplicación de PFM.

Cabe resaltar que la calidad de la información entregada por este modelo depende de la calidad de la clasificación realizada en la sección 2, razón por la cual se buscó hacer esa parte del ejercicio de forma muy rigurosa, y en futuros acercamientos se podrían implementar mecanismos adicionales para asegurar la clasificación de las transacciones, como el ya mencionado en el que se tome la clasificación final realizada por el usuario, y sea el insumo principal para el entrenamiento, lo cual permitiría además que el proceso de categorización sea personalizado para el cliente y no esté restringido exclusivamente a las categorías definidas por Bancolombia.

## **4. Análisis de Pagadores (Clientes)**

En esta sección se buscará:

1. Identificar cuáles son las principales características de los clientes, que determinan su probabilidad de realizar una transacción.
2. Crear clusters que permitan agrupar los clientes, de tal forma que se pueda generar valor para el banco mediante la implementación de estrategias de cross-selling, y para el cliente tener una base de comparación sobre la cual pueda recibir recomendaciones si su comportamiento de gasto es superior al de personas con características similares.

Para lograr esto el proceso que se seguirá incluye:

1. Analizar y preparar los datos de los pagadores (clientes)
2. Cruzarlos con el resumen de transacciones por cada tipo para cada cliente
3. Hacer los dos análisis que se abordarán.

### **4.1 Entendimiento de los datos de los pagadores (clientes)**

En esta sección se carga la información de los clientes, se hacen unos ajustes iniciales como por ejemplo cambiar las categorías por la descripción asociada, y luego hacer una exploración estadística y visual de los datos.

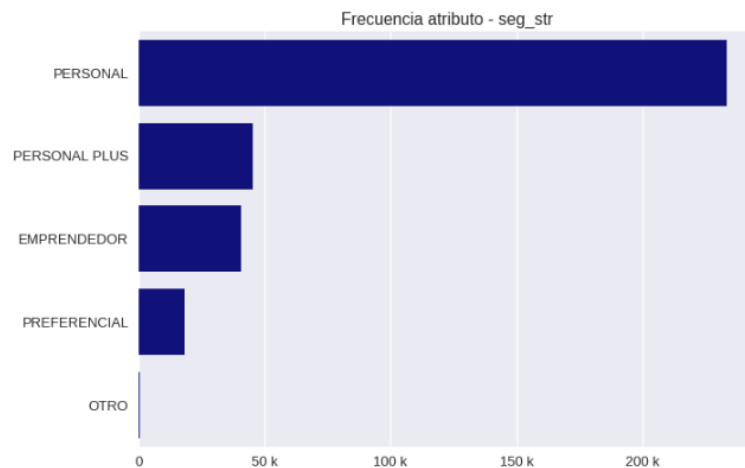
Al hacer una descripción de los datos se encuentra lo siguiente:

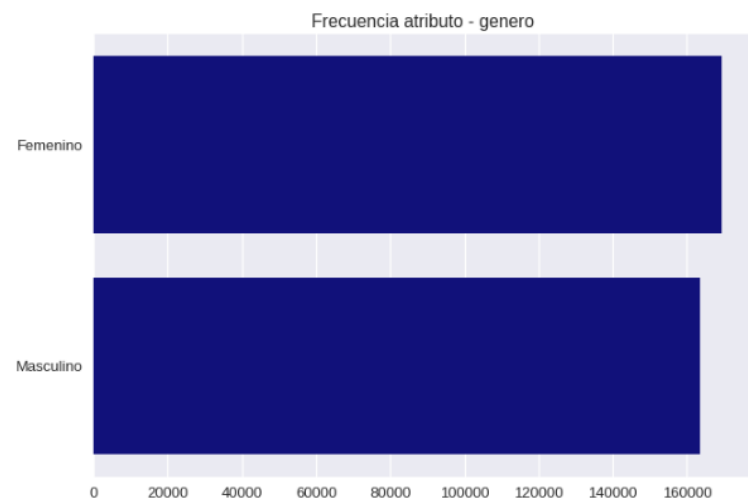
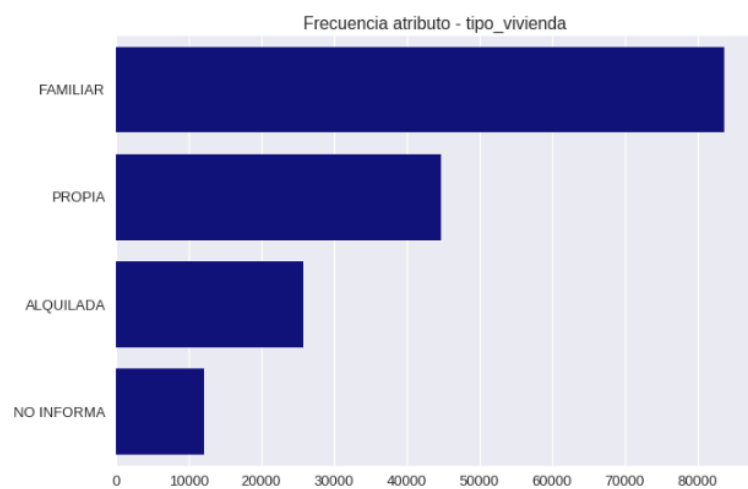
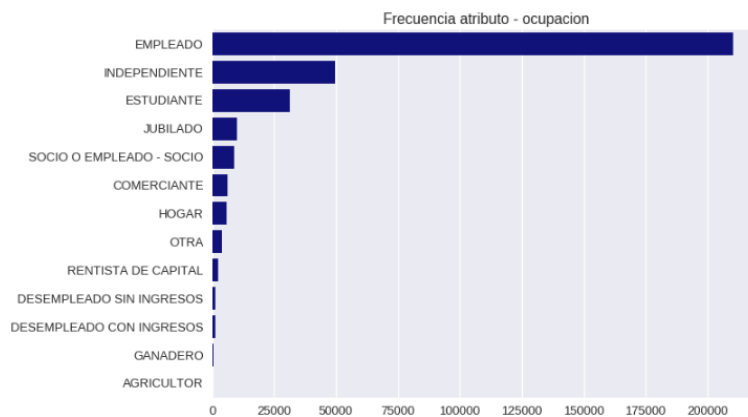
columna	count	unique	top	freq
id_cliente	338606	338606	195925	1
seg_str	338606	5	PERSONAL	233568
ocupacion	331769	13	EMPLEADO	210488
tipo_vivienda	166386	4	FAMILIAR	83736
nivel_academico	294313	8	UNIVERSITARIO	135754
estado_civil	332006	7	SOLTERO	175878
genero	332897	2	Femenino	169336
edad	338606	109	29	14001
ingreso_rango	338606	11	b. (1.1 2.2MM]	105099

