

NYC Taxicab Fact Book

Big Data Project – 2016

Vaibhav Aggarwal va771
Ginni Malik gm1908



Table of Contents

INTRODUCTION	3
EXPERIENCE.....	3
ISSUES	4
ANALYSIS	5
EXPERIMENTAL SETTING	25
INDIVIDUAL CONTRIBUTION.....	26
CONCLUSION.....	27

INTRODUCTION

The taxicabs of New York City are widely recognized icons of the city. They are *yellow* and *green*.

We have selected the project area – “Exploring NYC Taxi Trips” and have chosen to explore the yellow taxi data.

The Fact Book is a quick look at the state of the yellow taxi industry. It is a regular summary of taxi trends in New York City. It contains statistics on updated taxi trips and fares along with unique projection of trip patterns by boroughs.

We have studied yellow taxi data over three years, namely from: **2013-2015** and made some analysis.

Since the data is colossal, it is cumbersome to go through all the trip-sheets. The introduction of TPEP technology, the data has become easy to read and analyze as everything is electronic.

As the population of New York City is increasing, the number of yellow taxi cabs have increased constantly. The analysis of this data helps us find the various traveling patterns of New Yorkers. This can be useful in making certain critical decisions for the betterment of the community.

EXPERIENCE

This project has been a massive learning experience. Apart from our analytical skills, we built our technical and creative skills. We gained a new perspective on how data can be analyzed and used to derive important conclusions. Working as a team, posed certain issues like code collaboration, keeping track of changes by individual team members, etc. To counter this problem, we used version control software like **Git** which helped us maintain track of the changes being made to the code. Also at the time of failure, we could revert back to the previous version of the code. This proved to be a very efficient way of collaborating our code. Overall, we can say for sure that the way we perceive data has changed. We are able to successfully approach any problem and build an arsenal of data analytic methods.

ISSUES

Data used:

- TLC Trip Record Data- Yellow Cabs (2013-2015)
- Spatial Index to group data
- Used geoJson for coordinates of locations

1. **Incompatible Data:** The files for each year had different attribute names and some files had more number of attributes. For example, the 2015 data had additional attributes like extra, improvement surcharge, etc. This produced erroneous results for different years.

Solution: Generic map-reduce code was written such that it could work with data of all the years. The main/required attributes and their variations in the different files were considered and generic if-else conditions were included to check the data.

2. **Massive Datasets:** The total data that had to be analyzed was around **65GB** which is extremely large to analyze on our local machines. Therefore, we used **Amazon Web Services** to run the different map-reduce tasks and produce our outputs. The uploading of the datasets on the S3 bucket and the execution of the map-reduce tasks took a very long time, approximately **15hours**. Hence, it was very time consuming.

Solution: For the initial testing phases, the data was truncated using the **head command** and the code was run on a sample version of the file locally. This helped find errors in the code and rectify it. Once, the correctness of the code was assured, the final map-reduce tasks were run on the Amazon cluster to generate the final outputs. This proved to be more efficient.

3. **Spatial 2D-Indexing:** The data consisted of immense number of coordinates in terms of latitudes and longitudes. It was a complex task to map them into 2D-Index as searching the polygon is a complicated task and took a large amount of time due to the usage of nested loops.

Solution: Hence an index of all the possible neighborhood polygons was created using R-tree index. This reduced the time by half which was very helpful.

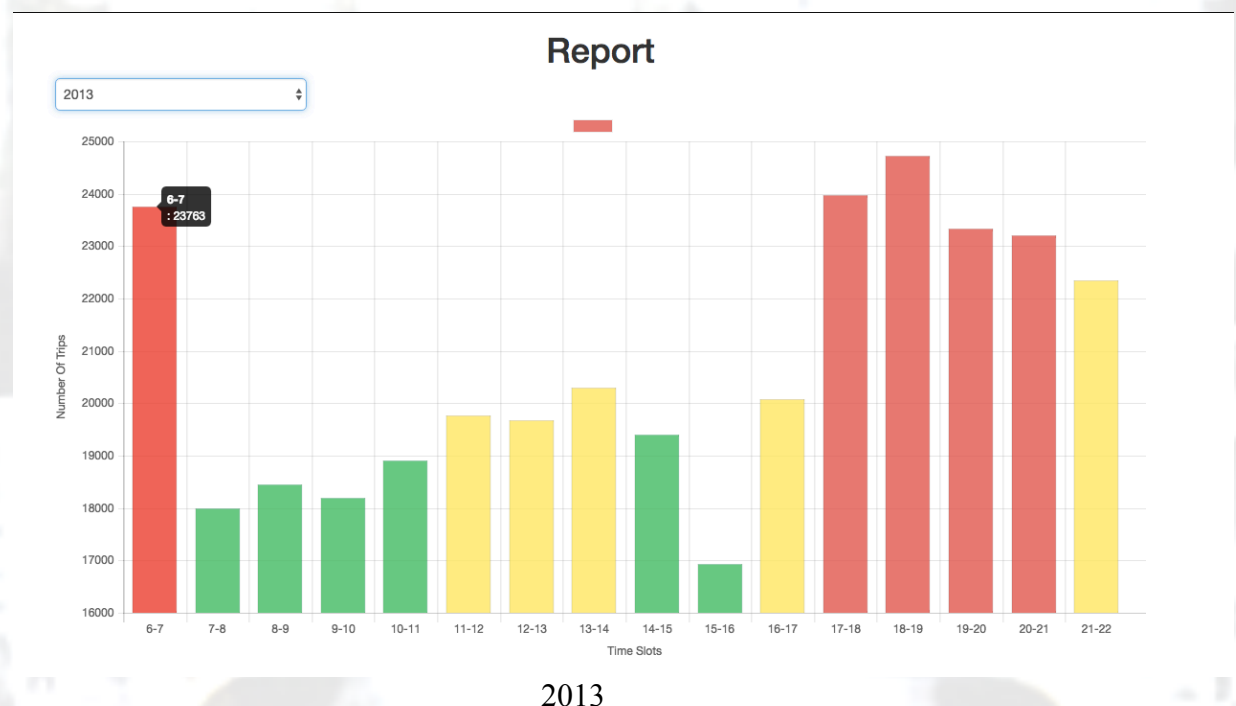
4. **Visualization in Multi-Dimension:** The output results of the map-reduce tasks had several dimensions which resulted in visualization problems. It is very difficult to consider all the dimensions of the data at once and comprehend.

Solution: We grouped the dimensions of the data and took less dimensions at a time. This made the results clearer, easy to analyze and derive conclusions.

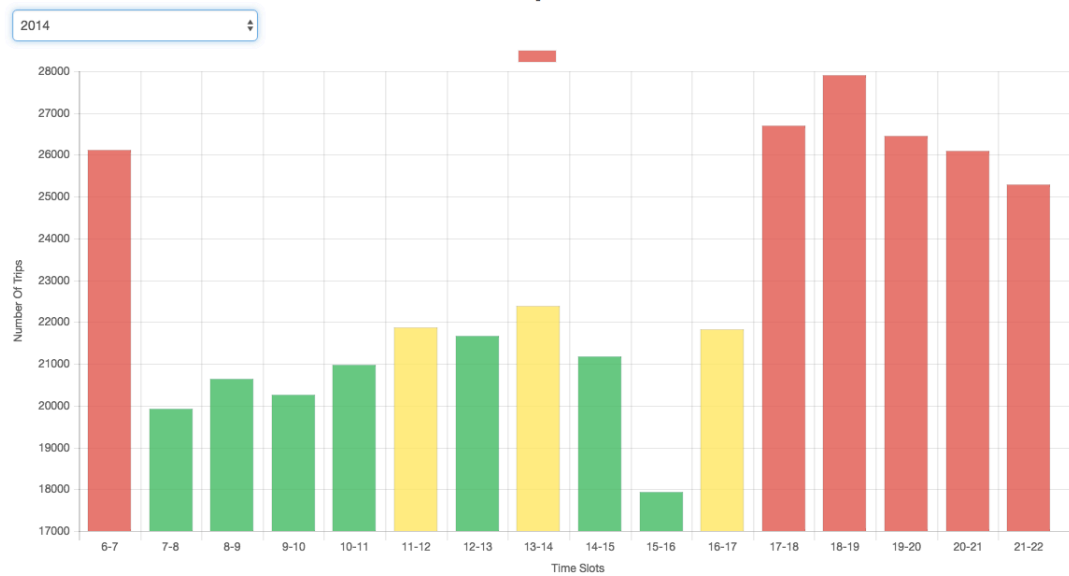
ANALYSIS

1. AIM: Traffic Analysis by hour(6am-10pm) per year

We have generated the number of taxi trips for each hour between 6am to 10pm in a day. This helped determine the peak traffic hours on a particular day. People can access this information and try to avoid the traffic hours by making planned trips to their destination.

