University of Illinois at Chicago

IDS 462

Citi Bikes Data Analysis

Submitted by:
Kalindi Deshmukh
Alejandro Rodriguez
Rupal Sinha

Overview

Citi Bike is New York City’s bike sharing system, and the largest in the nation. Like other bike sharing systems, Citi Bike has a fleet of specially designed and durable bikes that are available to anyone 24/7, 365 days a year. Citi Bikes can be unlocked from any station and returned to any other station in the city. Once you opt into Ride Insights, the Citi Bike app will use your phone's location to record the route you take between your starting and ending Citi Bike station to give exact mileage. This system is popular for those that commute to work, school, appointments and those that enjoy recreational bike riding. Citi Bike is operated by Motivate, global leader in bike share. It manages all of the largest bike sharing systems in the world, including Divvy (Chicago), Capital Bike Share (Washington D.C.), and more.

Introduction

Citibike offers many datasets from current and previous years that were useful in the planning stages of our research. The initial stage of our research was of course observing the available datasets and drafting ideas that may be useful in analyzing our datasets. Particularly, we each found datasets that we can merge with our Citibike datasets using zip codes as a common attribute to merge on. After we finalized the dataset that will be used to analyze Citi Bike, drafting research objective ideas was not as difficult as first expected. This was because we had been analyzing the datasets when merging them. We then came to the consensus that researching what factors affect revenue generated by Citi bikes will be important to Citi Bikes business model.

Data Collection

The main datasets were collected from Citi Bike’s official website. The System data is provided according to the NYCBS Data Policy. We extracted datasets from Citi Bike Trip History and Citi Bikes Ridership and Membership Data. The Citi Bike Trip History data gave details about the user and the trip for the year 2016.

Citi Bikes Ridership and Membership Data contains details about the types of passes purchased, memberships for January- April 2016. Some things noted about these datasets were that:

- Trip count and mileage estimates include trips with a duration of greater than one minute.
- Mileage estimates were calculated using an assumed speed of 7.456 miles per hour, up to two hours.
- Trips that began at publicly available stations.

For our analysis, we also tried to include the 2016 crime dataset from NYPD's official website. It contained details about the type of crime, longitude, latitude, etc. Further analysing, we found that the crime report showed very little correlation with variables in our data, so we decided to drop the Crime data and go ahead with the other datasets.

Research Topic

In this report we plan to analyze the factors that affect revenue generated by Citi bikes by using the following statistical methods:

- Univariate Analysis
- Bivariate Analysis
- Linear Regression Modelling
- Principal Component Analysis
- Clustering

Our data contains many important variables that are almost always useful in determining changes in revenue such as the age of a customer, whether a customer is a subscriber or one time user, weekday, and more. Thus, it is reasonable to believe revenue is higher on weekdays, revenue will be greater in some zip codes compared to others, and 24 hour passes will generate greater revenue than 3 day passes. New York City and its surrounding metropolitan areas are highly populated, and this means traffic is imminent during peak hours (work hours, events, etc.). Thus, all methods of transportation during highly active hours will be condensed with many people. With this said, we believe that Citi bikes provide the best possible method of transportation during weekdays to help customers get to work or other destinations during these highly active hours. Certain zip codes may also be populated more than other zip codes, so we believe that revenue generated in some zip codes will be greater than others that are not as popular. Citibike has many loyal customers that rely on its ease of access transportation method, so we believe that rather than buying constant single day passes, revenue generated by frequent customers that subscribe will be far greater than those who buy single passes. In the next section, we will analyze the data that guides our analysis and provides evidence behind the conclusions we make and the results we obtain to answer our research question.

Summary of Key Findings

Our analysis yielded insightful information about our dataset and important relationships that contribute to revenue generated by Citi bikes. Univariate and Bivariate analysis showed important relationships between revenue and the variables in our dataset. Linear regression modelling also generated important independent variables that contribute to increases (decreases) of total revenue. Principal component analysis determined that 6 components explained 84% of our data and returned useful loadings.

Description of Data

Our dataset contains 40 variables of many types. Listed are the variables that contributed the most to our analysis.

VARIABLE	TYPE	DESCRIPTION
Date	Date	Date of bike trip
Tripduration	Numeric	Trip duration
Starttime	Character	Start time of trip
Stoptime	Character	Stop time of trip

Start station latitude	Numeric	Start station latitude
Start station longitude	Numeric	Start station longitude
End station latitude	Numeric	End station latitude
End station longitude	Numeric	End station longitude
Usertype	Factor	User is subscriber or customer
Birth year	Numeric	Customer’s year of birth
Gender	Factor	Gender of user: 1 - Male 2 - Female
Trips.over.the.past.24.hours.midnight.to.11.59pm	Numeric	Number of trips in the past 24 hours
Cumulative.trips.since.launch	Numeric	Number of trips to date
Miles.traveled.today.midnight.to.11.59.pm	Numeric	Number of miles traveled on respective date
Miles.traveld.to.date	Numeric	Total number of miles traveled to date
Total.Annual.Members.All.Time	Numeric	Number of Annual Members of all time
X24.Hour.Passes.Purchased.midnight.to.11.59.pm	Numeric	Number of 24 hour passes purchased

X3.Day.Passes.Purchased.midnight.to.11.59.pm	Numeric	Number of Day passes purchased
Annual.members.added	Numeric	Number of annual members added on respective date
Revenue.from.annual.members	Numeric	Revenue from annual members
Revenue.from.24.hr.customer.passes	Numeric	Revenue from 24 hour customer passes
Revenue.from.3.day.customer.passes	Numeric	Revenue from 3 day customer passes
Total.Revenue	Numeric	Total revenue from passes and annual members
Weekday	Character	Day of the week
Distance	Numeric	Distance traveled
Speedkmh	Numeric	Average speed traveled (kmh)
Overage	Character	Did customer or subscriber exceed trip time limit?
Roundtrip	Character	Was the trip a roundtrip?
Timeslot	Character	Timeslot of trip
Age	Numeric	Age of customer/member

Zip code	Factor	Zip code of bike station

Data Preparation

As with all datasets used in statistical analysis, we first prepared our data by ‘cleaning’ it by removing missing values, converting variables to the correct data types, and removing unnecessary variables. Below are the steps used to reach the final data set.

- Merged datasets on the common variable ‘zip codes’
- Removed NA values
- Split the character variable ‘start time’ into two variables ‘start day’ and ‘start time’ and removed initial start time variable
- Generated zip codes using the revgeo( ) function (r-package:”revgeo “) from the latitude and longitude variables
- Used birth year variable to create a new ‘Age’ variable
- Removed level of 0 from the gender variable and kept levels 1 and 2
- Used the Citibike.com website to obtain prices to create three new variables (Revenue from annual members, 24hr customer passes, 3 day customer passes)
- Used the previous three variables to create a new variable ‘Total Revenue’ by added all three
- Created a new variable named ‘Weekday’ by using the weekdays ( ) function on the date variable
- Created two new variables (‘distance’ and ‘speedkmh’) by imap package and functions
- Created two new variables (‘overage’ and ‘roundtrip’) by using conditional statements on the start time and the latitude and longitude variables respectively.
- Converted variables with the wrong data types into the correct types

Analysis and Results

Univariate Analysis

Reporting univariate analysis statistics is important in our research because we want to understand each of the variables and find patterns that may be useful in our analysis. More particularly, we will understand each variable independently and/or with the dependent variable (Total Revenue). In reporting the univariate statistics, we used central tendency measures (mean, median and mode), standard deviation, histograms, and density plots to analyze the independent variables: The variables that showed significant results that were useful in understanding our data are shown in the appendix.

Total Revenue

In the Figure 1 from Appendix, we can see that the density plot for Total Revenue is not normally distributed. So, what we are able to do is scale the data so that is can become more evenly distributed by removing the outliers. And thus after scaling and removing outliers, we got Figure 2.

Miles Traveled

Miles traveled was seen to have a skewed and non-normal distribution as per Figure 3. We can scaled this data to analyze this more effectively. The Figure shows that the values are slightly more balanced than the previous graph.

Trip Duration

This density plot of Trip duration (Figure 5) is skewed. The reason behind this is because we have outliers with values of 10,000 and above. This may be because there was an error in the built in tracking system of the Citi bike use or it may have been stolen.

Zip Code and Days of the Week

The most popular bike stations are located in the 10016 and 10017 zip codes as seen in the plot (Figure 6). These zip codes are located near the center of New York City, so this is why these zip codes have relatively higher popularity compared to the rest. We can also see that there are significantly few riders on the weekends compared to the weekdays (Figure 7). We believe this may be because Citi bikes are used more among those that commute to work or meetings during the week rather than those who use the bikes recreationally.