De esto se concluye de forma preliminar que:

- En la columna cliente solo hay valores únicos
- Las columnas segmento, edad e ingreso\_rango no tienen nulos.
- Las demás columnas tienen algunos nulos
- La columna edad tiene 109 valores diferentes, lo cual es sospechoso para información de este tipo.

Posteriormente se hizo un análisis exploratorio de la información de los clientes, de los cuales se presentan algunos gráficos:





El detalle del resto de los campos se puede revisar en el Jupyter notebook.

## 4.2 Preparación de los datos

A partir de lo que se identifica en cada gráfica, se procede a hacer los siguientes ajustes iniciales:

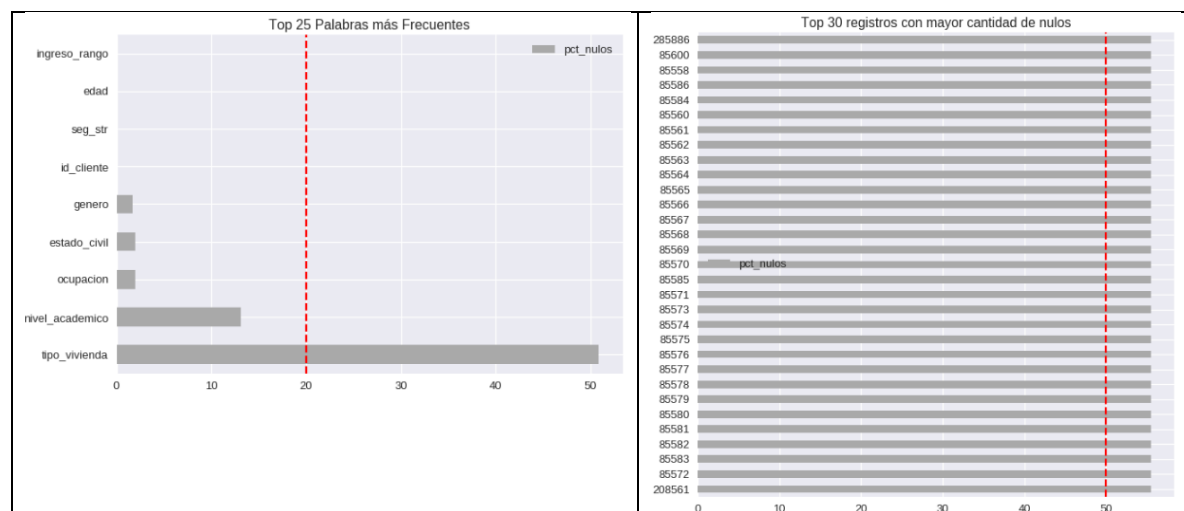
1. Para estado\_civil, unificar DESCONOCIDO con NO INFORMA, pues ambos tienen el mismo sentido, y en general los otros campos utilizan la categoría NO INFORMA.
2. Ajustar el formato de la categoría de ingresos, para que sea más fácil de entender.
3. Eliminar los valores de la variable edad que no sean numéricos.
4. Más adelante, se procederá a hacer una revisión de los valores nulos de cada variable con su respectiva eliminación/imputación.

Finalmente, se hará el tratamiento de los valores atípicos de la columna edad, para luego redefinir esta categoría y llevarla de números a rangos.

De esta forma, y luego de realizar los pasos 1, 2 y 3, se obtiene un set de datos con la siguiente estructura:

id_cliente	seg_str	ocupacion	tipo_vivienda	nivel_academico	estado_civil	genero	edad	ingreso_rango
18	PERSONAL PLUS	JUBILADO	PROPIA	UNIVERSITARIO	CASADO	Masculino	92	(4.4 5.5MM]
32	PERSONAL PLUS	SOCIO O EMPLEADO - SOCIO	FAMILIAR	TECNICO	CASADO	Masculino	80	(8.7 Inf)
41	EMPRENDEDOR	INDEPENDIENTE	PROPIA	NO INFORMA	VIUDO	Masculino	90	(1.1 2.2MM]
47	EMPRENDEDOR	GANADERO	NaN	NO INFORMA	NO INFORMA	Masculino	86	(2.2 3.3MM]
71	PERSONAL	JUBILADO	PROPIA	POSTGRADO	CASADO	Masculino	79	(4.4 5.5MM]
338486	PERSONAL	ESTUDIANTE	NaN	NaN	SOLTERO	Femenino	19	(1.1 2.2MM]
338512	PERSONAL	ESTUDIANTE	NaN	NaN	SOLTERO	Femenino	19	(2.2 3.3MM]
338567	PERSONAL	EMPLEADO	NaN	NaN	SOLTERO	Masculino	18	disponible
338578	PERSONAL	EMPLEADO	NaN	NaN	NO INFORMA	Masculino	18	(1.1 2.2MM]
338594	PERSONAL	EMPLEADO	NaN	NaN	SOLTERO	Masculino	18	(0 1.1MM]

