New York taxi data Analysis

Group 11

Pratik Gandhi

Rohit Shetty

Sankalp Sharma

Yan Yan

# Data Sources:

## Primary dataset:

<http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml>.

## This dataset includes trip records from all trips completed in yellow and green taxis in NYC for the 6 month period Jan - June 2015. Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

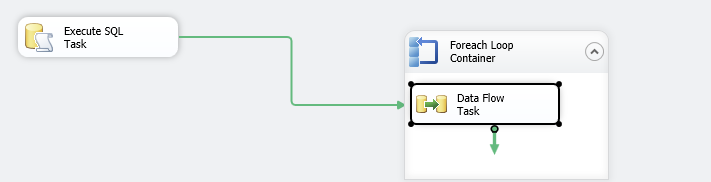
## Secondary dataset:

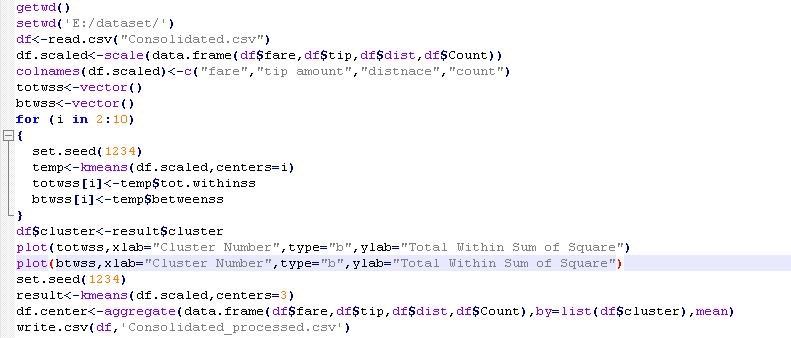
<https://nycopendata.socrata.com/widgets/uppf-z66u>

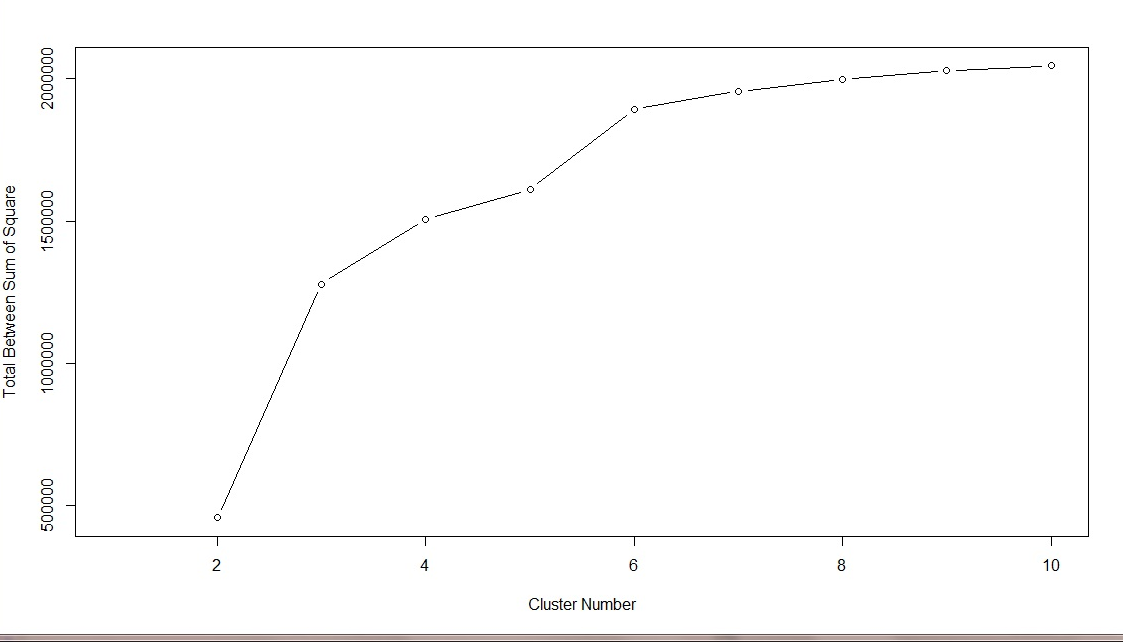
## This dataset contains New York taxi complaint data capturing fields such as incident date, incident address, type of complaint, city, Pick up location, latitude and longitudes for the period