Bivariate Analysis

The Bivariate analysis was done to find relations between Revenue generated and other variables that were important for our research. The analysis is as follows:

Revenue and Numeric Variables: For this analysis, we plotted the variables and checked the correlations with correlation tests (cor.test) and observed the p values (which were less for most of the cases). The higher the cor value, the more correlated.

Revenue ~ Variable Name	**Figure in Appendix**	**Cor and Comments**
Annual Membership Subscriptions	Figure 8	0.9110837=>Highly Correlated
Miles Travelled	Figure 9	0.6147264 => Correlated
24 hours Passes Purchased	Figure 10	0.8968604 =>Highly Correlated
3 Days passes Purchased	Figure 11	-0.1503073 => Less Correlation and inversely proportional
Speed	Figure 12	-0.009266251 => Very Less Correlation and inversely proportional
Trip Duration	Figure 13	0.04327174=> Very Less Correlation
Age	Figure 14	-0.05106508 => Less Correlation and inversely proportional

Revenue and Factor Variables: For this analysis, we plotted the variables with boxplots and checked the relations using Anova tests and Aggregate functions. We observed the asterisk from the results of the summary of the Anova model. The more the asterisks, the more correlation between the two variables. The results are tabulated as follows:

Revenue~ Variable Name	**Figure in Appendix**	**Anova Test Results**
Gender	Figure 15	Not Correlated
Zip Code	Figure 16	Correlated
Week Day	Figure 17	Highly Correlated
Time Slot	Figure 18	Highly Correlated

Findings from Bivariate Analysis:

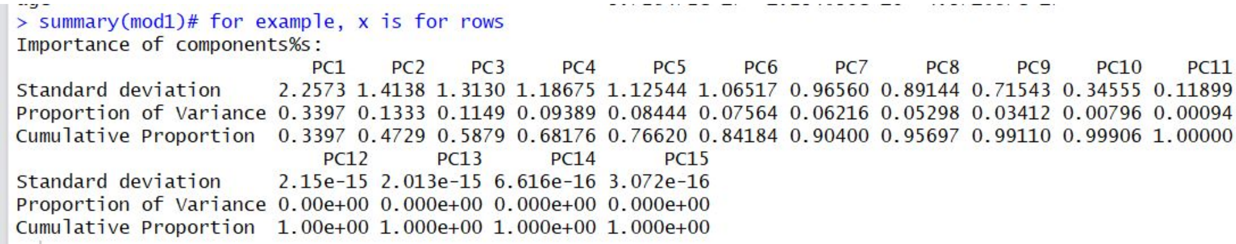
Annual Memberships, Miles Travelled, 24 hours passes purchased, Zip Codes, Week Day, Time Slots are correlated with the revenue generated. We can also see that, Time Slot 1 and weekends are the times when most of the revenue was generated.

Principal Component Analysis

To further wok on our analyses and cross check with the previous results, we used PCA.

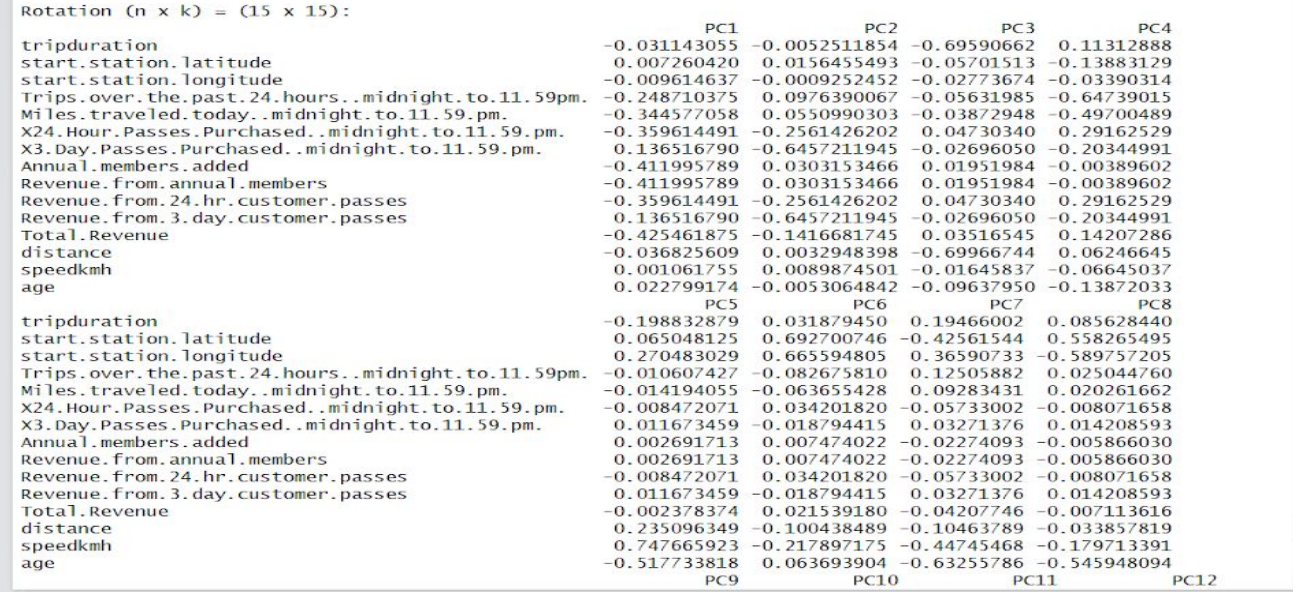
Principal components analysis (PCA) is one of the most popular dimension reduction techniques. This helps us understand the underlying structure of the data we have. We need it to identify the structure of the data and noise as well. We also seek to simplify our analysis visually with PCA. To start analyzing our data using PCA, we first dropped the futile variables. The variables that were retained were:

"tripduration" , "start station latitude" , start station longitude, Trips.over.the.past.24.hours..midnight.to.11.59pm., distance, speed, age, miles.traveled.today..midnight.to.11.59.pm, X24.Hour.Passes.Purchased..midnight.to.11.59.pm., Miles.traveled.today..midnight.to.11.59.pm, various passes purchased, and the amounts collected from them. Next, we determined how many principal components we will need by using the fa.parallel ( ) function in R, which returns a scree plot with a recommended number of principal components. As seen in the Scree plot (Figure 19), the plot recommends we use approximately six principal components. The first six components are able to explain about 84.18 % of variance in data.



```
> summary(mod1)# for example, x is for rows
Importance of components%s:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     2.2573 1.4138 1.3130 1.18675 1.12544 1.06517 0.96560 0.89144 0.71543 0.34555 0.11899
Proportion of Variance 0.3397 0.1333 0.1149 0.09389 0.08444 0.07564 0.06216 0.05298 0.03412 0.00796 0.00094
Cumulative Proportion  0.3397 0.4729 0.5879 0.68176 0.76620 0.84184 0.90400 0.95697 0.99110 0.99906 1.00000
                           PC12      PC13      PC14      PC15
Standard deviation     2.15e-15 2.013e-15 6.616e-16 3.072e-16
Proportion of Variance 0.00e+00 0.000e+00 0.000e+00 0.000e+00
Cumulative Proportion  1.00e+00 1.000e+00 1.000e+00 1.000e+00
```

Using principal component analysis function pcromp, we find the loadings of each variable.



```
Rotation (n x k) = (15 x 15):
                                                        PC1           PC2         PC3         PC4
tripduration                                   -0.031143055 -0.0052511854 -0.69590662  0.11312888
start.station.latitude                          0.007260420  0.0156455493 -0.05701513 -0.13883129
start.station.longitude                        -0.009614637 -0.0009252452 -0.02773674 -0.03390314
Trips.over.the.past.24.hours..midnight.to.11.59pm. -0.248710375  0.0976390067 -0.05631985 -0.64739015
Miles.traveled.today..midnight.to.11.59.pm.    -0.344577058  0.0550990303 -0.03872948 -0.49700489
X24.Hour.Passes.Purchased..midnight.to.11.59.pm. -0.359614491 -0.2561426202  0.04730340  0.29162529
X3.Day.Passes.Purchased..midnight.to.11.59.pm.  0.136516790 -0.6457211945 -0.02696050 -0.20344991
Annual.members.added                           -0.411995789  0.0303153466  0.01951984 -0.00389602
Revenue.from.annual.members                    -0.411995789  0.0303153466  0.01951984 -0.00389602
Revenue.from.24.hr.customer.passes             -0.359614491 -0.2561426202  0.04730340  0.29162529
Revenue.from.3.day.customer.passes              0.136516790 -0.6457211945 -0.02696050 -0.20344991
Total.Revenue                                  -0.425461875 -0.1416681745  0.03516545  0.14207286
distance                                       -0.036825609  0.0032948398 -0.69966744  0.06246645
speedkmh                                        0.001061755  0.0089874501 -0.01645837 -0.06645037
age                                             0.022799174 -0.0053064842 -0.09637950 -0.13872033
                                                        PC5          PC6         PC7          PC8
tripduration                                   -0.198832879  0.031879450  0.19466002  0.085628440
start.station.latitude                          0.065048125  0.692700746 -0.42561544  0.558265495
start.station.longitude                         0.270483029  0.665594805  0.36590733 -0.589757205
Trips.over.the.past.24.hours..midnight.to.11.59pm. -0.010607427 -0.082675810  0.12505882  0.025044760
Miles.traveled.today..midnight.to.11.59.pm.    -0.014194055 -0.063655428  0.09283431  0.020261662
X24.Hour.Passes.Purchased..midnight.to.11.59.pm. -0.008472071  0.034201820 -0.05733002 -0.008071658
X3.Day.Passes.Purchased..midnight.to.11.59.pm.  0.011673459 -0.018794415  0.03271376  0.014208593
Annual.members.added                            0.002691713  0.007474022 -0.02274093 -0.005866030
Revenue.from.annual.members                     0.002691713  0.007474022 -0.02274093 -0.005866030
Revenue.from.24.hr.customer.passes             -0.008472071  0.034201820 -0.05733002 -0.008071658
Revenue.from.3.day.customer.passes              0.011673459 -0.018794415  0.03271376  0.014208593
Total.Revenue                                  -0.002378374  0.021539180 -0.04207746 -0.007113616
distance                                        0.235096349 -0.100438489 -0.10463789 -0.033857819
speedkmh                                        0.747665923 -0.217897175 -0.44745468 -0.179713391
age                                            -0.517733818  0.063693904 -0.63255786 -0.545948094
                                                        PC9          PC10          PC11          PC12
```

