Project Information: Citi Bike Analysis

Data Source:

This is an internal data source. The data is owned by NYC OpenData and was obtained from Citi Bike's website. As government data, this is a trustworthy and reliable data source. The data is a combination of administrative and usage data that is collected in real-time. The data set contains information on bike trips of Citi Bike users in New York for the entire month of March in 2019. This data set meets all of the criteria specified in CareerFoundry's data requirements and is, therefore, relevant to my goal of building an interactive dashboard that will visually showcase well-curated results of an advanced exploratory analysis conducted in Python. I chose this data set because I wanted to explore NY Citi Bike's trip data set for March 2019 to better understand the behavior of NY Citi Bike's customer base and how they use NY Citi Bikes.

Raw Data Profile:

Variables and Data Types

Column	Description	Data Type
tripduration	trip duration (in seconds)	time invariant, structured, continuous variable
starttime	date and time trip started	time invariant, structured, ordinal variable
stoptime	date and time trip ended	time invariant, structured, ordinal variable
start station id	unique identifier of where trip began	time invariant, structured, nominal variable
start station name	name of station where trip began	time invariant, structured, nominal variable
start station latitude	latitude coordinate of where trip began	time invariant, structured, ordinal variable
start station longitude	longitude coordinate of where trip began	time invariant, structured, ordinal variable
end station id	unique identifier of where trip ended	time invariant, structured, nominal variable
end station name	name of station where trip ended	time invariant, structured, nominal variable
end station latitude	latitude coordinate of where trip ended	time invariant, structured, ordinal variable
end station longitude	longitude coordinate of where trip ended	time invariant, structured, ordinal variable
bikeid	unique identifier for bike used	time invariant, structured, nominal variable
usertype	whether user is a customer or subscriber	time variant, structured, nominal variable
birth year	rider's year of birth	time invariant, structured, ordinal variable
gender	rider's gender	time invariant, structured, nominal variable

Data Accuracy

	Statistics			
Column	Min	Max	Mean	Median
tripduration	61	2,969,781	882.654	539
birth year	1857	2003	Not applicable	1982
gender	0	2	Not applicable	1

Note:

- There were 71,577 records that had a value of 0 (unknown) for gender. Since there is no way to correct these values, we will remove records containing 0 for gender.
- The maximum trip duration is unusually large, but it's plausible that a Citi Bike user may borrow a bike for long periods of time, especially if the user is a subscriber. The interquartile range (IQR) method for detecting outliers found 66,070 trip duration values that are outliers. Removing that many values would result in losing more than 5% of the data, which could cause issues in our analysis. Therefore, we will keep the outliers.
- The minimum birth year is unusually low, implying a Citi Bike user could be as old as 162 years old! This is certainly not plausible. It does not seem plausible that someone over the age of 85 years

would be renting bikes to get around New York. For that reason, we will remove any records with a birth year that is lower than 1934.

Data Cleaning

Column	Issue	Action
start station id	10 missing values	Dropped rows with missing values
start station name	10 missing values	Dropped rows with missing values
end station id	10 missing values	Dropped rows with missing values
end station name	10 missing values	Dropped rows with missing values
gender	71,577 inaccuracies	Dropped rows with unknown gender
birth year	881 inaccuracies	Dropped rows with birth year < 1934 (dropped users older than 85)

