Clusterizacion Turistas de CABA

INTRO

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.

INPUT DATA

Datasets:

- **Resultados encuestas 2016**
- **Resultados encuestas 2017 y 2018**
- **Tipo de Cambio por Fecha**

Features Relevantes:

- **Fecha del Registro**
- **Centro de Atraccion turistica**
- **Cantidad pasajeros**
- **Pais de Origen**
- **Provincia**
- **Cantidad de Pernoctaciones**
- **Relacion Dolar - Peso**

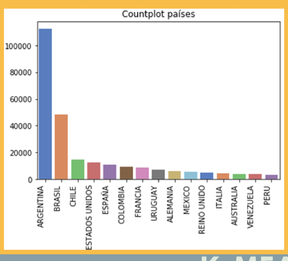
PREPROCESAMIENTO

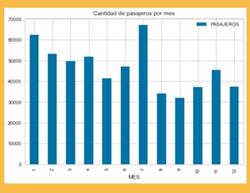
Dada la incompatibilidad de los datos del los diversos dataset, se realizo un trabajo para homogenizar los datos: tratamiento de tildes y mayúsculas, unificar lugares de atención turística, formato de fechas, números. Por otro lado, se agregaron variables como el valor del dolar, dia de las semana y mes del año para poder trabajar con mayor catidad de features.

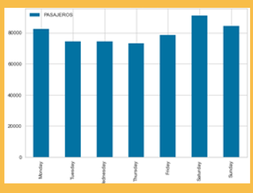
EXPLORATORY DATA ANALYSIS

ORIGENES

Haciendo un análisis de la procedencia de las personas, se pudo concluir que la mayoria de las mismas son Argentinas principalmente de la Provincia de Buenos Aires, con lo que respecta a los turistas extranjeros, se encuentra Brasil como el mayor origen de turistas extrajeros, seguido por chile y EEUU.







COMPORTAMIENTO

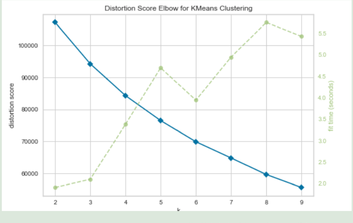
En cuanto a la movilización de los turistas, los mismos registran mas actividad en el primer semestre del año, principalmente en la época de vacaciones de verano. A su vez, se puede observar un pico en el mes de julio, coincidente con el receso escolar de invierno.
Analizando los días en los cuales prefieren ir a los principales centros turísticos de la Ciudad de Buenos Aires, los turistas eligen el fin de semana para recorrer la ciudad, seguido por los dias viernes y lunes.

K-MEANS

Consiste en un modelos no supervizado, sin la utilización de etiquetas, que permite detectar estructuras y propiedades de los datos permitiendo agrupar las muestras en "K" clusters. Los mismo son construidos de tal forma de que los datos del mismo cluster tenga mayor similaridad, mientras que los de que pertenecen a otros grupos uan mayor disimilaridad. Asignando un valor de cluster a cada muestra.

En cuanto al algoritmo, el mismo busca centroides de las muestras con tal de minimizar la distancia a las muestras asociadas al cluster correspondiente al centroide en cuestion. El proceso itera hasta poder minimizar la distancias totales.

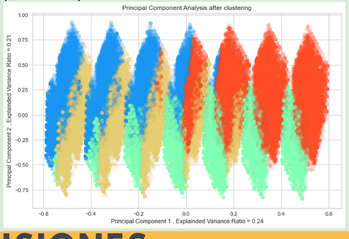
CLUSTERIZACION & VISUALIZACION



http://shabal.in/visuals/kmeans/5.html

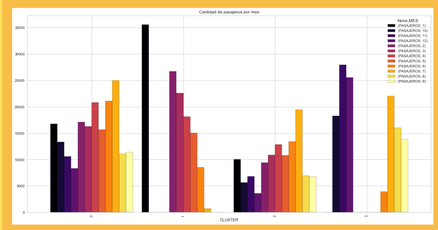
En el caso de este trabajo se optó por tomar K=4, coincidente con el cambio de pendiente en el gráfico del método Elbow.

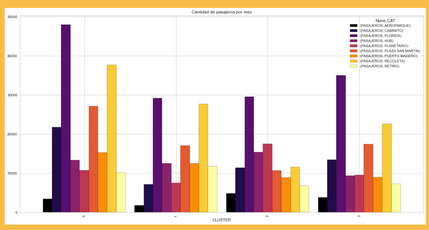
Para poder entender como estan distribuidos los cluster se procedio a visualizar mediante PCA con las 2 convinaciones lineales de las variables que mas representan la variabilidad de los datos



CONCLUSIONES

Luego de aplicado el modelo se calculo el Silhouette Score que en este caso dio 0,17, lo que significa que los clusters esta altamente superpuestos y no tiene una separación clara entre ellos. Sin embargo con los resultados obteneidos se puede dedusir las siguientes caracteristicas por clusters:





Cluster #0

Se caracteriza principalmente por turistas brasileros, con visitas principalmente los fines de semana con estadías estimadas de 7días. Los lugares que mas 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.

Autores: De Luca, German - Dobzewicz Kevin - Rodriguez Piro, Juan Darío