2017


# Analyzing NYC's Citi Bike System

MINERS 'R' US
PRAJAKTA TANDEL
LINDA ZENG
RAJ PAREKH
UDIT GUPTA

# 1. Introduction

Research Question: "What variables will most accurately predict whether it will be a busy day of Citi Bike?"

We believe that this question would be beneficial in optimizing distance between an existing bike station and any new stations. As well as planning for additional docks at existing bike stations.

Citi Bike is the most widely used bike-sharing program in New York City. Our team aggregated 2016 Citi Bike data and merged the weather data of that year in order to research the effects of weather on bike shares.

# 2. Data

## Data collection:

Our primary data comes from the publically available internal Citi Bike systems data at https://www.citibikenyc.com/system-data.

The secondary data comes from Kaggle by way of weather.gov, of the daily weather observations for all of 366 days of 2016, as measured from a Central Park weather station; https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016.

## Data Preprocessing:

A substantial time was spent on data preparation. The original Citi Bike data file listed every trip that occurred. We wrote scripts that would aggregate daily features to work with. We then merged daily weather and daily totals.

## Observations:

Citi Bike data: Observations are recorded in sequential order per ridership. We are using the 2016 data because it is the most recent complete annual data.

Weather data: Each observation is a daily record of weather data for NYC for the year 2016.

Combined data: Each observation depicts the ridership information for each day in 2016.

## Variables:

Citi Bike data: The original data has 15 variables. Categorical variables are start time, stop time, starting station name, ending station name, user type (there are 2 types; "subscribers" and "customers" who purchase short-term, 24-hour, 3- or 5-day passes) and gender.

Numerical variables include trip duration in seconds, starting station ID, starting station latitude, starting station longitude, ending station ID, ending station latitude, ending station longitude, bike ID, and birth year of users.

Weather data: The data has 366 observations with 7 variables. The sole categorical variable is date. Numerical variables are maximum temperature, minimum temperature, average

temperature (all in Fahrenheit), precipitation, snow fall, and snow depth (all in inches). The value T means trace of precipitation, not quantified.

Combined Dataset variables: This is our working data set. The aggregated data has 362 observations of 25 variables:

Variables of Interest:

| |
|---|
| Trips |
| Male |
| Female |
| Customer |
| Subscriber |
| Average Temperature |
| WeekendIndicator -new for project phase 2-Monday-Thursday is 0; Friday –Sunday is 1 |
| Morethan38K-new for project phase 2 |

2 variables will not be used (Unknown and Not Available).

**Date:** Date (Each Date of the month for Year 2016)
**Male:** Numerical (Count of Male Riders)
**Female:** Numerical (Count of Female Riders)
**Below 20:** Numerical (Number of Riders of Age Below 20)
**21 to 30:** Numerical (Number of Riders of Age between 21 and 30)
**31 to 40:** Numerical (Number of Riders of Age between 31 and 40)
**41 to 50:** Numerical (Number of Riders of Age between 41 and 50
**51 to 60:** Numerical (Number of Riders of Age between 51 and 60)
**Above 60:** Numerical (Number of Riders of Age above 60)
**Customers:** Numerical (Total Number of Customers)
**Subscribers:** Numerical (Total Number of Subscribers)
**Maximum temperature:** Numerical (Maximum Temperature for that day in Fahrenheit)
**Minimum temperature:** Numerical (Minimum Temperature for that day in Fahrenheit)
**Average temperature:** Numerical (Average Temperature for that day in Fahrenheit)
**Precipitation:** Numerical (Precipitation in inches)
**Snowfall:** Numerical (Snowfall on that day in inches)
**Snow depth:** Numerical (Depth of Snow in inches)
**Trips:** Total Number of Trips Per Day
**Miles traveled today:** Numerical (Total Number of Miles covered until midnight to 11:59 pm)
**Days:** Day of the Week
**WeekendIndicator:** 1-Friday, Saturday, Sunday, 0-Monday-Thursday
**Morethan38K:** Based on Trips variable, we classify whether each day's trip was more than 38,000 mile or less. 38,000 is approximately the mean and median of the Trips variable.

