Project: Capstone Project 2 – (EDA)

Project name: - Citi Bike– Repricing case study and Twitter Sentiment analysis for real time reputation management

Student Name: - Jitendra Agarwal

Course: - Springboard cohort Jan2 2018

Summary: - Citi Bike is the largest bike share program in us, with 10,000 bikes and 600 stations across Manhattan, Brooklyn, Queens and Jersey City. It was designed for quick trips with convenience in mind, and it's a fun and affordable way to get around town. Everyone knows that bike sharing is the answer to many environmental and urban transportation issues, yet it's not mainstream in US.

I am being asked by the senior executive team at Citi bike to use data science techniques to recommend 3 key action item to increase the company's' business.

Problem Statement: -

- Citi bike management is curious to know if there is any statistical way to find the reason in decline in trips observed few times last quarter.
- o How are user sentiments about a recent change in pricing and if Citi bike is really useful for users from time, cost and efficiency point of view compared to other transportation options?
- o How is the performance of unit economics per trip or per bike or per station? o What is the most common use of Citi bike? o What measure can be taken to increase user trips by 5% with existing users.

Project Goals: -

- User Sentiment analysis from twitter and identify most common customer issues and how to address them.
- o Analyze if any of the current plan can be repriced to get 5% increase on revenue with no customer impact.
- o Can we predict a right number of bike to stationed as a station? o Propose a new monthly pass pricing for office commuters.

About the DATA

Data: -

- Bike trip data provided by Citi bike: https://www.citibikenyc.com/system-data . We will use Q4 2017 data for this analysis
- Daily weather data from open sources

Data provider: - Citi bike NYC

The data includes:

- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station Name
- End Station Name
- Station ID
- Station Lat/Long
- Bike ID
- User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member)
- Gender (Zero=unknown; 1=male; 2=female)
- Year of Birth

This data has been processed to remove trips that are taken by staff as they service and inspect the system, trips that are taken to/from any of our “test” stations (which were being used more in June and July 2013), and any trips that were below 60 seconds in length (potentially false starts or users trying to re-dock a bike to ensure it's secure).

Data notes:

- Trip count and mileage estimates include trips with a duration of greater than one minute.
- Mileage estimates are calculated using an assumed speed of 7.456 miles per hour, up to two hours. Trips over two hours’ max-out at 14.9 miles.
- Data only include trips that begin at publicly available stations (thereby excluding trips that originate at our depots for rebalancing or maintenance purposes).

Some Stats: -

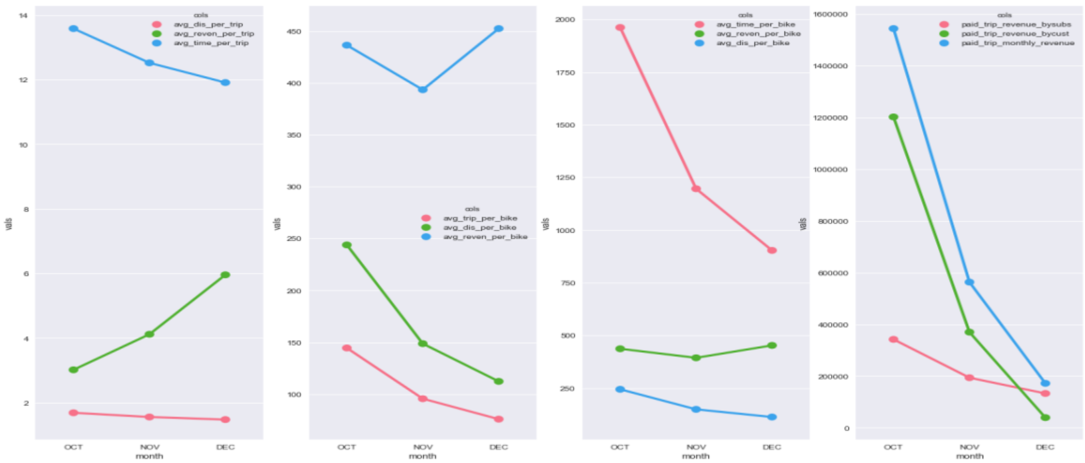
Data Analyzed:	Q4 2017 (Oct, Nov, December 2017)
Number of Trips Analyzed: -	4109518

