Segmentación de clientes de aerolíneas y cálculo del *Customer Lifetime Value*

Segmentation of airlines' customers and Customer Lifetime Value

III Master Data Science

Clara Pizarro Valle

1. Índice

2. Introducción....3
Objetivos del proyecto....3
Segmentación de clientes....3
3. Descripción de los datos de entrada....4
Carga y análisis de datos....5
Limpieza de datos....6
Creación de Base de Datos en PostgreSQL y tablas para la visualización....6
4. Metodología....7
5. Resultados y conclusiones....9
Modelo RFM....9
Clasificación por características sociodemográficas....12
Customer Lifetime Value....13
6. Visualización....14

2. Introducción

Objetivos del proyecto

El objetivo del proyecto es **determinar las características de los clientes de las aerolíneas de vuelos** y su relación con éstas, pudiendo agruparlos en base a condiciones sociodemográficas y comportamentales similares, así como poder **predecir el valor que aportarán a las empresas de vuelos.**

A partir de la segmentación del mercado realizada y el cálculo del valor del cliente a lo largo de su ciclo de vida, podrán construirse estrategias de marketing particularizadas para la captación de nuevos clientes, la fidelización de los actuales y el aumento de la satisfacción con el servicio.

Segmentación de clientes

Las empresas que se enfocan en "Gestionar las Relaciones con el Cliente" dividen el mercado en grupos uniformes más pequeños que tienen características y necesidades semejantes.

Estos grupos se denominan segmentos y se caracterizan por:

- Homogeneidad
- Heterogeneidad
- Rentabilidad
- Accesibilidad

Las empresas que segmentan el mercado definen su negocio y estructura organizacional en base a la segmentación, y dirigen la **estrategia de relacionamiento**, en función de ella.

Los datos son el input actual principal de la industria de las aerolíneas: cuanto más se conoce al usuario, pueden lanzarse más ofertas ajustadas a sus características y necesidades, aumentando las posibilidades de conversión.

El análisis de los datos y el empleo de técnicas de *machine learning* permite a la industria de las aerolíneas poder crear segmentaciones de usuarios. Esta **segmentación** puede realizarse bajo cualquier variable: **comportamiento, dispositivo, canal utilizado o producto comprado.** Así, por ejemplo, se puede incluir en el mismo segmento a todas aquellas personas que iniciaron el proceso de compra de un billete aéreo, pero lo abandonaron en el mismo momento.

A partir de la segmentación de los clientes que se efectúa en este proyecto, se obtendrán unos resultados que permitirán **implementar acciones específicas por cada tipo de cliente**, con el fin de ofrecer billetes con los precios más adecuados y atractivos para cada uno de ellos, fomentar la fidelización e incrementar su retorno.

En el marco del presente proyecto, se ha llevado a cabo **una segmentación de clientes** mediante la técnica de machine learning denominada ***K-Means***, a través de la cual, se han identificado una serie de patrones comportamentales comunes en los pasajeros, así como detectado características demográficas similares.

El resultado de esta segmentación ha permitido **diferenciar aquellos clientes que aportan mayor valor a las compañías, así como detectar los menos relevantes en términos económicos.**

Asimismo, se ha realizado el cálculo del *Customer Lifetime Value* (CLV), una variable que permite predecir el valor que el cliente puede aportar a la compañía, a lo largo de su relación con ésta. A través de la estimación del CLV pueden conocerse aquellos clientes más rentables para la compañía y obtener el *Return of Investment* (ROI) derivado de las acciones de fidelización y retención aplicadas sobre ellos.

3. Descripción de los datos de entrada

Los datos de entrada para el estudio parten de un archivo en formato texto de un volumen de 3GB, siendo las variables utilizadas, las siguientes:

Variable	Info
rloc	Código identificador del viaje
gender	Género del pasajero
age	Edad
date_of_birth	Fecha de nacimiento
document_number	Número del documento de identidad
Nationality	Nacionalidad
Cabin_code	Clase del asiento del pasajero: First F o C, Business J, Economy Premium W, Economy Y
departure_date_leg	Fecha de salida del vuelo
Quality_index	Calidad del dato

Variable	Info
creation_date	Fecha de compra del billete
advance_purchase	Tiempo transcurrido entre la fecha de reserva y fecha del vuelo
Booking_status_code	Código del estado de la reserva
Board_point	Código del aeropuerto de embarque
Board_lat	Latitud del punto de embarque
Board_lon	Longitud del punto de embarque
Board_country_code	Código del país de embarque
Board_continent_code	Código del continente de embarque
Off_point	Aeropuerto de final de trayecto
Off_lat	Latitud del punto de aterrizaje
Off_lon	Longitud del punto de aterrizaje
Off_country_code	Código del país de aterrizaje
Off_continent_code	Código del continente de aterrizaje
Distance_seg	Distancia en Km del segmento
Revenue_amount_seg	Ingresos
Route	Ruta

Carga y análisis de datos

Debido al volumen de datos con el que se trabaja, se han cargado y limpiado los datos con Spark y, en concreto, la API Spark SQL.

