Temporal and Spatial Analysis of NYC taxi data :A Visualization

Harish Malli Dhanarajan

15008991

National College of Ireland

M.Sc in Data Analytics

Abstract

Nowadays the usage of taxi are becoming more feasible to people not only in case of emergency but also for professional reasons. The development of mobile app to book taxi and 24 hours service have increased its attention towards the people. These type of taxi data that are taken from New York taxi records for the past 3 months and also along with the accidents that has taken place due to taxis. A top-down approach is undergone to accomplish this tasks that includes data cleaning, pre-processing, dimensionality reduction and followed by visualisation of NYC taxi details.This paper proposes the temporal and spatial analysis of NYC taxi details along with the accidents taken place.

Keywords – NYC Taxi, Accident, Visualisation.

1.Introduction

The usage of transport is becoming more nowadays with an increase in requirement to go anywhere at any specific time. Amoung all the transport we consider, taxis are one of the safest mode of road transport. The objective of this paper is the determine certain useful patterns of taxi movement in and around the New York city and in addition to the accidents that taxi has caused in the last 4 months. The research question are given below,

What are the factors that influenced more in taxi movement and its accidents around the city?

The usefulness of finding patterns are because of determining the negative hole in the taxi business if any or doing some minor changes by recording customers behaviour, in order to meet their requirement and so on. The rest of the paper gives a brief summary of usage of taxi and the factors that are most influencing accidents due to traffic. In the following section, section 1.1 comprises the related works, section 1.2 has the domain information and Objectives of the project with its Technical Background, section 1.3 has background of the dataset. Followed by section 1, the method and implementation are included in section 2, then result and conclusion is being included in section 3.

1.1 Literature Review

There are many several works done based on taxi data and the way of visualizing it. Some of such work are introduced in this section.

Jing Yang (2014) proposed a concepts based hidden themes in traffic movements with semantic transformation. This paper includes a visual analytics system to represent large amount of taxi trajectories. These techniques are implemented on huge amount of GPS sampling data. City Mobility pattern, Occupied and vacant taxi, Taxi based on time period that involves tour route analysis and city vs taxi analysis were performed in this proposal [2].

Nivian Ferreira [1] proposed a paper based on visual exploration of big-spatio temporal urban data which is a study of new York city taxi trips. Various kinds of visualization are made based on taxi data from the New York taxi department. The visuals were based location, time, date and so on.

Liao Binbin [3] proposed a paper based on GPS data which has undergone visual analytics followed by an anamoly detection. Similar to all work, this paper has also proposed about the visualization of GPS data which is a taxi data. Taxi data are computed using the GPS by tracking the movement of taxi in and around the city.

1.2 Domain Information

This section includes the domain information and the objective behind choosing this domain. At first, the taxi data are collected in such a way that, a device named GPS are installed in all the taxis to track the movement among them. New York city has three types of taxis running around namely yellow taxi , green taxi , uber taxi. Out of all these taxis, only green and yellow taxi are chosen for analysis. Coming back the point of domain information, we could say that the details that are noted down by the GPS are Pickup and drop Location (longitude, latitude), Time, date, number of passenger (entered by driver) , trip distance and so on. To add on, we could say that the yellow taxi are otherwise said to be medallion taxi and green taxi are said to be street hail vehicles. These taxis were brought into action by the year 1890 and still in its success path. All range of cars namely Nissan, dodge, volkswagon are used by green and yellow taxis. There around 13,437 taxi running in and around New York city. It typically travels around 700,000 miles throughout a year. Additional features such as hybrid-electric vehicles (60%), and wheelchair accessible (2%) are made possible. In addition, New York taxi has their own mobile application which is more obviously used for taxi booking.

Another side of investigation is the accidents that were caused by taxis around New York city. It is reported that, New York taxis are the only taxi to undergo a crash test before it is brought on road. But, the crash test is only for the taxis and not for the people around!!. In total amount of injury caused last year due to taxi is 19.37 % out of all the injuries (by other vehicles) and accidents that has happened last year (2015). Hence, an analysis of this is done in the further section.

1.2.**1Objective of the project**

The objective of this project is to determine patterns in various taxi movements with the help of data captured with GPS as mentioned above in the domain information. The patterns are as follows,

a. How many trips are made in a single day?
b. What Type trips are being made?
c. What Time does make a better profit?
d. When is the trip made?
e. Additionally, a visualization is done based on acccidents due to taxi based on the motive for reducing the accident in and around New York City and so on.

The usefulness of this patterns is to determine the customer behavior and bringing up in a relationship with the customer by providing better services. In case of the objective that lies behind accident data, we could say that the factors causing those accidents are well examined and drivers can be trained in such a way that they don't cause any such accident in future**. Note: Not only the above patterns, there are still more number of patterns which is being used in visualization in further section.**

1.3 **Background of Dataset.**

This section comprises the background of the data, variables used , number of rows and its relevant information. **The information explained in this sections are those before data cleaning and pre-processing.** The source of the dataset are https://nycopendata.socrata.com/ (New York Taxi) and https://www.data.gov/ (New York Taxi Accidents).

Structured data

- New York Taxi – It consists of 14 variables with 5,00,000 + rows. 5 variables say about the fare of trip, 6 about pickup and drop details, rest of the variables are about the payment mode, type of payment and other details such as pickup and drop down details. It had around 576 null.

- Accidents dataset consist of 6,50,000+ rows with 8 variables that includes factor that has caused the accident, type of vehicles that has caused the accidents, location of accident, time of the accident.

Unstructured data

The unstructed data is based on tweets which is extracted from twitter depending upon the tweets regarding taxi and its accidents.

Followed by this section, methods and implementation are explained in the next section that comprises design of the project, software, technology used, implementation related details.

2. Methodology and Implementation

This section has architecture of the project, software used, technology applied and implementation.

2.1 Architecture of the model

The architecture of the model is top-down approach as follows as given in figure 1.

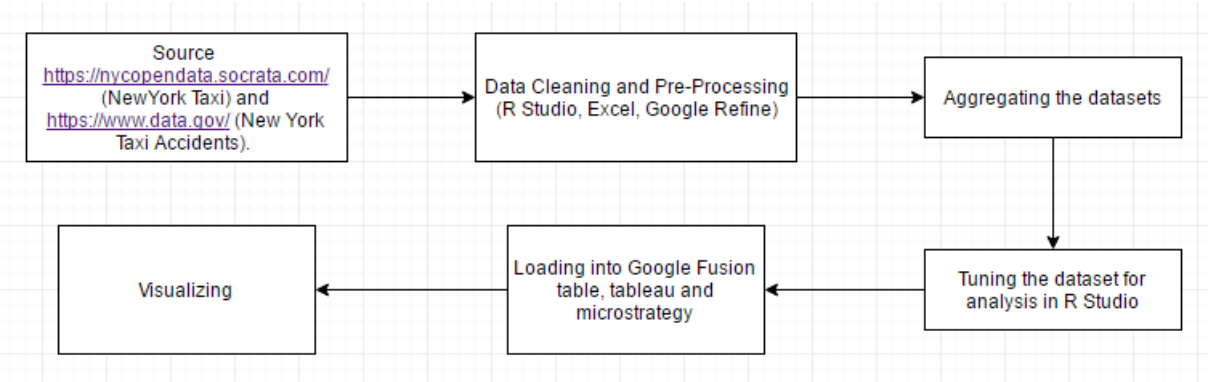

Figure 1. Architecture of the model

At first, the data is been taken from the specified website and they are cleaned according in various platform and again filtered using google refine. In total there are 7 dataset which are to be combined into a single aggregated dataset and this task is performed in R studio and partially in excel using a tool named ablebits data add-on. The aggregated dataset are again fine tuned by removing still more variables and rows and then made suitable for visualization purpose. After we get a fine tuned data, we import them into suitable visualization tool for finding patterns in them. Apart from this, sentimental ananlysis is performed to examine peoples opinion in a form of tweets.