## Scope of Inference-Generalizability:

The population of interest are local residents of the New York City area, as well as visitors who purchase day or weekly passes. Given the cyclical and seasonal nature of biking as a mode of

public transportation, we feel that this data has good generalizability as applied to year-to-year comparisons and predictions. Citi Bike users have come to rely on bike shares as a reliable mode of transit for work and leisure, thus we can make predictions given weather and variance data. While no personally identifying information is given for each trip except for birth year (and birth year is available for subscribers only), one potential bias is in assuming riders skew toward a younger age. From our exploratory analysis, we find that riders span all ages, with some surprising data about the age group that takes the longest rides.

### Scope of Inference-Causality:

One possible inference would be to analyze the causal link between Citi Bike usage and customer type. We want to find the variable(s) with highest level of causality with usage.

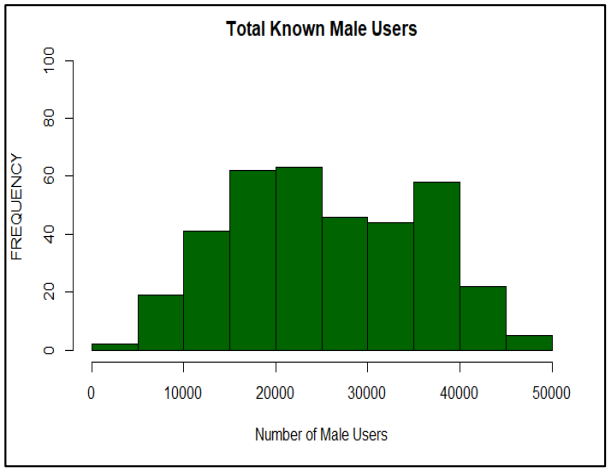
Furthermore, we will use the second weather data set to quantify the correlation between weather and ridership, because weather would predictably affect usage.

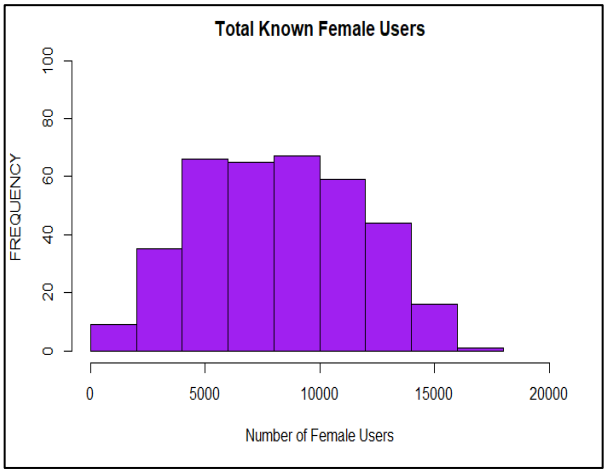
## 3. Exploratory Data Analysis

We aggregated data into daily records. Our initial phase goal is to aggregate each day's values into 1 row; we ultimately have a dataset having 362 records; January 23-26 are missing values due to the historic snowstorm, and we will address this gap based upon feedback.

### Descriptive Statistics and Exploratory Analysis Visualization:

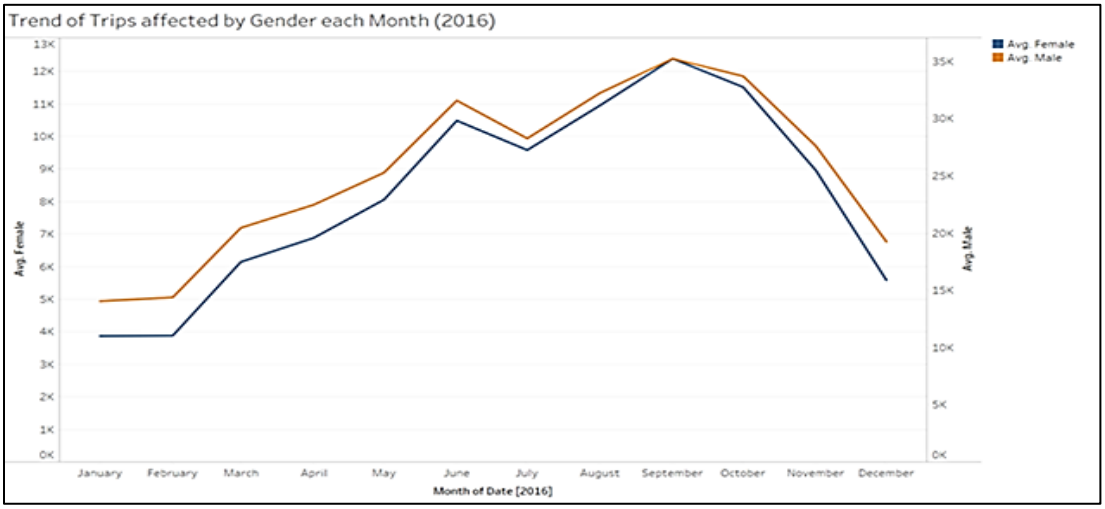
Both genders' usage are positively skewed:



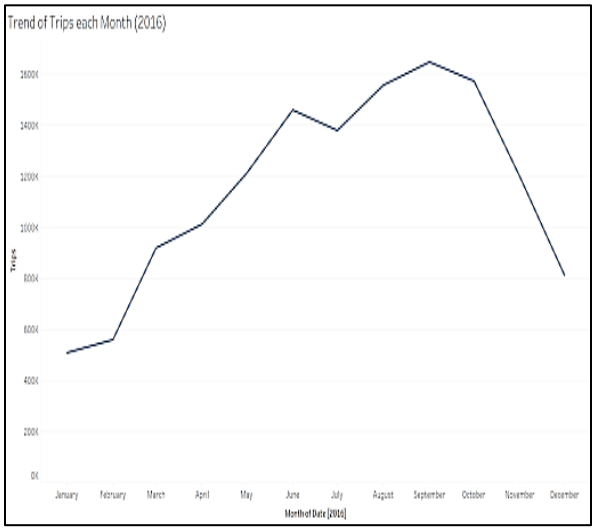



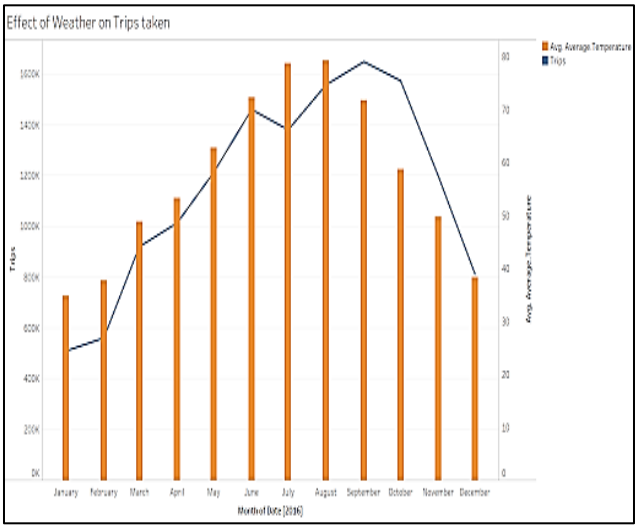
## Trend of Trips affected by Gender

Graph on left shows that males and females behave more or less the same but, there are fewer female riders than male.



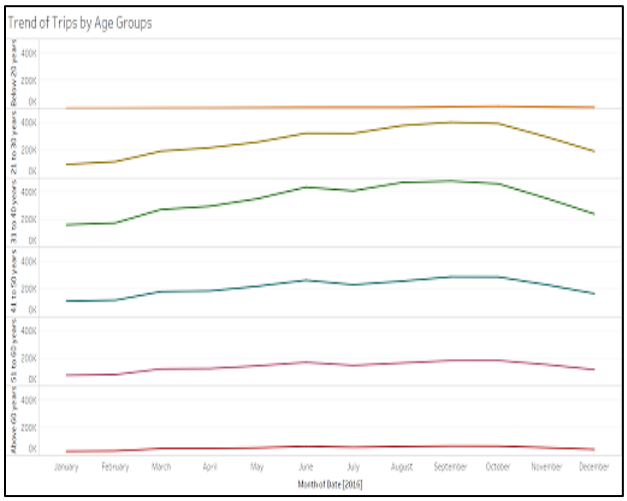

## Effect of weather on Trips

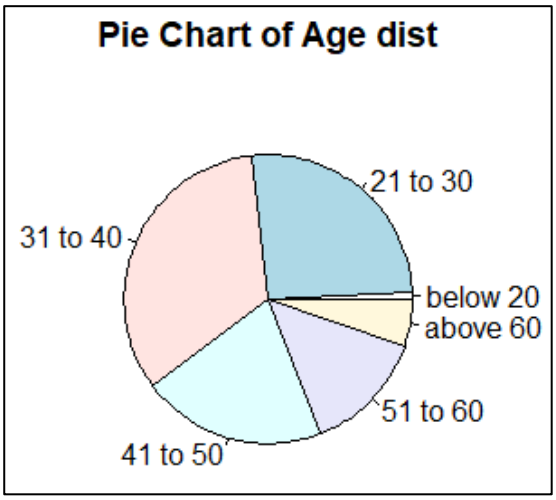



In graphs above, although ridership is predictably dependent on temperature, with a dip in usage in colder months, there are still large number of riders (eg, more than 400,000 trips in January alone) that use Citi Bikes year-round regardless of seasons.

