FIT5147-Data Visualization and Exploration

New York Traffic analysis and Taxi-fare Prediction

Student Name:-Chaithra Kumar
Student Id:- 29924170

Monash University semester-1 2019
Chai0005@student.monash.edu

TABLE OF CONTENTS

1.INTRODUCTION

As per the 2017 analysis, New York drivers averaged 91 peak hours stuck in traffic last year, tying with Moscow, Russia for second crowded place. Taxi-traffic has been increasing over the years in New York. What's even worse, four out of ten of the worst U.S. corridors are found right here in New York. For the third consecutive year, the eastbound section of the Cross Bronx Expressway (I-95) tops the list.
It is analysed that New York Traffic congestion will cost $100 million over next five years.

1.1 Main Findings

The main emphasis of this project is on analysing :-

- How the Taxi-traffic has been increasing over the years in different borough
- Identify which time the city is most crowded
- Find different correlation factors related to taxi-price , to predict the taxi-fare .
- Analyse all these factors related to borough and over the years.
- Identify how these factors various across the borough

1.2Intended Audience

The Intended audience are New York traffic police and crowd that travels in taxi often mainly to check fare price and crowded places.

2.Design

Sheet1:

In sheet 1 the main idea was to identify various factors related to taxi traffic congestion and different factors affecting the fare price .

- Main filters identified were borough, year for the leaflet map.
- Show trend of traffic years over the leaflet using facet wrap and onclick the trend in a region should be displayed
- For fare prediction include a correlation plot and scatter plot.
- Time-series graph for various region
- Main filters throughout would be categories borough ,Time of the day, travel time and distance travelled

Meta data :- (I have included this as a reference to better understanding related to visualization)

Column	Description
Pickup_datetime	Pick up Date and time of trip
Dropoff_datetime	Drop Date and time of trip
Pickup_longitude	Pick up latitude
Pickup_latitude	Pick up longitude
Dropoff_latitude	Drop off latitiude

Dropoff_longitude	Drop off Longitude
Passeneger_count	Number of passengers
Trip_distance	Trip distance in (miles)
Fare_amount	Fare amount in dollars
borough	Area name (Identified through clustering)
year	Year of trip
Travel_Time	Travel time of the trip

Sheet2:

Layout :- Include a leaflet plot with various filter like,year,borough,time-stamp .To show fare prediction include scatter plot with various factors effecting the price in form of a face warp.

Focus:- Based on filters selected ,total trips will be shown, Total trips of each location can be seen by clicking on markers on the leaflet.

Operations : - On click of marker count will shown , selecting required parameters from the drop down.

Disadvantage: -

- The main aim of this project is not achieved i.e. fare prediction.
- User has to select various filter .
- Trend over the years is not shown.

Advantage:- Simple design

Sheet3:-

Layout :- Include the bar chart to show taxi-trips in various location, only time filter. Plot the correlation matrix for identifying various factors related to fare amount. Include a filter to filter the factor ,Scatter plot for various factors. Provide filter for required factor

Focus:- On hover of each bar chart show the count of taxi trips ,year and borough name.

Operations : -:- On hover of each bar chart show the count of taxi trips, selecting required parameters from the drop down.

Advantage :- Various Trend across the borough and years are easily identified.

Disadvantage: - Location of the traffic is not specify. Correlation plot might be confusing for the users to understand.

Sheet4:-

Layout :- Include an interactive pie chart ,where each sector of pie chart indicates the percentage of taxi-trips based on the year and time filter. Fare prediction using scatter plot and fit different smoothers.

Focus:- On hover of each pie chart show the percentage of taxi trips ,year and borough name.

Operations : - On hover of each pie chart show the count of taxi trips, selecting required parameters from the drop down and parameters for different smoothers.

Advantage :- It's simpler and prediction parameters for taxi-fare can be selected , Various Trend across the borough and years are easily identified.

Disadvantage: - Location of the traffic is not specif Pie chart is difficult to read and doesn't go well with more than 3 sections.

Sheet5:-

Layout :- Include a dashboard kind of view ,the side menu gives option to select different borough ,based on the borough selected different tab menus can be selected for taxi-trip's (leaflet), most congested time of the day (Time Series)and prediction tab(Scatter Plot). The taxi-prediction tab has filter where user can select the distance and the predicted fare for that area is shown. Overview tab which shows trend of various factors across the years and borough (circular bar plot)

Focus:- On over of marker on leaflet it shows number of taxi trips and on over of bar chart the number of taxi-trips on a specific year is shown .

Operations : - On over of marker on leaflet it shows number of taxi trips and on over of bar chart the number of taxi-trips on a specific year is shown .Time series graph has slider which can be used for finding number of taxi-trips in a given time, Various parameters to determine taxi-fare can be selected.

Advantage :- As mentioned above the target audience are crowd who wish to travel by taxi, Hence providing a specific filter for borough initially would be great advantage to the user, to navigate through all the required information only in that borough. Overall trend is shown the Overview menu

Dependencies:- .

Algorithm:- Linear Regression to predict the Fare, Clustering

Design details:- leaflet marker for each location on map, points on location where taxi-trips starts. Circular chart with various colours for each borough to show trend across years

Libraries :- dplyr, shinydashboard, shiny,leaflet, digraph, ggplot, plotly, shiny themes.

Estimated Time to Build:- 1 week

Software Requirements:- R- shiny

3.Implementation

The project is implemented using R-shiny- **Shiny** is an R package that makes it easy to build interactive web apps straight from R.