2.2 Software and Technology

-Softwares used for this models are R studio, tableau, excel, microstrategy and online tools such as google refine and google fusion table are used.

-The package used in R studio for data reduction is pca and caret package in order to determine the most influential variables. Followed by this, an add-on tool were installed in micro-strategy for showing annotation tables.

2.3 Feature Engineering

Feature engineering is the process of building a model based on the domain of the data. Since, the data is a huge it is necessary to undergo data cleaning and pre-processing in order to get rid of noises in the dataset. In this section. the above architecture in section 2.1 is briefly explained. In data cleaning process, certain null values and NA values has to be removed from each and every variable by uploading all the seven datasets in R. Then, values were removed though certain messy special characters remained in the dataset. To remove these noises, google refine and excel were made use.

Once the cleaning process was over, the dataset has been analysed using a dimensionality reduction technique. The Dimensionality reduction technique used in this case is principal component analysis. Before performing the analysis, the dataset was converted into numeric data and in addition, the date, time and location columns were removed. The result was so obvious that all the variable in the data were given same set of importance excluding MTA tax, tips amount variable. **PCA was implemented only for taxi dataset and not for accident dataset.** After the data cleaning , an additional variable was added named as day of week which R studio has in-built for converting day from date and they are computed by using the formula below:

yellow_taxi1$day_of_week <- weekdays(as.Date(yellow_taxi1$Date.of.Trip))

The data cleaning is over, now the dataset is aggretated (6+1) into a single data**. Join function in R was performed i.e, Inner Join.** Furthermore, ablebits tool was used in excel to combine accident and aggregated taxi data into a single dataset. These dataset again underwent, cross checking of any noise inside the data. Now the dataset is ready for visualization. Software for visualization are tableau,

2.4 Implementation of Visualization

Before implementing the visualization, a simple pivot table visual that covers certain aspects of datasets are explained at first in figure 2 and 3.

In figure 2, The first set of data explains the number of trips made by green taxi and yellow taxi based on timing in the past 4 months. The second set of data explains about the number of trips made in weekdays and weekends and the third set of data explains the vendor details. The connectivity error in the third set of data is nothing but the connection lost in GPS that has occurred in the trip journey. It is made clear that the creative mobile technology vendor has made more number of connectivity error compared to other vendor.

Opinion – In the first set of data we can say that people have made enormously huge amount of trips in “On-Peak Morning Trips and in busy evenings”. It is obvious that people take taxi during peak periods and hence more number of taxi has to be active during peak time than the other timing. In second set of data, 1/3 of the trip in a whole week is done only on Saturdays and Sundays and hence to conclude we can say that taxi availability has to be made more in number even during weekends.

Timing	From	To	Green Taxi	Yellow Taxi
Midnight Trips	00:00:00	05:00:00	78685	66324
Early Morning Trips	05:00:00	08:00:00	46019	37888
On-Peak Morning Trips	08:00:00	11:00:00	69023	63075
Off-Peak Afternoon Trips	11:00:00	15:00:00	92376	96482
On-Peak Evening Trips	15:00:00	20:00:00	142190	141138
(Late) Night Trips	20:00:00	00:00:00	61434	76530

Day Of Week	Green Taxi	Yellow Taxi
Monday	78346	77622
Tuesday	90940	71503
Wednesday	57019	76267
Thursday	90901	64015
Friday	47667	61032
Saturday	72907	66629
Sunday	44570	64369

Driver Login	Green Taxi	Yellow Taxi	Connectivity Error (%)
Creative Mobile Technologie	106914	229309	5.31
VeriFone Inc.	375436	252129	0

Fig 2. Pivot Analysis of dataset

In the below figure the comparison is made between the sum of amount collected in both green and yellow taxi in the past three month. Figure 3 says that the sum of fare amount collected during evening peak time that in normal time.

Opinion – Though the number of taxi trips are more in both morning and evening peak times, it is more likely that the fare amount is collected more only during evening peak times. This might be because of people returning tired to home which makes them avoid other public transport and prefer taxi or it might be because of the fact that only small trip distance is made in morning peak time which in turn reduces the sum of amount but the number of trips still remains the higher. And in another way of comparison, we can conclude that yellow taxi has made a profit of over $ 5,00,000.

Timing	Sum of Total_amount	Descriptive	Timing	Sum of Total_amount
00:00:00	287891.12	Midnight-Early Morning	00:00:00	400478.35
01:00:00	374585.6		01:00:00	338809.15
02:00:00	294732.07		02:00:00	243851.07
03:00:00	231066.07		03:00:00	160709.27
04:00:00	207601.92		04:00:00	148079.82
05:00:00	164341.89		05:00:00	129430
06:00:00	153413.1	Start of the Day	06:00:00	167517.15
07:00:00	247526.71		07:00:00	260342.7
08:00:00	428041.35		08:00:00	399498.1
09:00:00	510962.69	On-Peak Morning	09:00:00	405511.27
10:00:00	463883.31		10:00:00	423892.3
11:00:00	414957.1		11:00:00	399338.87
12:00:00	407116.47	Off-Peak Afternoon	12:00:00	476037.63
13:00:00	406498.06		13:00:00	435660
14:00:00	444281.39		14:00:00	504716.01
15:00:00	512505.17	Busy Evening	15:00:00	536873.04
16:00:00	553490.46		16:00:00	526368.1
17:00:00	558185.39		17:00:00	491480.11
18:00:00	545869.16		18:00:00	631608.31
19:00:00	528167.44		19:00:00	619969.1
20:00:00	464338.56		20:00:00	573580.89
21:00:00	400564.73	Night Trips	21:00:00	500309.39
22:00:00	384932.18		22:00:00	578930.85
23:00:00	352422.9		23:00:00	498537.18
Grand Total	9337374.84	End of the day	Grand Total	9851528.66

Fig 3. Pivot Analysis of Fare Amount

The above two pivot analysis are implemented in visualization using tableau as represented in figure 4. Green Taxi stands top on Saturday, Tuesday and Thursday.

Opinion – Might be because of offers in green taxi once in two days and similary the day inbetween is filled up by yellow taxi.

Fig 4. Trips vs Day of Week and Timing vs Taxi

Comparison Yellow (Above) and Green (Below) taxi based on passenger count on day of week splits.

While looking at a prespective of timings, it is likely that the passenger count increases from 15:00:00 till 20:00:00 as represented in figure 5.

Opinion – Yes, obviously the passenger count is more during evening time but the main thing to be noticed is that, the trips are made more during weekends than weekdays during evening. This is because of people returning to city to enjoy the weekend or people who are going to railway station, airports in the evening time during weekend to spend their weekend outside the city.

