Citi Bike Ride Duration Analysis Project Report

Manh Tuong Nguyen

2024-11-18

0/ Introduction and Background

1/ Introduction

Bike-sharing programs like Citi Bike have become an integral part of New York City's urban transportation network, offering a sustainable and efficient travel option for residents and visitors alike. This project aims to enhance Citi Bike's operations by analyzing ride duration across various factors, including station locations, membership types, and bike categories. By investigating these patterns, the study seeks to identify the most influential factor affecting ride duration using statistical modeling techniques. The insights gained from this analysis can help optimize station placement, improve resource allocation, and better meet the evolving needs of Citi Bike customers.

2/ Background

Bike-sharing programs have emerged as a popular solution to urban transportation challenges, providing an eco-friendly, affordable, and efficient alternative for short-distance travel. Citi Bike, launched in 2013, is the largest bike-sharing program in New York City, offering thousands of bicycles across hundreds of stations. With millions of annual trips, Citi Bike generates vast amounts of data that capture user behaviors and system performance. Understanding ride duration, a key metric of system utilization, can provide valuable insights into user preferences and operational efficiency. By analyzing factors such as station locations, bike types, and membership statuses, this project seeks to explore how these variables influence ride duration. This background establishes the foundation for identifying ways to optimize Citi Bike's services and improve urban mobility in one of the busiest cities in the world.

3/ Data Importation and Pre-processing

I downloaded the data from the Citi Bike official website. The data was collected monthly from the start of the program until now. Each entry is a ride that is registered through the Lyft's App. For October 2024, the total data has 6 parts, so I only use a half of them for optimizing simplicity and performance. I first read in the data file and drop all the rows that has N/A values since the data is very large. Then, I perform reformatting to ensure all columns are in correct data types. (Appendix 0.1)

1/ Exploratory Data Analysis

1/ Shape and columns

```
## Rows: 2,339,743
## Columns: 10
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bik
e", …
## $ started_at         <dttm> 2024-10-13 17:38:36, 2024-10-06 00:58:33, 2024
-10-…
## $ ended_at           <dttm> 2024-10-13 17:45:22, 2024-10-06 01:02:46, 2024
-10-…
## $ start_station_name <chr> "Rutgers St & Henry St", "Washington Ave & E 16
7 St…
## $ end_station_name   <chr> "Lafayette St & Grand St", "College Ave & E 170
St"…
## $ start_lat          <dbl> 40.71332, 40.82990, 40.86175, 40.82990, 40.8617
5, 4…
## $ start_lng          <dbl> -73.99010, -73.90762, -73.89105, -73.90762, -73
.891…
## $ end_lat            <dbl> 40.72028, 40.83758, 40.87935, 40.80387, 40.8793
5, 4…
## $ end_lng            <dbl> -73.99879, -73.91049, -73.88534, -73.95593, -73
.885…
## $ member_casual      <chr> "member", "member", "member", "casual", "casual
", "…
```

Output 1.1.1: General views of rows and columns in the dataset (Appendix 1.1)

The cleaned data has 2,339,743 rows with 10 columns. The data type of all the columns are well-formatted. Among the features, the interested columns are rideable_type, started_at, start_lat, start_lng, and member_casual. When plotting the usage of by grouping the data by hour and minute, we have the plot:

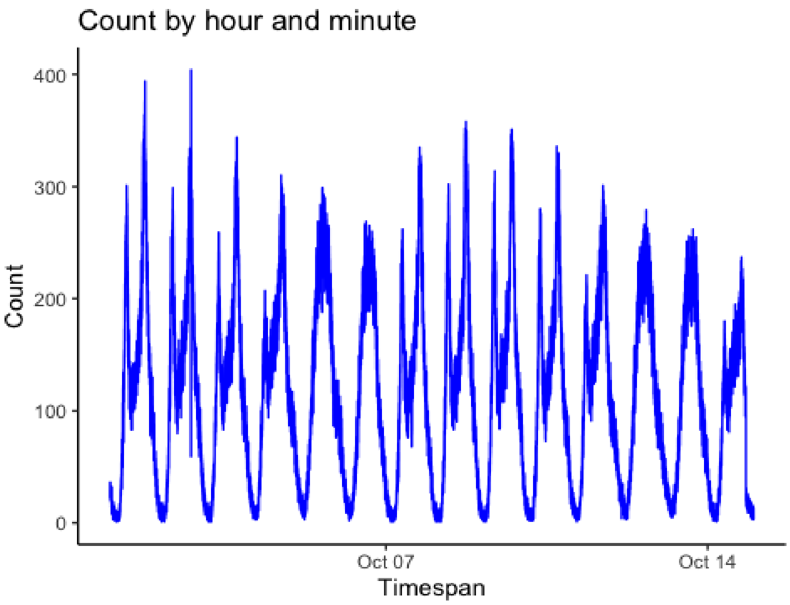
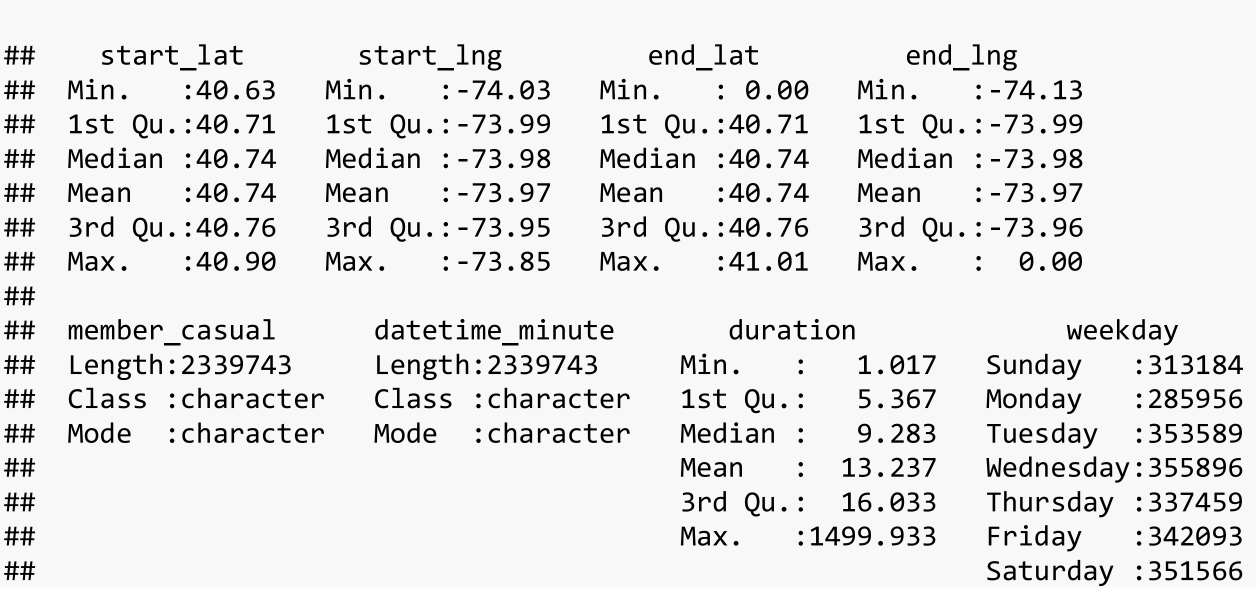


Figure 1.1.1: Timeseries plot of bike usage over time from October 1st , 2024 to Oct 14th , 2024 by hour and minute (Appendix 1.2)

The plot shows that bike usage has seasonality with weekly frequency. The usage is lowest on Monday and hit a medium peak on Wednesday and decrease. Finally, increase to the peak on Saturday then decrease again. This can easily be explained due to the habit of New Yorkers. During weekday, people mainly commute by cars. The usage of bikes in this period is mainly for commute. During weekend, the needs for exercising make the usage goes up to the top.