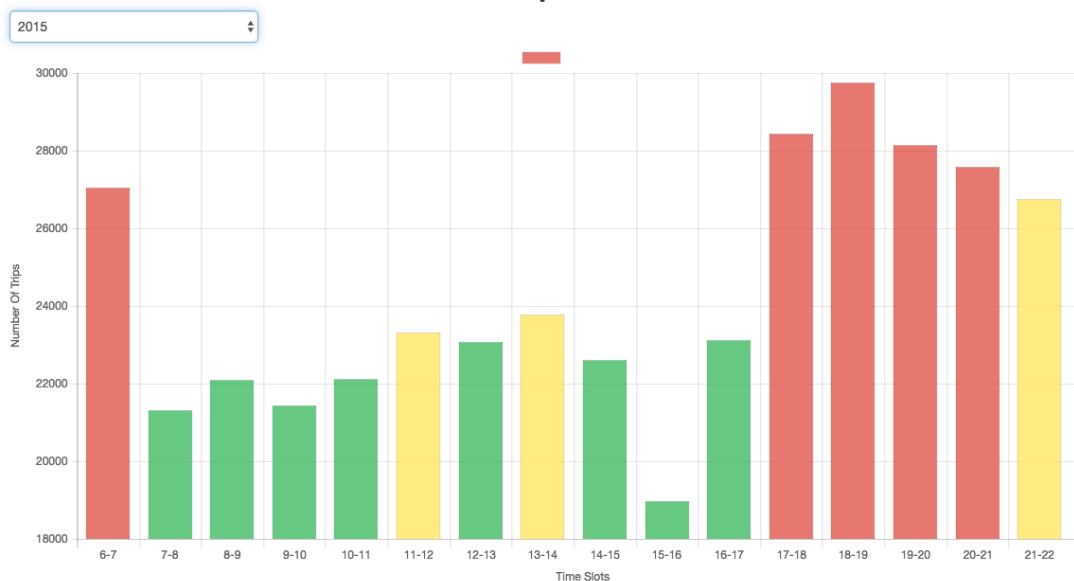


Report



2014

Report



2015

Conclusion:

For the sake of simplicity, we have empirically derived the values for the ranges of peak hours.

Red: Peak traffic $\geq 1.1 \times (\text{mean of the total number of trips})$

Yellow: Medium traffic

Green: Low traffic $< 0.9 \times (\text{mean of the total number of trips})$

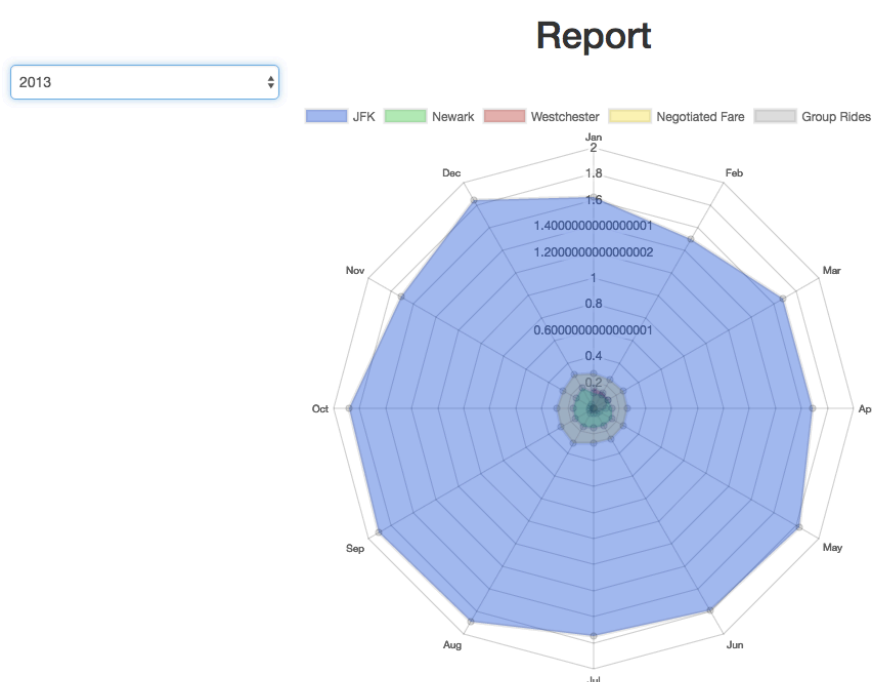
As you can see from the above results, the peak hours are generally between **6am-7am** and **5pm-9pm**.

This may be because, there might be many people traveling to work and kids going to school in the morning between 6am to 7am. Between 5pm-9pm, everyone is usually rushing home after their work. Some people might also be going out for casual outings.

Also, there has been a **constant growth** in the **number of trips** every year between all the time ranges. This could be because more people are moving into the city each year and hence we need more number of taxis to accommodate the population.

2. AIM: Rate code analysis

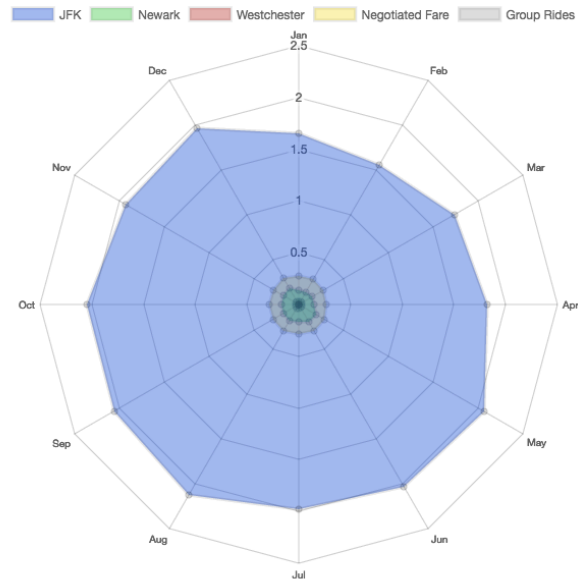
Based on the rate code data present in the files, we have generated a comparison between the most frequently visited airports around New York. Also, a comparison between rides with negotiated fare and group rides.



Comparison between JFK and other airport destinations 2013

Report

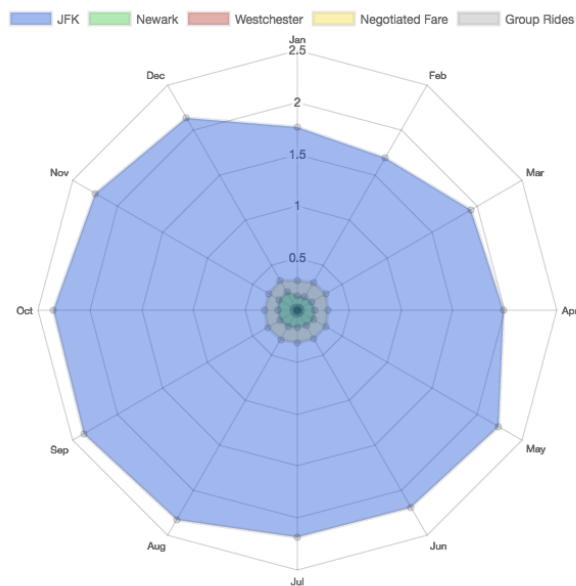
2014



Comparison between JFK and other airport destinations 2014

Report

2015



Comparison between JFK and other airport destinations 2015

As depicted in the graphs above, JFK is the most visited airport by the yellow taxis.

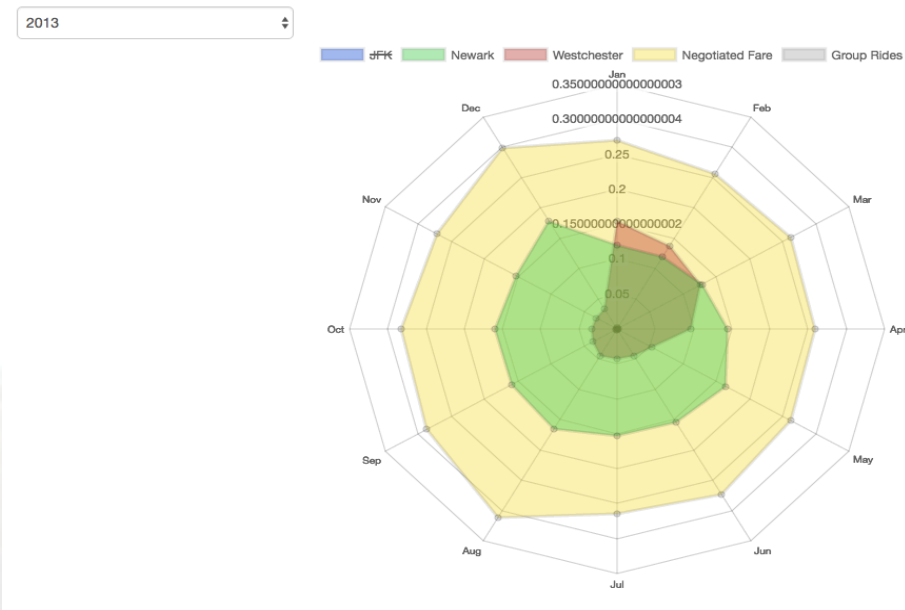
Conclusion:

JFK is the **busiest** airport.