Following the libraries used:-

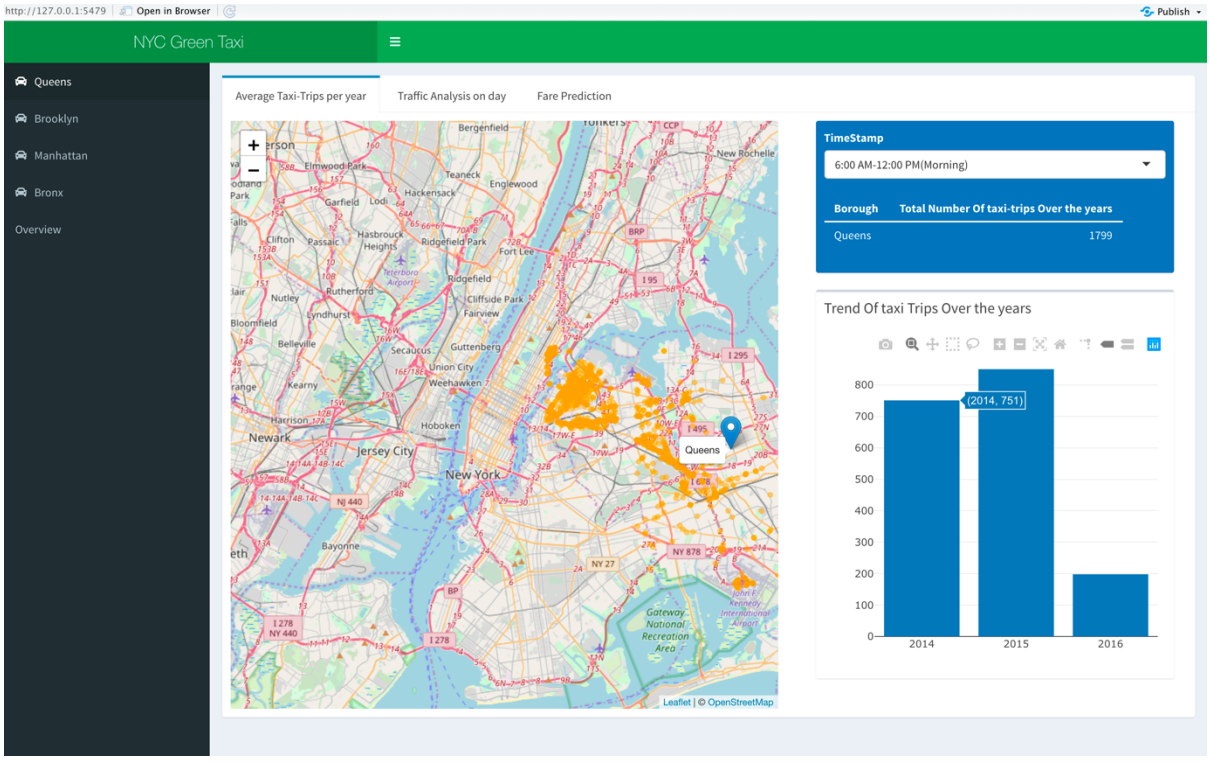
- dplyr (to filter data)
- shinydashboard (to get the interactive shiny dashboard)
- shinythemes(to change the colour of the dashboard)
- plotly(to get interactive map)
- ggplot(to plot graphs)

- dygraph(to plot time series graph)
- leaflet(to give interactive map)
- xts (to manipulate datetime object)

Implementation decisions :-

- Shiny is a inbuild framework to build we application directly from R .
- The main key feature is it has inbuild dashboard which I felt would be necessary for the outlook of visualization.
- To perform clusters and prediction R is the best Platform which is difficult to active through other platforms.

4.User Guide

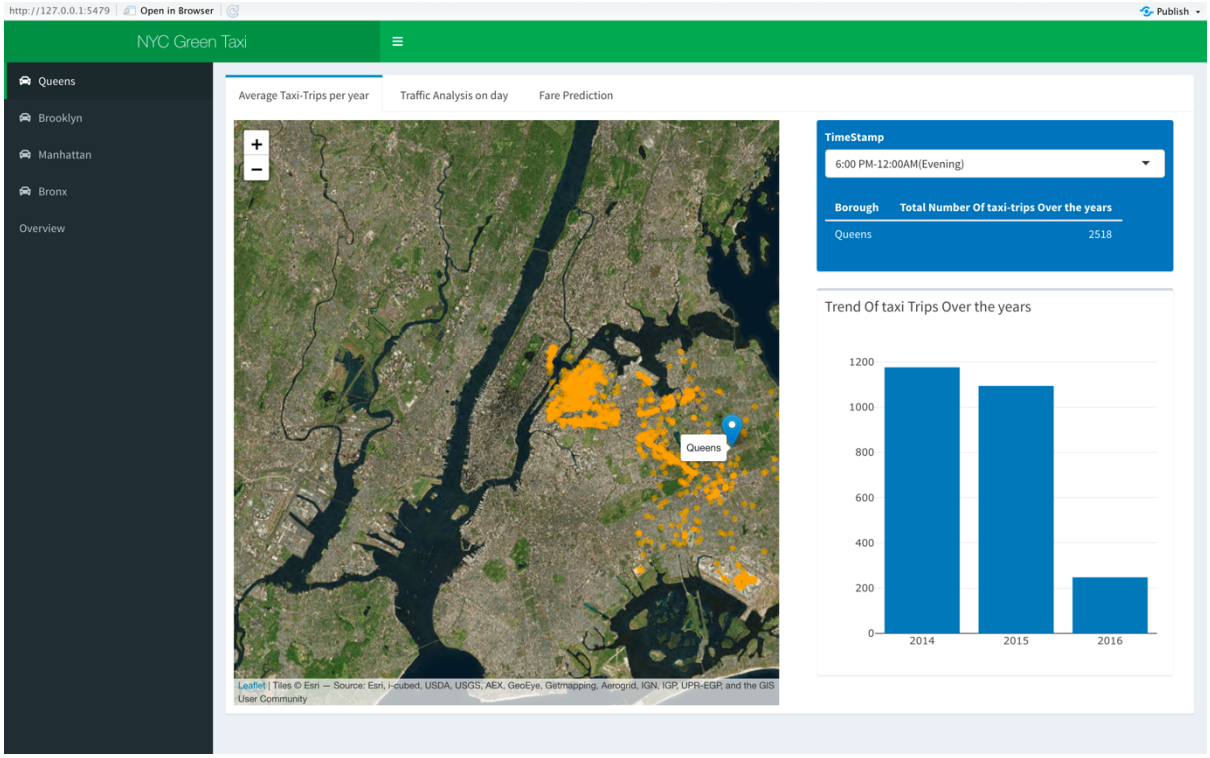


The above dashboard has side bar menus which on selection provides data related to particular borough.

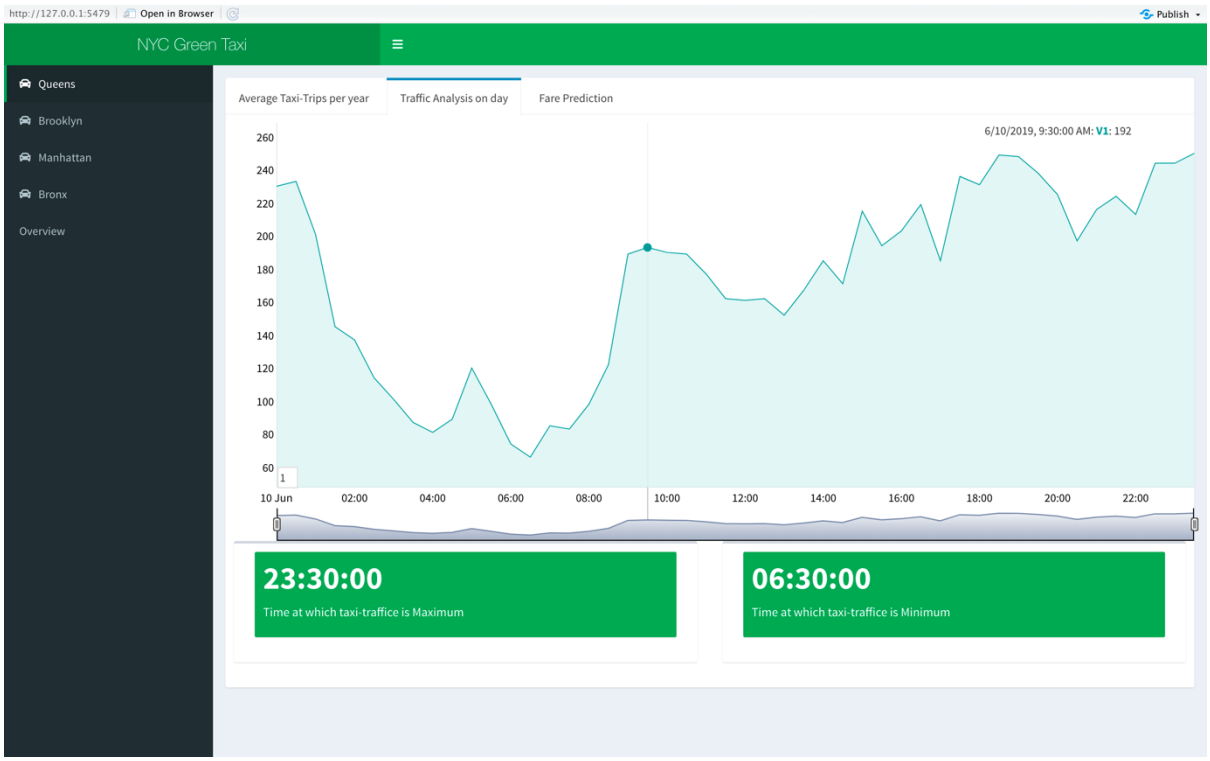
On selection of time stamp ,Number of trips in the borough is shown.

