New York City Taxi Trip Duration

Submitted to:
Prof. Douglas Jones

Group Members:

Divya Naidu
Hemakshi Shardha
Shwetha Krishna
Tarun Singh

December 03, 2018

Business Problem:

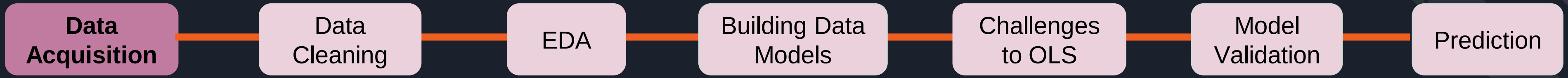
Build a model that predicts the total ride duration of taxi trips in New York City

Overview

Source: https://www.kaggle.com

Data Fields:

- Id
- Vendor_id
- Pickup_datetime
- Dropoff_datetime
- Passenger_count
- Pickup_longitude
- Pickup_latitude
- Dropoff_longitude
- Dropoff_latitude
- Store_and_fwd_flag
- Trip_duration



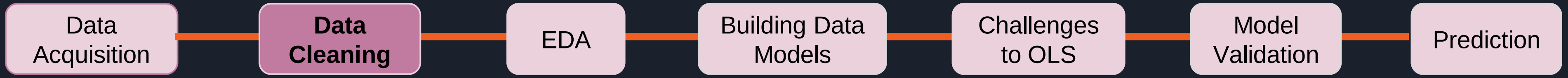
Data Acquisition → Data Cleaning → EDA → Building Data Models → Challenges to OLS → Model Validation → Prediction

Libraries Used

- library(dply)
- library(tibble)
- library(tidyr)
- library(stringr)
- library(forcast)
- library(lubridate)
- library(Amelia)
- library(mice)
- library(moments)
- library(ggplot2)
- library(rgdal)
- library(data.table)
- library(dplyr)
- library(geosphere)
- library(car)
- library(corrplot)
- library(DAAG)
- library(faraway)
- library(GGally)
- library(corrplot)
- library(gridExtra)

Data Manipulation

- Converted 0’s to NAs
- Converted categorical variable into a factor variable
- Converted the format of Data and time variables
- Used missmap to identify missing data.
- Used md.pattern to check the pattern of missing data
- Used MICE package for multiple imputation



Data Acquisition

Data Cleaning

EDA

Building Data Models

Challenges to OLS

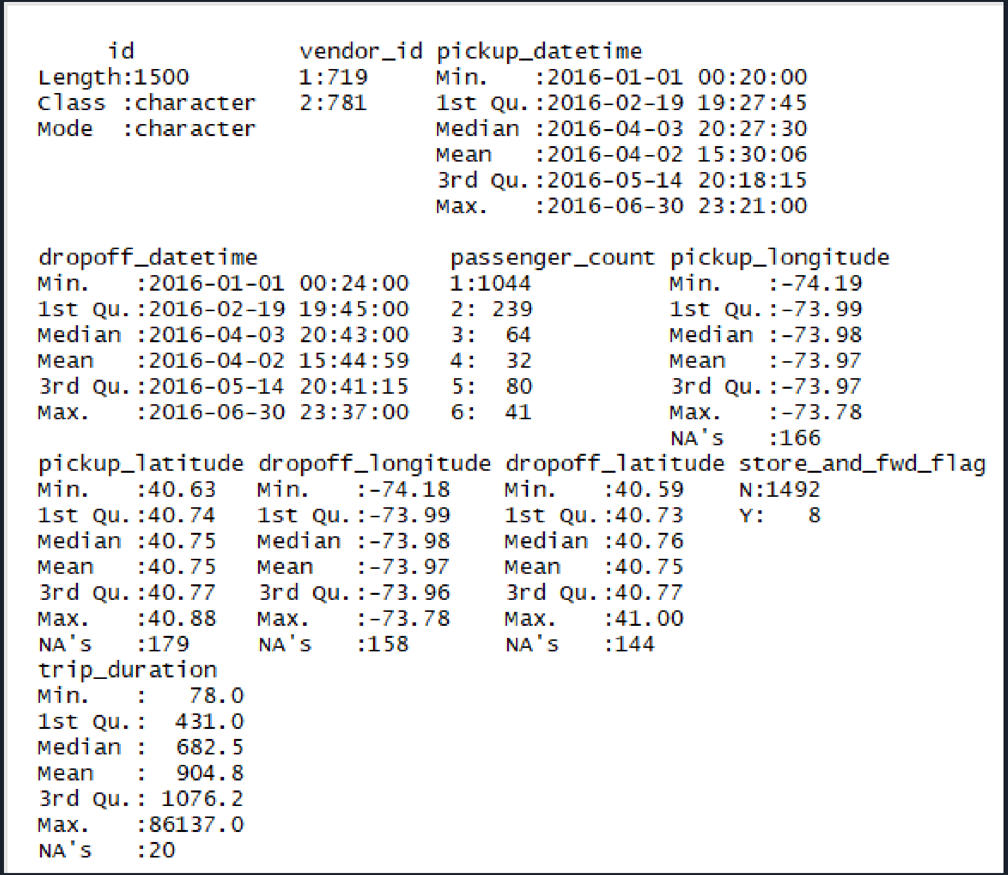
Model Validation

Prediction

Summary-Before MICE

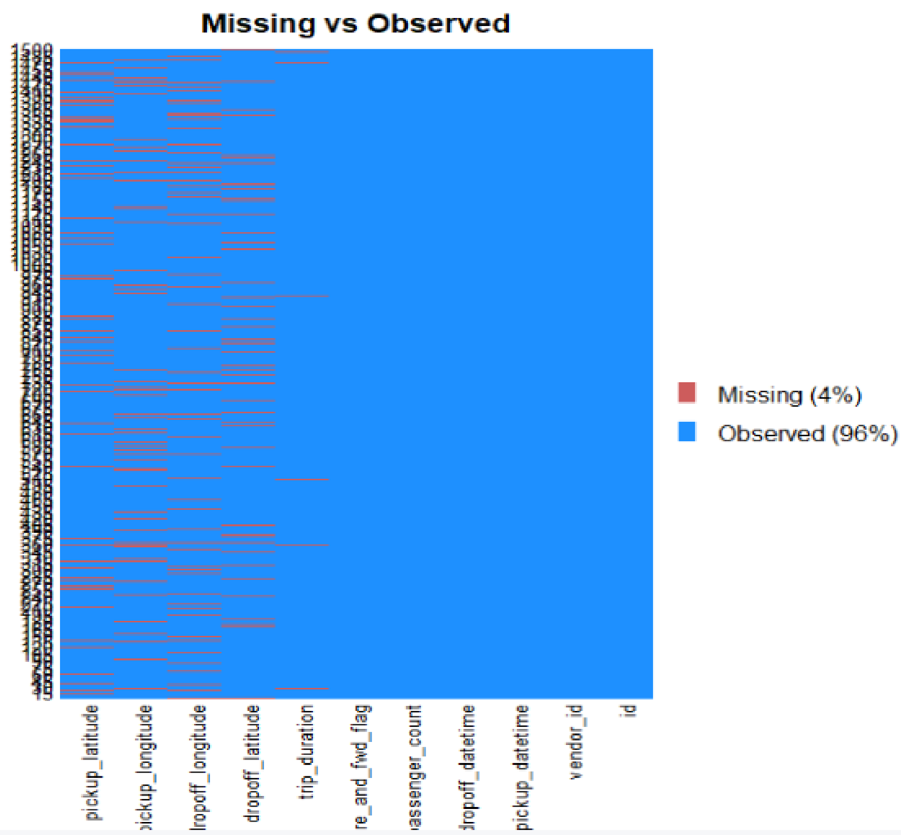
From summary, we can see that there are NAs in following columns:

- Pickup_Longitude
- Pickup_Latitude
- Dropoff_Longitude
- Dropoff_Latitude
- Trip_Duration

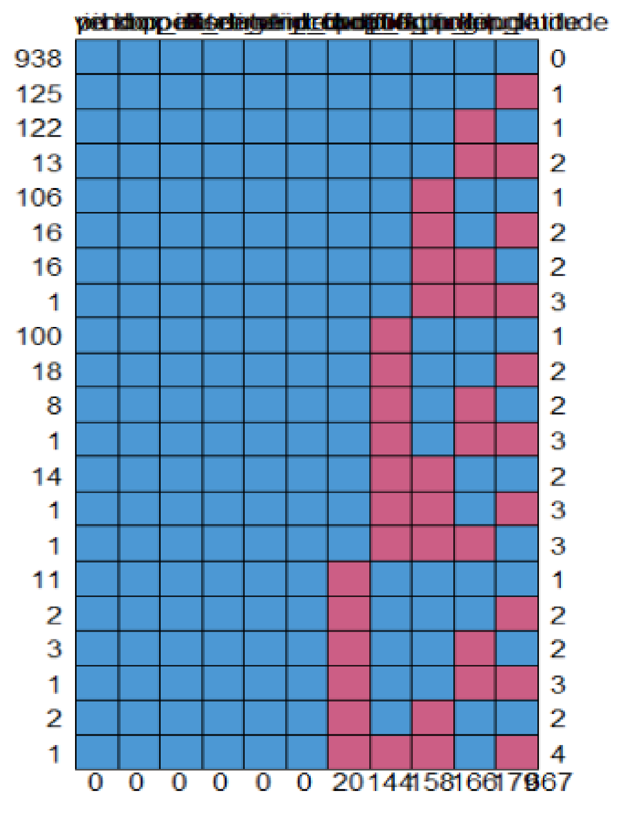


```
     id            vendor_id pickup_datetime
Length:1500        1:719     Min.   :2016-01-01 00:20:00
Class :character   2:781     1st Qu.:2016-02-19 19:27:45
Mode  :character             Median :2016-04-03 20:27:30
                             Mean   :2016-04-02 15:30:06
                             3rd Qu.:2016-05-14 20:18:15
                             Max.   :2016-06-30 23:21:00

dropoff_datetime              passenger_count pickup_longitude
Min.   :2016-01-01 00:24:00   1:1044          Min.   :-74.19
1st Qu.:2016-02-19 19:45:00   2: 239          1st Qu.:-73.99
Median :2016-04-03 20:43:00   3:  64          Median :-73.98
Mean   :2016-04-02 15:44:59   4:  32          Mean   :-73.97
3rd Qu.:2016-05-14 20:41:15   5:  80          3rd Qu.:-73.97
Max.   :2016-06-30 23:37:00   6:  41          Max.   :-73.78
                                              NA's   :166
pickup_latitude dropoff_longitude dropoff_latitude store_and_fwd_flag
Min.   :40.63   Min.   :-74.18    Min.   :40.59    N:1492
1st Qu.:40.74   1st Qu.:-73.99    1st Qu.:40.73    Y:   8
Median :40.75   Median :-73.98    Median :40.76
Mean   :40.75   Mean   :-73.97    Mean   :40.75
3rd Qu.:40.77   3rd Qu.:-73.96    3rd Qu.:40.77
Max.   :40.88   Max.   :-73.78    Max.   :41.00
NA's   :179     NA's   :158       NA's   :144
trip_duration
Min.   :   78.0
1st Qu.:  431.0
Median :  682.5
Mean   :  904.8
3rd Qu.: 1076.2
Max.   :86137.0
NA's   :20
```

