CITIBIKE - A DATA SCIENCE CASE STUDY

a little bit for EVERYONE

A City on Citibike

BIKE SHARING IS WORKING...

THE PROBLEM OVERVIEW

- How is the performance of unit economics per trip, per bike, and per station over the months?
- is there is any statistical way to find the reason for a decline in trips observed few times last quarter?
- How many bikes should be placed at a station so that customer always finds a bike when he needs?

THE GOAL

- Exploratory analysis to find meaningful patterns in data
- Predict the right number of bike to stationed at a station
- User Sentiment analysis from Twitter and identify most common customer issues and how to address them.

Data Acquisition and pre-processing

Read Data from CSV into data frame
Divide data into 2 part of monthly subscriber and Day pass holder
Filter data for user age > 100 as an outlier
Add a unique trip id column for each row as a unique key.
Extract Date and hour from timestamp column
Check if a trip is Free or paid, and calculate paid units and paid amount for each trip
Drop column which is not needed to reduce DF size
Get Weather data from CSV and store in Data Frame.

- Convert birth year to Int and calculate user age
- Calculate monthly revenue by adding different streams
- Calculate Monthly unit metrics such as
- o Average distance per trip
- o Average time per trip
- o Average distance per bike
- o Average time per bike
- o Average trip per bike
- o Average Revenue per bike
- o Average revenue per trip

THE BIG PICTURE

4th Quarter 2017

TIME AND DISTANCE ANALYSIS

PERFORMANCE

1.5 MM+ TRIPS

MONTHLY TRIPS

- 92% trips taken by Monthly subscribers
- Average 8 monthly trips by a subscriber. 20% utilization.
- Trips dropped in December to an average of 4 monthly trips only.

2.8 MM+ MILES

MONTHLY MILES

- October clocked 4 Million miles while in December only 2 million miles were traveled.
- Subscriber average trip is for 2 miles while day pass holder average trip is for 11 miles.

266 K HOURS

MONTHLY HOURS

- in October total 375K hours of trips clocked while December was less than half, total 167k hours
- The average trip time for subscribers is less than 15 mins while day pass holder it is 90 Mins. and 97% trips end within 1 hour.

700 Stations, 13K Bikes, 2.4 MM Subscribers, 1 Great City

STATS ARE INTERESTING !

150

Monthly average trips per bike, approx. 280 miles covered by a bike in a month

74 %

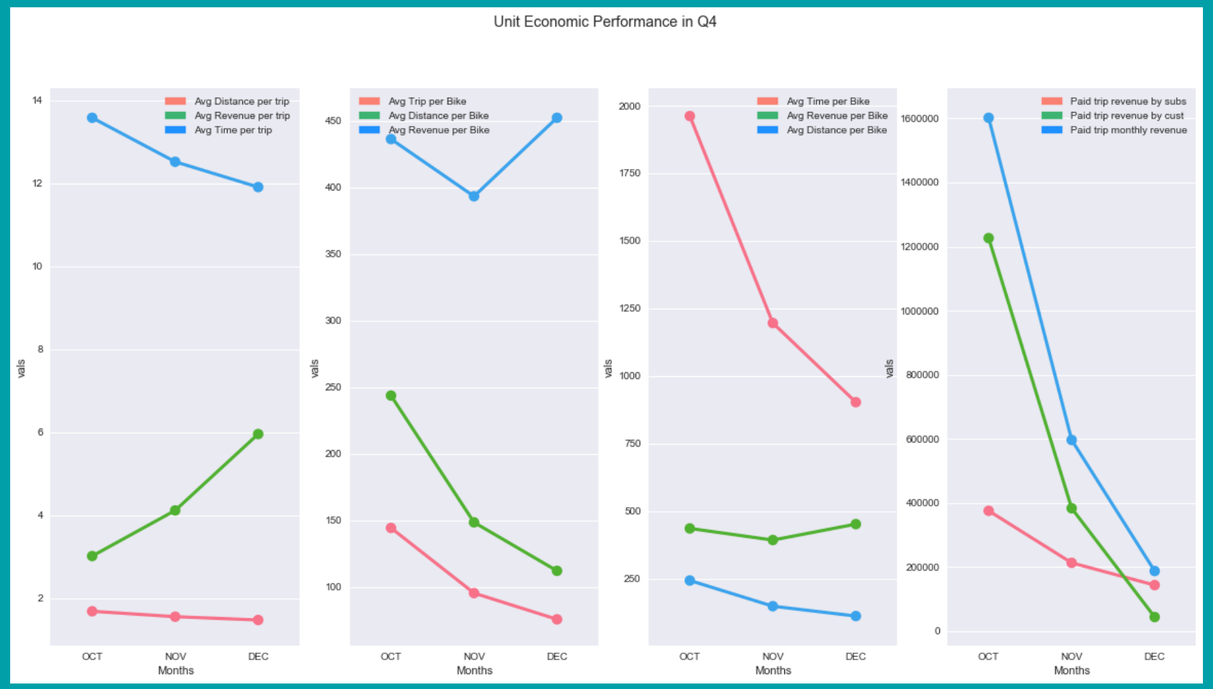
Trips were taken by male riders, only 24% were by Female riders and 2% are unknown.

99 %

Trips by Subscribers are under 45 Minutes limit. 83% trips by Day pass holder are under 30 minutes limit