Data Pre Processing and Data Wrangling

- Read Data from csv
- Divide data in 2 part of monthly subscriber and Day pass holder
- Convert birth year to Int and calculate user age
- Filter data for user age > 100 as a outlier
- Rename columns to remove spaces from their name
- Add a unique trip id column for each row as a unique key.
- Extract Date and hour from timestamp column
- Determine if a trip is Free or paid, and calculate paid units and paid amount for each trip
- Create a new DF with unique list of station list
- Drop column which are not needed to reduce DF size
- Get Weather data from csv and store in Data Frame.
- Calculate Monthly unit metrics such as
  - Total distance
  - Average distance per trip
  - Average time per trip
  - Average distance per bike
  - Average time per bike
  - Average trip per bike
  - Average Revenue per bike
  - Average revenue per trip
- Calculate monthly revenue by different streams

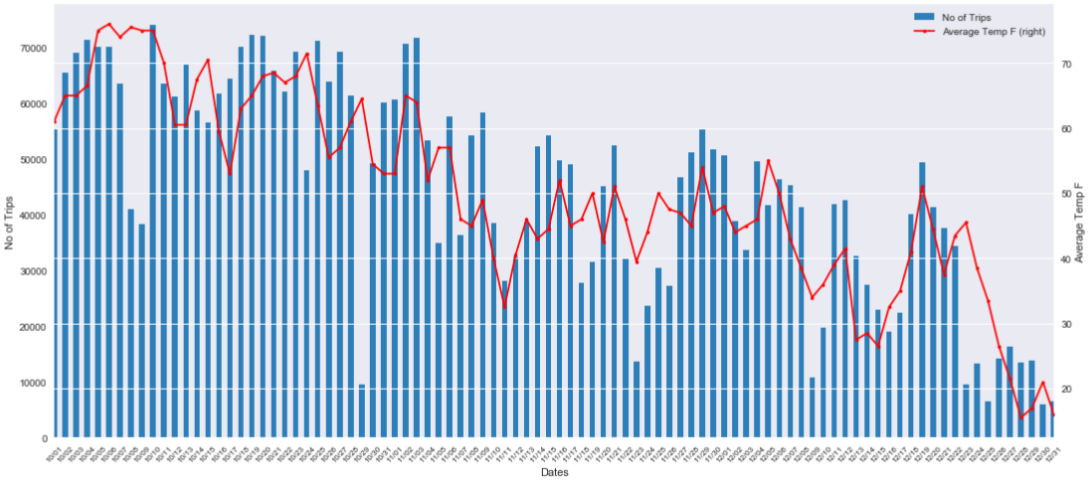
Data Visualization and Exploratory Data Analysis: -

1. Plot the Key unit metrics per trip and per bike for each month



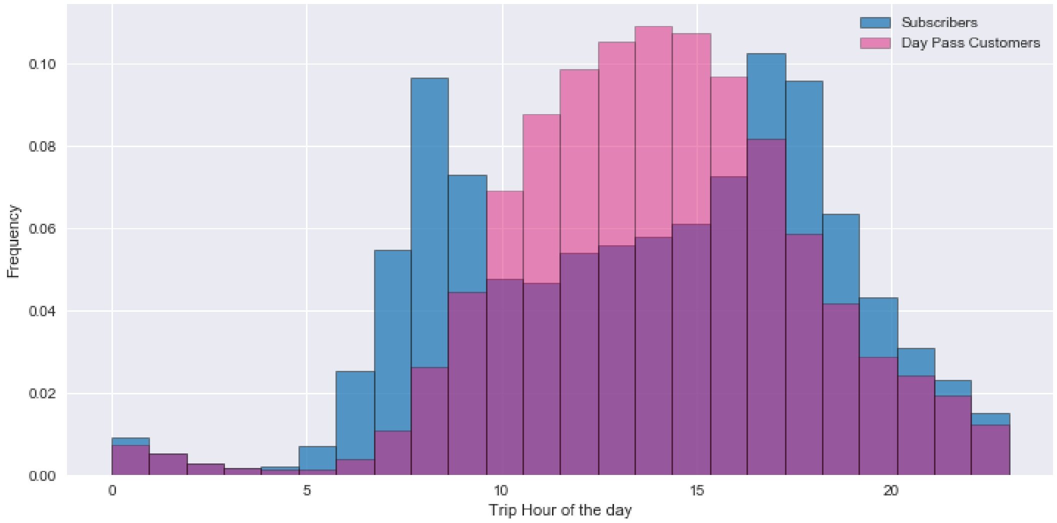
**December month have seen less number of trips and hence the trip and distance per bike decreased but revenue per bike has increased.