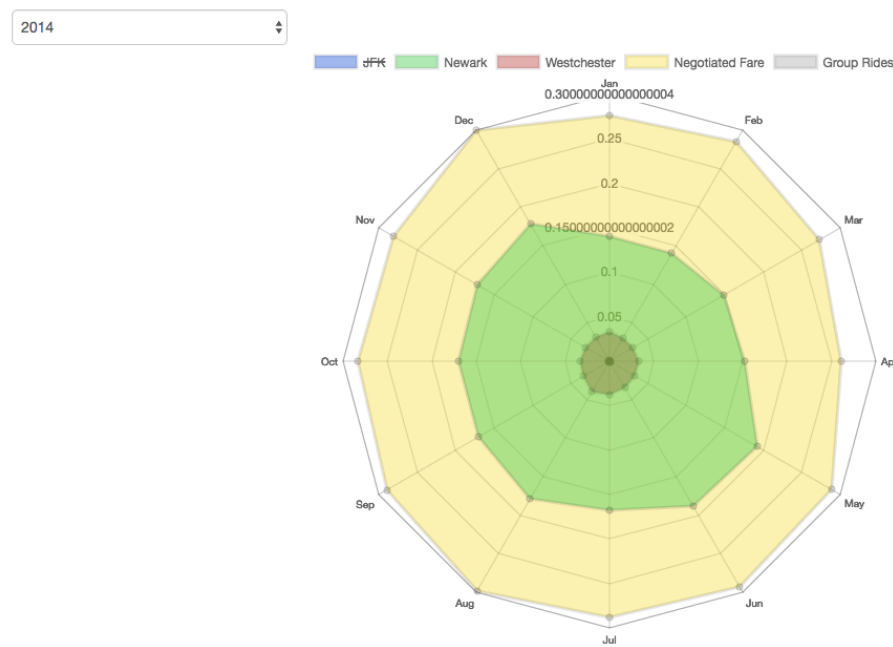
Also, we can see that from November to April, there is a decline in the number of trips to the airports. This shows that during these months, people prefer to stay within New York as it is one of the most famous tourist destinations in the world and has a lot of

activities during those months. Hence, people prefer staying within the city and celebrating.

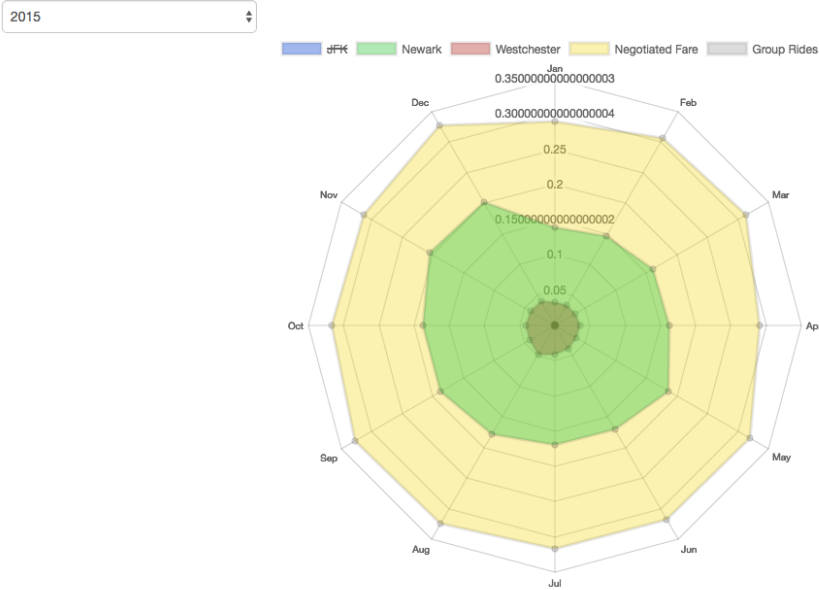
Report



Report



Report

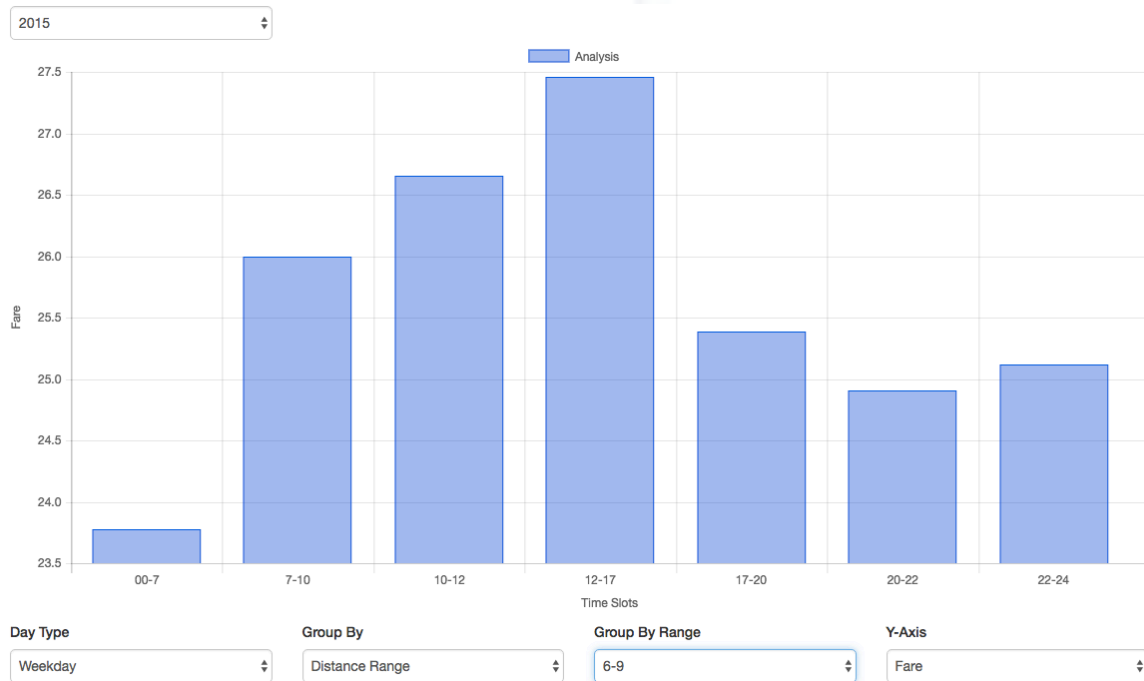


Comparison between Newark, Westchester as airport destinations and rides with negotiated fares

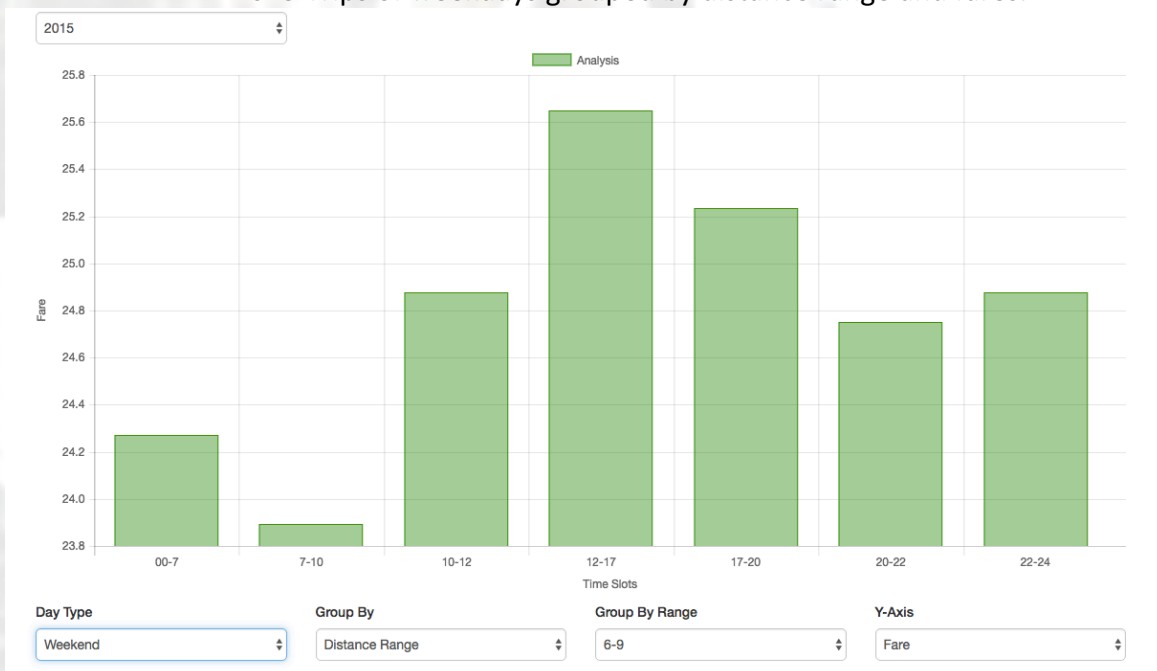
As observed, we can say that the preference for Westchester airport has decreased considerably over the years.

3. AIM: Day type Analysis (Week/Weekend) based on timeslots and distance range

We have analyzed the data to determine the fare for a given week day or weekend. This information can also be filtered for the time slots and distance ranges. The results provide us with the significant travel patterns of New Yorkers based on timings, distance ranges, trip fares and the days.



2015 Trips of weekdays grouped by distance range and fares.



2015 Trips of weekends grouped by distance range and fares.

The above graphs give the results for the Fares amount for a given distance range at various time intervals. We can point out that the fare on **weekends** during the hour **00-7** is **higher than** on **weekdays**.