Missmap from Amelia Library



Missing Data Pattern



Data Acquisition

Data Cleaning

EDA

Building Data Models

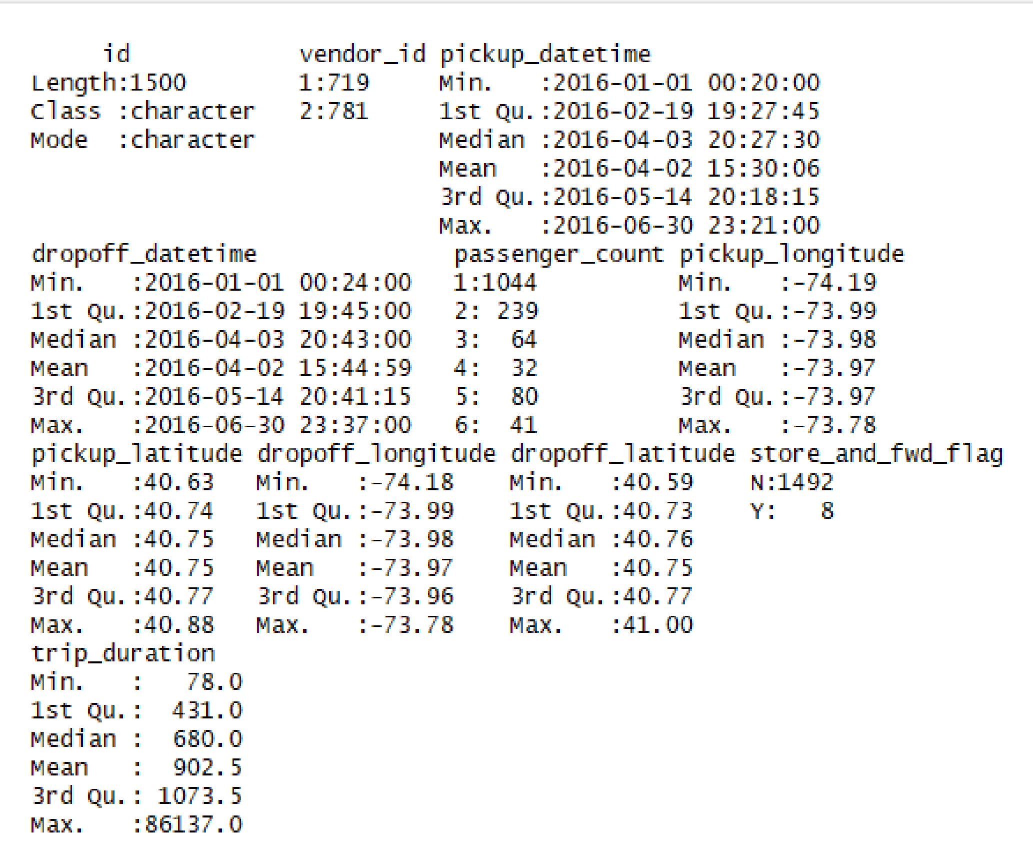
Challenges to OLS

Model Validation

Prediction

Summary-
After MICE

From summary, we can see that NAs are removed after multiple imputation.



```
      id              vendor_id pickup_datetime
 Length:1500        1:719     Min.   :2016-01-01 00:20:00
 Class :character   2:781     1st Qu.:2016-02-19 19:27:45
 Mode  :character             Median :2016-04-03 20:27:30
                              Mean   :2016-04-02 15:30:06
                              3rd Qu.:2016-05-14 20:18:15
                              Max.   :2016-06-30 23:21:00
 dropoff_datetime              passenger_count pickup_longitude
 Min.   :2016-01-01 00:24:00   1:1044          Min.   :-74.19
 1st Qu.:2016-02-19 19:45:00   2: 239          1st Qu.:-73.99
 Median :2016-04-03 20:43:00   3:  64          Median :-73.98
 Mean   :2016-04-02 15:44:59   4:  32          Mean   :-73.97
 3rd Qu.:2016-05-14 20:41:15   5:  80          3rd Qu.:-73.97
 Max.   :2016-06-30 23:37:00   6:  41          Max.   :-73.78
 pickup_latitude dropoff_longitude dropoff_latitude store_and_fwd_flag
 Min.   :40.63   Min.   :-74.18    Min.   :40.59    N:1492
 1st Qu.:40.74   1st Qu.:-73.99    1st Qu.:40.73    Y:   8
 Median :40.75   Median :-73.98    Median :40.76
 Mean   :40.75   Mean   :-73.97    Mean   :40.75
 3rd Qu.:40.77   3rd Qu.:-73.96    3rd Qu.:40.77
 Max.   :40.88   Max.   :-73.78    Max.   :41.00
 trip_duration
 Min.   :   78.0
 1st Qu.:  431.0
 Median :  680.0
 Mean   :  902.5
 3rd Qu.: 1073.5
 Max.   :86137.0
```

Feature Engineering

- **Neighborhoods Incorporation**:

  We have assigned a neighborhood to every pickup and dropoff location .

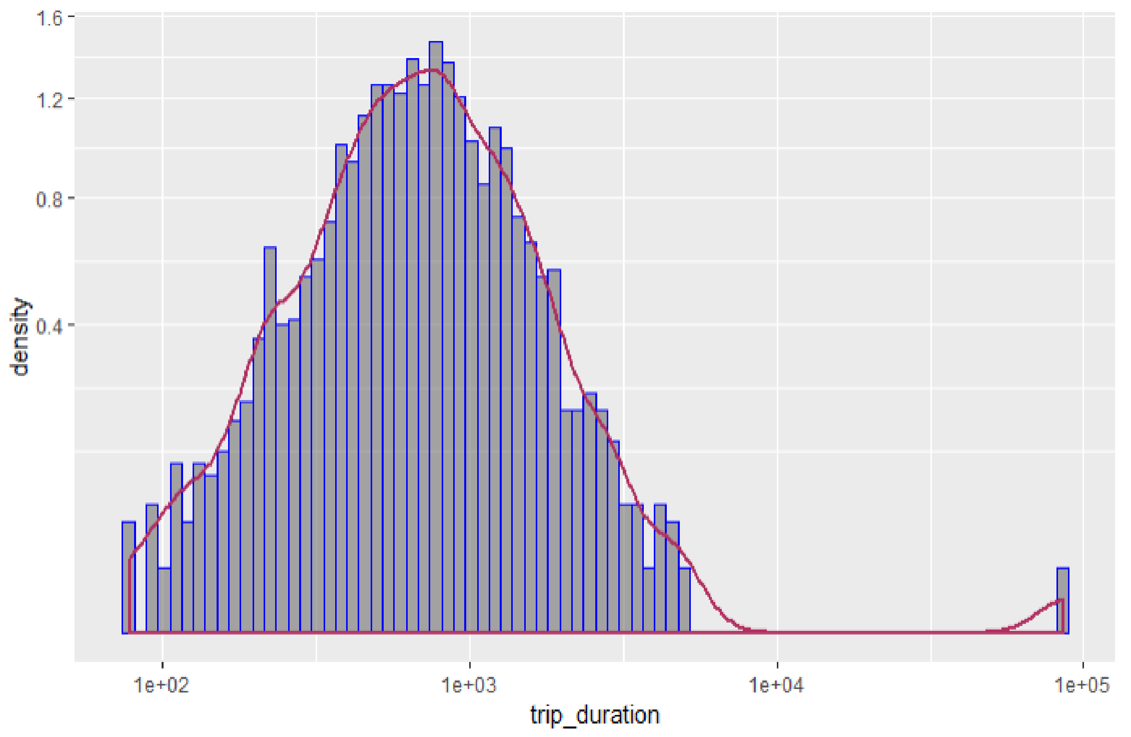
  In order to determine the neighborhoods of given locations, we have used the publicly available Zillow Neighborhood Boundary Shapefiles.

- **Distance Incorporation**:

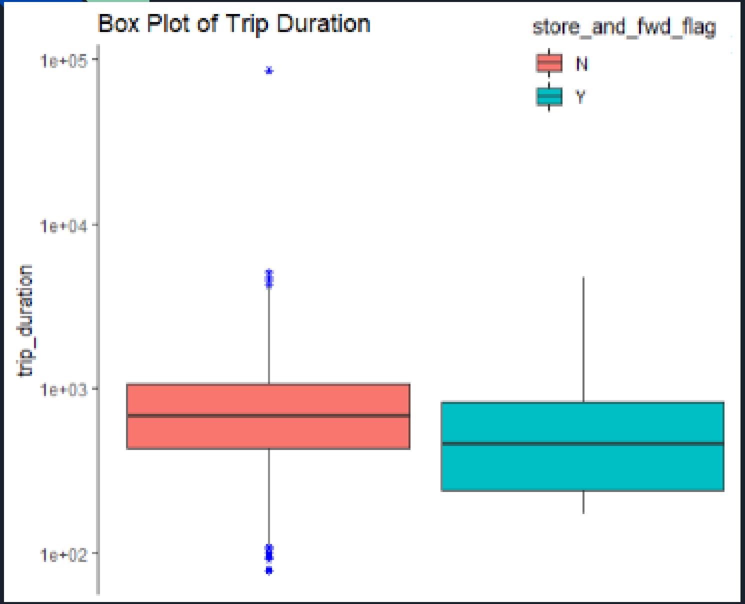
  We have calculated distance_miles using the Longitude and Latitude.

Skewness of Target Variable- Trip Duration

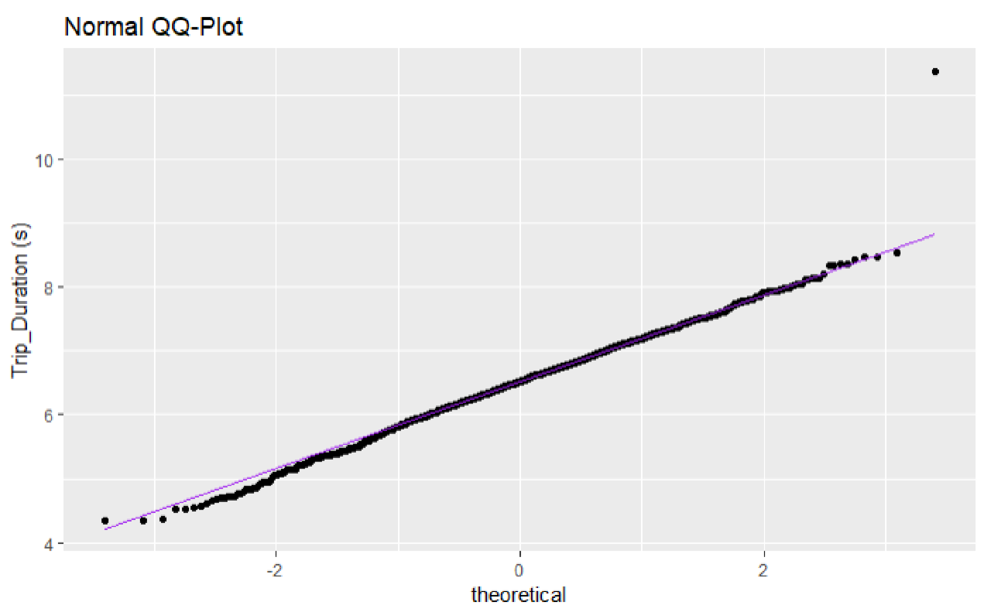
- Trip duration has almost log normal distribution.
- Most of the rides were less than 17 minutes.
- One potential outlier which is making our distribution bimodal i.e distribution with two peaks.
- It has edge-peak distribution i.e. a peak towards the edge of the distribution.