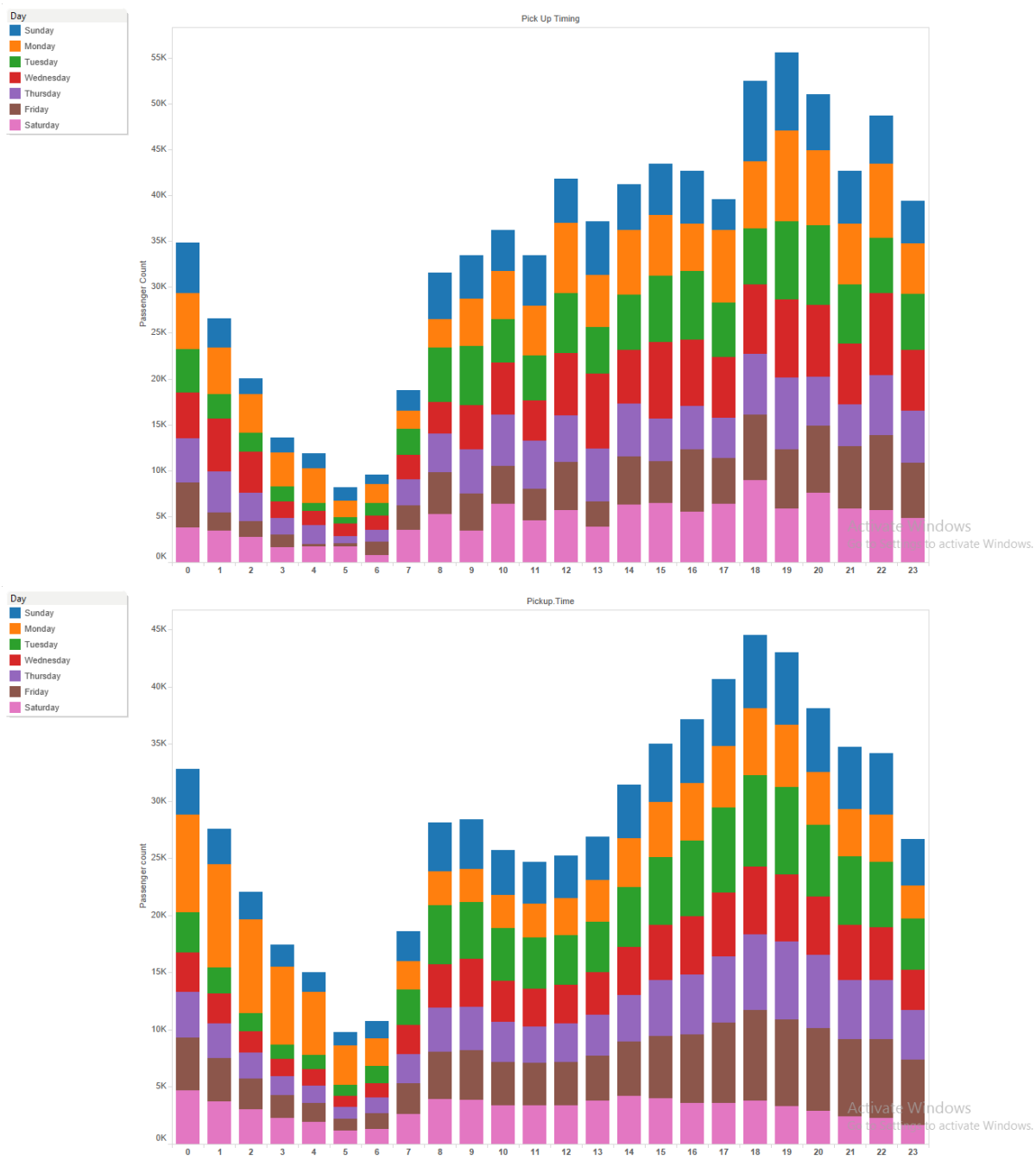

Fig 5. Comparison Yellow (Above) and Green (Below) taxi based on passenger count on day of week splits.

Health of Timing

The below graph represents the health of timing and day of week. i.e, Thursday and Tuesday has higher grid size which in turn represents that travel is made more. (Figure 6)

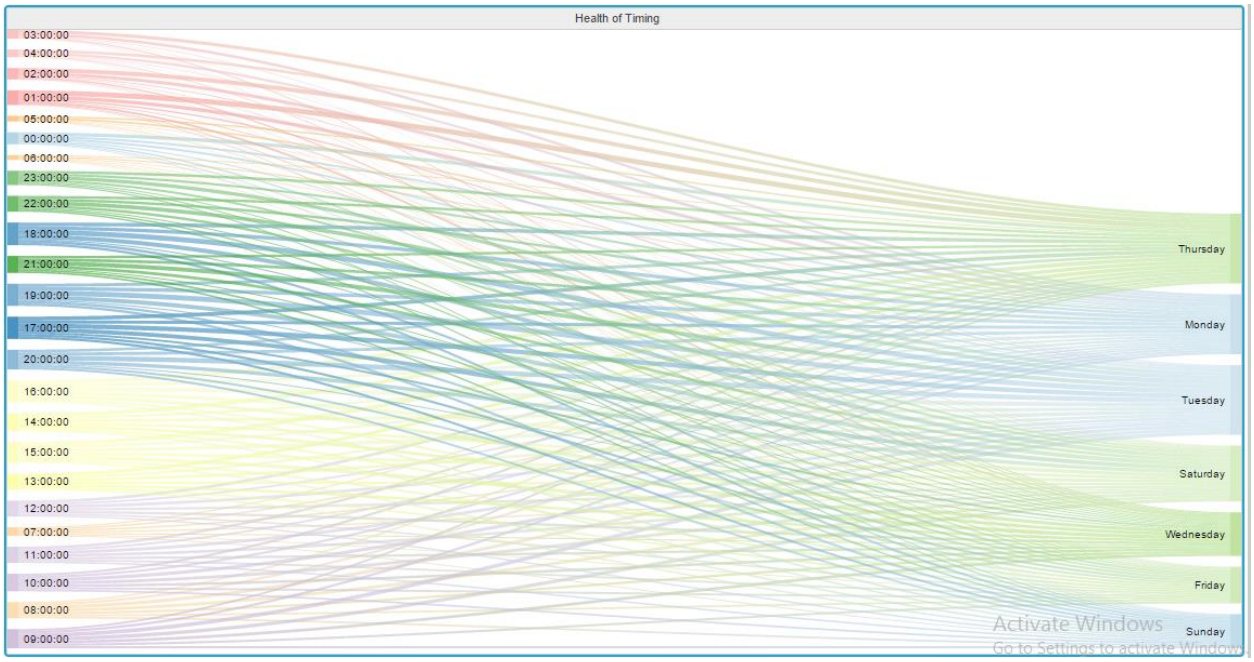

Fig 6. Health of Timing (Both Green and Yellow From aggregated dataset)

Fare Amount vs Day of Week

The sum of fare amount against that of day of week is plotted as below in figure 7. The motive behind this graph is to determine the fare amount collected on all the seven days of the week.

Opinion – Yes again! The fare amount collection is more on Friday, Saturday and Sunday. i.e, according to the graph, the fare amount collected on Saturday from 15,000 $ and 87,000 $. The least amount collected on 6am in the morning and the intermediate amount is collected in 11am,11pm and 7pm and more fare amount is collected at 6pm in the evening.