From the above results, it is clear that tripduration, trips in last 24 hours, miles travelled, annual members added, revenue from annual members added, revenue from 24 hour passes and total revenue are loading heavily in PC1.

In PC2, 24 hour passes purchased, revenue from 24 hour passes, 3 day passes purchased and total revenue generated are getting loaded heavily. From the plot in Figure 20, we can see that PC1 and PC2 are able to explain 47% of variance in data. The arrowheads having same direction indicate the similar variables . For example, Total Revenue and Passes purchased are facing towards the same side. This indicates that they are quite correlated.

Regression Modelling

An OLS regression modelling was performed on the data with respect to the revenue generated. OLS chooses the parameters of a linear function of a set of explanatory variables by minimizing the sum of the squares of the differences between the observed dependent variable (Revenue) in the given dataset and those predicted by the linear function. Around 12 models were built and we found out that 24 Hours passes, Miles Travelled and Annual Membership were the most important variables for Revenue. This was corroborated with the help of adjusted R squared values, which was 94.98% for our 11th Model (Figure 20, 21- Mod 11). So, we decided to go ahead with these findings.

Clustering

We have many important categorical variables, hence we need to include them for our cluster analysis. The way to perform such analysis is by using “Gower Distance”. Now, it is required to combine the scaled numerical variables along with other important categorical variables like: zipcodes, weekdays, timeslot, station names, usertype and gender. We are trying to group similar stations based on other variables, so that the group where performance is best is observed can be studied further and similar conditions can be applied to other groups. Hence, the data needs to be clustered on basis of station names. The gower distance is calculated using the daisy function. It ranges from 0.0054 to 0.707 (Figure 23).

In order to determine the optimum number of clusters, silhouette width is calculated by using PAM and plotted. The number of clusters can be seen in Figure 24. After plotting the Silhouette width, the best optimum number of clusters were found to be 6.

PAM is used on the gower distance to find the final fit. When the dendogram was plotted, it appeared to be very messy. Hence, for a neater and clearer plot of clusters, Rtsne package and functions are used.

Low dimensional embedding is done using Rtsne function and then clusters are allotted

Based on station names. The clusters are plotted using the ggplot function and is shown in the figure below.

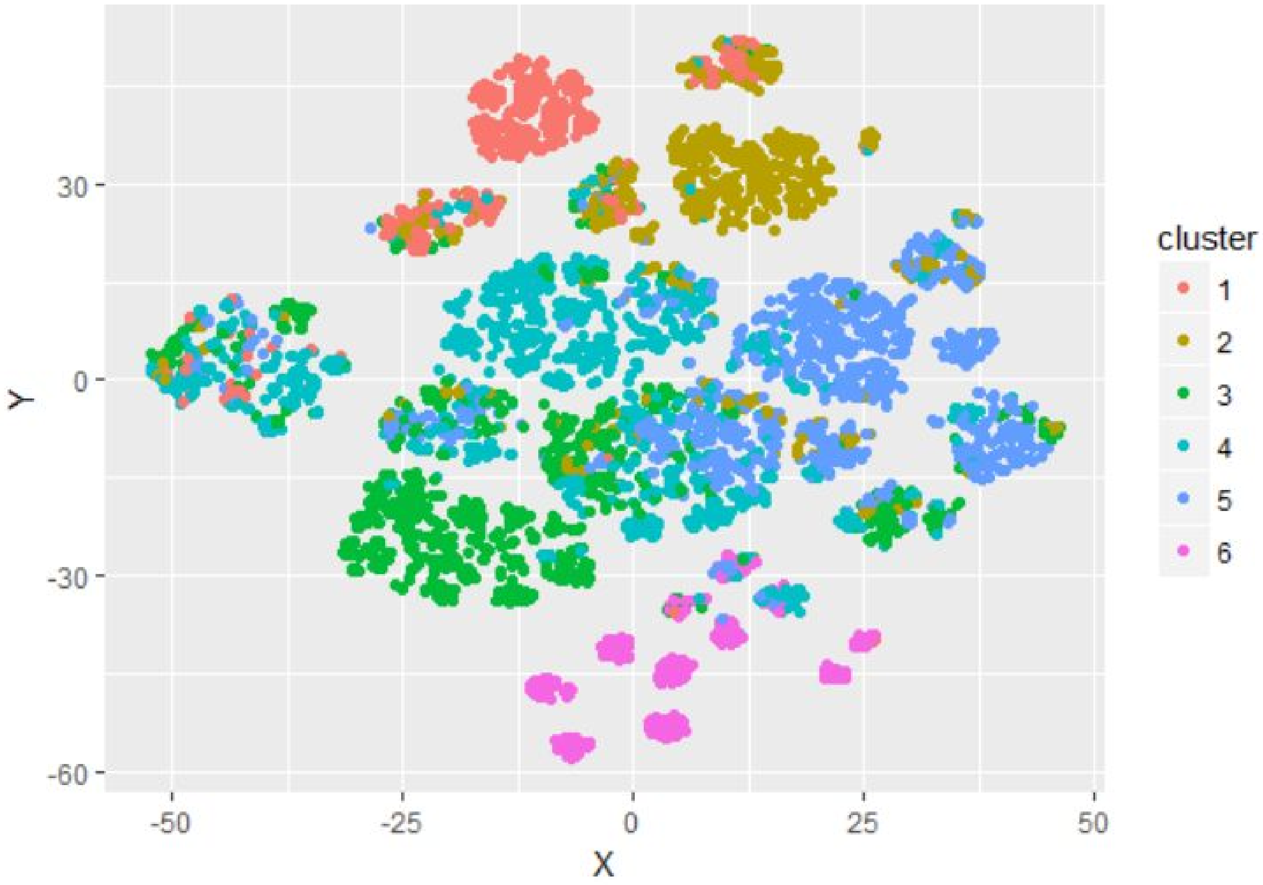
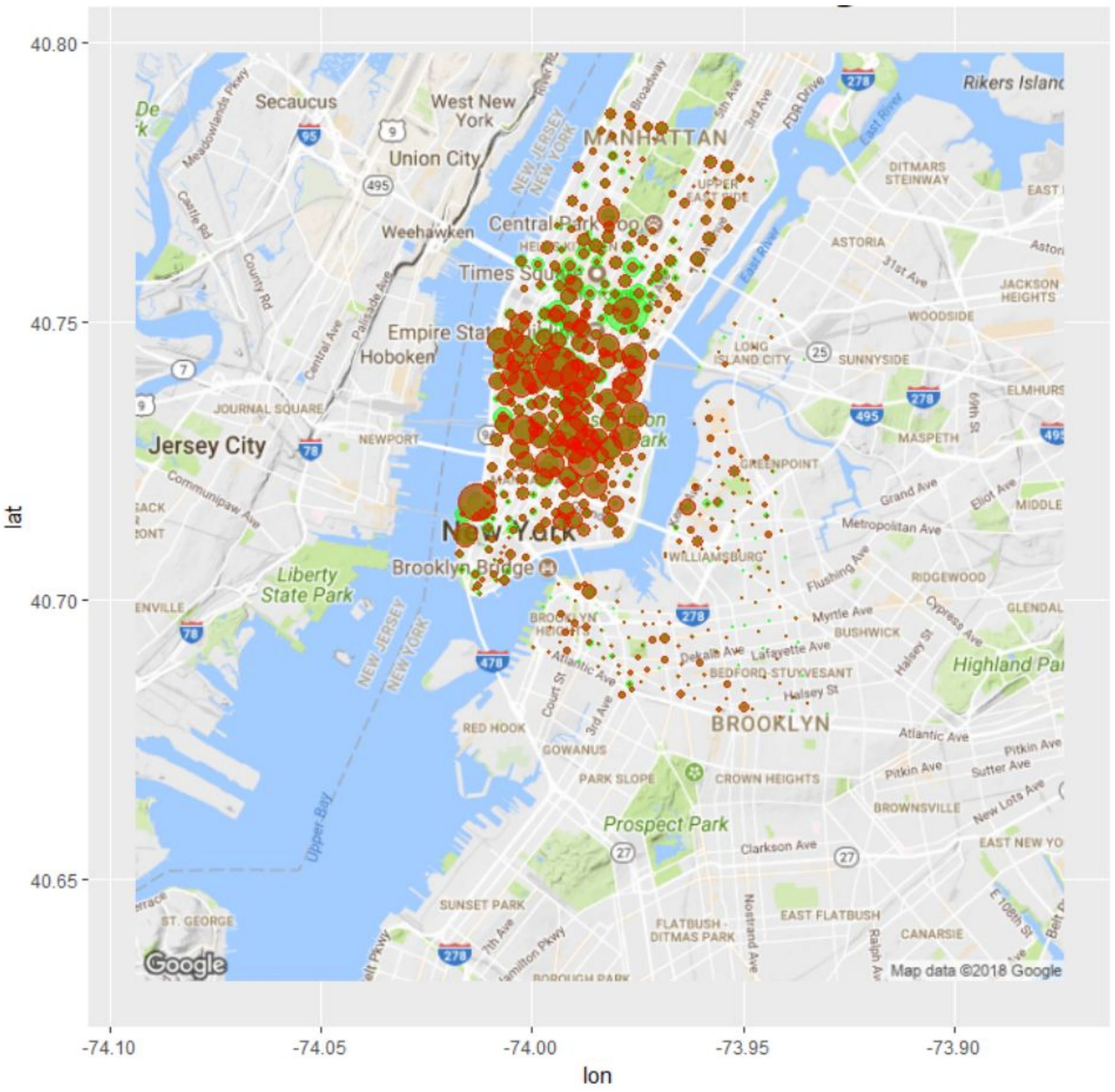