## Trend of trips demonstrated by Age Groups

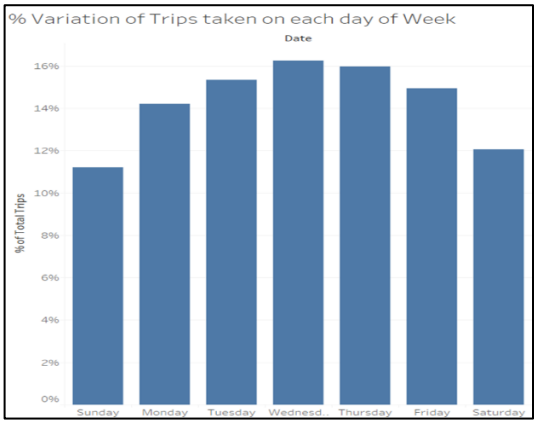


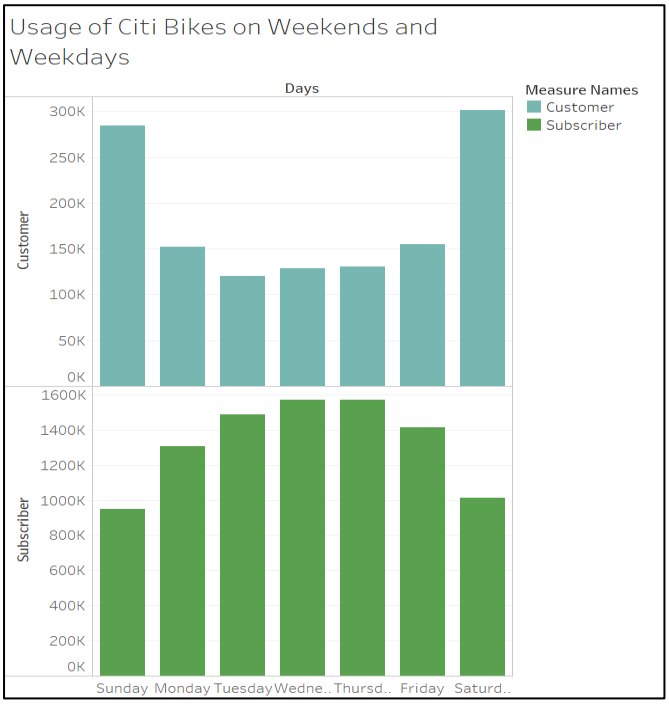




We see that for each additional year in age, the rider tends to be few seconds slower than on your average Citi Bike commuter.

## Variation of Trips on each day of week:

The bar chart shows the percentage of trips taken classified by days of the week. It is observed that highest usage of the bikes is in the mid-week. This can help the operators adjust their inventory of bikes so that more bikes are available in the mid-week.




Created using Tableau and RStudio, the graphs show us how variables such as gender, day of the week, and weather impact usage. For Phase 2 we will examine which set of variables can help us most accurately predict usage.

