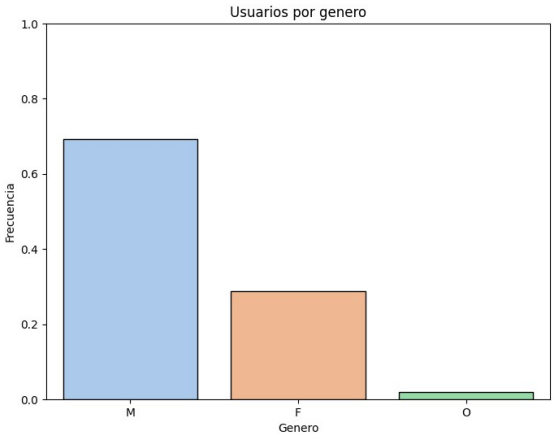
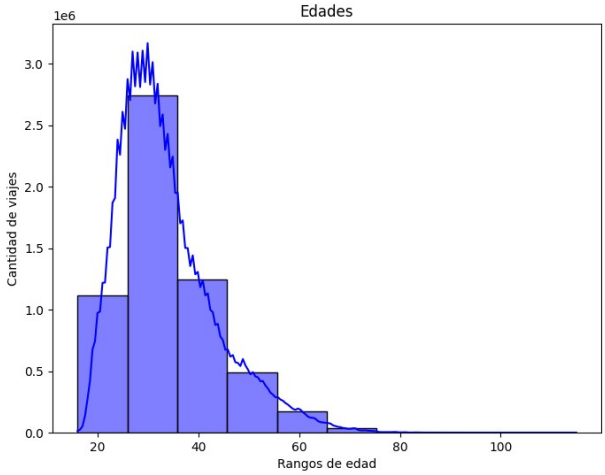
Test Científico de Datos

Ejercicio 1

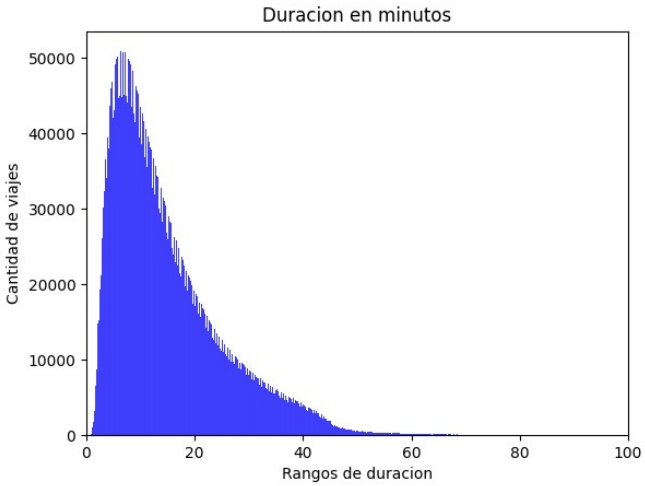
En este ejercicio se abordo el análisis exploratorio de los datos de viajes hechos en ecobici. Se tomo un histórico de tres meses de información de viajes, estos meses fueron desde agosto hasta octubre, el ultimo mes completo hasta esta fecha. Para iniciar el análisis exploratorio leemos los datos y tenemos con que el dataset cuenta con 5,798,107 registros y 9 columnas donde cada registro es un viaje de una estación a otra, están registrados los horarios de inicio y fin así como datos del genero del usuario y edad.
Posteriormente, calculamos el porcentaje de faltantes por columna y para la columna de genero sustituimos el valor '?' por un nan pues es un dato faltante. En total tenemos 678 estaciones de bicicleta y la frecuencia de viajes iniciados por genero de usuario es como se ve a continuación:



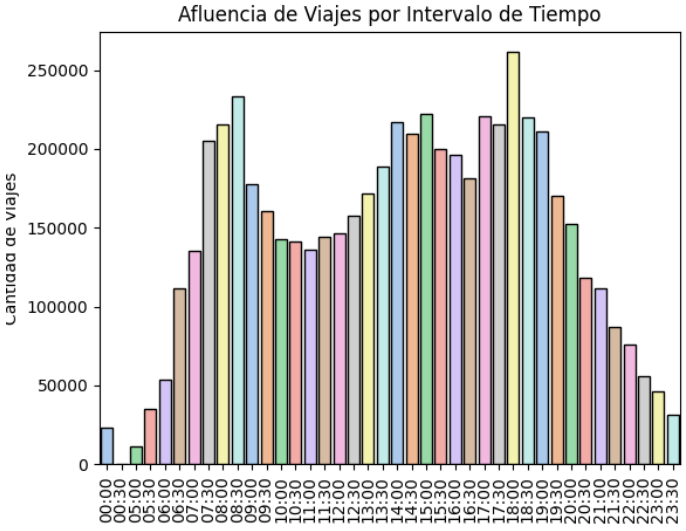
Otros datos adicionales son las 8246 bicicletas disponibles y que la edad promedio por genero es de 34 años para hombres, 32 para mujeres y 32 para otros. A continuación vemos la cantidad de viajes por rango de edad:



También se calculo el campo nuevo de ‘duracion_viaje’ que mide el tiempo transcurrido en minutos entre el inicio del recorrido hasta su final y vemos aquí su comportamiento:



donde vemos que la duración media esta entre los 10-20 minutos.
A continuación entramos en la sección del análisis relativa a identificar los niveles de afluencia por horarios y comportamientos atípicos. Un dato que me pareció interesante analizar es la ruta, es decir, los viajes que se hacen de una estación x a una estación y, con el fin de ver si aquí también hay rutas atípicas. Entonces se calculo este nuevo campo ‘viaje_inicio_destino’ uniendo los strings de la columna de la estación de inicio con las de fin. Ahora veremos los horarios de mayor afluencia:



Es inmediato el darse cuenta que los horarios con mas afluencia son los de la mañana, desde 6:30 hasta 9:00 y luego, por la tarde, desde las 17:00 hasta 19:00. Estos horarios son los de entrada y salida de trabajadores, que se mueven hacia sus lugares de trabajo y por la tarde de regreso a sus hogares; también se nota un incremento importante en los horarios del mediodía, tal vez, debido a la hora de comida.

Ahora pasamos a identificar aquellas estaciones con una afluencia atípica en el retiro de unidades, por tanto de inicio de viajes. Para hacer lo anterior, contamos la cantidad de apariciones de cada estación de retiro y calculamos el primer y tercer cuartil de estos conteos, con esto determinamos el rango intercuartil(iqr) que es la distancia entre los dos cuartiles anteriores y después generamos los limites:
limite inferior = q1 – 1.5*iqr
limite superior = q3 + 1.5*iqr
en resumen, calculamos los rangos de un boxplot y con esto determinamos las estaciones con mas viajes iniciados.

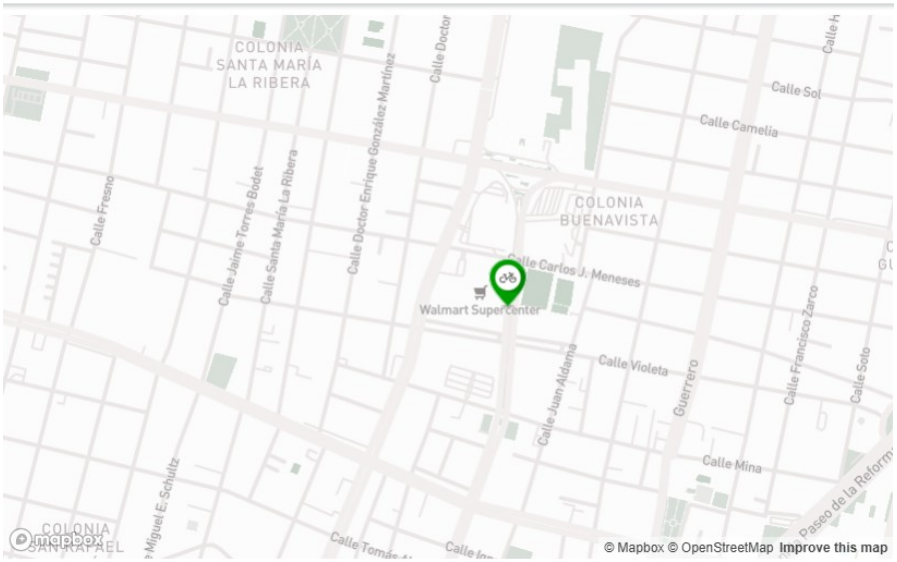
Entonces, con lo anterior se determinan las estaciones atípicas por entradas, arribos y también las rutas, las ubicaciones e estas estaciones se obtienen gracias al mapa de ecobici. Algunas de las estaciones con mas afluencia de entrada son:

Estacion	Localización
271-272	Cerca de estación Buenavista
027	Reforma, cerca de glorieta del Ahuehuete
064	Parque España, Condesa
237-238	Auditorio Nacional
014	Cerca de la estela de luz

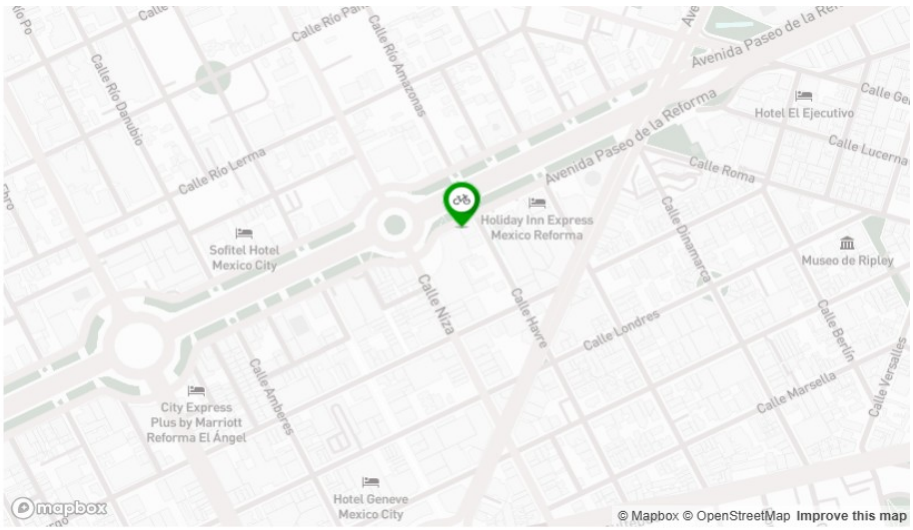
y con mas afluencia de arribos:

Estacion	Localización
271-272	Cerca de estación Buenavista
014	Cerca de la estela de luz
027	Reforma, cerca de glorieta del Ahuehuete
064	Parque España, Condesa

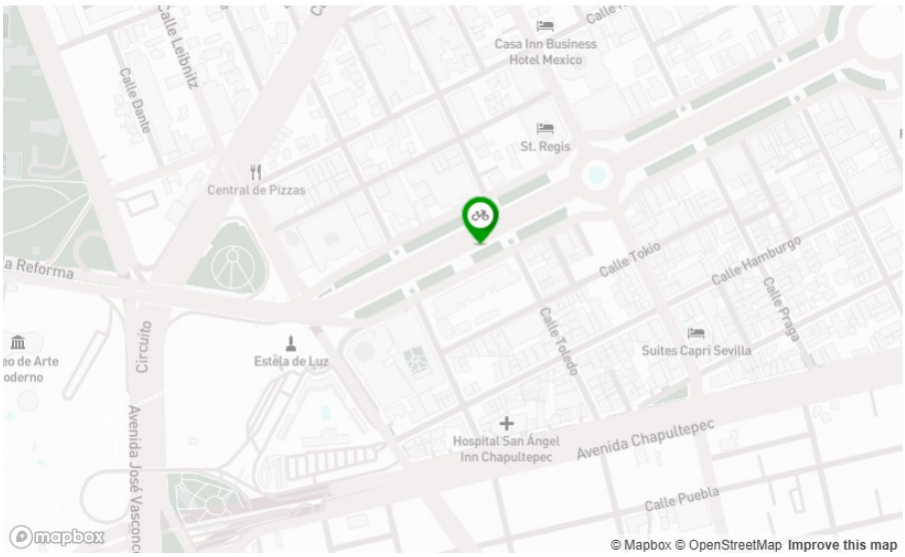
La ruta mas demandada es la que va de la estación 271-272 a la estación 014. A continuación las ubicación de algunas de estas estaciones, aquí esta la 271-272:



Estacion 027:



Estacion 014:



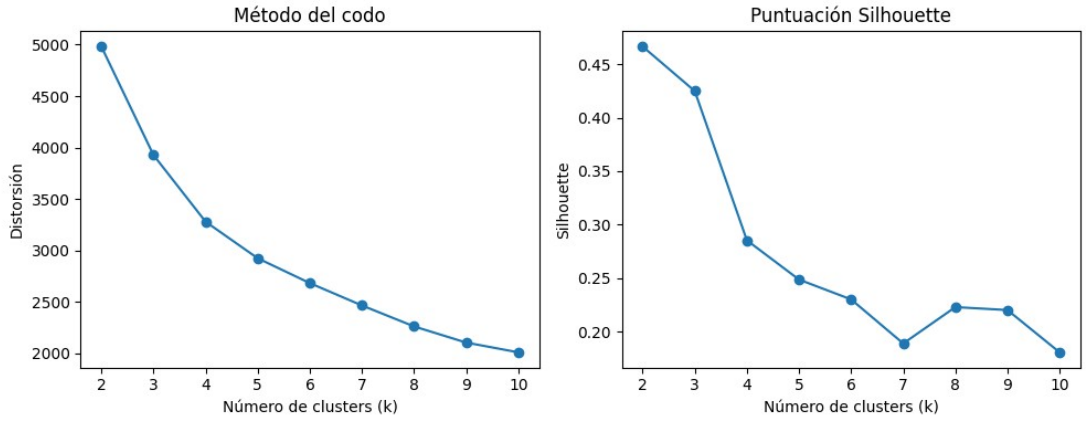
La razón por la que estas estaciones son identificadas como atípicas es por su ubicación; la estación 271-272 ubicada en Buenavista tiene, por mucho, la mayor afluencia de viajes de inicio y como destino. Buenavista es un cetram que tiene conexiones de metro, metrobus y principalmente del tren interurbano, aquí es donde gran parte de población del estado de México llega a la ciudad y a partir de ahí se mueven a sus destinos de trabajo, que, como podemos ver están en zonas de reforma como la glorieta del ahuehuete, la estela de luz y auditorio. Muchos de los viajes también inician en estas zonas de reforma pues una vez llegando a esa zona es mas fácil moverse en bici que en transporte publico o coche.
Respecto a las estaciones de arribo el comportamiento es muy similar, la estación 271-272 vuelve a ser la mas solicitada cuando la gente regresa a sus hogares y las estaciones de reforma reciben mucha afluencia pues la bici es un método muy solicitado de movilidad en la zona; con lo anterior se exlica muy bien porque la ruta 271-272 a la 014 es la mas utilizada.

Pasamos ahora a la formación de los clusters para ver si encontramos un perfil de uso. Se genero un dataset aparte con el fin de generar métricas por estación. Para cada estación se calculo la edad promedio de sus usuarios, duración promedio de los viajes, y el conteo de inicios de viaje y arribos por intervalos de tiempo; los intervalos de tiempo usados fueron de 05:00 a 10:00, 10:00 a 15:00, 15:00 a 20:00 y de 20:00 a 01:00, para el caso de arribos también se tomo en cuenta el intervalo de 01:00 a 04:59:59. Lo anterior con la idea de determinar si hay estaciones de mayor uso matutino, vespertino, etc. Antes de aplicar algoritmos se hizo un reescalamiento estándar entre -1 y 1 para evitar problemas con la escala de cada columna.

Los algoritmos que se probaron fueron kmeans y el método basado en densidad DBSCAN de scikit-learn. Para la elección del numero de clusters se hicieron algunos gráficos de la distorcion (SE) y el score de silhohuette. Antes de proceder diremos que el score de sihohuette es una medida para cada punto donde se toma su distancia promedio a cada otro punto dentro de su cluster (a_i) y la distancia promedio a cada otro punto fuera de su cluster (b_i), luego se calcula:

silhohuette score i = (b_i – a_i) / max(b_i, a_i)

Luego se promedia este score para todos los puntos y obtener el score total. La distorsión es el error cuadrático de cada punto con respecto al centroide de su cluster. A continuación las graficas para cada valor de k:



Con lo anterior se eligió k = 4. Para dbscan se eligieron los parámetros estándar y no dio resultados aceptables pues prácticamente toda la data fue identificada como ruido al asignarla al cluster con etiqueta 1. En el notebook viene un summary de los clusters creados; la estación 271-272 es un cluster por si misma (3) pues su comportamiento es muy diferente a las demás, el cluster 0 son las estaciones menos usadas pues parecen estar mas retiradas de los lugares con mas afluencia y de las zonas de trabajo, el cluster 1 son aquellas estaciones con gran afluencia (después de la 271-272) y están por regular en la zonas de reforma, condesa e incluso en alrededores de Buenavista, esto supongo, ante la saturación de la 271-272. Finalmente el cluster 2 son las estaciones con una afluencia regular, también parecen estar mas o menos alejadas de las principales zonas de movilidad.