Box Plot of Trip Duration

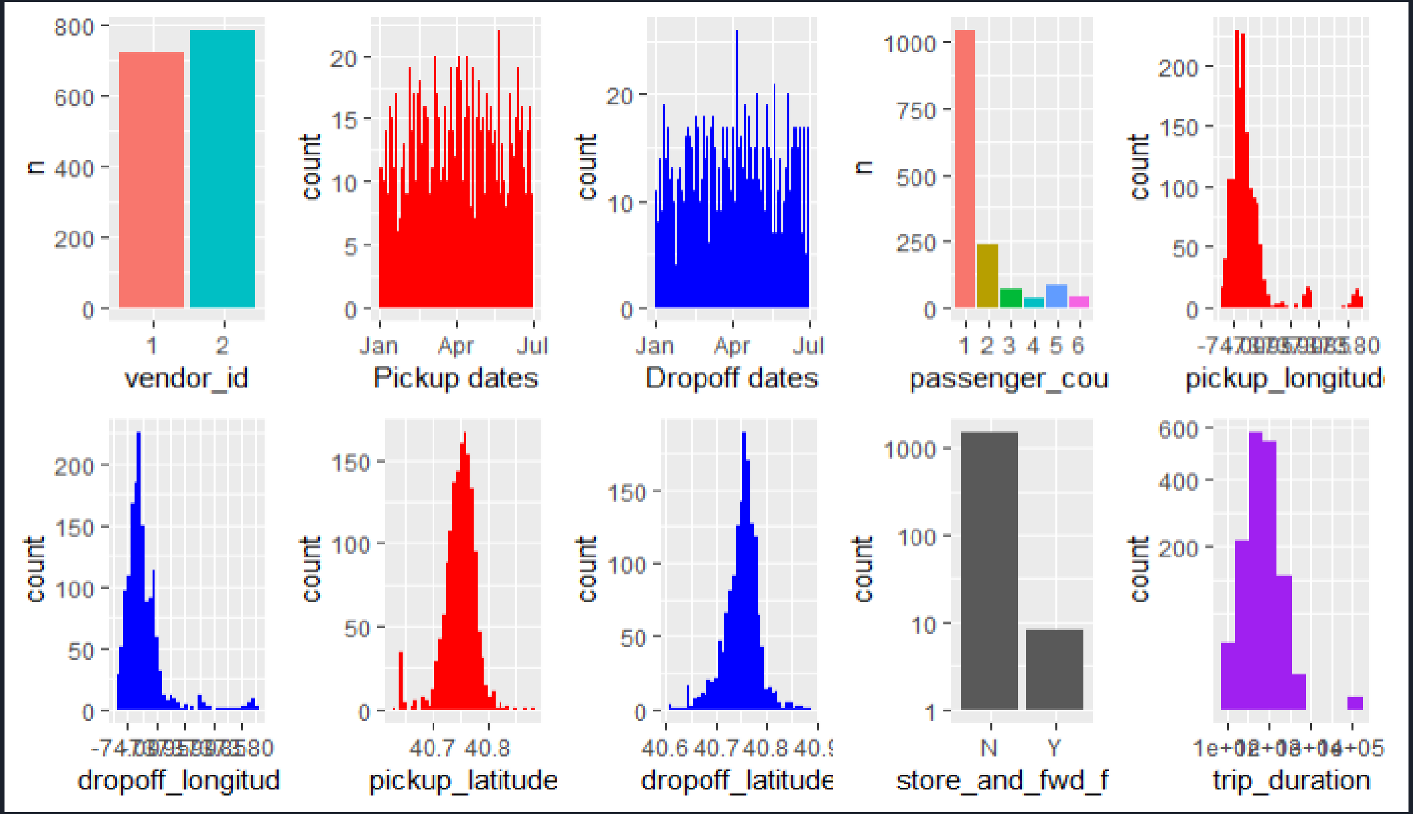


Normal QQ Plot of Trip Duration



Data Acquisition — Data Cleaning — **EDA** — Building Data Models — Challenges to OLS — Model Validation — Prediction

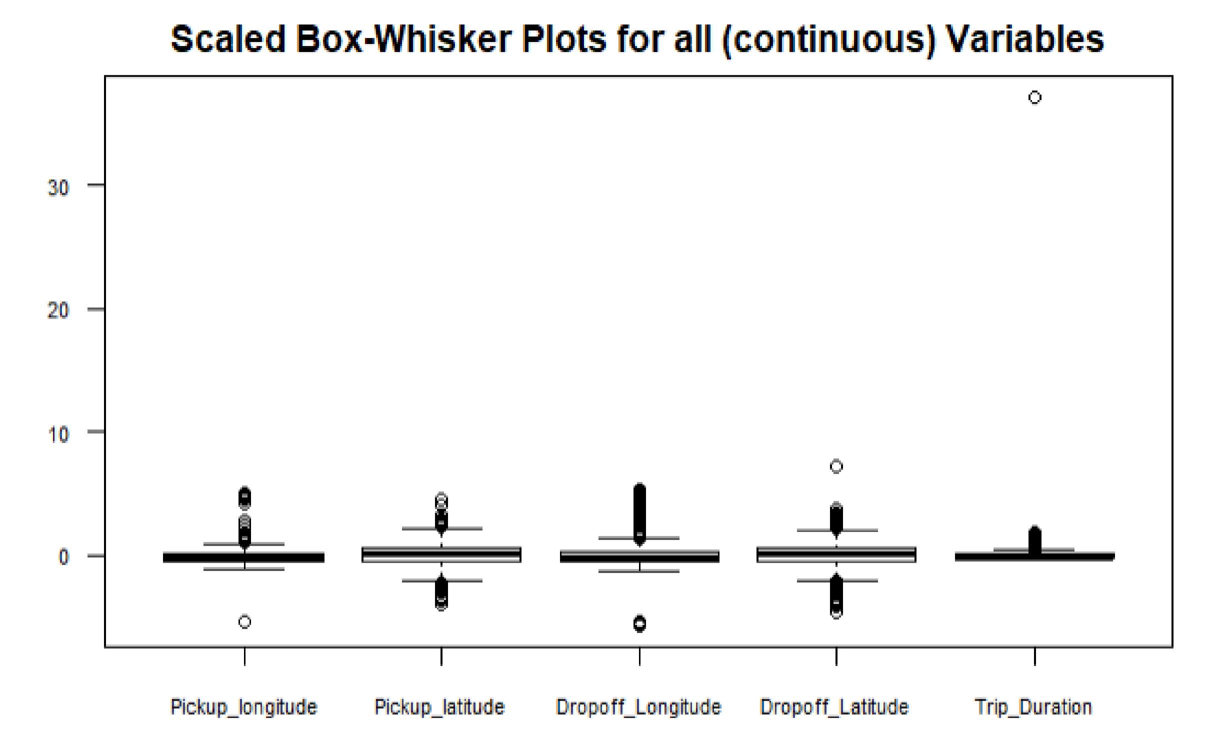
Skewness of Predictor Variables



Data Acquisition — Data Cleaning — **EDA** — Building Data Models — Challenges to OLS — Model Validation — Prediction

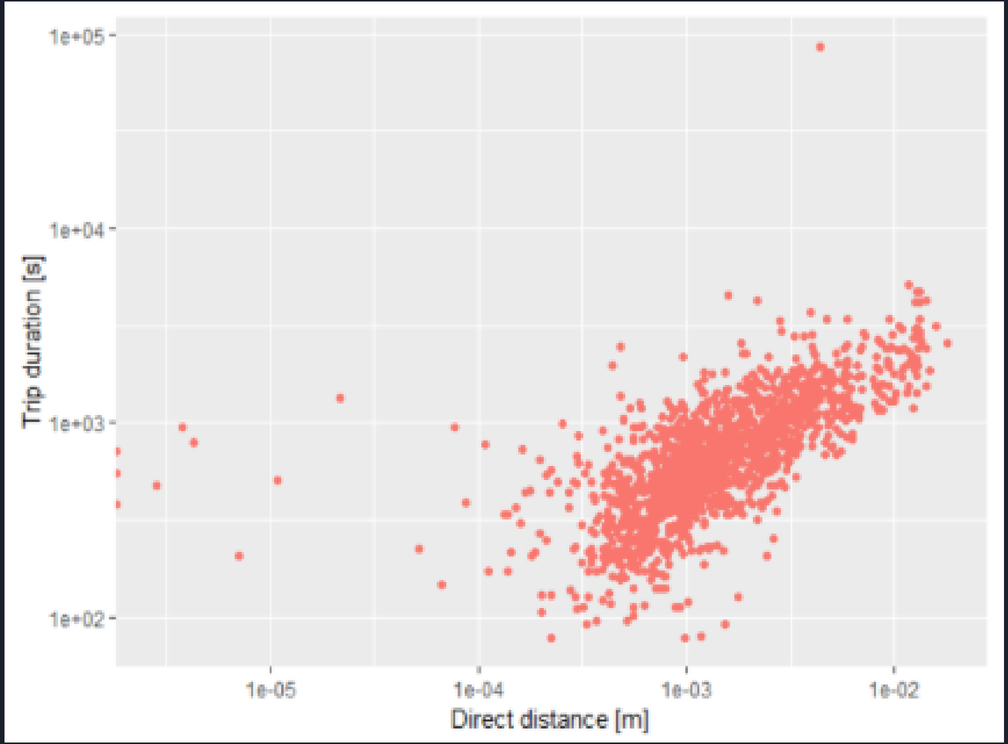
Skewness of Continuous Variables

- Except Trip duration, all the other variables are approximately normally distributed with few outliers.