Below the drop down ,the total trips over the years for the particular borough and selected time stamp is displayed.

The bar chart below shows number of trips in particular year on hovering.

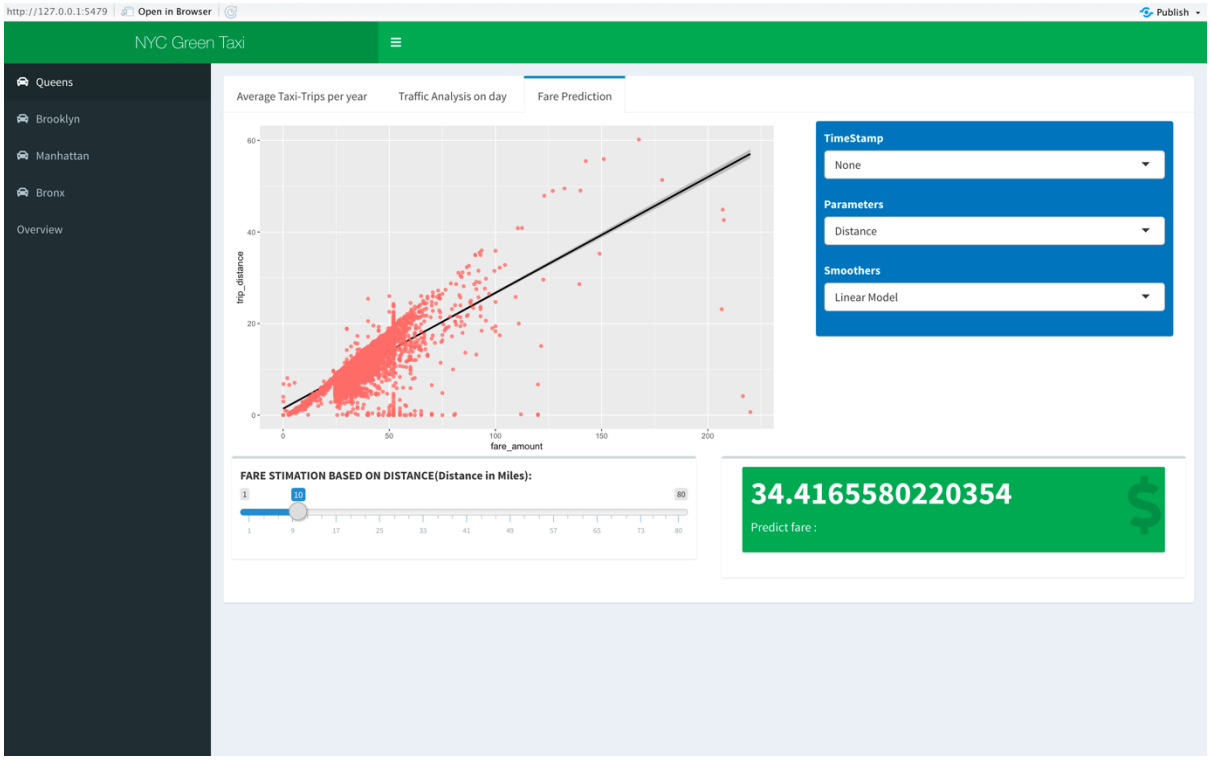


On change of the time stamp to evening or night, theme of the map changes.



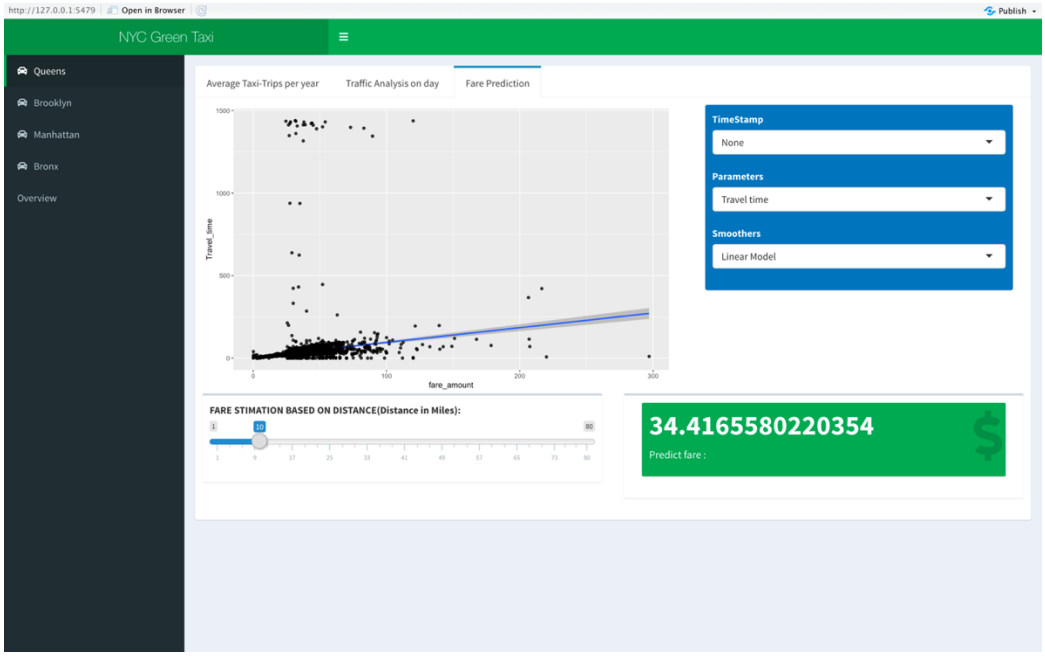
The above time series graph shows the number of taxi-trips at the given time of the day, on moving the slider along the time series ,count of trips to corresponding to time and borough is shown on the top.