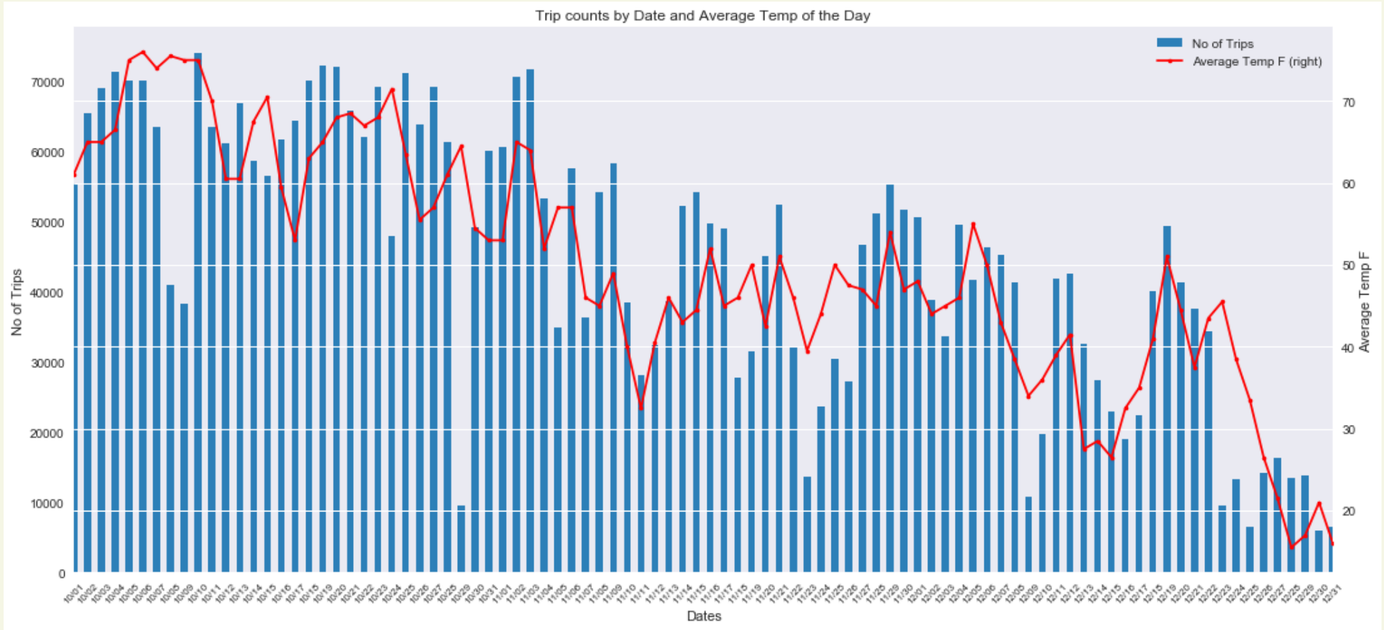
28-30

Most common age group of Rider. 63% trips were taken by rider between 25-42 years Age



-No of trips and other Key unit matrix went down for the month of November and December.
Average Bike utilization remained half in December compared to October.

-November and December are very cold months in NYC and bike rides are not a popular way to commute. Whenever the temperature drops, number fo trips drops significantly.

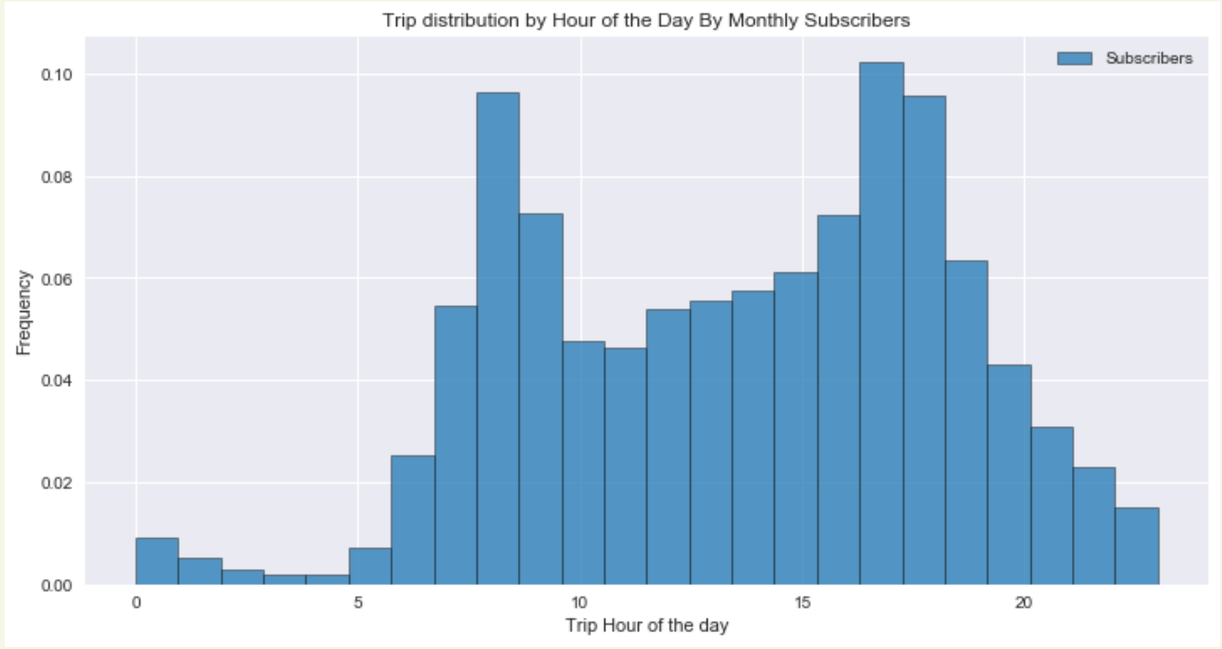


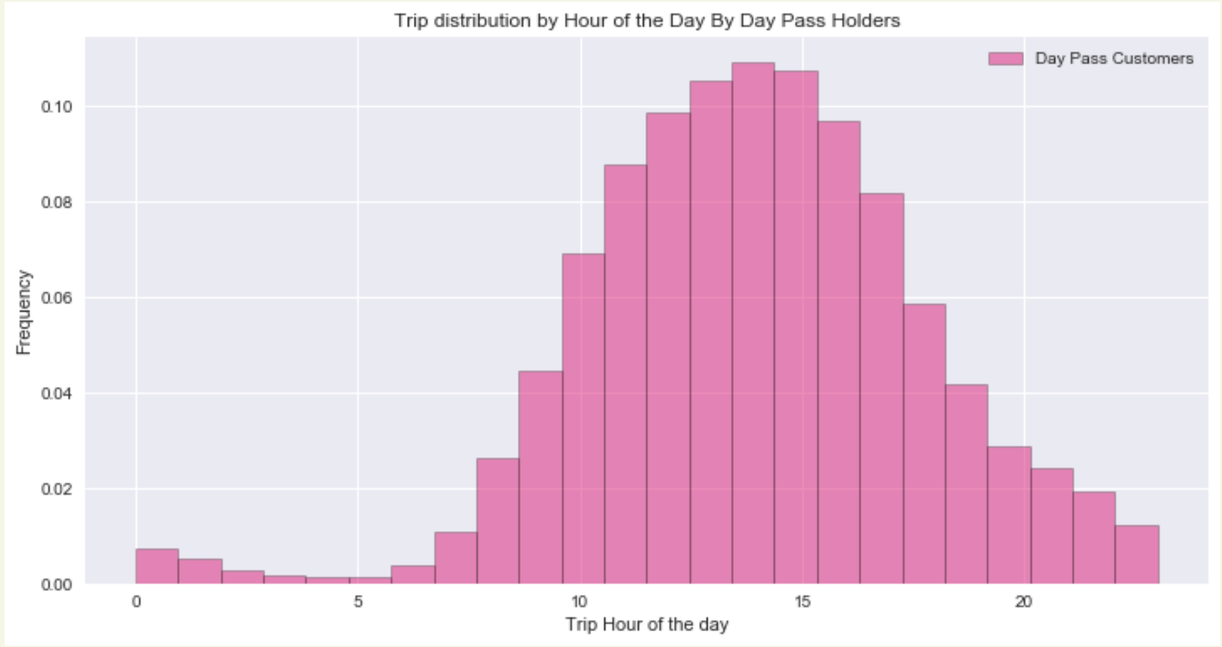
- 8AM-10 AM and 5PM-7PM is the most Popular time of trips for monthly subscribers.

