Detección de Clusters formados por turistas extranjeros y nacionales en la Ciudad Autónoma de Buenos Aires entre 2016 y 2018

Germán De Luca, Kevin Dobzewicz, Juan Dario Rodriguez Piro

Informe Final, Cluster AI

UTN FRBA 2020

Abstract: En este proyecto se analizarán datasets obtenidos de centros turísticos de la Ciudad de Buenos Aires, los cuales contienen información acerca de turistas de distintas nacionalidades.

Keywords: Turismo, CABA, Clusters, PCA.

Objetivo del proyecto

El objetivo del siguiente trabajo es analizar datos sobre turismo de la Ciudad de Buenos Aires con el fin de identificar grupos de turistas que presenten características similares. Se buscarán dichos grupos con el fin de encontrar comportamientos determinados dentro de cada uno de ellos.

1. Introducción

Miles de turistas de distintas nacionalidades eligen la Ciudad de Buenos Aires como destino vacacional debido a su historia y diversidad cultural. Estos son atraídos por distintos puntos de interés de la capital, quedando registradas sus consultas en distintos "Centros de atracción turística". Gracias a estas consultas, es posible identificar características en los distintos grupos, tales como su nacionalidad, número de integrantes y pernoctaciones. Asociando esas características junto al tipo de cambio vigente en las fechas coincidentes con los viajes, se pretende utilizar un modelo de aprendizaje no supervisado para evaluar la formación de Clusters, analizando la calidad de los mismos.

Para el análisis, se tomaron datasets provenientes de la Ciudad de Buenos Aires, las cuales fueron provistos por centros de atracción turística a lo largo de la ciudad, recolectando información de turistas. Dichos datos comprenden el periodo de tiempo entre 2016 y 2018.

2. Preprocesamiento de datos

Para el análisis se parte de ***dos*** datasets base, los cuales son resultados de encuestas realizadas en centros de atención turística de la Ciudad de Buenos Aires. El primer dataset contiene los resultados de encuestas en 2016, mientras que el segundo contiene datos de 2017 y 2018 [1].

Como primer paso del preprocesamiento, se precedió a homogeneizar los datos en ambos datasets. Tanto el archivo de 2016 como el de 2017/2018 tenían columnas similares, sin embargo, el formato de los datos comprendidos en las instancias variaba mucho en algunos casos. Las columnas (features) más interesantes para el análisis fueron:

- Fecha
- CAT (centro de atracción turística): esta columna indica el centro de atracción que recabó los datos de turistas (ejemplo: Retiro, Recoleta, Palermo, etc.).
- Pasajeros: El número total de integrantes en un grupo.
- País: el país de origen, pudiendo ser extranjero o argentino.
- Provincia: En caso de que el encuestado haya sido argentino, se computará la provincia de procedencia.

- Pernoctaciones

Para la homogeneización, se llevaron las features de ambos datasets al mismo formato. Para ello se verificaron los datos faltantes en cada columna, se eliminaron tildes para unificar el texto, se llevaron todos los textos a mayúsculas y se modificaron los tipos de variables (string a entero), entre otras operaciones.

Luego de homogeneizados los datos se crearon datasets reducidos que contengan las columnas de interés para el análisis, nombradas con anterioridad.

El siguiente paso fue la unificación de los sub-datasets de 2016 y 2017/2018 en uno solo. Para esto, se usó el comando pd.concat, uniendo ambos datasets por filas.

Se utilizó de forma auxiliar un ***tercer dataset***, el cual contenía el tipo de cambio [2] organizado por fechas. Usando el comando pd.merge, unimos los datasets de datos turísticos y tipo de cambio, utilizando la fecha a modo de "key".

Dado que se trabajaba con variables categóricas, estas debieron ser transformadas para el procesamiento posterior. Se utilizó label encoder para transformar las variables "PAIS", "PROVINCIA" y "CAT". Por otro lado, se transformó las fecha a formato de pandas para poder extraer los días y usarlos como feature. Se creó una nueva variable categórica llamada "DIA", la cual también se transformó con label encoder.

Inicialmente se tenían dos datasets. El dataset de 2016 tenía 106.942 filas y 52 columnas.