Distance (in Miles) Predictor

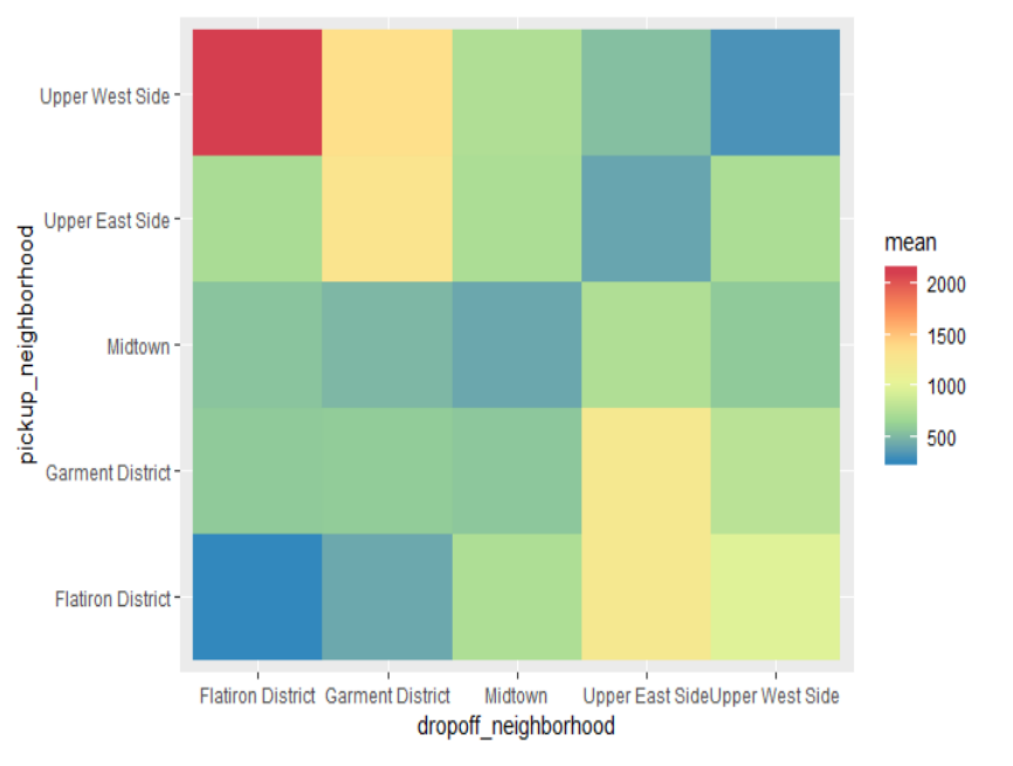
- Graph shows that Trip Duration has high correlation with Distance_miles.
- Due to the presence of an influential outlier, the graph is right skewed.



Heatmap based on Neighborhood

- The average trip duration is really low for rides within the neighborhood, it is logical.

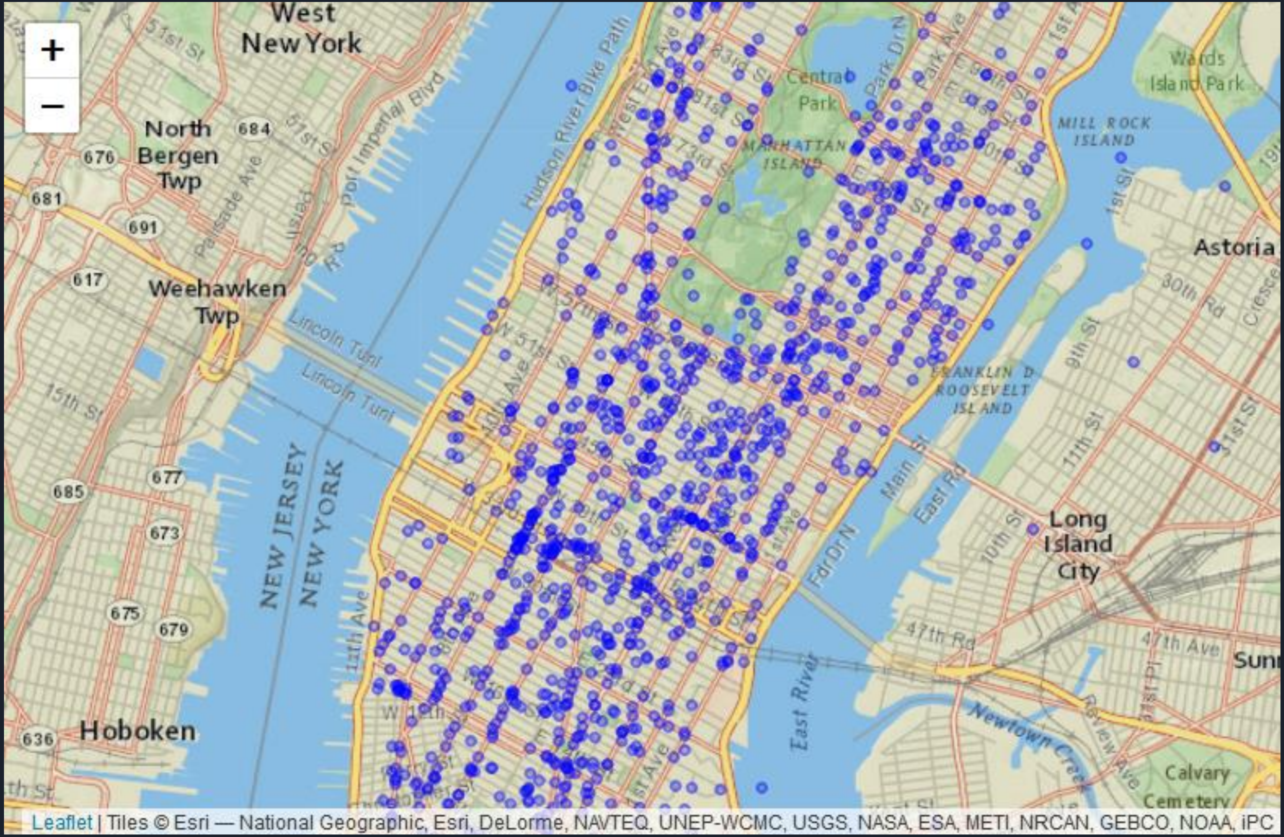
- The historical average trip duration between two neighborhood certainly informs about the trip duration between those two neighborhoods in the future - Strong Predictor!



Data Acquisition — Data Cleaning — **EDA** — Building Data Models — Challenges to OLS — Model Validation — Prediction

Leaflet

- A map of NYC and overlay a manageable number of pickup coordinates to get a general overview of the location and distances.



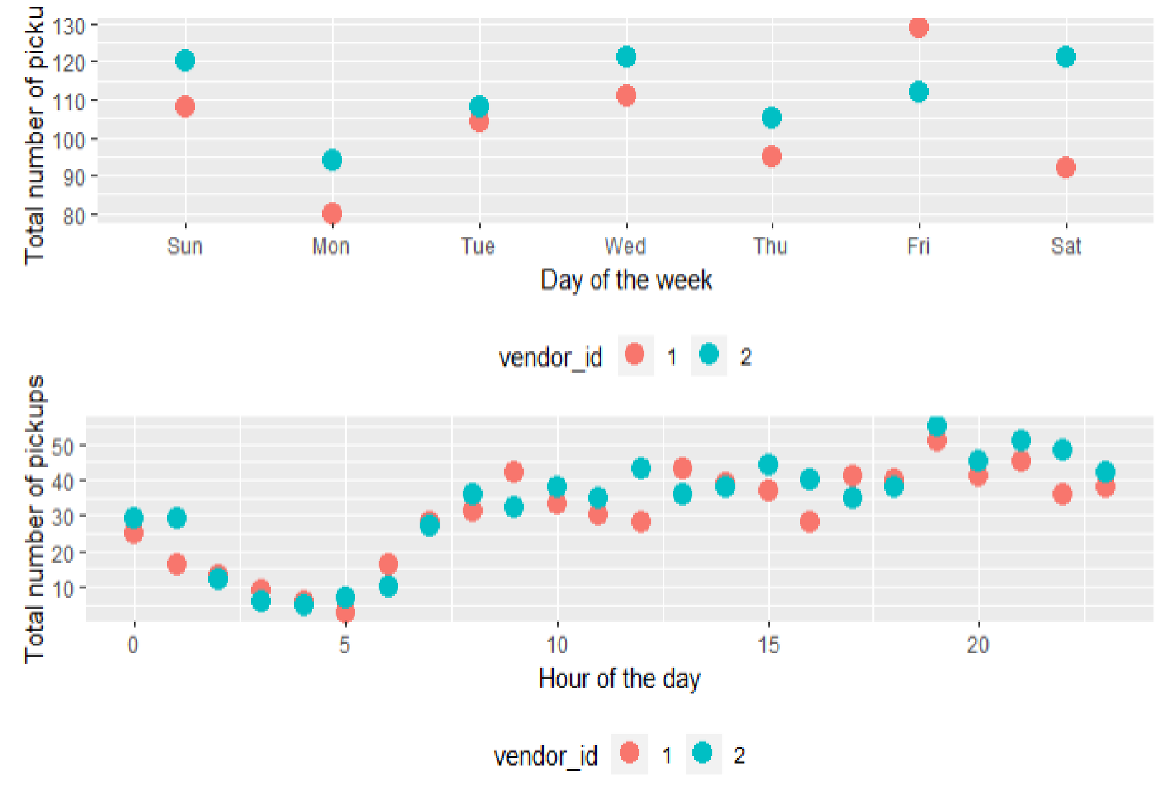
Data Acquisition — Data Cleaning — **EDA** — Building Data Models — Challenges to OLS — Model Validation — Prediction

Analysis- Number of Rides

- Almost all the days of the weeks and hours of the day, vendor 2 has more pickup as compared to vendor 1.

- Friday is the busiest day.

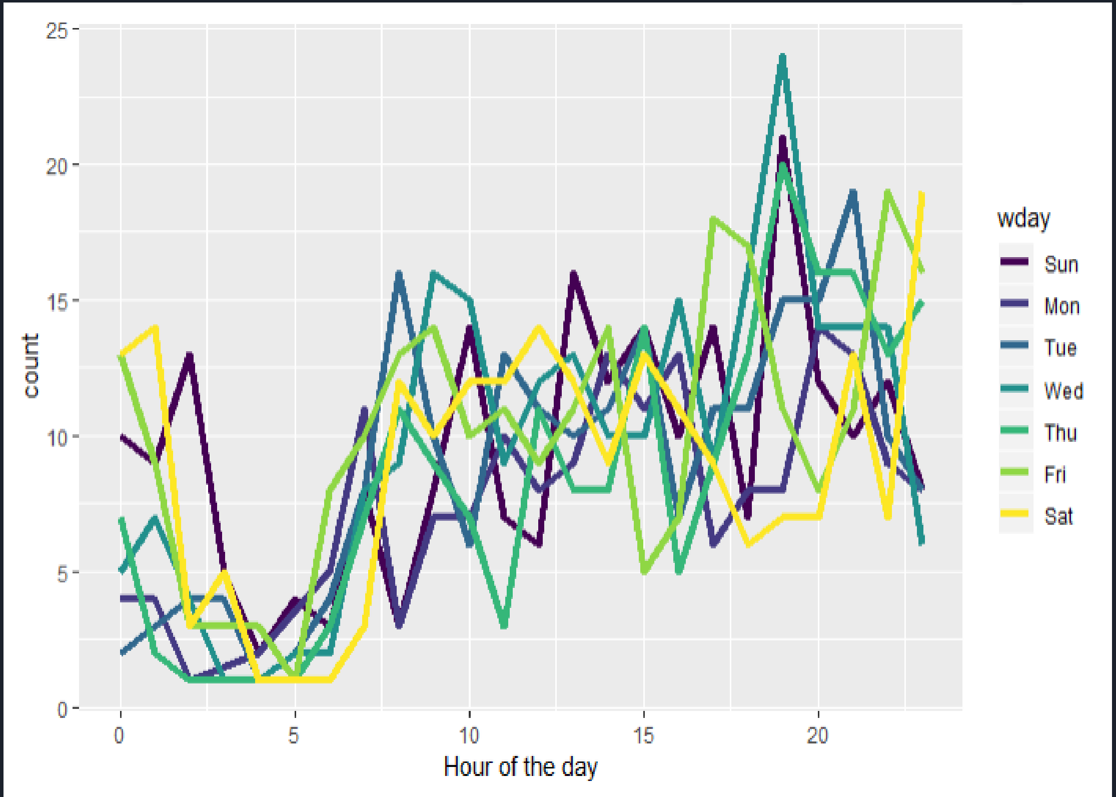
- Monday has the lowest number of rides.