Top Stations by foot traffic

- Pershing Square North
- Broadway & E 22 St
- W 21 St & 6 Ave
- E 17 St & Broadway
- 8 Ave & W 31 St

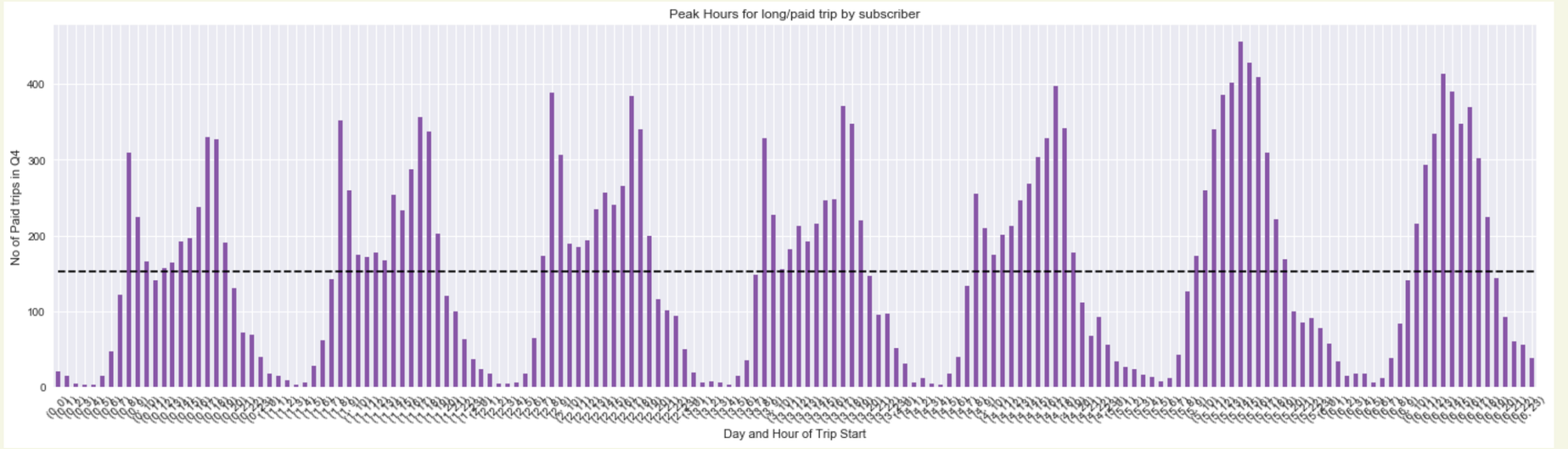
- For daily pass holders, the most Popular time of commute is 11 AM - 4 PM



