2020

GEOVISUALIZATION AND ADVANCED CARTOGRAPHY

ASSIGNMENT 2: GEOVISUALIZATION

HIRA ZAFAR-EMJMD CDE-INTAKE2

Table of Contents

Abstract

In this report, I will be trying to shed light on some of the final data visualizations of the chosen Blue bikes dataset. The report details each visualization technique used to display information about the dataset, data analysis and correlation among different attributes.

Dataset

Bluebikes is a public bike share system in Arlington, Boston, Brookline, Cambridge, Chelsea, Everett, Newton, Revere, Somerville, and Watertown. The company publish downloadable files of Bluebikes trip data each quarter. The dataset I selected for this assignment is of Feb 2020. The dataset consisted of 133236 records.

Source: Blue Bikes website

Website: https://www.bluebikes.com/system-data

Dataset variables:

- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station Name & ID
- End Station Name & ID
- Start Station latitude & longitude
- End Station latitude & longitude
- Bike ID
- User Type (Casual = Single Trip or Day Pass user; Member = Annual or Monthly Member)
- Birth Year
- Gender, self-reported by member (Zero=unknown; 1=male; 2=female)

Scope

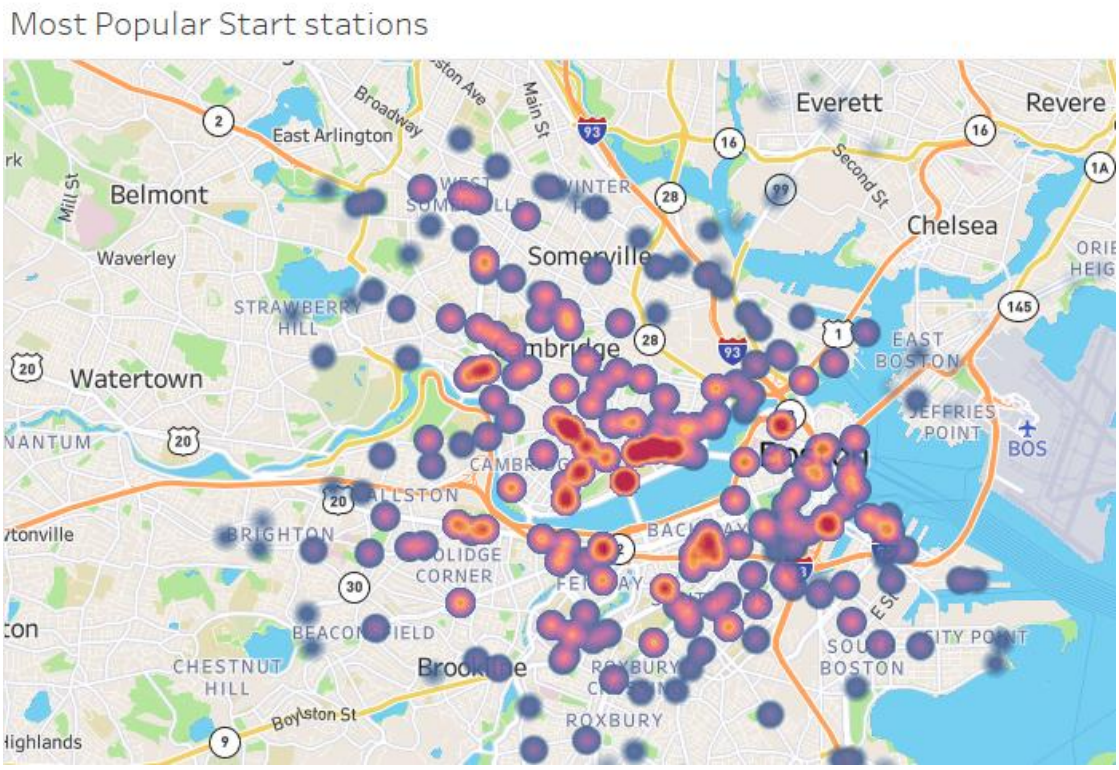
The task was to fulfill the objective of visualizing a dataset through the best visualization techniques. Since both mapping and geographical data is available in this dataset, multiple graphs that visualize the dataset statistically and geographically were created. The software used for this purpose is Tableau which provided a sufficient platform for our objective. Those multiple visualizations provide answers to the following questions in the clearest and accurate way:

- Which stations are most popular?
- How does the trip duration change by age?
- On what days of the week are most rides taken?
- How does the trip duration change by age?
- What are the peak hours on which bikes are used on different days of the week?
- What are the destinations of trips originating from same start point? (Origin-destination maps)
- Most popular trips of the month?

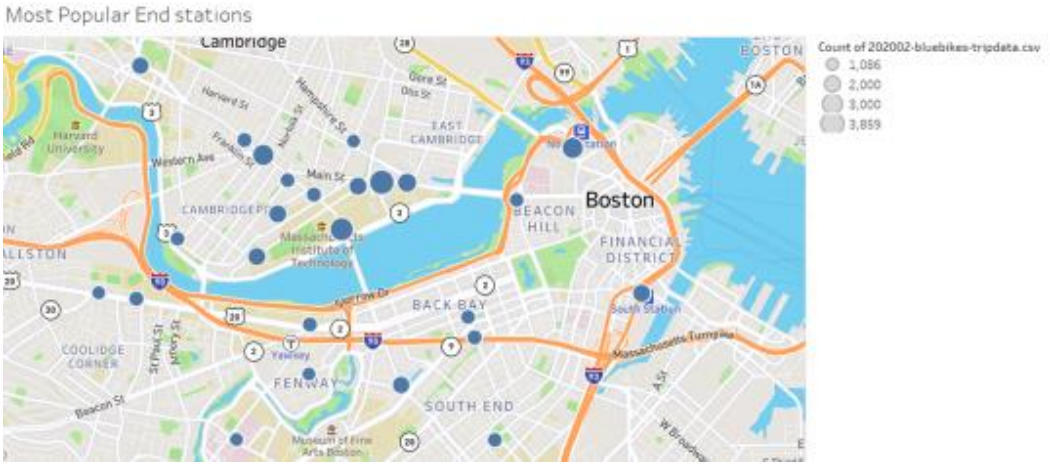
Final Visualizations

Most popular Start/End stations:

For this visualization Latitude and longitude which were geographic attributes of the data were mapped. Number of trips that originated and ended at those stations were considered for the popularity of the stations.



This is a density map based on start station latitude and longitude, colors show count of trips that originated from these stations, darker colors showing more number of trips.



Map based on End station latitude and longitude, size show count of trips that ended at these stations, darker, greater the size greater the number of trips.

Analysis

Upon analyzing the visualizations, the viewer is able to discern following piece of information:

- Top three start stations are:

Name	Number of trips
MIT at Mass Ave	3744
MIT Stata center	2958
Central square	2921

- Top three end stations are:

Name	Number of trips
Ames St at Main St	3859
MIT at Mass Ave	3425
Central square	2963

- Popular start and end stations are located at almost same places of the city.
- Most of the riders commute to and from the center of the city.
- The reason for the popularity of these stations is because popular points of interest that is MIT University, Harvard University and Central Square which is known for its wide variety of ethnic restaurants, churches, bars, and live music and theatre venues.
- As we move away from the city center the number of daily trips decreases.

Upon comparing the two types of visualizations used for showing popularity of stations, it can be inferred that type of visualization is very important for conveying the information to the viewer. Though both of the visualization type convey the popularity of stations one by the density and the other by the size of the circle, but the density map conveys the information in the very first look that which most popular stations exist at these places, on the other hand one has to focus on the sizes of circle to find most popular stations.

