NEW YORK CITY TAXI RIDES

Sarat Chandra Karasala
30098831
Lab 18 – Benjamin Lee

Table of Contents

Introduction

This Visualization tool has been created as a fun way of introducing basic statistics to school students. It uses a data set that has a little under half a million taxi ride details. Each record contains details like the pickup date and time, the distance, trip time, tip amount and a few other details.

Tipping is common in the US. So, our simple visualization tool takes this data and shows some basic statistics like the average tip amount for different NYC boroughs and the distribution of the tips for a given situation.

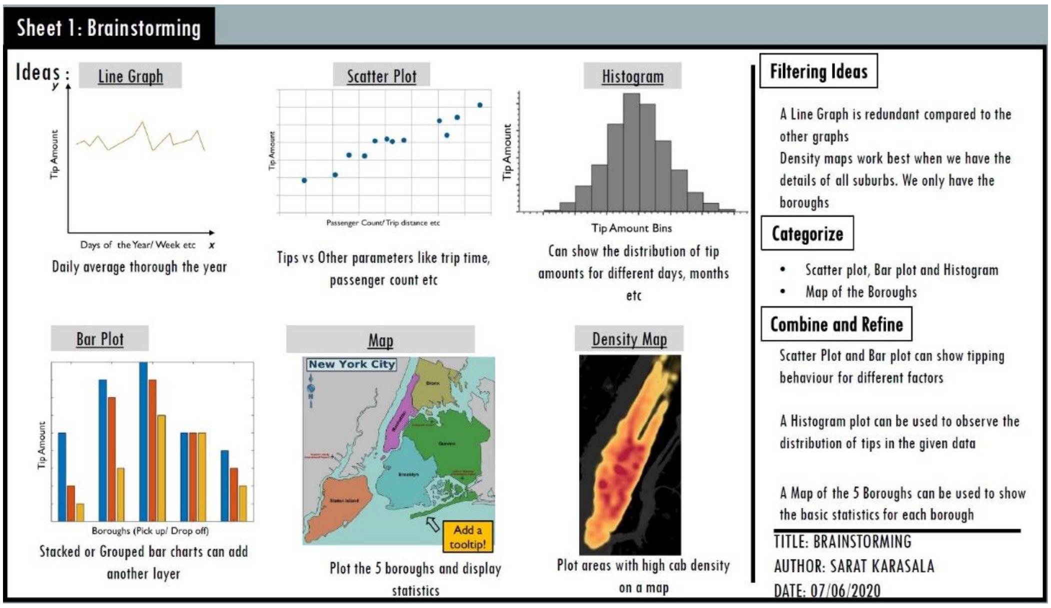
Our aim is to engage the student (user) with an interactive tool that can introduce them to basic visualization techniques and how to interpret them. The simplistic design is to ensure the user isn't overwhelmed with complex information. This tool could be a part of a larger edition of educational games or encyclopedias like the ones from DK.

Design

Five Sheet Design Methodology is an effective process that can be used to guide a visualization project. This could be for website design or making apps or any other project involving data and visualization. It consists of 5 sheets as the name suggests and the first sheet is called the **Ideas Sheet.**

Brainstorming

The first sheet is also called the brain storming sheet. Here we generate multiple ideas without going to deep into feasibility. The aim is to gather as many different ideas as possible that will make use of the available data.



Sheet 1: Brainstorming

Ideas :

Daily average thorough the year

Tips vs Other parameters like trip time, passenger count etc

Can show the distribution of tip amounts for different days, months etc

Stacked or Grouped bar charts can add another layer

Plot the 5 boroughs and display statistics

Plot areas with high cab density on a map

Filtering Ideas

A Line Graph is redundant compared to the other graphs

Density maps work best when we have the details of all suburbs. We only have the boroughs

Categorize

- Scatter plot, Bar plot and Histogram
- Map of the Boroughs

Combine and Refine

Scatter Plot and Bar plot can show tipping behaviour for different factors

A Histogram plot can be used to observe the distribution of tips in the given data

A Map of the 5 Boroughs can be used to show the basic statistics for each borough

