



FIT 5147

Data Visualization Project

NYC Taxi Statistics

By - Pradnya Alchetti
Student ID - 29595916

Table Of Contents

1. Introduction.....	2
2. Design.....	3
2.1 Sheet 1 - Brainstorming.....	3
2.2 Sheet 2 - Given Pickup and DropOff Zones get Fare Estimate.....	4
2.3 Sheet 3 - Taxi pickups trends from Airport.....	5
2.4 Sheet 4 - Revised Taxi pickups trends from Airport.....	6
2.5 Sheet 5 - Realization.....	7
3. Implementation.....	7
3.1 Explore tab 1 - NYC Airport Insights.....	8
3.2. Explore tab 2 - NYC Fare Estimator.....	11
4. User Guide.....	13
5. Challenges Faced.....	16
6. Conclusion.....	16
7. References.....	16
8. Appendix.....	17

1. Introduction

New York is known as the most busiest state in United States. Taxi has been an iconic part of New York. In New York there are two kinds of Taxi Services known as Green Taxi and Yellow Taxi serving day and night. People rely on them for everything, may it be a drive at 5am to the airport or a drive back home after the party.

New York city is served by several airports. The city's main three airports are John F. Kennedy International Airport and LaGuardia Airport in Queens, and Newark International Airport in Newark, New Jersey. These are known to be the busiest airport systems in America.

Taxi services would be highly required in these areas so that people can travel comfortably and safely may it be any time of the day.

Analysing the taxi trends over these airports would help the Business User to increase the profits by tying up with the airlines on these airports or by providing different discounts to the people.

New York city is divided into five Boroughs and each borough is further divided into zones. The Taxi data of these zones can help the customers to travel from one point to the other. On the customer point of view, a person always wants to travel with the least cost and minimum travel time. If he/she gets to know the travel cost and time prior to the journey, it would help them to plan their day accordingly.

The intended audience for this narrative visualization are Business Users and Customers. The narrative visualization uses NYC Green Taxi data for the year 2016 from October-December. This data set consists of 3.5M records.

For Business Users, it helps in analyzing the following trends:

- Monthly number of Pickups from the JFK, LGA and EWR airport.
- Payment Method used by the customers
- Average Tip percentage received with respect to day and time

For Customers, it helps to determine:

- Given the borough, pickup and dropoff zone average cost to travel from zone1 to zone2 at a particular time of the day
- Average trip time to reach the destination with respect to pickup time of the day.
- Number of trips from zone1 to zone2 which will help a person new to the city explore the zone with highest trips assuming that people mostly visit places where there are more Point of interests and this zone may be one of them.

Primary Message to be conveyed:

- Highest pickups from the airports are in the month of December which is the holiday period indicating people travel the most during this period and increasing number of taxis during this period would lead to a growth in the business.

- Additionally, allow the customers to estimate the fare rates for their travel.
- Customer can observe that average fare increases with respect to trip duration at particular hour of the day.

2. Design

Five Design Sheet Methodology(FDS) is used to create a prototype of the final design following a step by step process in each of the five sheets. This helps the designer to brainstorm and express his thoughts clearly.

In this narrative visualization, the FDS approach was used to design the final visualization. The description of each sheet is as follows:

2.1 Sheet 1 - Brainstorming

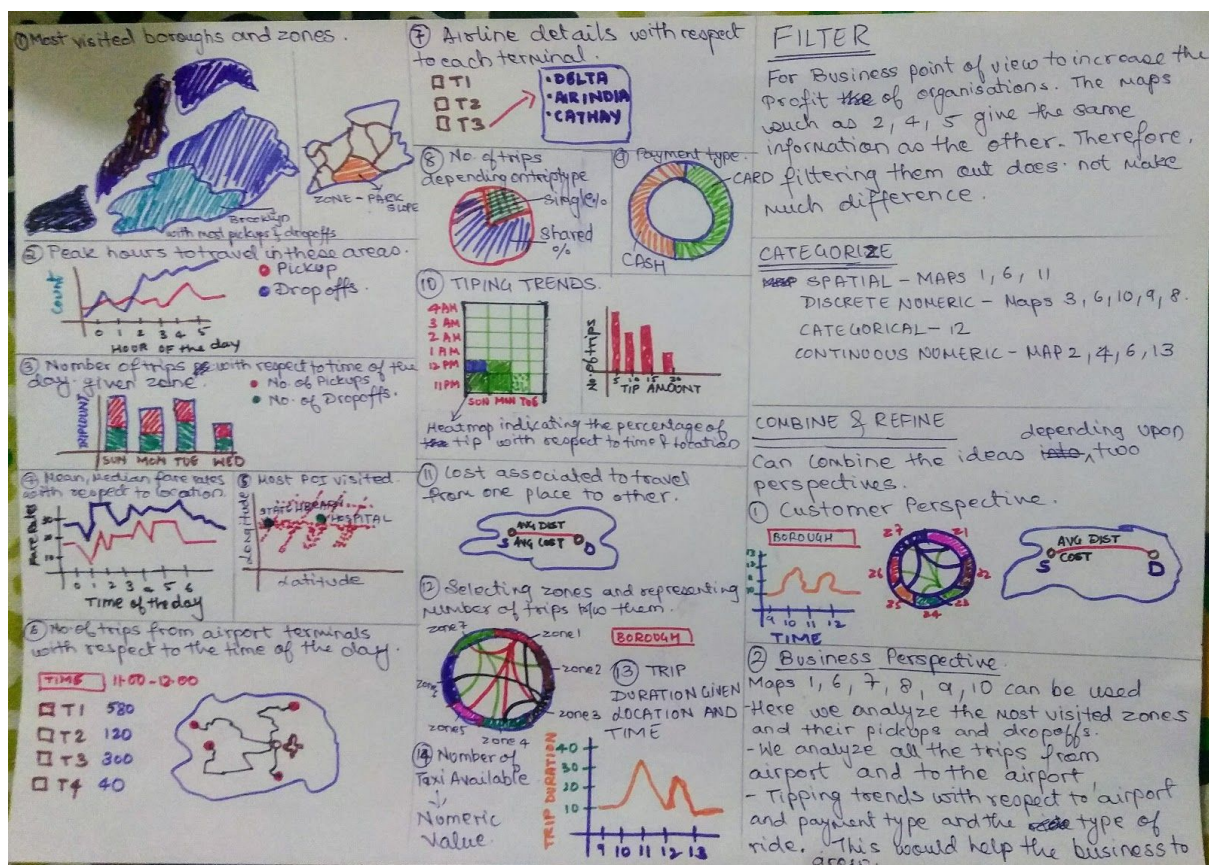


Figure1: Sheet 1-Brainstorming

This sheet represents all the possible ideas produced after brainstorming, that can be considered for final visualization such as Network Graphs, choropleth maps, heatmap, chord diagram.