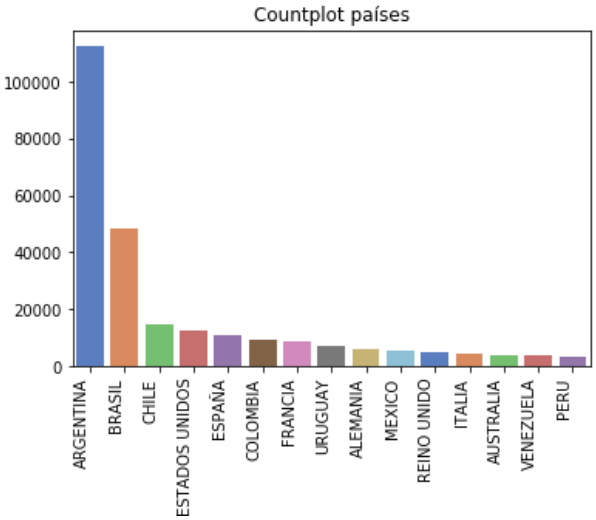
Por otro lado, el dataset de 2017 y 2018 tenía originalmente 174.805 filas y 23 columnas.

Luego de operar, se terminó con un dataset de 276.833 filas y 8 columnas.

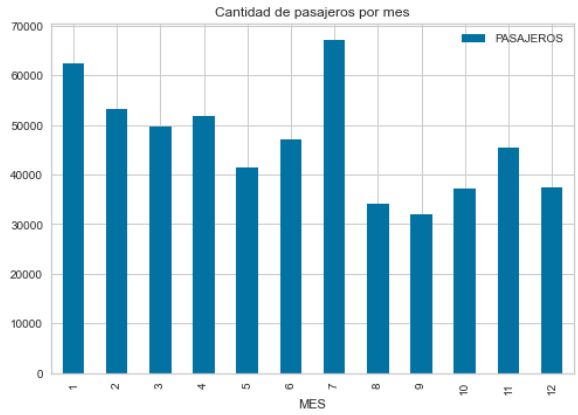
3. Análisis exploratorio de datos

A modo análisis, se realizaron una serie de gráficos para comprobar el comportamiento de algunas variables.

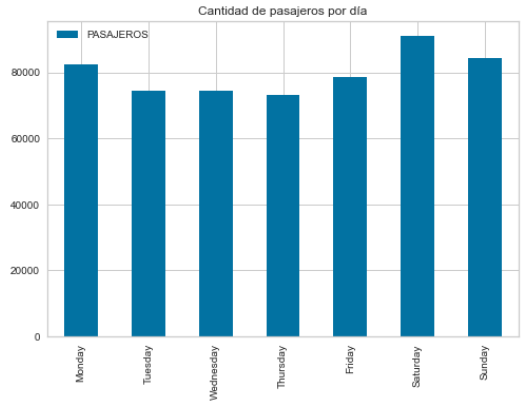
El primer análisis se realizó sobre la cantidad de registros de cada país, donde podemos observar que Argentina, Brasil y Chile están entre los primeros tres, seguidos por Estados Unidos.



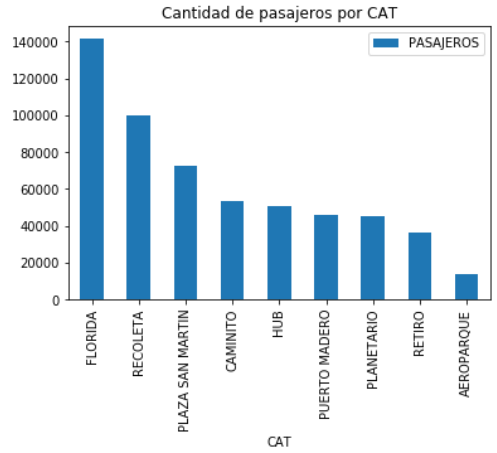
Luego, se llevó a cabo un conteo de pasajeros por mes, el cual muestra que los meses con más tránsito son enero, febrero, junio y julio.