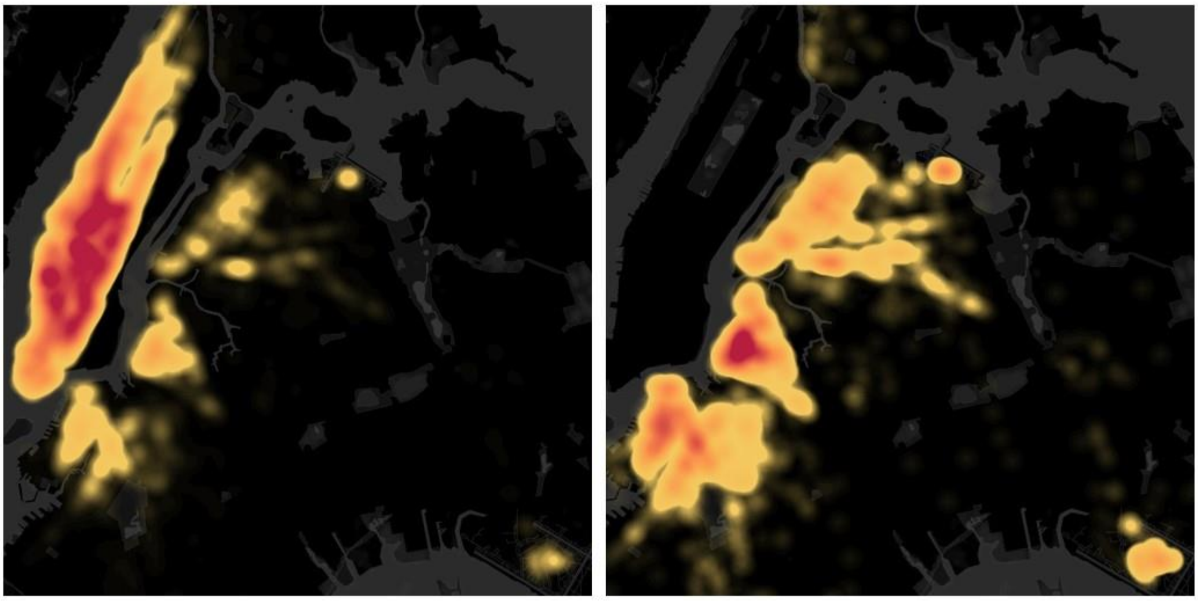
TITLE: BRAINSTORMING
AUTHOR: SARAT KARASALA
DATE: 07/06/2020

The taxi dataset has information like the pickup coordinates, date and time of ride, tip amount etc. Figure 1 shows the first sheet. The graphs considered were a line graph, scatter plot, histogram, bar chart, a geographic map and a heat map. The next step of brain storming is to **filter** these ideas. In this step we remove ideas that aren’t relevant or impossible to execute or unfeasible for any reason.

Filtering

The line graph is the most basic plot out of our ideas. It is only capable of showing one variable against a continuous variable like time. If we wanted to add more data, we would have to use multiple lines and risk making it too cluttered. So, a line graph is not the most informative compared to the other plots like bar, histogram and scatter plot which can show more data and information.

Density Maps show the concentration of dots on a map, where the difference in intensity is shown using a dual colour scheme. For example, as the number of cab rides starting in an area increases the area on the map moves from yellow to red. Density maps are a very good way to shows areas with high activity. The issue with our data is that about 90% of the taxi rides are in Manhattan. This leads to an interesting problem.



In the figure 2, the large land mass you see in the top left corner of the first image is Manhattan. The few scattered blobs of yellow represent the remaining 4 boroughs. But that is not accurate. Removing the Manhattan rides leads to the second image. In this case we are able to identify important details like the two airports marked. These features are hardly clear in the first image. The large difference in the amount of data between Manhattan and the remaining 4 boroughs makes this graph unusable. Even if it were implemented it would be too complicate for our target audience.

Categorize

In this step we categorize the mini ideas generated above into categories. The scatter plot, bar chart and histogram are like each other. They take tabular data and plot the points. Each graph shows the same information in slightly different ways which will be discussed in the following sections. The geographic map is also a good idea to present some statistics about each of the boroughs. It can be used to draw the target audience and engage them.

Combine and Refine

In this step we combine the ideas into bigger concepts. Maybe two graphs can complement each other and effectively present the information. For our project, the scatter plot and bar chart can show the tipping behaviour of users in different conditions. The scatter plot can show how tips vary with respect to any factor like time of day or trip distance etc. The bar chart can show how taxi riders from different boroughs tip.

The histogram shows us the distribution of the tip amounts. How many people tip 0 dollars? Or 1 or 2 and so on. The histogram coupled with one of the above graphs can show us more information than each graph by itself.

The map can be used to provide background information like little facts that can be memorized.

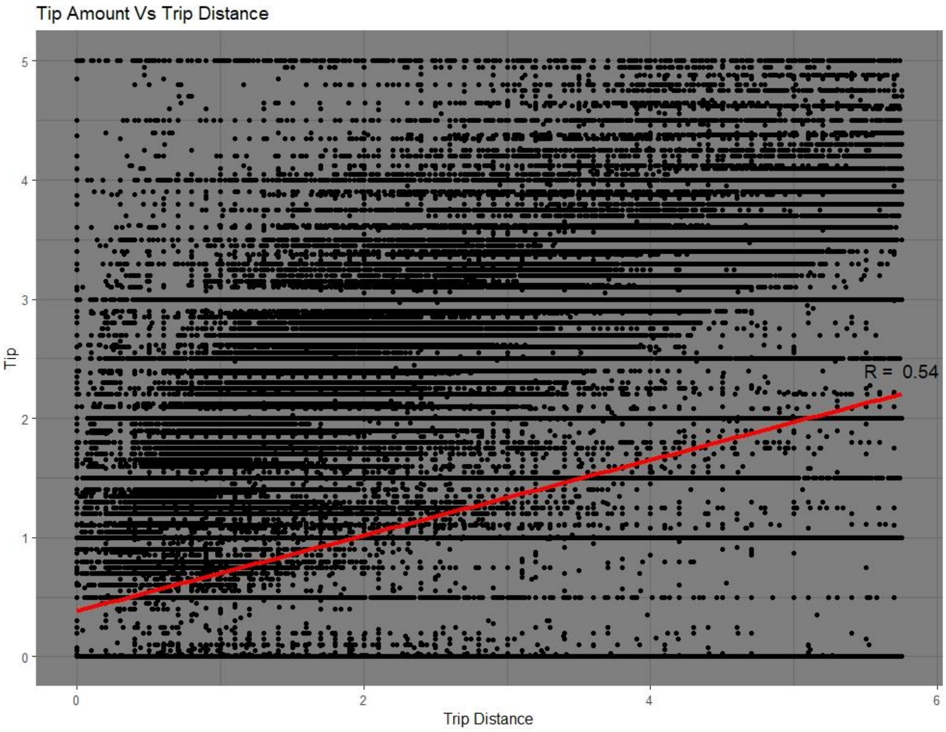
Sheets 2, 3 and 4

These are the design sheets. Here we refine the ideas we picked from the first step. So far we have shortlisted the scatter plot, bar chat, histogram and a geographic map.

Scatter plots

Scatter plots are very useful if you want to visualize a correlation between two variables. A cloud of points scattered all over the graph means no relation. A clearly identifiable line of data with a specific direction or a specific shape signifies a strong relationship between the variables.