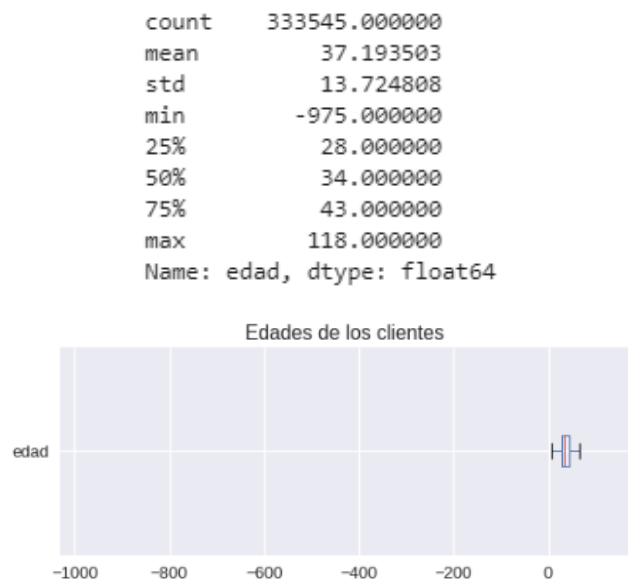
Y a partir de esto, se procede a hacer una revisión de nulos a nivel de columnas y de filas:



Se encuentra que hay varios campos nulos, para estos se procede de la siguiente forma:

- Eliminar la columna tipo\_vivienda pues tiene más del 50% de los datos nulos, y de los que no son nulos hay más de 10.000 con la categoría "NO INFORMA".
- Eliminar los 5.061 registros que tienen más del 50% de sus campos nulos, pues estos generan ruido en el procesamiento de la información.
- Reemplazar los datos nulos de las columnas nivel\_academico y estado\_civil por "NO INFORMA", pues esas columnas ya tienen esa categoría y un volumen importante de registros en ella.
- Reemplazar los datos nulos de las columnas ocupación y genero por la moda, pues estas columnas no tienen la categoría "NO INFORMA".

Luego de realizar estos ajustes, se procede a ajustar los valores de la columna edad, que es la única numérica dentro de la información de los pagadores (clientes). Para esto se hace primero un "describe" del campo y luego una exploración mediante boxplots.



Se encuentra que hay valores atípicos asociados a edades negativas o iguales a 0, por lo cual se procede a hacer un reemplazo de todas las edades menores a 5 años por la mediana.

Así mismo, se encuentra que hay múltiples registros con edad de 118 años, por lo cual se procede a cambiar también las edades mayores a 95 años también por la mediana. El nuevo boxplot luego de estos ajustes es el siguiente:



Se encuentra que las edades quedan ya en rangos adecuados para continuar con el análisis. Finalmente, se procede cambiar esta variable a categorías, para facilitar el entendimiento de los datos que se busca mediante este ejercicio. Para esto, se procede a crear una función que genera un texto para cada edad, indicando en qué rango está, y luego de aplicarla, junto con los demás ajustes asociados, se llega a un dataset del siguiente tipo:

id_cliente	seg_str	ocupacion	nivel_academico	estado_civil	genero	edad	ingreso_rango
18	PERSONAL PLUS	JUBILADO	UNIVERSITARIO	CASADO	Masculino	>=90	(4.4 5.5MM]
32	PERSONAL PLUS	SOCIO O EMPLEADO - SOCIO	TECNICO	CASADO	Masculino	80-89	(8.7 Inf)
41	EMPRENDEDOR	INDEPENDIENTE	NO INFORMA	VIUDO	Masculino	>=90	(1.1 2.2MM]
47	EMPRENDEDOR	GANADERO	NO INFORMA	NO INFORMA	Masculino	80-89	(2.2 3.3MM]
71	PERSONAL	JUBILADO	POSTGRADO	CASADO	Masculino	70-79	(4.4 5.5MM]
338486	PERSONAL	ESTUDIANTE	NO INFORMA	SOLTERO	Femenino	<20	(1.1 2.2MM]
338512	PERSONAL	ESTUDIANTE	NO INFORMA	SOLTERO	Femenino	<20	(2.2 3.3MM]
338567	PERSONAL	EMPLEADO	NO INFORMA	SOLTERO	Masculino	<20	disponible
338578	PERSONAL	EMPLEADO	NO INFORMA	NO INFORMA	Masculino	<20	(1.1 2.2MM]
338594	PERSONAL	EMPLEADO	NO INFORMA	SOLTERO	Masculino	<20	(0 1.1MM]

En el Jupyter, se puede encontrar la exploración visual de los datos luego de todo el procesamiento realizado hasta este punto.

