# Option B: Visualization

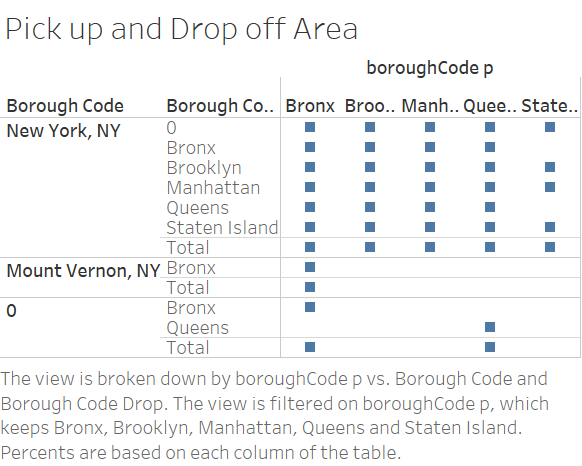
# Can you build a visualization (interactive or static) of the trip data that helps us understand intra- vs. inter-borough traffic? What story does it tell about how New Yorkers use their green taxis?

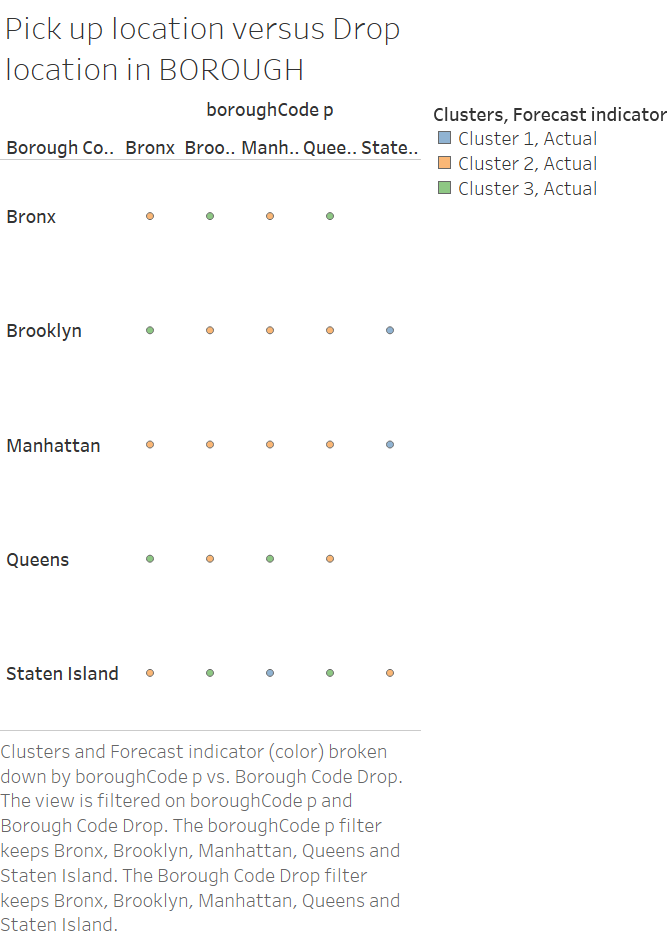
## Approach:

**The Data extracted in a CSV file and then used in Tableau to generate graphs and plots to dig out inferences.**

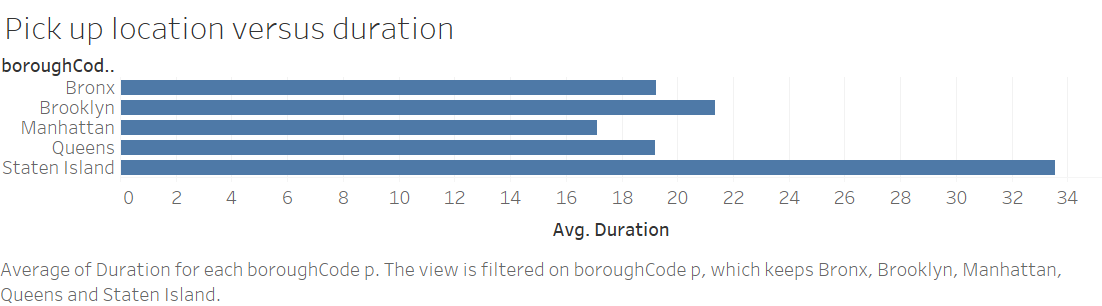
**Inter-borough - ‘Manhattan', 'Bronx', 'Brooklyn', 'Queens', 'Staten Island'**

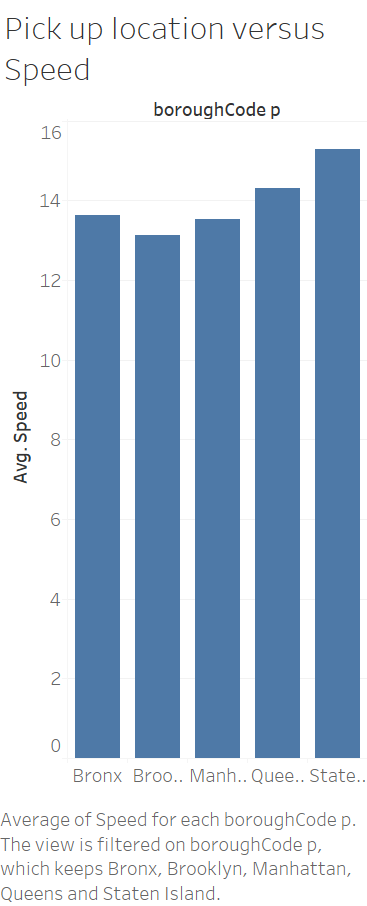
**The first step is to look into the two segment intra- vs. inter-borough. The chart signifies the pickup and drop off location for inter-borough. Only 'Staten Island location has no drop off location as Bronx and Queens for the green taxi trips. Rest all the 5 location has same inter pickup and drop off area.**

****

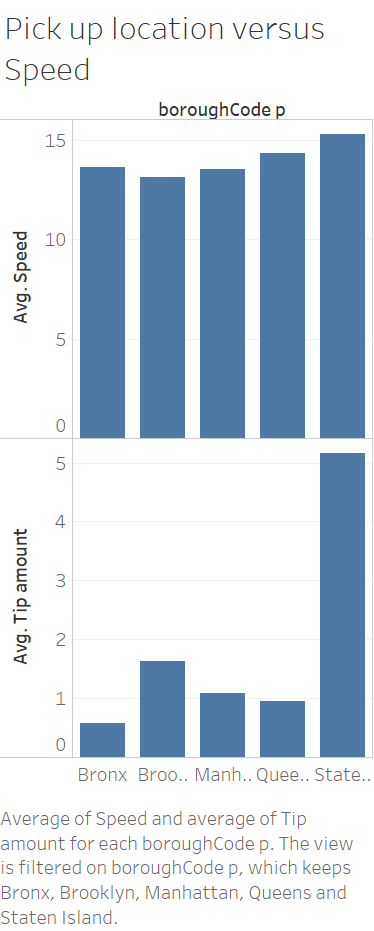
****

**The Staten Island takes the long trips as the duration is highest for it.**

****

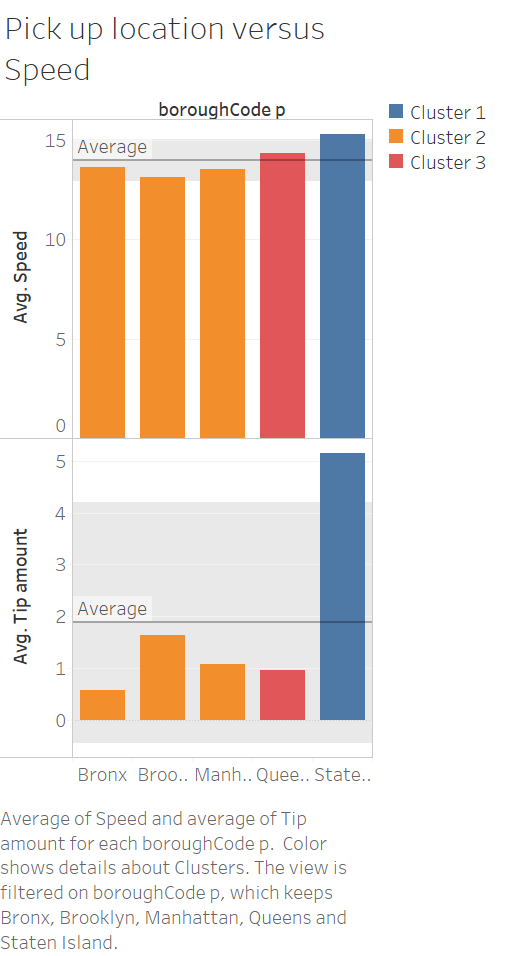
****

**The speed is almost same for all the 5 locations. Only in Staten Island the driver drives with highest speed as compared to other 5 locations. Hence we can infer that the duration is long hence the speed is highest.**

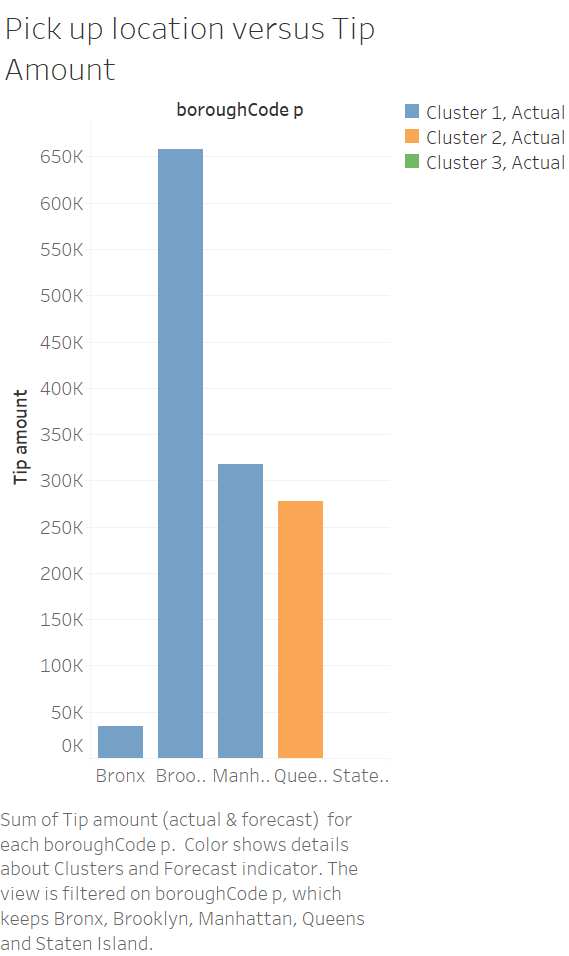
****

**The plot showcases an important insight that the tip amount given in Staten Island is much higher than the other location. Even the duration and speed is highest in this particular area.**