Since we have about half a million lines of data we are faced with another interesting problem. The graph below we plotted the tip amounts against the trip distance.

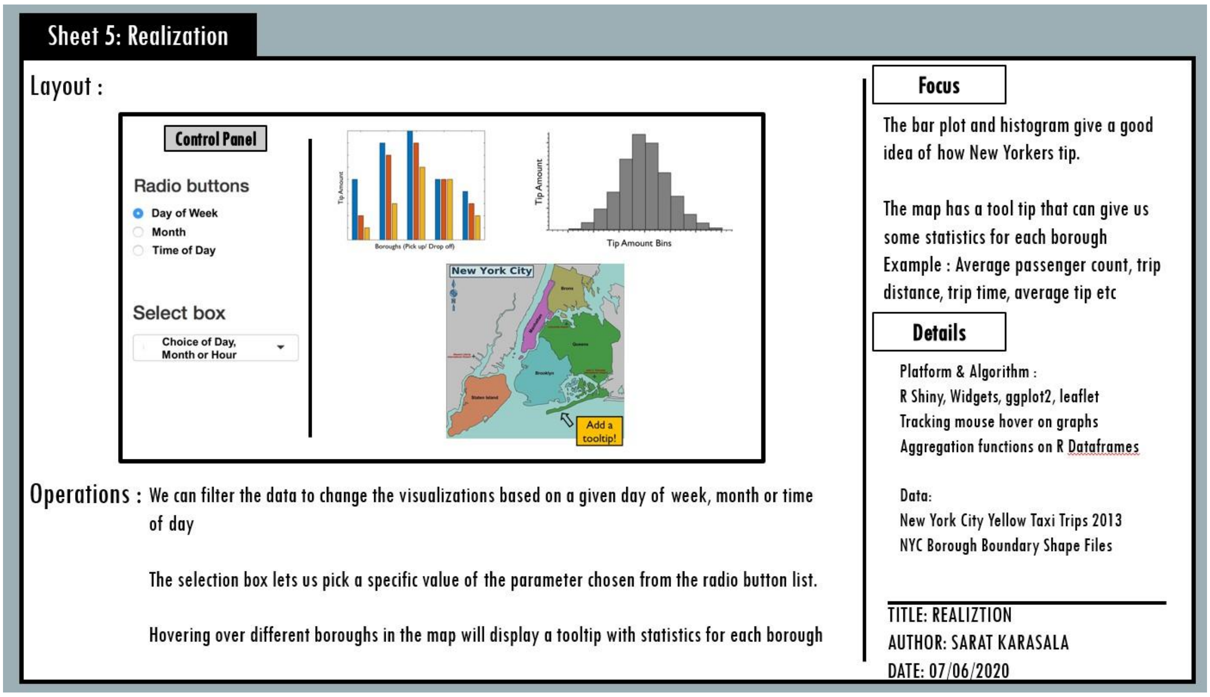


There is an interesting problem with the graph. The red line is a linear regression line with an R squared value of 0.54. And the line indicates where the points are the densest and hence leading to the least error. But when we look at the graph it appears as if the points are denser in the upper region of the graph far away from the regression line.

This demonstrates the misleading nature of a scatterplot when we have too many data points. This problem can be solved by the bar chart. With a bar chart we can show the average tips for different scenarios like different boroughs or different times of day etc. This way the thousands of points will be reduced to a single average value. But a bar graph alone would be insufficient. With the average tip amount, we have no indication of how the amounts are distributed. That is where the histogram comes in. The histogram will give us the frequency distribution of the various tip amounts. Together the bar graph and histogram give us a good visualization of the data.

Sheet 5

This is the final sheet. It is called the **Realization.** This represents the final design concept that will be implemented. The final implementation of our application will have 2 graphs that show the tip amount distribution and the map gives some more information.



Sheet 5: Realization

Layout :

Operations : We can filter the data to change the visualizations based on a given day of week, month or time of day

The selection box lets us pick a specific value of the parameter chosen from the radio button list.

Hovering over different boroughs in the map will display a tooltip with statistics for each borough

Focus

The bar plot and histogram give a good idea of how New Yorkers tip.

The map has a tool tip that can give us some statistics for each borough
Example : Average passenger count, trip distance, trip time, average tip etc

Details

Platform & Algorithm :
R Shiny, Widgets, ggplot2, leaflet
Tracking mouse hover on graphs
Aggregation functions on R Dataframes

Data:
New York City Yellow Taxi Trips 2013
NYC Borough Boundary Shape Files

TITLE: REALIZTION
AUTHOR: SARAT KARASALA
DATE: 07/06/2020

Implementation

Data

Two data sources were used for this application. The taxi ride data comes from the New York City Taxi and Limousine Commission (TLC). The full dataset has 165 million records. What we’re using is a much smaller dataset that was created by taking every two hundredth record. The other dataset is a shape file containing the boundaries of the boroughs in NYC. We use this data to classify the geographic coordinates of the taxi pick up into the 5 boroughs. This data was obtained from the NYC Open Data site.

Libraries used

A lot of different packages and libraries were used for the application. The important ones are shown below –

Shiny is an R package that creates simple interactive web applications.