2. Trips happened on each day of Q4 and plot it with the average temp for each day



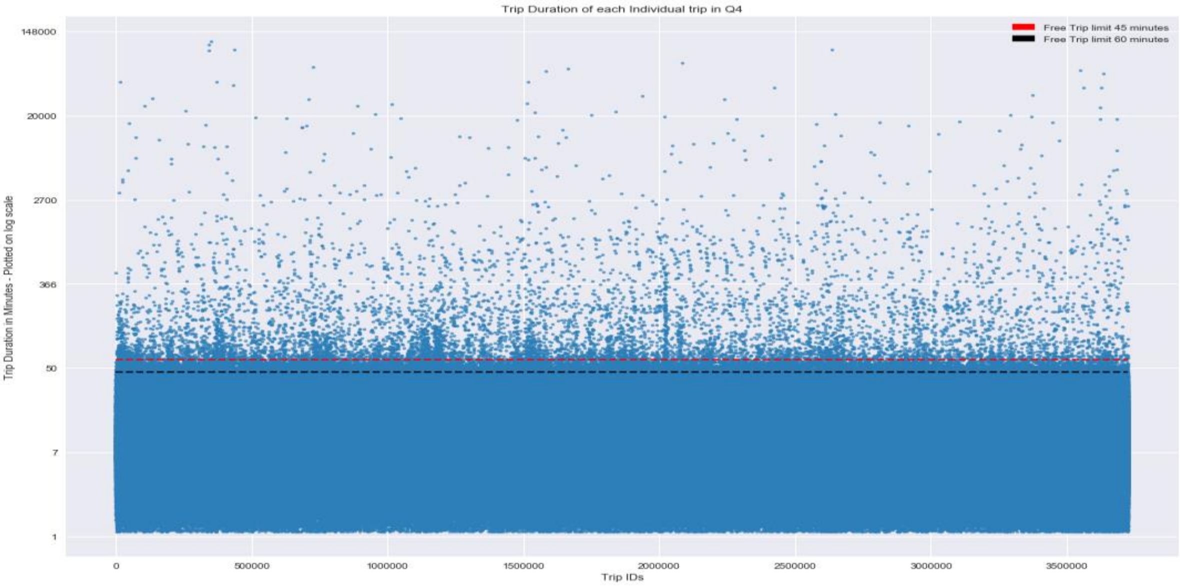
**Temperature is a big factor for people to take a bike trip. in the cold days the bike trips were dropped significantly.

