Project Premise

CitiBank has a New York City bike service where many bike stations are available to the general public. Any user may rent out a bike for a time to ride around the city. There is also a subscription service for these bikes where users can pay a monthly fee for an unlimited use per year. Many different people use this bike service. Are there unifying variables, which connect all these users? What users is this service more popular with and how can a new market of people be convinced to use this bike service? How can CitiBank market using deals or ads?

Data Source:

The data comes from a single dataset, Citi Bike Data by Kaggle user Ryan Cummings. The dataset can be viewed via Kaggle here. The Spotify and Youtube dataset includes data from various NYC Citi Bike users. The data is from October 2013, but the Kaggle page shows the page was updated four years ago. The dataset contains statistics from 50,000 different trips.

Cleaning Data:

Column	Issue	Action
trip_id	Unnecessary for analysis	Deleted
bike_id	Unnecessary for analysis	Deleted
weekday	Confusing column name	Changed name to 'day_of_week'
gender	Column values are not quick to interpret	Changed data type from int to str and then replaced the number values with the following: N/A, M, F

Understanding Data:

Variable	Description
trip_id	Bike trip's individual id
bike_id	Bike's individual id
weekday	Day of the week the bike trip occurred
start_hour	Hour of the day the bike trip occurred
start_time	Date and time the bike trip started

start_station_id	Station id of the station where the bike trip started
start_station_name	Station name where the bike trip started
start_station_latitude	Station latitude where the bike trip started
start_station_longitude	Station longitude where the bike trip started
end_time	Date and time the bike trip ended
end_station_id	Station id of the station where the bike trip ended
end_station_name	Station name where the bike trip ended
end_station_latitude	Station latitude where the bike trip ended
end_station_longitude	Station longitude where the bike trip ended
trip_duration	Trip's duration measured in seconds
subscriber	User's status on whether or not they are a CitiBike subscriber
birth_year	User's birth year, non-subscribers do not enter their birth year
gender	User's gender (0 = unknown, 1 = male, 2 = female)

Variable	Time-Variant or Invariant	Qualitative or Quantitative	Binary or Nominal or Ordinal	Discrete or Continuous
trip_id	Invariant	Qualitative	Nominal	N/A
bike_id	Invariant	Qualitative	Nominal	N/A
weekday	Invariant	Qualitative	Nominal	N/A
start_hour	Time Variant	Qualitative	Nominal	N/A
start_time	Time Variant	Qualitative	Nominal	N/A
start_station_id	Invariant	Qualitative	Nominal	N/A
start_station_name	Invariant	Qualitative	Nominal	N/A
start_station_latitude	Invariant	Qualitative	Nominal	N/A

start_station_longitude	Invariant	Qualitative	Nominal	N/A
end_time	Time Variant	Qualitative	Nominal	N/A
end_station_id	Invariant	Qualitative	Nominal	N/A
end_station_name	Invariant	Qualitative	Nominal	N/A
end_station_latitude	Invariant	Qualitative	Nominal	N/A
end_station_longitude	Invariant	Qualitative	Nominal	N/A
trip_duration	Time Variant	Quantative	N/A	Continuous
subscriber	Time Variant	Qualitative	Nominal	N/A
birth_year	Invariant	Qualitative	Ordinal	N/A
gender	Invariant	Qualitative	Binary	N/A

	bike_id	start_hour	start_station_id	start_station_latitude	start_station_longitude	end_station_id	end_station_latitude	end_station_longitude
count	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000
mean	17615.269360	14.145240	443.321500	40.734170	-73.991109	442.539700	40.733859	-73.991351
std	1675.407446	4.860541	356.559925	0.019911	0.012555	355.756022	0.019885	0.012569
min	14556.000000	0.000000	72.000000	40.680342	-74.017134	72.000000	40.680342	-74.017134
25%	16188.000000	10.000000	304.000000	40.720196	-74.000271	304.000000	40.720196	-74.001547
50%	17584.000000	15.000000	402.000000	40.735877	-73.990765	402.000000	40.735354	-73.991218
75%	19014.000000	18.000000	484.000000	40.750020	-73.981923	483.000000	40.749013	-73.982050
max	20642.000000	23.000000	3002.000000	40.770513	-73.950048	3002.000000	40.770513	-73.950048

trip_duration	birth_year	gender
50000.000000	43021.000000	50000.000000
838.982900	1975.627786	1.073540
573.663997	11.089001	0.589389
60.000000	1899.000000	0.000000
417.000000	1968.000000	1.000000
672.000000	1978.000000	1.000000
1112.000000	1984.000000	1.000000
2697.000000	1997.000000	2.000000

Limitations:

The dataset is current as of October 2013, so the data may not reflect current user trends. Variables such as user IDs and trip costs are not included. This excluded variables prevents any analysis of repeat customers, such as seeing how much they use the CitiBike service. Trip costs can be used to analyze how much cost factors into how users decide on how to ride on a CitiBike.

Ethics:

Data does not contain any HIPPA-related information. Data has been collected from CitiBike and does not make any mention of private information.

Key Questions:

- What variable matters the most when people decide to ride a Citi Bike?
- When is the busiest time for Citi Bike use?
- What is the busiest day for Citi Bike?
- What are the characteristics of an average Citi Bike user? (Age, gender, etc)
- How long do people usually ride for? Do different scenarios change ride times? (Day of the week, time, etc)
- Which stations are people leaving from and arriving at?
- How different are subscriber statistics vs non-subscriber statistics?