To prepare for the analysis, I created duration, day of the week, and start time (in hour) (Appendix 1.3) using the started_at data. In general, we have the 5 numbers summary as followed:

```
##  rideable_type        started_at
##  Length:2339743     Min.   :2024-10-01 00:00:01.00
##  Class :character   1st Qu.:2024-10-04 13:14:21.00
##  Mode  :character   Median :2024-10-07 19:02:31.00
##                     Mean   :2024-10-07 21:34:18.43
##                     3rd Qu.:2024-10-11 09:26:54.00
##                     Max.   :2024-10-14 23:59:42.00
##
##     ended_at                        start_station_name end_station_name
##  Min.   :2024-10-01 00:02:15.00   Length:2339743     Length:2339743
##  1st Qu.:2024-10-04 13:27:11.00   Class :character   Class :character
##  Median :2024-10-07 19:14:57.00   Mode  :character   Mode  :character
##  Mean   :2024-10-07 21:47:32.64
##  3rd Qu.:2024-10-11 09:38:02.00
##  Max.   :2024-10-15 16:59:06.00
##
```



```
##    start_lat       start_lng        end_lat         end_lng      
##  Min.   :40.63   Min.   :-74.03   Min.   : 0.00   Min.   :-74.13  
##  1st Qu.:40.71   1st Qu.:-73.99   1st Qu.:40.71   1st Qu.:-73.99  
##  Median :40.74   Median :-73.98   Median :40.74   Median :-73.98  
##  Mean   :40.74   Mean   :-73.97   Mean   :40.74   Mean   :-73.97  
##  3rd Qu.:40.76   3rd Qu.:-73.95   3rd Qu.:40.76   3rd Qu.:-73.96  
##  Max.   :40.90   Max.   :-73.85   Max.   :41.01   Max.   :  0.00  
##                                                                   
##  member_casual      datetime_minute       duration            weekday      
##  Length:2339743     Length:2339743     Min.   :   1.017   Sunday   :313184  
##  Class :character   Class :character   1st Qu.:   5.367   Monday   :285956  
##  Mode  :character   Mode  :character   Median :   9.283   Tuesday  :353589  
##                                        Mean   :  13.237   Wednesday:355896  
##                                        3rd Qu.:  16.033   Thursday :337459  
##                                        Max.   :1499.933   Friday   :342093  
##                                                           Saturday :351566  
```

Output 1.1.2: Five numbers summary of all features (Appendix 1.4)

The output shows that the availability of the data is from 0:00 Oct 1st , 2024 to 23:59 Oct 14th, 2024. The location of all the data are in the area of New York city using the mean. However, the min of end_lat and end_lng shows that we may have data at (0,0). This is an outliers so we will handle this in the Data Cleaning part. Our interest variable is duration, which has the mean of 13.237 minutes but median of 9.283. The maximum duration is 1499.933 minutes (~24.9 hours), which is of course an outlier. This observation may happened due to a case where the ride was not ended correctly or a bug in the system. We will check if there is any outliers in the data in the next part.

2/Data Cleaning

The statistical summary shows that the data has outliers. We need to take care them first before proceeding to the analysis

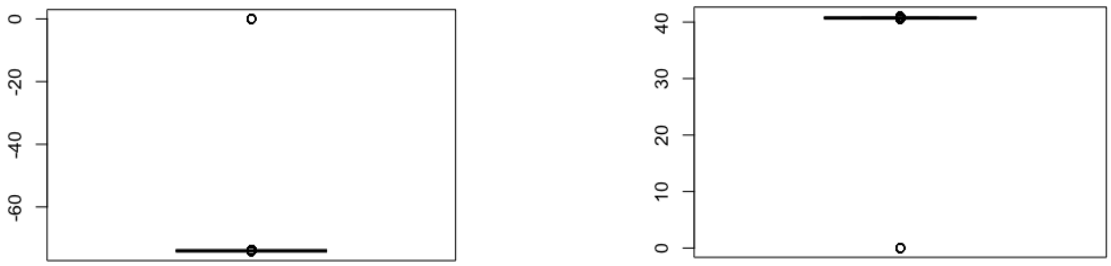


Figure 1.2.1: Box plot of end_lng (left) and end_lat (right) (Appendix 1.5)

The end lat and end_lon contains coordinations that are 0. Since there are only 15 entries looking like that, we are going to remove all of them.

Outliers in duration

As showing in the histogram of duration above, there are a lot of outliers in the dataset for duration. My method is to label the observations with duration is longer than 0.25 x InterQuartile Range from Q3 (Appendix 1.6). Why 0.25? Because I want to subset the data to consider data with significant representation in the dataset. Also, I want to subset as much as I can to improve the processing time of my machine.

By checking the proportion of outliers, we see that there are 17% of the data has outliers. Because the data is very large, we will discard those outliers to clean up the dataset and reduce the dataset's size as well.

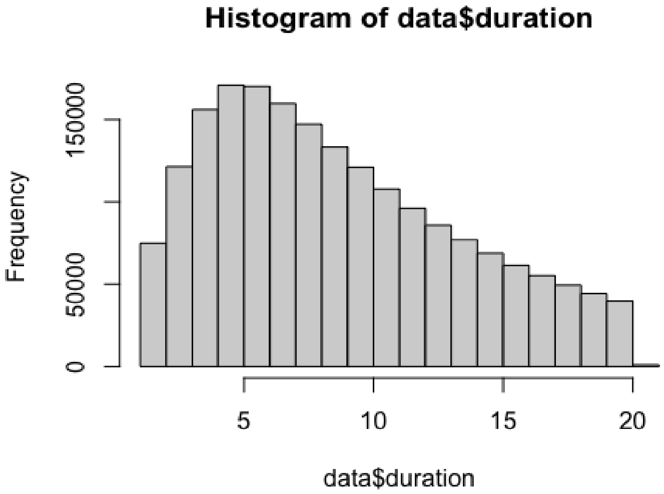


Figure 1.2.1a: Histogram of ride duration after trimming outliers (Appendix 1.6)

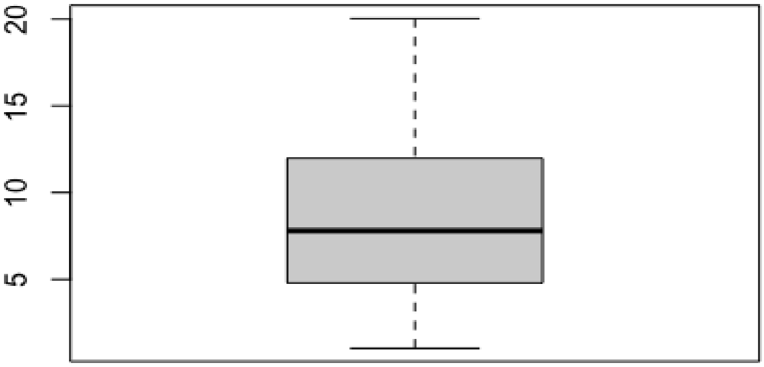


Figure 1.2.1b: Box plot of ride duration after trimming outliers (Appendix 1.6)

After the trim, we see that the duration data is heavily right-skewed. Most of the ride duration are under 10 minutes. The box plot shows that there are no outliers in our data anymore. Plus, the mean of ride duration is around 6-7 minutes per ride across the areas in New York City. This makes sense as geographically, New York is a big city with high volume of traffic. So, bike (specifically Citi Bike) is a popular mean of transportation for anyone who commute within a short distance. For far distance, New York have alternatives like subway, bus, taxi, or personal car that are more convenient and take less effort than bikes. With all the data ready, we will proceed to perform analysis.

In addition, I converted the start time of each ride to hour.minute numerical format to fit the model in the main question (Appendix 1.7)

2/ Subquestion 1: Do trips taken by subscribers differ in duration compared to trips taken by casual users?

1/ Assumption Check

1/ Normality

For this question, I want to perform t-test to test to confirm if the trip duration of rides taken by members differ from rides taken by casual drivers.