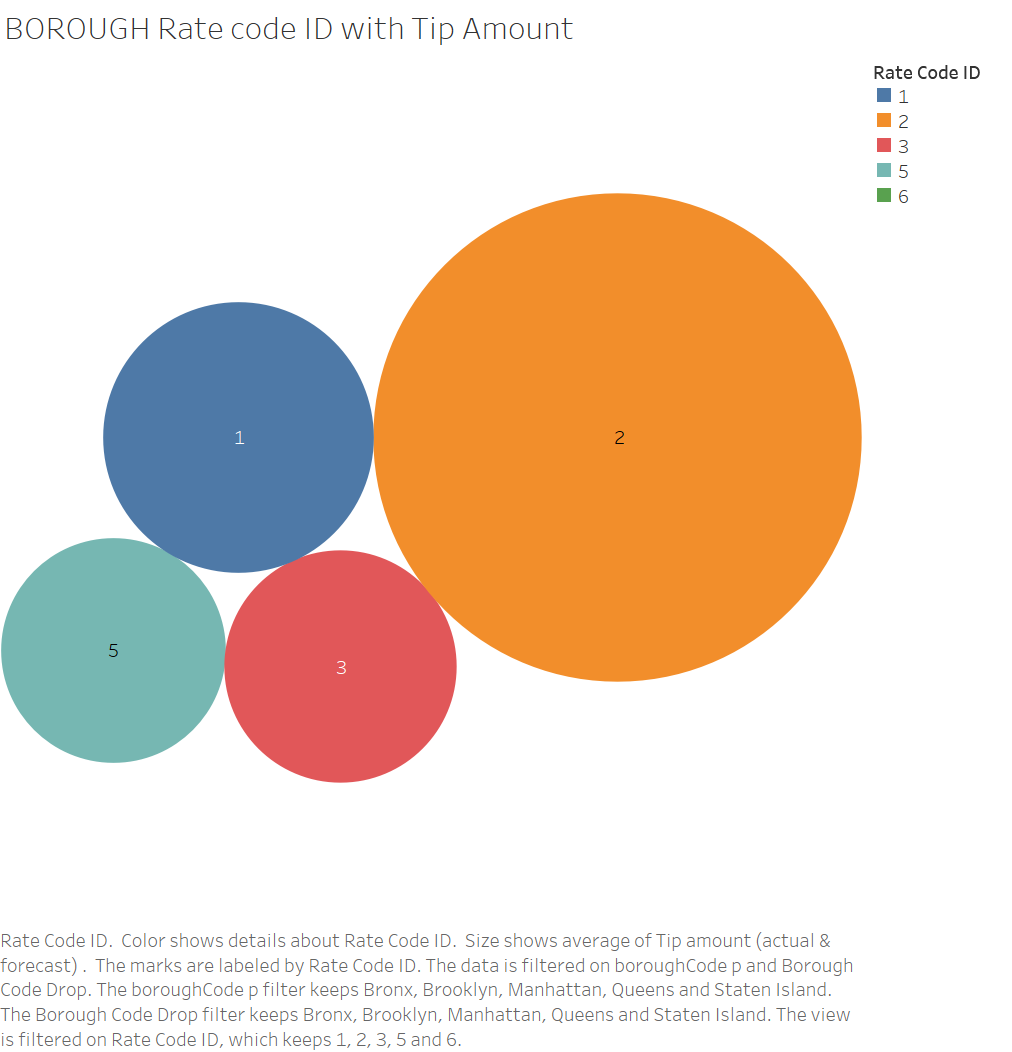
**The taxi driver will get extra earning in this area “Staten Island” but needs to drive for longer duration.**

****

**The data is divided into 3 clusters. This also highlights that in 'Brooklyn', the average speed is lowest but the driver earns a substantial amount of tip. It can infer that the passenger who travel in 'Brooklyn' are happy with low speed drive hence they give tips to the driver.**

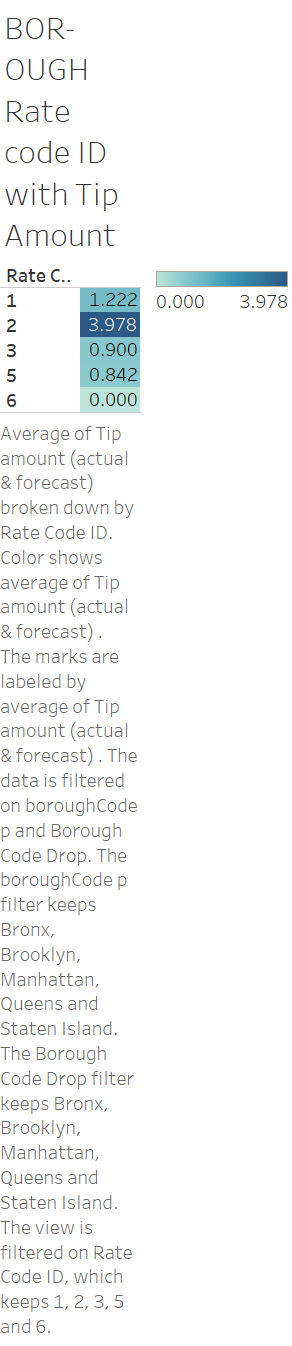
****

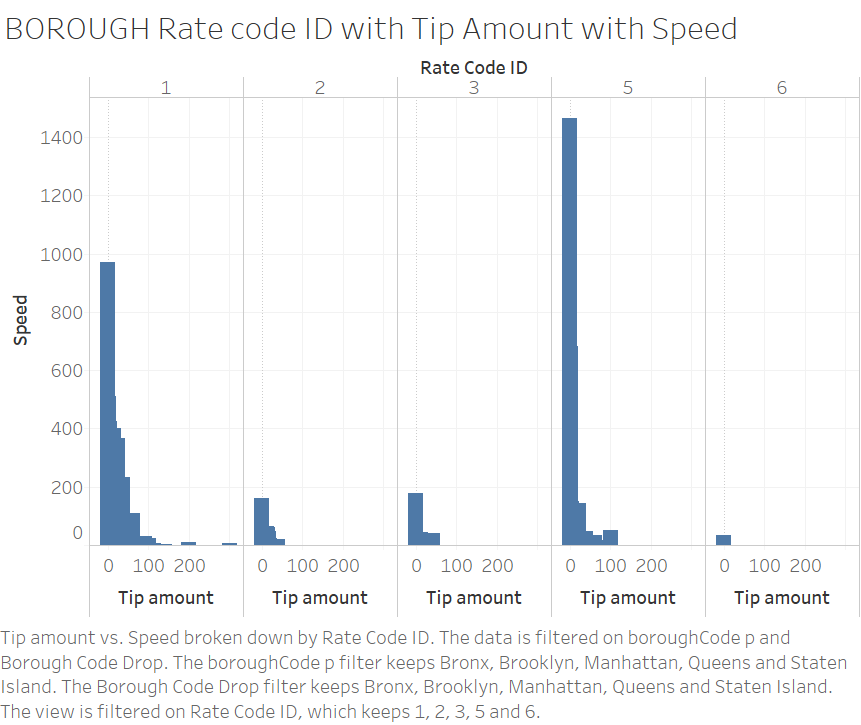
**The drivers in 'Brooklyn' earn a good amount of tip. Even they drive for a duration somewhat similar to other location except Staten Island with low Average speed.**

****

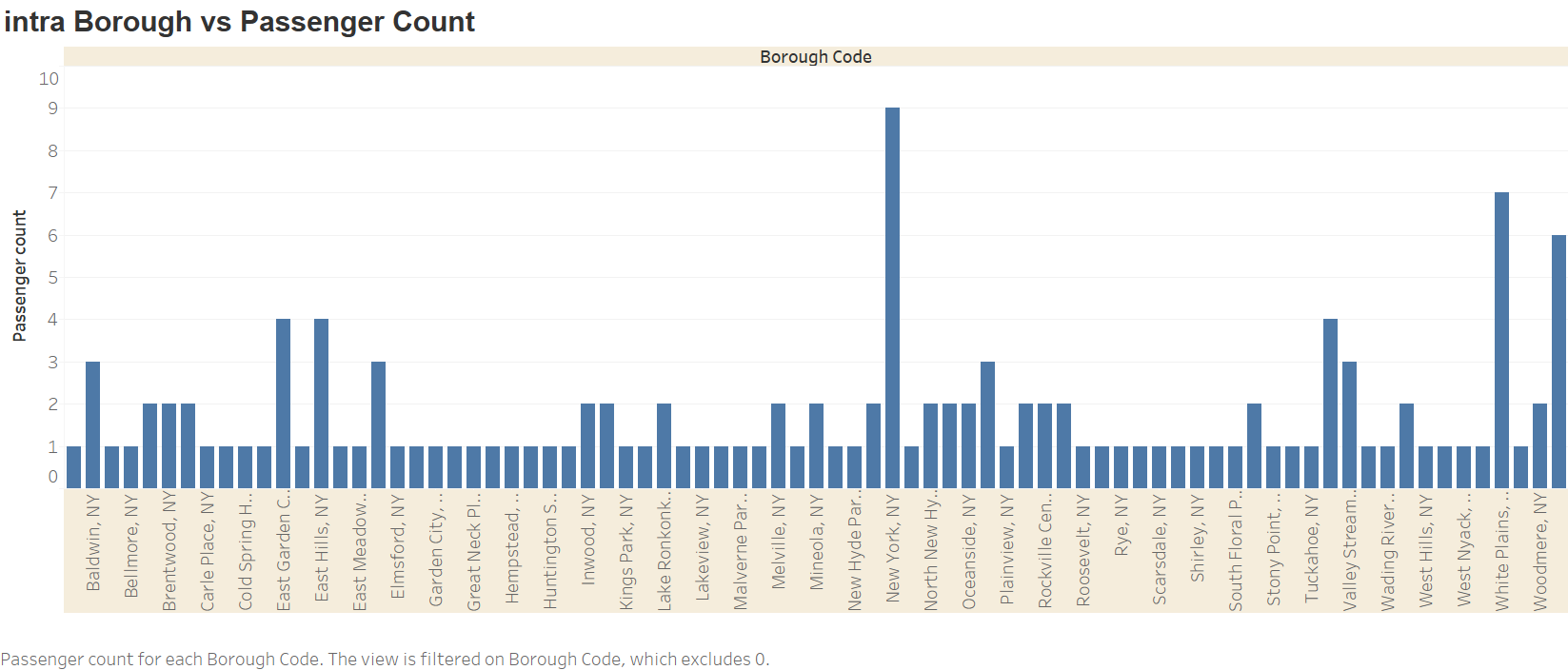
**The plot reflects an important insight that the passengers travelling to Airport JFK pays more tip as compared to other location.**