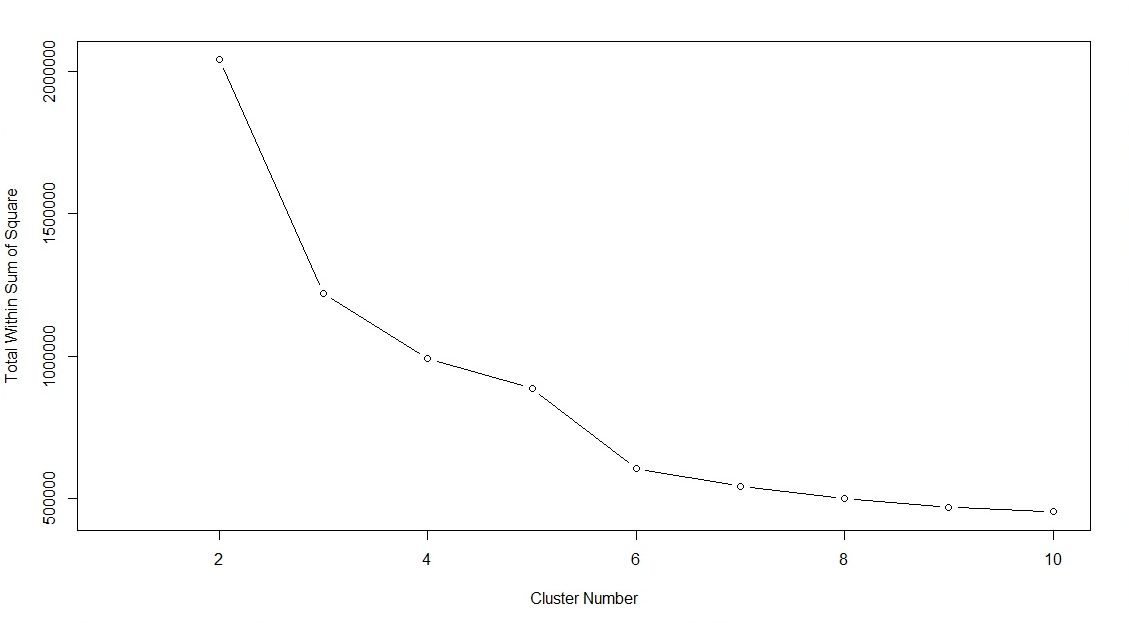
# Clustering and Pre-Processing:

## Since the total size of the dataset was more than 7 GB, we aggregated it using SSIS and exported it a new dataset. We aggregated the data based on unique pick locations and created clusters based on the fare amount, tip, count and trip distance in R. We added a new column in the dataset for cluster number to create a new dataset called “Consolidated\_Processed.csv









Following are the two hypothesis on which the analysis is being conducted

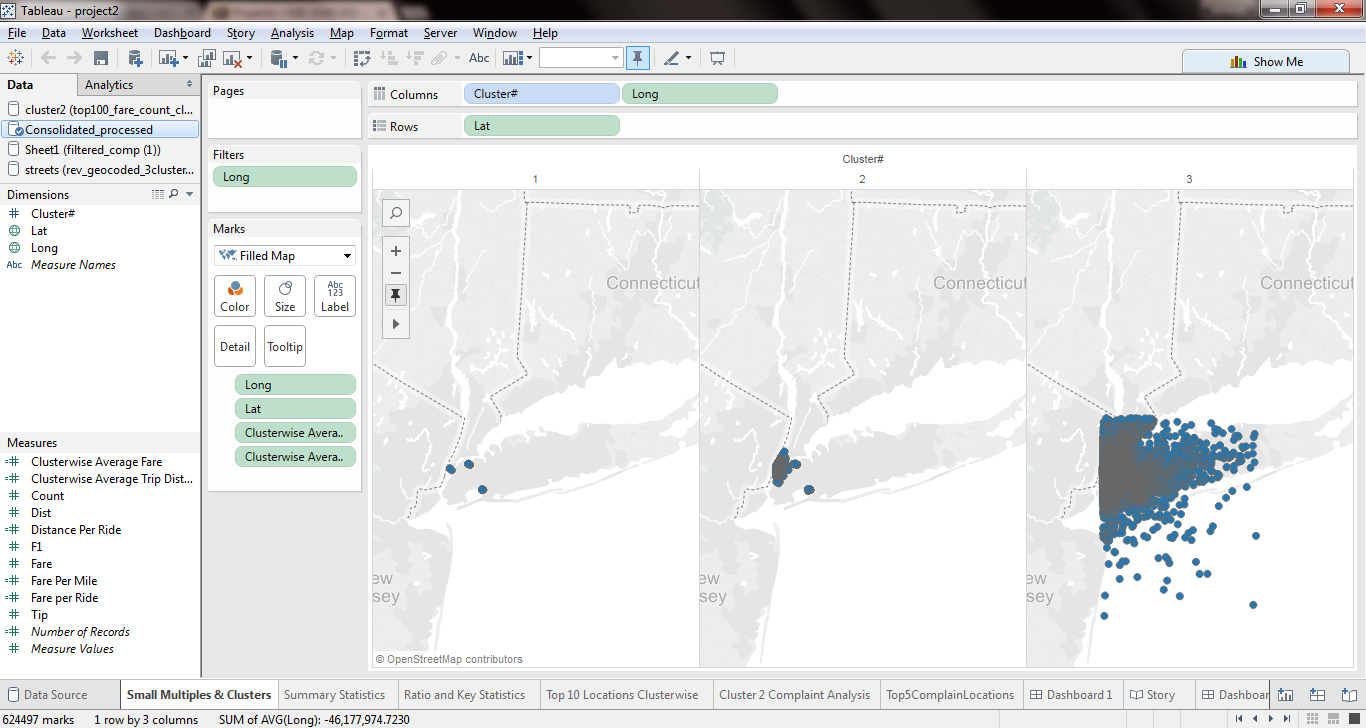
1. High Number of Rides lead to high revenue
2. High Revenue Area lead to high complaints

