MEMORIA EDA

En este EDA se quieren analizar los datos de uso del servicio de BiciMAD del Ayuntamiento de Madrid. Y relacionar si el uso de estas bicicletas eléctricas ha aumentado desde que se inició la pandemia en marzo de 2020. Se van a comparar datos del 2019, 2020 y 2021.

Introducción a los datos

Los datos se descargan en formato .json de manera mensual. Para algunos meses se descarga en bloque de cinco meses. Son archivos pesados que contienen cerca de 200.000 y 300.000 filas. La idea es sacar datos de tres años y comparar la serie temporal. Para el año 2021 sólo hay datos disponibles hasta junio, por lo que no se ha podido realizar la evolución completa de los tres años. El nombre de los archivos cambia a partir de julio de 2019, es importante tenerlo en cuenta al llamar a las funciones, ya que el argumento de la mayoría de ellas es el nombre del archivo. En general tratar estos datos ha sido laborioso en términos de computación por tener tantos registros. Al querer hacer una evolución temporal anual, no se ha usado un único mes, que hubiese sido más sencillo, por lo que es importante hacer un buen filtrado.

Existen muchos outliers en la columna de la duración del trayecto. Muchos valores menores a 3 min que no tienen sentido. En la columna del código postal del usuario existen direcciones de emails que no corresponden, NaN y códigos numéricos que no pertenecen a ningún código postal. En la columna de rango de edad hay una categoría de edad indeterminada que es la mayoritaria. Del mismo modo ocurre en la columna del tipo de usuario según si tiene abono anual o es ocasional. Existe una categoría indeterminada, mayoritaria también, que no se explica de dónde procede. En la misma variable del tipo de usuario, en los datos de 2021 existe una categoría que no se encuentra en las otras, la que se refiere al trabajador de BiciMAD. Algunos de estos datos contienen un .geojson con el track de cada trayecto. Al no estar disponible en todos los archivos se ha descartado la idea de usar esta variable. Aunque para más adelante estaría bien aplicarla para Machine Learning y predecir los trayectos.

Esta base de datos no tiene muchas variables numéricas, la mayoría son categóricas, por lo que la visualización se ha visto limitada a gráficos de línea, de barra, o de cajas y bigotes.

Estructura de carpetas

src/: archivo main.ipynb (notebook con el detalle de todos los pasos)
src/data: .json descargados de la web de BiciMAD
src/visualizaciones: plots creados en el notebook
src/utils.py: con las funciones usadas

Librerías utilizadas

- import requests
- import zipfile
- import numpy as np
- import pandas as pd
- import json
- import matplotlib as mpl
- import matplotlib.pyplot as plt
- import matplotlib.patches as mpatches
- import datetime
  from datetime import date
- import seaborn as sns
- import statistics as stats

Técnicas utilizadas

1. Hacer un request a la API del Ayuntamiento de Madrid donde se encuentran los datos de BiciMAD:
   - https://datos.madrid.es/portal/site/egob/
   - https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)

2. Guardar el archivo en .zip y descomprimir el contenido obteniendo un .json. En el EDA se usa esta manera de descargar datos combinada con una descarga manual.
3. Aplicar lo aprendido en clase sobre manejar bases de datos, filtrarlas, convertir columnas en datos más amigables (fecha), añadir y eliminar columnas y filas, localizar datos, eliminar NaN, agrupar por columnas, describir la estadística general del dataframe, escritura de datos en .csv, etc.
4. En general este EDA viene muy bien para practicar y ponerse al día de todo lo aprendido desde el inicio del bootcamp.

Paso a paso

Los pasos detallados vienen en el Notebook llamado main.ipynb con las celdas y los comentarios.

1- Crear funciones: se han usado seis funciones
   - Función para filtrar los datos.
   - Función que descarga el argumento url, hace la petición, guarda el archivo.zip y lo descomprime para ver su contenido.
   - Función que lee el argumento filename.
   - Función para pintar un boxplot.
   - Función para contar el número de registros agrupados por tipo de usuario.
   - Función para pintar un barplot del rango de edad de los ususarios con abono anual.

2- Análisis del número de registros por mes para ver si hay un impacto de la pandemia. Primero sólo para el año 2020, después desde el 2019 al 2021.

3- Análisis para saber por cuánto tiempo se utiliza la bicicleta en los tres años.

4- Filtro de outliers mediante visualización de boxplots.

5- Comparación de la duración del trayecto en bici en los tres años comparando mes de invierno y de verano,

6- Análisis por tipo de usuario y saber si hay más registrados en BiciMAD desde la pandemia.

7- Análisis por rango de edad. Estudiar qué edades hacen más usos del servicio y a qué horas.

8- Hacer un análisis separado, por meses, no por años, agrupando los datos por diferentes columnas y haciendo un conteo. Por ejemplo, para saber a qué franjas de edad se hacen trayectos más largos, el comportamiento de un usuario seleccionado al azar, saber si las estaciones de recogida de bicicletas son las mismas que las de devolución.
9- Hacer un análisis estadístico muy sencillo para conocer la duración de trayectos más frecuente, la hora y día en el que se hace mayor uso, el rango de edad que más utiliza el servicio, y las estaciones o el código postal más frecuentes.