The above graphs were then filtered and graphs that gave same information such as 2,4,5 were removed.

The graphs were then categorized based on two conditions:

1. Business Perspective - to increase the profit of organization
2. User Perspective - minimum cost and time to travel

From the business point of view, graphs such as 1,3,6,7,8,9,10 can be combined from which the business person can analyze the taxi trends in the most visited zones and airports. From user perspective, graphs 11,12,13,14 can be combined which will allow the user to estimate the fare and travel time, given two zones.

2.2 Sheet 2 - Given Pickup and DropOff Zones get Fare Estimate

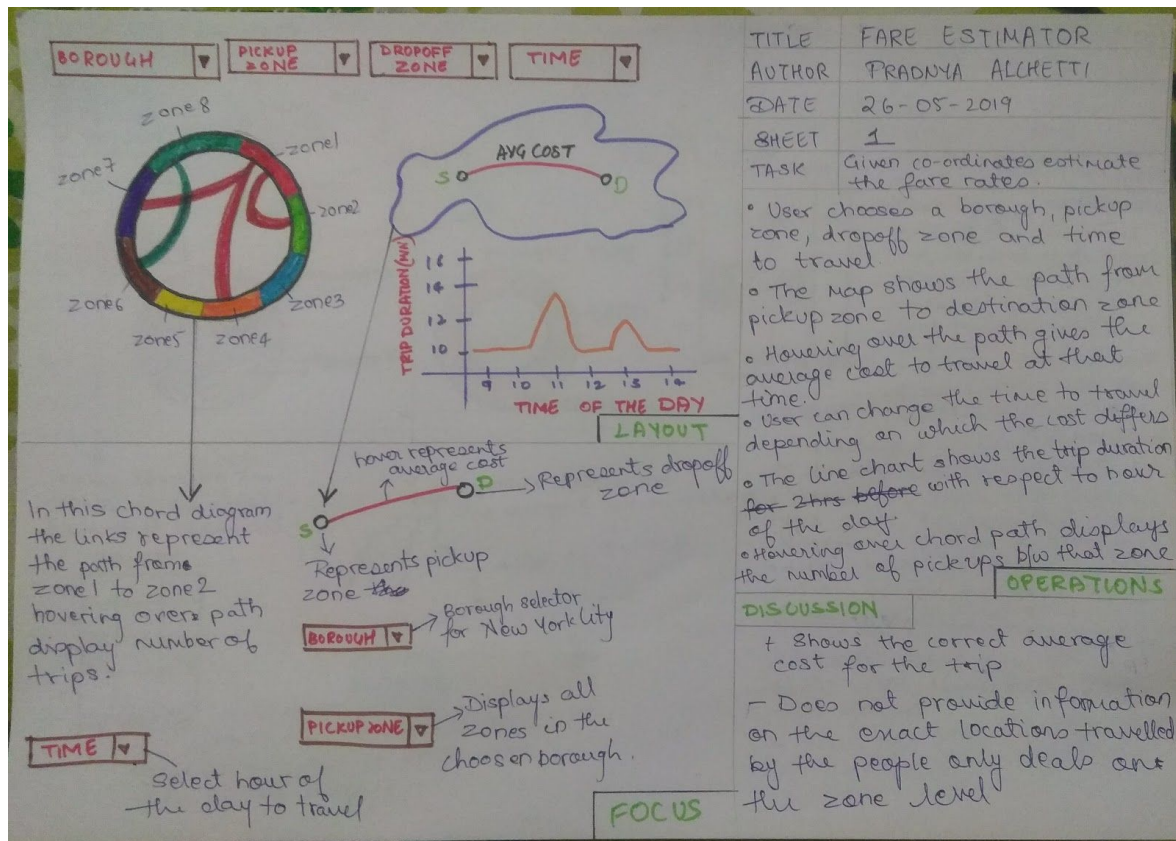


Figure 2: Sheet 2 - Fare estimate layout

The above sheet represents the layout of the Customer Perspective condition discussed above.

The layout consists of dropdowns to select Borough, depending on the selected borough respective zones in that borough appear in the dropdowns of Pickup_zone and Dropoff_zone.

The user then selects the pickup and dropoff zone and the time of the day to travel which renders a map with the pickup and dropoff location and path joining them, hovering over which gives the average cost to travel.

The line chart represents the time required to reach the destination with respect to the pickup hour of the day.

The chord diagram represents the connections between the zones and hovering over the links provide the number of trips between the two zones. This can help a new person in the city to explore the most visited places.