Creating Story

Establishment-

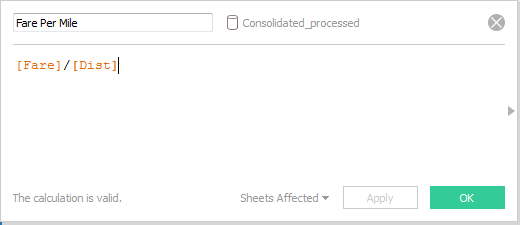
The analysis started off with understanding the summary values of key statistics wherein the sum of Fare and Count was analyzed for each cluster. Primary dataset was divided based on three clusters which was concluded from the graph of k-means **clustering** using “within sum of square and between sum of square” distance

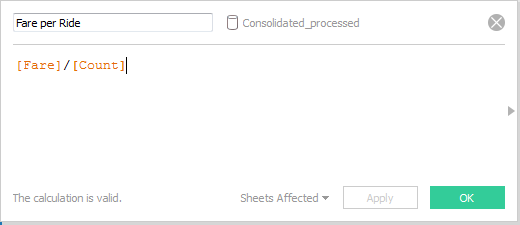
1. A further drill down was conducted on cluster level to analyze the pickup locations and find insights of the type of fare that are obtained from those pickup locations.

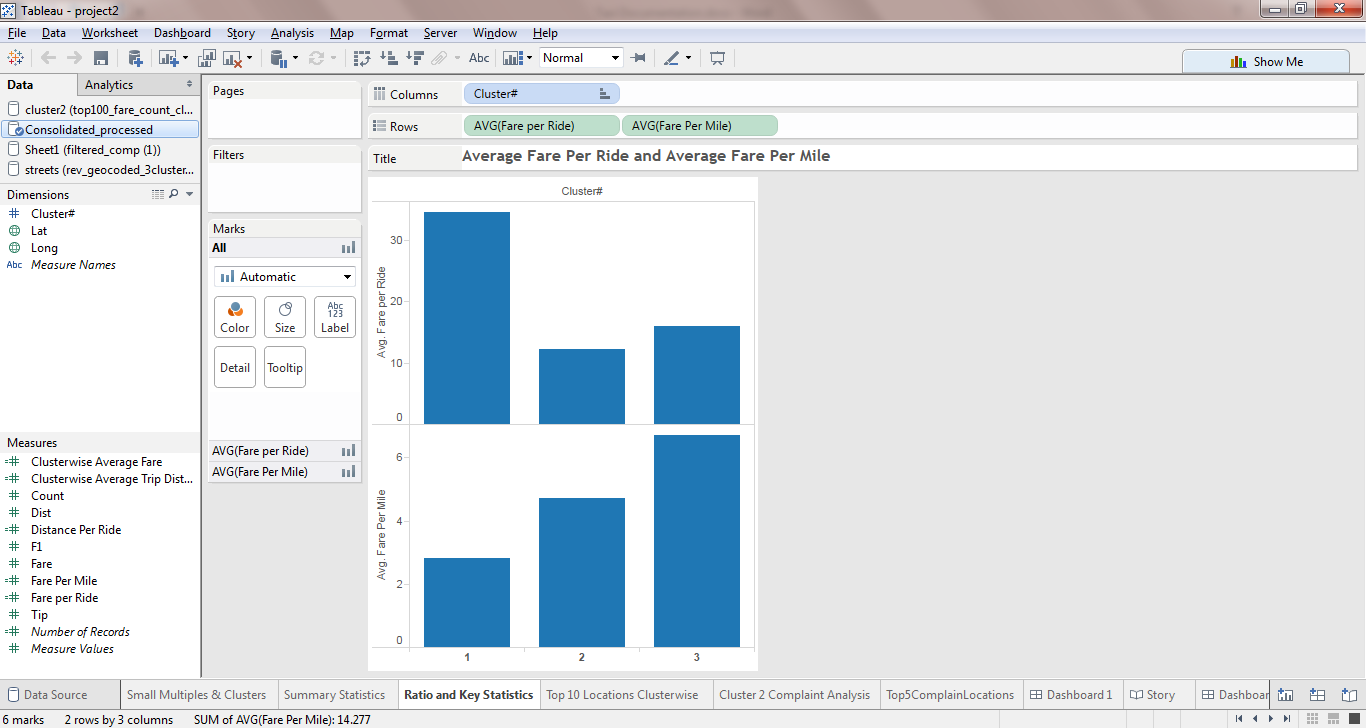


Cluster 2 is generating highest revenue but this cluster gets low fare rides whereas Cluster 1 is generating low fare but gets high fare rides. Also cluster 2 has got more trip distance when compared with cluster 1. Thus, cluster 2 although generates more revenue but is generated from low fare rides.

1. **Key ratios** were then studied to understand the characteristics of the clusters based on Average Fare per Mile and Average Fare per Ride.