Se **leen y analizan los datos** con Spark SQL:

- Se contabiliza el total de datos: 9.546.302
- Se obtiene el esquema de los mismos
- Se identifican las columnas y su tipología
- Se detectan valores NULL
- Se contabiliza el número de billetes existente (excluyendo los NULL): **2.636.677**

- Se obtienen las **fechas de los datos:**

```
+------------------------------+
|first(departure_date_leg, true)|
+------------------------------+
|           2012-06-27 00:00:...|
+------------------------------+

+------------------------------+
|last(departure_date_leg, true)|
+------------------------------+
|           2014-07-14 00:00:...|
+------------------------------+
```

- Se detectan las fechas desde y hasta las cuales hay un volumen de datos más homogéneo.

Limpieza de datos

A partir de la información obtenida, se empieza el proceso de limpieza de datos:

- Se escogen únicamente los datos no NULL y los NaNs se reemplazan por ceros
- Se escogen los datos cuyo estado de reserva es HK o confirmado
- Se escogen los datos cuya calidad es 1 (quality_index = 1)
- Se escoge el rango de fechas para el análisis:
  - 01-01-2012
  - 01-01-2014
- Se obtienen los datos de documentos únicos: 175.333
- Se obtienen los datos de clientes únicos: 523.541 (con *document_number*, *gender*, *date_of_birth* y *nationality* no duplicados).

Puede comprobarse el código en el archivo Cleaned dataset and RFM clustering.ipynb

Creación de Base de Datos en PostgreSQL y tablas para la visualización

A medida que se han ido generando tablas finales, éstas se han convertido al formato .csv y se han creado en una base de datos PostgreSQL, a la cual se conecta Tableau para la visualización.

4. Metodología

Para realizar la segmentación objeto del proyecto, se ha utilizado el modelo RFM sobre Python.

El modelo RFM es una metodología con la que se pueden realizar segmentaciones de clientes según características similares. Se creó hace más de 75 años, principalmente para los vendedores directos, para poder satisfacer a las finanzas mejorando los beneficios. Fue muy popular para los pioneros del marketing de base de datos (Stan Rapp, Tom Collins, David Pastor, Arthur Hughes, etc).

El modelo RFM contempla la Recencia, Frecuencia y Valor Monetario para cada cliente a partir del cual se puede determinar el comportamiento o evolución de compra de los mismos. Es un modelo fácil de entender, de explicar e implementar.

Así:

- **Recency:** representa cuándo fue la última vez que compró cada cliente. Concretamente, se miden los días que han pasado desde hoy (o cualquier fecha a futuro) hasta la fecha en que el cliente realizó su última compra.
- **Frequency:** representa cuántas veces ha comprado. En detalle, se mide el número de compras que ha realizado el cliente en total en el periodo estudiado.
- **Money:** cuánto dinero se ha gastado en total. Concretamente, es la suma total de cantidad de dinero que el cliente lleva gastado en sus compras

El RFM sigue la premisa de que "los más propensos a comprar son aquellos que han comprado más recientemente, con más frecuencia y gastan más dinero". Se basa en la conocida "Ley de Pareto" o del "80/20" enunciada por el economista italiano Vilfredo Pareto, en el siglo XIX. El RFM se aplica sobre esta "Ley de Pareto" y se refiere a que "el 80% de las compras las realizan el 20% de los clientes".

Para la segmentación de los clientes, se ha aplicado el **algoritmo de clustering *K-Means***.

K-Means divide las observaciones en un número (k) predefinido de clúster en los que cada observación pertenece al clúster con la media más cercana.

Una vez aplicado el algoritmo y obtenidos los clústeres, se han correlado con las características demográficas, aplicándose tanto el modelo **Decision Tree como Random**

Forest para estimar el grado de lealtad basado en estas características.