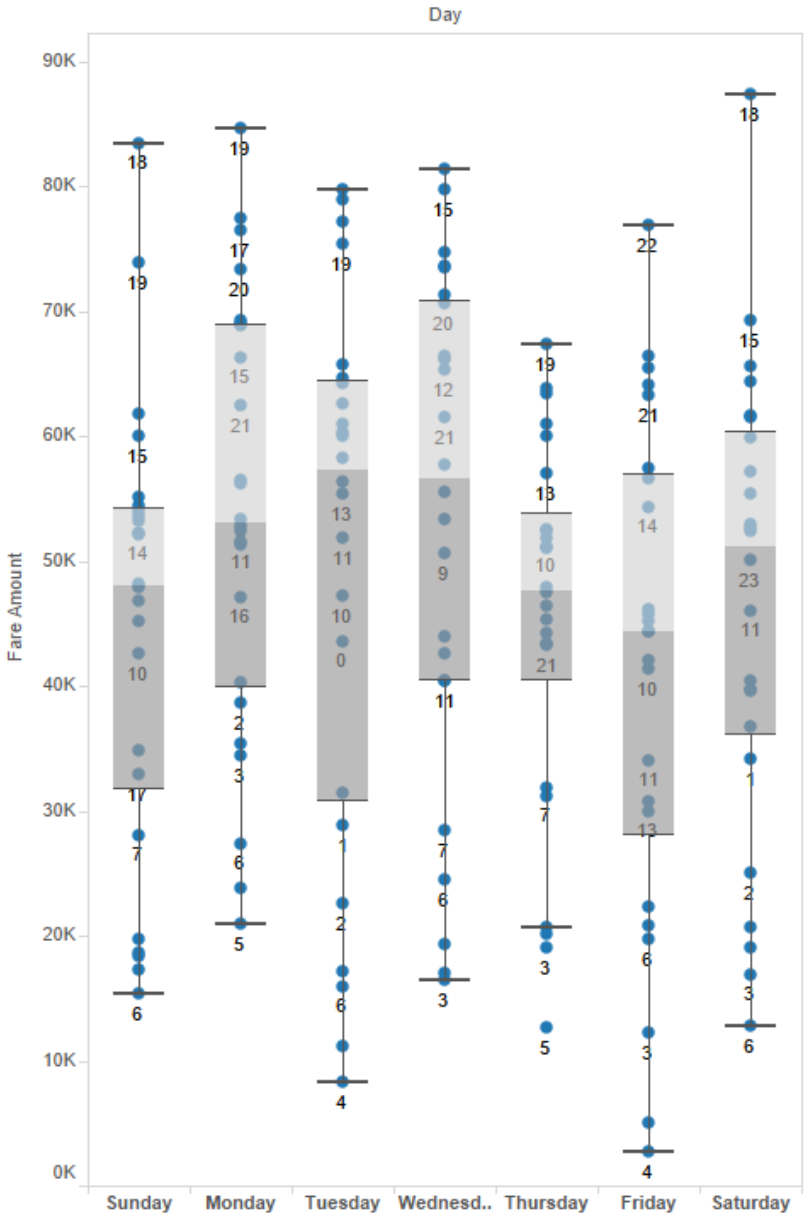

Fig 7. Fare Amount vs Day of Week

Rate Codes

The rate code of the trip is based on trips To & From LaGuardia Airport (Std. City Rate), To/From JFK and any location in Manhattan (JFK + Toll Rate), From JFK to other New York City destinations (Std. City Rate) , To Newark Airport (Newark Rate), To & From Westchester & Nassau (Std. Charge + Double Charge + MTA tax+Tolls). This is explained in figure 8.

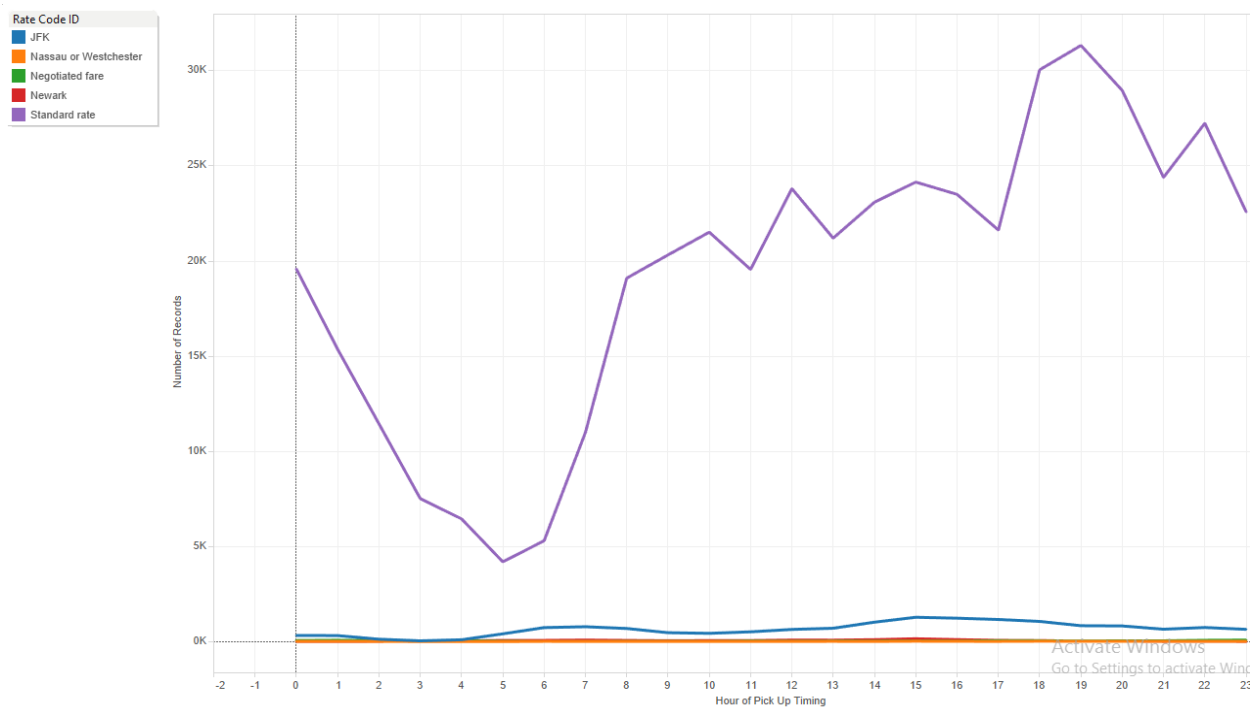

Figure 8. Rate Codes

Opinion – The standard rate is higher during midnight and increases along the day from 9am. It indicates that the trips are made mostly from To & From LaGuardia Airport and From JFK to other New York City destinations.

Extra Amount calculated based on pickup time

The below graph in figure 9 represents the times when was the extra amount collected for green taxi and yellow taxi. The extra amount vs time graph in both yellow and green taxi remains same. This is might be a company policy to get extra amount in particular period of timing.

The extra amount during 00:00:00 was 4.67% from the original fare and that goes on decrease all the way down and again touches the peak during 18:00:00 to 6.23% from the original fare. ***This percentage is calculated using the formula by SUM([Extra]) / TOTAL(SUM([Extra])) in rows and DATEPART('hour', [Pick Up Timing]) formula in columns part.***

Opinion – Might be a company policy to keep up extra fee during 18:00:00 to 00:00:00.