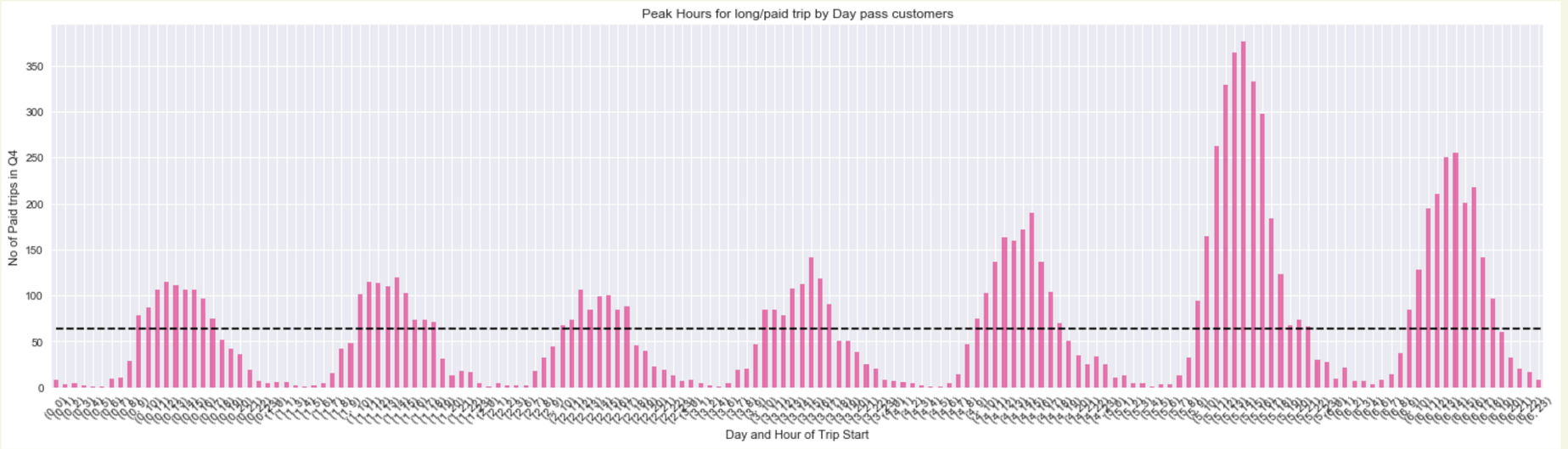


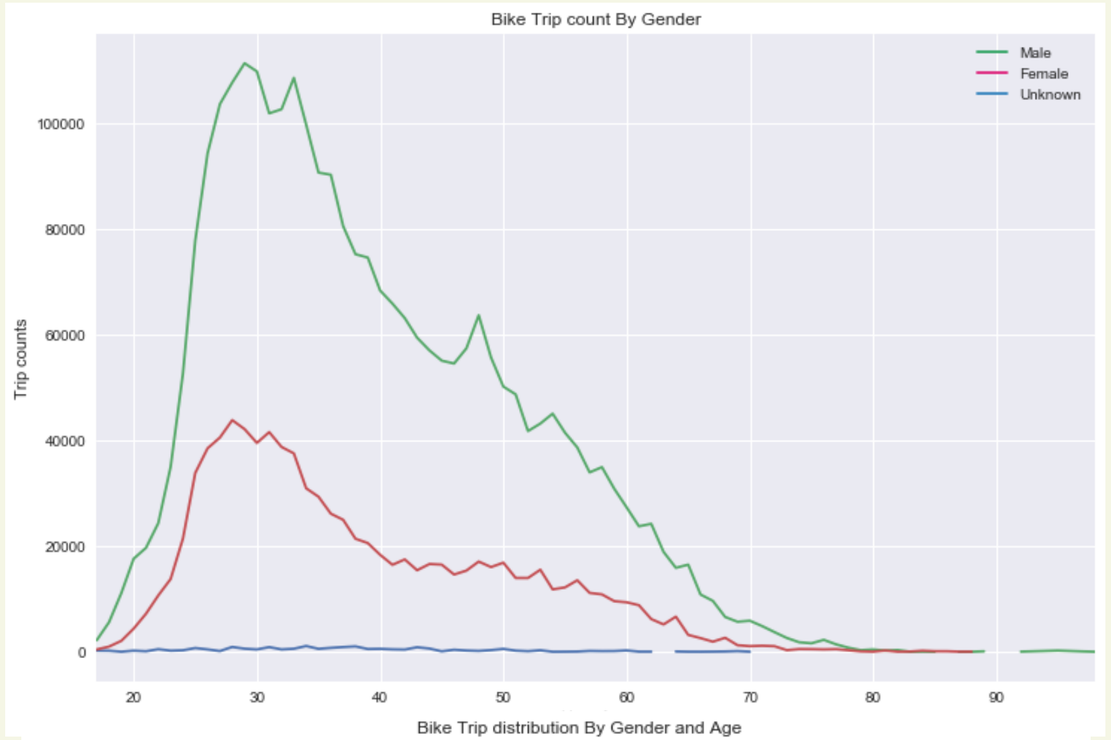
POPULAR TIME

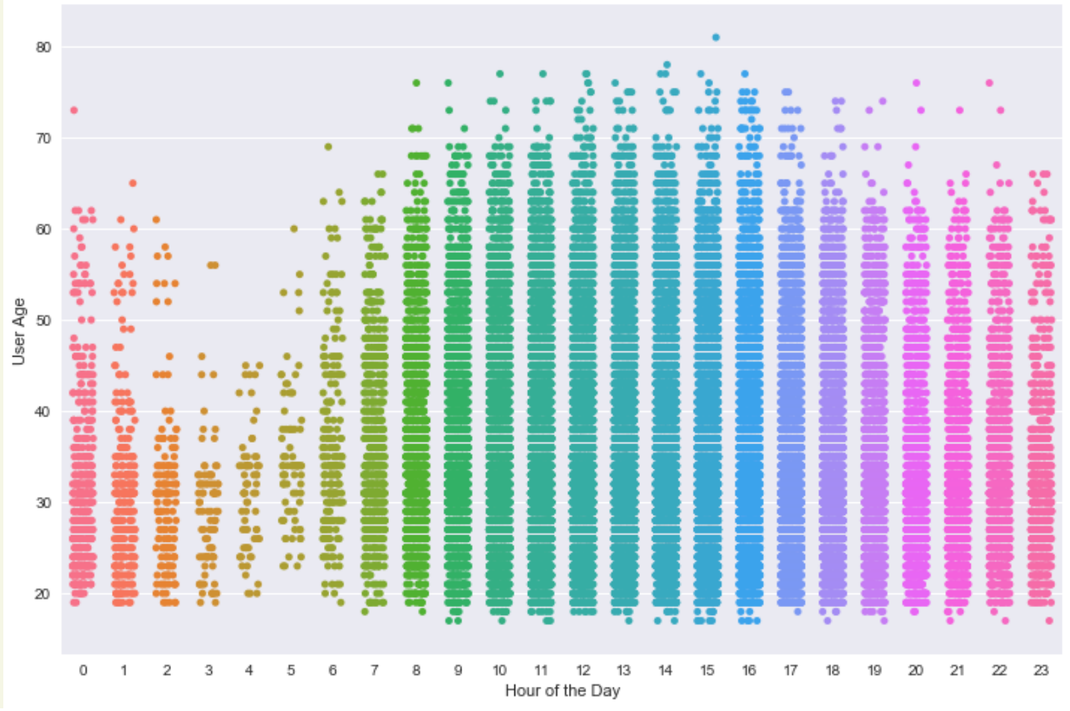
- Subscribers are taking most trips during morning and evening rush hours on weekdays.
- Weekend too are popular with subscribers during daytime hours.



- Day Pass holders are taking fewer trips during weekdays.
- Weekends are very popular with pass holders during daytime hours.