Para el **cálculo del CLV** se han empleado:

- El **modelo Beta Negative/NGB model,** que estima la probabilidad de una próxima compra y evalúa cuántos clientes están todavía activos dado su historial de compra. En este modelo se tiene en cuenta el tiempo que lleva siendo cliente, la recencia y la frecuencia. A partir de estas variables se estima la probabilidad de compra (p=1) o no (p=0).
- El **modelo Gamma-Gamma,** que predice el valor de las futuras compras.
  Este modelo se basa en tres principios:
  - que el valor monetario de la transacción determinada de un cliente varía al azar alrededor del valor medio de las transacciones
  - que los valores promedio de la transacción varían entre los clientes, pero a nivel de cada cliente este se mantiene estable a lo largo del tiempo
  - que la diferencia de la distribución de los valores medios de transacción de cada cliente es independiente de la transacción en sí.

5. Resultados y conclusiones

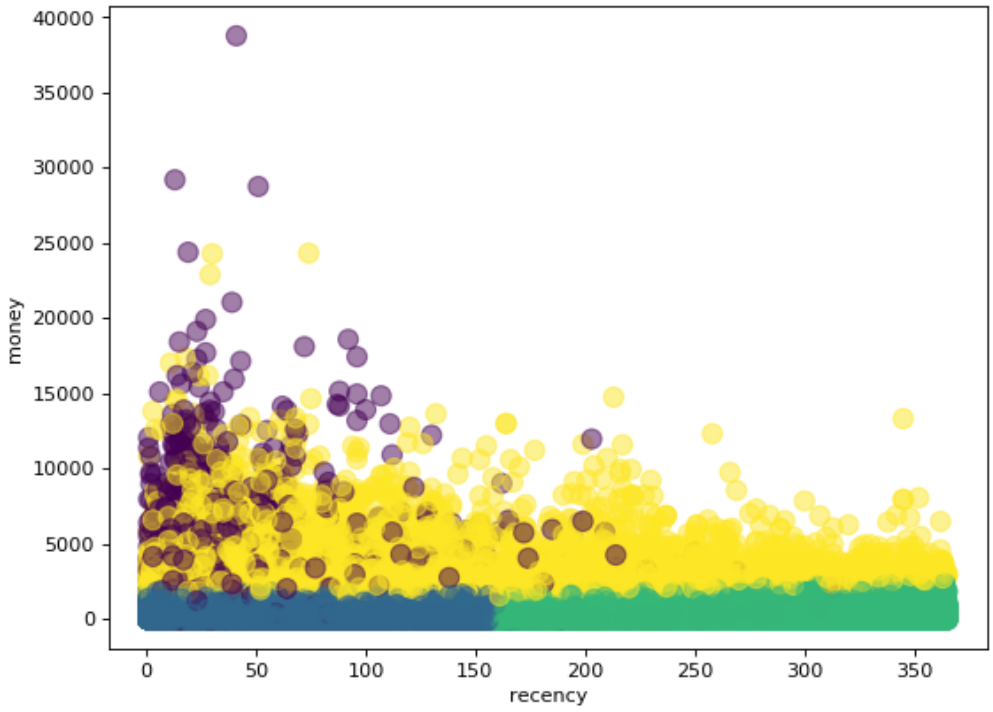
Los resultados han sido obtenidos a partir la aplicación de los algoritmos detallados en la metodología, con la librería SKLearn de Python. Para la visualización de los resultados se ha utilizado Tableau.

Modelo RFM

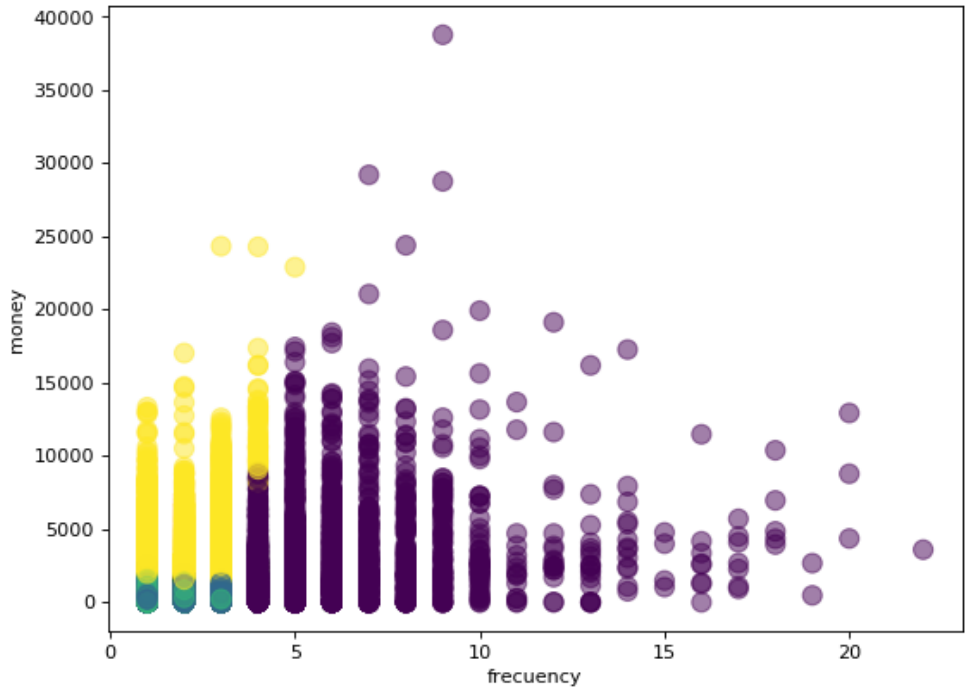
Se ha generado una tabla en Spark SQL con las siguientes variables, para aplicar posteriormente con SKLearn el algoritmo *K-Means* de segmentación de clientes:

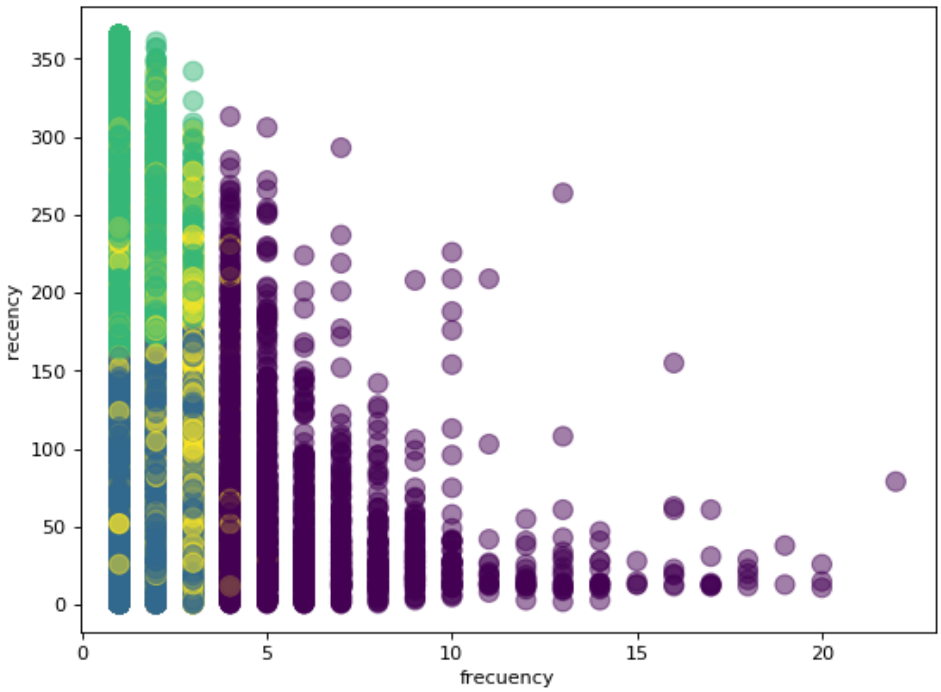
Variable	Info
Document_number	Código único identificador del cliente
Frequency	Frecuencia de compra de billetes en el histórico seleccionado
Recency	Tiempo transcurrido desde la compra del último billete
Money	Importe por viaje realizado

Se ha ejecutado con K-Means a través de un bucle que ha generado segmentaciones de 1 a 8 grupos. Al analizar la segmentación, se determina que la **segmentación de 4 grupos** es la que presenta una distribución de clientes más heterogénea.



Puede comprobarse el código en el archivo [Cleaned dataset and RFM clustering.ipynb]





A continuación, se presentan los datos principales de los 4 segmentos.

Cluster	Color	Name	Nº	Recency			Frequency			Money		
				Min	Max	Avg	Min	Max	Avg	Min	Max	Avg
0	**Purple**	Habituales	3.287	1	313	48,77	4	22	5,35	-	38.767,00	2.457,13
1	**Blue**	Nuevos	99.535	1	179	62,96	1	3	1,14	-	1.922,18	361,15
2	**Green**	Desvinculados	62.175	158	365	256,31	1	3	1,04	-	2.410,00	317,36
3	**Yellow**	Vacacionales	10.336	1	365	120,62	1	5	1,57	1.190,00	24.326,00	3.002,00

Segmento 0. Morado. Habituales

El segmento 0 está compuesto por clientes con una frecuencia de compra elevada, con frecuencia mínima de 4 y máxima de 22. Destacan por presentar **la mayor frecuencia y la menor recencia de todos los segmentos obtenidos,** con una media de 48,77 días y un gasto medio de 2.457 euros.