Figure 25

There are 460 stations, we cannot see the names of each and every station.

The below map plotted using ggmap shows the important stations in New York which are derived from the previous result.



Conclusion

Our analysis has shown that there are factors in our data that change revenue generated by Citi bikes. When using PCA, we found that six principal components explained 84.18% of variance in our data. In our first principal component PC1, tripduration, trips in last 24 hours,miles travelled,annual members added, revenue from annual members added, revenue from 24 hour passes and total revenue are loaded significantly, meaning these are important variables. When using linear regression modeling, we found that 24 Hours passes, Miles Travelled and Annual Membership were the most important variables for revenue with an adjusted R squared of

94.98% for our 11th Model. Aside from our analysis on numerical variables, we also had many categorical variables in our data that may also be able to show us important relationships among the values in each cluster. For our final output, we have six clusters that show us values that are clustered based on similarity, but we were not able to obtain the names of the values for better comprehension of our results. Using ggmap in place of the previous cluster plot returned a nice map that shows the important bike stations in New York City. Our analysis provided useful insight of what is changing CitiBike revenue and what recommendations we should make. However, there were a few limitations that we faced before starting our analysis. First, we wanted to incorporate a crime dataset that may have been useful in our analysis but was only 2016 data. So, we had to discard that data. Also, using some functions with such a large dataset occasionally caused R to crash and many hours were allocated to using the revgeo( ) function to generate zip codes.

Insights and Recommendations

1. 24 hour passes have a great share in revenue. Hence, one of the recommendation could be that even a slight increase in the 24 hours passes would generate a great revenue on a large scale. For instance, adding 50 cents for each pass wouldn't matter much to the customer, however, that would significantly boost the revenue with more number of passes.
2. The locations (Zipcode) of the Citi Bike stations largely contribute to the Revenue. Hence the company could use the clustered result for the more popular location and invest more with respect to locations.
3. The day of the week is also an important factor and it was noticed that the usage is very less on weekends compared to weekdays. So the company could introduce some offers or discounts for those days, because revenue follows the simple formula for number*price. If price is low, the number should increase to maintain the balance.

Appendix

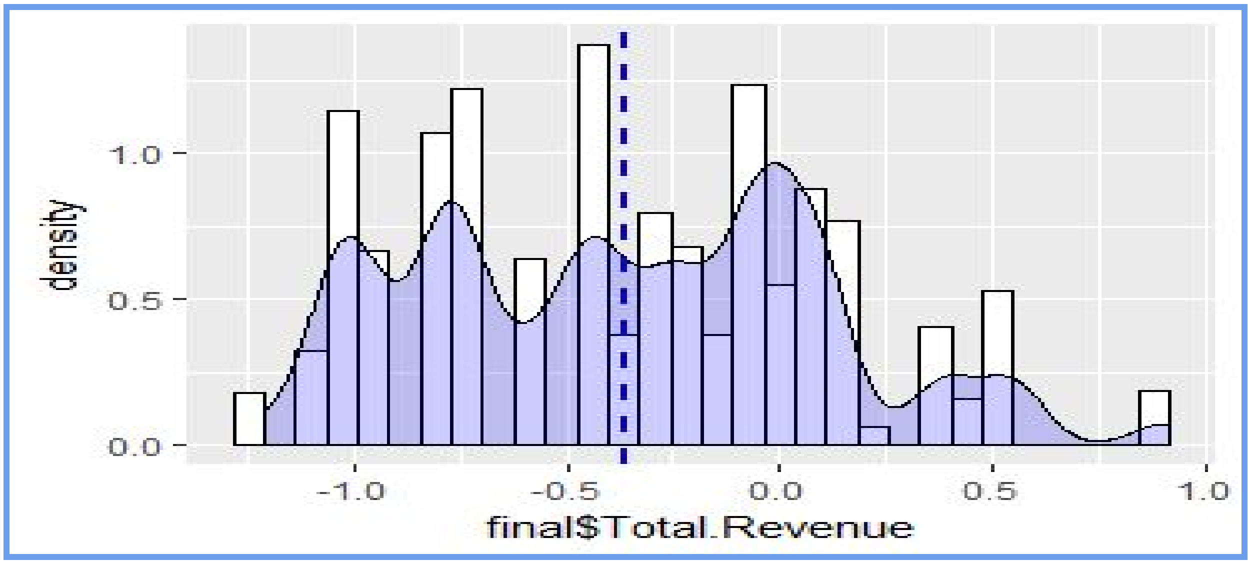


Figure 1: Univariate Analysis of Revenue

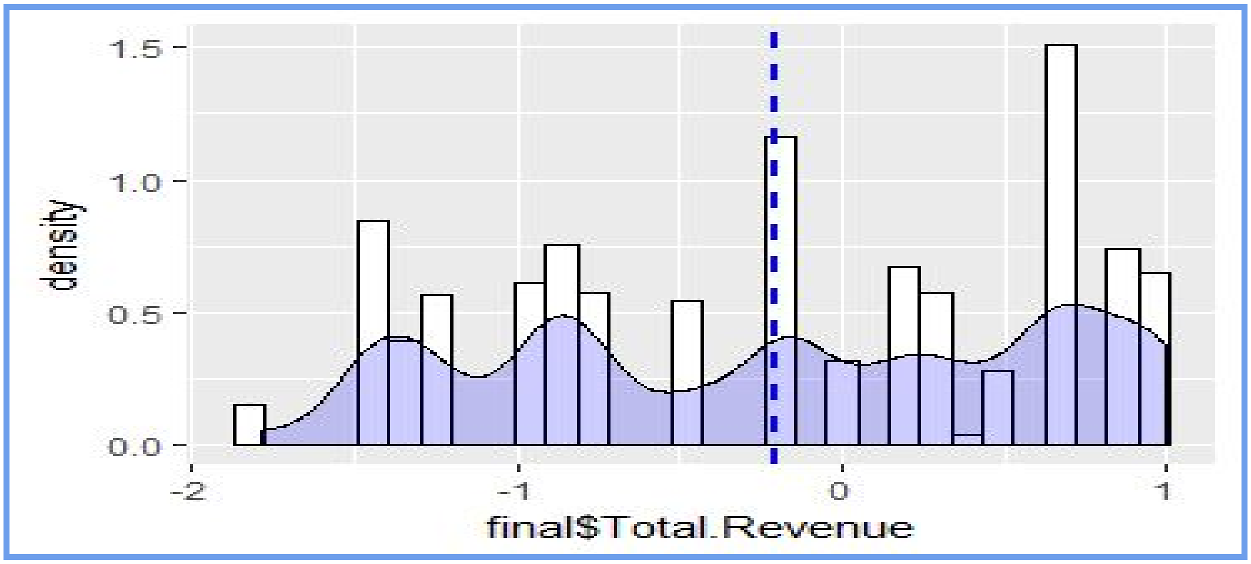


Figure 2: Univariate Analysis of Revenue (after scaling and removing the outliers)