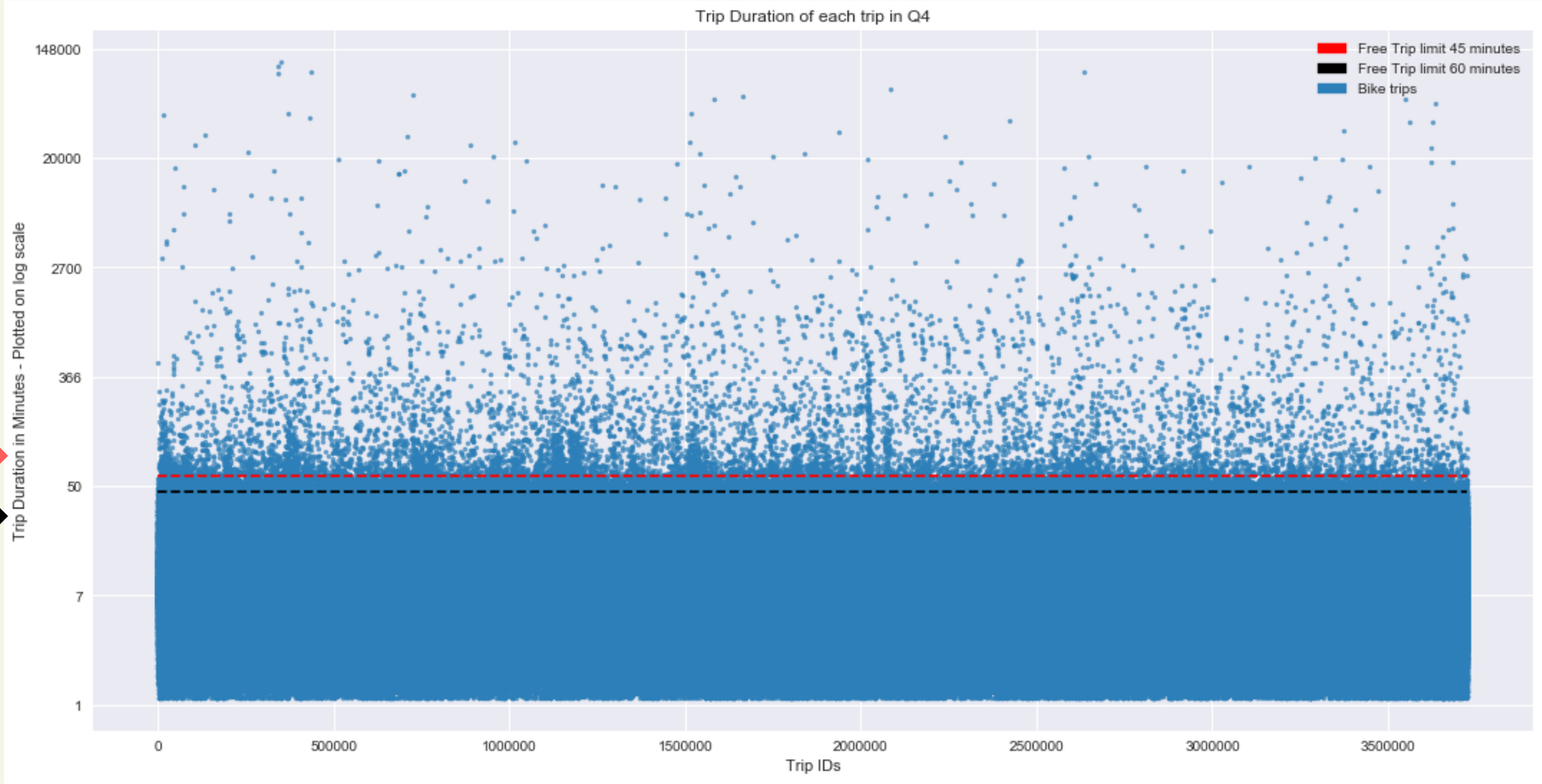
- Male Riders are taking 74% trips, and most common age is between 28-32 years.
- There are significant number of drivers with age 65+

- Bike rides are popular in all age group for the entire day, however, only younger riders are taking trips between 1 AM - 6 AM