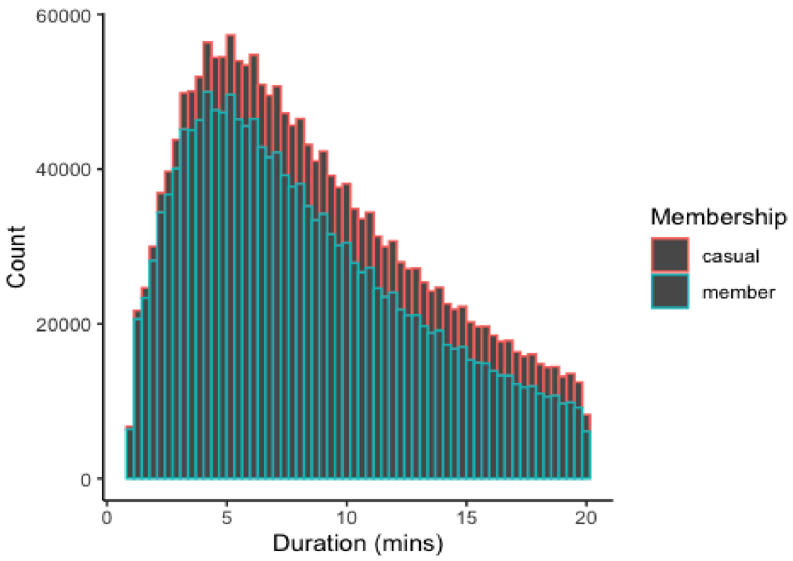


Figure 2.1.1: Histogram of ride duration under 20 minutes by membership (Appendix 2.1)

In the figure above, each bin represents one minute additional to the ride. We can see that casual users tend to rent the bike longer than member users. Let’s conduct t-test to see if memebership type has any effect on renting duration

Our hypothesis will be as followed:

- **Null Hypothesis (H0):** the mean of ride duration for member drivers and casual drivers are not different
- **Alternative Hypothesis (HA):** there is difference in the mean of ride duration for member and casual groups.

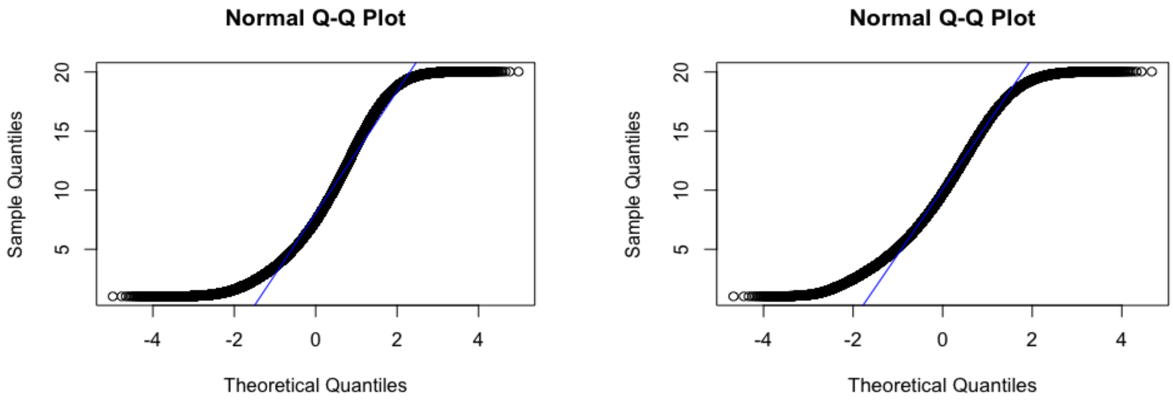


Figure 2.2.1: QQ-plot of ride duration for member users (left) and casual users (right) (Appendix 2.2)

The qq plots suggest the data are heavy-tailed. However, we can see that the middle range of the sample quantile is close to the normal line. Plus, since the data size is very big, we can use Central Limit Theorem to conclude the distribution of the mean is normal.

```
## Member's duration:

## Mean:  8.336255  Variance:  22.20394


## Casual's duration:

## Mean:  10.20369  Variance:  22.28267
```

Output 2.2: Mean and Variance of member group and casual group (Appendix 2.3)

So, we can conclude that member's duration and casual's duration satisfy the normality of mean assumption with the above mean and variance. Also, they potentially have different means and similar variance, too.

2/ Independence

The Citi Bike dataset is primarily composed of individual trips collected from a large number of users in New York City. While some users may have multiple entries, as they could try the service casually before subscribing to a membership, each trip is treated as a separate, random event. The trip durations are independent of each other because the data is randomly collected, and the likelihood of one bike ride's duration influencing another is minimal, which aligns with common sense. Observations in the dataset are independent since each trip represents a unique event, with distinct start and end times, durations, and routes, regardless of whether the same bike or station is involved. Although aggregate patterns, such as peak usage times, may emerge, they reflect group behaviors rather than interdependencies between individual trips. This assumption of independence ensures that statistical analyses can be conducted without bias from any interrelated data points. So, it is safe to conclude the observations are independent from each other based on what we have. Hence, all the assumptions for t-test are passed

3/ t-test

Now we have checked all the assumptions, let's proceed with 2 samples t-test.

```
##
##  Welch Two Sample t-test
##
## data:  duration by member_casual
## t = 207.76, df = 479268, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group casual and
group member is not equal to 0
## 95 percent confidence interval:
##  1.849816 1.885049
## sample estimates:
## mean in group casual mean in group member
##            10.203688             8.336255
```

Output 2.3: t-test result for ride duration in member group vs casual group (Appendix 2.4)

First, t-test's p-value (< 0.05) indicates that we have sufficient evidence to reject the null hypothesis. Hence, the average duration of bike rides of member and casual groups are different. With that being said, membership has influence on ride's duration.

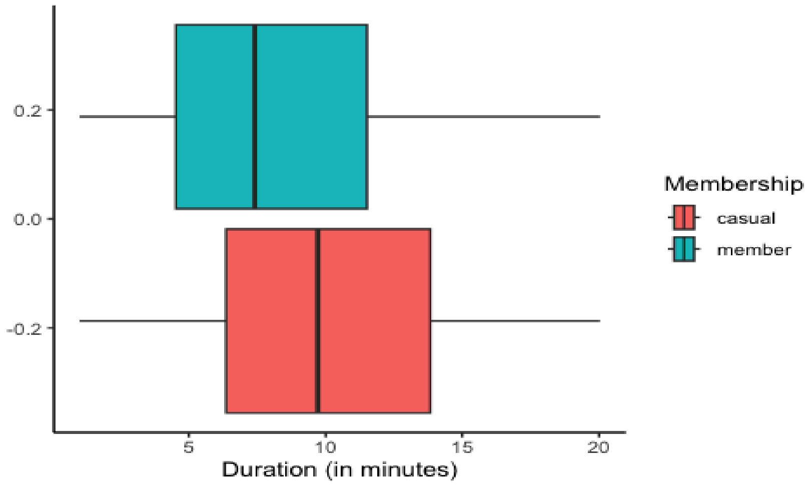


Figure 2.3: Boxplot of duration of casual group vs member group (Appendix 2.5)

Casual users tend to use the bike for longer duration than member drivers. This trend makes sense as people who subscribes to the membership plan usually commute regularly using bike (bike service from this program, specifically). If they commute regularly, it is more likely that the range of the trip is short so that they can save time to find parking or money on gas. For long trips, it is more likely that the bikes are used in emergency cases rather than regular commute. Also, maybe they just want to try out the service. As a result, this insight explain why membership affect the bike duration. Last but not least, membership should be included in our linear regression model for duration.

3/ Sub-question 2: What is the effect of bike type (electric vs. classic) on the trip duration, controlling for the day of the week?

1/ Assumption Check

E-bikes offer an effortless ride with the help of motors, while classic bikes require riders to use their pure strengths to move the bike. In New York City, the streets are often crowded, and bicycles have become a popular means of transportation for many residents, especially for commuting to work (NYC Department of Transportation, 2022). Due to the convenience, e-bikes' rates are significantly higher than that of classic bikes. So, it is interesting to see how the ride duration is for each type of ride with the consideration of the day of the week. This finding will tell about the economic effectiveness of electric bikes compared to classic bikes.

```
## # Groups:   rideable_type [2]
##    rideable_type weekday    count
##    <chr>         <fct>      <int>
##  1 classic_bike  Sunday     80631
##  2 classic_bike  Monday     83978
##  3 classic_bike  Tuesday   105597
##  4 classic_bike  Wednesday 103605
##  5 classic_bike  Thursday   96215
##  6 classic_bike  Friday     90116
##  7 classic_bike  Saturday   86007
##  8 electric_bike Sunday    163900
##  9 electric_bike Monday    160838
## 10 electric_bike Tuesday   196583
## 11 electric_bike Wednesday 200647
## 12 electric_bike Thursday  192549
## 13 electric_bike Friday    197507
## 14 electric_bike Saturday  185215
```

Output 3.1: Number of entries of classic bikes and e-bikes by day of the week (Appendix 3.1)

In all the groups, the numbers of entries is very large. Hence, we can use Central Limit Theorem like sub-question 1 analysis to confirm that the mean distribution of duration among all groups are normal. As confirmed in sub-question 1 analysis, the observations are already independent so we will skip this assumption. However, as we seen in the plot of usage timeseries (Figure 1.1.1), we can see that there is seasonality with weekly frequency in the data. So, if we group the data into day of the week (Monday, Tuesday, etc.), there is a potential that the data are not dependent. Thus, the setting must be comparing e-bike and classic bike groups 7 times corresponding to 7 days a week.