The below two box ,shows the time at which traffic is maximum and minimum for particular borough.

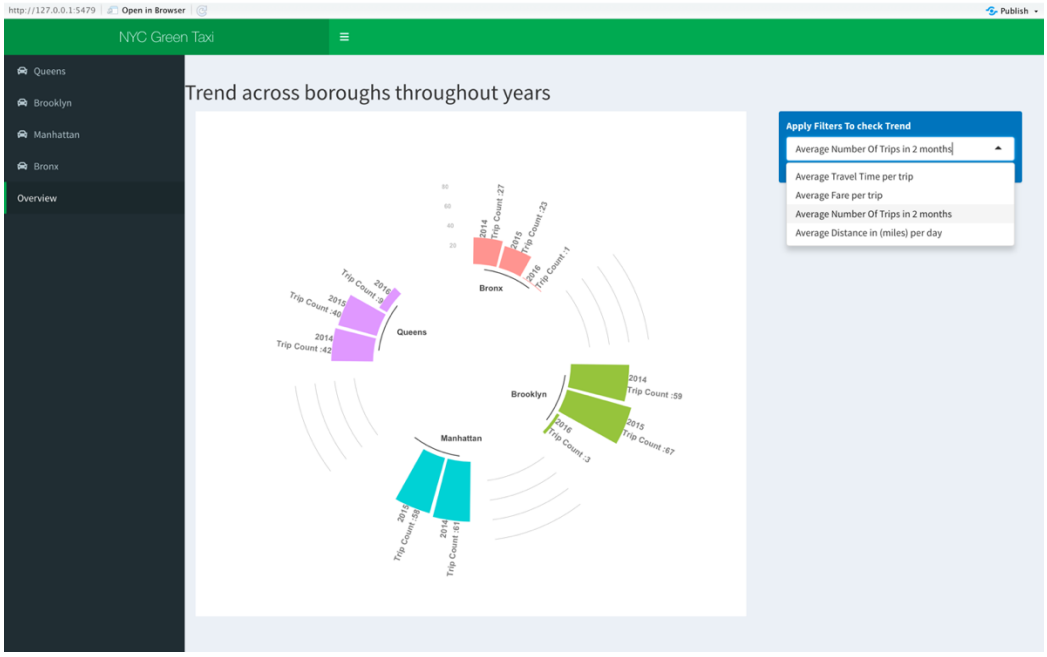


The scatter plot shows how fare amount varies across the parameter selected, here the parameter selected is Distance. The plot changes according to the filters selected.

On sliding to a value on slider, the fare for corresponding distance is shown in value box



The scatter plot shows how fare amount varies across the parameter selected, here the parameter selected is Travel Time



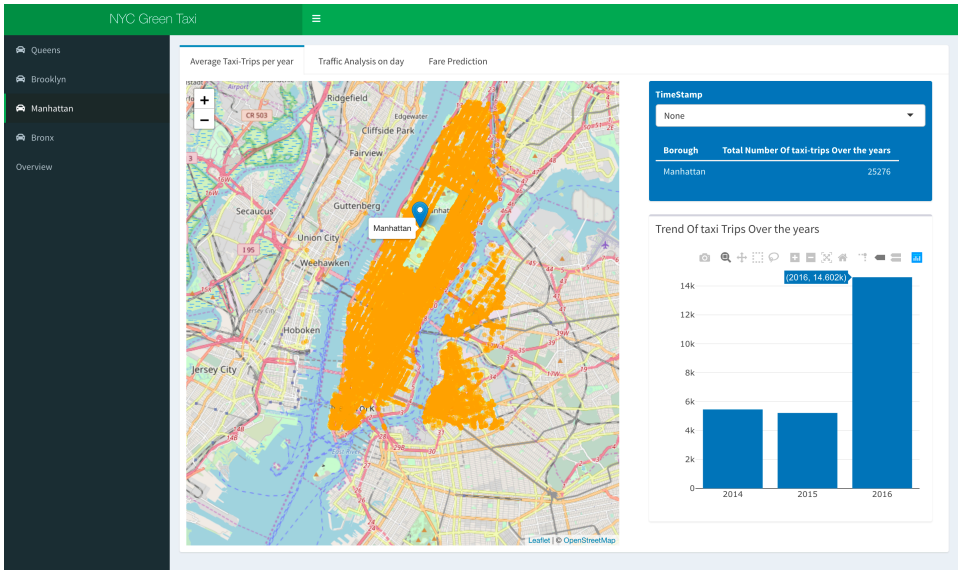
The above graph shows trend of select parameter throughout all boroughs over the years, This graph makes it very easy for comparison and to check the trend. Here the parameter selected is Average number of trips. The values varies as in the filter changes.

5.Conclusion

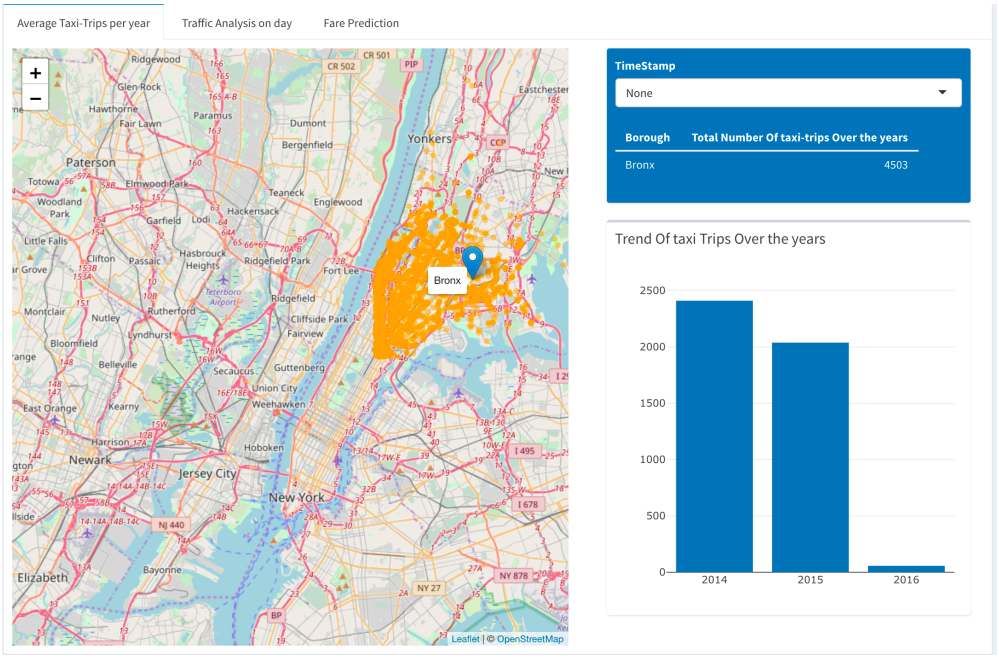
Brough	Total trips over the years
Manhattan	25,276
Brooklyn	11646
Queens	8245
Bronx	4503

From the above table it can be seen that Manhattan has highest number of trips over the years ,followed by Brooklyn.