Posteriormente y como en este ejercicio se busca identificar cuál es el perfil específico de los clientes que hacen cada tipo de transacción, el último proceso que se hace con los datos es generar dummy variables para cada categoría, de tal forma que posteriormente se puedan validar cuáles de las columnas dummy tienen mayor correlación con la variable buscada. Esto genera un dataset del siguiente tipo:

	segm_EMPRENDEDOR	segm_OTRO	segm_PERSONAL	segm_PERSONAL PLUS	segm_PREFERENCIAL	ocup_AGRICULTOR	ocup_COMERCIANTE
id_cliente							
18	0	0	0	1	0	0	0
32	0	0	0	1	0	0	0
41	1	0	0	0	0	0	0
47	1	0	0	0	0	0	0
71	0	0	1	0	0	0	0
338486	0	0	1	0	0	0	0
338512	0	0	1	0	0	0	0
338567	0	0	1	0	0	0	0
338578	0	0	1	0	0	0	0
338594	0	0	1	0	0	0	0

Finalmente, y como lo que se busca es revisar la relación entre los clientes y el tipo de transacciones que realiza, se crea un dataframe que agrupa para cada cliente la cantidad de transacciones que hace. Para esto se utiliza la función pivot\_table. Esto genera un dataset del siguiente tipo:

	Comida	Hogar	Cuidado personal	Entretenimiento	Educación	Transporte	Viajes	Ahorro	Pago de deudas	Mascotas	Moda	c
1	0.0	0.00	0.0	0.0	0.0	0.0	0.0	9942522.53	32588454.31	0.0	0.0	
10	0.0	3523278.59	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	
100	0.0	1456792.68	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	
1000	0.0	34112656.94	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	
10000	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	

Esta matriz se lleva a una booleana, que simplemente indique si un cliente ha realizado o no transacciones de cada categoría:

	Comida	Hogar	Cuidado personal	Entretenimiento	Educación	Transporte	Viajes	Ahorro	Pago de deudas	Mascotas	Moda	c
1	0	0	0	0	0	0	0	1	1	0	0	
10	0	1	0	0	0	0	0	0	0	0	0	
100	0	1	0	0	0	0	0	0	0	0	0	
1000	0	1	0	0	0	0	0	0	0	0	0	
10000	0	0	0	0	0	0	0	0	0	0	0	

Ya teniendo preparadas tanto la matriz de los pagadores (clientes) como la matriz de las transacciones por cliente, por valor y booleana, se procede a hacer el merge entre ambos, que genera un dataframe de 333545 registros por 71 columnas.

### 4.3 Modelamiento - Características de cliente por categoría de trx

Para el modelamiento de las principales características de cliente que se asocian a cada tipo de transacción, se crea una matriz de correlaciones a partir del dataframe que agrupa características de cliente y si ha realizado o no transacciones de cada tipo, de tal forma que luego se compare cuáles son los atributos que mayor correlación tienen.

El resultado de este análisis arroja lo siguiente:

Principales características para "Hogar"		Principales características para "Cuidado personal"		Principales características para "Entretenimiento"	
segm_PERSONAL PLUS	0.086	segm_PERSONAL PLUS	0.075	estu_UNIVERSITARIO	0.066
estu_UNIVERSITARIO	0.075	estu_UNIVERSITARIO	0.075	edad_30-39	0.052
ingr_(8.7 Inf)	0.066	eciv_CASADO	0.055	ocup_EMPLEADO	0.047
edad_30-39	0.065	ingr_(8.7 Inf)	0.052	segm_PERSONAL PLUS	0.044
estu_ESPECIALIZACION	0.056	estu_ESPECIALIZACION	0.052	eciv_SOLTERO	0.041
eciv_CASADO	0.055	gen_Femenino	0.05	estu_ESPECIALIZACION	0.041
ingr_(3.3 4.4MM]	0.054	edad_30-39	0.05	ingr_(3.3 4.4MM]	0.031
segm_PREFERENCIAL	0.052	segm_PREFERENCIAL	0.043	ingr_(4.4 5.5MM]	0.025
ingr_(4.4 5.5MM]	0.048	estu_POSTGRADO	0.041		
ocup_EMPLEADO	0.04	ingr_(4.4 5.5MM]	0.04		