Segmento 1. Azul. Nuevos

Este segmento es el que presenta mayor volumen de clientes (57%). Presentan una **frecuencia de compra de 1 vez;** sin embargo, la **recencia media se sitúa en 63 días.** Son clientes que han realizado su primera compra hace poco.

Segmento 2. Verde. Desvinculados

Este segmento presenta 62.175 clientes (33%), con una frecuencia media de 1, una recencia mínima de 158 días y 365 días de máxima. **Presentan el menor gasto de todos los segmentos obtenidos. Su frecuencia es baja** (de 1 a 3 compras), por lo que se definirá a este segmento como el de los clientes desvinculados.

Segmento 3. Amarillo. Vacacionales

A este segmento pertenecen los clientes que **han comprado hasta 5 veces** y presentan una **recencia media de 120 días (4 meses).** Estos clientes presentan un **gasto medio muy superior por viaje,** 3.002 €. Al presentar un comportamiento estacional, se podrían asignar la categoría de clientes vacacionales.

Clasificación por características sociodemográficas

Con el objetivo de poder clasificar a futuros clientes en uno de los 4 segmentos obtenidos, se han aplicado los modelos anteriormente mencionados: ***Decision Tree y Random Forest*.**

Puede comprobarse el código en el archivo Clustering new airline customers.ipynb

Se han utilizado, para la clasificación, las siguientes variables:

- edad
- género
- distancia recorrida
- antelación en la compra del billete
- tipo de cliente (*Business*, *Economy*, *Premium*)
- cliente con tarjeta de fidelización
- tipo de vuelo (Internacional o Nacional)
- canal de compra.

La variable dependiente ha sido calculada a partir de los percentiles de la variable Money obtenida en el modelo RFM.

Antes de estimar los modelos se ha realizado un análisis gráfico en el que se muestra la distribución de cada una de las variables independientes frente a la pertenencia de cada uno de los segmentos resultantes del *K-Means*, concluyéndose lo siguiente:

- No existen diferencias significativas en cuanto a género y sexo, por lo que probablemente estas dos variables no tienen un peso significativo en los modelos de clasificación.
- Se detectan distribuciones distintas en el tipo de cliente, en el canal de compra, en el tipo de vuelo, en la distancia y en la antelación de compra, siendo especialmente relevante en estas dos últimas.
- Para ambos modelos, en el código se muestra el ranking por importancia de cada una de las variables de clasificación, obteniéndose lo siguiente:
    - En el modelo *Decision Tree*, la principal variable para clasificar a un pasajero en un segmento u otro es si **vuela en Business**, seguida por la **distancia**, la **antelación en la compra** y, por último, si utiliza el **canal de la aerolínea para la reserva** del vuelo.
    - En el modelo *Random Forest*, se sitúa como **primera variable para clasificar, la distancia**

recorrida, seguida por si vuela en Business o no.

- El resto de variables no son relevantes.

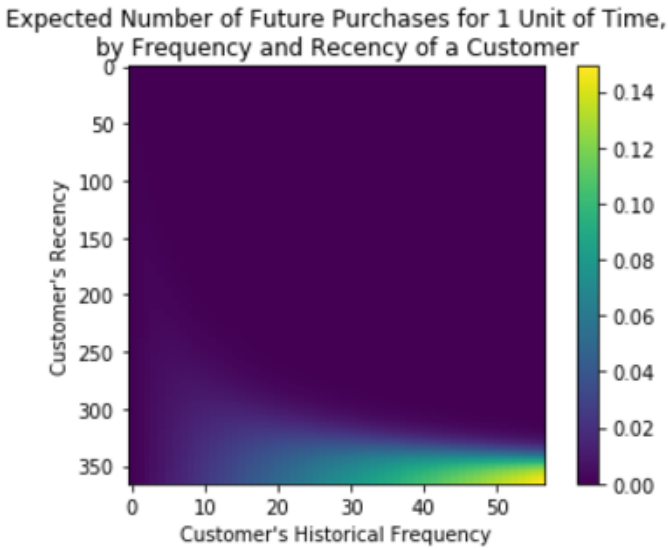
Customer Lifetime Value

Como antes se ha comentado, el objetivo de este modelo se sustenta en la **estimación del valor monetario del cliente** y en la **predicción de futuras compras.**

Puede comprobarse el código en el archivo Customer Lifetime Value.ipynb

Si se analiza la matriz Frecuencia–Recencia, se observa que si un cliente ha comprado con una frecuencia mayor a 50 veces y su última compra fue hace 355 días, **la probabilidad de que compre en los próximos instantes es muy elevada. Los clientes situados en la matriz en el área amarilla son los "mejores clientes".**

Los **"peores clientes" se sitúan en el extremo superior derecho;** compraron rápidamente pero no han vuelto a comprar en el periodo estudiado.



