Data Immersion

Exercise 6.1

9/25/23

Micky Smith

Sourcing Open Data

Source:

The data I'm utilizing came from Kaggle, and was directly pulled from Citi Bike. Kaggle doesn't provide to much of a description into how the data was gathered, however, Citi Bike does provide an archive of their monthly trip data. Since the data looks closely related to the data provided on the Citi Bike website, it would be reliable. Current data can also be found within Citi Bike System Data.

Collection:

The data was gathered through administrative and usage data. The information was gathered each time a user's app was linked to a Citi Bike around New York.

The app gathered information like where the bike was used and where it stopped. The time and duration of the bike ride.

The administrative data would consist of customer information when they created their profile and was obtained by the app on a mobile device.

Inaccuracies in the data would occur if the customer input incorrect personal information within the app during set up, or if the customer used their app to activate a bike for a different user.

Limitations:

The data is set with 50,000 rides, however, the creator in Kaggle did not elaborate on the data to provide an understanding towards why the limit was set for 50,000 or if the website only provides the last 50,000. There is no understanding towards if a rider had multiple trips or utilize the app more frequently than others.

Ethics:

The data is provided on Citi Bikes website and displayed for public use. Citi Bike states that its data follows the guidelines of NYCBS Data Use Policy, and provides it's data privacy policy within its website.

There should be no ethical concerns with this data.

Relevence:

The data set meets the requirements for the project. It's from an open source, includes geospatial components, and meets the size and variable requirements.

The data set isn't within 3 years, however, the data can be further pulled from Citi Bike to provide more recent information in comparison.

Content:

The data set contains trips taken on NYC Citi Bike from May 27th, 2013, to October 2013.

It provides some user information like if they are a subscriber, gender, and year of birth.

It provides the assigned bike ID, unique trip ID for each trip, and a station ID.

The data also provides the date with start time and end time of the trip, start and end location as well as longitude and latitude, the start and end hour, and the trip duration.

Data Profile:

Variable	Description	Time Variant/Invariant	Structured/Unstructured	Quantitative/Qualitative	Nominal/Ordinal/Discrete/Continuous
Trip_id	Unique identifier for trip	Invariant	Structured	Qualitative	Nominal
Bike_id	Unique identifier for trip	Invariant	Structured	Qualitative	Nominal
Weekday	Day of week for ride	Invariant	Structured	Qualitative	Discrete
Start_hour	Hour ride started	Invariant	Structured	Quantitative	Discrete
Start_time	Date and time of ride	Variant	Structured	Quantitative	Discrete
Start_station_id	Unique identifier for trip	Invariant	Structured	Qualitative	Nominal
Start_station_name	Name of station ride started	Invariant	Structured	Qualitative	Nominal
Start_station_latitude	Latitude of station ride started	Invariant	Structured	Quantitative	Continuous
Start_station_longitude	Longitude of station ride started	Invariant	Structured	Quantitative	Continuous
End_time	Date and time of end of ride	Variant	Structured	Quantitative	Discrete
End_station_id	Unique identifier for trip	Invariant	Structured	Qualitative	Nominal
End_station_name	Name of station for end of ride	Invariant	Structured	Qualitative	Nominal
End_station_latitude	Latitude of the station ride ending	Invariant	Structured	Quantitative	Continuous
End_station_longitude	Longitude of the station ride ending	Invariant	Structured	Quantitative	Continuous
Trip_duration	Duration of trip in seconds	Invariant	Structured	quantitative	Discrete
Subscriber	If the rider subscribes or not	Variant	Structured	Qualitative	Ordinal
Birth_year	Year of birth of rider	Invariant	Structured	Quantitative	Ordinal
Gender	Gender of rider	Invariant	Structured	Quantitative	Discrete

Data Cleaning

Column Rename	Column Type Change	Reason
Renamed weekdays to day_of_week		The name is easier to understand.
	Changed column organization	Made the data more organized
	Changed start_station_id and end_station_id to string/object	Though they are numbers, they are resemblances of places and not a count.

Missing Values	Inconducive information	Duplicates	Response
Missing birth_year for nonsubscribers.			**Keeping N/A because nonsubscribers could be necessary for data.**
		No duplicates	
	Not removing columns. All information could remain valuable		
	Removing anyone with birthdays prior to 1923.		Incorrect data by human error can cause incorrect data to be gathered.

Questions:

1. What percentage of Citi Bike users are unsubscribed?
2. What is the busiest day for Citi Bike?
3. What age group mostly uses Citi Bike?
4. Which stations tend to have the most traffic either leaving or incoming?
5. Do certain bike id’s get used more so then others?
6. What time of day is busiest for Citi Bike?