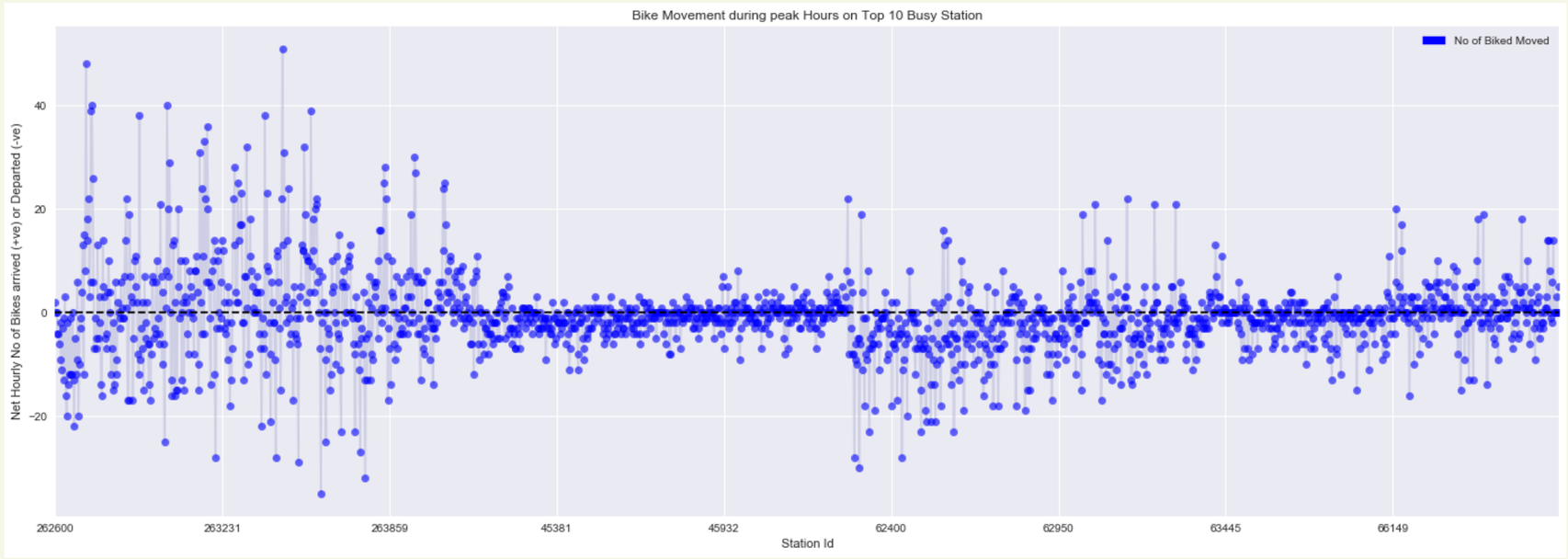
TRIP DURATION

Trip Duration: 60 Minutes

Trip Duration: 45 Minutes

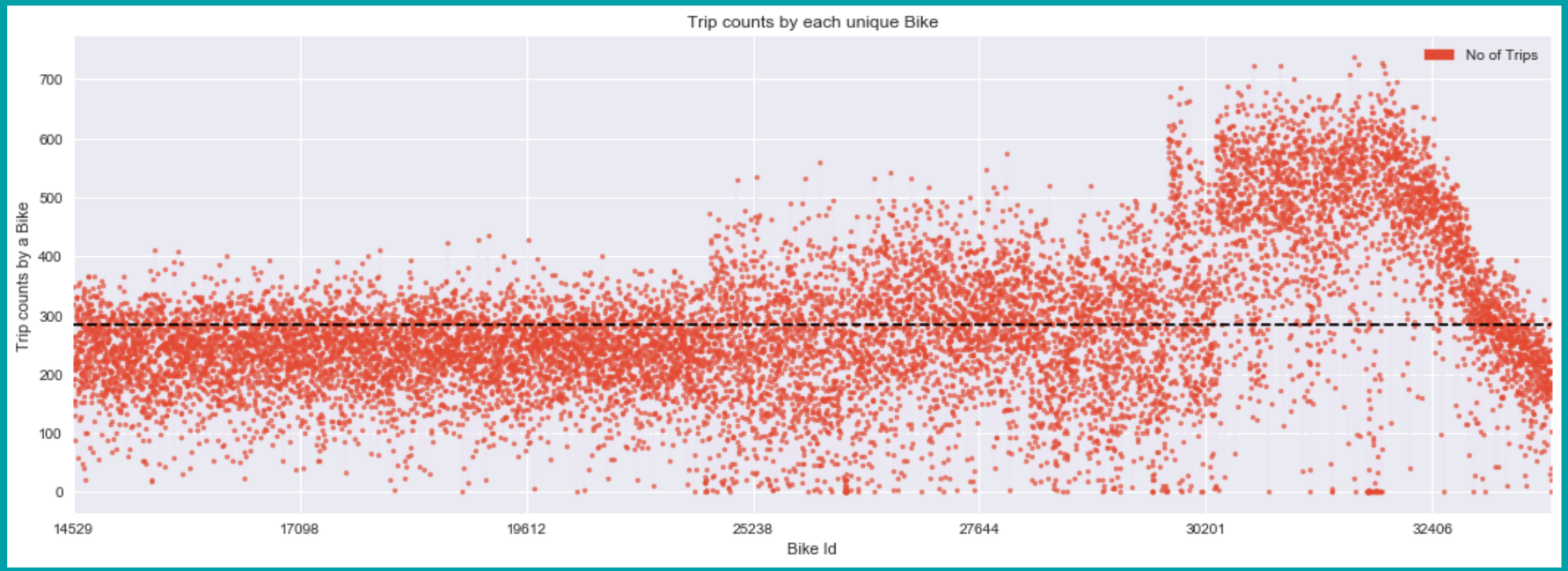


- 94% trips by Monthly Subscribers are ending under 45 minutes
- 99% trips by Monthly Subscribers are ending under 60 minutes
- Monthly Subscribers are taking trips even up to 10 hours

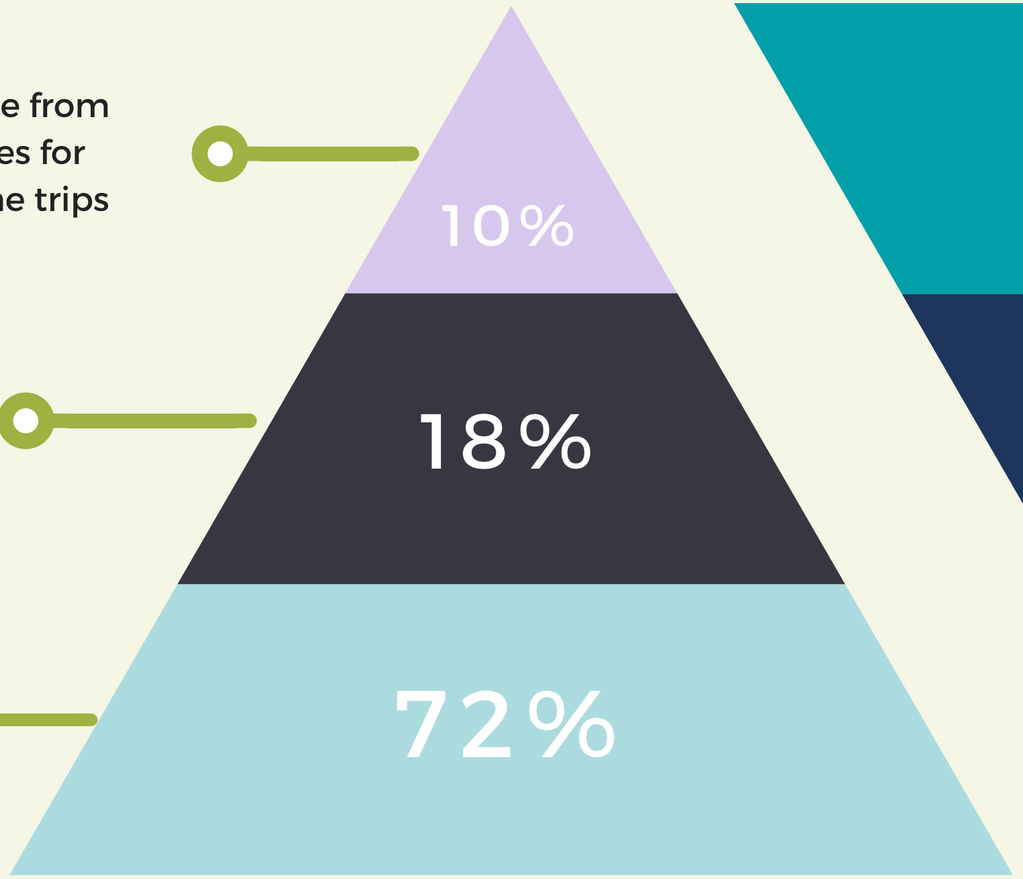


-Net Bike Movement for a station drives the need to reload bikes manually. For top 10 stations, net bike movement per hours is higher than other stations.

-No fo trips done by a bike is less than average for most of the bikes. however, few bikes are doing trips more than double the average. these might be due for a maintenance.