3. Plot trip distribution by hour of the day in entire Q4 for subscriber



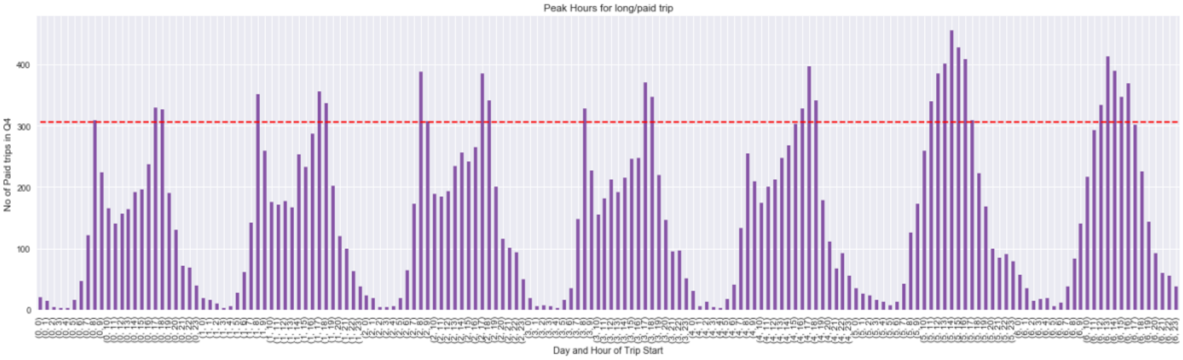
*** Monthly subscribers take most trip during Rush Hours while Day pass holder are taking most trips between 11 AM-4 PM*

4. Plot Trip Duration of each trip and show the threshold of free trip limit of 45 minutes for Subscribers



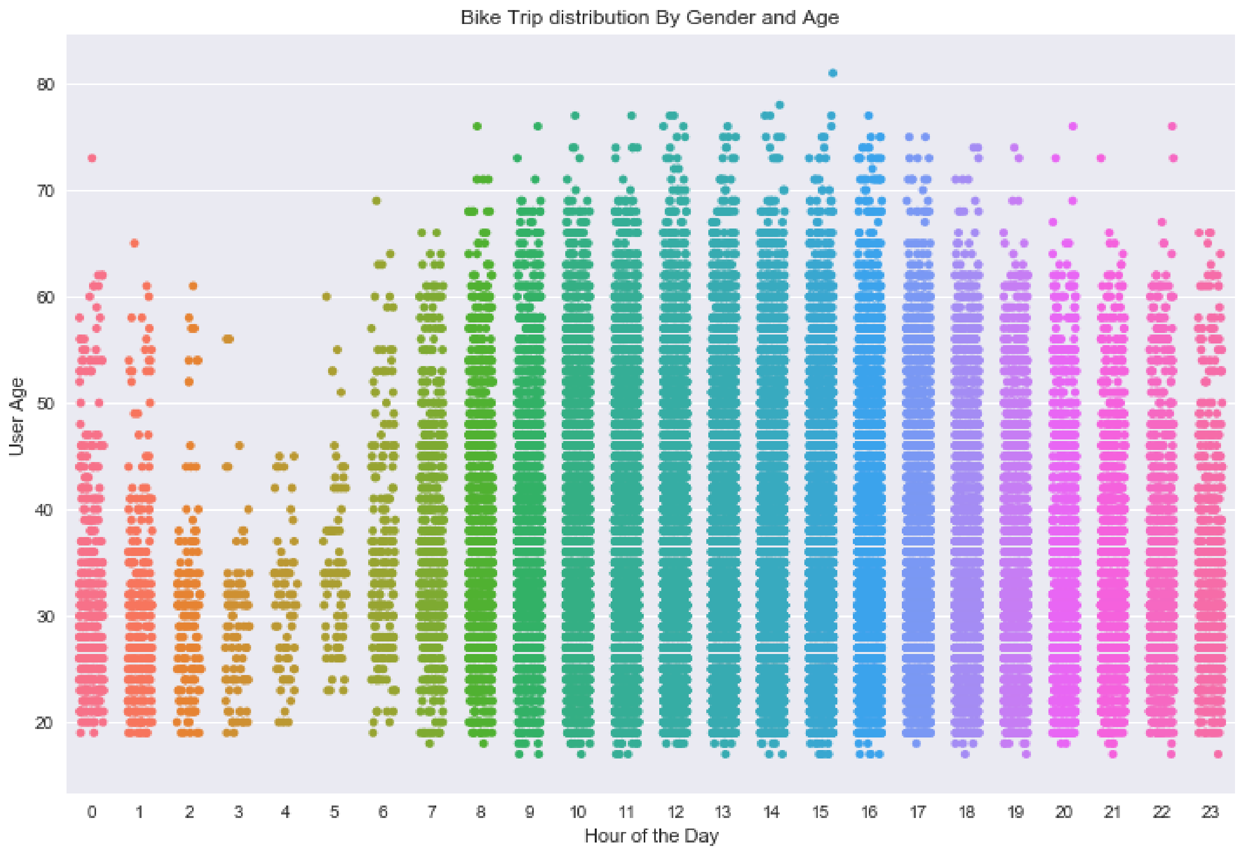
** We have used log scale to tune the long tail of trip duration. Area under 45 minutes and 60 minutes' limit is equally high density. there are many outliers of trips duration up to multiple Days.