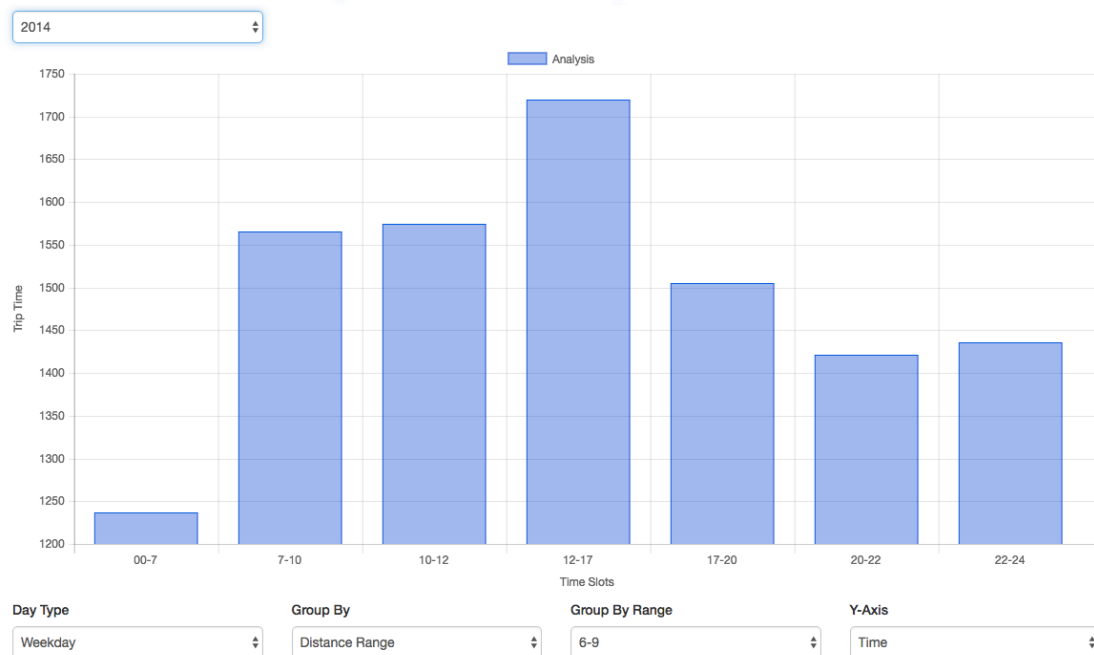
Conclusion: This may be due to the reason that a large number of people travel during weekend mornings for quick getaways. In order to provide weekend service, the fares are set high since there is a high demand for the taxis.

Also, during the **hours 7-10**, the fares on weekends are less as compared to weekdays

Conclusion: Less number of people are traveling and hence, there is less traffic on the streets.

The fares between 8pm to 12am on weekends have drastically increased

Conclusion: People go out for dinners and parties on weekend nights. This again leads to high demand of taxis and consequently the fares are raised.



2014 Trips of weekdays grouped by time slot and trip time.



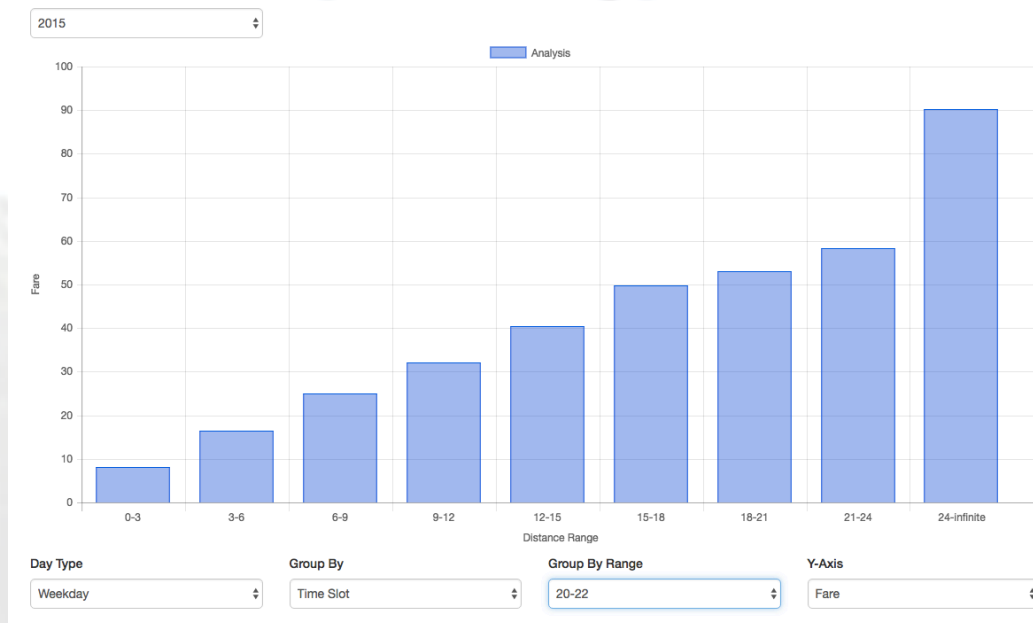
2014 Trips of weekends grouped by time slot and trip time.

The above graph shows that the time taken to travel 6-9 miles is higher on weekends between 12am - 7am than on weekdays.

Conclusion: This is because many people travel on weekend mornings for quick getaways.

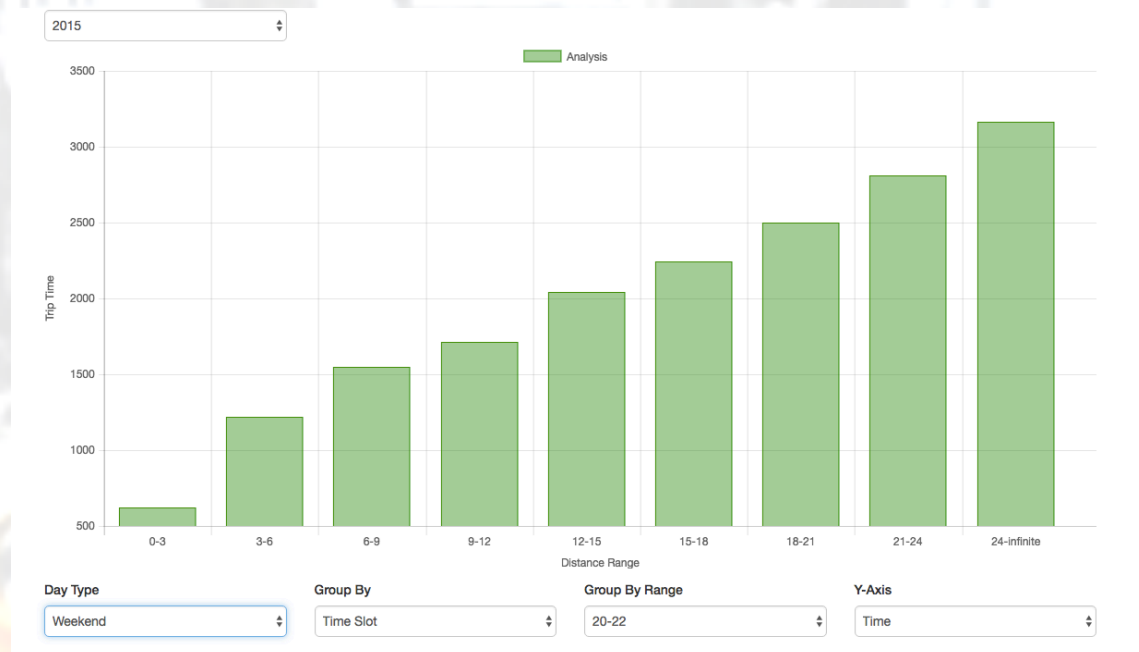
It takes longer to travel 6-9 miles between 12pm and 5pm on both weekdays and weekends.

Conclusion: More people traveling at the same time and hence traffic considerations.



2015 Trips of weekdays grouped by distance range and fare.

From the above graph we can say that the fares increase with increase in the distance traveled.