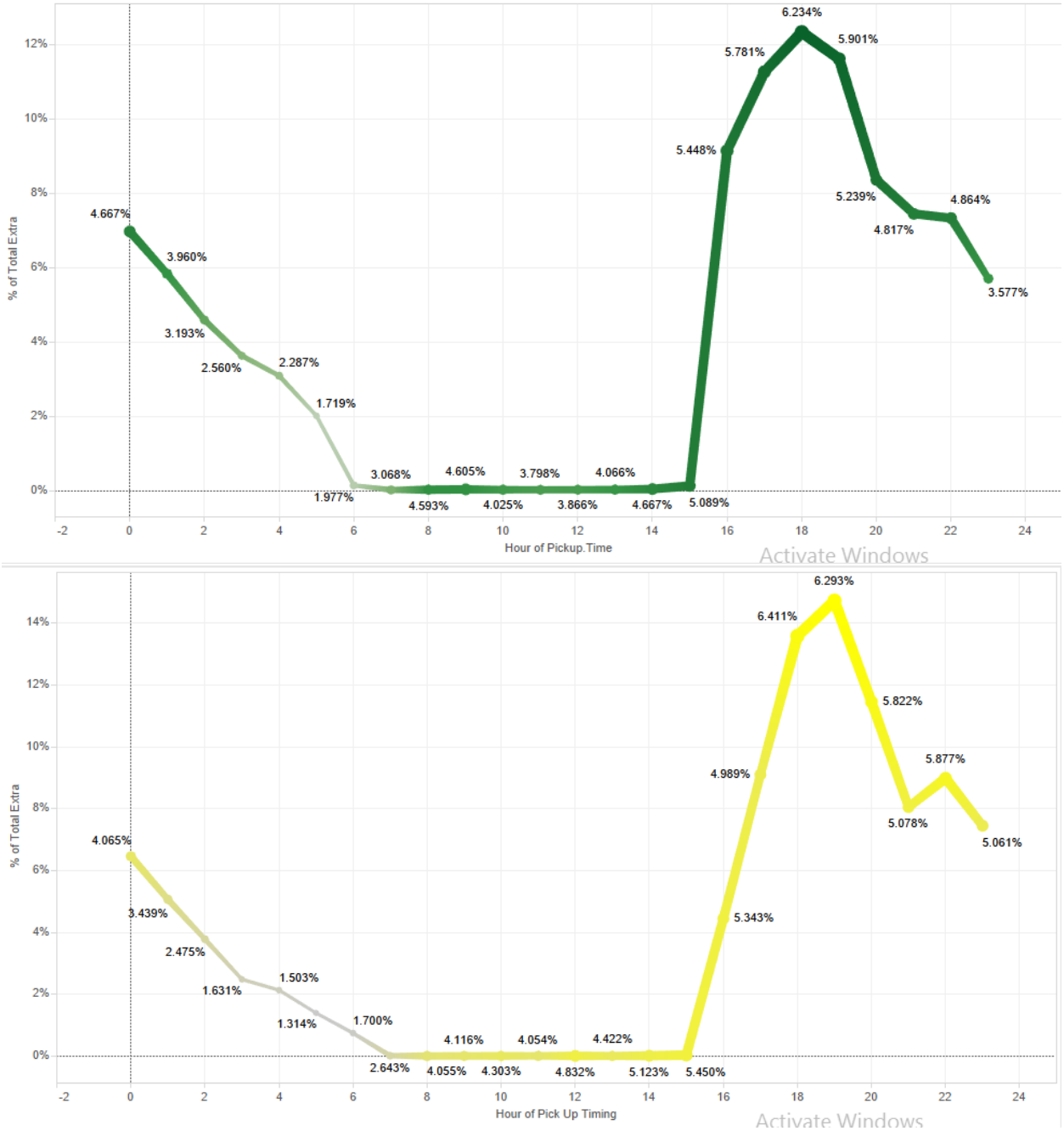

Fig 9. Extra Amount calculated based on pickup time

. ***Location of Pickups and Drops***

So far the visuals were based on day, time, passenger and trip count. Now, the below graph represents the latitude and longitude of pickup location in figure 10 for green trip and figure 11 for yellow taxi. These visuals states about the location of pickup and drops made.

Opinion – The Green taxi covers most of the main area in New York as well as outer area (Fig.10 Larger Oval) of New York in exceptional cases with an amazing service provided. In

case of yellow taxi, it covers only From / To Lower Manhattan and Upper Manhattan and less density of coverage in airport regions (Fig.11 Smaller circle). Comparitively, Green taxi has made a better coverage of area irrespective of other factors like date, time and day.

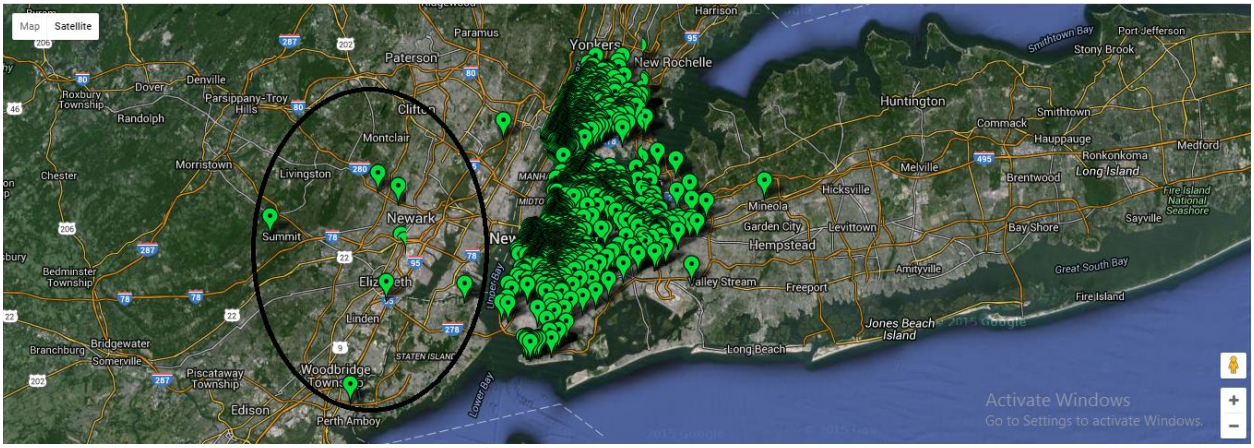

Fig 10. Location of Pickups and Drops (Green Taxi)

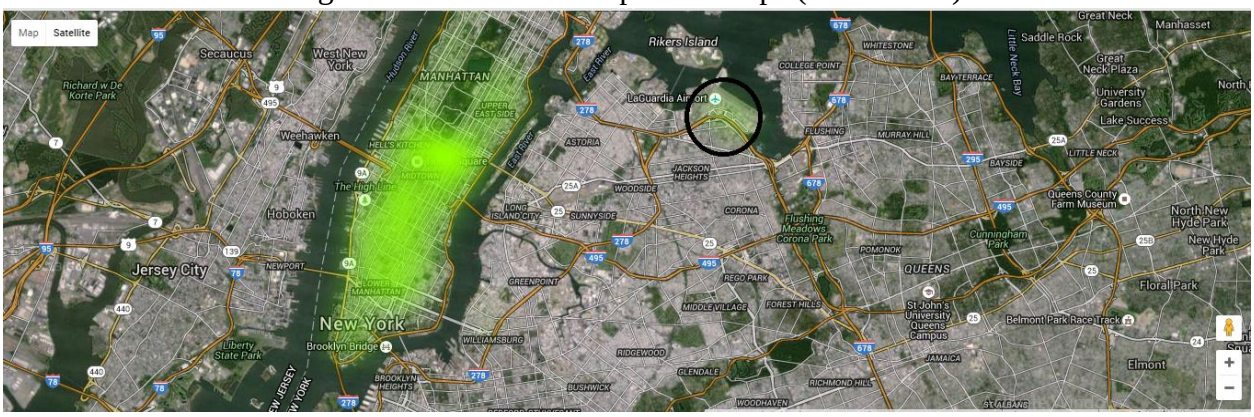

Fig 11. Location of Pickups and Drops (Yellow Taxi)

Overall Conclusion (Fig 12) represents that green taxi has performed well in midnight and early morning whereas yellow taxi has covered in other part of the day.