Trip duration change by age and gender

For this visualization Age was drawn on x-axis, total trip duration was displayed on y-axis and gender was used as a color classifier in tableau. The age attribute was not present in the data, only a birth year attribute was present. I calculated an Age column to make visuals easier to interpret using calculated field feature of tableau and the following formula:

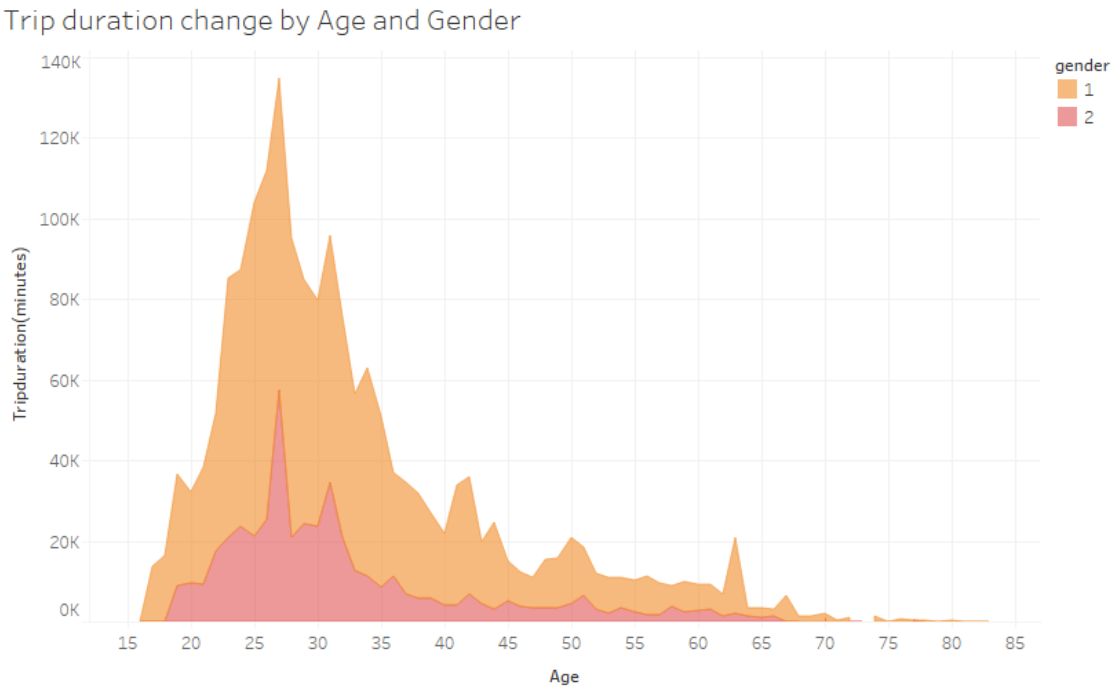
INT (DATEDIFF ('day', [birth year], today ()) / 365.25)

For better results we excluded those entries from this visualization whose gender was unknown.

Data Anomalies:

For Birth Year, there are some people born prior to 1960. I can believe some 60 year olds can ride a bike and that's a stretch, however, anyone "born" prior to that riding a Blue Bike is an anomaly and false data. There could be a few senior citizens riding a bike, but probably not likely.

My approach was to identify the age and after calculating this number, I removed the tail end of the data, age greater than 85.