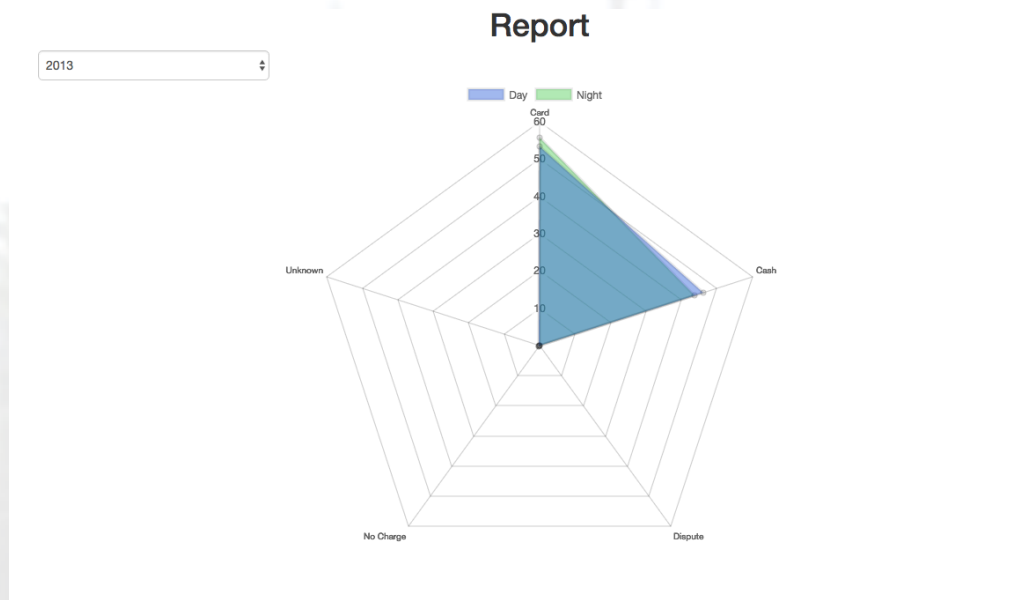


2015 Trips of weekends grouped by distance range and trip time.

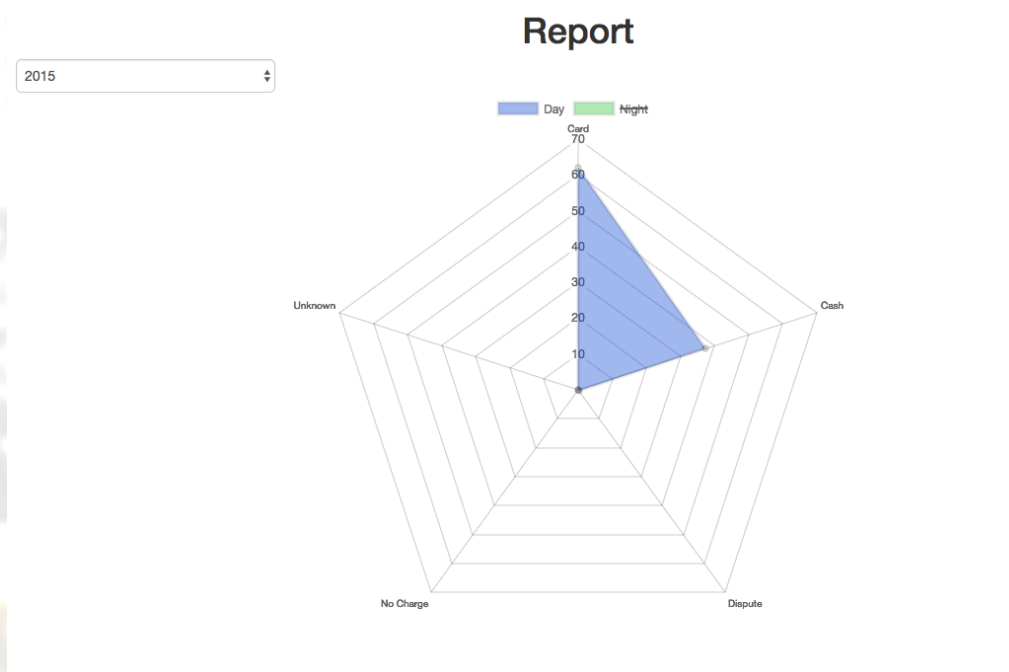
The result from the above graphs conclude that the trip time increases as the distance increases which is the expected result.

4. AIM: Cash and Credit usage comparison between Day and Night

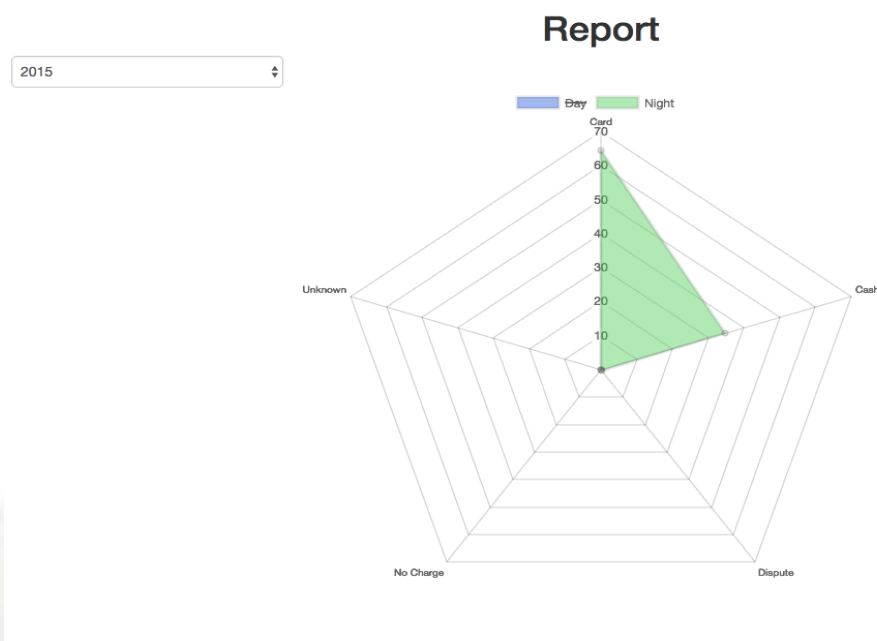
We have analyzed the payment methods of various trips and derived a relationship between the payment method and the time of the day.



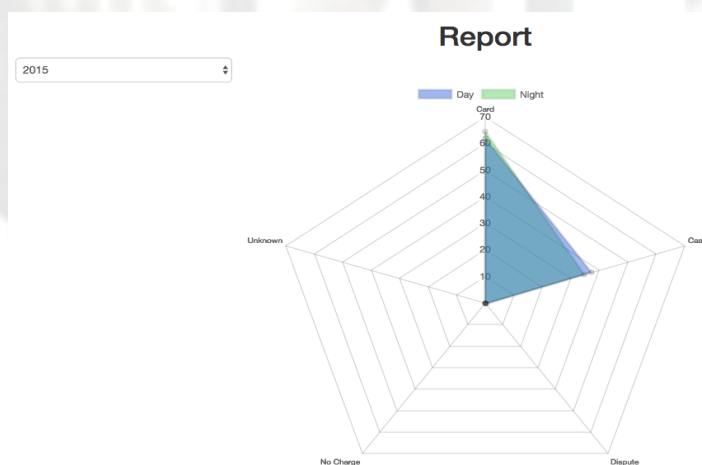
2013 report of payment methods



2015 report of payment method used during the day



2015 report of payment method used during the night



2015 report of one payment method as compared to the other

We can see from the above reports that, the payment through **card (61%)** is more preferred than **cash** payments (**37%**) regardless of day or night. Also, the payments made through card are much more in the night than the ones during daytime.

Conclusion: People avoid carrying cash during the night due to the fear of getting mugged.

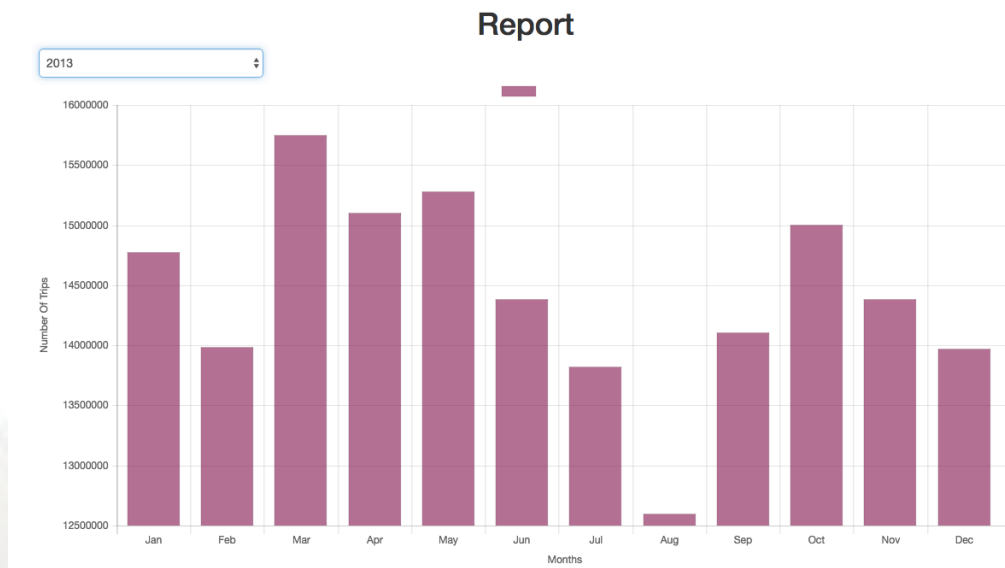
Also, there has been an increase in the amount of payments done through card since **2013 (55%)** to payments done through card in **2015(61%)**

Conclusion: This could be because of better algorithms and technologies used to ensure secure transactions. People have more faith in the system.