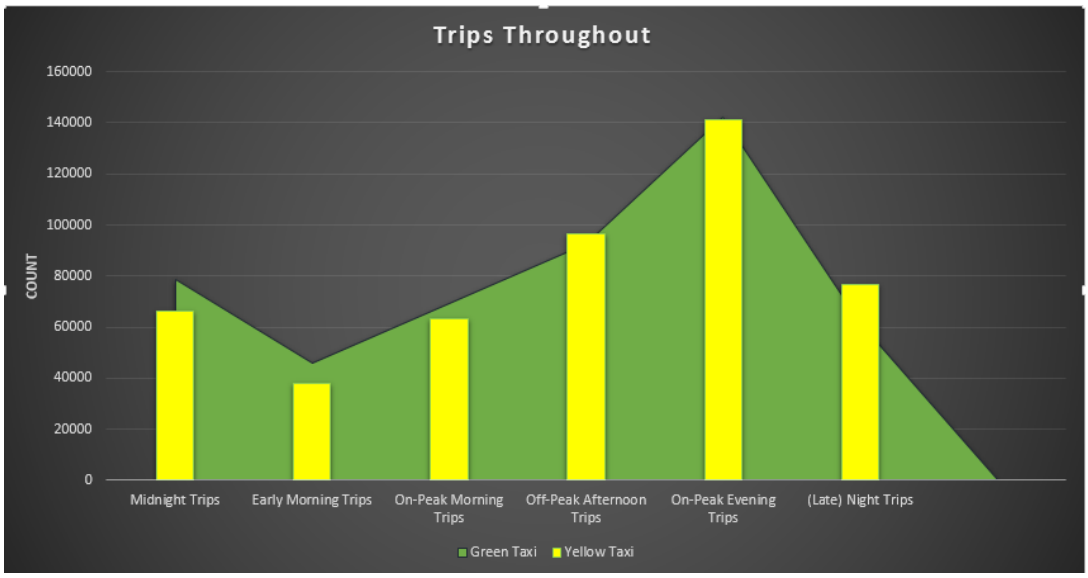

Fig 12. Trips Throughout

Taxi Accidents

Till now, a visualization of taxi movement over the New York City was examined, now the accidents due to those taxis in the past has been examined. The first visualization is based on maximum number of accidents by location in figure 13.

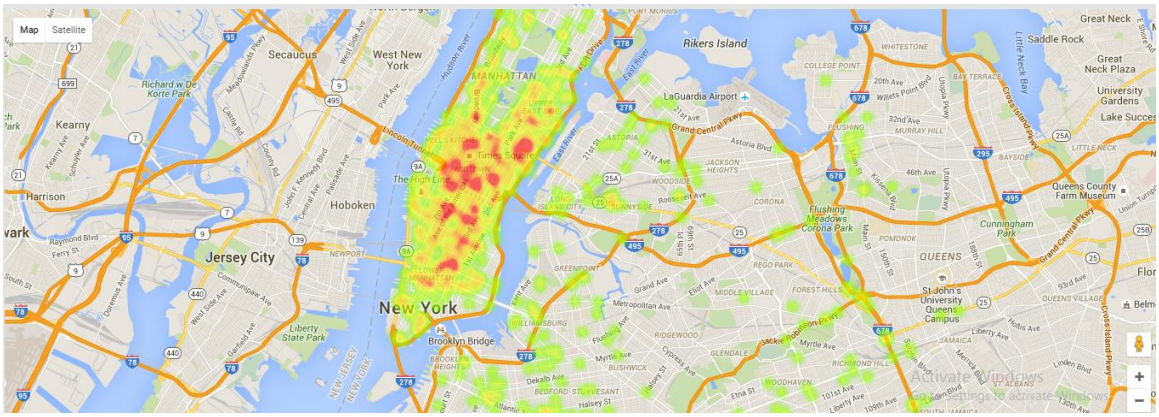

Fig 13. Accident by taxi.

Opinion – It is likely that more number of accidents are happening on middle of the New York city. This is due to traffic in the city or improper safety and so on. To rectify this, more number pre-cautions have to be taken, penality for rash driving has to be implemented and brought into action.

Persons Affected by Accidents VS Timing

The graph in figure 14 represents the number of injuries and number of persons affected due to it. The accidents has been happened during the peak time from 16:00:00 to 21:00:00, henceforth to conclude we can say that injury and person died be only because of traffic.

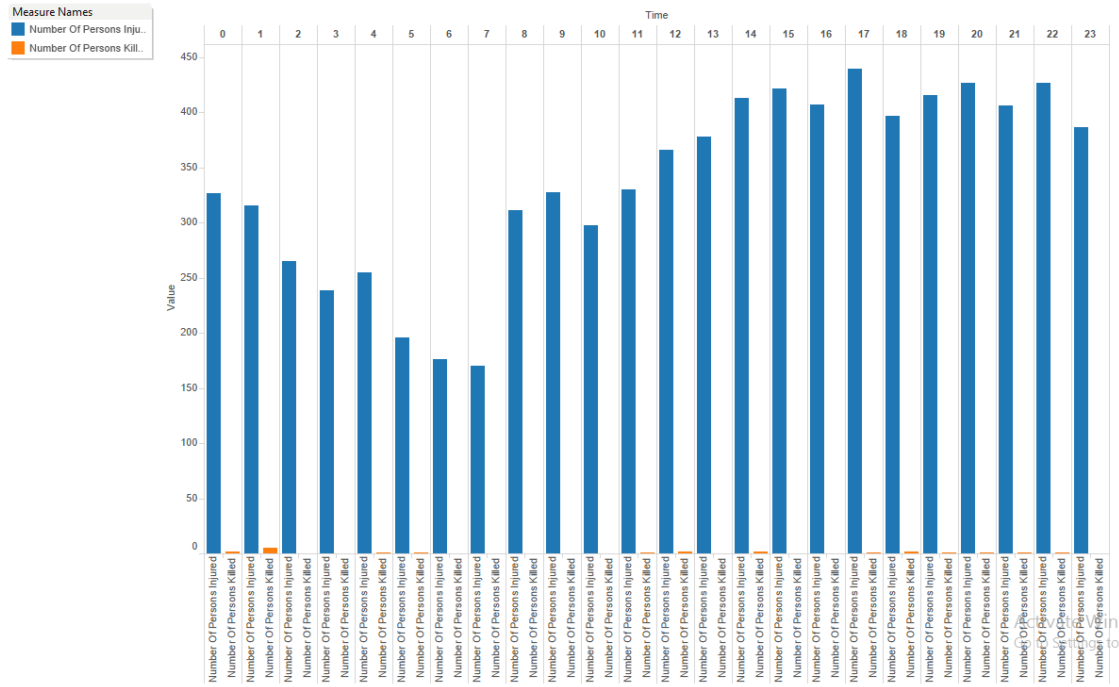

Fig. 14 Persons Affected by Accidents VS Timing

The next graph says about the factors that influence the accidents in figure 15 and figure 16 also represents the same set of information but in a level wise and counts. The top opinion about the graph is that taxi accidents are caused because of “Other Factors” ,”Fatigue/Drowsy”, “Failure to keep right”, “Cell Phones / Handsfree” , “Aggressive Driving” and “Lost Consiousness”.