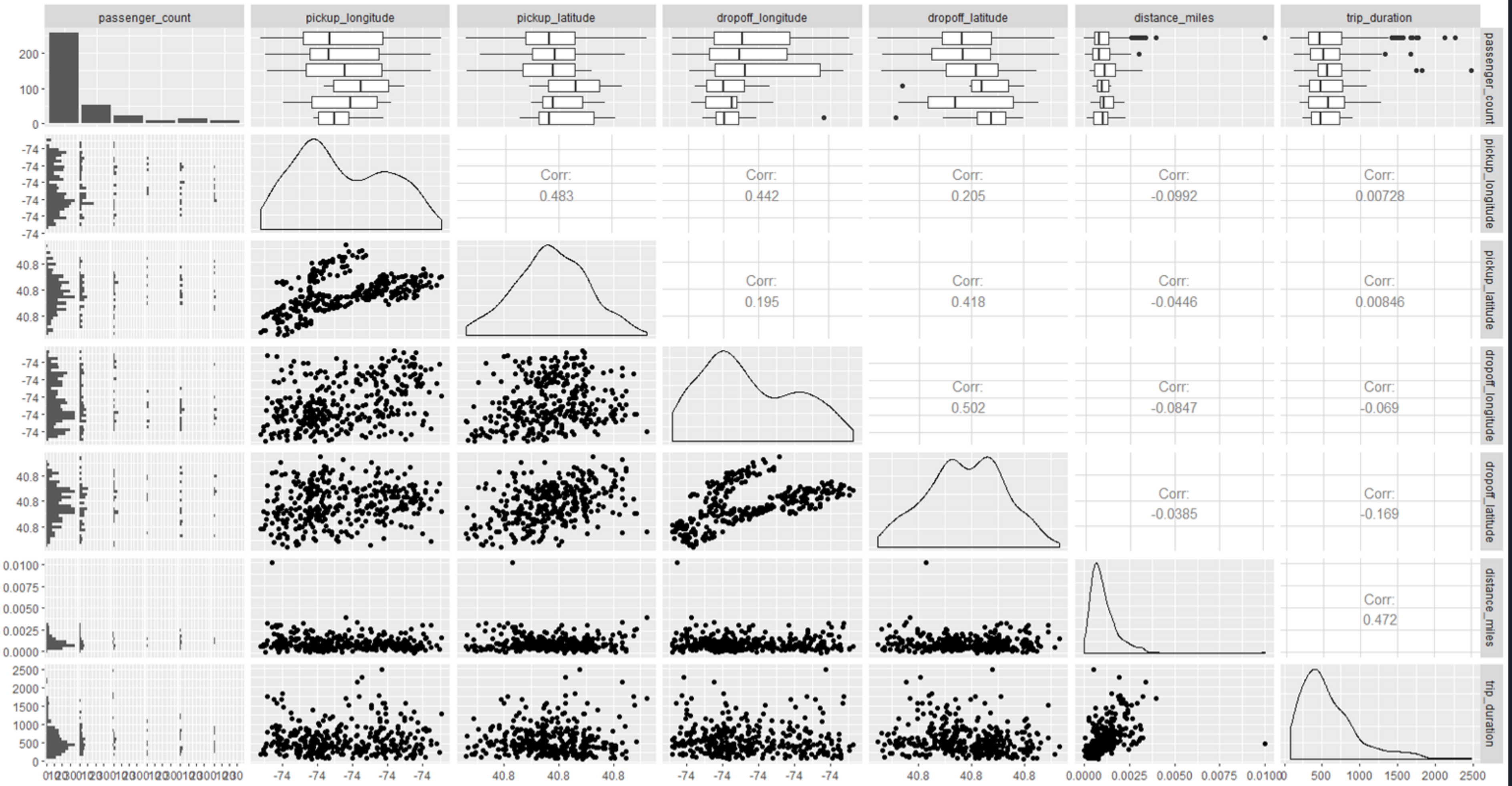


Analysis- Number of Rides based on Hour of the Day

- On friday, Saturday and Sunday, we have more trips during early morning hours on the contrary, trips are low in between 5 to 10.

- One possible reason for this distinction could be the contrast between the lifestyle of people on business days and night life.




passenger_count
pickup_longitude
pickup_latitude
dropoff_longitude
dropoff_latitude
distance_miles
trip_duration
Corr:
0.483
Corr:
0.442
Corr:
0.205
Corr:
-0.0992
Corr:
0.00728
Corr:
0.195
Corr:
0.418
Corr:
-0.0446
Corr:
0.00846
Corr:
0.502
Corr:
-0.0847
Corr:
-0.069
Corr:
-0.0385
Corr:
-0.169
Corr:
0.472
0
100
200
-74
40.8
0.0000
0.0025
0.0050
0.0075
0.0100
0
500
1000
1500
2000
2500
0123
0.0000
0.0025
0.0050
0.0075
0.0100
0
500
1000
1500
2000
2500

GGpairs- Two-at-time Redundancy and Outliers

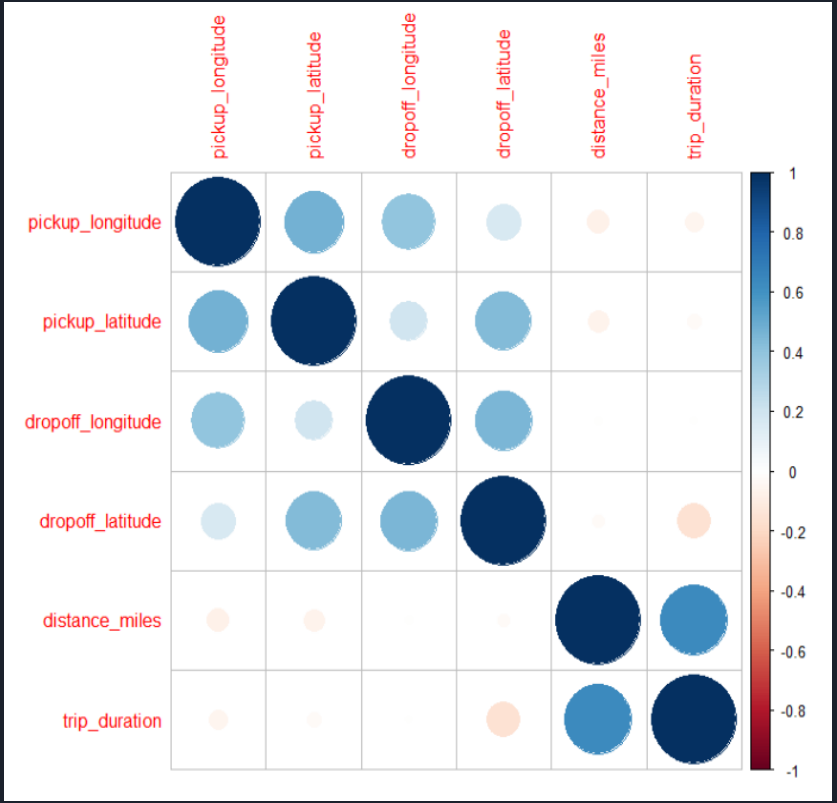
- No two predictors are redundant.
- Distance_miles is right skewed due to presence of an influential outlier.
- Pickup_Latitude and Dropoff_Latitude has approximately normal distribution as from the leaflet, we can see that most of the rides has their pickup location of Manhattan.
- Pickup_longitutde and Dropoff_longitutde has a right tailed distribution. This is consistent with the fact that most of the rides were for Manhattan.

Corrplot- Strong Predictor Checked

- Trip Duration has highest correlation with Distance_miles

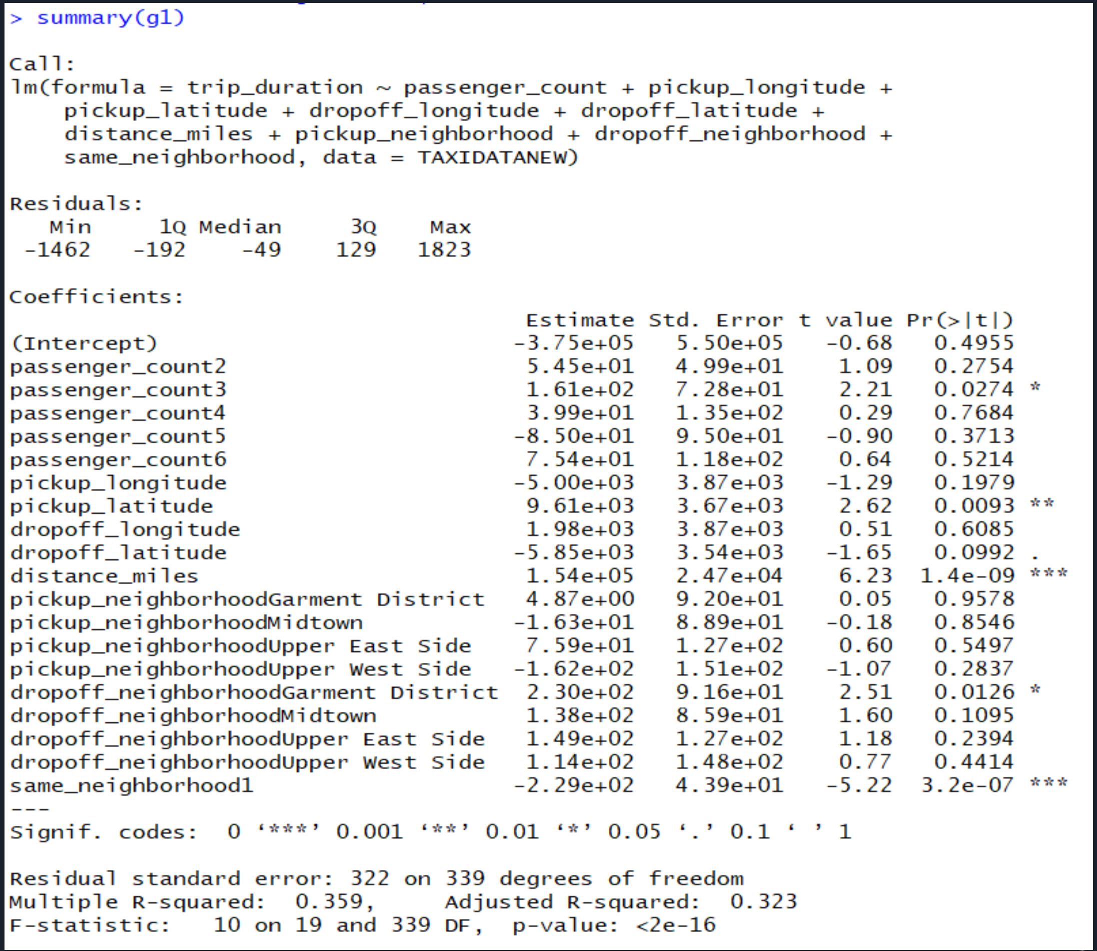
- Although correlation of neighborhood (categorical variable) and Trip duration is not possible, the data reveals that there is correlation between the two.