<div>Principales características para "Educación"</div> <table><tr><td>ingr_(8.7 Inf)</td><td>0.054</td></tr><tr><td>segm_PREFERENCIAL</td><td>0.052</td></tr><tr><td>ocup_ESTUDIANTE</td><td>0.044</td></tr><tr><td>estu_UNIVERSITARIO</td><td>0.042</td></tr><tr><td>segm_PERSONAL PLUS</td><td>0.042</td></tr><tr><td>eciv_CASADO</td><td>0.041</td></tr><tr><td>edad_40-49</td><td>0.037</td></tr><tr><td>estu_ESPECIALIZACION</td><td>0.033</td></tr><tr><td>estu_POSTGRADO</td><td>0.032</td></tr></table>	ingr_(8.7 Inf)	0.054	segm_PREFERENCIAL	0.052	ocup_ESTUDIANTE	0.044	estu_UNIVERSITARIO	0.042	segm_PERSONAL PLUS	0.042	eciv_CASADO	0.041	edad_40-49	0.037	estu_ESPECIALIZACION	0.033	estu_POSTGRADO	0.032	<div>Principales características para "Transporte"</div> <table><tr><td>ingr_(8.7 Inf)</td><td>0.088</td></tr><tr><td>segm_PERSONAL PLUS</td><td>0.083</td></tr><tr><td>estu_UNIVERSITARIO</td><td>0.07</td></tr><tr><td>segm_PREFERENCIAL</td><td>0.062</td></tr><tr><td>estu_ESPECIALIZACION</td><td>0.044</td></tr><tr><td>estu_POSTGRADO</td><td>0.04</td></tr><tr><td>ingr_(3.3 4.4MM]</td><td>0.033</td></tr><tr><td>eciv_CASADO</td><td>0.033</td></tr><tr><td>ingr_(4.4 5.5MM]</td><td>0.032</td></tr><tr><td>ingr_(5.5 6.6MM]</td><td>0.031</td></tr></table>	ingr_(8.7 Inf)	0.088	segm_PERSONAL PLUS	0.083	estu_UNIVERSITARIO	0.07	segm_PREFERENCIAL	0.062	estu_ESPECIALIZACION	0.044	estu_POSTGRADO	0.04	ingr_(3.3 4.4MM]	0.033	eciv_CASADO	0.033	ingr_(4.4 5.5MM]	0.032	ingr_(5.5 6.6MM]	0.031	<div>Principales características para "Viajes"</div> <table><tr><td>eciv_SOLTERO</td><td>0.05</td></tr><tr><td>estu_UNIVERSITARIO</td><td>0.046</td></tr><tr><td>edad_30-39</td><td>0.037</td></tr><tr><td>segm_PERSONAL PLUS</td><td>0.036</td></tr><tr><td>ingr_(3.3 4.4MM]</td><td>0.034</td></tr><tr><td>ocup_INDEPENDIENTE</td><td>0.031</td></tr><tr><td>ingr_(4.4 5.5MM]</td><td>0.027</td></tr><tr><td>edad_20-29</td><td>0.025</td></tr><tr><td>ingr_(5.5 6.6MM]</td><td>0.023</td></tr><tr><td>segm_EMPRENEDEDOR</td><td>0.022</td></tr></table>	eciv_SOLTERO	0.05	estu_UNIVERSITARIO	0.046	edad_30-39	0.037	segm_PERSONAL PLUS	0.036	ingr_(3.3 4.4MM]	0.034	ocup_INDEPENDIENTE	0.031	ingr_(4.4 5.5MM]	0.027	edad_20-29	0.025	ingr_(5.5 6.6MM]	0.023	segm_EMPRENEDEDOR	0.022		
ingr_(8.7 Inf)	0.054																																																													
segm_PREFERENCIAL	0.052																																																													
ocup_ESTUDIANTE	0.044																																																													
estu_UNIVERSITARIO	0.042																																																													
segm_PERSONAL PLUS	0.042																																																													
eciv_CASADO	0.041																																																													
edad_40-49	0.037																																																													
estu_ESPECIALIZACION	0.033																																																													
estu_POSTGRADO	0.032																																																													
ingr_(8.7 Inf)	0.088																																																													
segm_PERSONAL PLUS	0.083																																																													
estu_UNIVERSITARIO	0.07																																																													
segm_PREFERENCIAL	0.062																																																													
estu_ESPECIALIZACION	0.044																																																													
estu_POSTGRADO	0.04																																																													
ingr_(3.3 4.4MM]	0.033																																																													
eciv_CASADO	0.033																																																													
ingr_(4.4 5.5MM]	0.032																																																													
ingr_(5.5 6.6MM]	0.031																																																													
eciv_SOLTERO	0.05																																																													
estu_UNIVERSITARIO	0.046																																																													
edad_30-39	0.037																																																													
segm_PERSONAL PLUS	0.036																																																													
ingr_(3.3 4.4MM]	0.034																																																													
ocup_INDEPENDIENTE	0.031																																																													
ingr_(4.4 5.5MM]	0.027																																																													
edad_20-29	0.025																																																													
ingr_(5.5 6.6MM]	0.023																																																													
segm_EMPRENEDEDOR	0.022																																																													