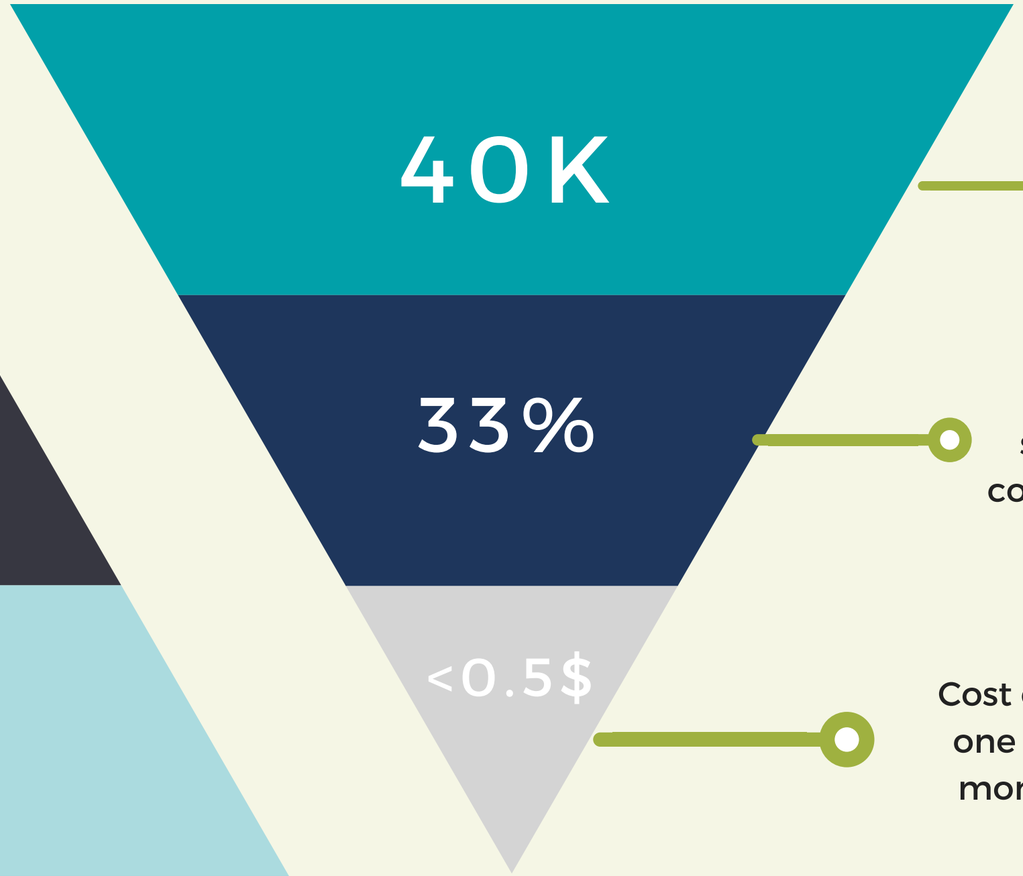
REVENUE ANALYSIS



Revenue from charges for overtime trips

Revenue from Day pass sales

Revenue from Monthly Subscriber

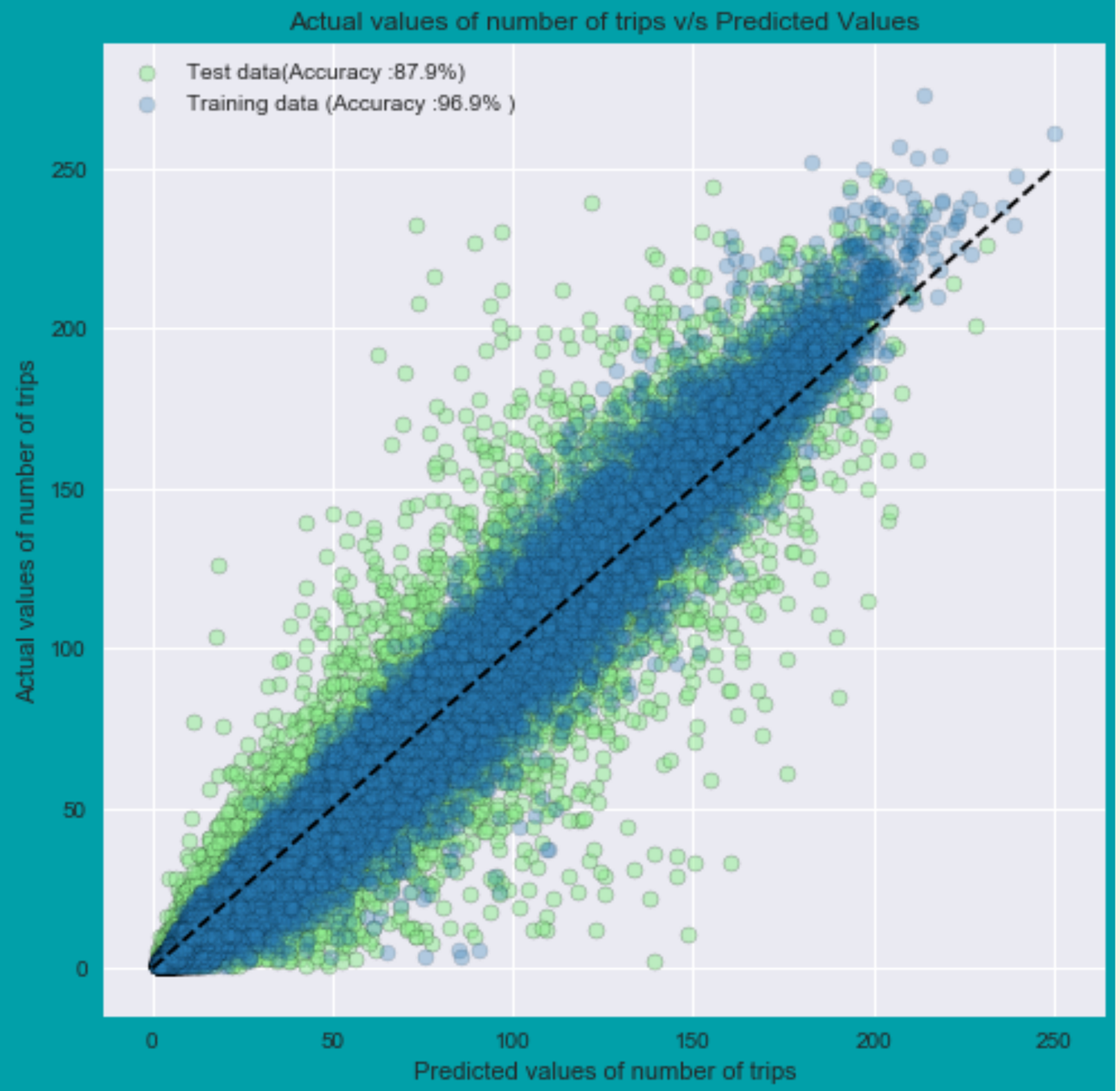


40K Trips are taken per day by Subscribers

Trips by subscriber are completed within 45 minutes

Cost of a trip for users if one takes 40 trips per month up-to 4 miles

PREDICTED V/S ACTUAL VALUES



MACHINE LEARNING OUTCOME

- Based on Machine Learning exercise we can predict the no of trips from a station with 87% accuracy.
- The no of trips can be predicted by providing the values of Station id, Average temperature, station density score and Day of the week.
- Most important feature for prediction is Station id.