- Strong Predictors - Neighborhood and [illegible]iles



Model Fitting

- g1 : Fitted biggest model using Linear Regression



```
> summary(g1)

Call:
lm(formula = trip_duration ~ passenger_count + pickup_longitude +
    pickup_latitude + dropoff_longitude + dropoff_latitude +
    distance_miles + pickup_neighborhood + dropoff_neighborhood +
    same_neighborhood, data = TAXIDATANEW)

Residuals:
   Min     1Q Median     3Q    Max
 -1462   -192    -49    129   1823

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                         -3.75e+05   5.50e+05   -0.68   0.4955
passenger_count2                     5.45e+01   4.99e+01    1.09   0.2754
passenger_count3                     1.61e+02   7.28e+01    2.21   0.0274 *
passenger_count4                     3.99e+01   1.35e+02    0.29   0.7684
passenger_count5                    -8.50e+01   9.50e+01   -0.90   0.3713
passenger_count6                     7.54e+01   1.18e+02    0.64   0.5214
pickup_longitude                    -5.00e+03   3.87e+03   -1.29   0.1979
pickup_latitude                      9.61e+03   3.67e+03    2.62   0.0093 **
dropoff_longitude                    1.98e+03   3.87e+03    0.51   0.6085
dropoff_latitude                    -5.85e+03   3.54e+03   -1.65   0.0992 .
distance_miles                       1.54e+05   2.47e+04    6.23  1.4e-09 ***
pickup_neighborhoodGarment District  4.87e+00   9.20e+01    0.05   0.9578
pickup_neighborhoodMidtown          -1.63e+01   8.89e+01   -0.18   0.8546
pickup_neighborhoodUpper East Side   7.59e+01   1.27e+02    0.60   0.5497
pickup_neighborhoodUpper West Side  -1.62e+02   1.51e+02   -1.07   0.2837
dropoff_neighborhoodGarment District 2.30e+02   9.16e+01    2.51   0.0126 *
dropoff_neighborhoodMidtown          1.38e+02   8.59e+01    1.60   0.1095
dropoff_neighborhoodUpper East Side  1.49e+02   1.27e+02    1.18   0.2394
dropoff_neighborhoodUpper West Side  1.14e+02   1.48e+02    0.77   0.4414
same_neighborhood1                  -2.29e+02   4.39e+01   -5.22  3.2e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 322 on 339 degrees of freedom
Multiple R-squared:  0.359,	Adjusted R-squared:  0.323
F-statistic:   10 on 19 and 339 DF,  p-value: <2e-16
```

Data Acquisition

Data Cleaning

EDA

Building Data Models

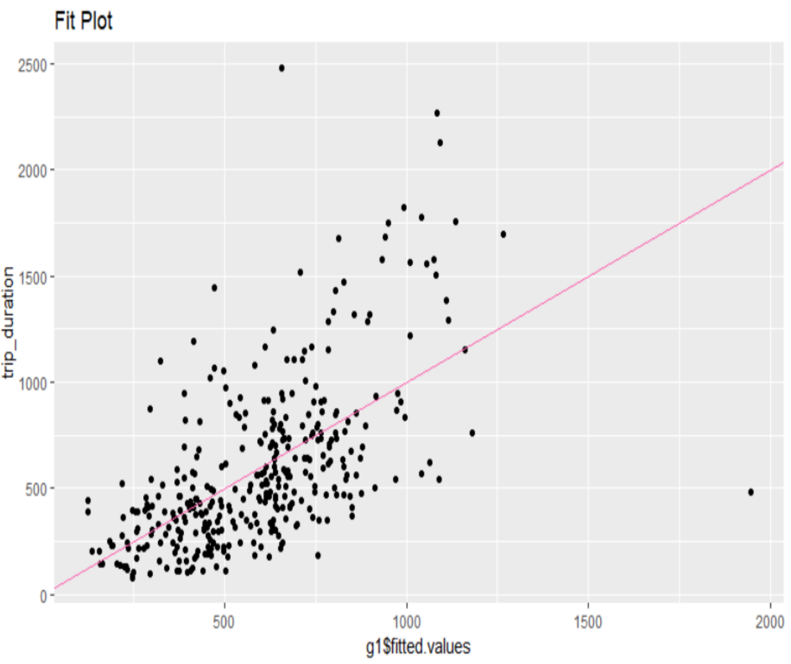
Challenges to OLS

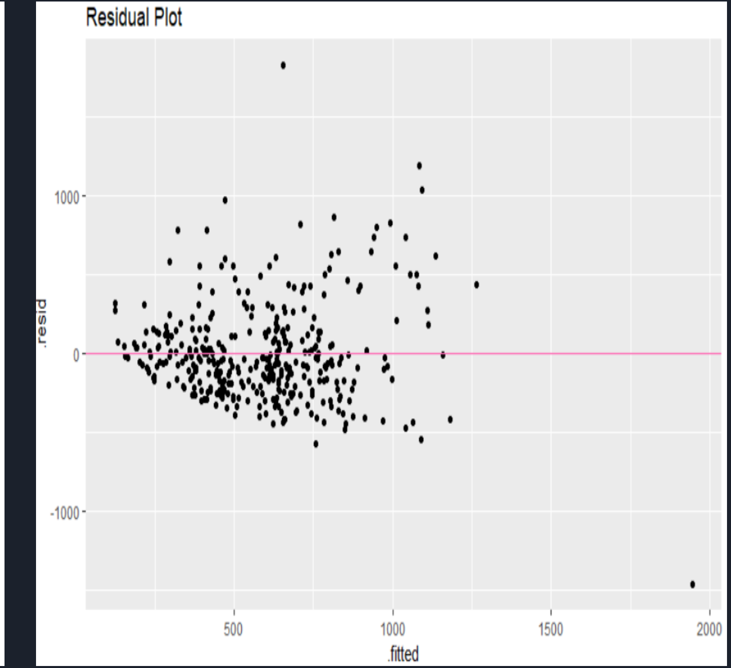
Model Validation

Prediction

Fit Plot and Residual Plot

- g1 is not a very good model as points are scattered.
- Residual plot shows a strong pattern

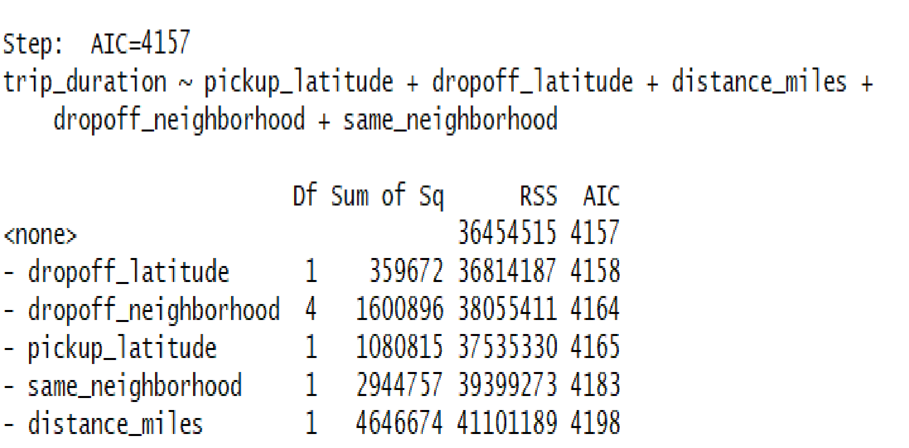




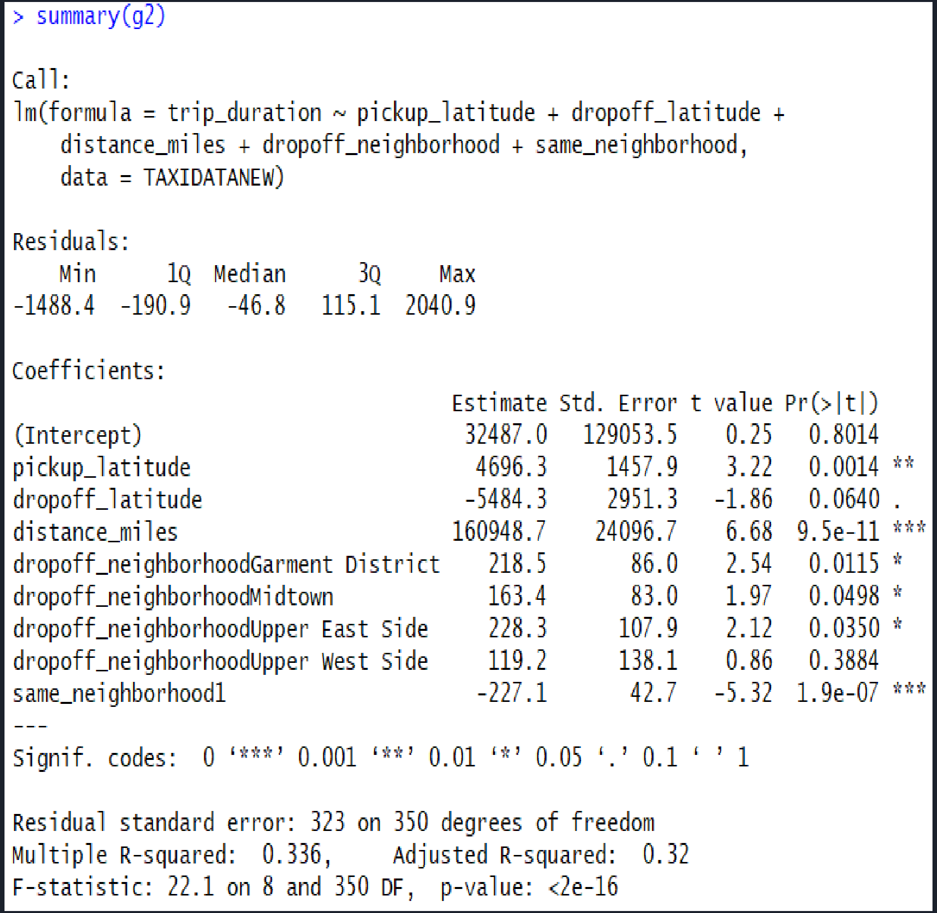
Data Acquisition — Data Cleaning — EDA — **Building Data Models** — Challenges to OLS — Model Validation — Prediction

Model Fitting

- g2 : Applied Stepwise Regression on the biggest model (g1).