2.3 Sheet 3 - Taxi pickups trends from Airport

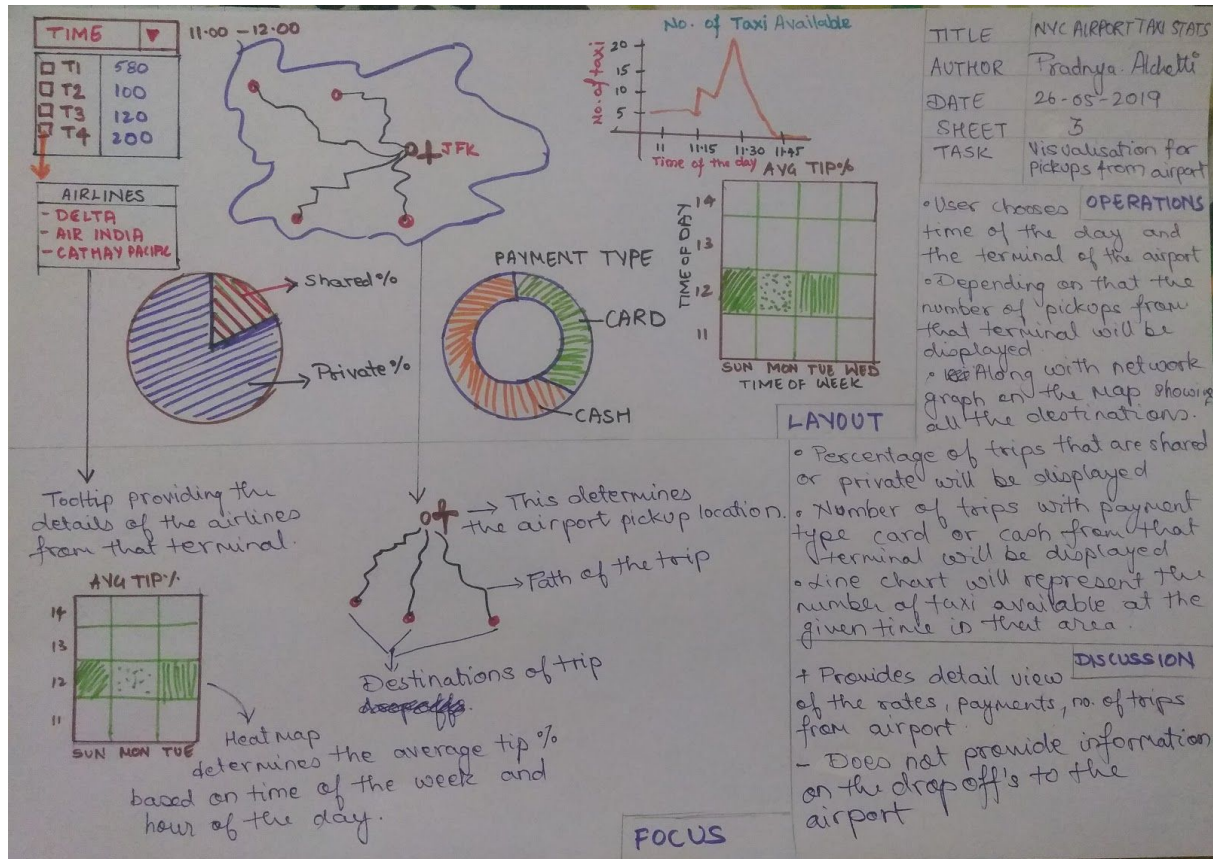


Figure 3: Sheet 3- Airport Trends

As the New York airports are the most busiest in America, from business perspective the business user would be interested in knowing taxi trends in these areas.

Thus graphs for most visited boroughs, zones and number of pickups and dropoffs in each zone were removed.

In the above sheet user can select the time by clicking on the Time dropdown and can select the terminal of the airport for which he wishes to see the stats.

Upon selecting the terminal the Map displays a network diagram which show all the trips from the selected terminal. Hovering over the airplane icon which displays the terminal on the map gives the information about the total number of pickups from that terminal at that time.

The pie chart represents how many of these trips were shared or private and the donut chart represents the number of these trips that are paid by cash or card.

The heatmap, average tip percentage with respect to time of the day would help the business user to get insights as to how the tips value change with respect to time and day.

2.4. Sheet 4 - Revised Taxi pickups trends from Airport

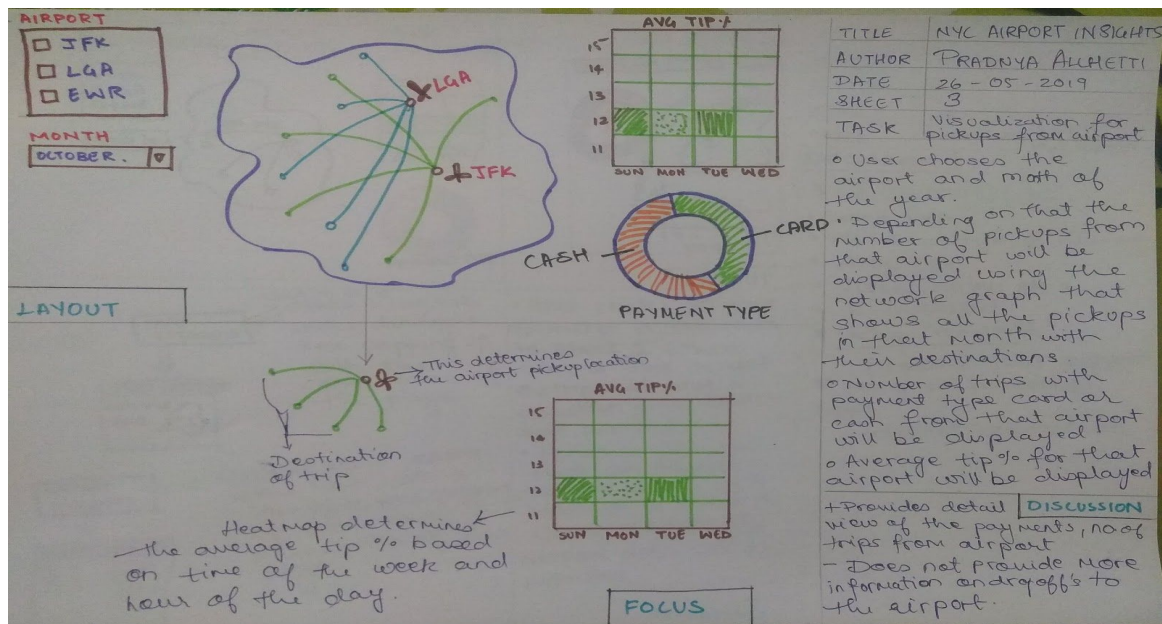


Figure 4: Sheet 4 - Airport Trends Revised

In the above the individual airport terminal checkboxes are replaced with overall airport. This was changed as the taxi data produced duplicate records for the terminals that is, if record A is valid for terminal 1 then it was also valid for terminal 2. In order to avoid this condition overall airport is considered.

The time dropdown changed to month as the business user would be more interested in monthly trends rather than daily trends.

Pie chart for shared or private ride was removed as there is no explicit column present in the data that gives this information.

Thus the final layout consists of dropdowns for time, checkbox for airport selection, Map with network diagram representing all the trips from particular airport and graphs for payment type and tip trends.