**The passengers travelling to Newark Airport pay less tip to the drivers.**

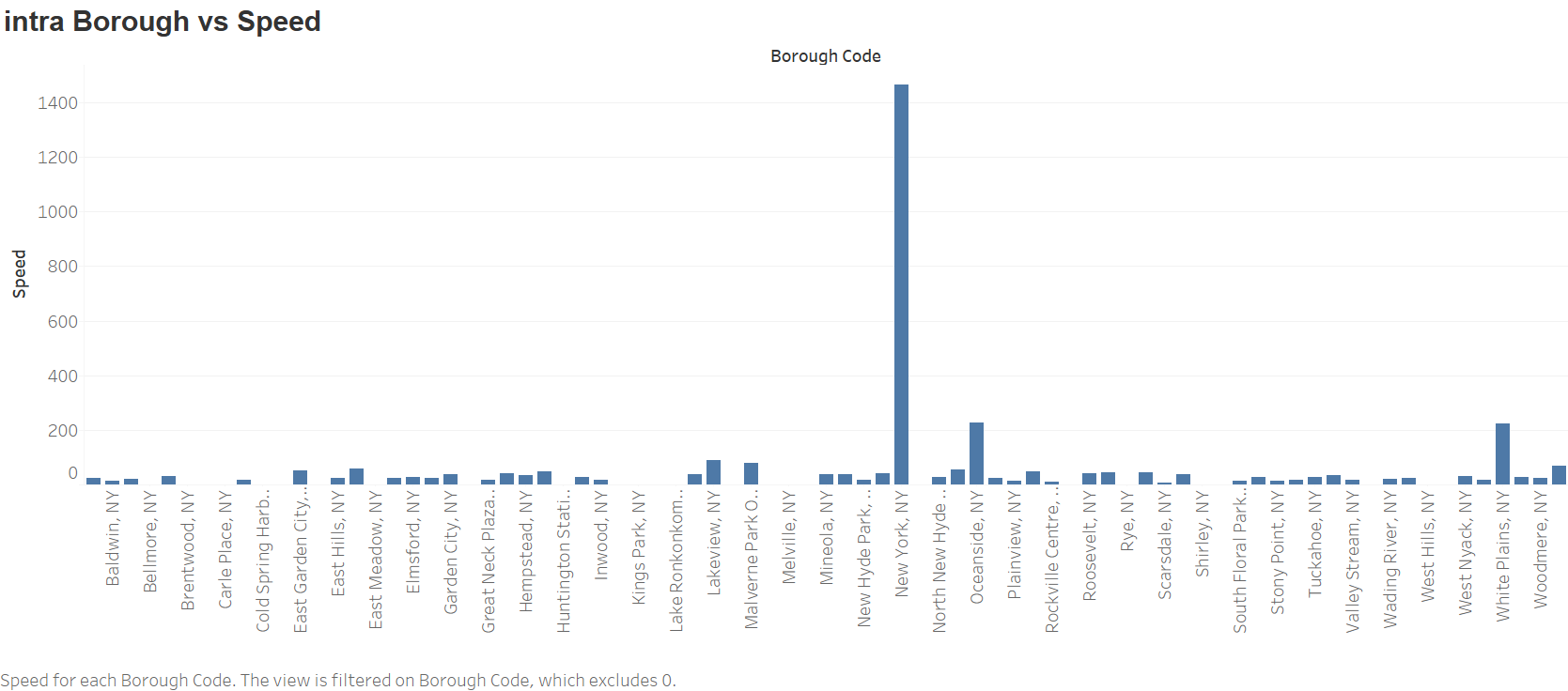
****

****

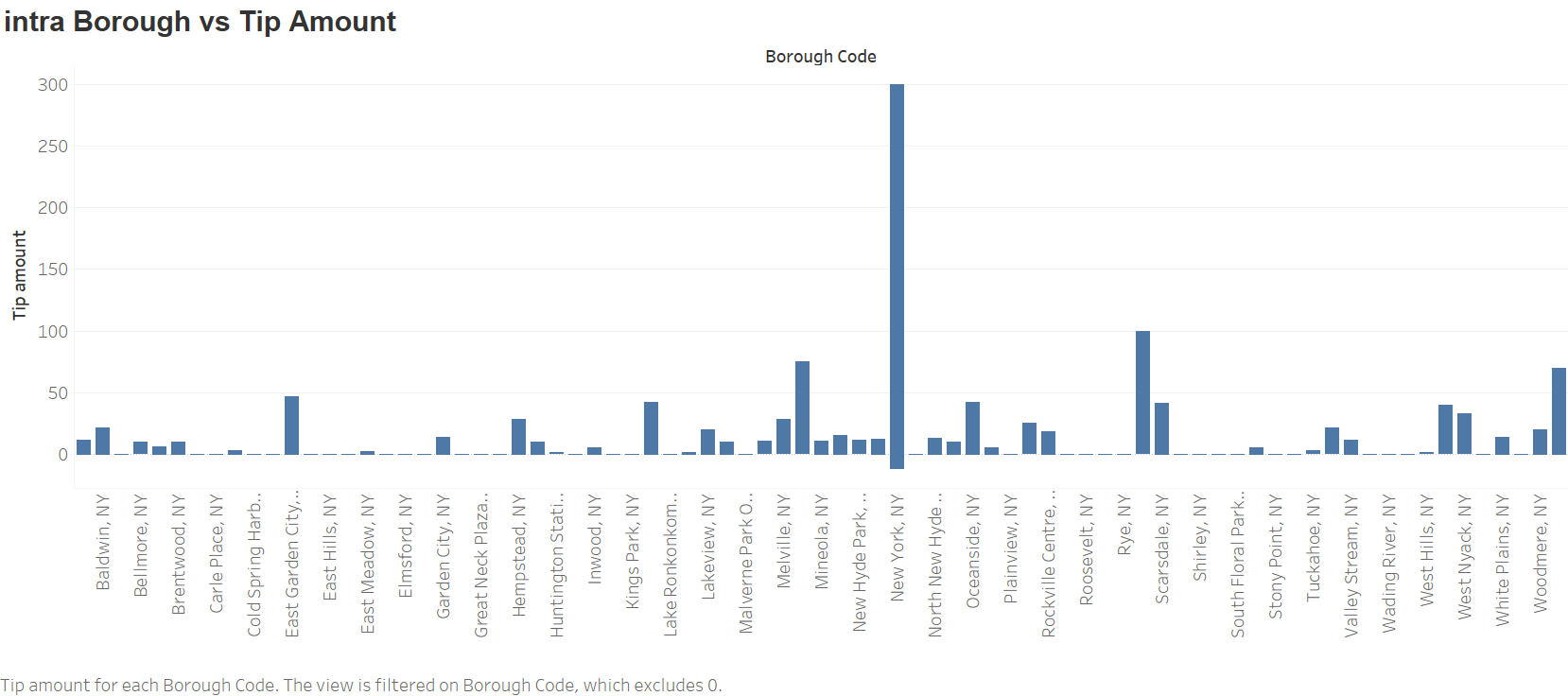
**For rate code Id = 4 “Negotiated Fare” there are cases where driver drives at very high speed and have hardly got ant tip from the passenger. The same applies to standard rate too. The plot shows that the passengers have given tips when the driver is driving with a nominal speed. Even passengers appreciate speed as per the rules.**

****

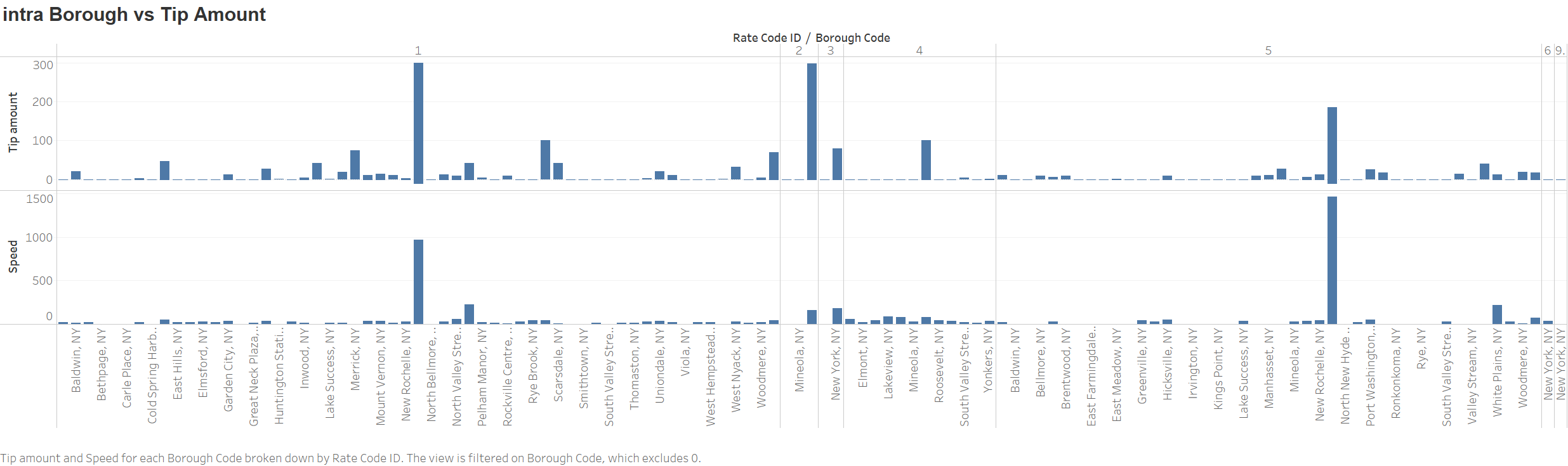
**The maximum passengers travelling are from New York, White Plans and Woodmere.**