Taxi traffic has been decreasing over the years in queen and Bronx, by Manhattan shows high traffic increase over the years



The above graph shows the taxi trip for Manhattan over the years has increased by double and the data points overlap a lot in that region which shows it has highest number of pickup points and taxi-traffic

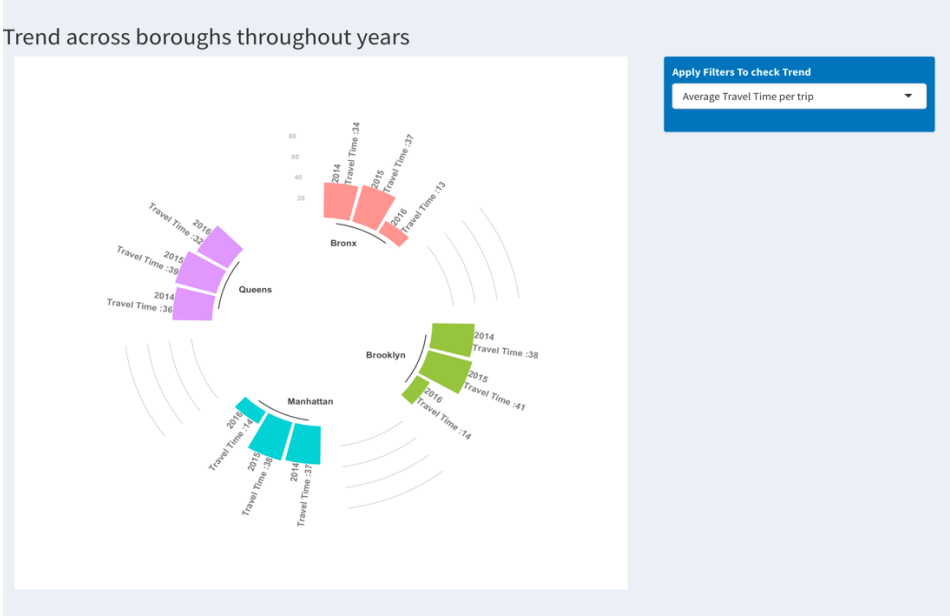


Bronx has lowest number of taxi trips and the taxi-trips has decreased abruptly in 2016.

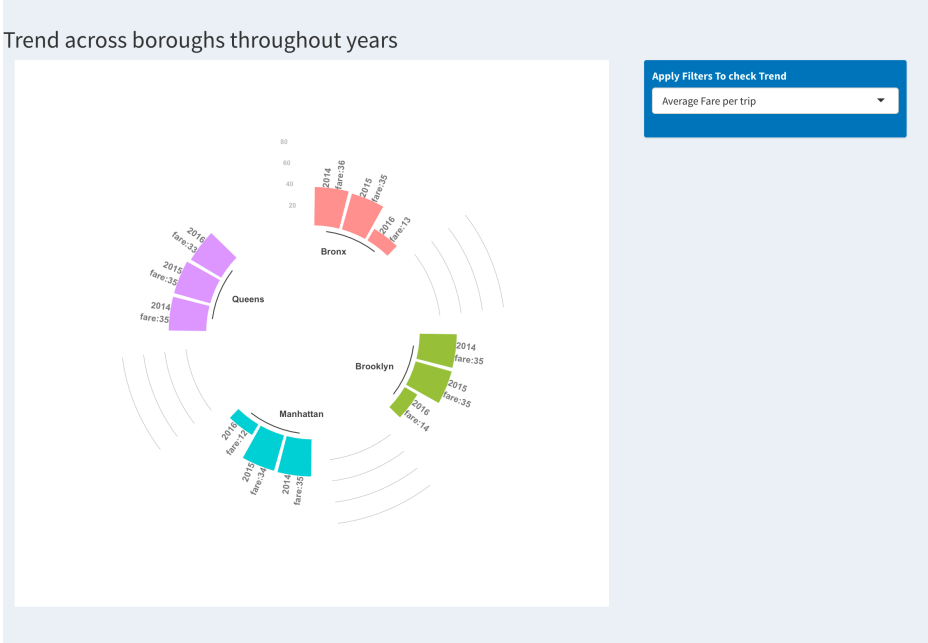
Taxi traffic is usually high at evening or night around 6-10 PM in all the regions ,whereas remains low in the morning around 3-6 am.

Borough	Predict fare for 10 Miles
Queens	34.416
Brooklyn	35.19
Bronx	34.677
Manhattan	29.84

The above table shows fare price in all the boroughs, for 10 miles highest is in Brooklyn whereas lowest where most taxi trips is taken i.e. Manhattan



The above circular bar chart shows per ride on average what is the travel time, Manhattan Travel time per ride has decrease may be due to short trips(Total fare amount will be reduced with increase in congestion), queens has average travel time of 35 minutes over the years.



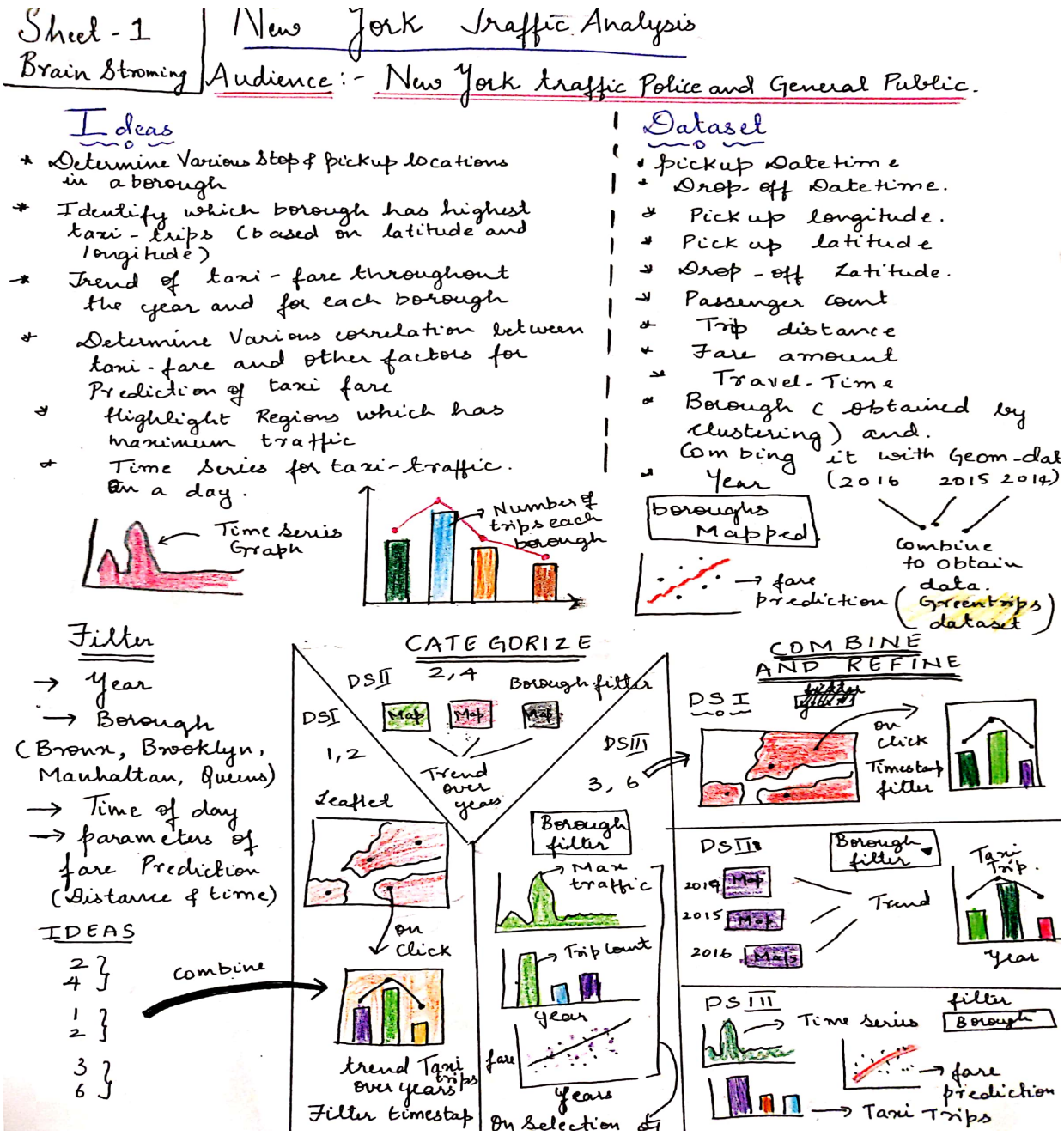
The conclusion made from above graph hold true, fare amount for Brooklyn has decrease due to short trips, whereas as in Queens it remains similar.