Data Wrangling

Column Derived/Dropped/Renamed/Data Type Changed	Comment/Reason
Derived trip_duration column: trip_duration = round(tripduration/60 , 2)	Converted trip duration to minutes
Derived age column: age = 2019 – birth year	Necessary for analysis
Derived age_group column from age column	Necessary for analysis
Derived start_hour column from starttime column	Necessary for analysis
Derived start_day column from starttime column	Necessary for analysis
Derived start_day_name column from starttime column	Necessary for analysis
Dropped tripduration column	Duplicate column
Dropped start station id column	Not relevant for analysis
Dropped end station id column	Not relevant for analysis
Dropped bikeid column	Not relevant for analysis
Dropped birth year column	Not relevant for analysis
Dropped starttime column	Not relevant for analysis
Dropped stoptime column	Not relevant for analysis
Renamed tripduration to trip_duration	Proper naming convention
Renamed starttime to start_time	Proper naming convention
Renamed stoptime to end_time	Proper naming convention
Renamed start station name to start_station	Proper naming convention
Renamed start station latitude to start_latitude	Proper naming convention
Renamed start station longitude to start_longitude	Proper naming convention
Renamed end station name to end_station	Proper naming convention
Renamed end station latitude to end_latitude	Proper naming convention
Renamed end station longitude to end_longitude	Proper naming convention
Renamed usertype to user_type	Proper naming convention
Changed age data type to 8-bit unsigned integer	More optimal data type
Changed trip_duration data type to 32-bit float	More optimal data type
Changed start_day data type to 8-bit unsigned integer	More optimal data type
Changed start_hour data type to 8-bit unsigned integer	More optimal data type
Changed start_latitude data type to 32-bit float	More optimal data type
Changed start_longitude data type to 32-bit float	More optimal data type
Changed end_latitude data type to 32-bit float	More optimal data type
Changed end_longitude data type to 32-bit float	More optimal data type

Clean Data Profile:

Variables and Data Types

Column	Description	Data Type
user_type	whether user is a customer or subscriber	time variant, structured, nominal variable
gender	gender of Citi Bike user	time invariant, structured, nominal variable
age	age of Citi Bike user	time variant, structured, discrete variable
age_group	age group to which Citi Bike user belongs	time variant, structured, ordinal variable

trip_duration	trip duration (in seconds)	time invariant, structured, continuous variable
start_day_name	name of day when bike trip started	time invariant, structured, nominal variable
start_day	day in March when bike trip started	time invariant, structured, ordinal variable
start_hour	hour of day when bike trip started	time invariant, structured, ordinal variable
start_station	name of station where trip began	time invariant, structured, nominal variable
start_latitude	latitude coordinate of where trip began	time invariant, structured, ordinal variable
start_longitude	longitude coordinate of where trip began	time invariant, structured, ordinal variable
end_station	name of station where trip ended	time invariant, structured, nominal variable
end_latitude	latitude coordinate of where trip ended	time invariant, structured, ordinal variable
end_longitude	longitude coordinate of where trip ended	time invariant, structured, ordinal variable

Data Accuracy

Column	Statistics			
	Min	**Max**	**Mean**	**Median**
age (years)	16	85	39.05	36
trip_duration (min)	1.02	49,496.35	13.597	8.73
start_day	1	31	Not applicable	18
start_hour	0	23	Not applicable	15

Ethics & Limitations:

- The data does not include a rider ID, which isn't necessarily a limitation, but if it had been included, analysis on the riders could have been made. It's likely that rider ID was not included as it could be personally identifiable information.
- Data by the user has to be input manually, so there is chance for the data to be mistakes caused by human error; for instance, mistyping their year of birth or selecting the wrong gender.
- There is a chance for the bike that tracks data in real-time to malfunction, causing the data such as trip duration to be inaccurate.
- Citi Bike is transparent with their users the information they collect from them and how it's used, shared, stored and protected, and what their rights and choices are regarding their data. Also, the data they provide for analysis is according to the NYCBS Data Use Policy, so there does not appear to be any ethical concerns with the data.

Questions to explore:

- What day and time of hour do Citi Bike users rent the least/most frequently? How does this vary across age group? User type? Gender?
- What are the most popular pick-up locations across the city for Citi Bike users?
- What are the most popular drop-off locations across the city for Citi Bike users?
- How does the average trip duration vary across age group? User type? Days of the week? Gender?
- Which age group rents bikes least/most frequently?
- How does bike rental vary across user type/age group/gender on different days of the week?
- Does user's age impact the bike trip duration?