****

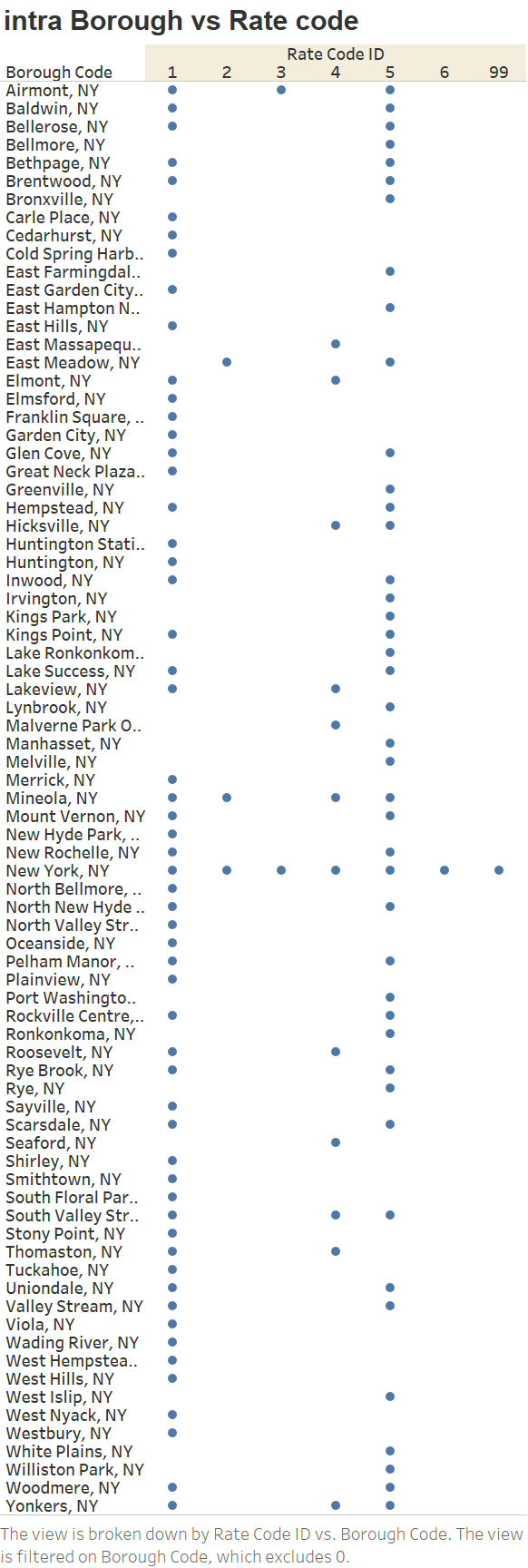
**The driver drives very fast in New York.**

****

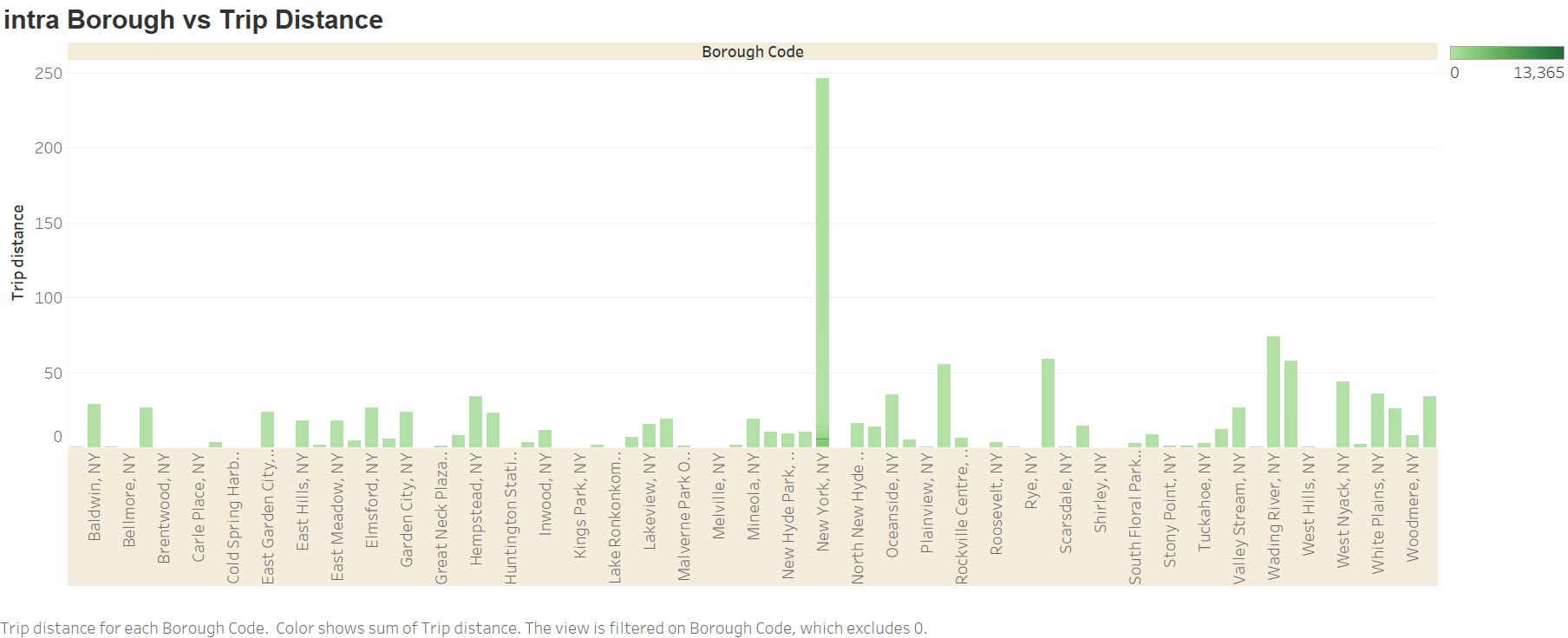
**Highest Tips are earned from New York, Rye, Woodmere and Melville**

****

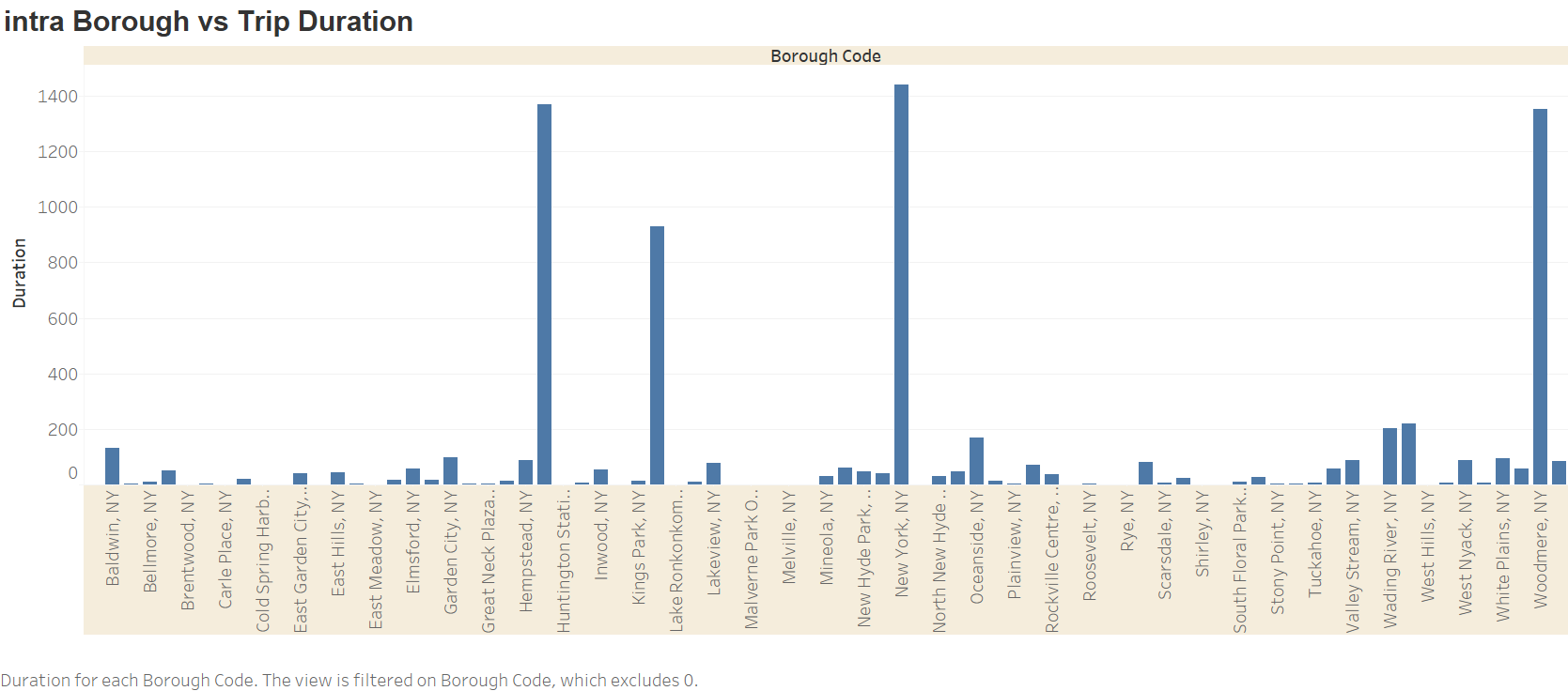
**The highest tips are given for JFK is highest from Mineola, NY then New Rochelle for standard rate trips.**