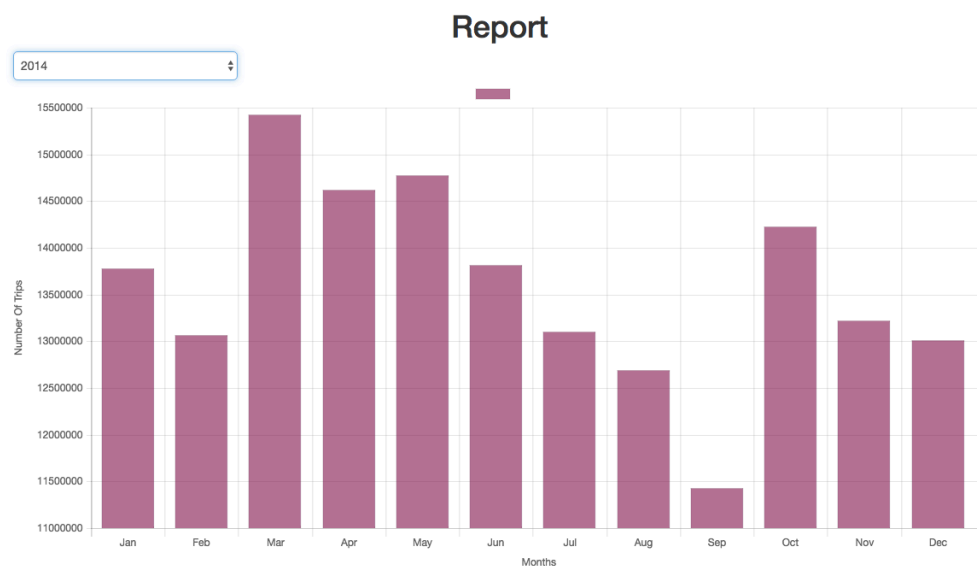
5. AIM: Number of trips distributed over the various months of each year

We have compared the number of trips for each month per year. This analysis can be used to determine months where people use the taxis the most. This in turn is helpful

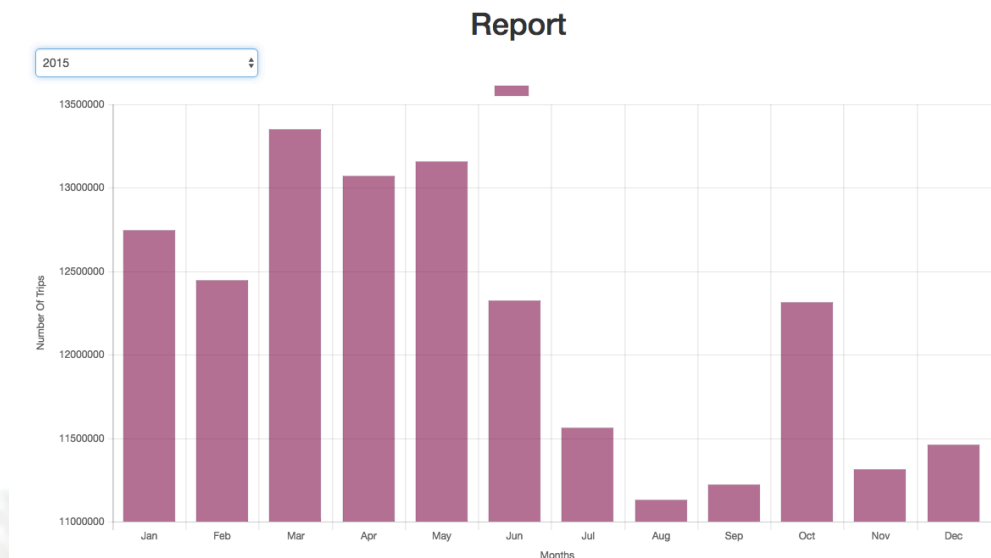
in determining busy months of the city which can be due to various events taking place.



2013 report of number of trips for each month



2014 report of number of trips for each month



2015 report of number of trips for each month

As we can see from the above graphs, the months with high taxi usage are March, May and April.

Conclusion: This could be due to various activities taking place in the city. March is the month when spring commences. New York is one of the main tourist attractions in the world, has pleasant weather during these months and many tourist events are held. This increases the influx of tourists into the city and hence number of trips increase during these months.

Also, as we can see the **number of trips** have **decreased** over the years from **2013 to 2015**.

Conclusion: This can be because of the rise of other cheap modes of transport over the last 3 years. For example, Subway trains, Uber and Lyft taxis which provide tempting offers on rides. Between June-September, the trips are comparatively lesser than other months as it is the summer and fall seasons and people prefer traveling on CitiBikes which may have impacted the taxi rides.

6. AIM: Days of the year with abnormal taxi usage.

We have computed the average number of trips for each year and compared it to the number of trips of each day. Discrepancies have helped us determine important days of the year.

Report

2013

Date	Day	Num Trips
2013-12-25	Christmas	245723.0
2013-10-31	Halloween	489417.0
2013-01-01	New Year	412630.0
2013-11-28	Thanksgiving	331610.0
2013-11-29	Black Friday	373808.0
2013-12-31	New Year	467587.0
2013-02-14	Valentines Day	527476.0
2013-12-24	Christmas Eve	369596.0
2013-07-04	Independence day	326471.0
Average Trips Per day		474465.093151

2013 report of important days and taxi usage

Report

2015

Date	Day	Num Trips
2015-02-14	Valentines Day	501397.0
2015-12-31	New Year	339939.0
2015-12-25	Christmas	188254.0
2015-12-24	Christmas Eve	297167.0
2015-07-04	Independence day	254353.0
2015-10-31	Halloween	441582.0
2015-11-26	Thanksgiving	242393.0
2015-11-27	Black Friday	275078.0
2015-01-01	New Year	382014.0
Average Trips Per day		400309.558904

2015 report of important days and taxi usage

The average number of trips per day for the year 2013 and 2015 are approximately 474,465.09 and 400,309.55.

Halloween and Valentine's Day have clocked **more than average** trips.

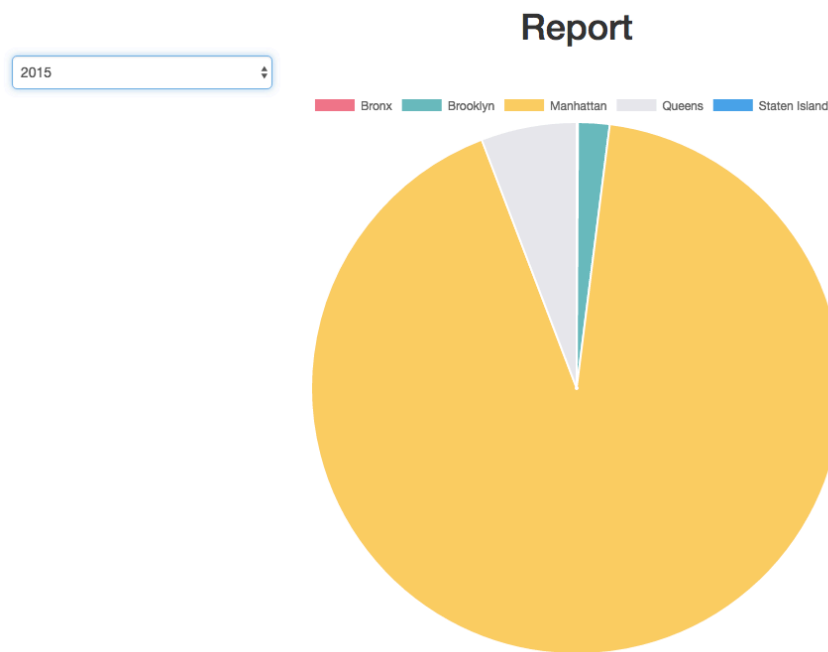
Conclusion: This is because, on Halloween, lot of people go out to party and there is a Halloween parade in the city for which many people come from all over the world to attend. On Valentine's Day, many couples go out on dates to celebrate their love for each other.

New year day, Christmas, Thanksgiving and Independence day have clocked **below than average** trips.

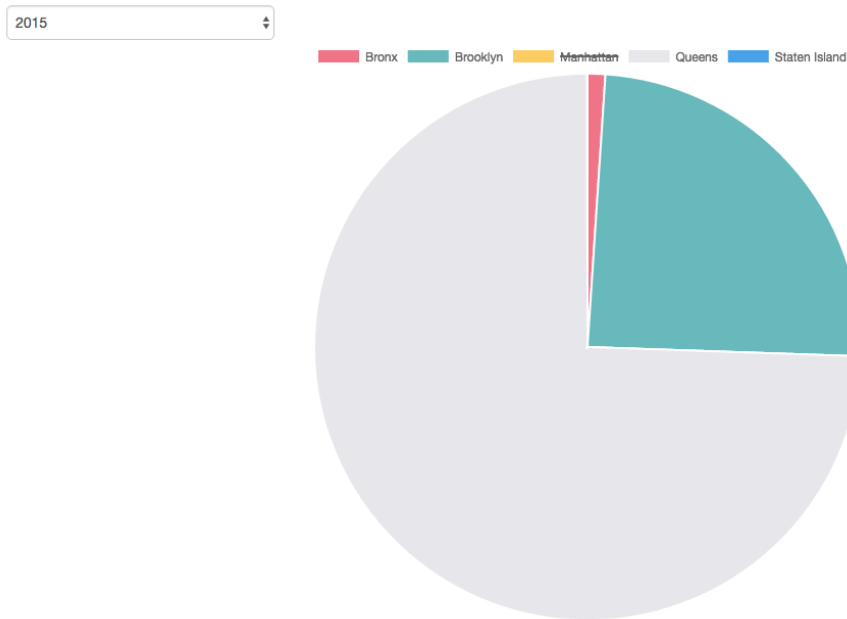
Conclusion: This is because, these festivals bring families and their loved ones together. People generally prefer staying home and spending time with each other.