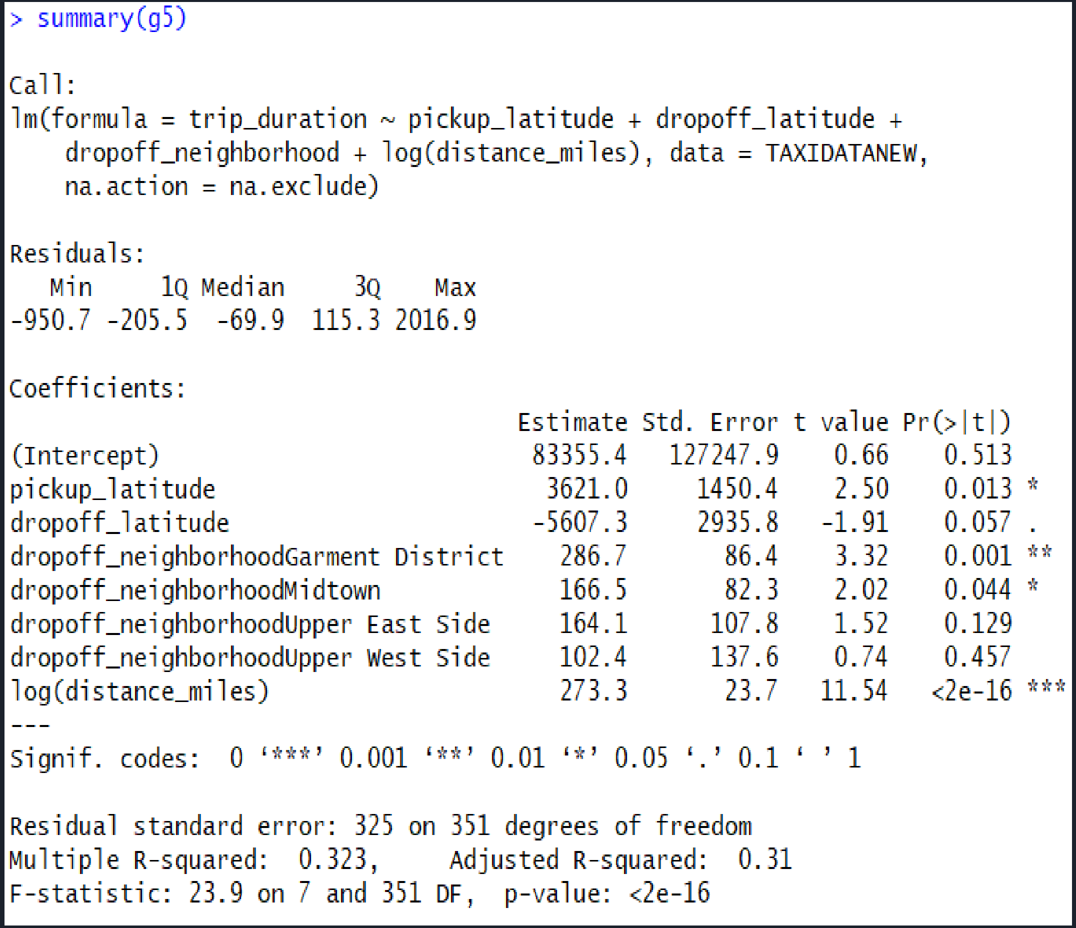
```
Step:  AIC=4157
trip_duration ~ pickup_latitude + dropoff_latitude + distance_miles +
    dropoff_neighborhood + same_neighborhood

                       Df Sum of Sq      RSS  AIC
<none>                              36454515 4157
- dropoff_latitude      1    359672 36814187 4158
- dropoff_neighborhood  4   1600896 38055411 4164
- pickup_latitude       1   1080815 37535330 4165
- same_neighborhood     1   2944757 39399273 4183
- distance_miles        1   4646674 41101189 4198
```



```
> summary(g2)

Call:
lm(formula = trip_duration ~ pickup_latitude + dropoff_latitude +
    distance_miles + dropoff_neighborhood + same_neighborhood,
    data = TAXIDATANEW)

Residuals:
    Min      1Q  Median      3Q     Max
-1488.4  -190.9   -46.8   115.1  2040.9

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                         32487.0   129053.5    0.25   0.8014
pickup_latitude                      4696.3     1457.9    3.22   0.0014 **
dropoff_latitude                    -5484.3     2951.3   -1.86   0.0640 .
distance_miles                     160948.7    24096.7    6.68  9.5e-11 ***
dropoff_neighborhoodGarment District  218.5       86.0    2.54   0.0115 *
dropoff_neighborhoodMidtown           163.4       83.0    1.97   0.0498 *
dropoff_neighborhoodUpper East Side   228.3      107.9    2.12   0.0350 *
dropoff_neighborhoodUpper West Side   119.2      138.1    0.86   0.3884
same_neighborhood1                   -227.1       42.7   -5.32  1.9e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 323 on 350 degrees of freedom
Multiple R-squared:  0.336,     Adjusted R-squared:  0.32
F-statistic: 22.1 on 8 and 350 DF,  p-value: <2e-16
```

Model Fitting

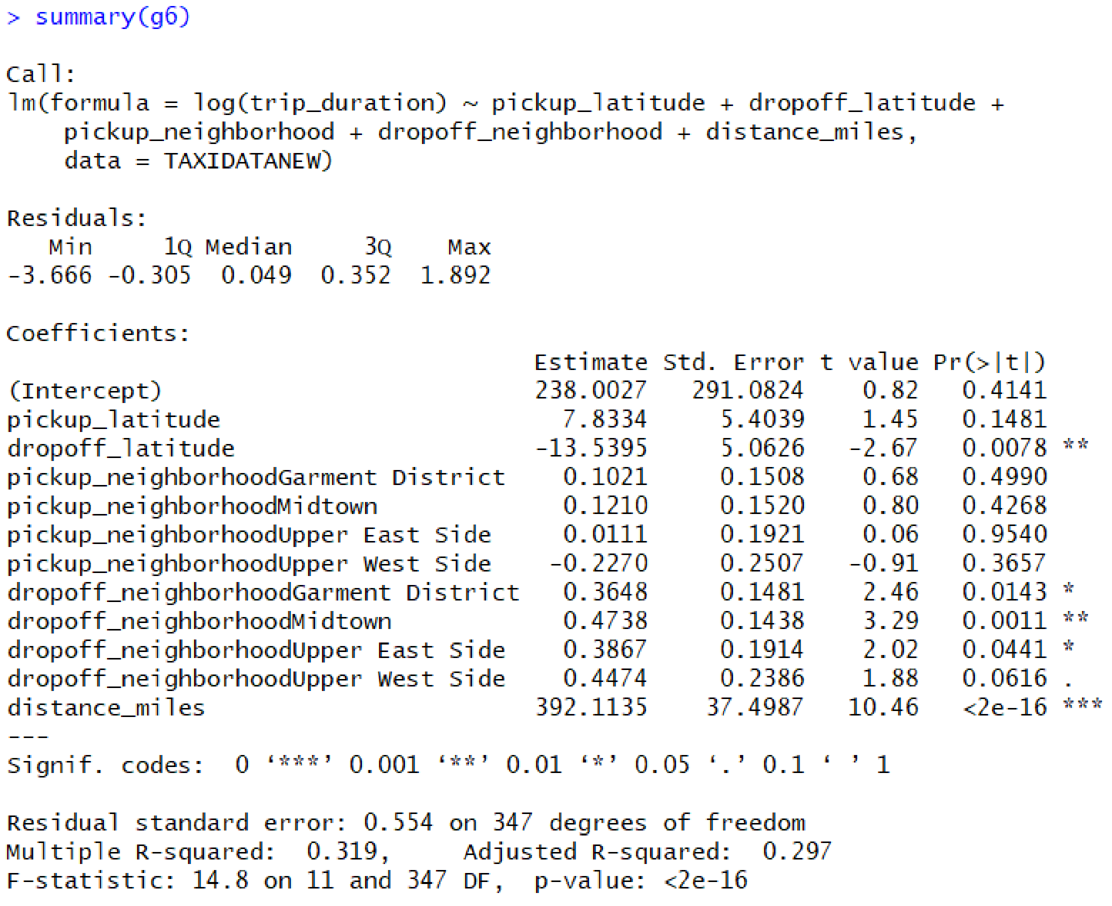
- g5 : Applied Stepwise Regression with one predictor variable as log transformed.



```
> summary(g5)

Call:
lm(formula = trip_duration ~ pickup_latitude + dropoff_latitude +
    dropoff_neighborhood + log(distance_miles), data = TAXIDATANEW,
    na.action = na.exclude)

Residuals:
   Min     1Q Median     3Q    Max
-950.7 -205.5  -69.9  115.3 2016.9

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        83355.4   127247.9    0.66    0.513
pickup_latitude                     3621.0     1450.4    2.50    0.013 *
dropoff_latitude                   -5607.3     2935.8   -1.91    0.057 .
dropoff_neighborhoodGarment District  286.7       86.4    3.32    0.001 **
dropoff_neighborhoodMidtown          166.5       82.3    2.02    0.044 *
dropoff_neighborhoodUpper East Side  164.1      107.8    1.52    0.129
dropoff_neighborhoodUpper West Side  102.4      137.6    0.74    0.457
log(distance_miles)                  273.3       23.7   11.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 325 on 351 degrees of freedom
Multiple R-squared:  0.323,     Adjusted R-squared:  0.31
F-statistic: 23.9 on 7 and 351 DF,  p-value: <2e-16
```

Model Fitting

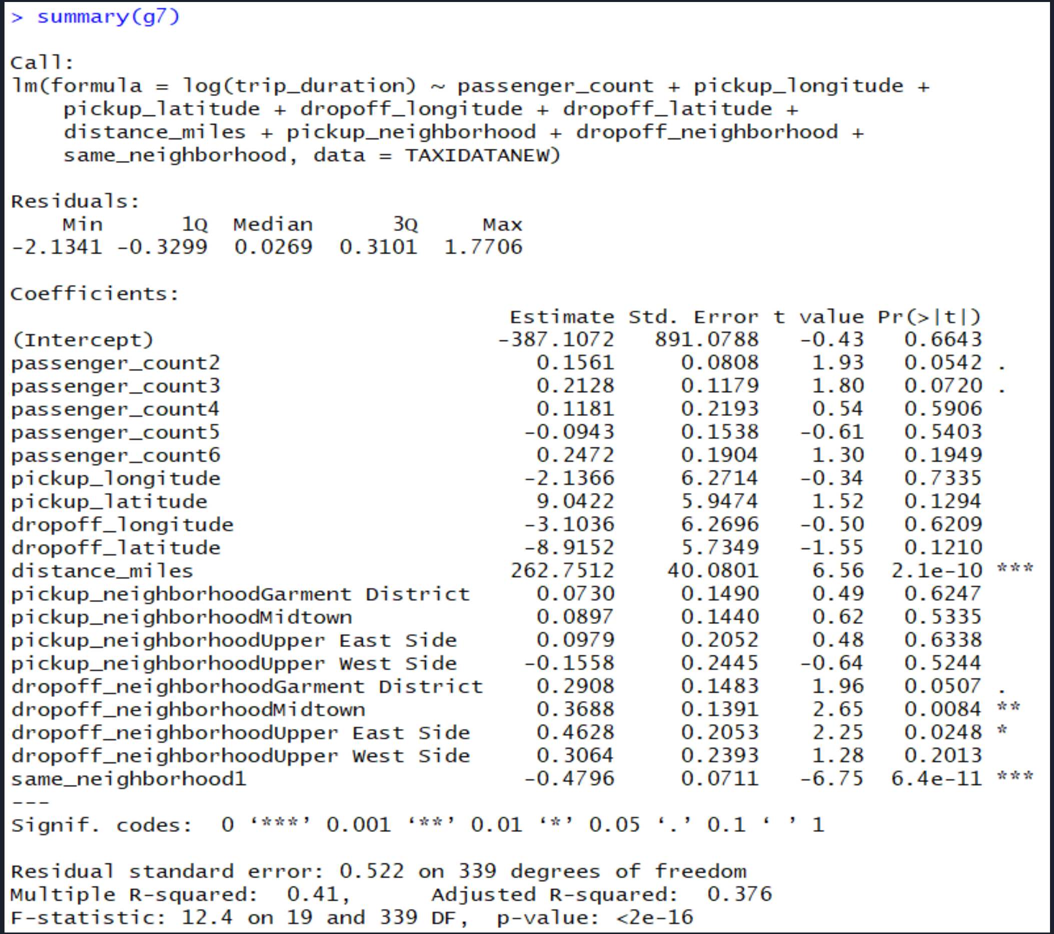
- g6 : Applied Stepwise regression with target variable as log transformed.



```
> summary(g6)

Call:
lm(formula = log(trip_duration) ~ pickup_latitude + dropoff_latitude +
    pickup_neighborhood + dropoff_neighborhood + distance_miles,
    data = TAXIDATANEW)

Residuals:
   Min     1Q Median     3Q    Max
-3.666 -0.305  0.049  0.352  1.892

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         238.0027   291.0824    0.82   0.4141
pickup_latitude                       7.8334     5.4039    1.45   0.1481
dropoff_latitude                    -13.5395     5.0626   -2.67   0.0078 **
pickup_neighborhoodGarment District   0.1021     0.1508    0.68   0.4990
pickup_neighborhoodMidtown            0.1210     0.1520    0.80   0.4268
pickup_neighborhoodUpper East Side    0.0111     0.1921    0.06   0.9540
pickup_neighborhoodUpper West Side   -0.2270     0.2507   -0.91   0.3657
dropoff_neighborhoodGarment District  0.3648     0.1481    2.46   0.0143 *
dropoff_neighborhoodMidtown           0.4738     0.1438    3.29   0.0011 **
dropoff_neighborhoodUpper East Side   0.3867     0.1914    2.02   0.0441 *
dropoff_neighborhoodUpper West Side   0.4474     0.2386    1.88   0.0616 .
distance_miles                      392.1135    37.4987   10.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.554 on 347 degrees of freedom
Multiple R-squared:  0.319,     Adjusted R-squared:  0.297
F-statistic: 14.8 on 11 and 347 DF,  p-value: <2e-16
```