Shiny Dashboard is a package built on shiny that makes it easy to build an interface with a neat side bar that can change tabs and other features.

dplyr is an R library for data manipulation on any dataframes or similar data types.

ggplot2 is a package for making graphs like the bar chart and histogram we have

Leaflet is a java script enabled package that can build interactive maps using R

rgdal and **sf** are packages for working with geospatial data like the borough shape files

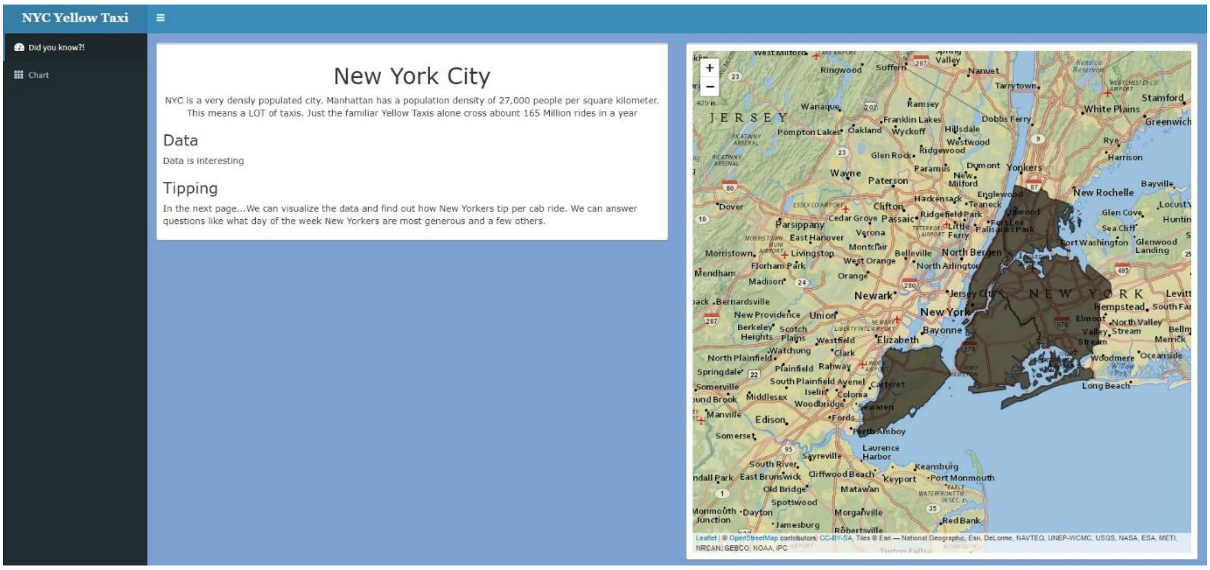
htmltools was used for formatting the tool tip in the maps

User Interface

The app has a two-tab interface. The first tab has some interesting facts and visuals that will draw the user in, and the second tab has the main visualization. This is possible using the Shiny Dashboard which makes building the interface very easy. More detail about the interface is given in the next section.

User Guide

User Interface



NYC Yellow Taxi

Did you know?!

Chart

New York City

NYC is a very densly populated city. Manhattan has a population density of 27,000 people per square kilometer. This means a LOT of taxis. Just the familiar Yellow Taxis alone cross abount 165 Million rides in a year

Data

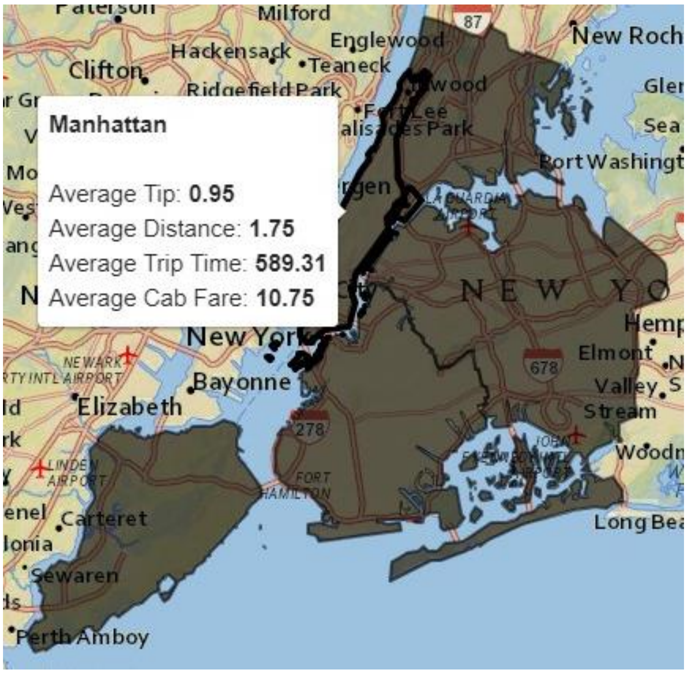
Data is interesting

Tipping

In the next page...We can visualize the data and find out how New Yorkers tip per cab ride. We can answer questions like what day of the week New Yorkers are most generous and a few others.

+
−
JERSEY
NEW YORK
Leaflet | © OpenStreetMap contributors, CC-BY-SA, Tiles © Esri — National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, IPC