7. AIM: Boroughs and their taxi usage.

We have compared boroughs based on their taxi usage. This can be useful to determine the popular boroughs and locations of New York. This analysis can be used to make business decisions and know requirement of number of taxis for each borough.



Report



2015 report of taxi usage distributed between Queens, Brooklyn and Bronx

Report



2015 report of taxi usage distributed between Bronx and Staten Island

Manhattan has the **maximum** number of taxi trips followed by **Queens** and the **least** in **Staten island**.

Conclusion: Manhattan is the most visited place by the taxis as it popular locations in terms of restaurants, tourist places, shopping and offices. Queens also has a high

percentage of taxi visits as it has two very busy airports, JFK and La Guardia. Staten Island is not a very commercial area. Thus concluding that there are very less taxi trips to this borough.

8. AIM: Favorite locations visited per year:

We have grouped certain locations as source and destination based on the the number of trips for that pair. This can help people to travel and explore popular places if they are new to the city.

Report

2014

Source	Destination	Num Trips
Upper East Side	Upper East Side	7941550
Midtown	Midtown	6417079
Upper East Side	Midtown	4346675
Upper West Side	Upper West Side	4302908
Midtown	Upper East Side	4261791
Chelsea	Midtown	2475961
Midtown	Chelsea	2285728
Upper West Side	Upper East Side	2263819
Upper East Side	Upper West Side	2256302
Chelsea	Chelsea	2080526

2014 report of top 10 pair of frequent locations

Report

2015

Source	Destination	Num Trips
Upper East Side	Upper East Side	7621876
Midtown	Midtown	5701414
Upper West Side	Upper West Side	4161192
Midtown	Upper East Side	3880630
Upper East Side	Midtown	3854145
Chelsea	Midtown	2222499
Upper West Side	Upper East Side	2074540
Upper East Side	Upper West Side	2072439
Midtown	Chelsea	2059371
Chelsea	Chelsea	1919207

2015 report of top 10 pair of frequent locations

Looking at the above charts, it is certain that the most popular locations visited are Upper East side, Upper West side, Midtown and Chelsea.

Conclusion: This is because, these locations have lots of offices, restaurants, pubs, clubs, etc. which people go to.

9. AIM: Locations capturing the nightlife of New York based on months for each year.

We have analyzed the various locations visited in the night between 8pm-12am and determined the famous top 10 locations visited based on months for each year. This can be useful for people interested in opening up restaurants, bars, etc.

BigData Home Visualize Fare Time Calculator										
2013										
Month/Rank	1	2	3	4	5	6	7	8	9	10
Jan	Midtown,Manhattan	Upper East Side,Manhattan	East Village,Manhattan	Chelsea,Manhattan	West Village,Manhattan	Upper West Side,Manhattan	Hell's Kitchen,Manhattan	Lower East Side,Manhattan	Greenwich Village,Manhattan	Williamsburg,Brooklyn
Feb	Midtown,Manhattan	Upper East Side,Manhattan	East Village,Manhattan	Chelsea,Manhattan	Upper West Side,Manhattan	West Village,Manhattan	Hell's Kitchen,Manhattan	Lower East Side,Manhattan	Greenwich Village,Manhattan	Williamsburg,Brooklyn
Mar	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	West Village,Manhattan	Upper West Side,Manhattan	Hell's Kitchen,Manhattan	Lower East Side,Manhattan	Williamsburg,Brooklyn	Greenwich Village,Manhattan
Apr	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	West Village,Manhattan	Hell's Kitchen,Manhattan	Lower East Side,Manhattan	Williamsburg,Brooklyn	Greenwich Village,Manhattan
May	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	West Village,Manhattan	Hell's Kitchen,Manhattan	Lower East Side,Manhattan	Williamsburg,Brooklyn	Greenwich Village,Manhattan
Jun	Midtown,Manhattan	Chelsea,Manhattan	Upper East Side,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	Hell's Kitchen,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Williamsburg,Brooklyn	Theater District,Manhattan
Jul	Midtown,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper East Side,Manhattan	Hell's Kitchen,Manhattan	West Village,Manhattan	Upper West Side,Manhattan	Lower East Side,Manhattan	Theater District,Manhattan	Williamsburg,Brooklyn
Aug	Midtown,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper East Side,Manhattan	West Village,Manhattan	Hell's Kitchen,Manhattan	Upper West Side,Manhattan	Lower East Side,Manhattan	Williamsburg,Brooklyn	Theater District,Manhattan
Sep	Midtown,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper East Side,Manhattan	West Village,Manhattan	Hell's Kitchen,Manhattan	Upper West Side,Manhattan	Lower East Side,Manhattan	Williamsburg,Brooklyn	Greenwich Village,Manhattan
Oct	Midtown,Manhattan	Chelsea,Manhattan	Upper East Side,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	West Village,Manhattan	Hell's Kitchen,Manhattan	Lower East Side,Manhattan	Williamsburg,Brooklyn	Greenwich Village,Manhattan
Nov	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	West Village,Manhattan	Hell's Kitchen,Manhattan	Lower East Side,Manhattan	Williamsburg,Brooklyn	Greenwich Village,Manhattan
Dec	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	Hell's Kitchen,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Williamsburg,Brooklyn	Greenwich Village,Manhattan

2013 report of famous locations capturing nightlife

2015										
Month/Rank	1	2	3	4	5	6	7	8	9	10
Jan	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	Hell's Kitchen,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Greenwich Village,Manhattan	Kips Bay,Manhattan
Feb	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	Hell's Kitchen,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Greenwich Village,Manhattan	Theater District,Manhattan
Mar	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	Hell's Kitchen,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Theater District,Manhattan	Greenwich Village,Manhattan
Apr	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	Hell's Kitchen,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Williamsburg,Brooklyn	Greenwich Village,Manhattan
May	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	Hell's Kitchen,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Theater District,Manhattan	Williamsburg,Brooklyn
Jun	Midtown,Manhattan	Chelsea,Manhattan	Upper East Side,Manhattan	East Village,Manhattan	Hell's Kitchen,Manhattan	Upper West Side,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Theater District,Manhattan	Greenwich Village,Manhattan
Jul	Midtown,Manhattan	Chelsea,Manhattan	Upper East Side,Manhattan	East Village,Manhattan	Hell's Kitchen,Manhattan	Upper West Side,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Theater District,Manhattan	Greenwich Village,Manhattan
Aug	Midtown,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper East Side,Manhattan	Hell's Kitchen,Manhattan	West Village,Manhattan	Upper West Side,Manhattan	Lower East Side,Manhattan	Theater District,Manhattan	Greenwich Village,Manhattan
Sep	Midtown,Manhattan	Chelsea,Manhattan	Upper East Side,Manhattan	East Village,Manhattan	Hell's Kitchen,Manhattan	Upper West Side,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Theater District,Manhattan	Greenwich Village,Manhattan
Oct	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	Hell's Kitchen,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Theater District,Manhattan	Greenwich Village,Manhattan

Nov	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	Hell's Kitchen,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Theater District,Manhattan	Greenwich Village,Manhattan
Dec	Midtown,Manhattan	Upper East Side,Manhattan	Chelsea,Manhattan	East Village,Manhattan	Upper West Side,Manhattan	Hell's Kitchen,Manhattan	West Village,Manhattan	Lower East Side,Manhattan	Theater District,Manhattan	Greenwich Village,Manhattan

2014 report of famous locations capturing nightlife

The above mentioned places will be the most expensive locations to live in. The rent would be high due to famous restaurants, bars and pubs around it. This would also help business owners planning to open new bars and restaurants.

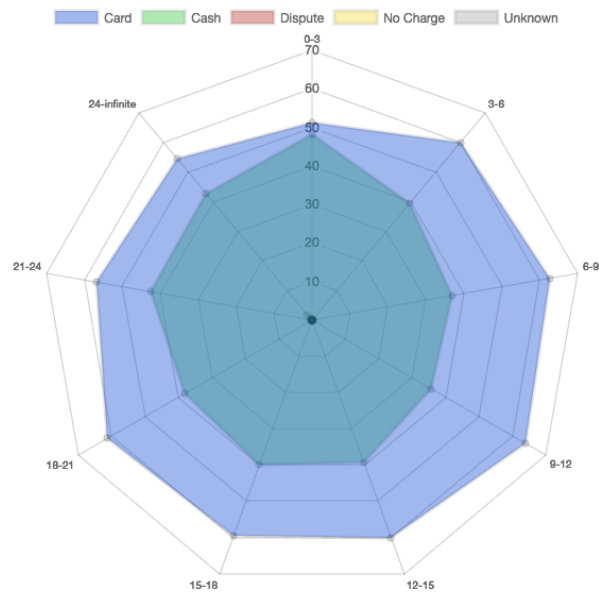
Conclusion: This is because, these locations have lots of offices, restaurants, pubs, clubs, etc. which people go to.

10. AIM: Payment method used for various distance ranges

We have analyzed the type of payment method used for different ranges. This will help determine what payment method is preferred by people.