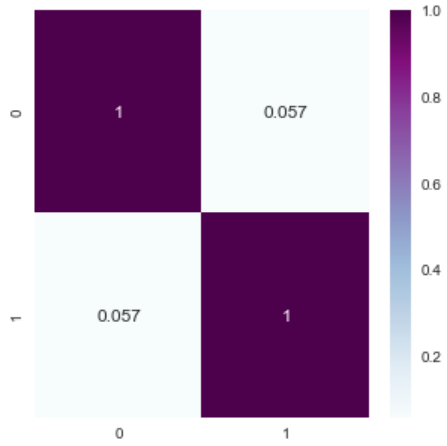
También se realizó un conteo de días más transitados, del cual podemos ver que los días con más circulación de turistas son sábado, domingo y lunes.



Otro punto a chequear fue la cantidad de consultas en cada CAT, siendo Florida, Recoleta, y Plaza San Martín los centros con mayor cantidad de consultas



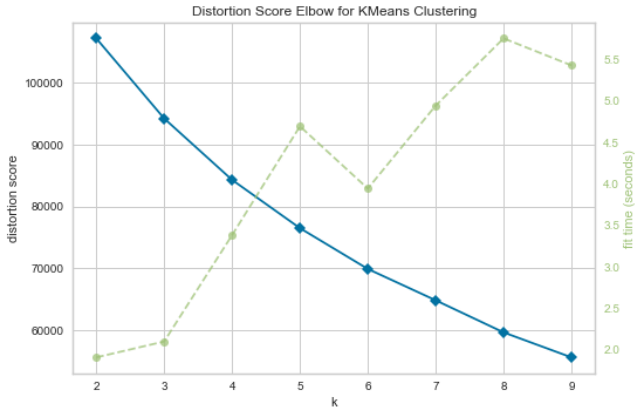
Por último, se chequeo correlación lineal entre tipo de cambio y cantidad de pasajeros, pero este análisis arrojó como resultado que no hay una correlación lineal.



5. Clusterización

En esta etapa del proyecto, se implementó un modelo de clusterización utilizando el algoritmo K – Means. El método consiste en la definición clusters y asignación de muestras según la distancia entre dicha muestra y un centroide que identifica a cada cluster. La distancia euclidiana cuadrática se usa para realizar la asignación de cada muestra a un cluster. Este algoritmo es iterativo, durante el proceso se recalculan los centroides y la pertenencia de las muestras a los clusters.

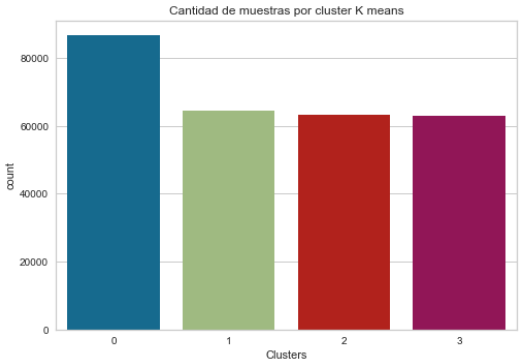
Para utilizar este algoritmo, se debe definir como hiper parámetro la cantidad de clusters. Para ello, se utilizó el método "elbow" [3].



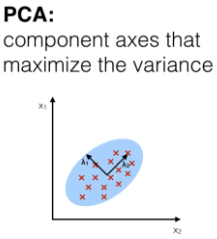
Se definió el hiperparámetro K (nº de clsuters) =4.

Definida la cantidad de clusters, se procedió a calcular los centroides de cada uno y a asignar las muestras.

Se obtuvieron los clusters 0, 1, 2 y 3 [4].

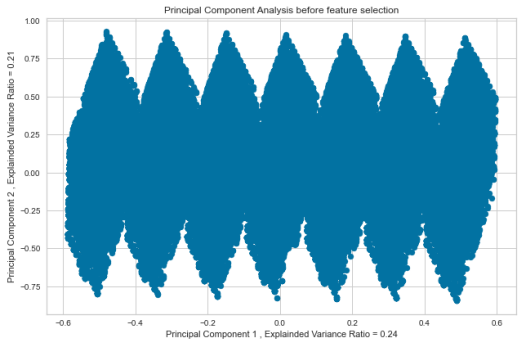


Luego, se utilizó **PCA** para visualizar los clusters. El método consiste en la descomposicion de la matriz de covarianza en sus autovectores y autovalores asociados.Luego se construye una matriz utilizando los “p” autovectores asociados a los “p” autovalores más grandes. Dicha matriz se utiliza para proyectar los datos en las componentes principales.