<div>Principales características para "Pago de deudas"</div> <table><tr><td>estu_UNIVERSITARIO</td><td>0.093</td></tr><tr><td>segm_PERSONAL PLUS</td><td>0.084</td></tr><tr><td>edad_30-39</td><td>0.069</td></tr><tr><td>ingr_(8.7 Inf)</td><td>0.058</td></tr><tr><td>ocup_EMPLEADO</td><td>0.057</td></tr><tr><td>estu_ESPECIALIZACION</td><td>0.056</td></tr><tr><td>eciv_CASADO</td><td>0.053</td></tr><tr><td>ingr_(3.3 4.4MM]</td><td>0.051</td></tr><tr><td>ingr_(4.4 5.5MM]</td><td>0.048</td></tr><tr><td>segm_PREFERENCIAL</td><td>0.039</td></tr></table>	estu_UNIVERSITARIO	0.093	segm_PERSONAL PLUS	0.084	edad_30-39	0.069	ingr_(8.7 Inf)	0.058	ocup_EMPLEADO	0.057	estu_ESPECIALIZACION	0.056	eciv_CASADO	0.053	ingr_(3.3 4.4MM]	0.051	ingr_(4.4 5.5MM]	0.048	segm_PREFERENCIAL	0.039	<div>Principales características para "Tecnología y comunicaciones"</div> <table><tr><td>ocup_EMPLEADO</td><td>0.083</td></tr><tr><td>estu_UNIVERSITARIO</td><td>0.079</td></tr><tr><td>edad_30-39</td><td>0.072</td></tr><tr><td>eciv_SOLTERO</td><td>0.052</td></tr><tr><td>estu_ESPECIALIZACION</td><td>0.051</td></tr><tr><td>ingr_(3.3 4.4MM]</td><td>0.043</td></tr><tr><td>ingr_(2.2 3.3MM]</td><td>0.041</td></tr><tr><td>edad_20-29</td><td>0.035</td></tr><tr><td>segm_PERSONAL PLUS</td><td>0.03</td></tr><tr><td>ingr_(4.4 5.5MM]</td><td>0.024</td></tr></table>	ocup_EMPLEADO	0.083	estu_UNIVERSITARIO	0.079	edad_30-39	0.072	eciv_SOLTERO	0.052	estu_ESPECIALIZACION	0.051	ingr_(3.3 4.4MM]	0.043	ingr_(2.2 3.3MM]	0.041	edad_20-29	0.035	segm_PERSONAL PLUS	0.03	ingr_(4.4 5.5MM]	0.024	<div>Principales características para "Gobierno e impuestos"</div> <table><tr><td>ingr_(8.7 Inf)</td><td>0.15</td></tr><tr><td>segm_PERSONAL PLUS</td><td>0.15</td></tr><tr><td>estu_UNIVERSITARIO</td><td>0.14</td></tr><tr><td>segm_PREFERENCIAL</td><td>0.13</td></tr><tr><td>eciv_CASADO</td><td>0.13</td></tr><tr><td>estu_POSTGRADO</td><td>0.09</td></tr><tr><td>edad_40-49</td><td>0.086</td></tr><tr><td>estu_ESPECIALIZACION</td><td>0.083</td></tr><tr><td>ingr_(3.3 4.4MM]</td><td>0.075</td></tr><tr><td>ingr_(4.4 5.5MM]</td><td>0.073</td></tr></table>	ingr_(8.7 Inf)	0.15	segm_PERSONAL PLUS	0.15	estu_UNIVERSITARIO	0.14	segm_PREFERENCIAL	0.13	eciv_CASADO	0.13	estu_POSTGRADO	0.09	edad_40-49	0.086	estu_ESPECIALIZACION	0.083	ingr_(3.3 4.4MM]	0.075	ingr_(4.4 5.5MM]	0.073
estu_UNIVERSITARIO	0.093																																																													
segm_PERSONAL PLUS	0.084																																																													
edad_30-39	0.069																																																													
ingr_(8.7 Inf)	0.058																																																													
ocup_EMPLEADO	0.057																																																													
estu_ESPECIALIZACION	0.056																																																													
eciv_CASADO	0.053																																																													
ingr_(3.3 4.4MM]	0.051																																																													
ingr_(4.4 5.5MM]	0.048																																																													
segm_PREFERENCIAL	0.039																																																													
ocup_EMPLEADO	0.083																																																													
estu_UNIVERSITARIO	0.079																																																													
edad_30-39	0.072																																																													
eciv_SOLTERO	0.052																																																													
estu_ESPECIALIZACION	0.051																																																													
ingr_(3.3 4.4MM]	0.043																																																													
ingr_(2.2 3.3MM]	0.041																																																													
edad_20-29	0.035																																																													
segm_PERSONAL PLUS	0.03																																																													
ingr_(4.4 5.5MM]	0.024																																																													
ingr_(8.7 Inf)	0.15																																																													
segm_PERSONAL PLUS	0.15																																																													
estu_UNIVERSITARIO	0.14																																																													
segm_PREFERENCIAL	0.13																																																													
eciv_CASADO	0.13																																																													
estu_POSTGRADO	0.09																																																													
edad_40-49	0.086																																																													
estu_ESPECIALIZACION	0.083																																																													
ingr_(3.3 4.4MM]	0.075																																																													
ingr_(4.4 5.5MM]	0.073																																																													
<div>Principales características para "Otros servicios financieros"</div> <table><tr><td>segm_PERSONAL PLUS</td><td>0.085</td></tr><tr><td>ingr_(8.7 Inf)</td><td>0.081</td></tr><tr><td>estu_UNIVERSITARIO</td><td>0.08</td></tr><tr><td>segm_PREFERENCIAL</td><td>0.066</td></tr><tr><td>estu_ESPECIALIZACION</td><td>0.063</td></tr><tr><td>eciv_CASADO</td><td>0.054</td></tr><tr><td>estu_POSTGRADO</td><td>0.042</td></tr><tr><td>ingr_(4.4 5.5MM]</td><td>0.04</td></tr><tr><td>ingr_(5.5 6.6MM]</td><td>0.038</td></tr><tr><td>ingr_(3.3 4.4MM]</td><td>0.038</td></tr></table>	segm_PERSONAL PLUS	0.085	ingr_(8.7 Inf)	0.081	estu_UNIVERSITARIO	0.08	segm_PREFERENCIAL	0.066	estu_ESPECIALIZACION	0.063	eciv_CASADO	0.054	estu_POSTGRADO	0.042	ingr_(4.4 5.5MM]	0.04	ingr_(5.5 6.6MM]	0.038	ingr_(3.3 4.4MM]	0.038	<div>Principales características para "Almacenes de cadena"</div> <table><tr><td>estu_UNIVERSITARIO</td><td>0.077</td></tr><tr><td>segm_PERSONAL PLUS</td><td>0.067</td></tr><tr><td>edad_30-39</td><td>0.057</td></tr><tr><td>eciv_CASADO</td><td>0.053</td></tr><tr><td>ingr_(8.7 Inf)</td><td>0.051</td></tr><tr><td>ingr_(3.3 4.4MM]</td><td>0.05</td></tr><tr><td>estu_ESPECIALIZACION</td><td>0.045</td></tr><tr><td>ingr_(4.4 5.5MM]</td><td>0.043</td></tr><tr><td>segm_PREFERENCIAL</td><td>0.039</td></tr><tr><td>ingr_(5.5 6.6MM]</td><td>0.035</td></tr></table>	estu_UNIVERSITARIO	0.077	segm_PERSONAL PLUS	0.067	edad_30-39	0.057	eciv_CASADO	0.053	ingr_(8.7 Inf)	0.051	ingr_(3.3 4.4MM]	0.05	estu_ESPECIALIZACION	0.045	ingr_(4.4 5.5MM]	0.043	segm_PREFERENCIAL	0.039	ingr_(5.5 6.6MM]	0.035	<div>Principales características para "Seguros"</div> <table><tr><td>ingr_(8.7 Inf)</td><td>0.083</td></tr><tr><td>segm_PERSONAL PLUS</td><td>0.082</td></tr><tr><td>segm_PREFERENCIAL</td><td>0.064</td></tr><tr><td>estu_UNIVERSITARIO</td><td>0.06</td></tr><tr><td>eciv_CASADO</td><td>0.054</td></tr><tr><td>estu_ESPECIALIZACION</td><td>0.045</td></tr><tr><td>edad_30-39</td><td>0.043</td></tr><tr><td>estu_POSTGRADO</td><td>0.04</td></tr><tr><td>ingr_(4.4 5.5MM]</td><td>0.039</td></tr><tr><td>ingr_(5.5 6.6MM]</td><td>0.036</td></tr></table>	ingr_(8.7 Inf)	0.083	segm_PERSONAL PLUS	0.082	segm_PREFERENCIAL	0.064	estu_UNIVERSITARIO	0.06	eciv_CASADO	0.054	estu_ESPECIALIZACION	0.045	edad_30-39	0.043	estu_POSTGRADO	0.04	ingr_(4.4 5.5MM]	0.039	ingr_(5.5 6.6MM]	0.036
segm_PERSONAL PLUS	0.085																																																													
ingr_(8.7 Inf)	0.081																																																													
estu_UNIVERSITARIO	0.08																																																													
segm_PREFERENCIAL	0.066																																																													
estu_ESPECIALIZACION	0.063																																																													
eciv_CASADO	0.054																																																													
estu_POSTGRADO	0.042																																																													
ingr_(4.4 5.5MM]	0.04																																																													
ingr_(5.5 6.6MM]	0.038																																																													
ingr_(3.3 4.4MM]	0.038																																																													
estu_UNIVERSITARIO	0.077																																																													
segm_PERSONAL PLUS	0.067																																																													
edad_30-39	0.057																																																													
eciv_CASADO	0.053																																																													
ingr_(8.7 Inf)	0.051																																																													
ingr_(3.3 4.4MM]	0.05																																																													
estu_ESPECIALIZACION	0.045																																																													
ingr_(4.4 5.5MM]	0.043																																																													
segm_PREFERENCIAL	0.039																																																													
ingr_(5.5 6.6MM]	0.035																																																													
ingr_(8.7 Inf)	0.083																																																													
segm_PERSONAL PLUS	0.082																																																													
segm_PREFERENCIAL	0.064																																																													
estu_UNIVERSITARIO	0.06																																																													
eciv_CASADO	0.054																																																													
estu_ESPECIALIZACION	0.045																																																													
edad_30-39	0.043																																																													
estu_POSTGRADO	0.04																																																													
ingr_(4.4 5.5MM]	0.039																																																													
ingr_(5.5 6.6MM]	0.036																																																													
<div>Principales características para "Comida"</div> <table><tr><td>gen_Femenino</td><td>0.047</td></tr><tr><td>ocup_EMPLEADO</td><td>0.035</td></tr><tr><td>edad_30-39</td><td>0.03</td></tr><tr><td>estu_UNIVERSITARIO</td><td>0.029</td></tr></table>	gen_Femenino	0.047	ocup_EMPLEADO	0.035	edad_30-39	0.03	estu_UNIVERSITARIO	0.029	<div>Principales características para "Ahorro"</div> <table><tr><td>estu_UNIVERSITARIO</td><td>0.02</td></tr></table>	estu_UNIVERSITARIO	0.02	<div>Principales características para "Moda"</div> <table><tr><td>gen_Femenino</td><td>0.048</td></tr></table>	gen_Femenino	0.048																																																
gen_Femenino	0.047																																																													
ocup_EMPLEADO	0.035																																																													
edad_30-39	0.03																																																													
estu_UNIVERSITARIO	0.029																																																													
estu_UNIVERSITARIO	0.02																																																													
gen_Femenino	0.048																																																													