# 4. Predictive Data Analysis

## Decision Trees

**Sampling data:** Using Decision Tree models, we were able to select which variables best predict usage (MoreThan38K).

### Model 1:

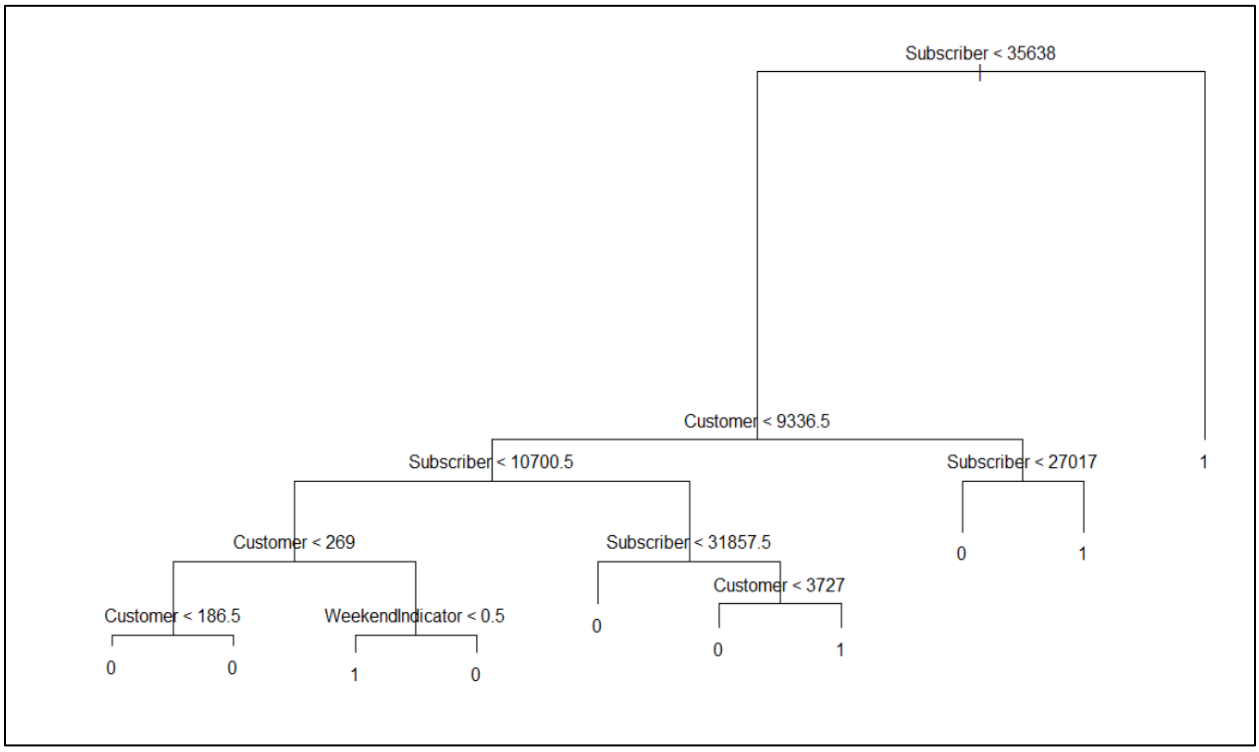
Input Variables: Customer + Subscriber + WeekendIndicator (no Average Temperature, since it is the best indicator of usage, quantitatively and makes real-life sense)

Label: MoreThan38K

Significance of Label: Morethan38K indicates that a given day would be above both the mean (38237) and median (38519) of the entire aggregated data for 2016.

Findings: after sampling the data and applying decision tree model, we find that Customer, Subscriver, and Day of the Week best predicted usage. However, just to be sure, we ran a second decision tree model to account for Average Temperature (another variable) and we find that the accuracy if only improved by 2%, from 86% here to 88% with Avg Temp added.

```
Classification tree:
tree(formula = train_labels ~ Customer + Subscriber + WeekendIndicator,
    data = train_data)
Number of terminal nodes:  10
Residual mean deviance:  0.2387 = 68.04 / 285
Misclassification error rate: 0.06102 = 18 / 295
```