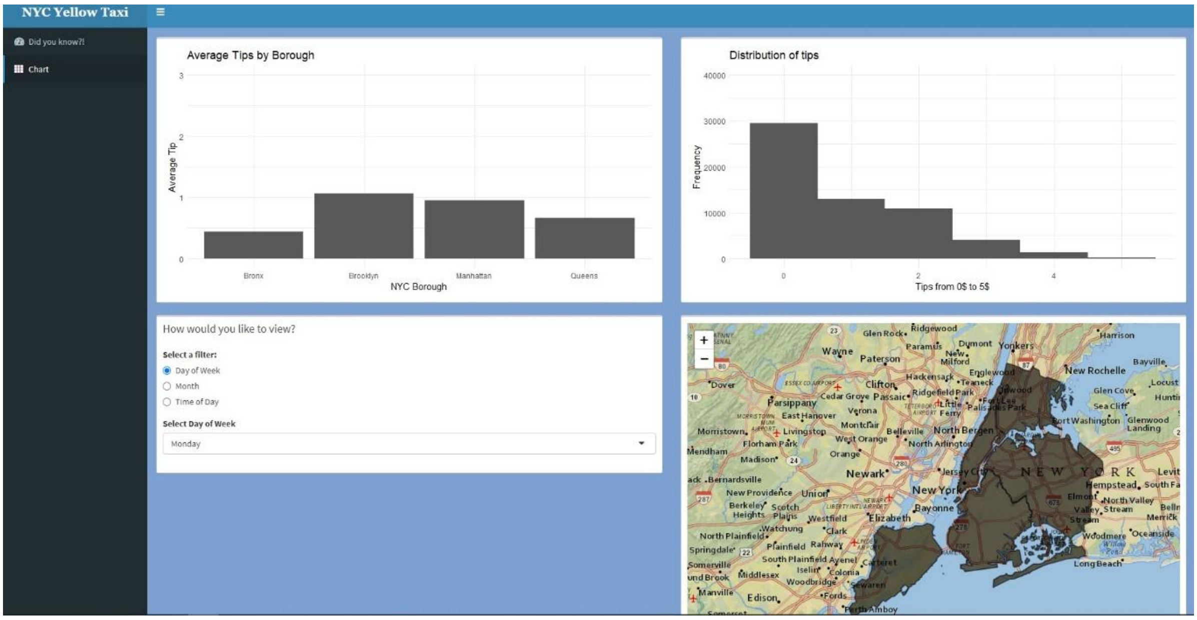
The figure above shows the opening page of the application. This is where the user will start. The darker section on the left side is called the side bar. It has the tab selection options. This application has two tabs. From here you can navigate to the second tab which is shown below. The box on the left presents a few facts about New York City and the map shows the 5 boroughs' boundaries.



Manhattan

Average Tip: **0.95**
Average Distance: **1.75**
Average Trip Time: **589.31**
Average Cab Fare: **10.75**

Hovering over each of the boroughs with a mouse will show a tool tip with a few basic details about the borough. The tooltip is shown in the figure below. For each borough it displays the average tip, distance, time and cab fare.



How would you like to view?

Select a filter:

● Day of Week

○ Month

○ Time of Day

Select Day of Week

Monday ▾

This is the main page of the app. You can see 4 sections. One of them is the selection box. There are two controls for the user to operate. The filter is a choice between **Day of Week, Month, Time of Day.** This lets the user pick how he/she wants to filter the data. For example, if we want the statistics of tipping in the month of August, we pick month in the first selection.

The second selection box is a list of whatever the user selects above. A list of months or hours of day or the days of a week. The remaining 3 graphs reflect what the user selects. The bar graph shows the average tips of each borough and the histogram shows the frequencies of different tip amounts. The map is like the one on the main page. The tool tip shows the same statistics for the respective filters.

Conclusion

We can observe a few trends after spending some time with the application. For example, riders starting from Bronx are the most fluctuating tippers. The graphs below show the distribution for Friday and Sunday. Throughout the week Brooklyn and Manhattan remain relatively stable but Bronx shows the widest fluctuation of tips offered with Sunday being the least at $0.07 and Friday highest at $0.79.

Brooklyn is where the tips are consistently on top with Manhattan a close second. Another obvious trend is that most of the riders offer no tip. This can be seen from the histogram. No matter what the filter is and the second selection the number of 0 tips is always highest.

Inferences from the project

Working on this project introduced me to two new packages. **Shiny and Leaflet**. What I've used is very little compared to what they offer. There were a few issues with the dataset that caused a few issues.

After the pre-processing of data there was only 1 record left that started in Staten Ssland. That is the reason in our visualizations the graphs mostly don't show Staten Island and even when it is visible (When the data is shown for Tuesday the graph shows Staten Island valued at 0) the value is usually not available (In the map)

There were a lot of records whose geo coordinates weren't translated into boroughs. This could be due to a problem with the code or the coordinates being invalid. Finding the cause of this problem will be tricky.

Working with Shiny Dashboard and Leaflet was a very good experience and I want to explore this further. Maybe finding a dataset related to a field I'm fascinated by will give me the push to explore these packages further.

Bibliography

Data Sources

[1] ipython-books/minibook-2nd-data. (2020). Retrieved 28 April 2020, from https://github.com/ipython-books/minibook-2nd-data/blob/master/nyc_taxi.zip

[2] Borough Boundaries. (2020). Retrieved 28 April 2020, from https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm

References

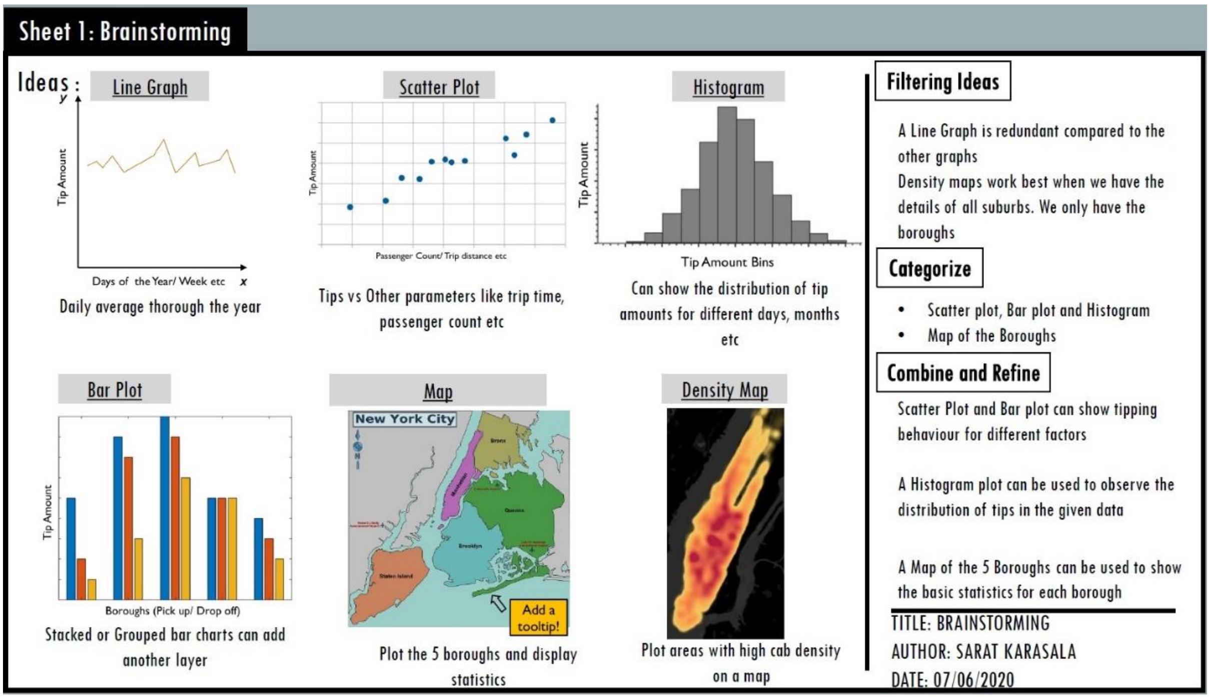
[3] Shiny Dashboard. (2020). Retrieved 21 June 2020, from https://rstudio.github.io/shinydashboard/

[4] 7 Fun Facts About NYC Taxis You Might Not Know. (2020). Retrieved 21 June 2020, from https://www.purewow.com/entertainment/fun-facts-about-nyc-taxis

[5] Leaflet for R - Introduction. (2020). Retrieved 21 June 2020, from https://rstudio.github.io/leaflet/

Appendix

Sheet 1



y
Tip Amount
Days of the Year/ Week etc x
Tip Amount
Passenger Count/ Trip distance etc
Tip Amount
Tip Amount Bins
Tip Amount
Boroughs (Pick up/ Drop off)
New York City
Add a tooltip!

Sheet 1: Brainstorming

Ideas :

Line Graph

Daily average thorough the year

Scatter Plot

Tips vs Other parameters like trip time, passenger count etc

Histogram

Can show the distribution of tip amounts for different days, months etc

Bar Plot

Stacked or Grouped bar charts can add another layer

Map

Plot the 5 boroughs and display statistics

Density Map

Plot areas with high cab density on a map

Filtering Ideas

A Line Graph is redundant compared to the other graphs

Density maps work best when we have the details of all suburbs. We only have the boroughs

Categorize

- Scatter plot, Bar plot and Histogram
- Map of the Boroughs

Combine and Refine

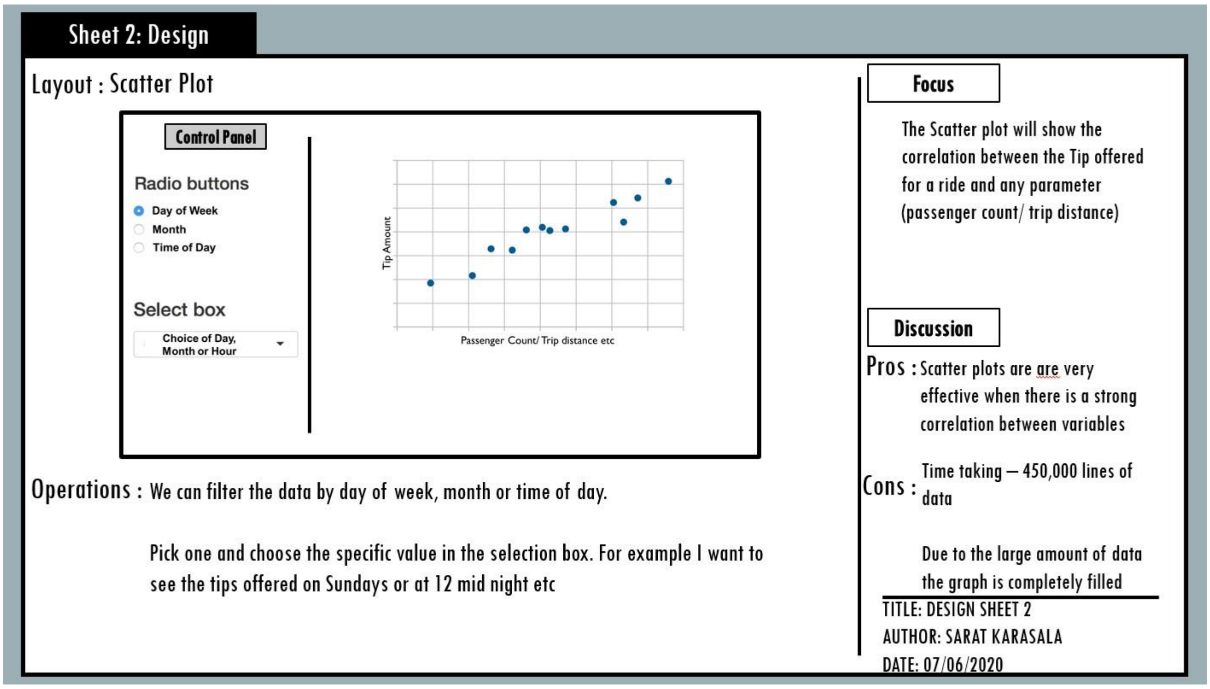
Scatter Plot and Bar plot can show tipping behaviour for different factors

A Histogram plot can be used to observe the distribution of tips in the given data

A Map of the 5 Boroughs can be used to show the basic statistics for each borough

TITLE: BRAINSTORMING
AUTHOR: SARAT KARASALA
DATE: 07/06/2020

Sheet 2



Sheet 2: Design

Layout : Scatter Plot

Operations : We can filter the data by day of week, month or time of day.