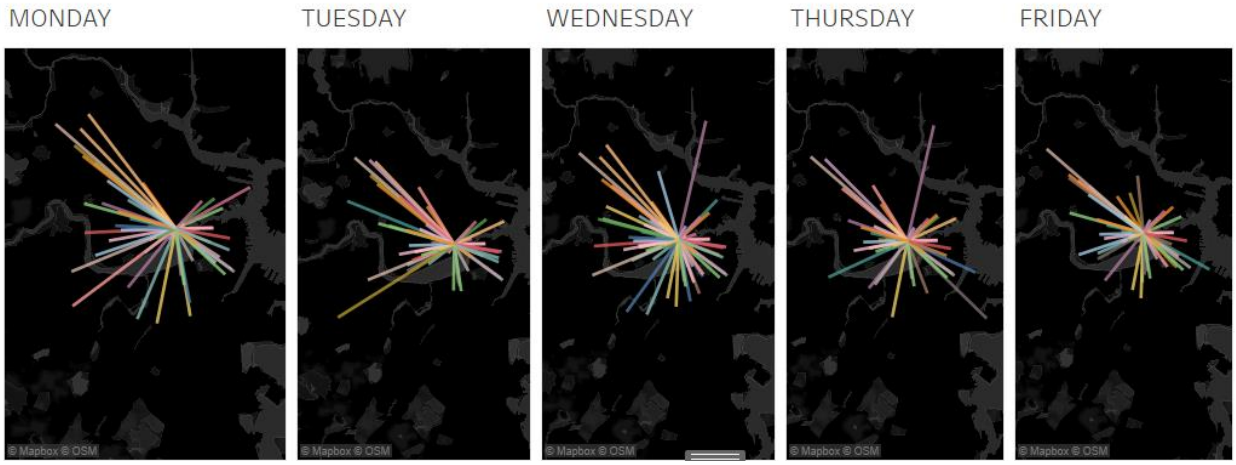
Analysis:

The visualization shows that trip duration of male riders is almost three times greater than that of female candidates. Trip duration changes by age trend is same for male and females. Most of the riders start from late teen age and trip duration kept on increasing untill their late twenties. Both male and female riders between the age 25 and 30 covered the longest trips. The trip duration starts declining from late thirties and reach the lowest point when the riders were in late 60's and above. This shows a strong correlation between age and trip duration variables.