****

**Mostly people travel with standard rates or negotiated fare in INTRA- BOROUGH Areas.**

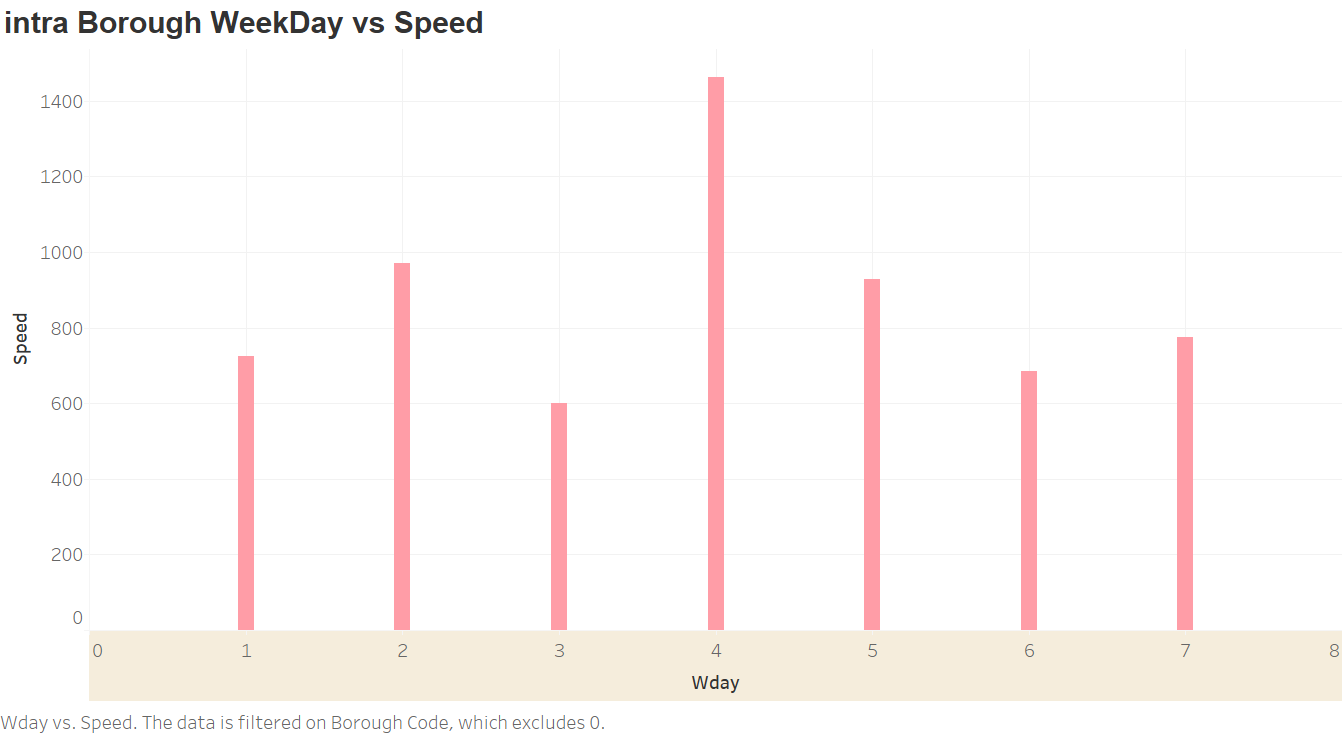
****

**The highest Trip distance is in New York. NY**

****

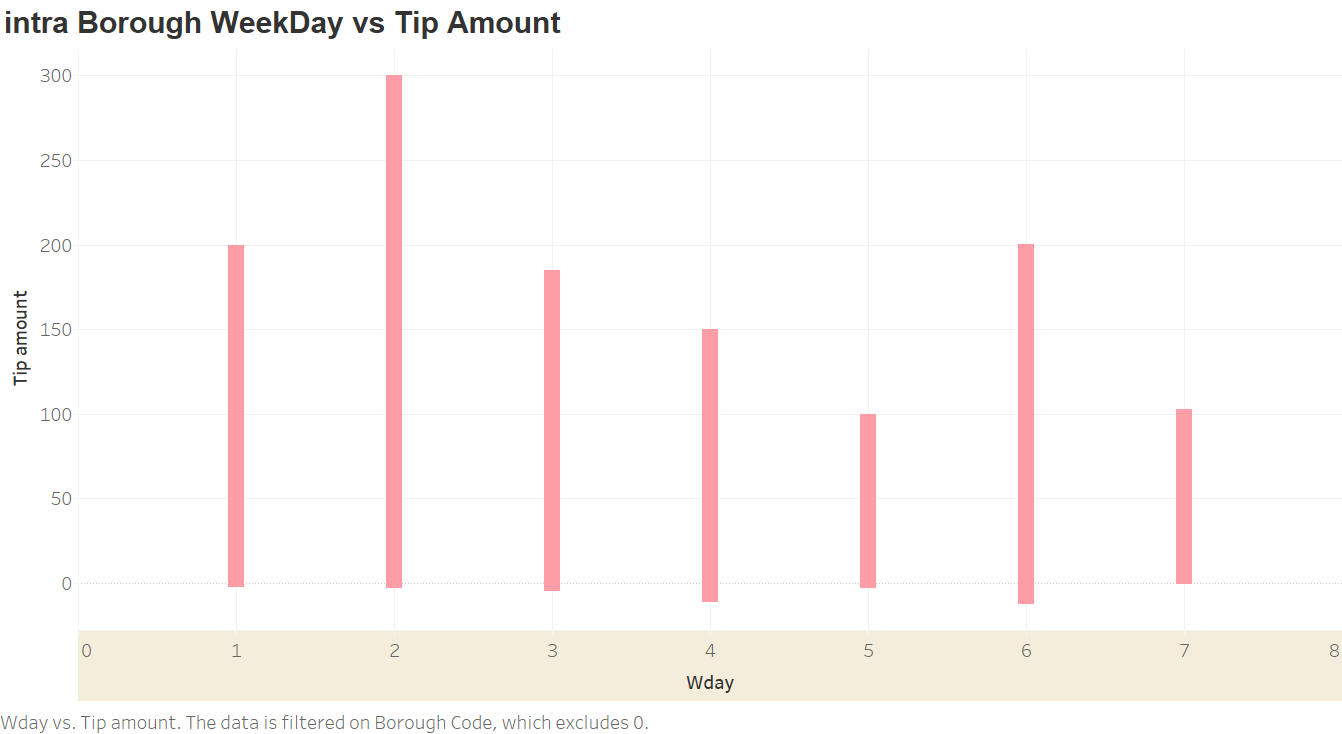
**The highest Trip duration is in New York. NY as the trip distance is more.**

**But in Hempstead, Woodmere and Kings Park the duration is more and Trip distance is less. This shows that these are more traffic prone areas as the trip take more time with less distance travelled. Even passengers are not willing to pay Tip amounts as compared to other areas.**

****

**Mostly 4th day of the week, the drivers drive fast hence less traffic on road. This can also help us conclude that people may take work from home on the 4th day of the week and therefore no traffic for office.**

**If I had time so would have done analysis on the duration of time – morning, afternoon, evening patterns and its effect on the traffic movement, trip duration and Tip amount.**

****

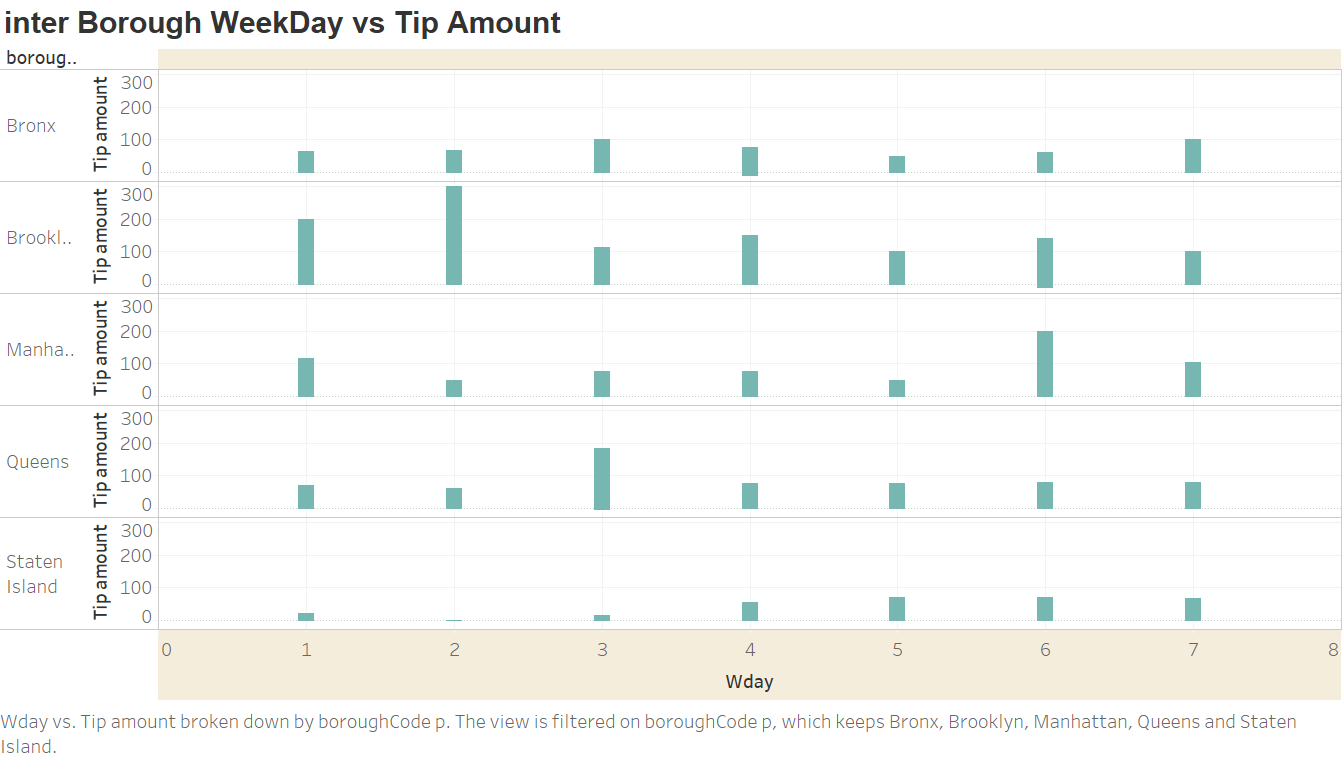
**Passengers mostly give High Tip amounts on 2nd and 6th day of the week. 2nd when they are on work mode and 6th when there are on party(holiday) mode.**

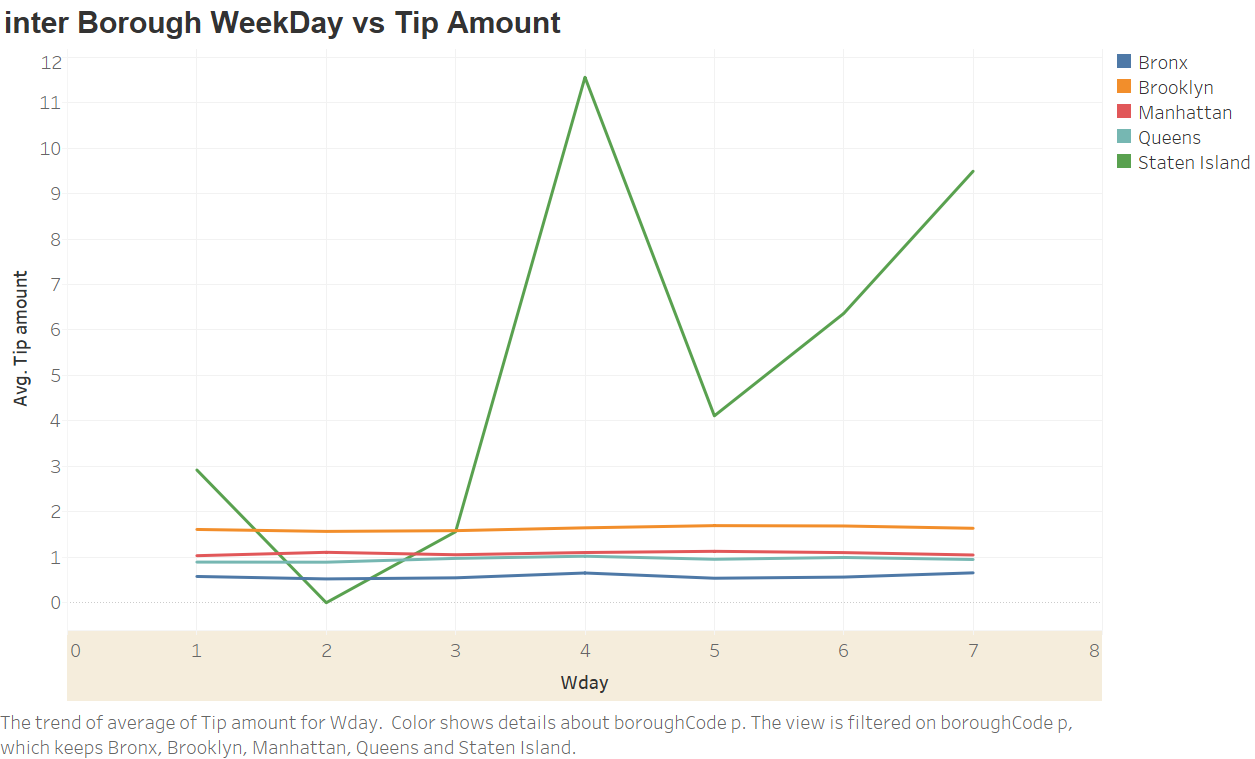
**The Graph shows that no tip is given on 2nd day of the week in Staten Island.**