2.5. Sheet 5 - Realization

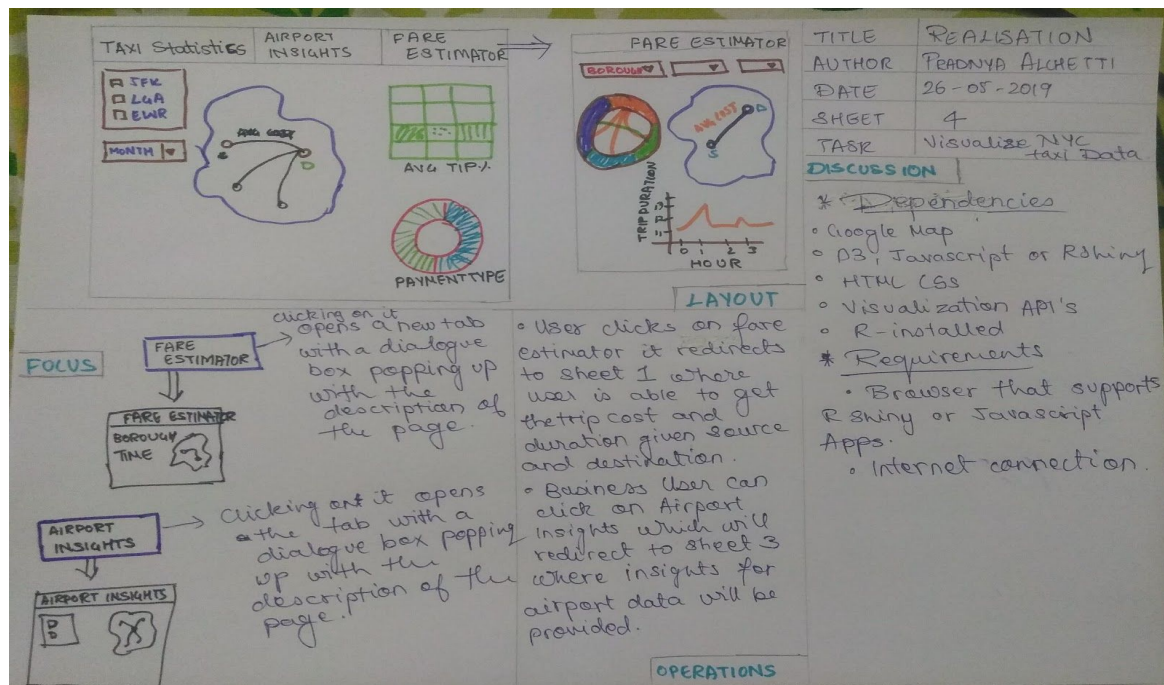


Figure 5: Sheet 5 - Realization

The above sheet represents the final layout as to how the final visualization would look like. The webpage would have two tabs one as Airport Insights and other as Fare Estimator. The Airport Insights is designed keeping in mind the Business User perspective as discussed above and the Fare Estimator represents the Customer perspective. On loading the application, Airport Insight would be displayed by default with a dialog box popping up describing what the page contains and if you are a customer then please visit the next tab. On clicking on Fare Estimator tab a dialog box would pop up describing the customer page. The Airport Insights final layout is as Sheet-4 and Fare Estimator Final Layout is as Sheet-2.

3. Implementation

Visualizations were implemented using Rshiny with R version 3.6.0 and shiny(library) version 1.3.2 with usage of CSS.

The visualization is performed on NYC Green Taxi Data 2016 for the months October to December.

As the dataset was huge about 3.5M records. For the purpose of Visualization only 8L dataset was sampled.

For airport insights, the airport data was then extracted by retrieving all the records within 1000m of range from JFK or LGA or EWR airport.

On loading the application, the page loads the following tab panel

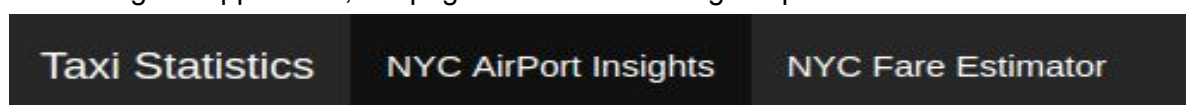


Figure 6: Tab panel

The first column represents the page name that is "Taxi Statistics" while the other two columns represents the tab names. Switching between the tabs can be achieved by clicking on any one of the two tabs NYC Airport Insights or NYC Fare Estimator. Clicking on any of the tab loads the respective dialog box describing what information can be retrieved from the particular page. The following is the dialog box rendered on clicking on NYC Airport Insights.