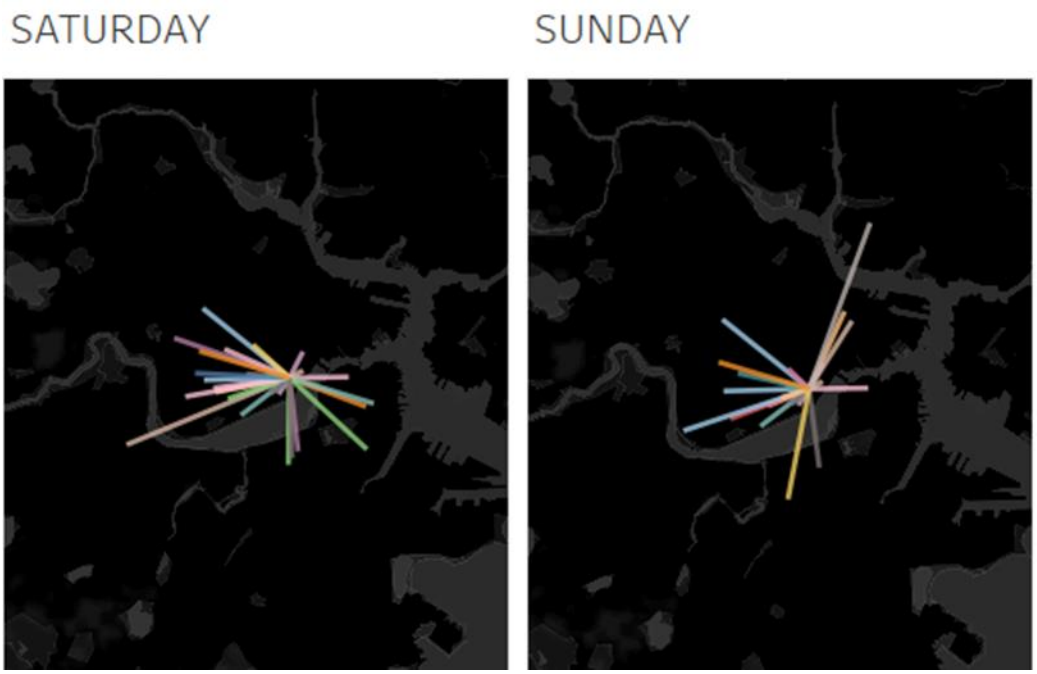
Origin-Destination Map

For this visualization a map was made to display information about where do blue bikes riders go from different start stations on different days of the week. The segments drawn on the map are pointed at destinations of trips from a single start station. Filters of start station and weekday were applied for this visualization and trend analysis of trips on different days of the week. The weekday attribute was not present in the data; it was calculated using formula: **DATENAME("weekday", [Starttime])**

Looking horizontally the maps show bikes usage on different days of the weekday.



Blue Bike trips on Weekdays from 75 Binney St station



A key design decision while making the above visualization was how to display the trip segments. Showing all start stations and their trip segments in one map was a jumbled mess that yielded

minimal useful information. An attempt was made to include all segments and thinning the line size but that still was very cluttered. The line segments were classified based on path names. Another attempt that I made at initially was to show number of trips to a destination by scaling the line thickness but this again resulted in clutter and I did not want a small number of stations to dominate the visualization. It is very important to intelligently think and include only those variables in the visualization that serve the purpose, too much information in one visualization sometimes ruin the main objective.

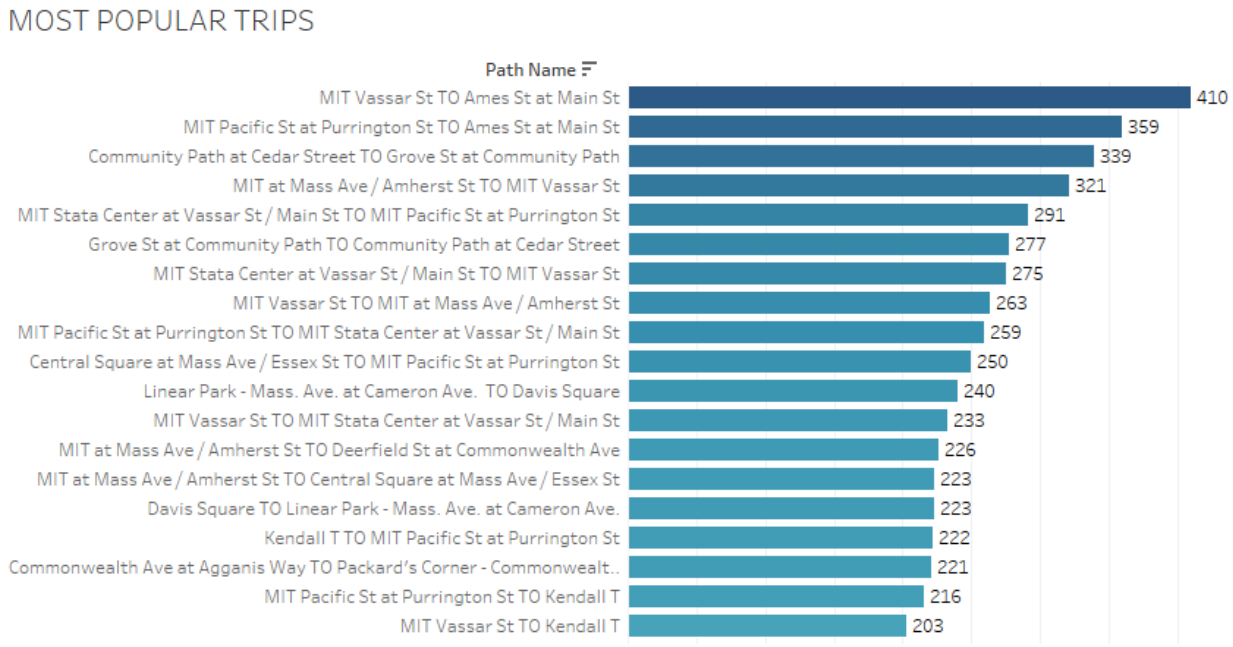
Analysis:

Some of the obvious information that can be inferred from Data analysis of above visualization are:

- Weekdays are the busiest days on which blue bikes are used as compared to weekend.
- The origin destination paths show that on weekdays most of the trips are destined at same places however this trend changes on weekend days. It can be seen that weekend trips are destined at different and far off places compared to weekday trips.
- On weekdays most of the trips are destined at Harvard university, MIT, public library and playground.
- Weekend is highly concentrated along the northern side, beach points, hill stations, parks, gardens, museum, riverside and other tourist locations.

Most Popular Trips

To find the most popular trips of Feb 2020, the most convenient way was to group the data by their paths and total number of the trips and tableau do it very efficiently and intelligently. The number of trips were displayed on x-axis and paths were displayed on y-axis.

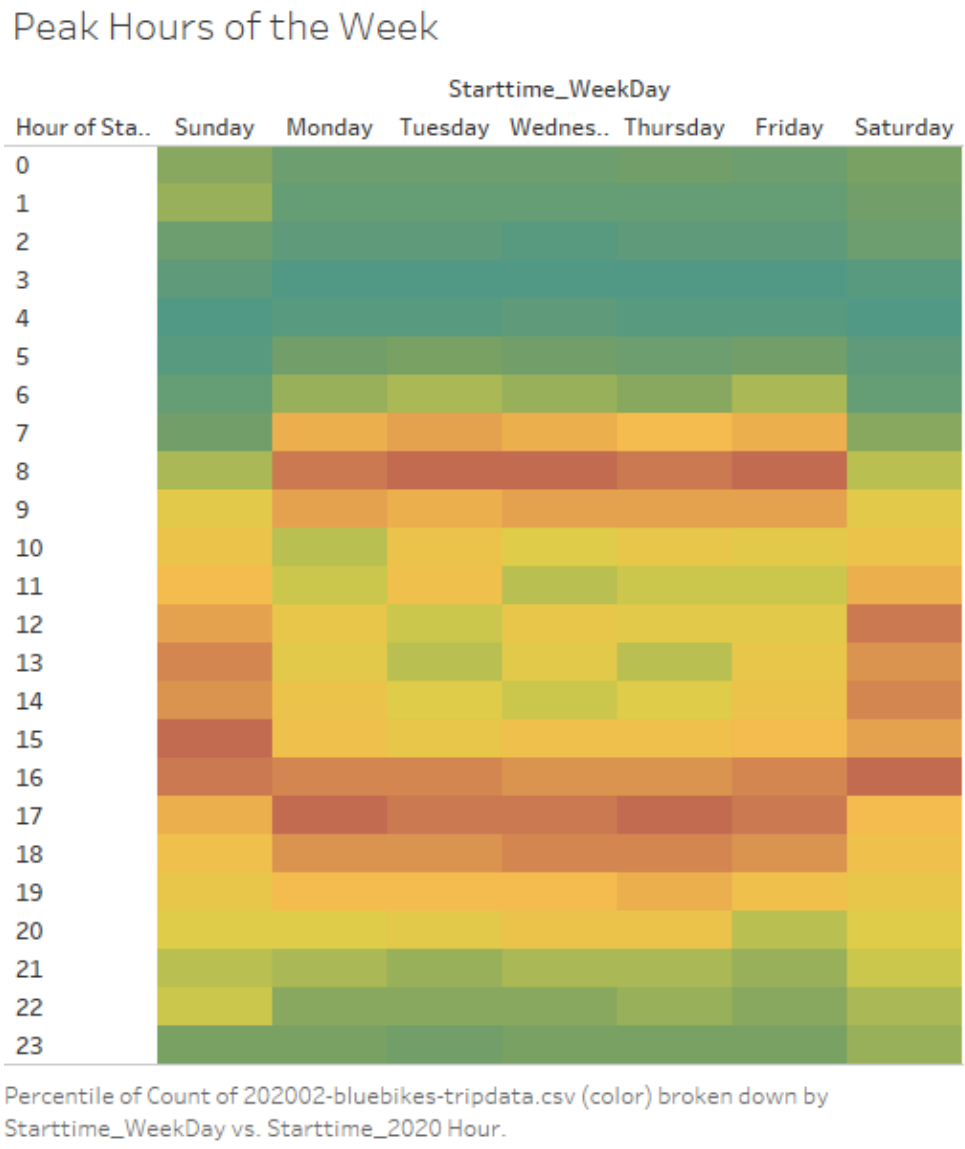


Analysis:

The visualization shows that most popular trips either started from MIT and were destined at Main St station or from Central square and ended that MIT. In general, most of the trips were in the neighborhood of MIT and central square. This can help Blue Bike determine where to keep bikes and where to have docks available based on the time of day. The assumption being riders travel in one direction to work in the morning and in the opposite direction to get home in the evening. It seems that more people use the bikes to return from work instead of going to work. Probably, people avoid arriving to work sweating and tired and therefore they prefer using other mode of transportation.

Blue Bike Peak/Rush hours