Our hypothesis are as followed:

- **Null hypothesis (H0):** There is no different in trip's average duration between ride type - weekday groups
- **Alternative hypothesis (HA):** There exist a ride type - weekday group that has difference trip's average duration compared to others

2/ Multiple t-test

```
## $Sunday
## 
##  Welch Two Sample t-test
## 
## data:  duration by rideable_type
## t = -40.586, df = 156161, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group classic_bik
e and group electric_bike is not equal to 0
## 95 percent confidence interval:
##  -0.8989706 -0.8161435
## sample estimates:
##  mean in group classic_bike mean in group electric_bike 
##                    8.410639                    9.268196 
## 
```

```
##
## $Monday
##
##  Welch Two Sample t-test
##
## data:  duration by rideable_type
## t = -41.114, df = 167281, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group classic_bik
e and group electric_bike is not equal to 0
## 95 percent confidence interval:
##  -0.8651223 -0.7863919
## sample estimates:
##  mean in group classic_bike mean in group electric_bike
##                    7.792347                    8.618104
##
##
## $Tuesday
##
##  Welch Two Sample t-test
##
## data:  duration by rideable_type
## t = -49.264, df = 212099, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group classic_bik
e and group electric_bike is not equal to 0
## 95 percent confidence interval:
##  -0.9234978 -0.8528272
## sample estimates:
##  mean in group classic_bike mean in group electric_bike
##                    7.951186                    8.839348
##
##
## $Wednesday
##
##  Welch Two Sample t-test
##
## data:  duration by rideable_type
## t = -52.621, df = 206373, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group classic_bik
e and group electric_bike is not equal to 0
## 95 percent confidence interval:
##  -0.9828281 -0.9122428
## sample estimates:
##  mean in group classic_bike mean in group electric_bike
##                    7.900765                    8.848301
##
##
## $Thursday
##
##  Welch Two Sample t-test
##
```

```
## data:  duration by rideable_type
## t = -45.48, df = 188256, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group classic_bik
e and group electric_bike is not equal to 0
## 95 percent confidence interval:
##  -0.8836596 -0.8106426
## sample estimates:
##  mean in group classic_bike mean in group electric_bike 
##                    7.909806                    8.756957 
## 
## 
## $Friday
## 
##  Welch Two Sample t-test
## 
## data:  duration by rideable_type
## t = -41.71, df = 171748, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group classic_bik
e and group electric_bike is not equal to 0
## 95 percent confidence interval:
##  -0.8330136 -0.7582404
## sample estimates:
##  mean in group classic_bike mean in group electric_bike 
##                    8.045209                    8.840836 
## 
## 
## $Saturday
## 
##  Welch Two Sample t-test
## 
## data:  duration by rideable_type
## t = -44.218, df = 163068, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group classic_bik
e and group electric_bike is not equal to 0
## 95 percent confidence interval:
##  -0.9402505 -0.8604351
## sample estimates:
##  mean in group classic_bike mean in group electric_bike 
##                    8.581836                    9.482178
```

Output 3.2: Results of all 7 t-tests of ride duration of classic bikes and e-bikes corresponding to 7 days in a week (Appendix 3.2)

Among all the test results, the p-value are always less than 0.05 suggesting that we can reject the null hypothesis and conclude that the trip's average duration between classic bike and electric bike are diffrent controlling by day of the week. Furthermore, the mean of ride duration of e-bikes is always higher than classic bikes. Hence, it means that the e-bike is mainly picked if people want to go for a long ride while classic bikes are more suitable to short rides. Let's plot them all on a plot to identify the weekly trend.

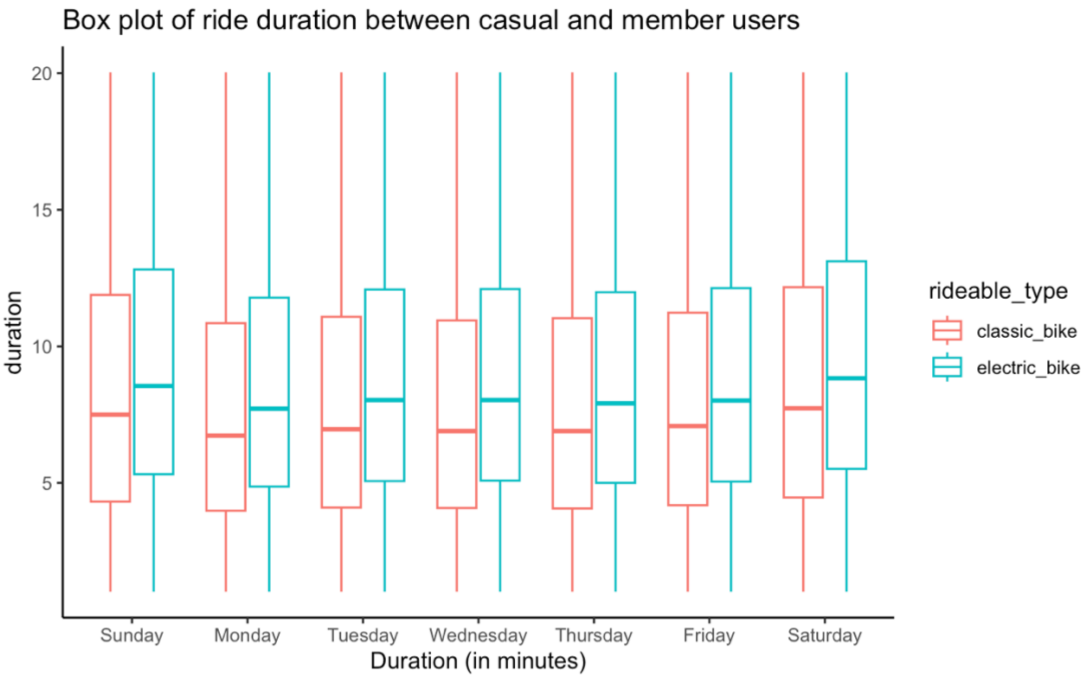


Figure 3.2: Box plot of ride duration on each day of the week by ride types (Appendix 3.3)

Averagely, the duration of rides of both classic bikes and e-bikes peak during weekend days. This can be explained because people who want to use the bikes as a way to enjoy the fresh air or exercise after a busy week contribute to the surge of demand. As a result, Citi Bike should have a routine distribution of bike, meaning that they should store bikes in warehouse during the week and release them fully during the weekend to minimize the chance of bikes become damaged due to frequent usages. Also, we can conclude that rideable type and day of the week are significant predictor for the duration linear regression model.

4/ Sub-question 3: How does ride duration differ at different starting locations?

For this question, I want to check which region tend to have longer rides compare to others at different time a day, in other words, the ranking of demand at different regions. Knowing this will help decision makers gain more insights on where to ship newly charged bikes or schedule more frequent maintenance to meet the demand.

First, I created the regions using grouping. I will focus on five different regions: Downtown Manhattan, Central Park, Uptown New York + The Bronx, Brooklyn, and Queens. Those are the areas that represented in the dataset. The New York main island part was separated into three parts to increase classification accuracy as it is very broad and dense.