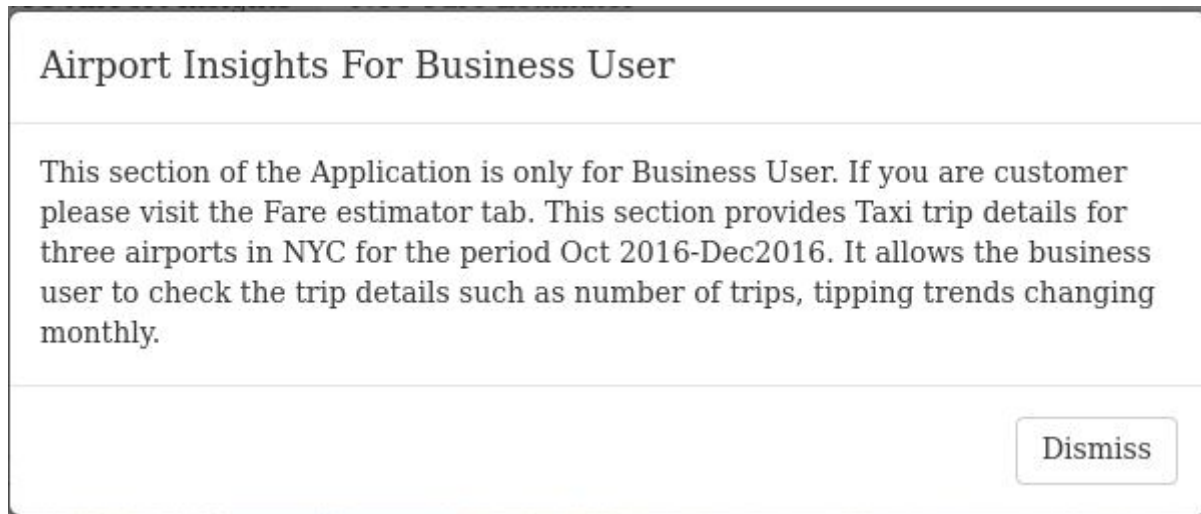


Figure 7: Dialogue Box

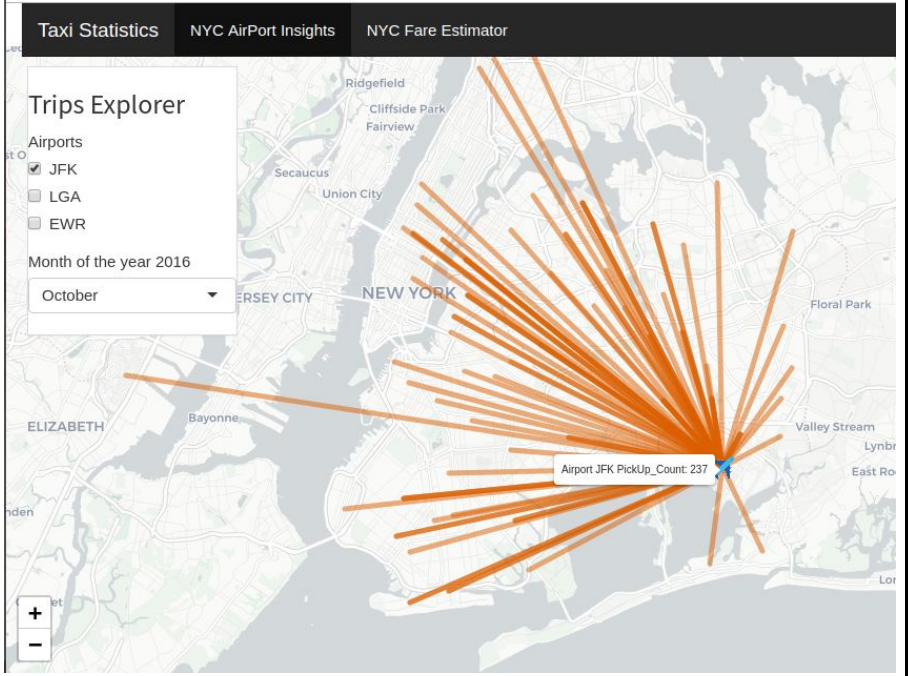
This will allow the application user to understand what information a particular tab would provide. User is then supposed to click on Dismiss to continue.

3.1 Explore tab 1 - NYC Airport Insights

The graphical elements used in visualizations are described below along with the interactions and features.

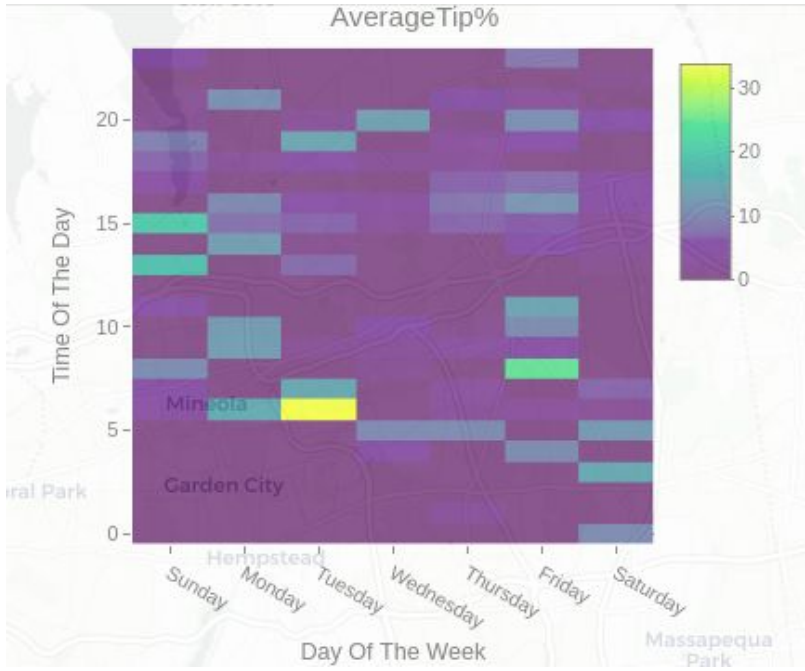
- Leaflet with Flow Graph

Visual Element	Leaflet with Flow Graph
Purpose	To provide the number of pickups from the selected airport and to determine in which zone most of the destinations lie. Depending upon the number of pickups the business user can increase the taxi frequency at that airport in that month

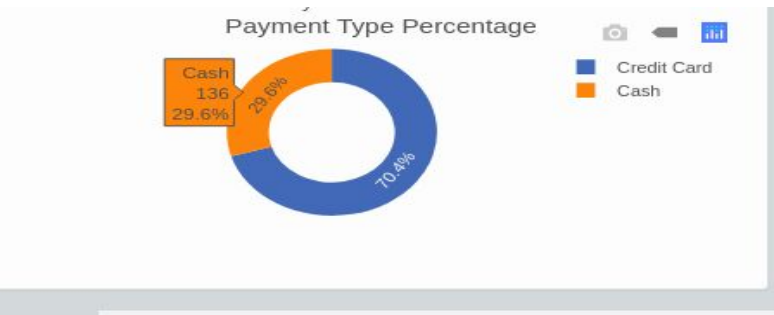
Visualization	
Interactivity	Hovering over the plane icon displays the total number of pickups from that airport
Filters	Filtering is applied based on the Airport and Month of the Year
Libraries	leaflet(render map), geosphere(determine intermediate points), RColorBrewer (for color palette)
References	Walker, K. (2019). Interactive flow visualization in R. Retrieved from http://personal.tcu.edu/kylewalker/interactive-flow-visualization-in-r.html

- Heatmap

Visual Element	Heatmap
Purpose	To provide information on which day and at which hour of the day the average tip is highest given the airport and month. This will allow the business users to know about the highest performed drivers at that time.

Visualization	
Interactivity	Hovering over the cells determine the average percentage of tip received with respect to particular day and time
Filters	Filtering is applied based on the Airport and Month of the Year
Library	plotly
References	Heatmaps. (2019). Retrieved from https://plot.ly/r/heatmaps/

- Donut Chart

Visual Element	Donut Chart
Purpose	To show the percentage of trips with payment type as Card or Cash. The business user would be able to provide more facilities depending upon which payment type is highest and would be able to keep a track of cash.
Visualization	
Interactivity	Hovering over the specific color on the ring will display the number of trips and the percentage

Filters	Filtering is applied based on the Airport and Month of the Year
Library	plotly
References	Pie Charts. (2019). Retrieved from https://plot.ly/r/pie-charts/