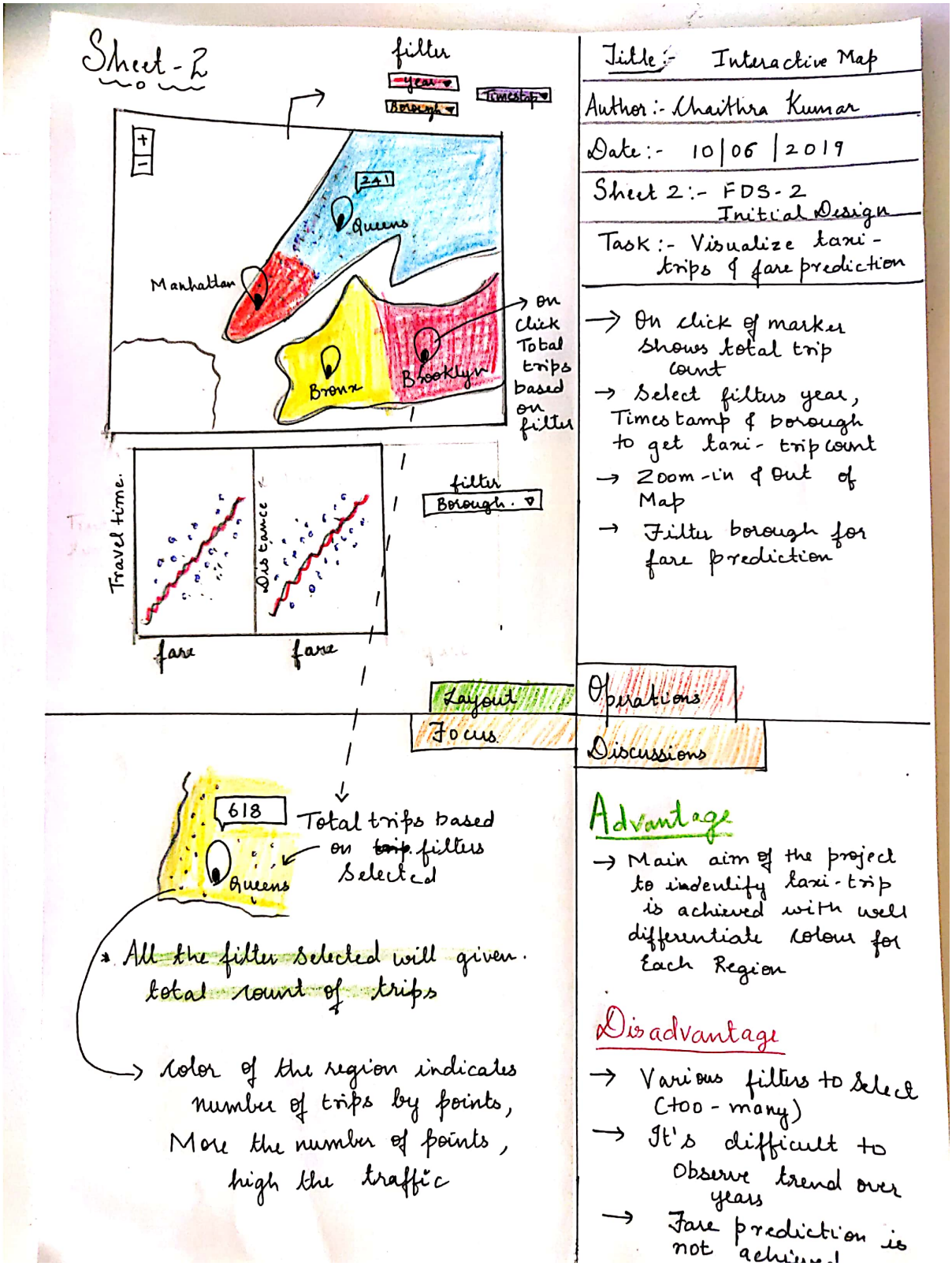
Learning Outcomes:- Various ways to visualize data ,apart from bar and simple charts and Five sheet methodology.