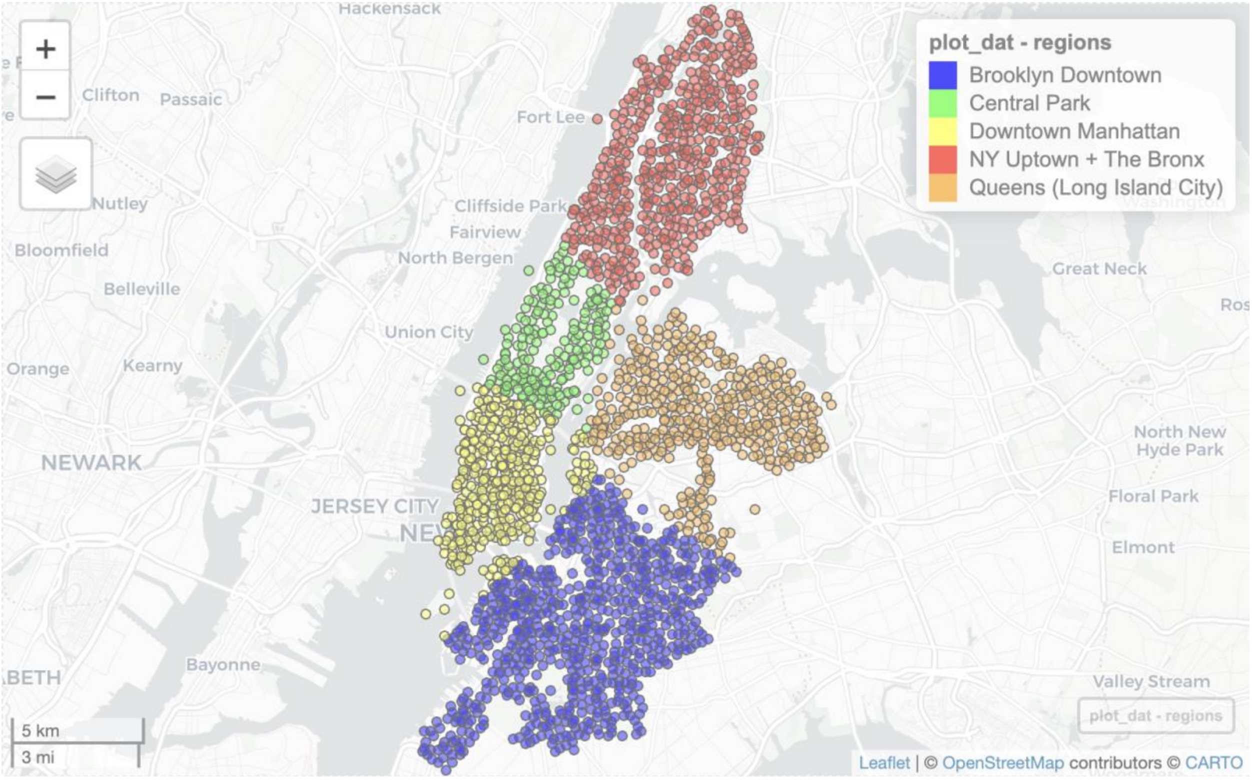


Figure 4.1: Map of all the start station regions in New York City using Mapview (Appendix 4.1)

Due to the size of the data set, it will take a lot of time to use API to map the start stations coordination to the exact region. So, I created five central coordinates corresponding to five big areas and performed clustering using Euclidean distance (Appendix 4.2). In Figure 4.1, we can see that there are some mismatches of Downtown Manhattan stations near Brooklyn Downtown. It is because of the points that we used as the center. However, the grouping shows the distinction between the areas so we are good to go.

1/ Assumption check

The plan is to use Analysis Of VAriance (ANOVA) to compare the average ride duration at five regions. Again, since each entry is a random ride made by a random user, we can safely assume the independence assumption is satisfied. Also, due to the data size is large, we can use Central Limit Theorem like sub-question 1 analysis to conclude that the distribution of average duration in all the five regions is normal.

```
## Mean and variance of different regions:

## Downtown Manhattan : Mean:  8.641411  Variance:  21.66763
## Brooklyn Downtown : Mean:  8.770029  Variance:  23.78392
## NY Uptown + The Bronx : Mean:  7.892681  Variance:  22.57655
## Central Park : Mean:  9.117114  Variance:  23.42285
## Queens (Long Island City) : Mean:  7.892869  Variance:  22.03885
```

Output 4.1: Mean and Variance of ride duration between the five New York City regions (Appendix 4.2)

Next up, we can see that there is potential different in the means of all 5 groups. Plus. the variance of all groups are very close to each other. Hence, the data passed the potential mean different and homogeneity assumption. So, the data already passed all the assumptions, we will proceed with the ANOVA analysis with the hypothesis:

- Null hypothesis (H0): The mean of duration at all 5 regions are similar to each other
- Alternative hypothesis (HA): There is at least one region that has average ride duration different from other regions.

2/ ANOVA

```
##                   Df   Sum Sq Mean Sq F value Pr(>F)    
## start_region       4   232495   58124    2573 <2e-16 ***
## Residuals    1943383 43905781      23                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output 4.2a: ANOVA of ride duration between the five New York City regions (Appendix 4.3)

Since ANOVA result in p-value < 0.05, we have sufficient evidence to reject the null hypothesis and conclude that there is at least one region that has average ride duration different from other regions. In addition, I will perform Tukey’s post hoc test to identify the differences between groups in pair.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = duration ~ start_region, data = data)
## 
## $start_region
##                                                          diff         lwr
## Central Park-Brooklyn Downtown                   0.3470852498  0.31811850
## Downtown Manhattan-Brooklyn Downtown            -0.1286182285 -0.15195355
## NY Uptown + The Bronx-Brooklyn Downtown         -0.8773483230 -0.91650173
## Queens (Long Island City)-Brooklyn Downtown     -0.8771604336 -0.91880569
## Downtown Manhattan-Central Park                 -0.4757034784 -0.50193021
## NY Uptown + The Bronx-Central Park              -1.2244335728 -1.26537614
## Queens (Long Island City)-Central Park          -1.2242456834 -1.26757733
## NY Uptown + The Bronx-Downtown Manhattan        -0.7487300944 -0.78590211
## Queens (Long Island City)-Downtown Manhattan    -0.7485422051 -0.78833035
## Queens (Long Island City)-NY Uptown + The Bronx  0.0001878893 -0.05052255
##                                                         upr p adj
## Central Park-Brooklyn Downtown                   0.37605200     0
## Downtown Manhattan-Brooklyn Downtown            -0.10528291     0
## NY Uptown + The Bronx-Brooklyn Downtown         -0.83819491     0
## Queens (Long Island City)-Brooklyn Downtown     -0.83551518     0
## Downtown Manhattan-Central Park                    -0.44947674     0
## NY Uptown + The Bronx-Central Park                 -1.18349100     0
## Queens (Long Island City)-Central Park             -1.18091404     0
## NY Uptown + The Bronx-Downtown Manhattan           -0.71155808     0
## Queens (Long Island City)-Downtown Manhattan       -0.70875406     0
## Queens (Long Island City)-NY Uptown + The Bronx  0.05089833     1
```

Output 4.2b: Tukey HSD of ride duration between the five New York City regions (Appendix 4.4)

The result of Tukey HSD told us that all the groups are distinct using p-value. For the Queens - NY Uptown + The Bronx pair, since p-value is not under 0.05, we conclude that the 2 groups are not different in term of ride duration. This result means the average duration of rides taken in the five major areas of New York City are generally different. Let's corporate boxplots to compare the groups to give the ranking of average duration

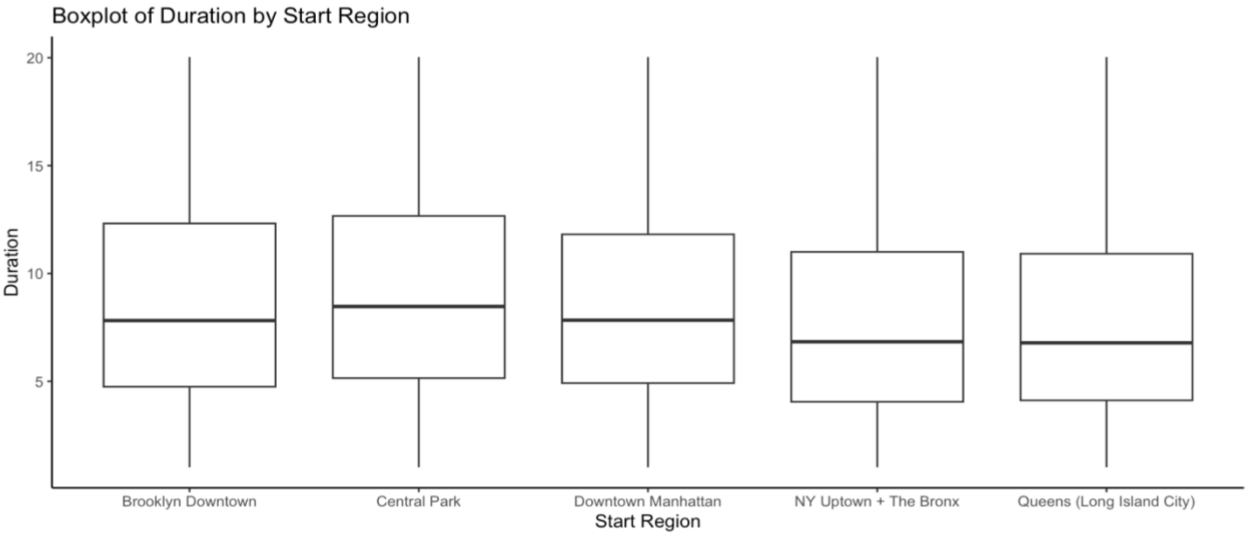


Figure 4.3: Box plot of ride duration between the five New York City regions (Appendix 4.5)