3.2. Explore tab 2 - NYC Fare Estimator

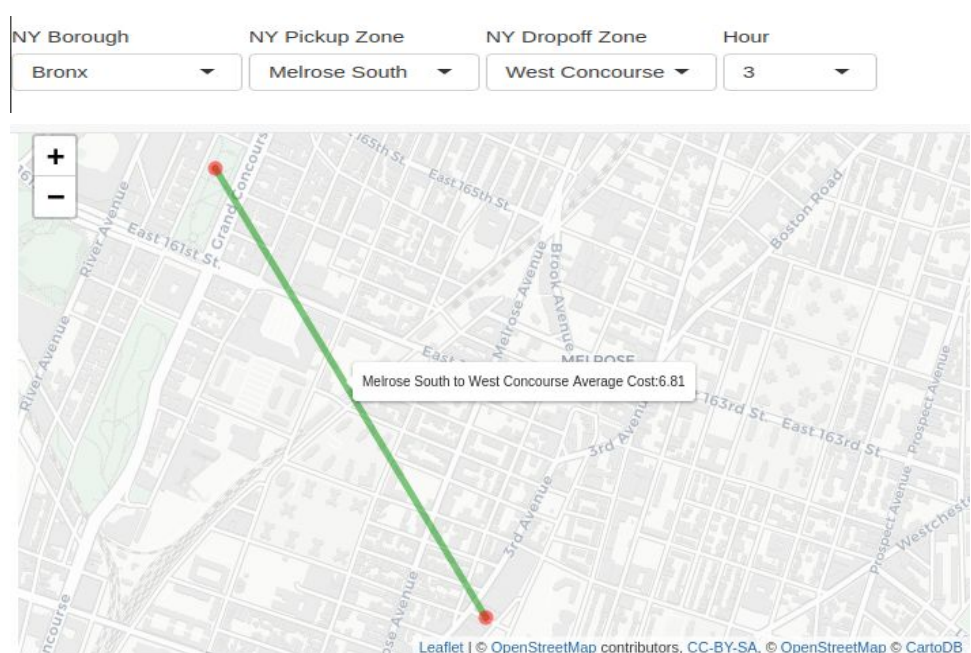
The graphical elements used in visualizations are described below along with the interactions and features.

- Chord Diagram

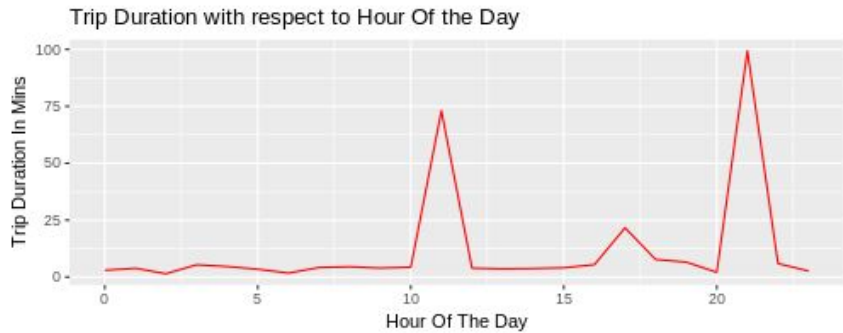
Visual Element	Chord Diagram
Purpose	Provides information on zone 1 is connected to how many other zones and number of trips between them. This information can be useful for the Customer who is new in the city and wants to explore. People tend to travel more to those places where there are more point of interests and this diagram can be helpful for the customer to find such places.
Visualization	<p>The visualization is a chord diagram representing travel data between different zones in the Bronx. The interface includes filters for 'NY Borough' (Bronx), 'NY Pickup Zone' (Mott Haven/Port Morris), 'NY Dropoff Zone' (West Concourse), and 'Hour' (0). The chord diagram shows connections between various zones, with a tooltip highlighting the 'Soundview/Bruckner' link, indicating 108 trips.</p>
Interactivity	Hovering over the chord link highlights the path between the zones and displays the number of trips between zone 1 and zone 2
Filters	Filtering is applied based on the borough. On selecting a different borough, corresponding zones to that borough are displayed on the chord

Library	library(chorddiag)
References	Interactive Chord Diagrams in R/Shiny. (2019). Retrieved from https://datascience-enthusiast.com/R/Interactive_chord_diagrams_R.html

- Leaflet with nodes and path

Visual Element	Leaflet with nodes and path
Purpose	Provides a path between the selected pickup and dropoff zone and displays the average cost required to reach the destination at a particular hour of the day. This allows the customer to estimate the average fare for travelling from one zone to the other.
Visualization	 <p>The visualization shows a web interface with four dropdown menus at the top: 'NY Borough' (set to Bronx), 'NY Pickup Zone' (set to Melrose South), 'NY Dropoff Zone' (set to West Concourse), and 'Hour' (set to 3). Below these is a map of a portion of NYC. A green line represents the travel path between two red circular markers. A tooltip box above the path displays the text 'Melrose South to West Concourse Average Cost: 6.81'. The map includes street names like River Avenue, East 161st St, Grand Concourse, and Melrose Avenue. Map controls like zoom in (+) and zoom out (-) are visible on the left. The bottom of the map has attribution for Leaflet, OpenStreetMap, and CartoDB.</p>
Interactivity	<ul style="list-style-type: none"> - Hovering over the path renders the average cost to travel. - Hovering over the nodes displays the pickup and dropoff zone names respectively. - If the pickup and dropoff zones are same then a circle marker over that zone is rendered and hovering over the circle, displays the name of the zone and the average cost
Filters	Filtering is applied based on the borough. Selecting the borough changes the dropdowns for pickup and dropoff zones. Selecting the pickup and dropoff zones renders the corresponding map
Library	leaflet(render map), geosphere(determine intermediate points)
References	Smart Taxi Vis: Interactive Data Visualization for NYC Taxi Data by Qiru Wang. (2019). Retrieved from https://www.youtube.com/watch?v=HE3Nva0lkyQ

- Line Chart

Visual Element	Line Chart
Purpose	To provide information about the trip duration in mins with respect to the pickup hour of the day. This would help the Customer to plan their travel as to at what hour of the day should they leave and schedule their day accordingly
Visualization	
Filters	Filtering is applied based on the pickup and dropoff zones. Changing the zones renders the corresponding trip duration chart.
Library	ggplot2

Miscellaneous Libraries: dplyr, shiny, shinythemes

4. User Guide

The Taxi Statistics Visualization can be viewed using a Rstudio and standard web browser. The files to run the code are present in the submitted folder.

- In order to run the app the required files are ui.R, server.R and styles.css.
- Load the ui.R in Rstudio and click on runApp.
- This will display the application in Rstudio panel. Please wait for around a minute to load the data
- Click on “view on browser” to view the application on the browser

This will display the first tab of the page “NYC Airport Insights ” .

- A dialogue box will appear rendering the description of the page. Click on dismiss.
- The following page should be rendered.