**In Brooklyn the tip is given mostly on the 2nd day of the week**

**In queens the tip is given mostly on the 3rd day of the week**

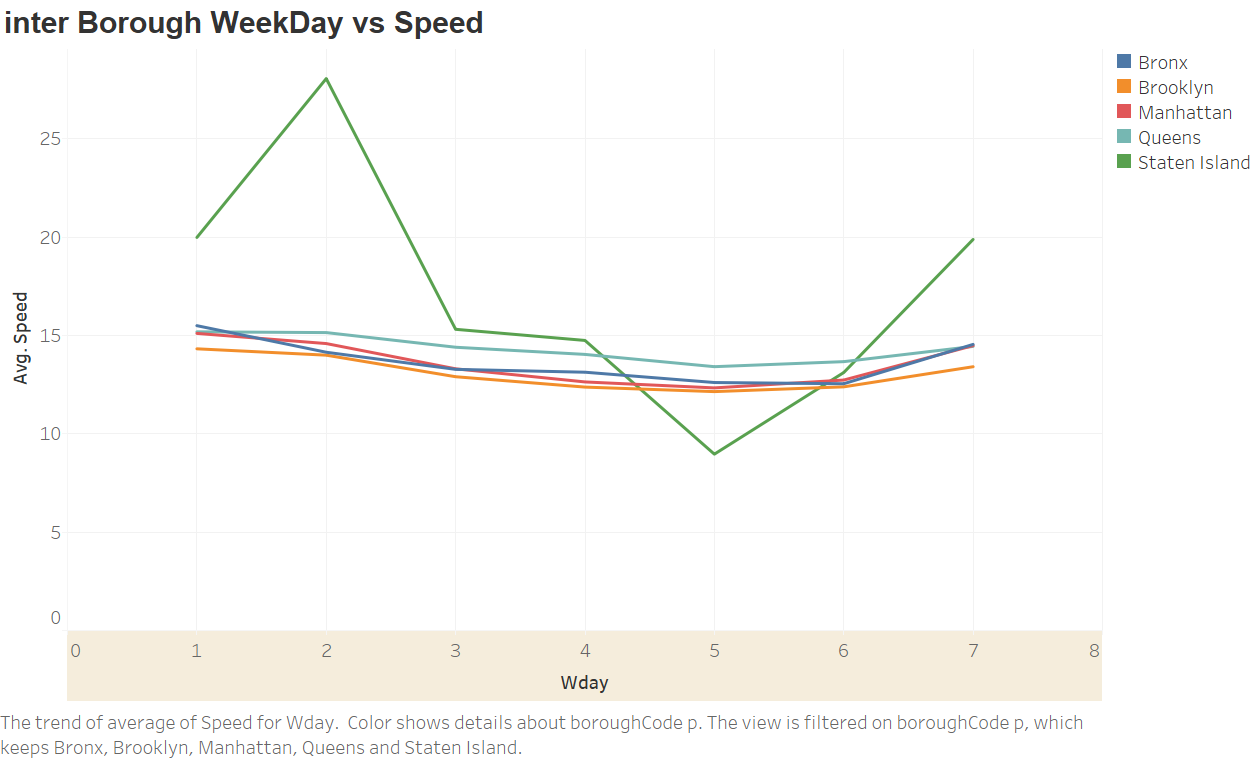
**On the 6th day that is Saturdays, the tip amount is more than the average, hence passengers pay more tips on Saturdays.**

****

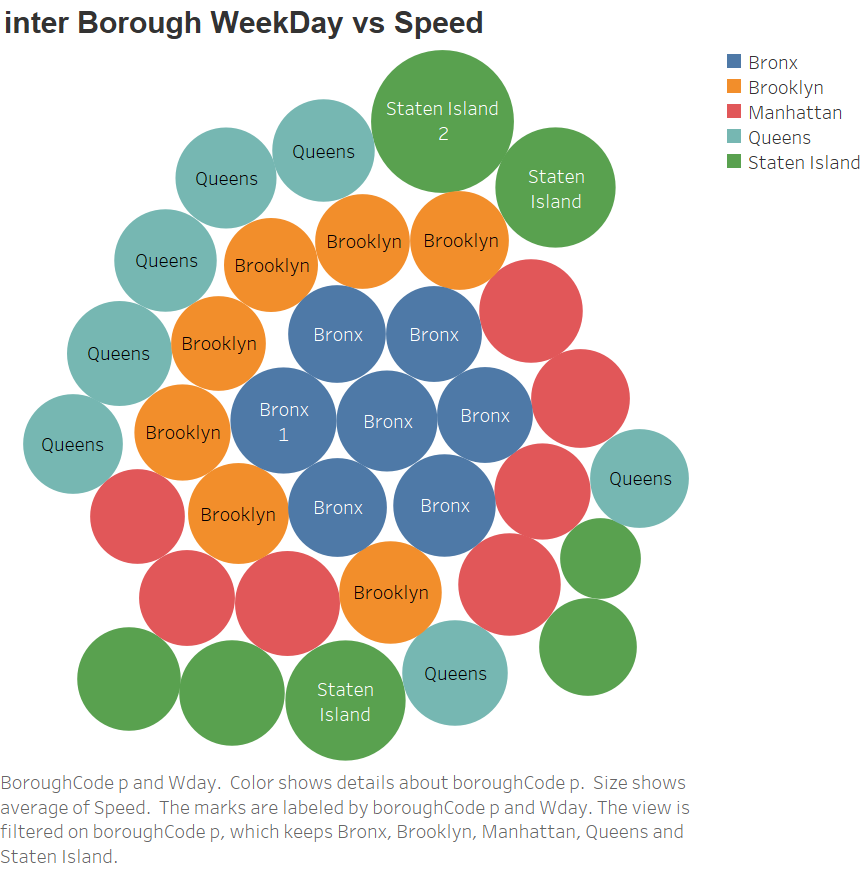
****

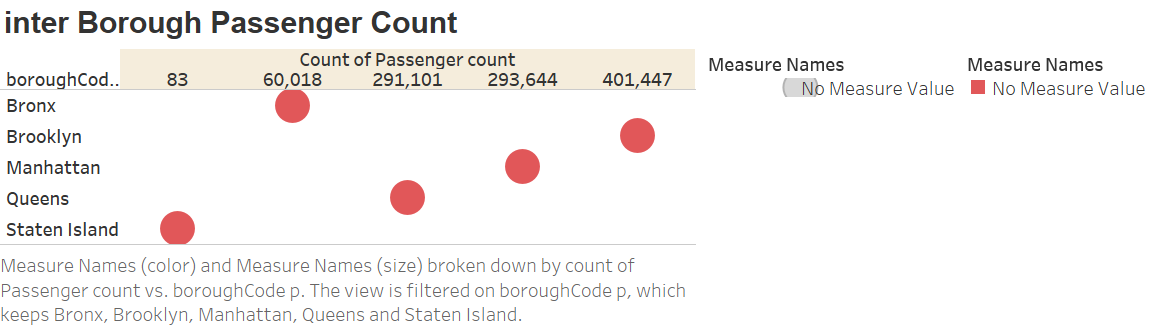
**The average Tip amount is somewhat same for all the regions except the Staten Island.**

**The highest is on 4th day the followed by 7th day.**

****

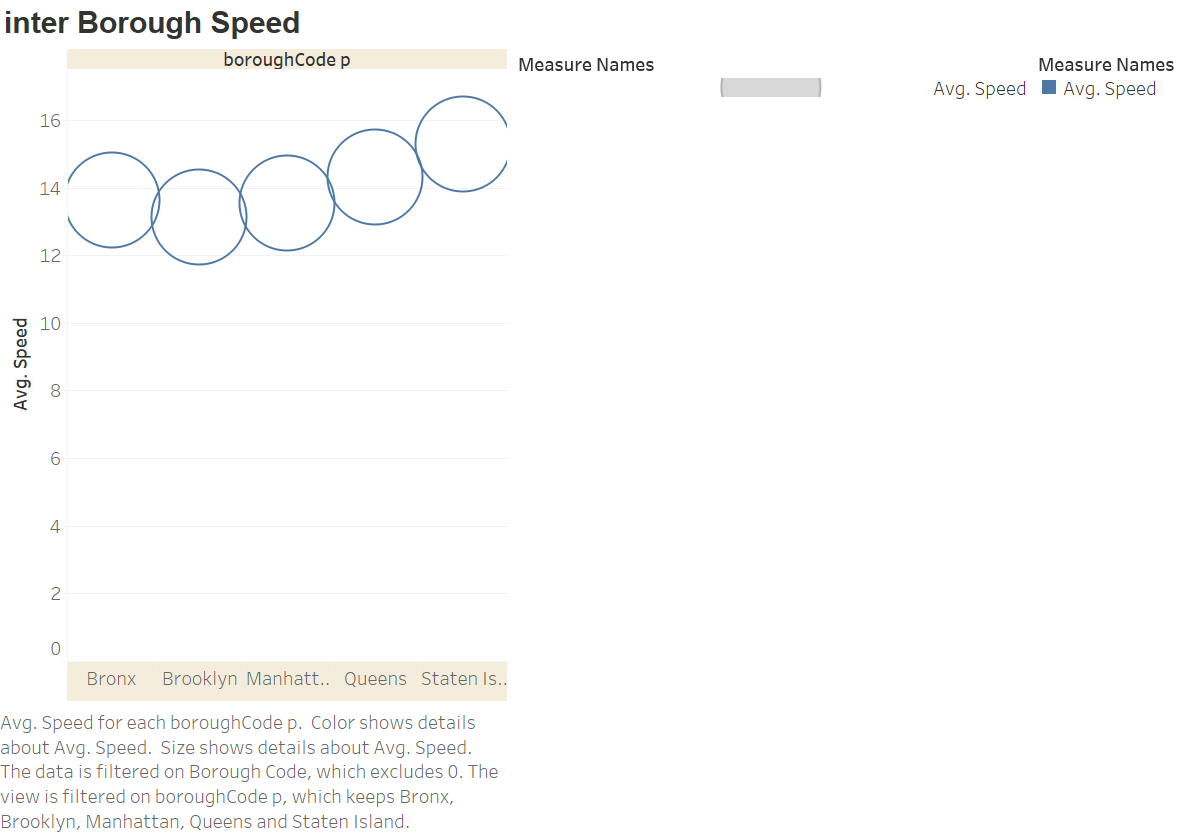
**On second day of the week, in Staten Island, the Average speed increases.**

