Open Project Plan: Citibike Data

Data Source

This dataset from Kaggle contains trip information for Citibike customers in New York City. I chose this dataset because I am interested in the possibility of going into retail analytics/business intelligence.

Link: https://www.kaggle.com/datasets/ryanmcummings/citi-bike-data

Data Cleaning

- Found 6979 missing values in the “birth year” column. Imputed the median to fill the missing values.
- Deleted 23 rows that had a birth year prior to 1913.
- Dropped the gender column.

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

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 Citibike. We are not provided with any information on the customer other than whether they are Citibike 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

- What are the most popular days for rentals?
- What are the most popular times of day for rentals?
- Which stations are the most popular to rent from?
- Which stations are the most popular destinations?
- How long is the average trip duration?
- Are most customers subscribers or casual riders?
- What is the age of the average customer?
- Which age group rents bikes the most?