- Random Forest Regressor was used as Machine learning model
- After tuning, train data prediction accuracy was 97% and test data prediction accuracy was 87%
- CitiBike trip data volume is huge and it has high variance in distribution as not all places have an equal footprint of bike riders.
- Naturalized mean Square error value was only 0.06770262

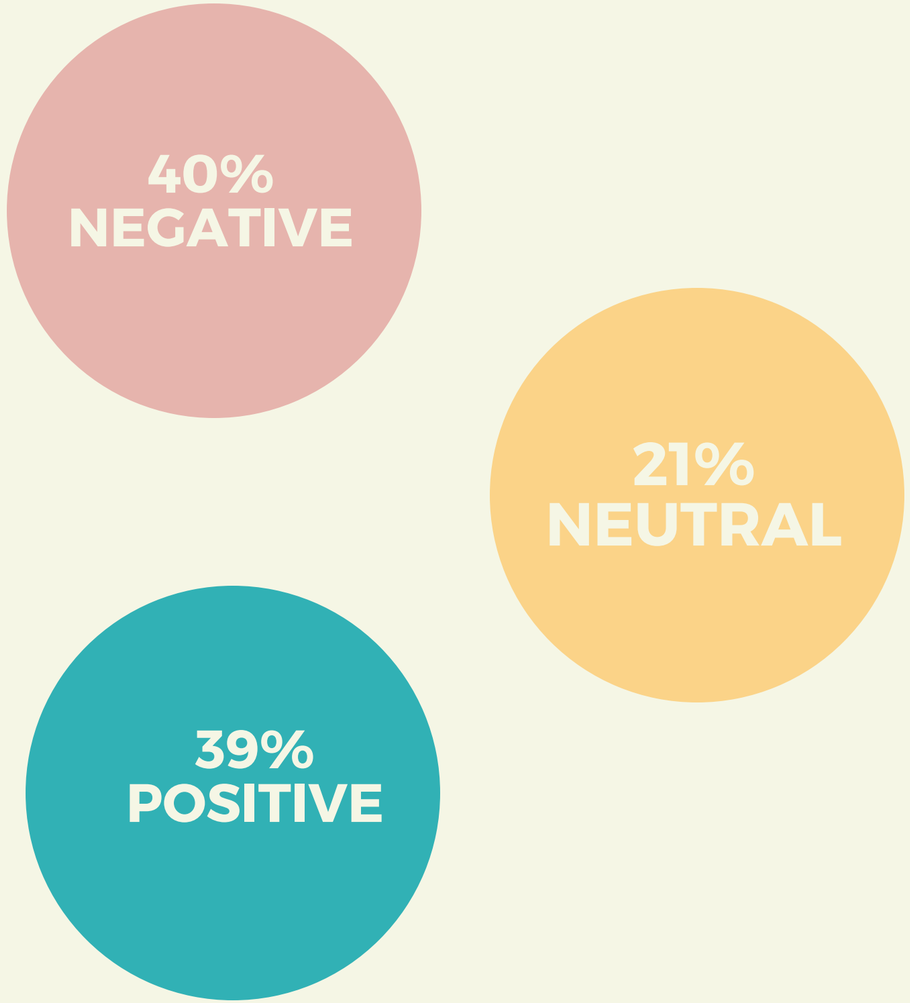
USER SENTIMENT ANALYSIS

- We analyzed 140 tweets from last 20 days for top 2 hashtags for Citibike '#citibike', '#bikenyc'

- We also analyzed latest 300 yelp public reviews for Citibike for last 1 month

- **45% tweets are with positive sentiment, 42% with neutral and 13% tweets with negative sentiment**
- **78% yelp reviews are having negative sentiment**
- **The combined sentiment is 39% positive and is towards negative due to high negative Yelp reviews.**
- **NLP machine learning model accuracy was 96%**

- **Most common 10 words on twitter are bikenyc, nyc, lane, bike, citibike, transalt, protected, nypd, like, th**




KEY BENEFITS :-
• SAVE MONEY
• SAVE TIME
• HAVE FUN
• GET EXERCISE
• GO GREEN
Secaucus
West New York
Union City
Hoboken
Jersey City
Croxton
Pulaski Skyway
Communipaw
Greenville
Colonels Row
Statue of Liberty
LINCOLN HARBOR
Randalls and Wards Islands
Rikers Island
LGA
COLLEGE POINT
CORONA
ELMHURST
Flushing Meadows-Corona Park
FRESH POND JUNCTION
BUSHWICK
BROOKLYN MANOR
CYPRESS HILLS
OZONE PARK
PARKSIDE
Fulton St
Myrtle Ave
Jamaica Ave
Queens Blvd
80th St
New Jersey Turnpike
County Ave
Tonnelle Ave
FDR

RECOMMENDATIONS

#1 Increase Stations in low-income areas and high residential areas to encourage more women to take rides.

#2 Dock surfing is an issue. Introduce a 5 min wait time between check-in and check-out from the same dock, especially for day pass holders.

#3 A user activity analysis is recommended using user's data to predict when a particular user will take a trip.

#4 Pricing is perfect from the user point of view but a $1 premium option to reserve a bike in up to 1 hr in advance can be looked at.

References:

- Trip data source:- https://www.citibikenyc.com/system-data campaign (**CitiBike**)
- Weather data:- http://w2.weather.gov/climate/
- yelp data:- Yelp public reviews open source web scraping.
- Tweet Data:- twitter API
- Inspirations 1:- https://github.com/toddwschneider/nyc-citibike-data\
- Inspirations 2:- http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/
- Inspirations 3:- http://toddwschneider.com/posts/taxi-vs-citi-bike-nyc/