Práctica integradora
Aprendizaje Estadístico

Ejercicio 0:

Importe los datos de recorridos realizados en las bicicletas públicas de CABA durante el año 2018 y las bases de usuarios de 2015 a 2018 inclusive (Se encuentran disponibles en el siguiente link)

Ejercicio 1:
Agrupe todas las tablas de usuarios en un único dataframe y realice un EDA sobre los 2 dataframes resultantes: el de recorridos de 2018 y el de todos los usuarios registrados a la fecha.

Ejercicio 2:
Produzca visualizaciones univariadas de las variables que le resulten de interés de ambos datasets. Adicionalmente, genere una columna calculada en el dataframe de recorridos para determinar el tiempo de cada recorrido y observe su distribución. Qué sucede si le aplicamos una escala logarítmica?

Ejercicio 3:
Estime el total de recorridos generados cada día, elaborando una serie de tiempo de frecuencia diaria y desarrolle una visual con esos datos. Repita con una frecuencia mensual en vez de diaria y usando barplots.

Ejercicio 4:
Genere una nueva columna en el dataset de recorridos, determinando el día de la semana del origen del recorrido. Use esta nueva columna para representar gráficamente las duraciones de los viajes de cada día y la frecuencia de los mismos.

Ejercicio 5:
A partir del criterio de rango inter-cuartil (IQR) identifique los outliers u observaciones extremas del vector de tiempos de duración creado, para aquellos valores que son extremos en la cola derecha de la distribución. Grafique en histograma.

Ejercicio 6:
Agrupando las observaciones por día de la semana, estime el intervalo de confianza para una confianza del 95% de la cantidad de recorridos que se realizan. Desarrolle una visualización usando un Cleveland dot plot u otra visual que permita representar esa información de manera clara. ¿Puede arribar a alguna conclusión?

Ejercicio 7:
Elabore un mapa donde puedan observarse todas las estaciones de bicicletas públicas presentes en el dataset de recorridos.

Ejercicio 8:
Con los datos de usuarios, discretizar la variable edad, armando rangos etarios que vayan de 15 a 100 años en intervalos de 5 años. Puede resultarle de utilidad la función cut().

Ejercicio 9:
Elabore una pirámide poblacional que represente la edad y composición etaria de los usuarios de las bicicletas de la ciudad, puede usar como guía el presente documento, o alguna librería especializada.

Ejercicio 10:
Realice un left join del dataset de recorridos con el de usuarios y conserve únicamente registros en que se conoce la edad del usuario y esta no es mayor a 100 años. Filtre los datos para quedarse únicamente con los registros que se originan en la estación 177 y se dirigen a la estación 5 de los cuales se conoce el tiempo de viaje. Por último, conserve las variables de genero, edad, dia de la semana, horario de alta (expresado en minutos), horario de inicio del recorrido (expresado en minutos) y tiempo del recorrido.

Ejercicio 11:
Realice un train test split sobre los datos del punto anterior, haga una validación de modelos lineales con Best subset selection, buscando predecir el valor del tiempo del recorrido. Entrene un modelo con todas las variables y evalúe la performance en el test set. ¿Qué variables que no tiene cree que en caso de disponerlas podría mejorar la performance del algoritmo?