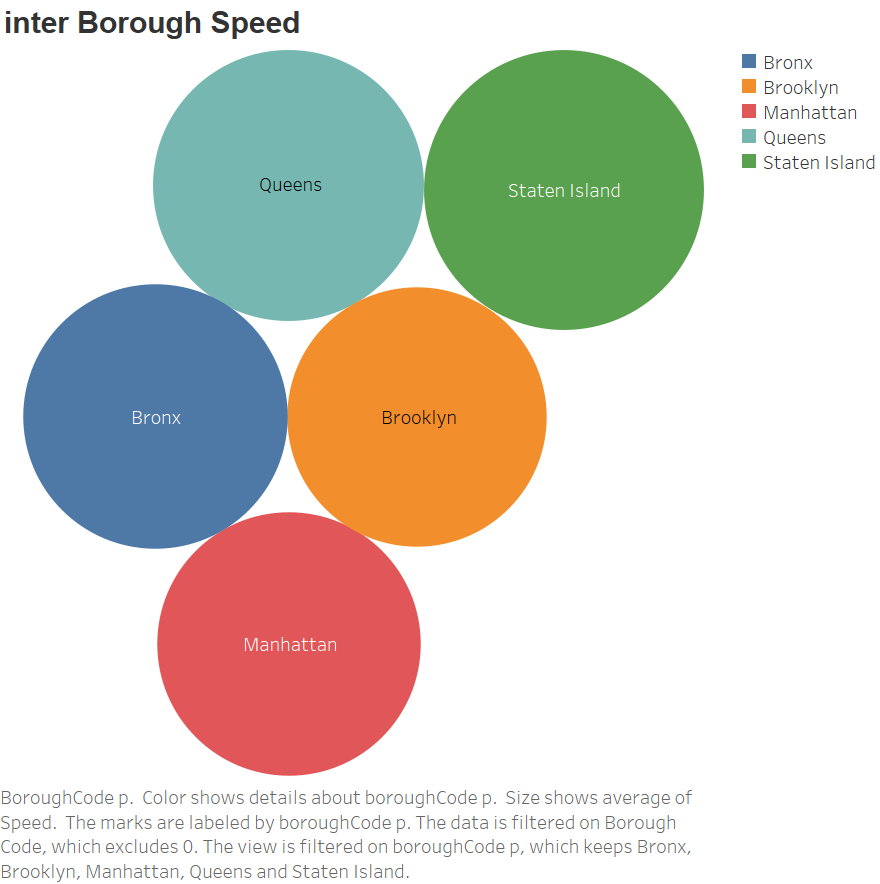
****

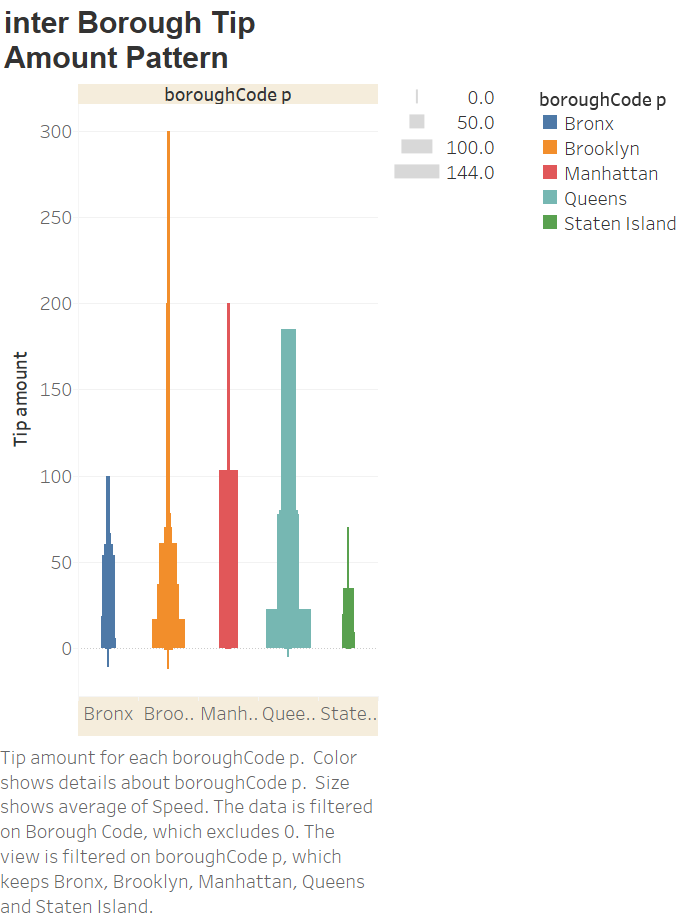
****

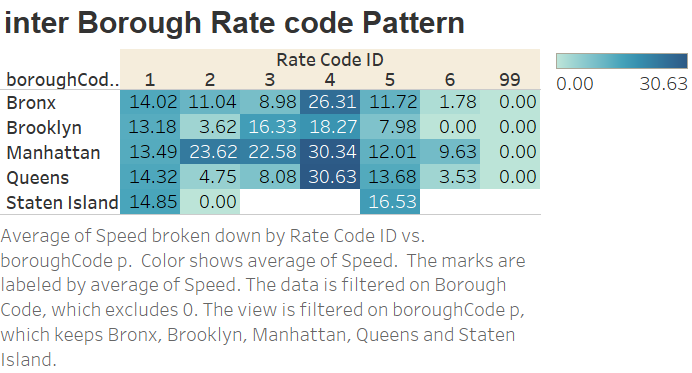
**The highest passengers are from Brooklyn and even the tip amount is more. Hence this shows that in Brooklyn people mostly travel by green taxi and they like slow speed that means relaxed lifestyle.**