This visualization highlights the utilization time of the Blue Bikes. The heatmap below plots the "hours of the day" in the y-axis and "days of the week" in x-axis. The count of the bikes is represented through the heatmap matrix. Colors of the heatmap varies between: (Green—Yellow—Red) in response to the bikes "count" levels which I believe clearly draw our eyes to the peak hours plotted in the temperature diverging colors of the heatmap. The design consideration for this visualization were tricky including colors, data cluster, scale and axis.



Data Analysis:

The viewer can discern following pieces of information from the above visualization:

- Usage is much higher during the week

- Weekday usage has a bimodal distribution with peaks during morning and evening commuting times.
- Weekend usage has a unimodal distribution centered in the early afternoon till night.
- More commuters use the bike in the evening than the morning.
- It appears that the bikes' highest demand during the weekdays move along the rush hours (7:00 –8:00 and 16:00—18:00)
- On weekends the bike demand trend changes and most of demand is during 11:00—16:00.
- Peak hours trend is same throughout the weekdays.
- There are almost no trips from 23:00 –5:00.

Conclusion:

Data analysis is very fast with Tableau and the visualizations created are in the form of dashboards and worksheets. The data that is created using Tableau can be understood by professional at any level in an organization. It even allows a non-technical user to create a customized dashboard. The great thing about Tableau software is that it doesn't require any technical or any kind of programming skills to operate.

The best features of Tableau are

- Excellent, mind-blowing variety of visualization wizards
- Data Blending
- Real time analysis
- Variety of data sources supported by its connectors, from plain file, Excel, RDBMS, and Hadoop to NoSQL databases.
- Creation of dashboards, charts, report generation, visualization

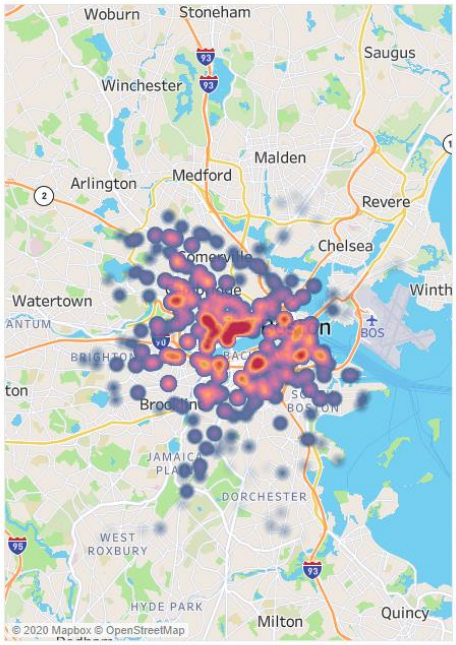
Following were the weakness of the software which I observed:

- Padding and working with null or missing values.
- Slow rendering when the data volume is huge.