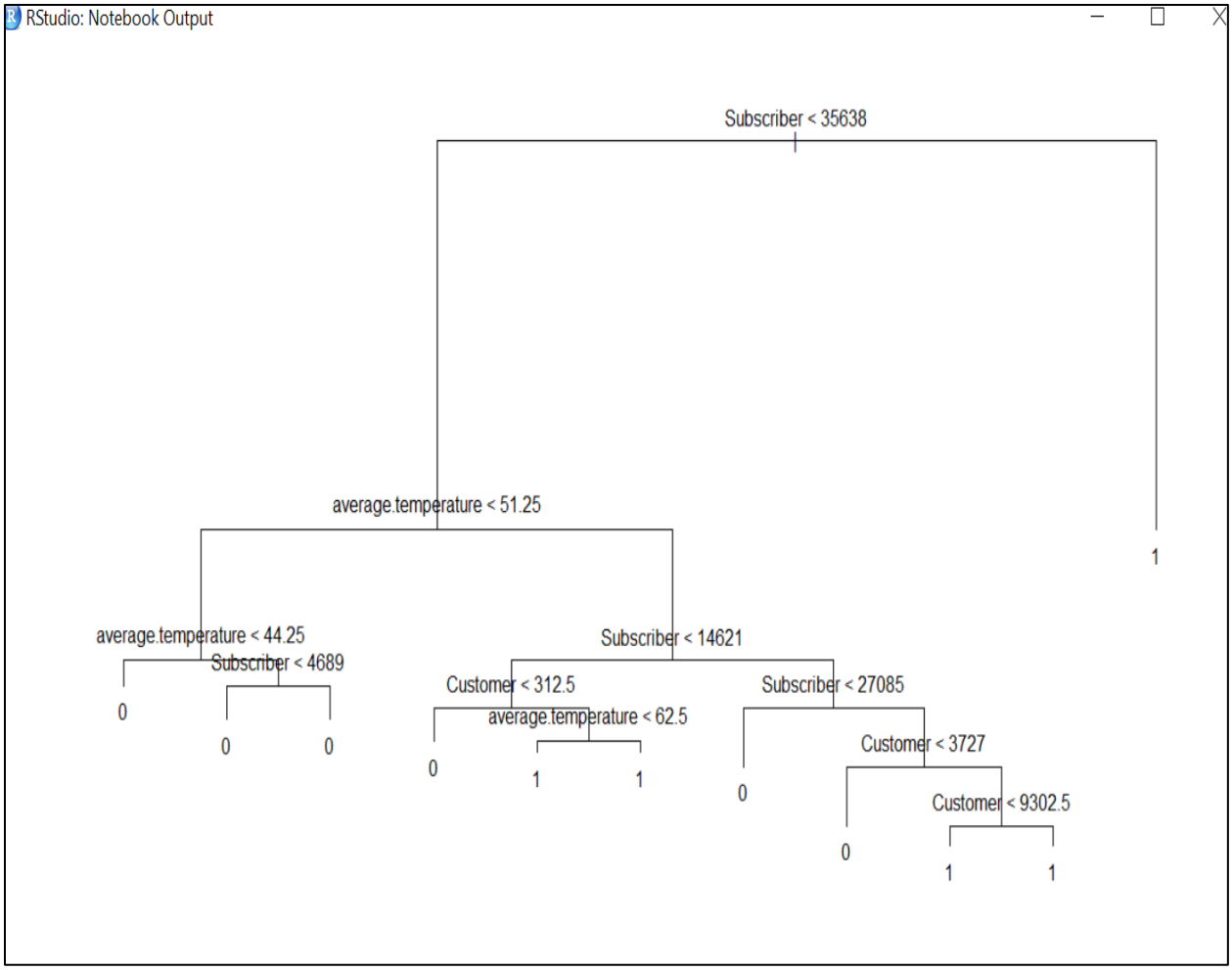
## Model 2:

How will accuracy change when accounting for customer, subscriber, temperature and knowing whether it's a weekday or weekend?