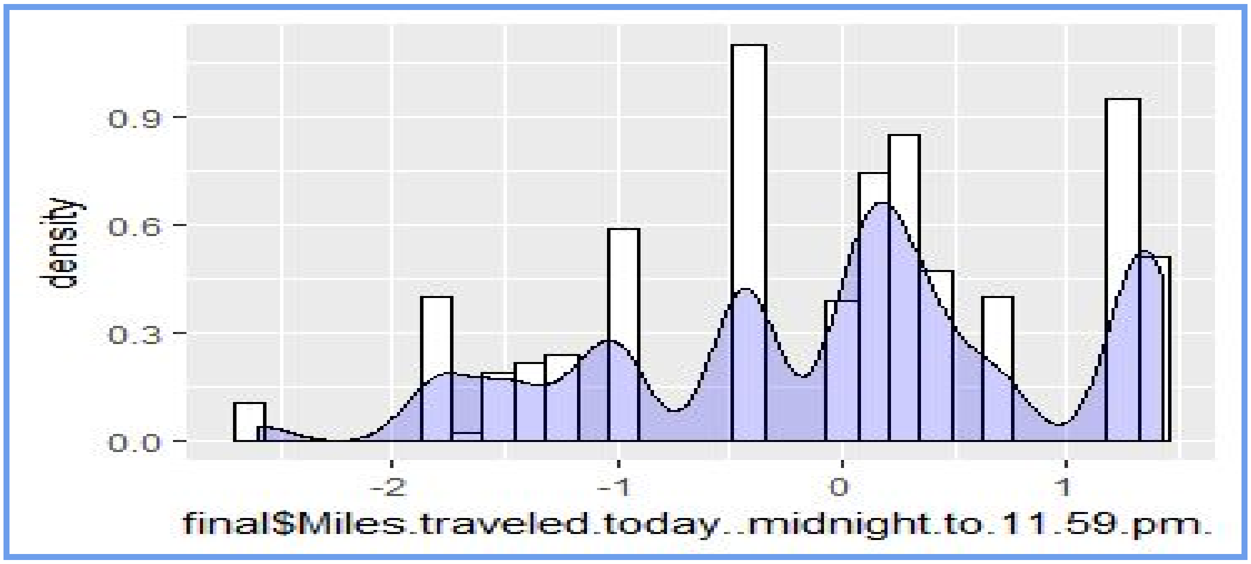


Figure 3: Univariate Analysis of Miles Travelled

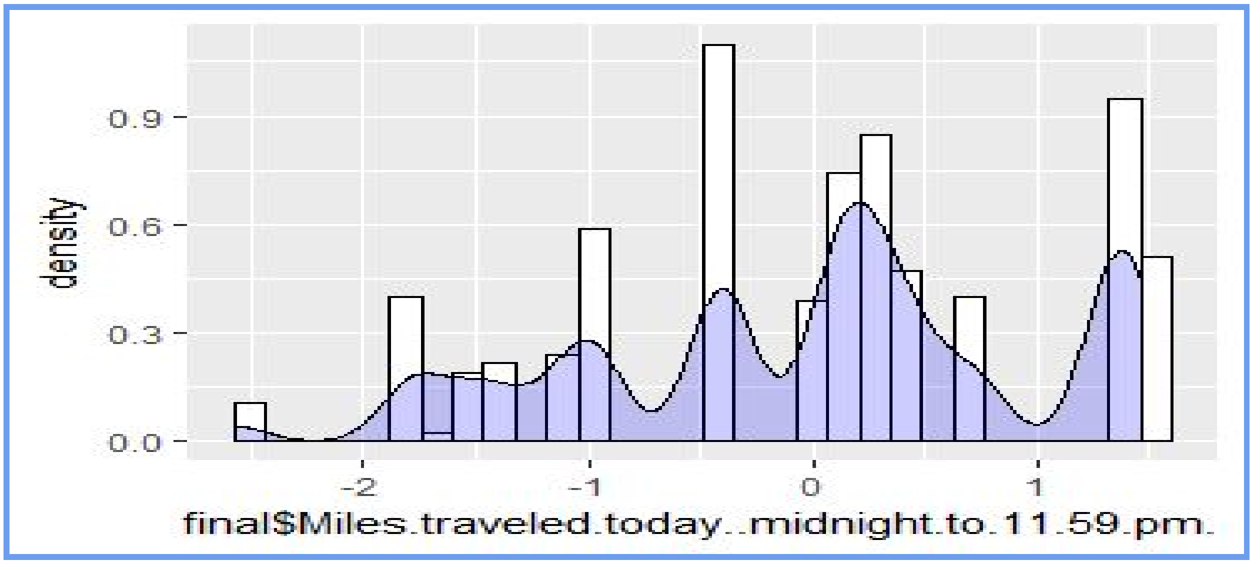


Figure 4: Univariate Analysis of Miles Travelled after scaling

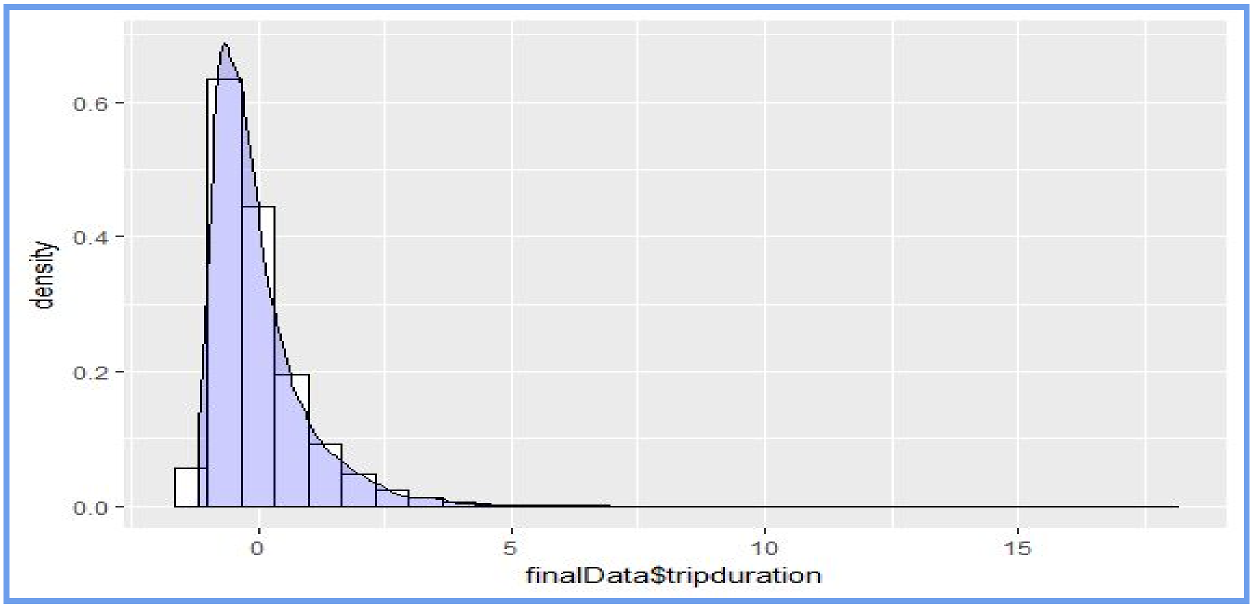


Figure 5: Univariate Analysis of Trip Duration

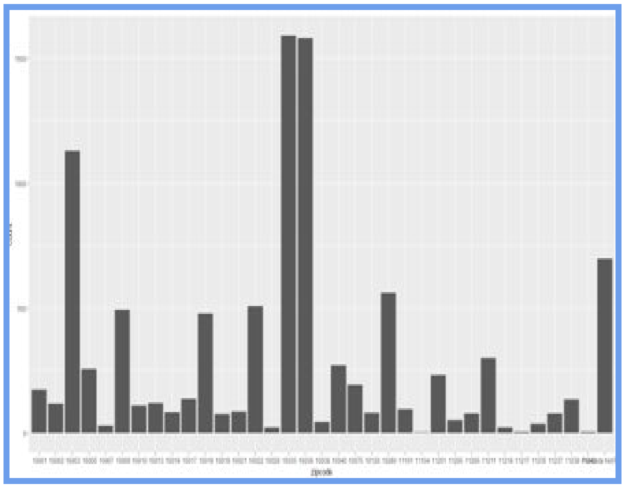


Figure 6: Univariate Analysis of Zip Codes

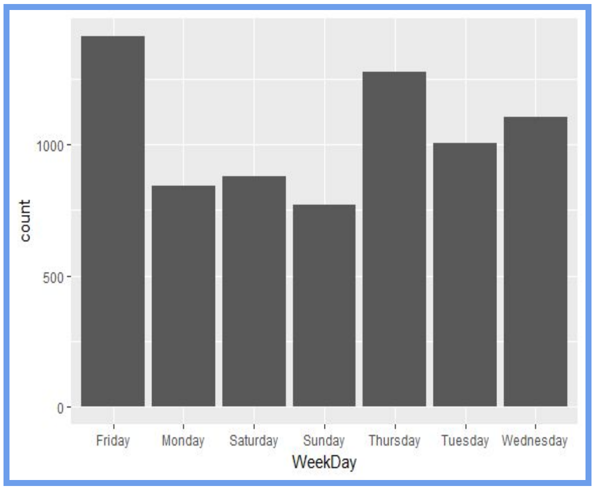


Figure 7: Univariate Analysis of Zip Week Days

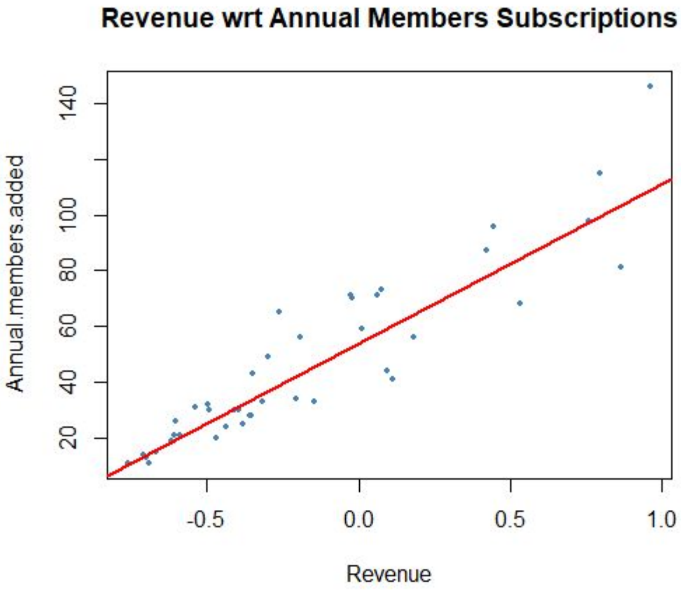


Figure 8.

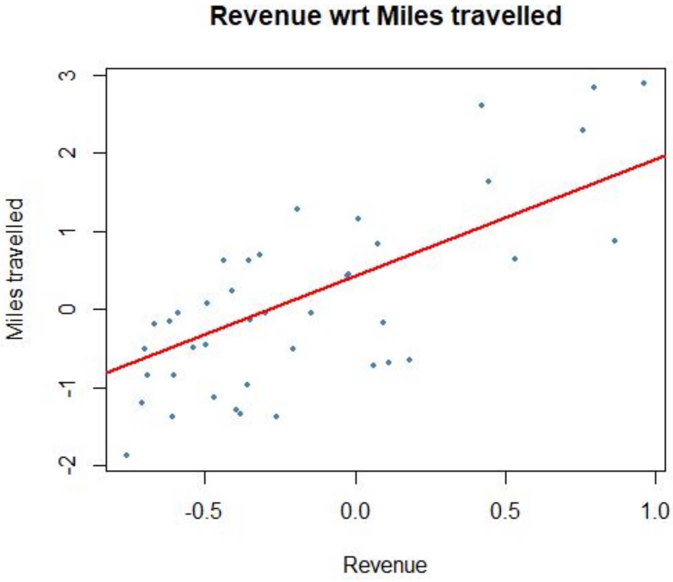


Figure 9.