El modelo Gamma–Gamma permite asignar un valor a cada una de las compras futuras.

Customer ID	Valor
1	24.658622
2	18.911496
3	35.170995

Como conclusiones, el **CLV** puede ser utilizado para **modelar patrones de fugas de clientes,** para mejorar el rendimiento

del servicio de atención al cliente, para optimizar la identificación del público objetivo de cada una de las campañas de marketing y aplicar técnicas de *Cross Selling*.

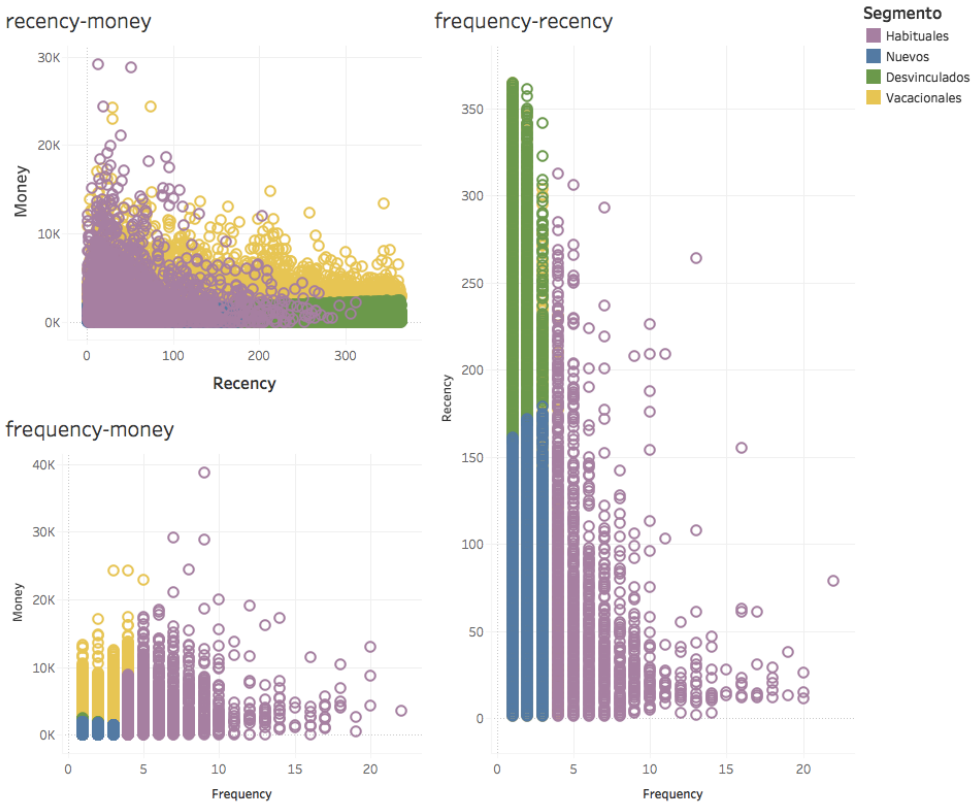
6. Visualización

Aun habiendo realizado visualizaciones en Python, se ha empleado Tableau para llevar a cabo *Dashboards* con el objetivo de poder filtrar y analizar de manera dinámica los resultados alcanzados.

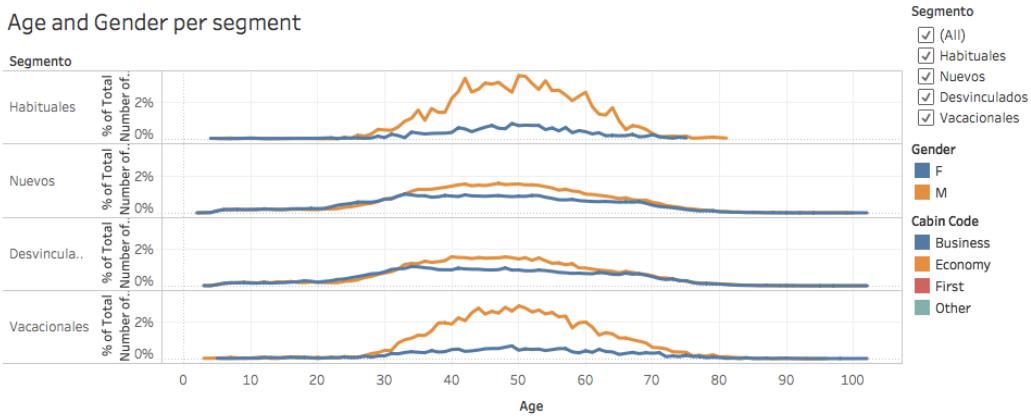
Los datos a partir de los cuales se realiza la visualización se encuentran en el servidor PostgreSQL. No obstante el código se puede comprobar en Tablas PostgreSQL

La primera visualización que se ejecuta es la resultante de la segmentación obtenida tras ejecutar K-Means. De esta manera podemos percibir cómo se relacionan las variables Frequency, Recency y Money de la segmentación RFM.