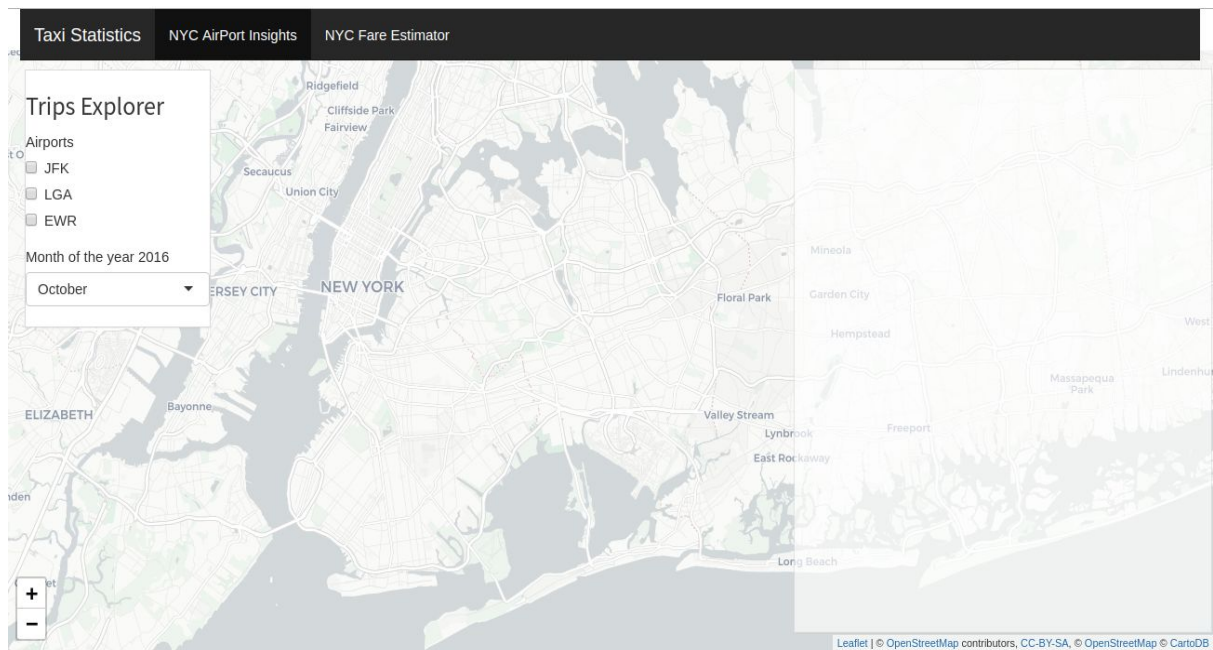


Figure 8: NYC Airport Insights Screen 1

- Select any of the checkbox and month of the year
- It should render the following

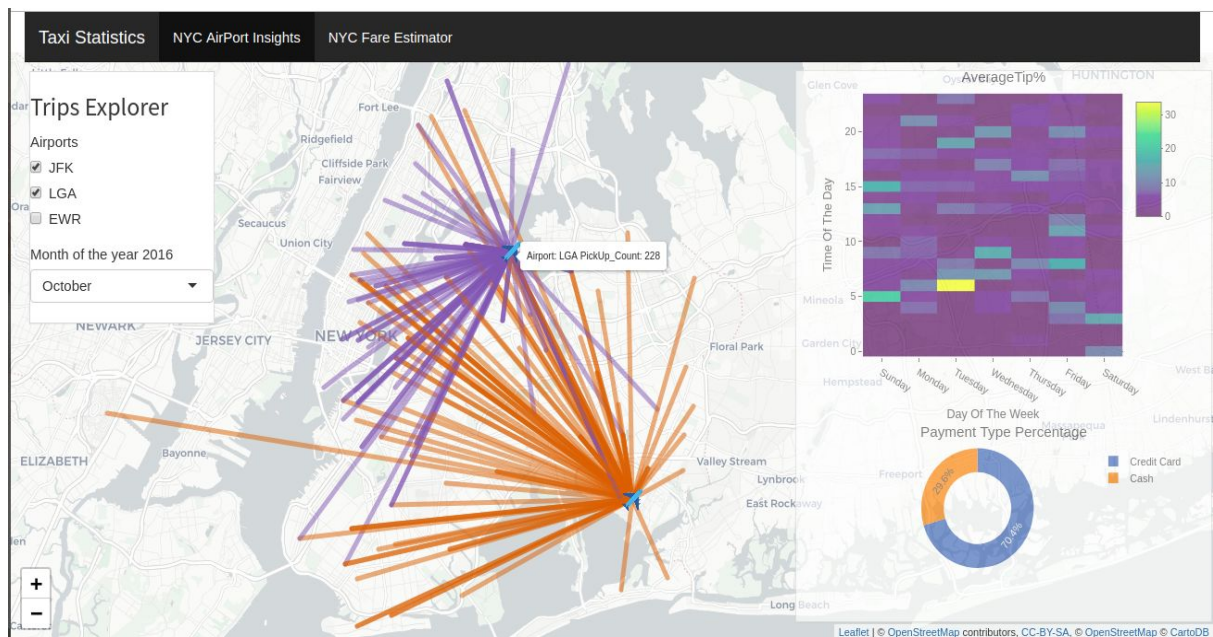


Figure 9: NYC Airport Insights Screen 2

- Hovering over the planes will display the total number of pickups from that airport
- The side panel on the right displays the average tip percentage for the selected airports and month while pie chart represents the percentage of trips that have paid by card or cash.
- Clicking on the next tab “Fare Estimator” should render the following