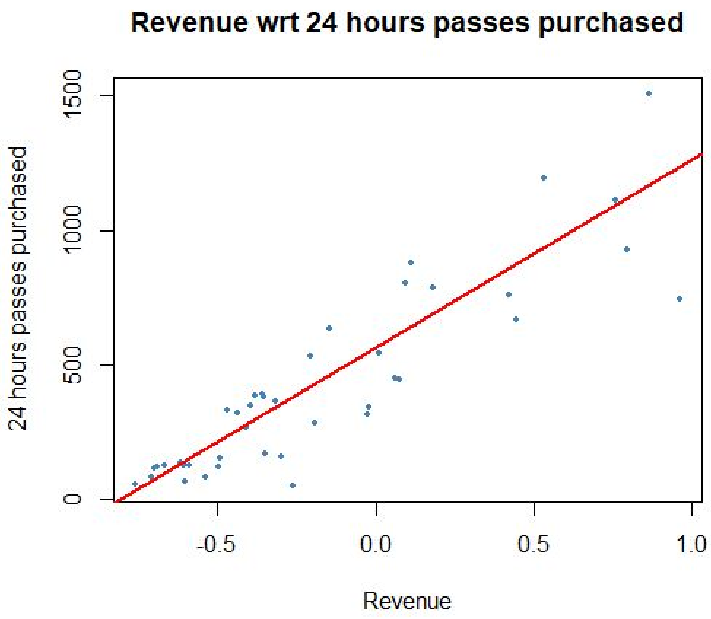


Figure 10

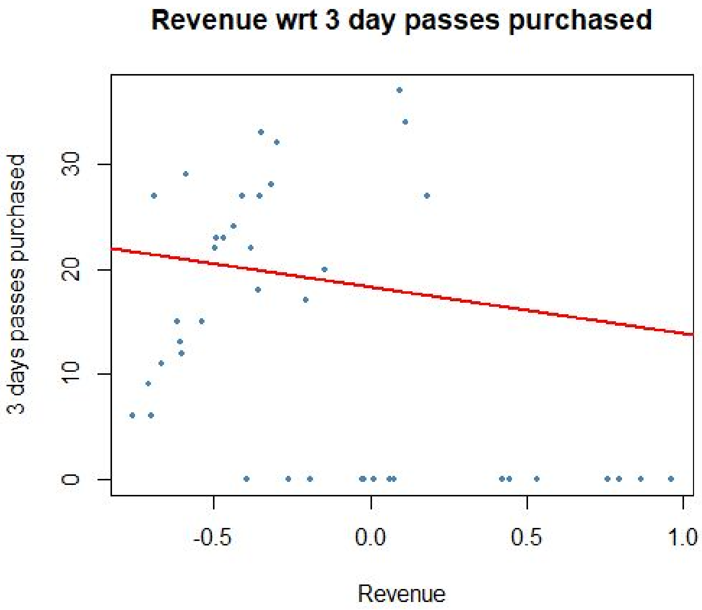


Figure 11

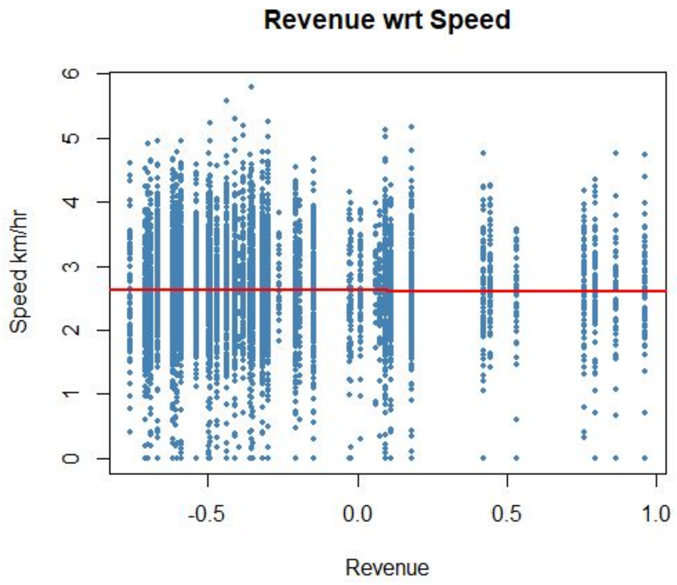


Figure 12

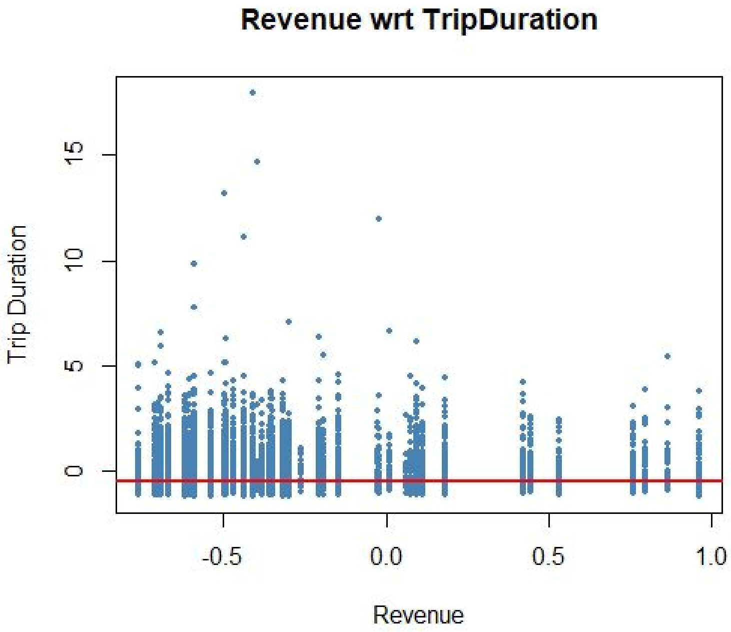


Figure 13

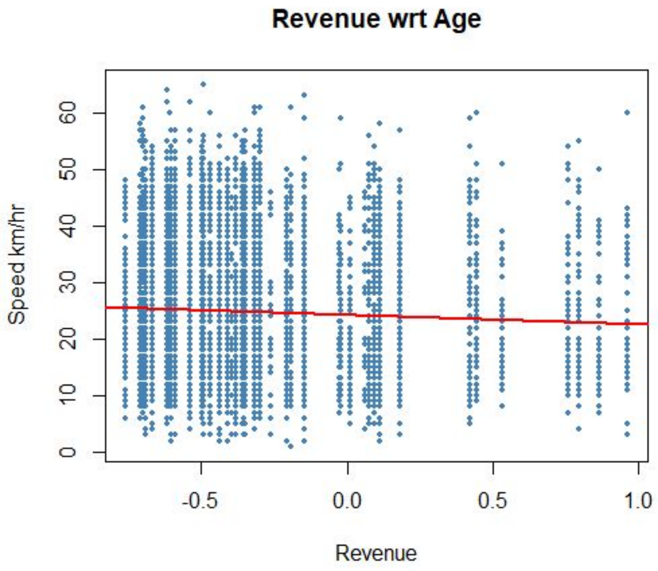


Figure 14

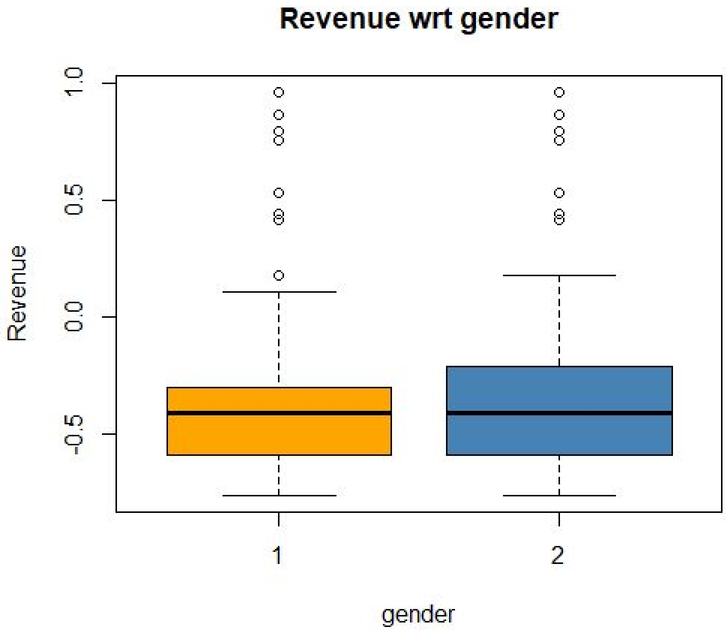


Figure 15

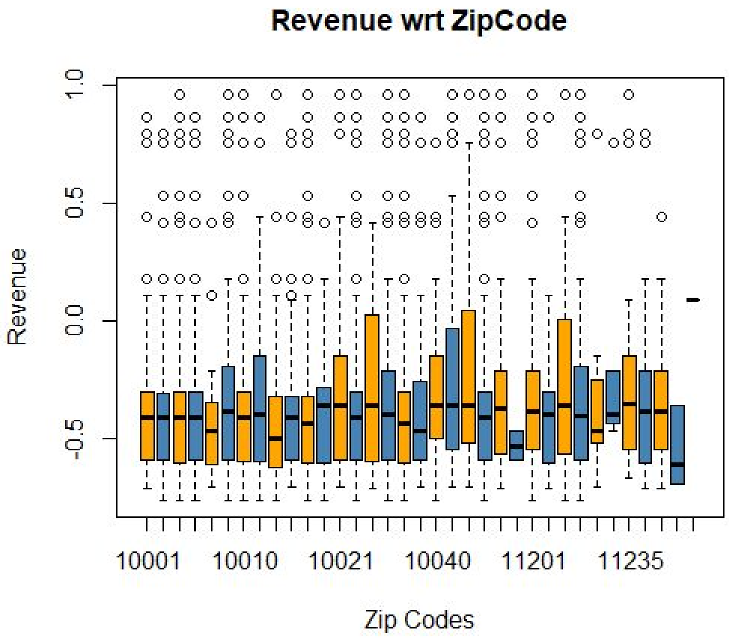


Figure 16

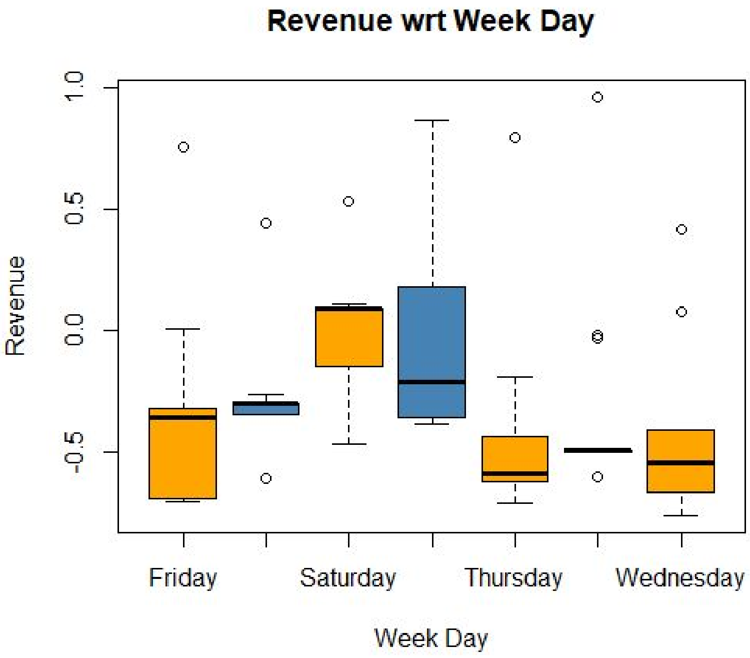


Figure 17

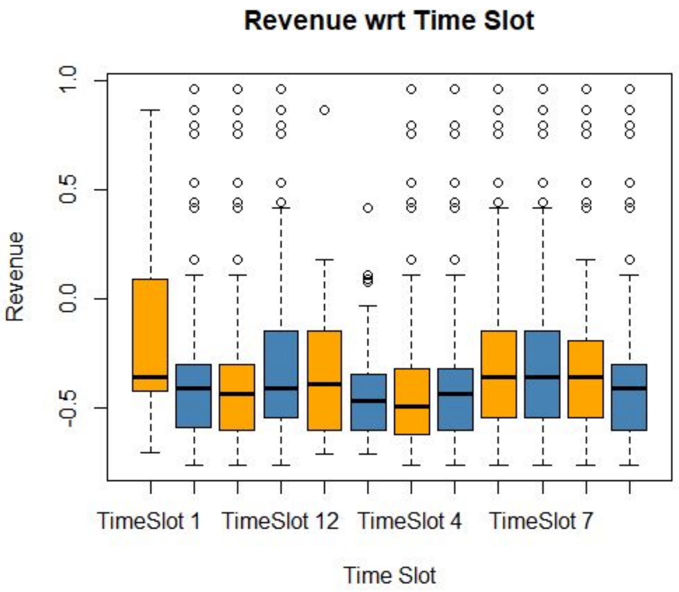


Figure 18

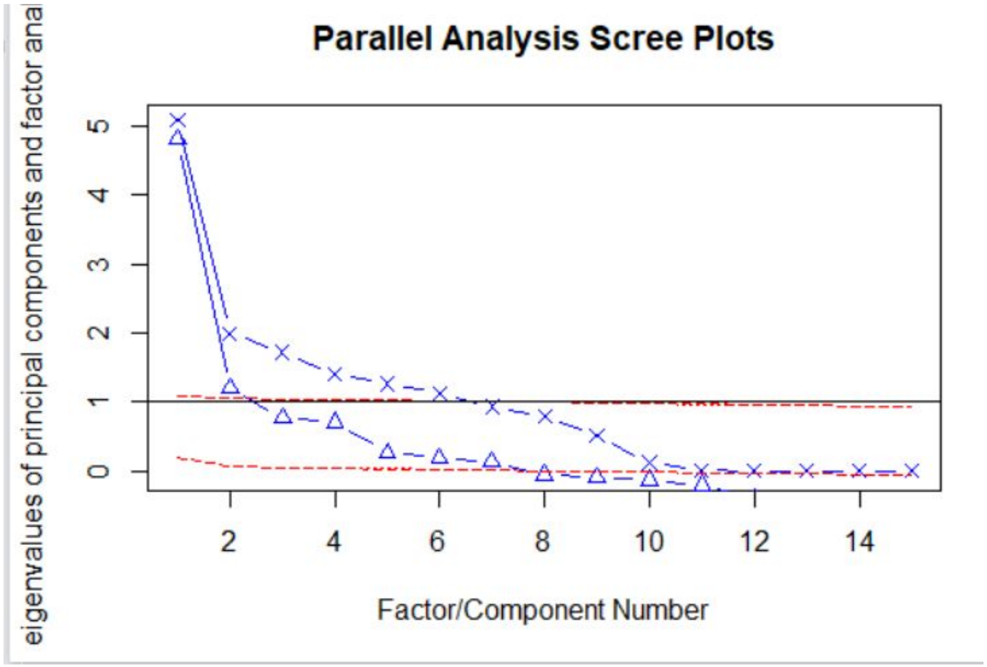


Figure 19

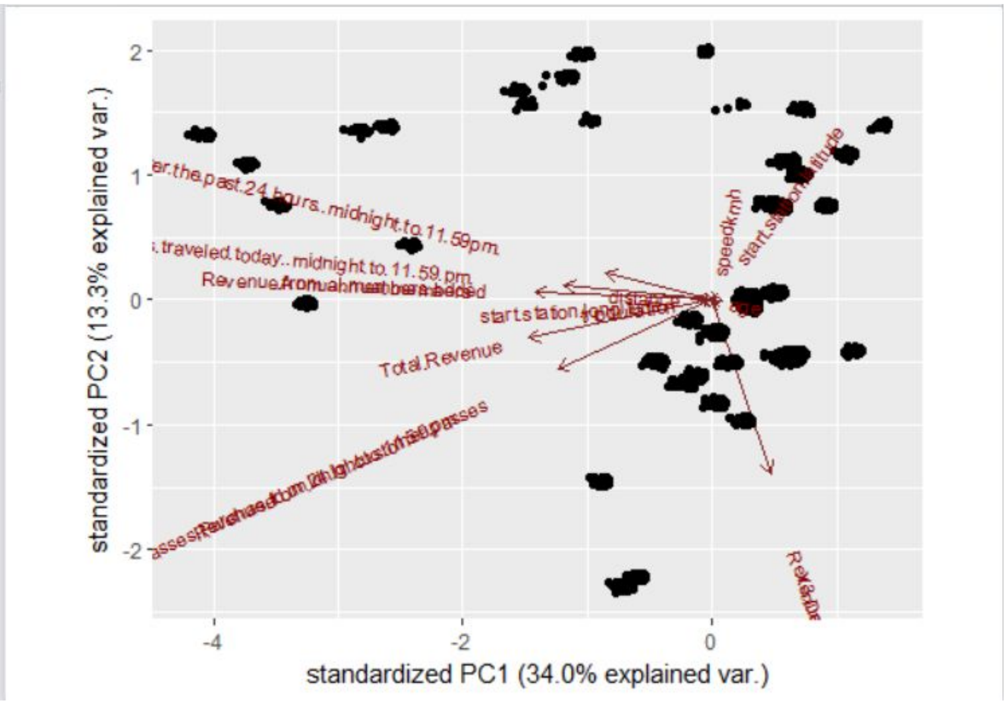
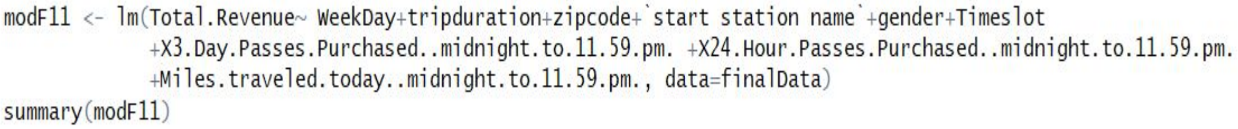