Pick one and choose the specific value in the selection box. For example I want to see the tips offered on Sundays or at 12 mid night etc

Focus

The Scatter plot will show the correlation between the Tip offered for a ride and any parameter (passenger count/ trip distance)

Discussion

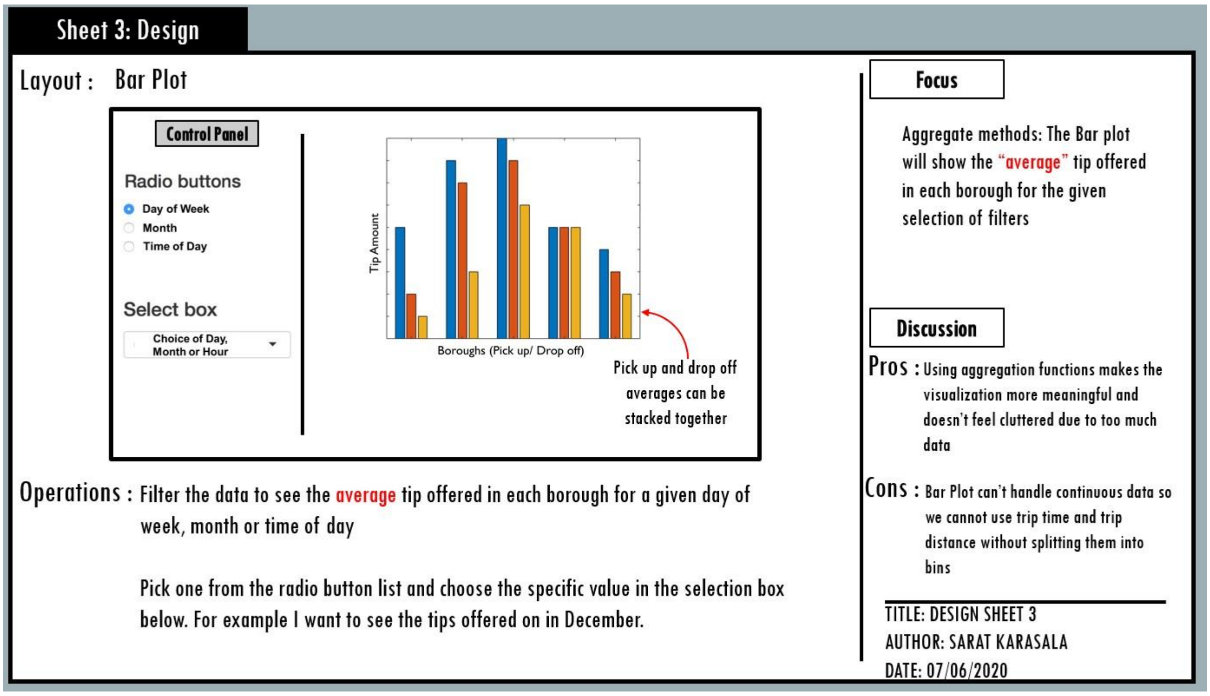
Pros : Scatter plots are are very effective when there is a strong correlation between variables

Cons : Time taking – 450,000 lines of data

Due to the large amount of data the graph is completely filled

TITLE: DESIGN SHEET 2
AUTHOR: SARAT KARASALA
DATE: 07/06/2020

Sheet 3



Sheet 3: Design

Layout : Bar Plot

Control Panel

Radio buttons

● Day of Week
○ Month
○ Time of Day

Select box

Choice of Day, Month or Hour ▾

Pick up and drop off averages can be stacked together

Operations : Filter the data to see the average tip offered in each borough for a given day of week, month or time of day

Pick one from the radio button list and choose the specific value in the selection box below. For example I want to see the tips offered on in December.

Focus

Aggregate methods: The Bar plot will show the "average" tip offered in each borough for the given selection of filters