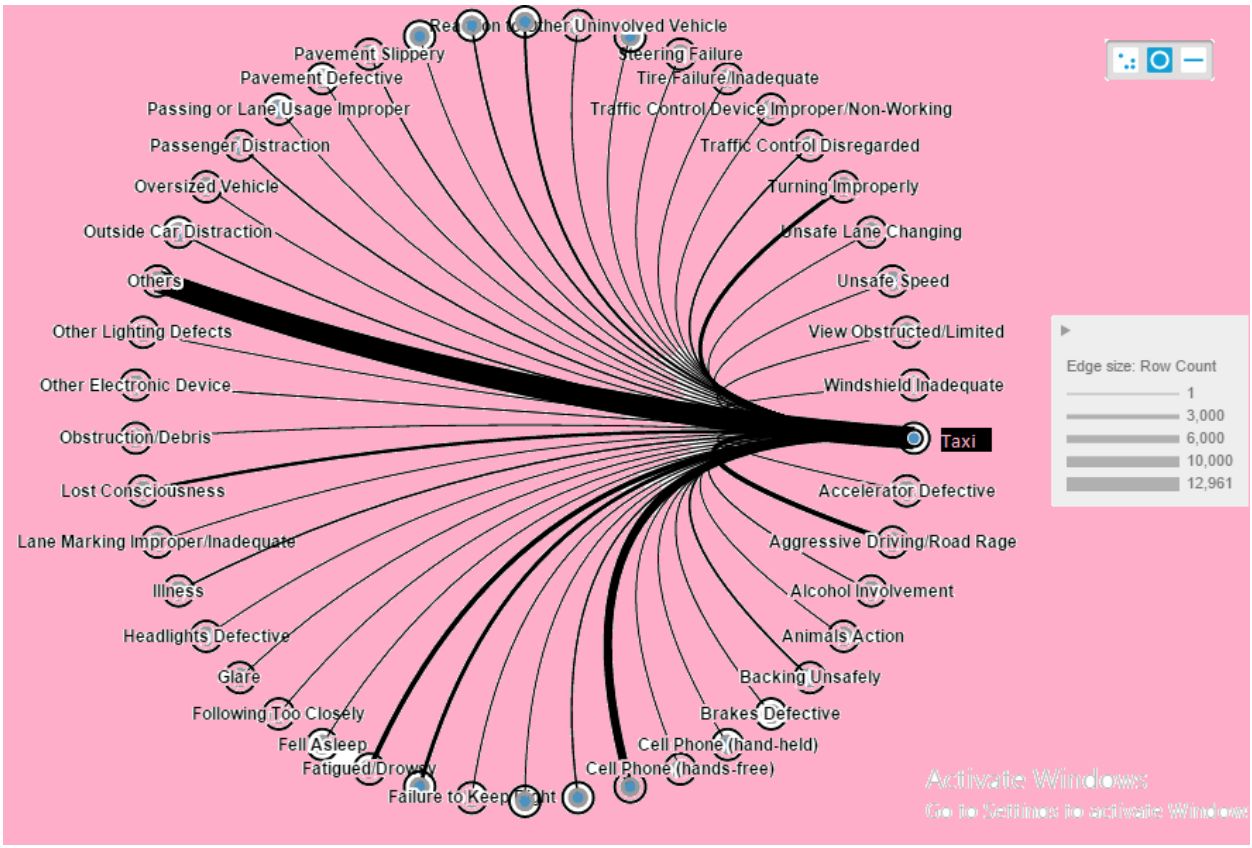

Fig. 15 Factors influencing accidents by taxi

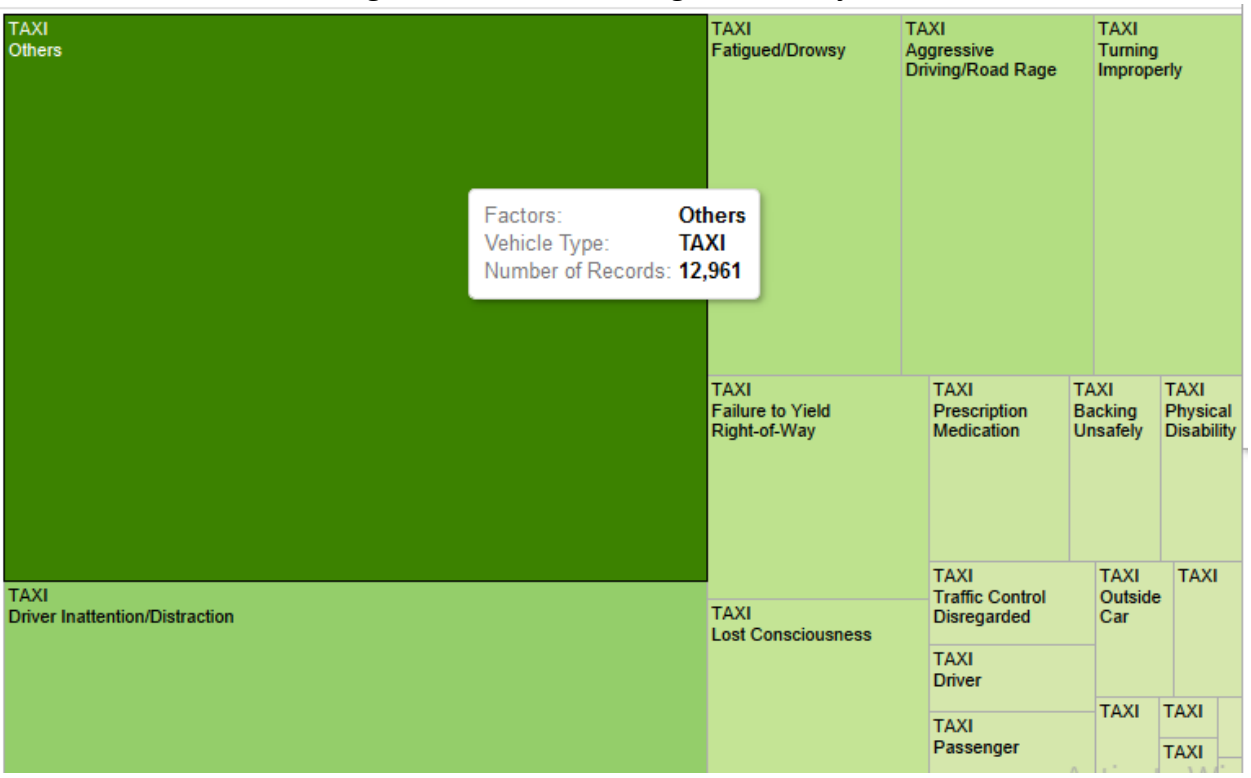

Fig. 16 Factors influencing accidents by taxi

The top five factors are taken in charge and percentage of person affected in determined in figure 17. This is computed by using the calculated field as below,

% of person affected = SUM([Number Of Persons Injured]) / TOTAL(SUM([Number Of Persons Injured]))

