Citi Bike is a bike share system in New York City that provides residents and visitors with a convenient transportation option. There are thousands of bikes at hundreds of stations. Pick up a bike at any station, then ride and return to any station.

Data:

Source:

- Open-source dataset downloaded from Kaggle.
- Original source: https://www.citibikenyc.com/system-data
- Open data from New Jersey City Website: New Jersey CitiBike Stations geojson
- Distance from PATH Stations data set generated using GoogleMaps and above Kaggle dataset: CitiBike_PATH_distance.csv

Contents:

- Data set contains Citi Bike trip details for January 2020 to December 2020, and January 2021 to April 2021.

Data Profile:

Data Cleaning and Statistical Checks:

- No mixed-type data
- No missing values
- 13,451 duplicate values removed
- Errors in 'birth year' column. Replaced values of 1888 with 1988, and value of 1900 with 1980 (mean). Records with error make up .005% of data.
- Data shape before cleaning and statistical checks: (393312, 15)
- Data shape after cleaning and statistical checks: (379861, 15)

Data Profile:

Variables	Description	Variable Type	Descriptive Statistics
Trip Duration	Trip duration in seconds	Quantitative/Continuous	max: 3261756 min: 61 mean: 1554
Start Time	Date and time trip began	Qualitative/Ordinal	
Stop Time	Date and time trip ended	Qualitative/Ordinal	

Start Station ID	ID number for start station	Qualitative/Ordinal	
Start Station Name	Name of start station	Qualitative/Nominal	
Start Station Latitude	Latitude for start station location	Quantitative/Continuous	
Start Station Longitude	Longitude for start station location	Quantitative/Continuous	
End Station ID	ID number for end station	Qualitative/Ordinal	
End Station Name	Name of end station	Qualitative/Nominal	
End Station Latitude	Latitude for start station location	Quantitative/Continuous	
End Station Longitude	Longitude for end station location	Quantitative/Continuous	
Bike ID	ID number for bike	Qualitative/Ordinal	
User Type	Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member	Qualitative/Nominal	
Birth Year	Year customer was born	Qualitative/Ordinal	max: 2004 min: 1920 mean: 1980
Gender	Customer gender: 0=unknown 1=male 2=female	Qualitative/Nominal	

Data Limitations:

- Historical data that only includes records for 16 months.
- Possible inaccuracy for Birth Year variable if reported incorrectly by customer.
- 24% of Gender values = 0 (unknown).

Questions to Explore:

- Which months are busiest? Slowest?
- Which days of the week are busiest? Slowest?
- Which times of day are busiest? Slowest?
- Is there seasonality?
- Which locations experience the most traffic? Least?
- Explore customer behavior based on age, gender, and user type.