Based on all the results, we can see that the average ride duration in Central Park is longest. Runner up is Brooklyn Downtown then Downtown Manhattan. And the final place are both NY Uptown + The Bronx and Queens. Using this insight, we can conclude two things:

- As the result from sub-question 2, Citi Bike should arrange the most e-bikes in Central Park and promote their use, as customers there are more likely to go for long trips and find them more appealing, while reducing the number of classic bikes. The classic bikes should be put at Queens and NY Uptowns + The Bronx areas because of the low average duration. For short rides, it is more likely that people use it for short commute or exercise. Hence, classic bikes will be more practical for users in those 2 regions

- To improve customer satisfaction, Citi Bike can focus the maintenance team into servicing the bikes in areas with high average ride duration as the bikes there will more likely to travel at long distance so it will more likely to have issues than in regions with low average ride duration.

In general, region of start_station is a significant predictor for duration in our linear regression model.

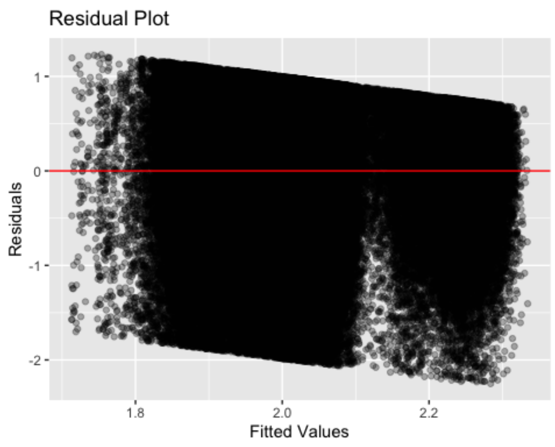
5/ Main question: What factors significantly influence the duration of Citi Bike trips in New York City Central Park?

At first, I wanted to do the whole city scaled, but it is tedious to diagnose the linear model with such large dataset. Hence, I will focus the most on the Central Park region as this is a unique location where bikes can be used for various purposes (commuting, exercising, etc.). We can possibly see the effect of weekday and start time here using the variability in purpose of bike use.

1/ Assumption Check

Firstly, as stated before, since the data was collected using random sampling as each ride does not relate to each other. So, we can safely confirm the data is independence.

So, let's begin with assumption checking by first fitting the model



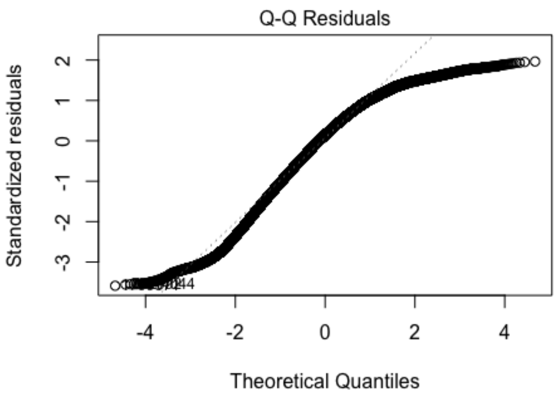


Figure 5.1a: Residuals plot of ride duration (Appendix 5.1)

Figure 5.1b: QQ-plot of residuals (Appendix 5.1)

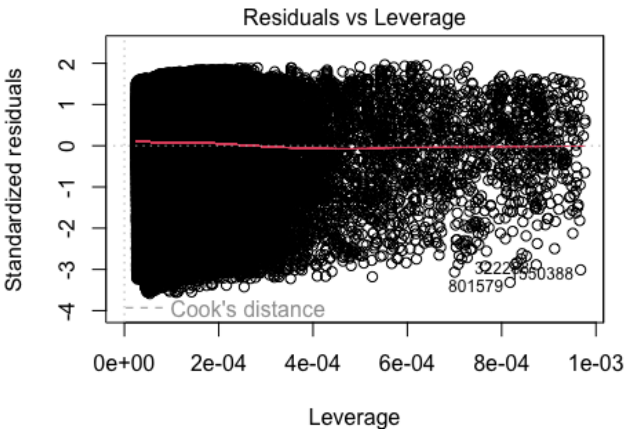


Figure 5.1c: Residual vs Leverage plot of residuals (Appendix 5.2)

For the residual, I used log transformation to reduce the skewness of duration. If we look at the QQ-plot, the upper part turn away from the qq line, which shows that there is heavy skewness in one side of the errors. In contrast, the lower part align very close to the normal line. We cannot use CLT here as it only applies to the sample mean not the individual data points. So, using this result we can say the residuals are not completely normally distributed. Also, this means that the model can be over-performed comparing to using the original duration data. Moreover, the prediction will change to log-scaled duration. If the model perform well, we need to perform e^{x} transformation to get the exact value for duration.

By applying dim on points in the residual plot with low appearances (low alpha on a point, so points with more appearances will be darker), we see that there is no general pattern. The scatter points cluster around -0.5 for major of the fitted values. Thus, we can conclude that the error has constant variance This indicates that there is some correlation between at least one predictor and the dependent variable. However, since the shape of the residual plot is not a horizontal band, the linearity assumption between duration and the predictors is not satisfied.

In the residual vs leverage plot, we don't see any observations nor red dashed lines, hence, there are not any influential points in the dataset. In general, the assumptions for the linear regression are not completely satisfied. Particularly, the linearity and normality of residuals are violated. Thus, the result may not be correct.