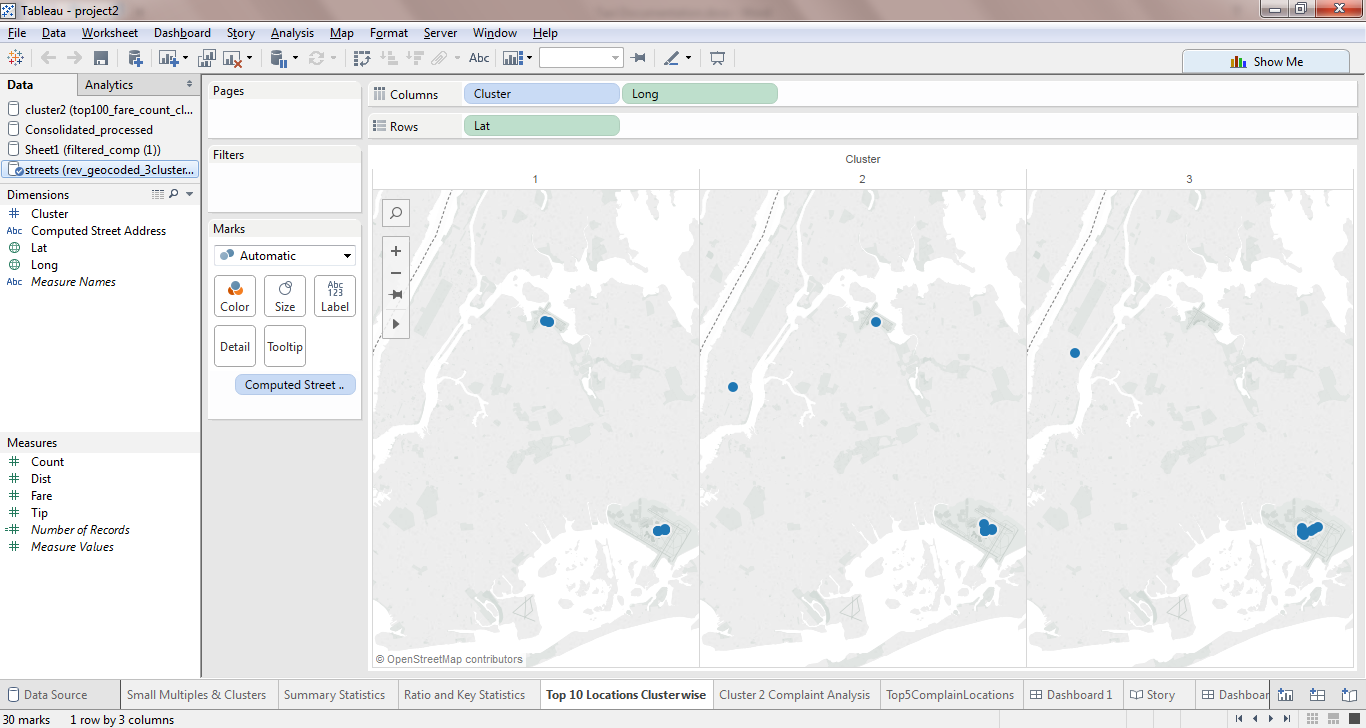




Cluster 1 generates the highest revenue per ride. Thus, we can conclude that these pick locations generates rides with highest earning potential though the total number of rides is considerably low in this cluster compared to other clusters.