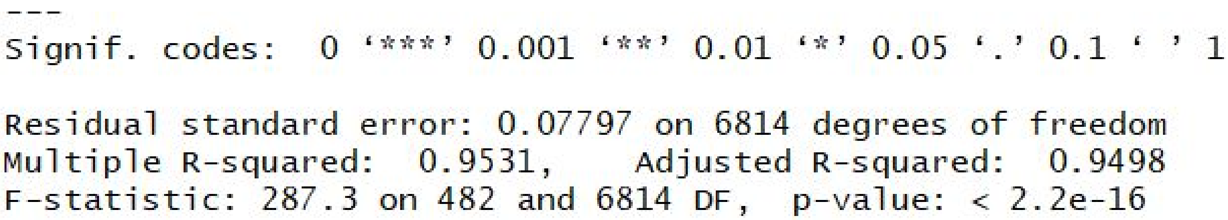


Figure 20



```
modF11 <- lm(Total.Revenue~ WeekDay+tripduration+zipcode+`start station name`+gender+Timeslot
            +X3.Day.Passes.Purchased..midnight.to.11.59.pm. +X24.Hour.Passes.Purchased..midnight.to.11.59.pm.
            +Miles.traveled.today..midnight.to.11.59.pm., data=finalData)
summary(modF11)
```

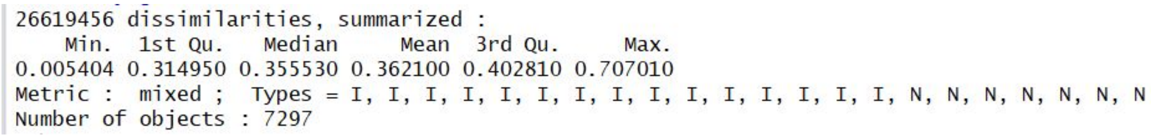
Figure 21



```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07797 on 6814 degrees of freedom
Multiple R-squared:  0.9531,	Adjusted R-squared:  0.9498 
F-statistic: 287.3 on 482 and 6814 DF,  p-value: < 2.2e-16