2/ Regression Model Result

```
##
## Call:
## lm(formula = log(duration) ~ member_casual + (start_lat:start_lng) +
##     rideable_type * (start_time * weekday), data = tmp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25111 -0.39637  0.09178  0.48996  1.23215
##
## Coefficients:
##                                                           Estimate Std. Err
or
## (Intercept)                                             2.1439766  0.01959
01
## member_casualmember                                    -0.2143362  0.00300
35
## rideable_typeelectric_bike                              0.0388972  0.02258
80
## start_time                                              0.0026678  0.00128
53
## weekdayMonday                                          -0.2107224  0.02590
58
## weekdayTuesday                                         -0.1235877  0.02454
85
## weekdayWednesday                                       -0.1758381  0.02442
49
## weekdayThursday                                        -0.1777110  0.02494
38
## weekdayFriday                                          -0.1379058  0.02543
94
## weekdaySaturday                                        -0.0143317  0.02700
15
## start_lat:start_lng                                    13.6965373  2.27339
35
## start_time:weekdayMonday                                0.0116978  0.00174
21
## start_time:weekdayTuesday                               0.0079928  0.00163
24
## start_time:weekdayWednesday                             0.0117747  0.00162
70
## start_time:weekdayThursday                              0.0097938  0.00166
03
## start_time:weekdayFriday                                0.0075186  0.00168
42
## start_time:weekdaySaturday                             -0.0002123  0.00178
21
## rideable_typeelectric_bike:start_time                   0.0015745  0.00148
75
## rideable_typeelectric_bike:weekdayMonday                0.2069375  0.03031
51
## rideable_typeelectric_bike:weekdayTuesday               0.1180951  0.02879
10
## rideable_typeelectric_bike:weekdayWednesday             0.1736903  0.02867
37
## rideable_typeelectric_bike:weekdayThursday              0.1600782  0.02924
08
## rideable_typeelectric_bike:weekdayFriday                0.0458158  0.02966
```

```
45
## rideable_typeelectric_bike:weekdaySaturday              0.0063548  0.03139
57
## rideable_typeelectric_bike:start_time:weekdayMonday    -0.0125786  0.00203
25
## rideable_typeelectric_bike:start_time:weekdayTuesday   -0.0063845  0.00190
79
## rideable_typeelectric_bike:start_time:weekdayWednesday -0.0107329  0.00190
28
## rideable_typeelectric_bike:start_time:weekdayThursday  -0.0091451  0.00193
75
## rideable_typeelectric_bike:start_time:weekdayFriday    -0.0023480  0.00195
87
## rideable_typeelectric_bike:start_time:weekdaySaturday   0.0010099  0.00206
29
##                                                        t value Pr(>|t|)
## (Intercept)                                            109.442  < 2e-16 **
*
## member_casualmember                                    -71.362  < 2e-16 **
*
## rideable_typeelectric_bike                               1.722 0.085065 .
## start_time                                               2.076 0.037924 *
## weekdayMonday                                           -8.134 4.16e-16 **
*
## weekdayTuesday                                          -5.034 4.80e-07 **
*
## weekdayWednesday                                        -7.199 6.07e-13 **
*
## weekdayThursday                                         -7.124 1.05e-12 **
*
## weekdayFriday                                           -5.421 5.93e-08 **
*
## weekdaySaturday                                         -0.531 0.595577
## start_lat:start_lng                                      6.025 1.70e-09 **
*
## start_time:weekdayMonday                                 6.715 1.89e-11 **
*
## start_time:weekdayTuesday                                4.896 9.76e-07 **
*
## start_time:weekdayWednesday                              7.237 4.59e-13 **
*
## start_time:weekdayThursday                               5.899 3.66e-09 **
*
## start_time:weekdayFriday                                 4.464 8.04e-06 **
*
## start_time:weekdaySaturday                              -0.119 0.905165
## rideable_typeelectric_bike:start_time                    1.058 0.289837
## rideable_typeelectric_bike:weekdayMonday                 6.826 8.73e-12 **
*
## rideable_typeelectric_bike:weekdayTuesday                4.102 4.10e-05 **
```

```
*
## rideable_typeelectric_bike:weekdayWednesday              6.057 1.38e-09 **
*
## rideable_typeelectric_bike:weekdayThursday               5.474 4.39e-08 **
*
## rideable_typeelectric_bike:weekdayFriday                 1.544 0.122476    
## rideable_typeelectric_bike:weekdaySaturday               0.202 0.839596    
## rideable_typeelectric_bike:start_time:weekdayMonday     -6.189 6.07e-10 **
*
## rideable_typeelectric_bike:start_time:weekdayTuesday    -3.346 0.000819 **
*
## rideable_typeelectric_bike:start_time:weekdayWednesday  -5.640 1.70e-08 **
*
## rideable_typeelectric_bike:start_time:weekdayThursday   -4.720 2.36e-06 **
*
## rideable_typeelectric_bike:start_time:weekdayFriday     -1.199 0.230622    
## rideable_typeelectric_bike:start_time:weekdaySaturday    0.490 0.624447    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6274 on 342211 degrees of freedom
## Multiple R-squared:  0.02404,    Adjusted R-squared:  0.02396 
## F-statistic: 290.7 on 29 and 342211 DF,  p-value: < 2.2e-16

## [1] "MSE: 0.394"
```

Output 5.2: Multivariate Linear Regression Result (Appendix 5.3)

a/ Formula:

To create the linear regression model for the duration at stations in Central Park region, I used membership, start location coordinate and start time (in hour), day of the week, and ride type. Since we did not investigate the relationship between membership and other variables, I let it be the standalone predictor (no interaction). The same applied for start station coordinate as we have not investigated its relationship with weekday or start time. Hence I did not include the interaction for this variable. For ride type, we saw that it varied among day of the week. So, we will include that interaction and improve the detail by including ride type interaction with start time in the formula to represent the purposes of bike usage. In general, our formula is:

```
log(duration) ~ member_casual + (start_lat : start_lng) + rideable_type *
start_time * weekday
```

Noted that I have performed grid search by trying different combinations of all the variables and using F-statistics to indicate the good of fitment, Mean adjusted R-squared (R2) and Mean Squared Error (MSE) to indicate the accuracy. The above formula is the best one. Noted that although high F-statistics is good, we must beware of extremely high values as it indicates that the model is over-fit if R^2 is low and MSE is high. Moreover, I also performed mean centering to the start coordinate of start station for better interpretation.

This means that the intercept will account for cases where the station is in the center of the Central Park region cluster. Then the other station will be represented by the coordinate different from the center (not distance).

b/ Model fitment:

The F-statistics of 290.7 with p-value < 0.05 shows that this model is well-fitted. The MSE of 0.391 is relatively small comparing to the range of values (1-3), however, it is not good enough. Those 2 metrics show that the model is balanced in terms of error and prediction accuracy, which mean it may not be over-fit or under-fit. Looking at adjusted-R2, we can see that the value is around 0.024. It means that this multivariate linear regression model can only explain 2.4% of log-transformed duration. This is extremely bad already. If we apply e^x transformation to revert to original scale, the performance will get worse. Hence, this model is not suitable to predict the ride duration of stations in Central Park region. As a result, all the coefficients are meanless.

6/ Conclusion and future questions

By checking all the evaluation metric, we see that the model cannot represent the characteristic of ride duration in Central Park region. Moreover, not all the assumptions are met. Hence, the coefficients and p-value of all the predictors are not correct. Hence, we cannot identify the most influential factor to ride duration using this approach. However, it is possible to model the ride duration as the results from the three sub-questions show that the difference in ride types, membership, weekday, and start station location can result in different values of ride duration. Hence, we can use other approaches such as analyzing and comparing feature importance from Decision Tree based algorithms to answer this question. Moreover, we can also use other Machine Learning algorithms create a predictive model for ride duration. This model will be very helpful for Citi Bike decision makers to assess and decide “which station should be upgraded to meet the demand of customer?”, or “where should the new station locate?”. Also, we can perform analysis on the year-scaled data to analyze the usage of stations and corporate that information to the above analysis to integrate more in-depth information to better answer those questions.

About the dataset, since this is a big data problem, my machine cannot handle it properly so I had to limit the range of data to only half a month. Different results might appear if we analyze those questions again on the data that capture a wider time-span. It will require big data capable tools like Spark and a stronger machine to handle.

In conclusion, we found out that:

- People who subscribes to the Citi Bike membership tends to have make shorter rides compare to people who do not.
- E-bikes is more preferred for longer trips and classic bikes is the most popular choice for short trips.

- The ride duration varies at different regions in New York City.

Using the above findings, decision makers can change the operation to better serve the customers and reduce the chance of damaging the vehicles by having a better distribution plan.

Code Appendix

Libraries in use: `tidyverse`, `ggplot2`, `sf`, `mapview`, `tidygeocoder`

Part 0: Data Importation

```
temp <- list.files(path='/Users/eddie/MATH167R/Data', pattern="\\.csv$", all.
files=TRUE, full.names = TRUE) # Create a list of .csv files in the clarified
path

myfiles <- lapply(temp, read.csv) # Read all the files using csv and make a l
ist to store all data.frames

data <- do.call(rbind, myfiles) # Bind all the data.frames into one

rm(myfiles) # Erase the list from memory to free space

data <- drop_na(data = data) #Drop all rows with N/A values

data <- subset(data, select =  -c(end_station_id, start_station_id, ride_id))
# Drop columns that contains IDs

data <- data.frame(data) # Ensure the data is stored as data.frame

data$started_at <- as.POSIXct(data$started_at, format = "%Y-%m-%d %H:%M:%S")
# Reformat the data as POSIXct date

data$ended_at <- as.POSIXct(data$ended_at, format = "%Y-%m-%d %H:%M:%S")

data <- data[(data$started_at > as.POSIXct("2024-10-01 00:00:00")) & (data$st
arted_at < as.POSIXct("2024-10-14 23:59:59")),] # Ensure the data has rides s
tarted from 0:00:00 Oct 10, 2024 to 23:59:59 Oct 14, 2024
```

Appendix 0.1: Importing and pre-processing dataset

Part 1: Exploratory Data Analysis

```
glimpse(data)
```

Appendix 1.1: Creating summary of dataset information

```
data$datetime_minute <- format(data$started_at, "%Y-%m-%d %H:%M") # Extract d
ate-time up to minutes
group_by(data, datetime_minute) |>
  summarise(count = n()) |>
  ggplot(aes(x=as.POSIXct(datetime_minute, format='%Y-%m-%d %H:%M'), y=count)
) +
  geom_line(color='blue') + labs(title='Count by hour and minute', x='Timespa
n', y='Count') +
  theme_classic()
```