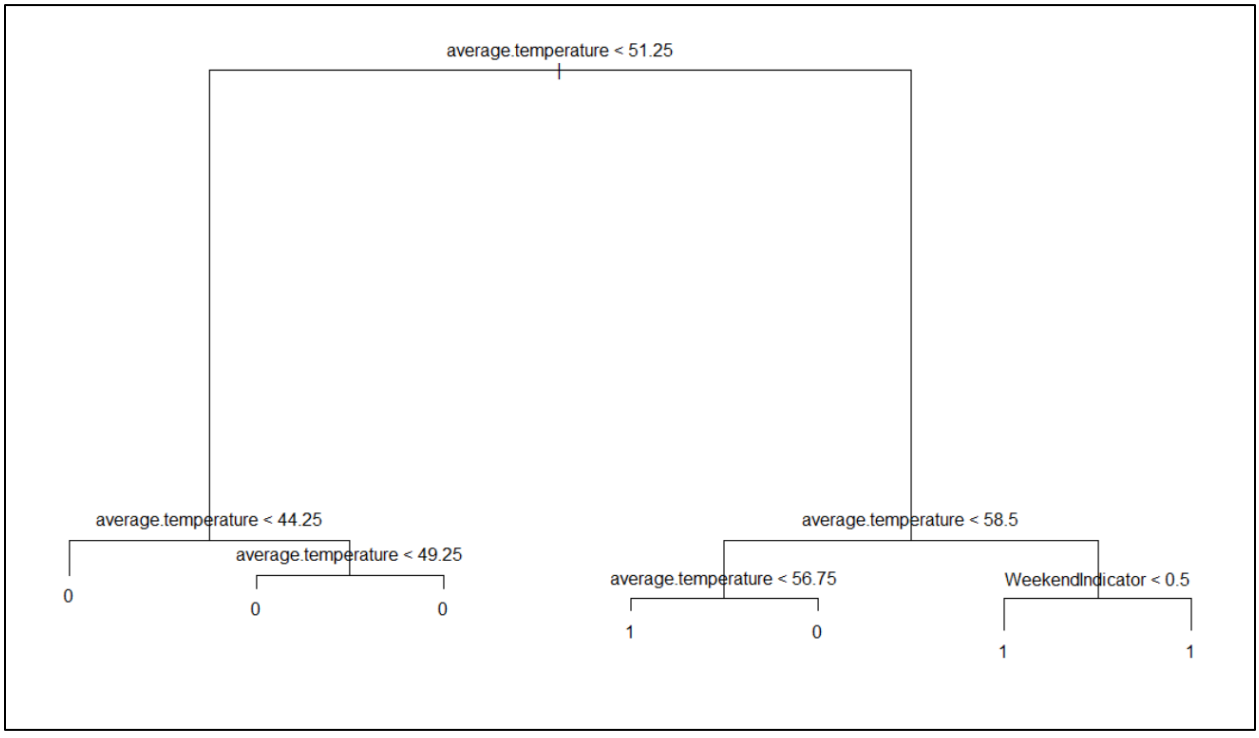
Input variables: Customer + Subscriber + Average Temperature + WeekendIndicator

Label: MoreThan38K

```
my.predictions  0  1
             0 24  6
             1  2 35
Confusion Matrix and Statistics

              test_data.test_labels
my.predictions  0  1
             0 24  6
             1  2 35

               Accuracy : 0.8806
                 95% CI : (0.7782, 0.947)
    No Information Rate : 0.6119
    P-Value [Acc > NIR] : 1.097e-06

                  Kappa : 0.7555
 Mcnemar's Test P-Value : 0.2888

            Sensitivity : 0.9231
            Specificity : 0.8537
         Pos Pred Value : 0.8000
         Neg Pred Value : 0.9459
             Prevalence : 0.3881
         Detection Rate : 0.3582
   Detection Prevalence : 0.4478
      Balanced Accuracy : 0.8884

       'Positive' Class : 0
```


RStudio: Notebook Output
Subscriber < 35638
average.temperature < 51.25
1
average.temperature < 44.25
Subscriber < 14621
Subscriber < 4689
Customer < 312.5
Subscriber < 27085
average.temperature < 62.5
Customer < 3727
Customer < 9302.5
0
0
0
0
1
1
0
0
1
1

```
             test_data.test_labels
my.predictions  0  1
             0 19  8
             1  7 33
Confusion Matrix and Statistics

             test_data.test_labels
my.predictions  0  1
             0 19  8
             1  7 33

               Accuracy : 0.7761
                 95% CI : (0.6578, 0.8689)
    No Information Rate : 0.6119
    P-Value [Acc > NIR] : 0.003304

                  Kappa : 0.5319
 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.7308
            Specificity : 0.8049
         Pos Pred Value : 0.7037
         Neg Pred Value : 0.8250
             Prevalence : 0.3881
         Detection Rate : 0.2836
   Detection Prevalence : 0.4030
      Balanced Accuracy : 0.7678

       'Positive' Class : 0
```



# Logistic Regression

Goal is to try and predict if the uasge of Citi Bikes will be more than 38K (which is the mean of the Trips taken in the 2016 dataset) using Logistic Regression based on demographic variables available in the CitiBike data.

```
Call:
glm(formula = Morethan38K ~ WeekendIndicator + average.temperature +
    Subscriber + Customer, family = binomial(link = "logit"),
    data = trainingData)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.86730  -0.55718   0.06765   0.57432   1.94940

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -7.583e+00  1.027e+00  -7.385 1.52e-13 ***
WeekendIndicator1  -8.260e-01  4.189e-01  -1.972   0.0486 *
average.temperature 1.474e-01  2.105e-02   7.001 2.54e-12 ***
Subscriber         -5.409e-06  1.432e-05  -0.378   0.7057
Customer           -1.048e-04  6.752e-05  -1.553   0.1205
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 354.89  on 255  degrees of freedom
Residual deviance: 205.78  on 251  degrees of freedom
AIC: 215.78

Number of Fisher Scoring iterations: 5
```

## Performance of Logistic Regression Model

- **AIC (Akaike Information Criteria) –**
  The analogous metric of adjusted $R^2$ in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.
  In our model when we used the variables WeekendIndicator and AverageTemperature the AIC value was 220.39. When we changed the variables to WeekendIndicator, AverageTemperature, Subscriber and Customer the AIC value lowered to 215.78.