In the graph, you can see that Cluster 3 has highest fare per mile distance followed by Cluster 2 and Cluster 3 in that order. Thus, Cluster 3 contains the most lucrative pick-up locations in terms of mileage.

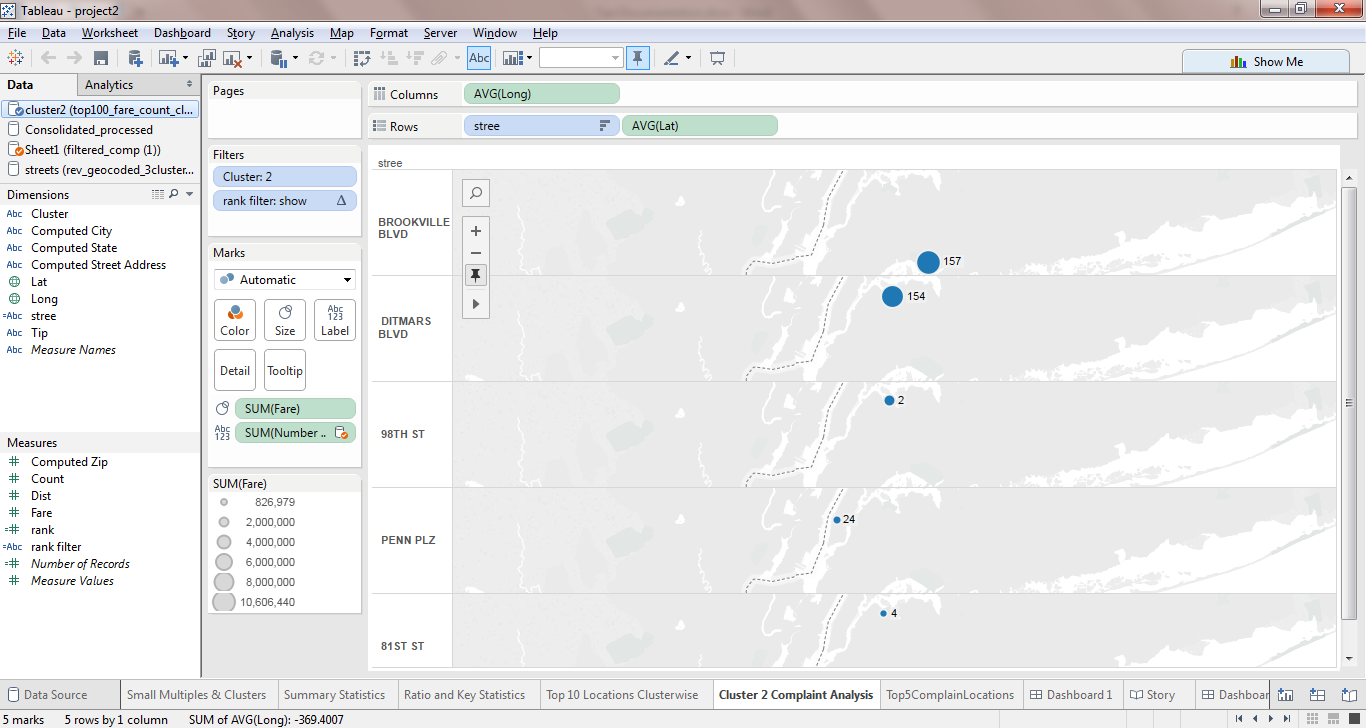
1. Top 10 pickup locations of each clusters are then visualized.