Appendix 1.2: Plotting timeseries of usage count

```
data$duration <- round(as.numeric(difftime(data$ended_at, data$started_at, un
its='secs'))) # use difference between end time and start time in seconds the
n convert to numeric and round up
data$duration <- data$duration / 60 # convert to minutes to make the data bec
ome continuous

weekdays <- factor(c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday",
"Friday", "Saturday"),
                   levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Th
ursday", "Friday", "Saturday")) # Creating day of the week as factor to repre
sent the order of the day


data$weekday <- weekdays[wday(as.POSIXct(data$started_at))] # Map week date d
ata from POSIXct date (1,2,3,etc.) to the above factor
```

Appendix 1.3: Creation of duration and day of the week features

```
summary(data)
```

Appendix 1.4: Creating five numbers of statistics

```
boxplot(data$end_lng)

boxplot(data$end_lat)

count(data |> filter(end_lat == 0))

##    n
## 1 15

count(data |> filter(end_lng == 0))

##    n
## 1 15
```

Appendix 1.5: Creating boxplot of end_lat and end_lng then check number of entries that h as outliers

```
#Calculate the outlier initial value
```

```r
duration_quantiles <- quantile(data$duration, probs=c(.25, .5, .75), type=1)
# Compute quantiles

iqr <- duration_quantiles[3] - duration_quantiles[1] # Calculate IQR

duration_outlier_val <- duration_quantiles[3] + 0.25 * iqr # Get the first va
lue as outlier.

# Filter values

prop_true <- sum(data$duration >= duration_outlier_val) / nrow(data)

cat("Proportion of TRUE:", prop_true, "\n")

## Proportion of TRUE: 0.1694011

data <- data |> filter(data$duration < duration_outlier_val)

hist(data$duration)
```

Appendix 1.6: Identify quantiles and IQR then get the first value that is outlier. Finally, identify the proportion of outliers then filter the data and plot histogram of duration

```r
convert_hour <- function(time) {
  hour <- as.numeric(format(time, "%H"))
  minute <- as.numeric(format(time, "%M"))

  return(hour + minute/60)
}

data$start_time <- sapply(data$started_at, convert_hour)
```

Appendix 1.7: Convert start time to numerical using hour + minute/60

Part 2: Sub-question 1

```r
member_duration <- data[data$member_casual == 'member',]$duration
casual_duration <- data[data$member_casual == 'casual',]$duration

ggplot(data=data[data$duration < 60,], aes(x=duration, color=member_casual, g
roup=member_casual)) + geom_histogram(bins = 60) + theme_classic() + labs(tit
le='Histogram of ride duration under 60 minutes', x='Duration (mins)', y='Cou
nt', color='Membership')
```

Appendix 2.1: Creating histogram of ride duration by member group and casual group

```
for (membership in c('member', 'casual')){

  qqnorm(data[data$member_casual == membership,]$duration)
  qqline(data[data$member_casual == membership,]$duration, col='blue')
}
```

Appendix 2.2: Creating qq-plot for ride duration of member group and casual group

```
mean_and_variance <- function(data) {
  nu <- mean(data, na.rm = TRUE) # Calculate mean
  variance <- var(data, na.rm = TRUE) # Calculate variance
  cat('Mean: ', nu, ' Variance: ', variance)
}
```

Appendix 2.3: Creating function for calculating mean and variance of a list/vector/series

```
t.test(duration ~ member_casual, data=data)
```

Appendix 2.4: Performing t-test for ride duration of member vs casual

```
ggplot(data=data[data$duration < 60,], aes(x=duration, fill=member_casual, gr
oup=member_casual)) +
  geom_boxplot(outlier.shape = NA) + labs(title='Box plot of ride duration be
tween casual and member users', x='Duration (in minutes)', fill='Membership')
+ theme_classic()
```

Appendix 2.5: Creating boxplot of ride duration between members vs casuals

Part 3: Sub-question 2

```
data |> group_by(rideable_type, weekday) |> summarise(count = n())
```

Appendix 3.1: Counting entries after grouping data by ride types and day of the week

```
results <- list()

for (day in weekdays) { # Loop through each weekday

  tmp <- data[data$weekday == day, ] # Filter the data for the current day

  test <- t.test(duration ~ rideable_type, data = tmp) # Run t-test
```

```r
  results[[day]] <- test # Save the summary of the t-test in the results list
}

print(results)
```

Appendix 3.2: Performing t-test of e-bikes vs classic bikes for seven days of the week

```r
ggplot(data = data, aes(x=weekday, y=duration, colour = rideable_type)) + geom_boxplot(outlier.shape = NA) + labs(title='Box plot of ride duration between casual and member users', x='Duration (in minutes)', fill='Ride type') + theme_classic()
```

Appendix 3.3: Creating box plot of ride duration of each day of the week by ride types

Part 4: Sub-question 3

```r
locations <- data[, c('start_lat', 'start_lng')]

find_nearest_region <- function(lat, lon, regions) {

  distances <- sqrt((regions$lat - lat)^2 + (regions$lng - lon)^2) # Calculate great-circle distances using the Euclidean distance formula

  nearest_region <- regions[which.min(distances), ] # Find the region with the minimum distance

  return(nearest_region$region) # Return the nearest region's name
}

regions <- data.frame(
  region = c(
    "Downtown Manhattan",
    "Brooklyn Downtown",
    "NY Uptown + The Bronx",
    "Central Park",
    "Queens (Long Island City)"
  ),
  lat = c(
    40.732724,  # Downtown Manhattan
    40.689089,  # Brooklyn Downtown
    40.820514,  # The Bronx
    40.778834, # Central Park
    40.762762  # Queens (Long Island City)
  ),
  lng = c(
    -73.991779, # Downtown Manhattan
    -73.957245, # Brooklyn Downtown
    -73.935164, # The Bronx
    -73.973304, # Central Park
    -73.920252 # Queens (Long Island City)
  )
)

locations$regions <- mapply(find_nearest_region, locations$start_lat, locatio
ns$start_lng, MoreArgs = list(regions)) # Creating region data using by apply
ing the above function

data$start_region <- locations$regions

plot_dat <- locations |>
  distinct(start_lat, start_lng, regions) |>
  st_as_sf(coords = c("start_lng", "start_lat"), crs = 4326) # Convert coordi
nate data to sf coordinate for plotting

mapview(plot_dat, cex = 3, label = "regions", zcol = "regions", col.regions =
c("blue", "green", "yellow", "red", "orange")) # Plotting points using region
name as hover label, map color to separate regions
```

Appendix 4.1: Mapping NYC regions and plot using mapview

```r
for (region in regions$region) {
  cal_data <-  data %>% filter(start_region == region) %>% drop_na() # Filter
region
  cal_data <- cal_data$duration
  cat(region, ': ')
  mean_and_variance(cal_data) # Reuse the function from Appendix 2.3
  cat('\n')
}
```

Appendix 4.2: Calculating mean and variance of each region

```r
model <- aov(duration ~ start_region, data=data)
summary(model)
```

Appendix 4.3: Performing ANOVA between the five NYC regions

```r
TukeyHSD(model, conf.level = 0.95)
```

Appendix 4.4: Performing Tukey HSD between the five NYC regions

```r
ggplot(data, aes(x = start_region, y = duration)) +
  geom_boxplot() +
  labs(title = "Boxplot of Duration by Start Region",
       x = "Start Region", y = "Duration") +
  theme_classic()
```

Appendix 4.5: Creating box plot of ride duration between the five NYC regions

Part 5: Main question

```r
data$start_lat <- scale(data$start_lat, scale = FALSE)
data$start_lng <- scale(data$start_lng, scale = FALSE)
```

Appendix 5.0: Centering coordinate data

```r
model <- lm(log(duration) ~ member_casual + (start_lat : start_lng) + rideabl
e_type * (start_time * weekday) , data=tmp)
plot(model)
```

Appendix 5.1: Fitting linear regression model and plot diagnostics plots

```r
ggplot(data = data.frame(fitted = fitted(model), residuals = residuals(model)
),
       aes(x = fitted, y = residuals)) +
  geom_point(alpha = 0.3) + # Dim the points (adjust opacity)
  geom_hline(yintercept = 0, color = "red") + # Highlight y=0
  labs(x = "Fitted Values", y = "Residuals", title="Residual Plot")
```

Appendix 5.2: Creating residuals plot with adjusted opacity

```r
summary(model)
```

Appendix 5.3: Show model summary