5. Trip Distribution by Day/time of a week: When are the paid trips happening: -



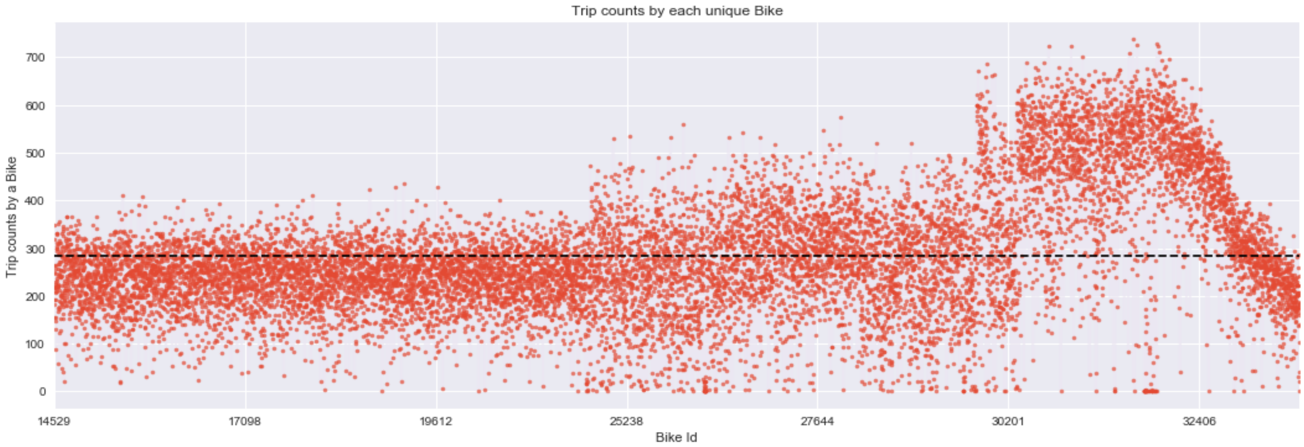
** Subscribers trips count is little more than average during Rush Hours in weekdays and afternoon in weekend. Hence less Deviation.

6. Plot Bike Trip Distribution by Gender and Age



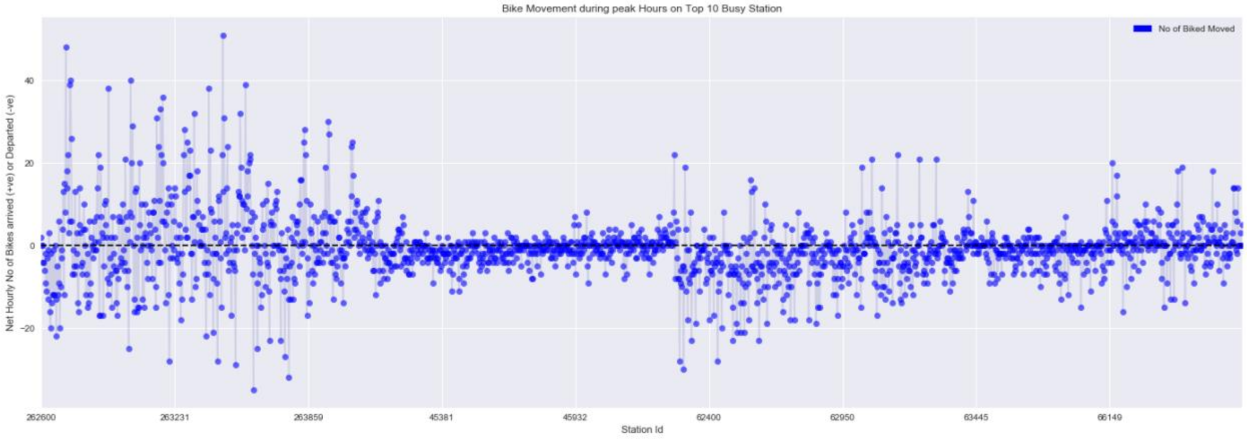
** Male from 25-35 is most common user segment for Subscribers. Younger users are taking trips during late night hours also.