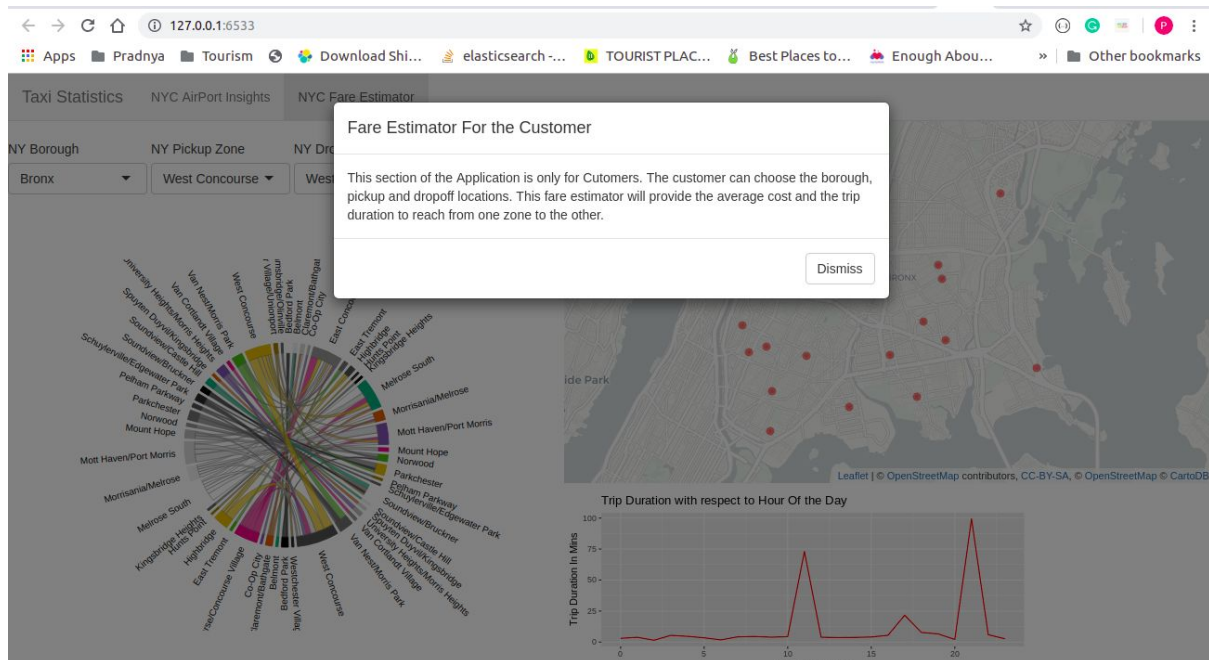


Figure 10: Fare Estimator Screen 1

- Click on Dismiss
- Select the borough of your choice and also the pickup and dropoff zone and hour of the day
- The following should be rendered on the page

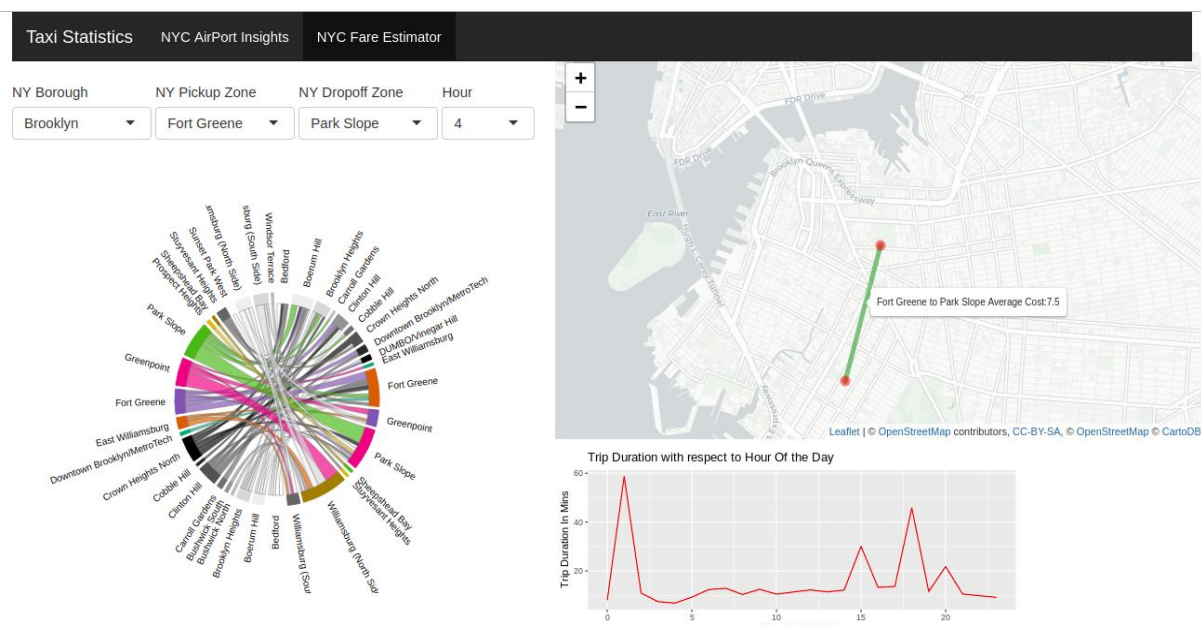


Figure 11: Fare Estimator Screen 2

- Hovering over the path displays the average cost estimated for the trip at the selected time
- Changing the time will also change the cost
- Hovering over the links of the Chord diagram highlights the link and displays the trip count between the selected zones.

- Try selecting the same pickup and dropoff zone, it should render a circle marker on the map indicating the zone and hovering over that displays the zone name and the average travel cost.

5. Challenges Faced

- The dataset was huge and was crashing the Rstudio. Hence, sampled the data to 8L records.
- Making the entire page as leaflet and adding additional graphs on it by using reactive in shiny.
- Faced animation issues while plotting the flow lines on the leaflet
- As the dataset was too large manipulating the data for chord diagram was a challenge.

6. Conclusion

The implementation of the narrative visualization in Rshiny was challenging. The ability to synchronise the interactivity and transitions adds complexity in programming as well as in visual presentation to achieve the final outcome.

Learnings from this exercise are as follows:

- Learning R and Rshiny coding to manipulate data and plot graphs.
- Implementing reactive and observe events in Rshiny.
- Linking CSS and Javascript with Rshiny.
- Creating an attractive UI by positioning the panels correctly.

This project helped me to improve my visual analytics skills by making use of different visualisations. I have other ideas such as in depth interactions with the chord diagram that is on clicking on the link it should render the corresponding path directly on the map. This will help the Customer to easily select the pickup and dropoff options. Also the fare estimator can be considered on a granular level such as exact location instead of zones.

7. References

Shiny - SuperZip example. (2019). Retrieved from
<http://shiny.rstudio.com/gallery/superzip-example.html>

Ognyanova, K. (2019). Static and dynamic network visualization with R. Retrieved from
<https://kateto.net/network-visualization>

app, C., Beer, P., & Alexander, J. (2019). Change the default position of zoom control in leaflet map of Shiny app. Retrieved from
<https://stackoverflow.com/questions/35543814/change-the-default-position-of-zoom-control-in-leaflet-map-of-shiny-app/43257393>

Shiny - modalDialog. (2019). Retrieved from
<https://shiny.rstudio.com/reference/shiny/latest/modalDialog.html>

8. Appendix