15468 Brooklyn Blvd has got pickup locations that generate high average fare per ride and low average fare per ride.

Tension-

1. Cluster 2 has highest fare so we focused on complaints related to pickup locations for cluster 2

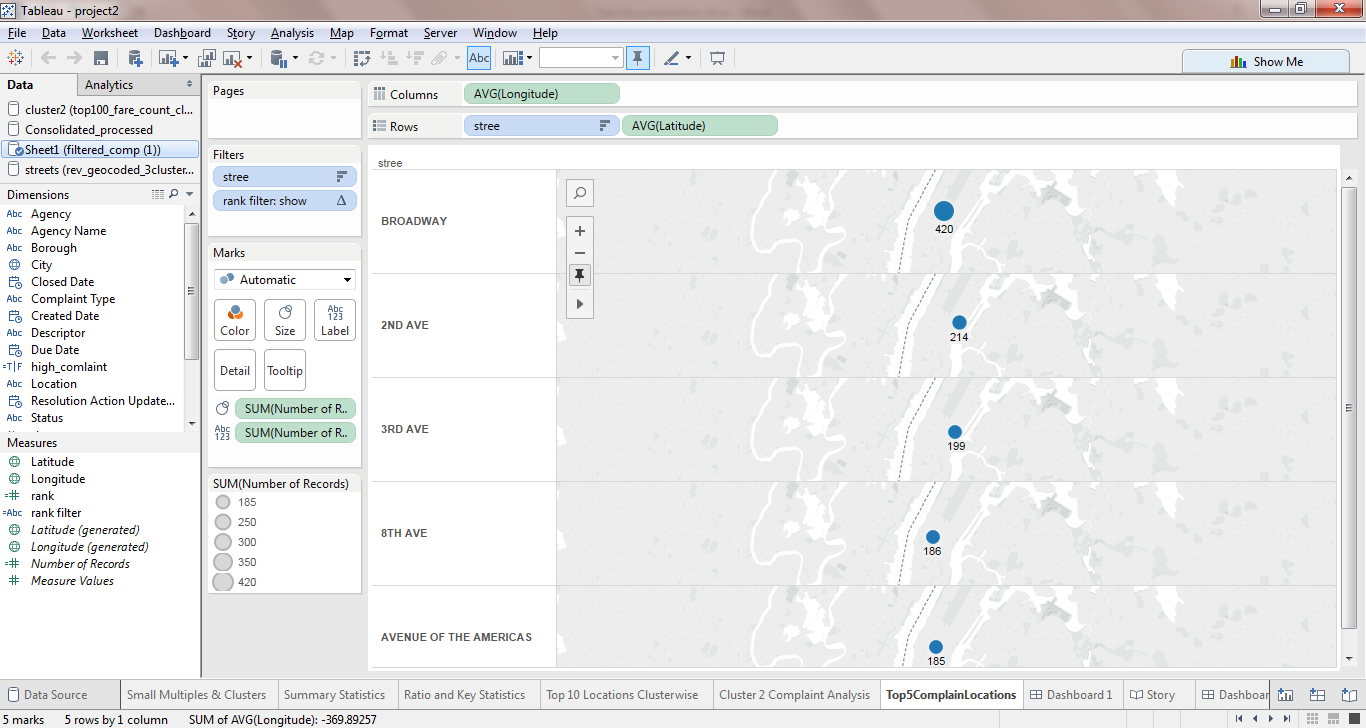


Size of circles represents fare and labels in it represents total number of complaints for that pickup location.

Brooklyn has got pickup locations which generate all types of fare but has got more of complaints.

Climax

1. The above visualized locations were then cross verified with the locations generating higher number of complaints irrespective of the cluster.



1. Our second hypothesis was thus rejected (high revenue generating area leads to higher complaints), as cluster 2 members were not amongst the top 5 complaint locations.

Insights

* Pickup locations can be broadly categorized into high average fare, medium and low average fare.
* High average fare pickup locations generate low total revenue
* Low average fare pickup locations generate high total revenue
* Brookville Blvd provides pickup locations for small and high fare
* Top locations generating high revenue area does not have high complaints

Recommendation

* Increase number of rides in locations in cluster 1 (areas generating more average fare per ride)
* Improve services in areas such as Broadway which is a major tourist hub