```

Figure 22



```
26619456 dissimilarities, summarized :
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
0.005404 0.314950 0.355530 0.362100 0.402810 0.707010 
Metric :  mixed ;  Types = I, I, I, I, I, I, I, I, I, I, I, I, I, I, I, N, N, N, N, N, N, N 
Number of objects : 7297
```

Figure 23

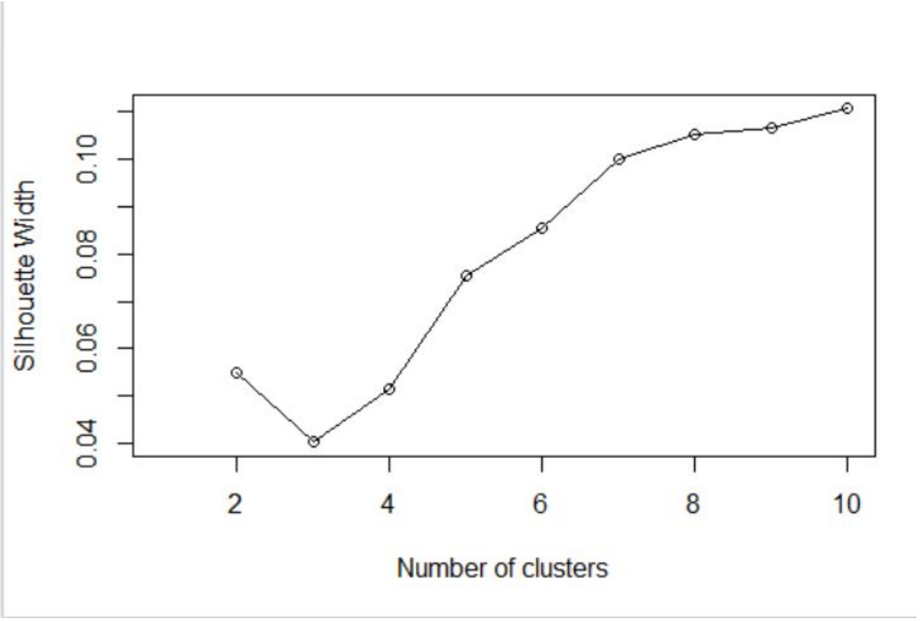


Figure 24

References

1. https://www.citibikenyc.com/system-data
2. https://datawrapper.dwcdn.net/tqNn6/2/
3. https://s3.amazonaws.com/tripdata/index.html
4. http://www1.nyc.gov/site/nypd/stats/crime-statistics/crime-statistics-landing.page