6.1: Sourcing Open Data

BY Bilel Yahyaoui

Data Source

This dataset sourced from Kaggle and provides a detailed Rider information since the launch of Citi Bike until the beginning of October 20ti3 in New York City. it contains 18 columns and 50000 rows.
The data can be access here
Link: https://www.kaggle.com/datasets/ryanmcummings/citi-bike-data

Why This Data

The NY Citi Bikes dataset from Kaggle is an excellent choice for a data analysis project due to its rich and diverse data, offering numerous opportunities for temporal, spatial, and user segmentation analyses. It is particularly relevant for research on urban mobility and public health. The dataset is accessible and well-documented, with support from the Kaggle community. Insights from this dataset can help optimize resources, improve bike-sharing services, and provide valuable information for public policy.

Data cleaning summary

- **Consistency Checks :**
  - No mixed_type data founded
  - 6979 missing values on ‘birth_year’, I imputed the median to fill the missing values
  - No duplicates
  - Drop 23 records of ‘birth year’ with birth years prior to 1913, it seems no logic that someone who have more than 100 years can ride a bike.

- **Data Wrangling :**
  - Convert start-time & end_time to datetime values
  - Convert id_columns like : 'bike_id', 'start_station_id', 'end_station_id', 'gender' to object.
  - changing column name from weekday to day_of_week.

Data Profile

Variables	Time-variant/ Time-invariant	Structured/ Unstructured	Qualitative/ Quantitative	Qualitative: Nominal/ Ordinal Quantitative: Discrete/ Continuous
Trip_id	Time-invariant	Structured	Qualitative	Nominal
Bike_id	Time-invariant	Structured	Qualitative	Nominal
Weekday	Time-variant	Structured	Qualitative	Nominal
Start_hour	Time-variant	Structured	Quantitative	Discrete
Start_time	Time-variant	Structured	Quantitative	Continuous
Start_station_id	Time-invariant	Structured	Qualitative	Nominal
Start_station_name	Time-invariant	Structured	Qualitative	Nominal
Start_station_latitude	Time-invariant	Structured	Quantitative	Continuous
Start_station_longitude	Time-invariant	Structured	Quantitative	Continuous
End_time	Time-variant	Structured	Quantitative	Continuous
End_station_id	Time-invariant	Structured	Qualitative	Nominal
End_station_name	Time-invariant	Structured	Qualitative	Nominal
End_station_latitude	Time-invariant	Structured	Quantitative	Continuous
End_station_longitude	Time-invariant	Structured	Quantitative	Continuous
Trip_duration	Time-invariant	Structured	Quantitative	Continuous
Subscriber	Time-invariant	Structured	Qualitative	Nominal
Birth_year	Time-invariant	Structured	Quantitative	Continuous
Gender	Time-invariant	Structured	Qualitative	Nominal

**** A descriptif analysis can be founded on the Jupiter Notebook.***

Limitations and Ethics

Ethically speaking, the data is as anonymous as possible since it does not contain any PII. The individual trips are labeled with a trip_id and bike_id that are assigned by Citi Bikes. We are not provided with any information on the customer other than whether they are Citi bikes subscriber. I am not sure the ages are accurate. I imputed the median birth years into the blank records to retain as much data as possible, however I am not sure how accurate any of the information in that column is, since it seems to be the only field that is entered by the customer.

Questions to explore

- **Temporal Analysis :**
  1. **Peak Hours**: What are the peak hours for bike usage each day of the week?
  2. **Daily Variations**: How do bike trips vary throughout the days of the week (e.g., are there differences between weekdays and weekends)?

- **Spatial Analysis :**
  1. **Popular Stations**: What are the most popular start and end stations?
  2. **Frequent Trips**: What are the most common trips between different stations?
  3. **Geographic Analysis**: How do the latitude and longitude of stations influence trip patterns (e.g., are trips more frequent in certain geographic areas)?

- **User Analysis :**
  1. **Subscriber Behavior**: How do subscriber behaviors differ from casual users?
  2. **User Demographics**: What is the demographic profile (age and gender) of bike-sharing users?

- **Performance and Duration :**
  1. **Trip Duration**: What is the average trip duration and how does it vary by time of day or day of the week?
  2. **Bike Performance**: Is there a difference in performance between different bikes (Bike_id)?

- **Optimization and Logistics :**
  2. **Bike Distribution**: How can the distribution of bikes between stations be optimized to meet demand?
  3. **Wait Times and Availability**: Which stations experience shortages or surpluses of bikes at different times of the day?