7. Plot Trip counts by Bike Ids

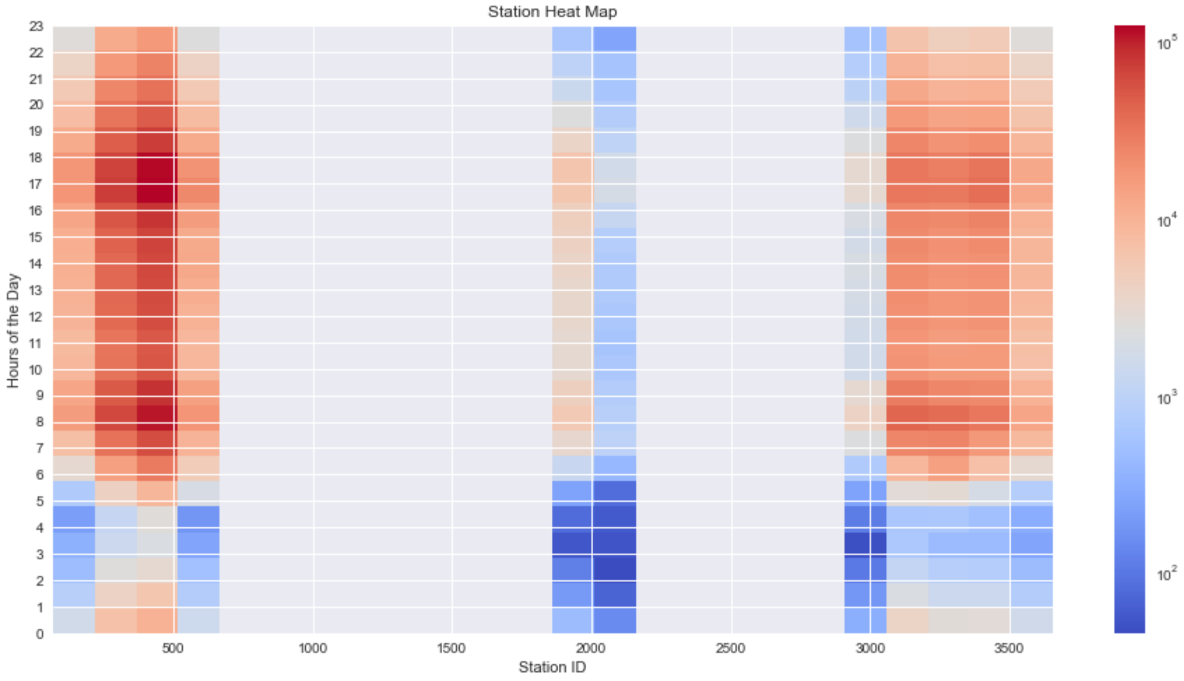


** While most bikes are taking less trips then overall average, some bikes are taking way more trips than average (almost double). These may be due for major upgrades.

8. Track and Plot Bike Movement in Peak Hours



9. Create a Heat map to show which station is busy against the hour of the day.


Station Heat Map
Hours of the Day
Station ID
10^5
10^4
10^3
10^2