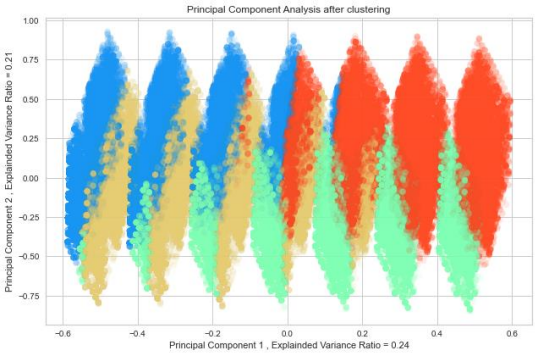
[5]

Se realizó PCA para visualizar los datos en dos dimensiones principales. Utilizando dos dimensiones, solamente se logró retener el 45% de la variabilidad en estas.



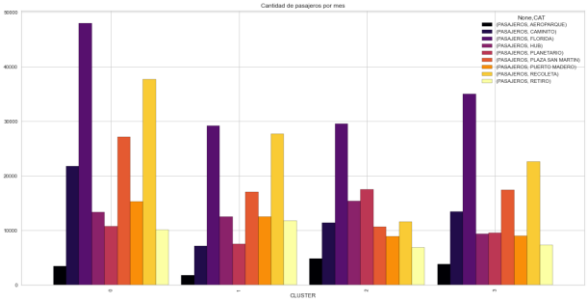
A simple vista, se puede ver que las clases no son linealmente separables.

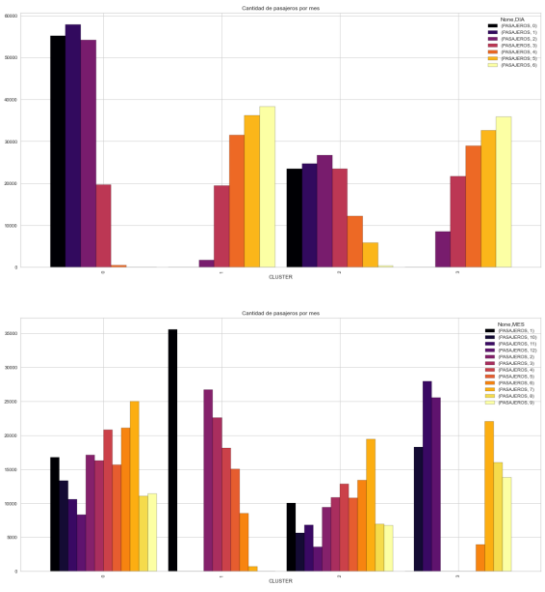
El Silhouette Score obtenido en la clusterización fue de 0,17, por lo que se puede intuir que se presentan clusters superpuestos.



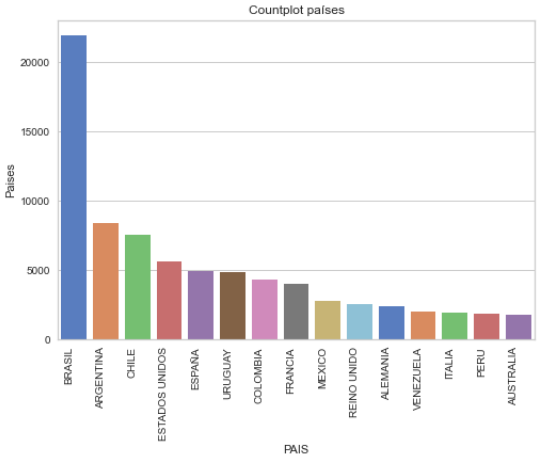
6. Conclusiones

Si bien se obtuvo un bajo Silhouette Score, lo que denota superposición de clusters, se pueden llegar a las siguientes conclusiones acerca de estos:

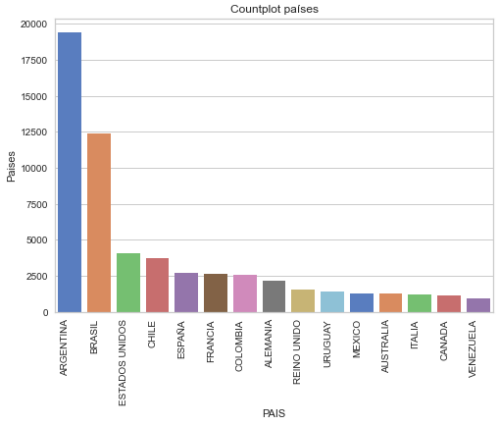




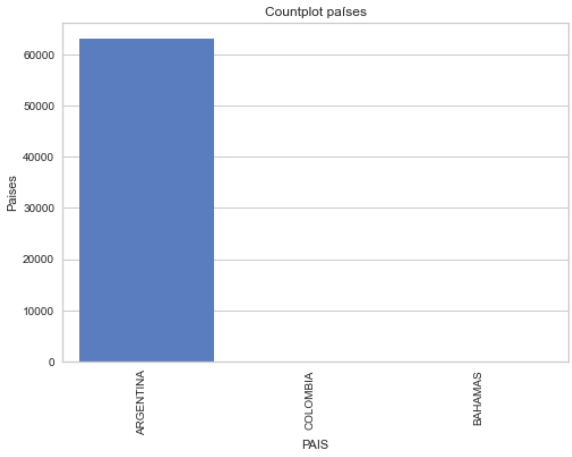
Cluster 0: Se caracteriza principalmente por turistas brasileros, con visitas principalmente los fines de semana con estadías estimadas de 7 días. Los lugares que más frecuentan son: Florida, Plaza San Martín y Recoleta. Suelen tener visitas constantes durante todo el año y con un leve aumento en invierno.



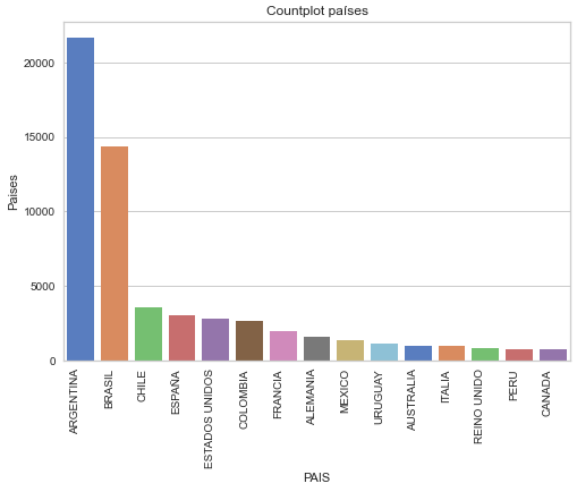
Cluster 1: Suelen frecuentar los lugares turísticos en los días de la semana, principalmente en los meses de enero, febrero y marzo, los orígenes suelen variar pero tiene un alto grado de incidencia turista de Santa Fe y Mendoza. Tiene una baja tendencia de estar en aeroparque.



Cluster 2: Integrado casi por si totalidad por personas de nacionalidad argentina, con una duración promedio de viaje de 2 a 3 días y suelen tener un comportamiento parejo durante todo el año y en los días de la semana. Los lugares con mayor concurrencia de este cluster en diferencia del resto son Aeroparque y Planetario, pero no suelen visitar la zona de Recoleta.



Cluster 3: Con un Viaje promedio de 6 días de duración, se concentran principalmente en las vacaciones de invierno y verano.



Bibliografía adicional

[1] Ciudad de Buenos Aires - Datasets de 2016 y 2017-2018

[2] Infra Datos - Dataset de tipo de cambio

[3] Elbow Method

[4] Hamerly & Elkan – Learning the k in k - means

[5] Sebastian Raschka – Linear Discriminant Analysis

Curso Cluster AI 2020 – UTN FRBA