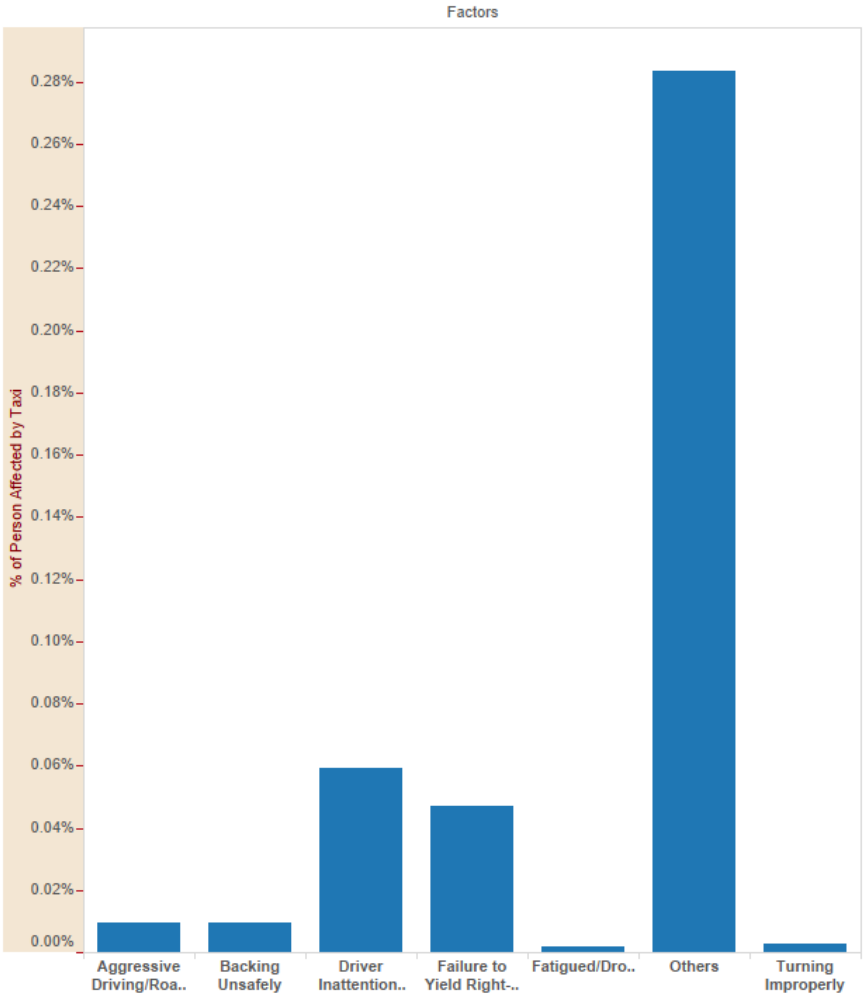

Fig 17 . % of person affected.

Explanation of above graph – If a person is injured then the probability of getting injured by other factor is 0.25%, aggressive driving would have caused 0.01 % of person affected and 0.06 % of driver in attention.

The following section consists of sentimental analysis of tweets.

Sentimental Analysis

Apart from the analysis done above, peoples opinions has also has to be considered. The only open source is the social media and the most feasible was the tweets from twitter based on NYC taxi and the accidents caused. The tweets were imported into R studio by hooking up R

with twitter and the basic packages such as geoffjentry/twitter, devtools and curl. A set of positive and negative words were first uploaded into twitter from PC and they were compared to the tweets and a sentimental analysis was performed using those tweets.

Opinion – People have neither tweeted negative nor positive i.e, Only 1/3rd of the tweets were positive and negative. Most of the tweets are neutral. These tweets are categorized into subjective and objective type. So, to conclude we can say that taxi tweets are more neutral in nature.

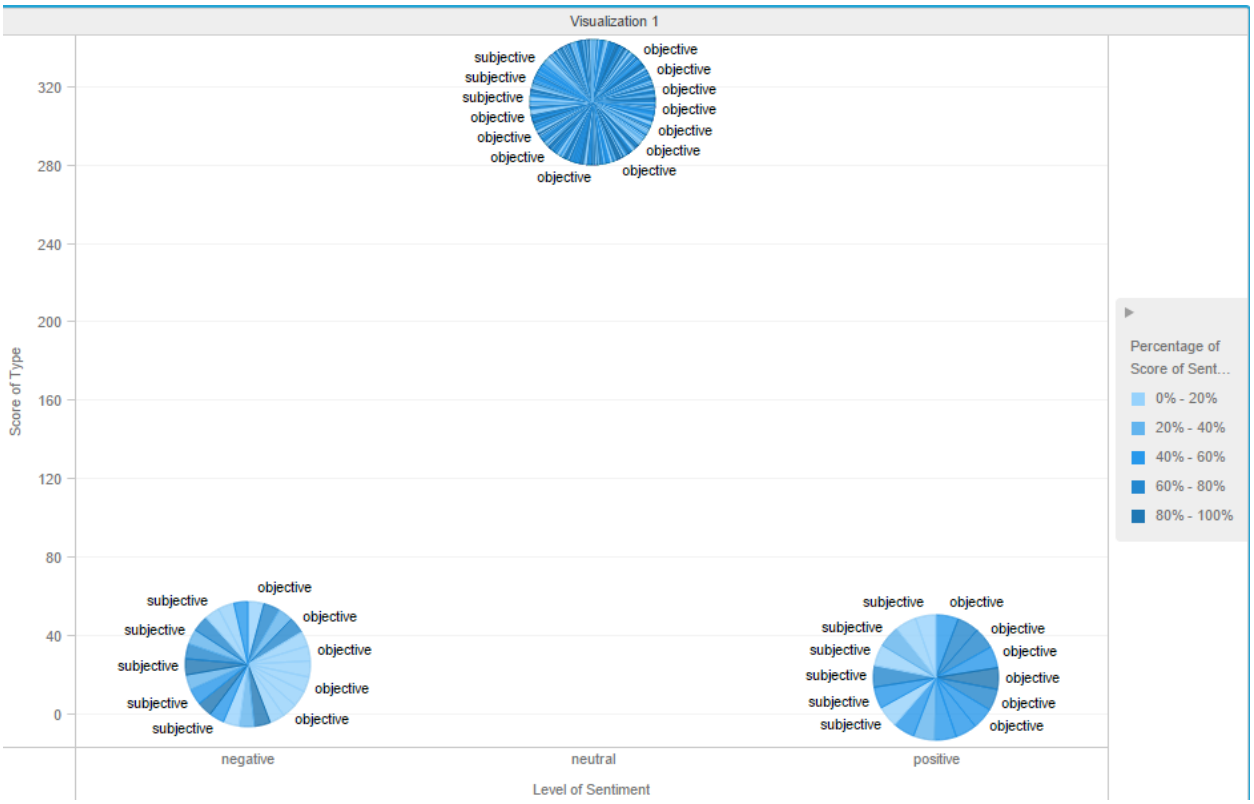

Fig 18. Sentimental Analysis

3.Conclusion and future work

To conclude, we could have certain factors in control which we have analysed so far in visualization from this paper. In this paper several techniques such as dimensionality reduction, data cleaning, data preprocessing and visualization of spatial and temporal taxi data was proposed. In future,the factors that can be considered in future is that,

1. Increase of taxi in weekends.
2. Increase of taxi during peak periods.
3. Precautions during peak time to avoid accidents.
4. Penality for driver even if the commit a small things that leads to accidents which inturn may reduce the accidents in NYC.

5. Increase of taxi in outer city from NYC especially for yellow taxis.
6. Driving rash at cities has to be fined and penalized inorder to reduce accidents caused by taxi drivers.
7. Others factors based on the above visualization can also be made.

References

1. Nivan Ferreira, Jorge Poco, Huy T. Vo, Juliana Freire, and Cl´audio T. Silva (2013), *Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips,* NYU CUSP.
2. Ding, C., Jing, Y. and Zheng, M. (2014), *Visualizing Hidden Themes of Taxi Movement with Semantic Transformation,* IEEE Journal

3. Laio, B (2010) , *ANOMALY DETECTION IN GPS DATA BASED ON VISUAL ANALYTICS.*