- **Null Deviance and Residual Deviance –**
  Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.
  In our model when we used the variables WeekendIndicator and AverageTemperature the Residual value was 209.26. when we changed the variables to WeekendIndicator, AverageTemperature, Subscriber and Customer the AIC value lowered to 205.78.

- **Confusion Matrix**

| Actual \ Prediction | 0 | 1 |
|---|---|---|
| 0 | TN | FP |
| 1 | FN | TP |

## Accuracy of Model = (TN + TP) / (TN+TP+FN+FP)

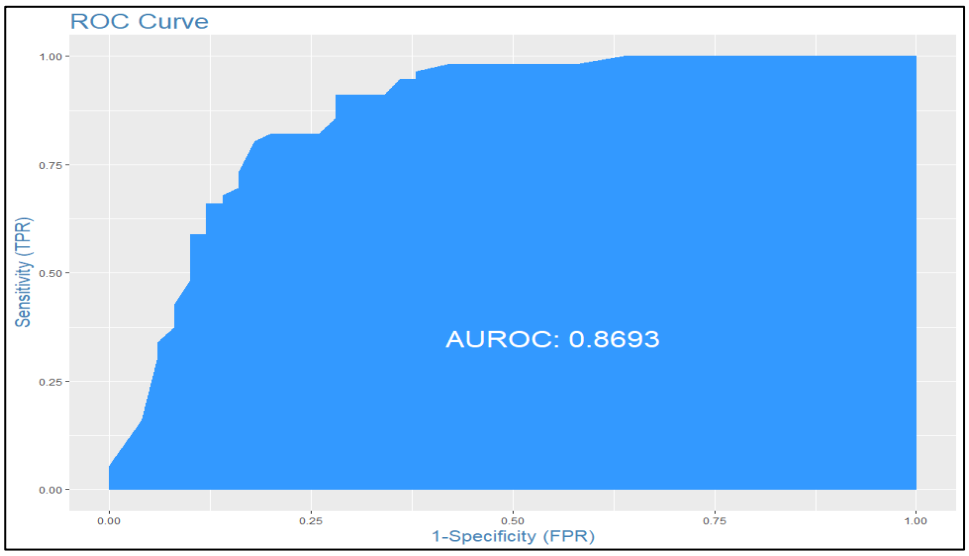
```
> confusionMatrix(testData$Morethan38K,predicted, threshold = optCutOff)
   0  1
0 36  5
1 14 51
```

## Accuracy = (36+51)/(36+51+14+5)

## Accuracy = 0.8207547

- **ROC Curve**

  Receiver Operating Characteristic(ROC) summarizes the predictive power for all possible values of p > 0.5. The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model. Below is a sample ROC curve. The ROC of a perfect predictive model has TP equals 1 and FP equals 0. This curve will touch the top left corner of the graph. The below model has area under ROC curve 86.93%, which is pretty good.



- **Specificity and Sensitivity**
  Sensitivity (or True Positive Rate) is the percentage of 1's (actuals) correctly predicted by the model, while, specificity is the percentage of 0's (actuals) correctly predicted. Specificity can also be calculated as 1 − False Positive Rate.

```
accuracy(testData$Morethan38K,predicted,threshold= optCutOff)
threshold       AUC omission.rate sensitivity specificity prop.correct     Kappa
0.3833166 0.8153571    0.08928571   0.9107143        0.72    0.8207547 0.6368554
```

The above numbers are calculated on the validation sample that was not used for training the model. So, a truth detection rate of 91% on test data is good.

# Conclusion

Using Exploratory Data Analysis and Predictive Models, we find that user type (whether someone has a membership or just a short-term user), and whether it is a weekday or weekend, are the best indicators of usage for a given day. With this information, Citi Bike can optimize their maintenance resources to improve stations and bikes during anticipated higher-usage days.