**In Staten island, very few people travel in green taxi.**

****

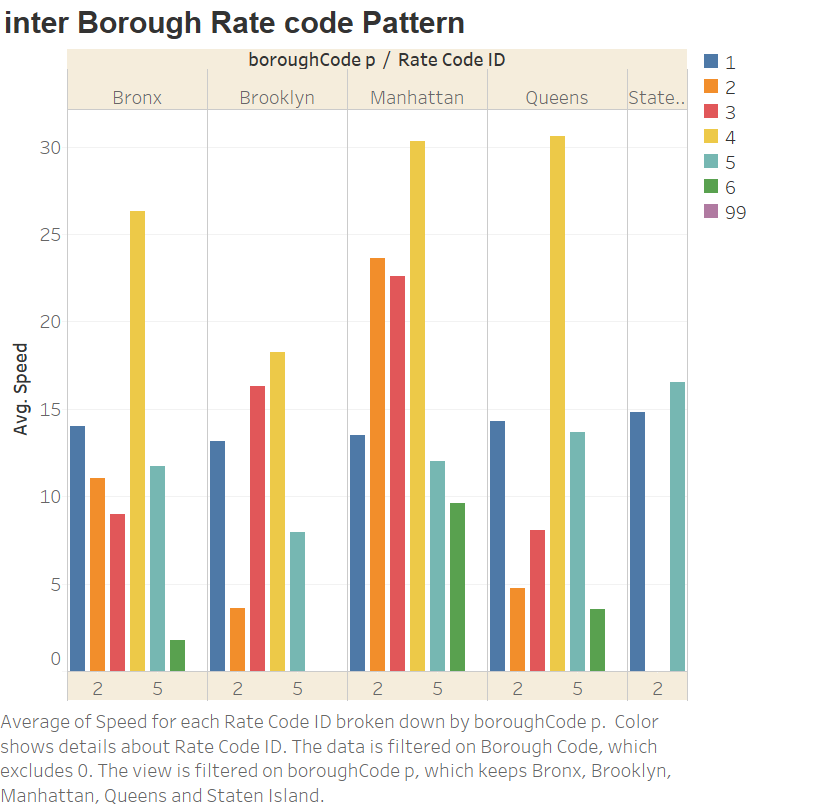
****

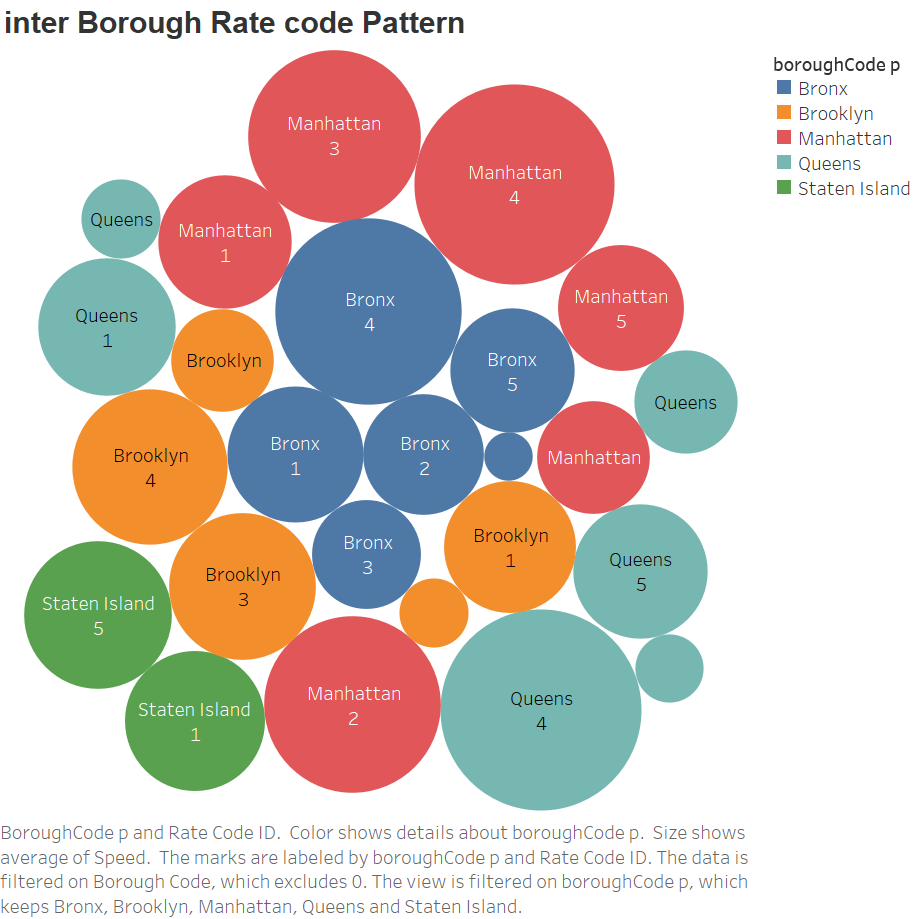
****

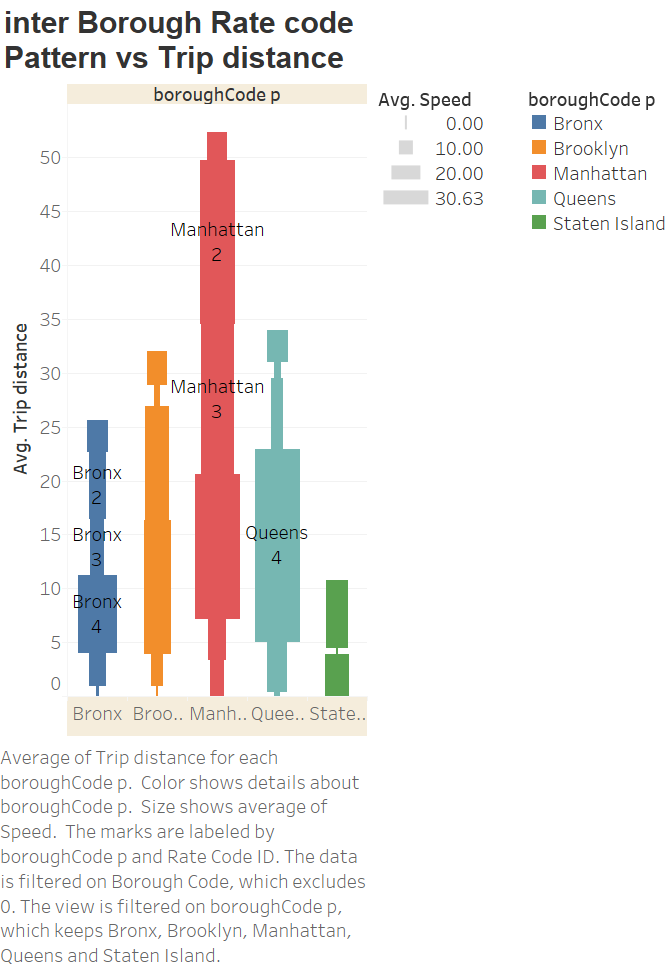
****

**Interesting facts are from Staten Island that no one travels to airport from green taxi and very few which travels from green taxi are local travelers with standard rates.**

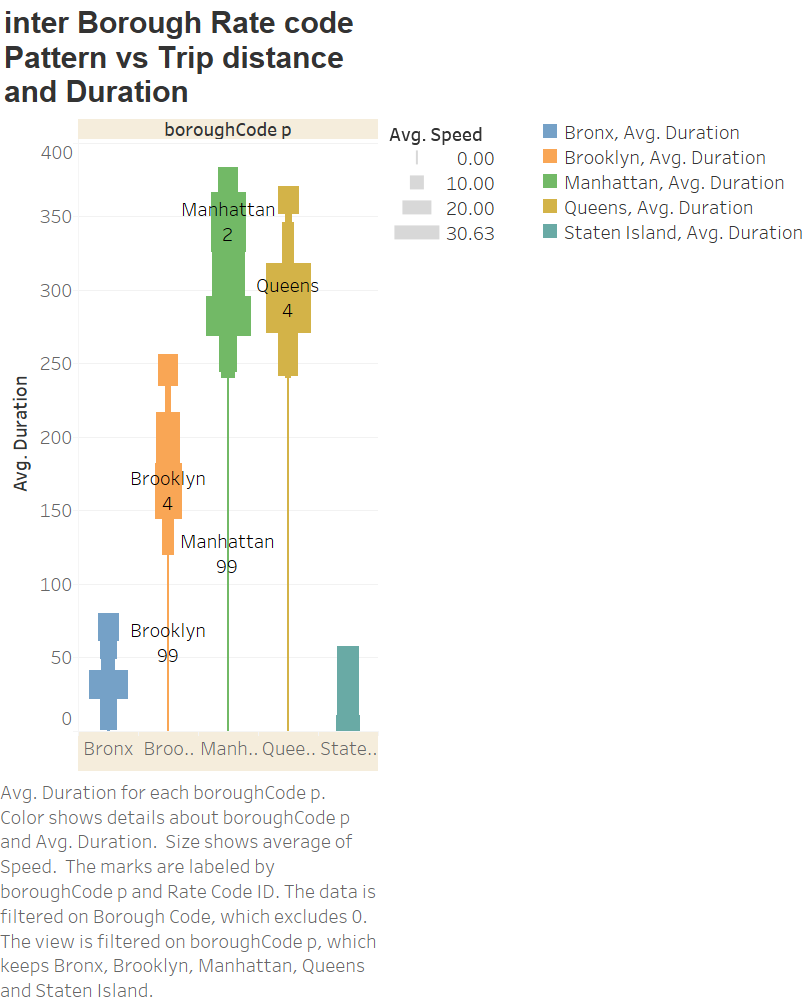
**The maximum travelers for Airport in JFK and Newark are from Manhattan and they give substantial Tip amount. 293644 number of Travelers from Manhattan are a good amount. The taxi driver should focus more on Manhattan Travelers.**

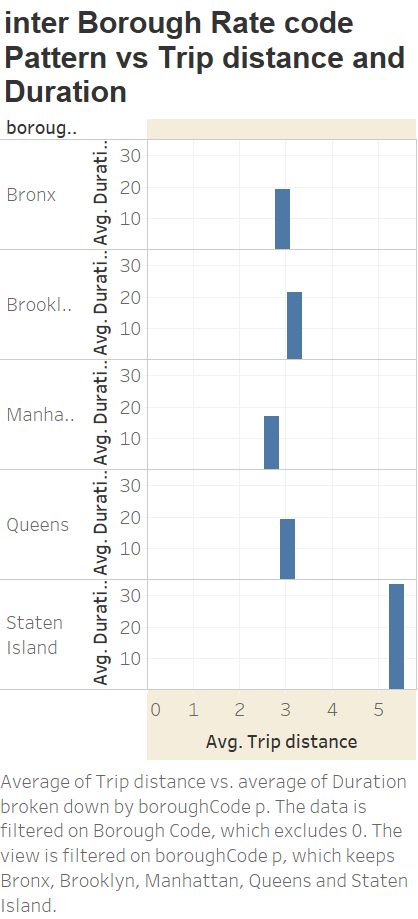
****

****

****

**The trip distance for Manhattan are more mainly for airports and that’s why it’s a profitable area for Green Taxi.**

****

****

# Code to extract the excel file for Visualization:

**library(data.table)**

**library(spatial)**

**library(jsonlite)**

**library(geojsonio)**

**#Read the file to extract Borough code**

**nycjson<-geojson\_read("https://raw.githubusercontent.com/dwillis/nyc-maps/master/boroughs.geojson",what='sp')**

**#Only Borough Code Mapping**

**Boroughs = c('Manhattan', 'Bronx', 'Brooklyn', 'Queens', 'Staten Island')**

**summary(Boroughs)**

**nycZone<-nycjson**

**nycZone[,2:3]<-NULL #only keep the borough code**

**#Reading the Green Taxi sept 2015 file**

**taxiGreen <- fread('https://s3.amazonaws.com/nyc-tlc/trip+data/green\_tripdata\_2015-09.csv',stringsAsFactors = F)**

**taxiGreen$dummy1 <- NULL #handle the excess ',' in all the rows of the csv files**

**taxiGreen$dummy2 <- NULL**

**names(taxiGreen) = c("VendorID" , "Pickup\_datetime" ,"Dropoff\_datetime" ,"Store\_and\_fwd\_flag" ,**

**"RateCodeID" , "Pickup\_longitude" , "Pickup\_latitude" , "Dropoff\_longitude" ,**

**"Dropoff\_latitude" , "Passenger\_count" , "Trip\_distance" , "Fare\_amount" ,**

**"Extra" , "MTA\_tax" , "Tip\_amount" , "Tolls\_amount" ,**

**"Ehail\_fee" , "improvement\_surcharge" ,"Total\_amount" , "Payment\_type" ,**

**"Trip\_type" )**

**summary(taxiGreen)**

**library(sp)**

**library(rgdal)**

**#To calculate pickup and drop from longitude and Latitude**

**PickupArea<-SpatialPoints(cbind(taxiGreen$Pickup\_longitude,taxiGreen$Pickup\_latitude))**

**PickupArea@proj4string <- nycjson@proj4string**

**pickupBoroughCodes<-PickupArea %over% nycZone**

**summary(pickupBoroughCodes)**

**taxiGreen$boroughCode\_p <- pickupBoroughCodes$BoroName**

**DropoffPts<-SpatialPoints(cbind(taxiGreen$Dropoff\_longitude,taxiGreen$Dropoff\_latitude))**

**DropoffPts@proj4string <- nycjson@proj4string**

**dropoffBoroughCodes<-DropoffPts %over% nycZone**

**taxiGreen$boroughCodeDrop<-dropoffBoroughCodes$BoroName**

**#To extract month/weekday/hour from the data given**

**taxiGreen$month <- month(taxiGreen$Pickup\_datetime)**

**taxiGreen$wday <- wday(taxiGreen$Pickup\_datetime)**

**taxiGreen$hour <- hour(taxiGreen$Pickup\_datetime)**

**# Change the format of datetime from string to POSIXct objects**

**taxiGreen$Pickup\_datetime <- as.POSIXct(taxiGreen$Pickup\_datetime,format='%Y-%m-%d %H:%M:%S')**

**taxiGreen$Dropoff\_datetime <- as.POSIXct(taxiGreen$Dropoff\_datetime,format='%Y-%m-%d %H:%M:%S')**

**#To calculate the duration of the trip**

**taxiGreen$duration <- floor(as.double(taxiGreen$Dropoff\_datetime-taxiGreen$Pickup\_datetime)/60.0)**

**#To calculate Tip amount as a percentage of total fare amount**

**taxiGreen$percent <- taxiGreen$Tip\_amount/taxiGreen$Total\_amount \* 100.0**

**#To calculate Speed**

**taxiGreen$speed <- taxiGreen$Trip\_distance/taxiGreen$duration \*60**

**#Filter speed**

**taxiGreen$speed[taxiGreen$speed=="NaN"]<-0**

**# future usage to extract csv file and used for visualization in Tableau**

**write.csv(taxiGreen, file = "greenLocationBoro.csv", row.names = TRUE)**

**str(taxiGreen)**