Report

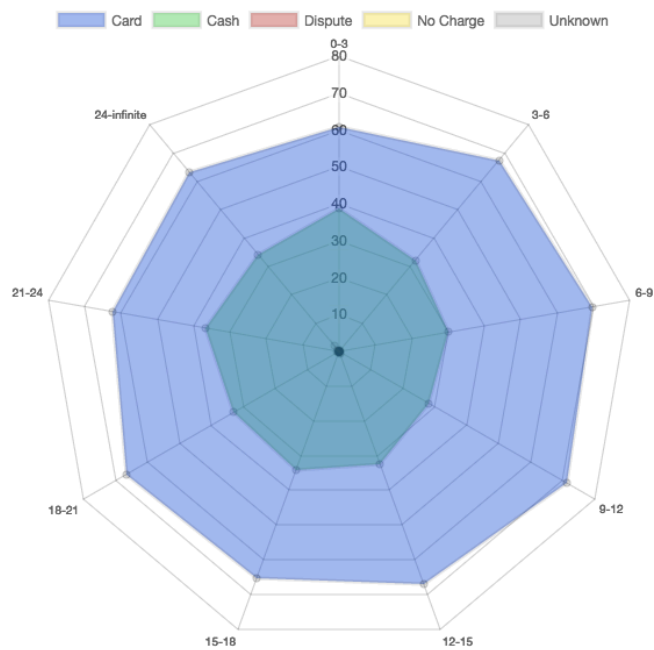
2013



2013 Report for Cash vs. Credit card usage for various distance range

Report

2015

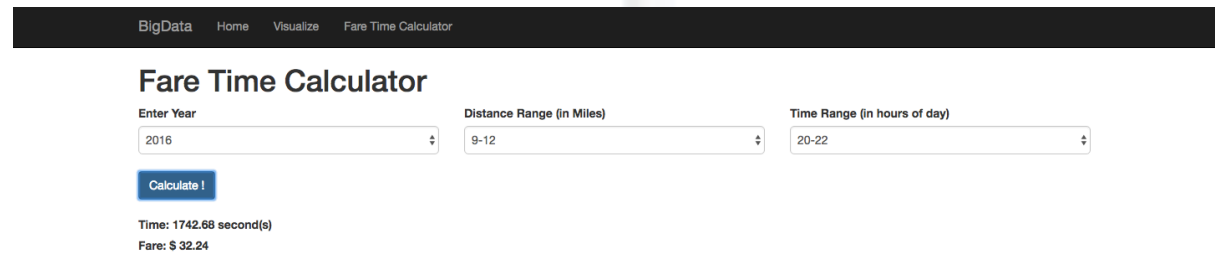


2015 Report for Cash vs. Credit card usage for various distance range

Conclusion: We see that Credit cards are mostly used. Also, the number of credit cards used have increased in 2015 as compared to 2013.

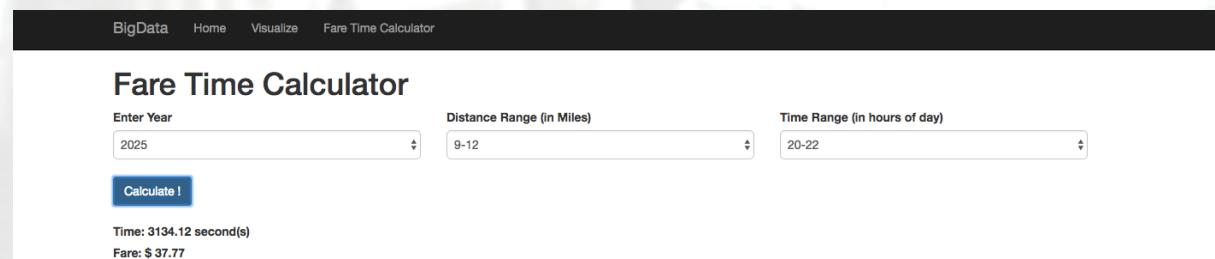
11. AIM: Fare Time Calculator

The Fare Time calculator predicts the fare and trip time using the year, distance range and time as the parameters. We used linear regression analysis to accomplish this.



The screenshot shows a web application titled "Fare Time Calculator" with a navigation bar containing "BigData", "Home", "Visualize", and "Fare Time Calculator". The main form has three dropdown menus: "Enter Year" (set to 2016), "Distance Range (in Miles)" (set to 9-12), and "Time Range (in hours of day)" (set to 20-22). A blue "Calculate !" button is below the inputs. The results displayed are "Time: 1742.68 second(s)" and "Fare: \$ 32.24".

2016 fare and Trip time prediction



The screenshot shows the same "Fare Time Calculator" web application, but with the "Enter Year" dropdown set to 2025. The "Distance Range" and "Time Range" remain 9-12 and 20-22 respectively. The "Calculate !" button is present. The results displayed are "Time: 3134.12 second(s)" and "Fare: \$ 37.77".

2016 fare and trip time prediction

We have given the option of calculating and predicting the future fares and trip time. This will help in determining the budgets for the yellow taxis. This will help in making efficient business decisions for the yellow taxis.

Conclusion: The fares are increasing in the future for a particular distance. The trip time for the given distance range also increases in the future as there would be more traffic in the future.

EXPERIMENTAL SETTING

Cluster Configuration

We used a standard **Amazon AWS m3.xlarge** cluster for running all map reduce tasks. The default configuration was **7 Reducers** that we mostly used, and for some specific tasks we had to use one reducer.

Tools Used

- We used python to write all map reduce tasks, which was executed using Hadoop streaming provided by Amazon AWS.
- For spatial mapping we used rtree index python library.

- For visualization we created a Web Application using Django Framework and ChartJS for visualization.

Detailed information:

1. Python: We used python 2.7. Link -<https://www.python.org/downloads/>
2. Django-Framework: We used django-framework 1.9.6 for web application, and get interactive charts using JS. Link - <https://www.djangoproject.com/>
3. Hadoop Streaming Hadoop provides a UI interface to configure Hadoop streaming
4. ChartJS : For creating interactive and colorful charts on HTML pages. Link- <http://www.chartjs.org/>
5. Python Libraries:
 - a. Shapely: For finding polygons for neighborhoods Link- <https://pypi.python.org/pypi/Shapely>
 - b. Rtree Index: For indexing polygons (neighbourhoods) to quickly query for points Link- <https://pypi.python.org/pypi/Rtree/>
6. WEKA : For regression analysis for Fare Time Calculator. Link - <http://www.cs.waikato.ac.nz/ml/weka/>

Performance

On an average the **running time of map reduce** code was **one hour initially**. For optimization we then **added combiners** which boosted the speed and took **35 minutes on an average**.

One of the major challenge was mapping Latitudes and Longitudes to area names. Initially we tried a **basic polygon code** and it took **3 hours for 1%** of the task. Then we **added Rtree** index on the polygons that really boosted the speed and we got complete **results in 5 hours**. To install rtree we had to ssh into each amazon node and install library dependencies and rtree (the bootstrap functionality of Amazon AWS was failing due to libspatialindex requirement)

Running The Project

The project repository has all the Map Reduce tasks in their respective folders. All the tasks can be run on Amazon AWS by giving input as the folder having yellow taxi data.

We transferred the results to WebApp so that we can visualize the results.

To run the WebApp we have used **standard Django Server**, by running - **python manage.py runserver** it would host the website on port 8000.

INDIVIDUAL CONTRIBUTION

All the members actively participated in the project and showed great spirit as a team. Each member contributed to the ideas for the tasks on the datasets. The various map-reduce tasks were equally distributed among three of the us. The report was prepared by all the members.

Bhagya Lakshmi Gummalla-

- Worked on the dataset of the year 2015 Yellow taxi
- Uploaded data and created bucket
- Wrote the map-reduce python programs:
 - Rate code analysis,

- Month wise trips
 - Traffic Data Analysis
- Ran the outputs for the dataset of 2015
- Worked on the development of the web app.
- Compiled the report
- Prepared the Poster

Ginni Malik-

- Worked on the dataset of the year 2015 Yellow taxi
- Uploaded and created bucket
- Wrote the map-reduce python programs:
 - Credit Cash Time
 - Credit Cash Distance
 - Day Wise Trips
- Ran the outputs for the dataset of 2015
- Worked on the development of the web app
- Compiled the report
- Prepared the poster

Jay Dharmendra Solanki-

- Worked on the dataset of the year 2015 Yellow taxi
- Uploaded and created bucket
- Wrote the map-reduce python programs:
 - Traffic Analysis by hour
 - Night Life Analysis
 - Borough Wise Trips
 - Favorite trips
- Ran the outputs for the dataset of 2015
- Worked on the visualization of all the outputs
- Code Debugging
- Regression program for Fare-time calculation
- Prepared the poster

CONCLUSION

Data analysis can be used to study many interesting facts that can be determined from the data sets. The data on the yellow taxis helped us figure out various correlated things about New York City other than the travelling patterns. For example, the most popular areas of the city in terms of restaurants and bars. The analysis also revealed about the busy hours of the city through the traffic hours. This analysis can be useful in the improvement of business and also the services provided by the yellow cab by allocating enough cabs in a borough or at a popular location.