Se encuentra mediante este ejercicio que es posible identificar algunas de las características principales de los usuarios que realizan cada tipo de transacción, lo cual puede ser de gran utilidad para el Banco o terceros con los cuales este se quiera aliar, pues se puede identificar qué tipos de clientes son más propensos a realizar transacciones de los tipos específicos.

Por ejemplo, la categoría “Seguros”, suele ser pagada a través de PSE (posiblemente son los que adquieren este tipo de productos), principalmente por personas de ingresos altos, del segmento preferencial plus o superior, casados, y con título universitario o superior.


### 5. Aplicación – Proactive Financial Management

Como se ha comentado, lo que se busca con este ejercicio es lograr aplicar los diferentes modelos propuestos, para lo cual se diseñaron una serie de mockups que sirvan como base para una definición de cómo llevar algunos de los modelos diseñados a la aplicación MFP del cliente.

El siguiente link brinda acceso al mockup diseñado. También es posible descargar el aplicativo y evaluar de manera preliminar su potencial de funcionamiento y apropiación de cara a los clientes.  
<https://marvelapp.com/3831cd3>

Es importante resaltar en este punto que una de las causas por las cuales los aplicativos de PMF no suelen tener mucho éxito es el ingreso manual por parte de los clientes de la información correspondiente a las transacciones.

Es así como se evidencia una gran oportunidad si consideramos la posible integración de los servicios actualmente ofrecidos por el Banco y otras plataformas e información que puede ser recaudada.

Imagen	Descripción
	<p><b>Histórico de datos:</b> por medio de esta aplicación es posible brindar al usuario (quien en un principio no deberá ingresar ningún tipo de información relativa a sus gastos e ingresos) la información de comportamiento financiero para diferentes periodos. Adicionalmente es posible brindarle información de la distribución de sus gastos y visualizarlos en el aplicativo.</p>

TIGO LTE

9:55 p. m.

46 %

Atrás

Bancolombia

Salir

Productos

ZENaida GONZÁLEZ

Su último ingreso fue el: 2018/10/28 21:31:12

IP: 177.252.244. 98

Mis Finanzas - Este Mes

2018/10/28

TIGO - COLOMBIA MOVIL S.A. ESP

\$ -49,900.00

Tecnología y comunicaciones

2018/10/28

PAGO PSE TIGO - COLOMBIA MOVI

\$ -49,900.00

Tecnología y comunicaciones

2018/10/27

COMPRA EN DOMINOS PI

\$ -19,400.00

Comida

2018/10/24

COMPRA EN CAFETERIA

\$ -20,000.00

Comida

2018/10/21

PAGO PSE Empresas Publicas de

\$ -388,888.00

Hogar

Productos

Transferencias

PFM

Seguridad

Cada uno de los movimientos que se hagan ya sea vía medio de pago Redeban o PSE, pueden ser categorizados y dispuestos para hacer uno de ellos mediante los diferentes modelos y aplicaciones ya expuestos

TIGO LTE

9:55 p. m.

46 %

Atrás

Bancolombia

Salir

Productos

ZENaida GONZÁLEZ

Su último ingreso fue el: 2018/10/28 21:31:12

IP: 177.252.244. 98

Mis Finanzas - Este Mes

Categoría	Monto
Almacenes De Cadena	\$ 0
Comida	\$ 100.000
Cuidado Personal	\$ 0
Educación	\$ 0
Entretenimiento	\$ 0
Gobierno E Impuestos	\$ 0

Productos

Transferencias

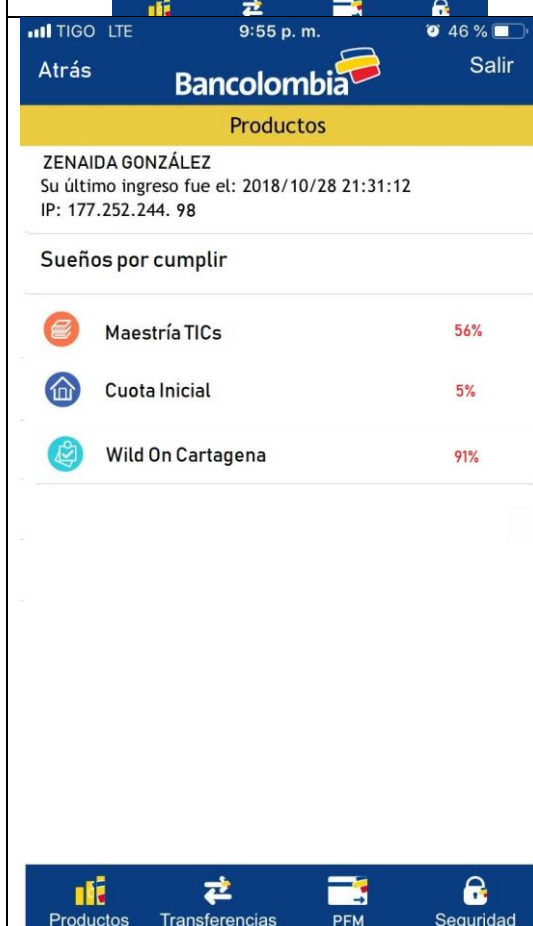
PFM

Seguridad

**Presupuesto:** Definición y seguimiento.



**Perfilamiento de clientes:** en ocasiones, la mejor manera de aprender es viendo como otras personas lo realizan. Es así pues, como basado en la información de los clientes y en la implementación de diferentes técnicas de Inteligencia artificial, machine learning, entre otras, es posible realizar una comparación del usuario, con otras personas que tengan características similares a las suyas e incluso tener acceso a diferentes combinaciones de presupuestos que le puedan orientar.



**Sueños:** reconocer en los clientes sus metas y aspiraciones mayores siempre será importante. Este tipo de aplicativos, permite tener una relación de mayor cercanía con el cliente, quién verá en el banco, más que el lugar en el cual deposita su dinero, un aliado estratégico.