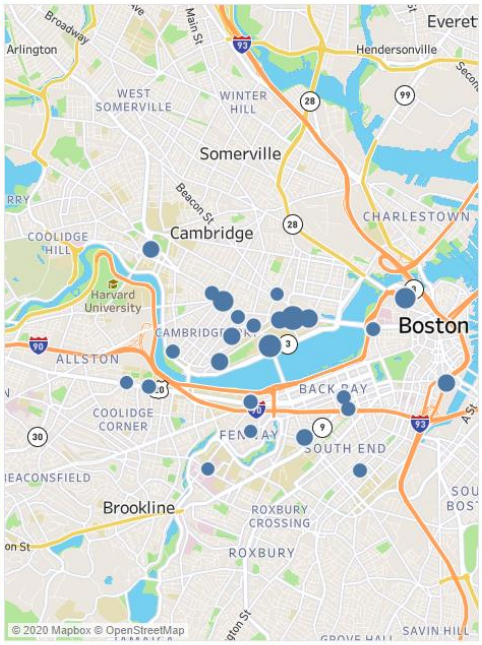
Dashboard:

I created the following dashboard, which is a very good feature of tableau that allows you to arrange individual worksheets in the form of dashboard and share it with public. I tried sharing it with public using tableau online but it gave me error **Not authorized,** I assume that happened because I was using tableau desktop with my student account and not tableau public and desktop version doesn't allow you to publish your data and visualizations to public. I didn't use tableau public because it does not allow you to save your workbooks locally.

Most Popular Start stations



Most Popular End stations



Count of 202002-bluebike..
To Null

Start_time_weekday
- Sunday
- Monday
- Tuesday
- Wednesday
- Thursday
- Friday
- Saturday

Start Station Name
75 Binney St

Count of 202002-bluebike..
From 1,072

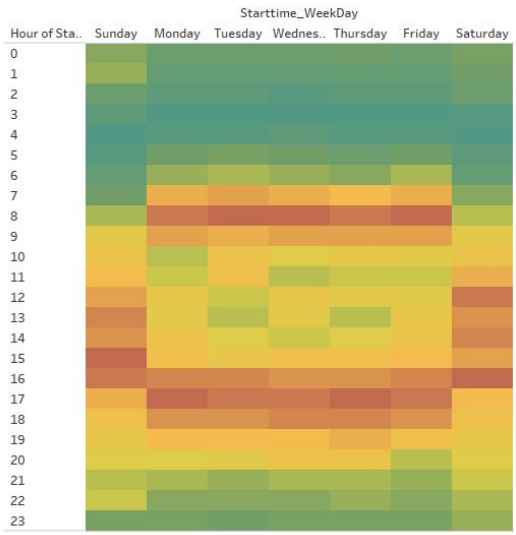
Count of 202002-bluebike..
- 1,086
- 2,000
- 3,000
- 3,859

Percentile of Count of 20..
0.0% 100.0%

gender
- 1
- 2

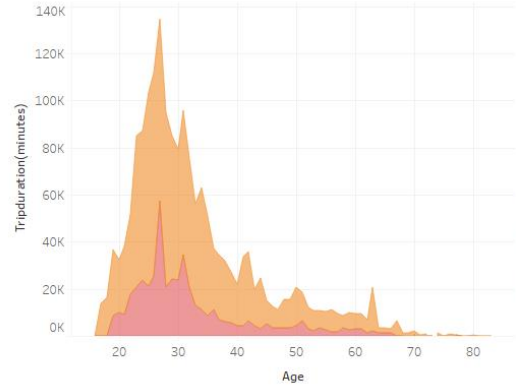
Count of 202002-bluebike..
1 410

Peak Hours of the Week



Origin Destination Maps

Trip duration change by Age



MOST POPULAR TRIPS