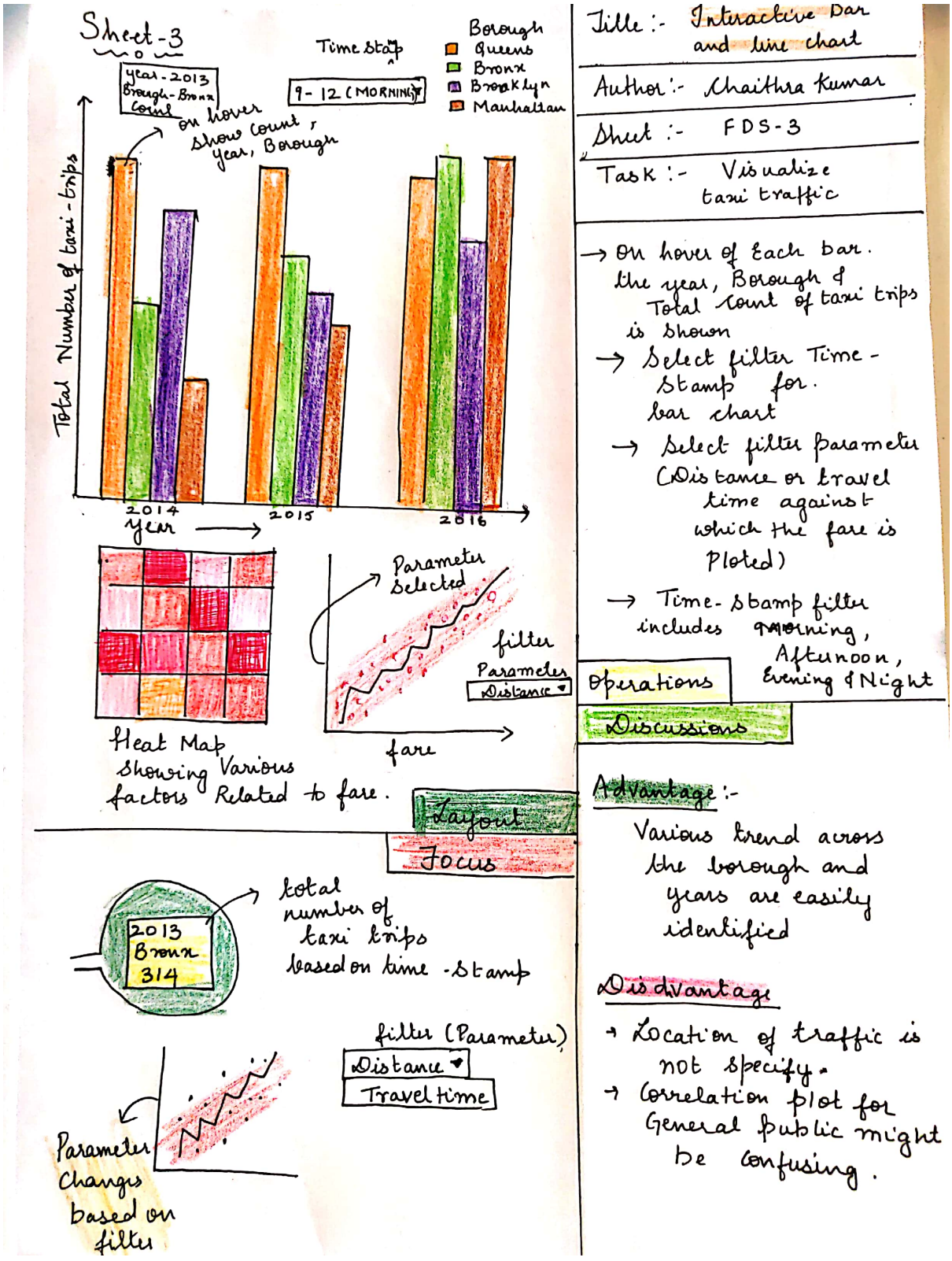
6.References

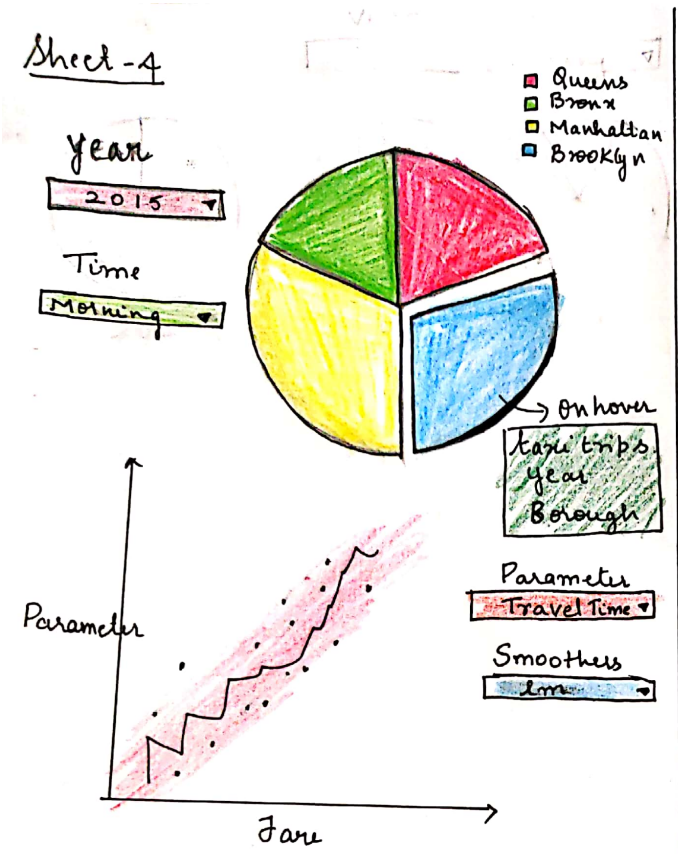
1. https://shiny.rstudio.com/gallery/
2. https://www.r-graph-gallery.com/interactive-charts/
3. https://www.statmethods.net/advstats/cluster.html
4. http://r-statistics.co/Linear-Regression.html

7.Appendix:




Sheet - 2
filter
year
Borough
Timestamp
241
Queens
Manhattan
Bronx
Brooklyn
On click Total trips based on filter
Travel time.
Distance
fare
fare
filter
Borough
Layout
Operations
Focus
Discussions
618
Queens
Total trips based on filters Selected
All the filter selected will given total count of trips
Color of the region indicates number of trips by points, More the number of points, high the traffic
Title :- Interactive Map
Author :- Chaithra Kumar
Date :- 10/06/2019
Sheet 2 :- FDS-2 Initial Design
Task :- Visualize taxi-trips & fare prediction
On click of marker shows total trip count
Select filters year, Timestamp & borough to get taxi-trip count
Zoom-in & Out of Map
Filter borough for fare prediction
Advantage
Main aim of the project to indentify taxi-trip is achieved with well differentiate colour for each Region
Disadvantage
Various filters to Select (too-many)
It's difficult to Observe trend over years
Fare prediction is not achieved


Sheet-3
Time Stamp
9-12 (MORNING)
Borough
Queens
Bronx
Brooklyn
Manhattan
year-2013
Borough-Bronx
Count
on hover show count, year, Borough
Total Number of taxi-trips
2014
2015
2016
year
Heat Map showing Various factors Related to fare.
Parameter Selected
filter Parameter
Distance
fare
Layout
Focus
2013
Bronx
314
total number of taxi trips based on time-Stamp
filter (Parameter)
Distance
Travel time
Parameter Changes based on filter
Title :- Interactive Bar and line chart
Author :- Chaithra Kumar
Sheet :- FDS-3
Task :- Visualize taxi traffic
→ On hover of each bar. the year, Borough & Total count of taxi trips is shown
→ Select filter Time-Stamp for. bar chart
→ Select filter parameter (Distance or travel time against which the fare is Plotted)
→ Time-Stamp filter includes Morning, Afternoon, Evening & Night
Operations
Discussions
Advantage :-
Various trend across the borough and years are easily identified
Disadvantage
→ Location of traffic is not specify.
→ Correlation plot for General public might be confusing.



Title :- Interactive pie-chart and fare prediction.

