DAI Exercise 6.1
Katherine Valdivia
08/22/23

Data Source

The dataset is from the Citi bike website but can be downloaded via Kaggle which is where I downloaded it from. This source is from the Citi bike website so I would consider it reliable.

Data Collection Method:

A combination of the bikes GPS and the Citi bikes app help collect data from users. Once the user uses the app to unlock a Citi bike that's when data collection starts like when and where the user took the bike and when and where they the left it when they were done.

Data Relevancy:

This data meets all the requirements for this project.

Data Contents:

There is only one data set, and it contains 18 columns with 50,000 rows. The data set contains information on Citi bike users from May 2013, when the bikes were launched to October 2013. Each trip is given a unique ID with information about the bike the user used (bike ID), date, day, start and end times of their bike usage, the location of the station using longitude and latitude where they picked up and left off the bike, the names of the stations they picked up and left off the bike and the duration of the ride. It also contains some basic user information like whether the user is a subscriber or not, their gender, and birth year.

Citibike.csv Table:

Column Name	Description
Trip_id	Unique trip identifier
Bike_id	Unique bike identifier
Weekday	Weekday that trip occurred
Start_hour	Hour that the trip started
Start_time	Time trip started
Start_station_id	Unique station identifier for start of trip
Start_station_name	Name of the station at the start of the trip
Start_station_latitude	Station latitude of the start of the trip
Start_station_longitude	Station longitude of the start of the trip
End_time	End time of trip
End_station_id	Unique station identifier for end of trip
End_station_name	Name of the station at the end of the trip
End_station_latitude	Station latitude at the end of the trip
End_station_longitude	Station longitude at the start of the trip
Trip_duration	How long the trip lasted in seconds

Subscriber	If the user is subscribed or not
Birth_year	Birth year of user
Gender	Gender of user

Data Profile

Understanding Data:

Column Name	Time variant/ Invariant	Structured/ Unstructured	Quantitative/ Qualitative	Nom/Ord/Discrete/ Continuous
Trip_id	Invariant	Structured	Qualitative	Nominal
Bike_id	Invariant	Structured	Qualitative	Nominal
Weekday	Invariant	Structured	Quantitative	Discrete
Start_hour	Invariant	Structured	Quantitative	Discrete
Start_time	Invariant	Structured	Quantitative	Discrete
Start_station_id	Invariant	Structured	Qualitative	Nominal
Start_station_name	Invariant	Structured	Qualitative	Nominal
Start_station_latitude	Invariant	Structured	Quantitative	Continuous
Start_station_longitude	Invariant	Structured	Quantitative	Continuous
End_time	Invariant	Structured	Quantitative	Discrete
End_station_id	Invariant	Structured	Qualitative	Nominal
End_station_name	Invariant	Structured	Qualitative	Nominal
End_station_latitude	Invariant	Structured	Quantitative	Continuous
End_station_longitude	Invariant	Structured	Quantitative	Continuous
Trip_duration	Invariant	Structured	Quantitative	Discrete
Subscriber	Variant	Structured	Qualitative	Ordinal
Birth_year	Invariant	Structured	Quantitative	Ordinal
Gender	Invariant	Structured	Qualitative	Discrete

Cleaning Data:

Columns dropped	Column data type change	Reasoning
trip_id		Unnecessary for data analysis
bike_id		Unnecessary for data analysis
	Gender (string)	Changed to string since I changed values to names
	Start_time (datetime)	Change data type
	End_time (datetime)	Change data type

Column value change	Missing values dropped	Duplicates	Reasoning
Gender			Wanted full name of gender
	Birth_year (6979)		6979 null birth years
		None	none

Data Limitations:
The data filled out by users (gender and birth year) could be subject to human error. There is also no rider id in this data set to see how often a specific Citi bike user utilizes their subscription and if they don't have subscriptions and are using Citi bikes often this information could have been good to market to those specific users or provide incentives.

Ethics:
I don't think there is sensitive personal information about users (full name, address, full birthday, and SSN) that could be used to identify someone, so I don't see any ethical concerns as this data is also publicly available to anyone and is in accordance with the NYCBS Data Use Policy as they state on their website.

Key Questions:

- What time of the day is Citi bike the busiest/least busy?
- What day(s) are Citi bike the busiest/least busy?
- What station are most popular/least popular?
- What age group uses Citi bike the most/least?
- Are there more subscribers than unsubscribes?
- What is the average duration for a Citi bike ride?
- What gender tends to use Citi bike the most/least?