Objective

Citi Bike is a public bicycle sharing system serving the New York City boroughs of the Bronx, Brooklyn, Manhattan, and Queens, as well as Jersey City and Hoboken, New Jersey.

The goal is to perform an in-depth exploratory analysis of data, reveal significant trends, derive insights about NY Citi Bike's customer behavior and suggest strategies for better segmentation based on the provided criteria.

Data profile

Data Sourcing: External open data source provided by Kaggle website which is formed of a community of data scientists and developers. It is the world's largest data science community with over one million registered users. Medium trustful source.

Data Collection: Administrative, collected from the company City Bike, so should be dependable. Because where is no clear information about data collection process, meaning that we cannot assume that data totally complete and accurate.

Data limitation and Ethics: The dataset does not contain any PII-related information. **Time lag:** data set is only for period 1 month.

Data Contents: The data contains 1 month bikes rides information in the New York city.

Counts are broken into categories: *trip id, bike id, weekday, start hour, start time, start station id, start station name, start station latitude, start station longitude, end time, end station id, end station name, end station latitude, end station longitude, trip duration, subscriber, birth year, gender.*

Data Relevance: This data source is necessary to assess objectives and research of my hypothesis. Includes a geospatial component and meets the size and variable requirements. The project is a little older than 3 years old but was found in the career foundry brief. This is relevant source of information for my project which I can use.

Data profile

Variables	Data Types			
	time -variant/-invariant	structured/ unstructured	qualitative/ quantitative	qualitative: nominal/ordinal quantitative: discrete/continuous
trip id	time -invariant	structured	qualitative	nominal
bike id	time -invariant	structured	qualitative	nominal
weekday	time -invariant	structured	qualitative	ordinal
start hour	time -invariant	structured	quantitative	discrete
start time	time -invariant	structured	quantitative	continuous
start station id	time -invariant	structured	qualitative	nominal
start station name	time -invariant	structured	qualitative	nominal
start station latitude	time -invariant	structured	quantitative	continuous

start station longitude	time -invariant	structured	quantitative	continuous
end time	time -invariant	structured	quantitative	continuous
end station id	time -invariant	structured	qualitative	nominal
end station name	time -invariant	structured	qualitative	nominal
end station latitude	time -invariant	structured	quantitative	continuous
end station longitude	time -invariant	structured	quantitative	continuous
trip duration	time -invariant	structured	quantitative	discrete
subscriber	time -invariant	structured	qualitative	nominal
birth year	time -invariant	structured	quantitative	nominal
gender	time -invariant	structured	quantitative	discrete

Data Integrity						
Data Accuracy (numeric columns)	*Birth Year*	***Start hour***		*Trip duration*	*Gender*	
minimum	1899	0		60	0	
maximum	1997	23		2697	2	
mean	do not apply	do not apply			do not apply	
	* Variable Birth Year is starts from 1899, y continues with 1900, 1901, 1910, 1917. Probably contains manual errors as in this age no possible to use bikes.					

Data consistency

No duplicates found.		
Count of Birth Year	43021	Dropped missing value (NA)
Counts from rest variables	50000	

Data cleaning

Data Cleaning/Renaming/Reformatting		
	Variables	Changes
	Trip_ID	Dropped full column
	Birth Year	Dropped missing values (NA) 6979 rows Dropped years 1899-1930 31 rows
	Trip duration	Converted from seconds to minutes
	Bike_id	Changing data type to string
	Start_station_id	Changing data type to string
	End_station_id	Changing data type to string

Key questions

What the busiest days of the week and hours of the day (i.e., days and times with the most rides)? What is ride trend overtime?

What are the most popular pick-up and drop-off locations across the city for NY Citi Bike rental?

Which age group rents the most bikes? How their customer habits vary (days, hours)? How does the average trip duration vary across different age groups?

How does the average trip duration vary across hours of day? Do busiest days/hours impact the average bike trip duration?