Mark Regini

Task 6.1

Sourcing Open Data

**Data Source**

The dataset I have chosen for this achievement is about the New York Citi Bike system, sourced from Kaggle.com at [this](https://www.kaggle.com/datasets/ryanmcummings/citi-bike-data) location. The data was collected and aggregated using the Citibike service and information collected from its use. This is an internal data source, as it is collected, used, and shared by the New York Citibike service for analysis. The data was collected by usage, as the trips taken with the bike rental service are recorded as they occur. The data set lists all the trips purchased within the service, and include the time and day of the trip, start and end locations, trip duration, and basic customer demographic information like birth year, gender, and whether they are a subscriber to the service or not.

**Data Cleaning**

*Mixed-type data:* The column for ‘gender’ was a mixed-typing after import to Jupyter. After reviewing the column, it has a key for gender instead of the gender written out. I have changed the column to a string-type for this reason.

*Missing data:* The ‘birth\_year’ information is missing from the dataset in 6,979 records. This accounts for almost 14% of the records, taking out the possibility of simply removing them. I will keep them blank for the time being, as it should not affect analysis.

*Duplicated data:* No duplicate records were found.

**Data Understanding**

*Variables*

* Trip\_id – unique code to indicate trip
* Bike\_id – identifying label for each individual bike
* weekday – day of the week of the trip
* start\_hour – hour the trip has started
* start\_time / end\_time – exact time and date of the trip start and end
* start\_station\_id / end\_station\_id – ID number of the bike station the trip started and ended on
* start\_station\_name / end\_station\_name – name of the starting and ending location
* start\_station\_latitude / start\_station\_longitude / end\_station\_latitude / end\_station\_longitude – latitude and longitude of the starting and ending station
* trip\_duration – total time of trip, from start to end, in seconds
* subscriber – indicates whether the trip was done by a subscriber to the service or not
* birth\_year – birth year of the customer
* gender – indicates gender of the customer; 0 is Unknown, 1 is Male, 2 is Female

*Descriptive Analysis*

* Summary
  + The dataset is 50000 records long, with 18 variables
* Descriptive Analysis

Graphical user interface

Description automatically generated

Graphical user interface, application, website

Description automatically generatedTable

Description automatically generated with medium confidence

* Qualitative Data
  + Trip ID
    - Values count: 50,000
    - Mode: N/A
  + Weekday
    - Values count: 7
    - Mode: Monday
  + Start station name
    - Values count: 330
    - Mode: W 20th St and 11th Ave
  + Subscriber
    - Values count: 2
    - Mode: Subscriber
  + Birth Year
    - Values count: 77
    - Mode: NA
  + Gender
    - Values count: 3
    - Mode: 1
* Data Integrity
  + Everything appears to be within logical parameters

**Limitations and Ethics**

One limitation that immediately pops out is the accuracy of the demographic information. Being entered by the customer, not only is it prone to being incorrectly entered but also runs the risk of them not giving the correct answer for the sake of security (the birth year, for example).

Ethically, I don’t see any information that is PII, or could connect any of the trips to an individual person. However, even though the data set does not have direct information, it does provide exact coordinates of the bike and time/date information of the trip. If put into the wrong hands, this could lead to some issues with security. As long as the service has a term of service which shares that they collect location information, we should be ok legally.

**Questions to Explore**

Some possible questions that can be explored through the data:

* Marketing: What times and days are customers, both subscribers and non-subscribers, most active? Are there certain areas within New York that see the most usage of the service? What is the age demographic of the most common customers? Are holidays a popular time for customers to use the service?
* Operations: Where are the most trips starting? Where are most bikes being left at the end of the trip? Is there a correlation between age and how long the trip is? Are there certain bikes that are used much more than others? Is there seasonality with the trips taken?