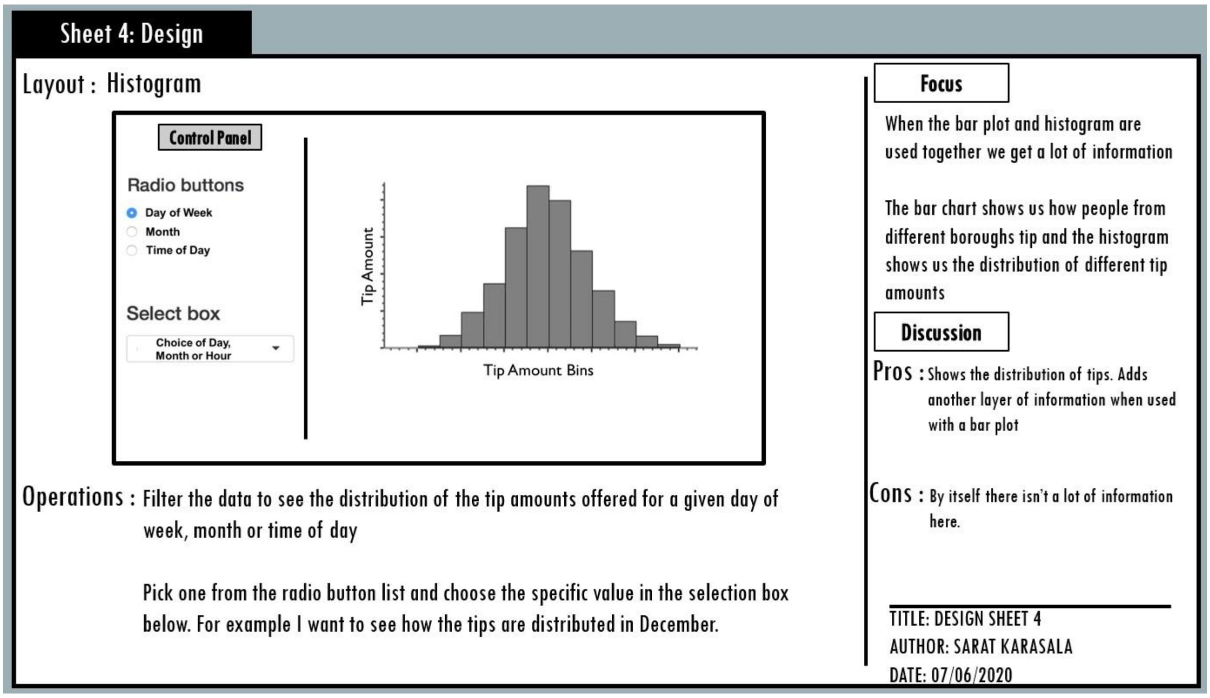
Discussion

Pros : Using aggregation functions makes the visualization more meaningful and doesn't feel cluttered due to too much data

Cons : Bar Plot can't handle continuous data so we cannot use trip time and trip distance without splitting them into bins

TITLE: DESIGN SHEET 3
AUTHOR: SARAT KARASALA
DATE: 07/06/2020

Sheet 4



Sheet 4: Design

Layout : Histogram

Control Panel

Radio buttons

● Day of Week
○ Month
○ Time of Day

Select box

Choice of Day, Month or Hour ▾

Operations : Filter the data to see the distribution of the tip amounts offered for a given day of week, month or time of day

Pick one from the radio button list and choose the specific value in the selection box below. For example I want to see how the tips are distributed in December.

Focus

When the bar plot and histogram are used together we get a lot of information

The bar chart shows us how people from different boroughs tip and the histogram shows us the distribution of different tip amounts

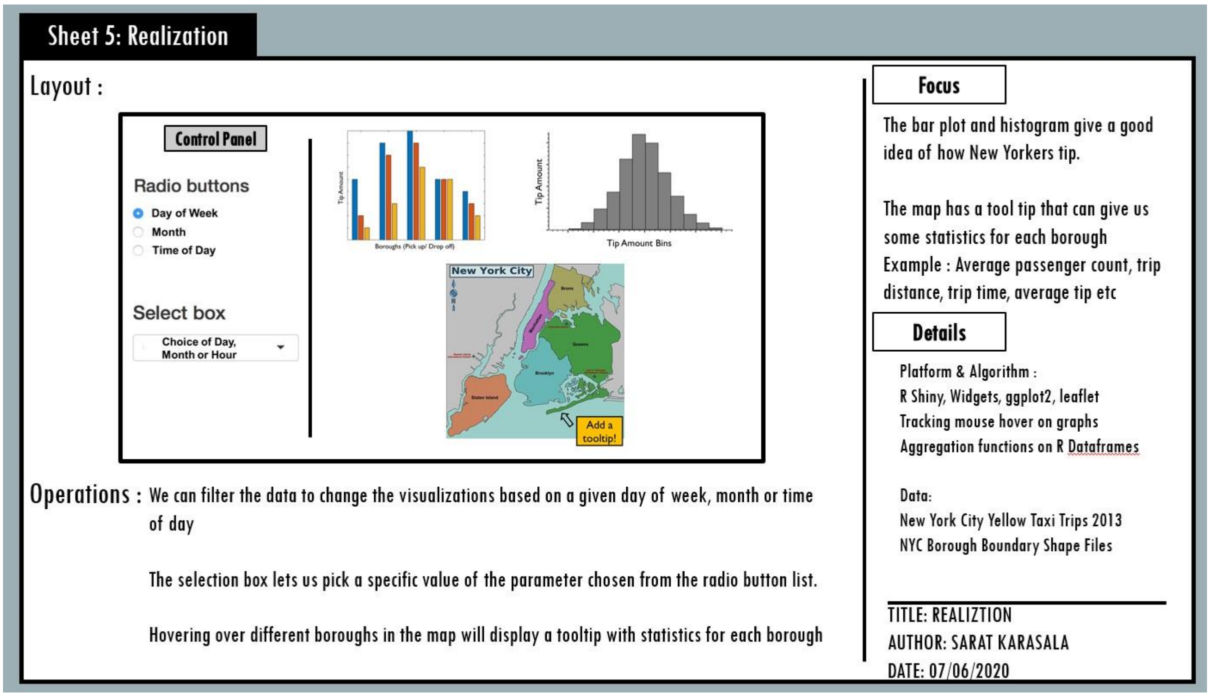
Discussion

Pros : Shows the distribution of tips. Adds another layer of information when used with a bar plot

Cons : By itself there isn't a lot of information here.

TITLE: DESIGN SHEET 4
AUTHOR: SARAT KARASALA
DATE: 07/06/2020

Sheet 5



Sheet 5: Realization

Layout :

Operations : We can filter the data to change the visualizations based on a given day of week, month or time of day

The selection box lets us pick a specific value of the parameter chosen from the radio button list.

Hovering over different boroughs in the map will display a tooltip with statistics for each borough

Focus

The bar plot and histogram give a good idea of how New Yorkers tip.

The map has a tool tip that can give us some statistics for each borough
Example : Average passenger count, trip distance, trip time, average tip etc

Details

Platform & Algorithm :
R Shiny, Widgets, ggplot2, leaflet
Tracking mouse hover on graphs
Aggregation functions on R Dataframes

Data:
New York City Yellow Taxi Trips 2013
NYC Borough Boundary Shape Files

TITLE: REALIZTION
AUTHOR: SARAT KARASALA
DATE: 07/06/2020