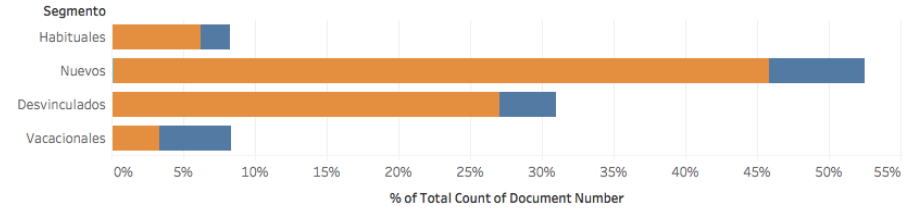
Se puede obtener la visualización en el siguiente archivo: Segmentación.twb



Age and Gender per segment



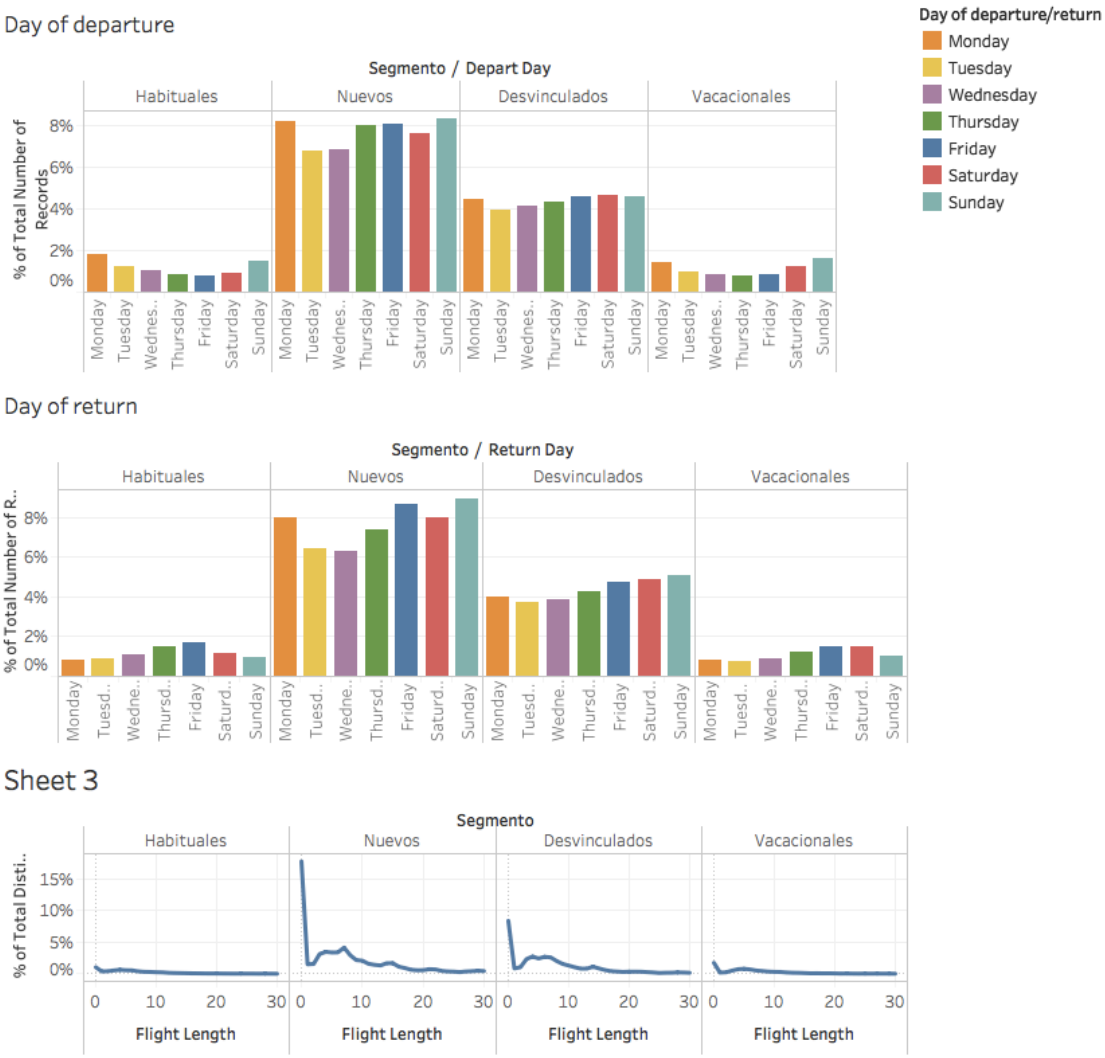
Cabin code per segment



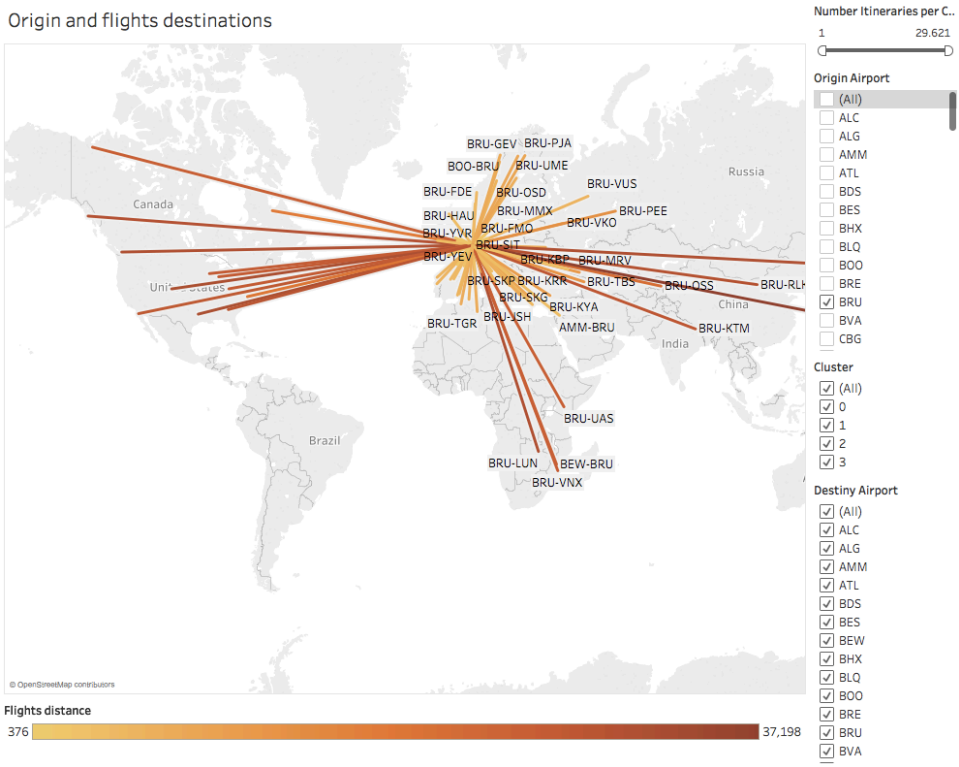
La siguiente visualización muestra la frecuencia de vuelo por segmento y día de la semana, así como el día en el que más se viaja (ida/vuelta).

Se puede obtener la visualización en el siguiente archivo:

Journey lenght.twb



Finalmente se ha generado un *dashboard* en formato mapa en el que se puede **filtrar por origen y destino de los vuelos, aeropuertos, distancia recorrida y número de itinerarios realizados por segmento.** Se puede visualizar en el siguiente archivo: Airports_visualization.twb


Origin and flights destinations
BRU-GEV
BRU-PJA
BOO-BRU
BRU-UME
BRU-VUS
BRU-FDE
BRU-OSD
BRU-HAU
BRU-MMX
BRU-PEE
BRU-YVR
BRU-FMO
BRU-VKO
BRU-SLT
BRU-YEV
BRU-KBP
BRU-MRV
BRU-SKP
BRU-KRR
BRU-TBS
BRU-OSS
BRU-SKG
BRU-KYA
BRU-JSH
BRU-TGR
AMM-BRU
BRU-KTM
BRU-UAS
BRU-LUN
BEW-BRU
BRU-VNX
Canada
Russia
China
India
Brazil
© OpenStreetMap contributors
Flights distance
376
37,198
Number Itineraries per C..
1
29.621
Origin Airport
(All)
ALC
ALG
AMM
ATL
BDS
BES
BHX
BLQ
BOO
BRE
BRU
BVA
CBG
Cluster
(All)
0
1
2
3
Destiny Airport
(All)
ALC
ALG
AMM
ATL
BDS
BES
BEW
BHX
BLQ
BOO
BRE
BRU
BVA