Author :- Chaithra Kumar

Sheet :- FDS-4

Visualize taxi-trips

→ On hover of each segment of pie-chart taxi-trip count, year and Borough is displayed

→ Year & Time filter for Pie-chart

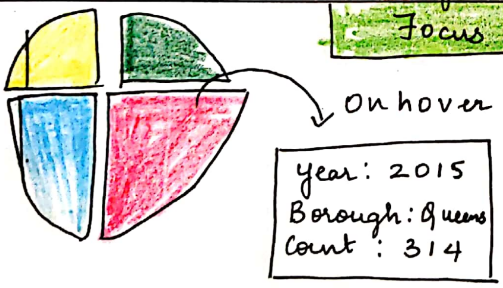
→ Parameter and Smoothers filter for fare-Prediction

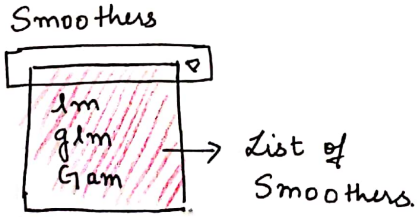
Operations

Layout

Discussions

Focus





Advantage

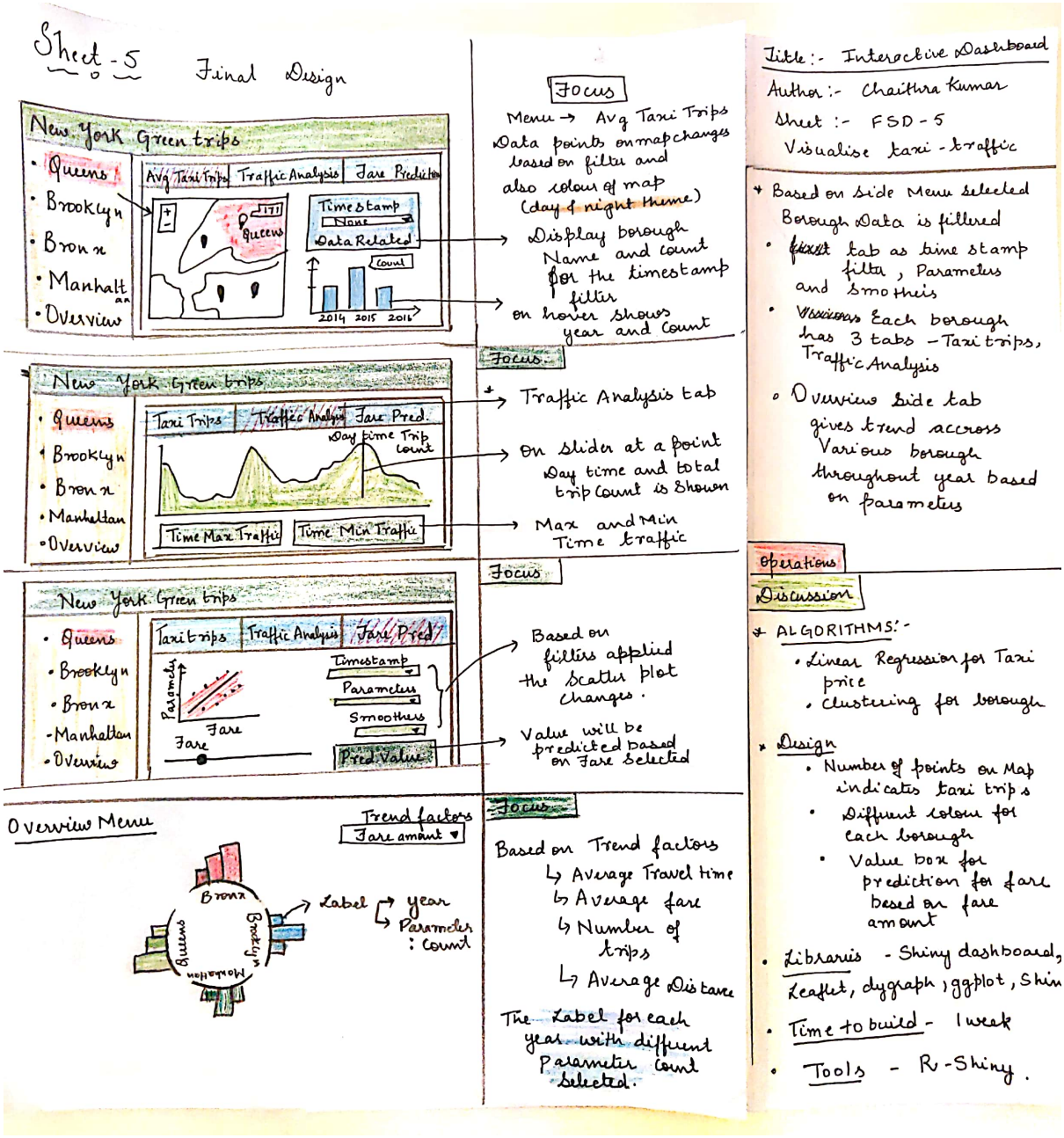
→ It's simple and prediction parameters with Smoothers can be selected

→ Trend across borough in a year can be easily identified.

Disadvantage

→ Location is not specific

→ Pie-chart is difficult to read and doesn't go well with more than


Sheet-5
Final Design
New York Green trips
Queens
Brooklyn
Bronx
Manhatt
Overview
Avg Taxi Trips
Traffic Analysis
Fare Prediction
Timestamp
None
Data Related
Count
2014 2015 2016
Focus
Menu → Avg Taxi Trips
Data points on map changes based on filter and also colour of map (day & night theme)
Display borough Name and count for the timestamp filter
on hover shows year and count
Title :- Interactive Dashboard
Author :- Chaithra Kumar
Sheet :- FSD-5
Visualise taxi-traffic
Based on side Menu selected Borough Data is filtered
tab as time stamp filter, Parameters and Smoothens
Each borough has 3 tabs - Taxi trips, Traffic Analysis
Overview side tab gives trend accross Various borough throughout year based on parameters
Taxi Trips
Traffic Analysis
Fare Pred.
Day time Trip count
Time Max Traffic
Time Min Traffic
Focus
Traffic Analysis tab
On slider at a point Day time and total trip count is shown
Max and Min Time traffic
Parameter
Fare
Smoothens
Pred. Value
Focus
Based on filters applied the scatter plot changes.
Value will be predicted based on Fare Selected
Operations
Discussion
ALGORITHMS:-
Linear Regression for Taxi price
Clustering for borough
Design
Number of points on Map indicates taxi trips
Diffrent colour for each borough
Value box for prediction for fare based on fare amount
Libraries - Shiny dashboard, Leaflet, dygraph, ggplot, Shin
Time to build - 1 week
Tools - R-Shiny.
Overview Menu
Trend factors
Fare amount
Bronx
Brooklyn
Queens
Manhattan
Label
year
Parameter : Count
Focus
Based on Trend factors
Average Travel time
Average fare
Number of trips
Average Distance
The Label for each year with diffrent Parameter Count Selected.