Model Fitting

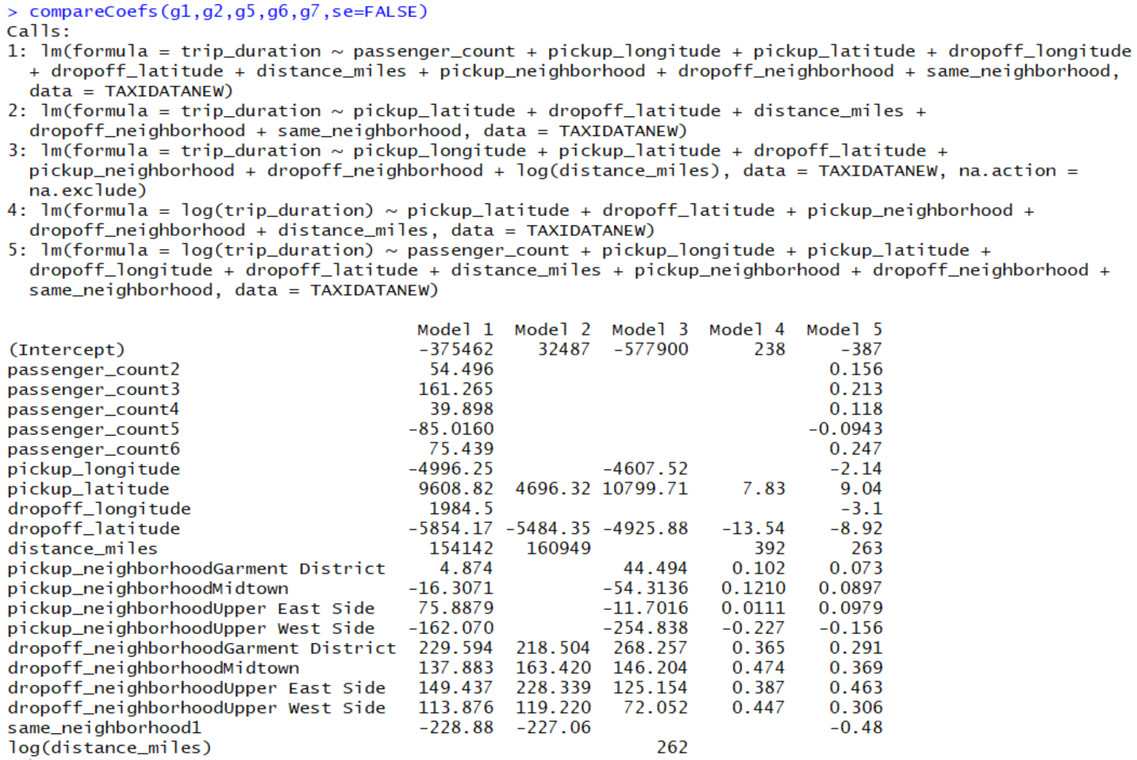
- g7 : Applied Stepwise regression on the big model with target variable as log transformed.



```
> summary(g7)

Call:
lm(formula = log(trip_duration) ~ passenger_count + pickup_longitude +
    pickup_latitude + dropoff_longitude + dropoff_latitude +
    distance_miles + pickup_neighborhood + dropoff_neighborhood +
    same_neighborhood, data = TAXIDATANEW)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1341 -0.3299  0.0269  0.3101  1.7706

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        -387.1072   891.0788   -0.43   0.6643
passenger_count2                      0.1561     0.0808    1.93   0.0542 .
passenger_count3                      0.2128     0.1179    1.80   0.0720 .
passenger_count4                      0.1181     0.2193    0.54   0.5906
passenger_count5                     -0.0943     0.1538   -0.61   0.5403
passenger_count6                      0.2472     0.1904    1.30   0.1949
pickup_longitude                     -2.1366     6.2714   -0.34   0.7335
pickup_latitude                       9.0422     5.9474    1.52   0.1294
dropoff_longitude                    -3.1036     6.2696   -0.50   0.6209
dropoff_latitude                     -8.9152     5.7349   -1.55   0.1210
distance_miles                      262.7512    40.0801    6.56  2.1e-10 ***
pickup_neighborhoodGarment District   0.0730     0.1490    0.49   0.6247
pickup_neighborhoodMidtown            0.0897     0.1440    0.62   0.5335
pickup_neighborhoodUpper East Side    0.0979     0.2052    0.48   0.6338
pickup_neighborhoodUpper West Side   -0.1558     0.2445   -0.64   0.5244
dropoff_neighborhoodGarment District  0.2908     0.1483    1.96   0.0507 .
dropoff_neighborhoodMidtown           0.3688     0.1391    2.65   0.0084 **
dropoff_neighborhoodUpper East Side   0.4628     0.2053    2.25   0.0248 *
dropoff_neighborhoodUpper West Side   0.3064     0.2393    1.28   0.2013
same_neighborhood1                   -0.4796     0.0711   -6.75  6.4e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.522 on 339 degrees of freedom
Multiple R-squared:  0.41,	Adjusted R-squared:  0.376
F-statistic: 12.4 on 19 and 339 DF,  p-value: <2e-16
```

Compare Coefficients

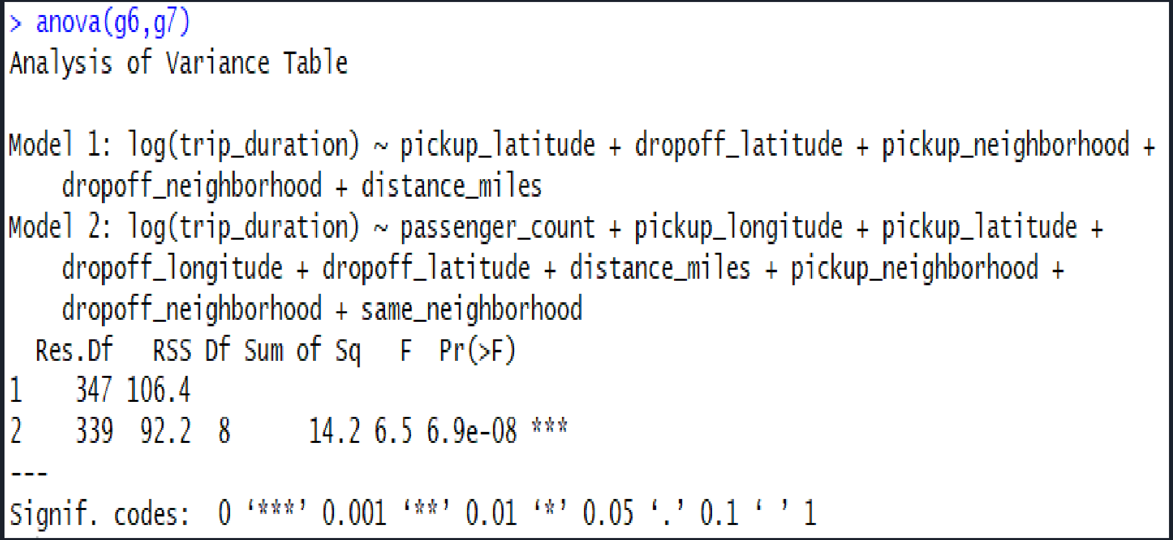
- Comparing the coefficients, we realize that model 4 i.e. g6 is better for predicting Trip_duration



```
> compareCoefs(g1,g2,g5,g6,g7,se=FALSE)
Calls:
1: lm(formula = trip_duration ~ passenger_count + pickup_longitude + pickup_latitude + dropoff_longitude
  + dropoff_latitude + distance_miles + pickup_neighborhood + dropoff_neighborhood + same_neighborhood,
  data = TAXIDATANEW)
2: lm(formula = trip_duration ~ pickup_latitude + dropoff_latitude + distance_miles +
  dropoff_neighborhood + same_neighborhood, data = TAXIDATANEW)
3: lm(formula = trip_duration ~ pickup_longitude + pickup_latitude + dropoff_latitude +
  pickup_neighborhood + dropoff_neighborhood + log(distance_miles), data = TAXIDATANEW, na.action =
  na.exclude)
4: lm(formula = log(trip_duration) ~ pickup_latitude + dropoff_latitude + pickup_neighborhood +
  dropoff_neighborhood + distance_miles, data = TAXIDATANEW)
5: lm(formula = log(trip_duration) ~ passenger_count + pickup_longitude + pickup_latitude +
  dropoff_longitude + dropoff_latitude + distance_miles + pickup_neighborhood + dropoff_neighborhood +
  same_neighborhood, data = TAXIDATANEW)

                                      Model 1  Model 2  Model 3  Model 4  Model 5
(Intercept)                           -375462    32487  -577900      238     -387
passenger_count2                       54.496                               0.156
passenger_count3                      161.265                               0.213
passenger_count4                       39.898                               0.118
passenger_count5                     -85.0160                             -0.0943
passenger_count6                       75.439                               0.247
pickup_longitude                     -4996.25          -4607.52             -2.14
pickup_latitude                       9608.82  4696.32 10799.71     7.83     9.04
dropoff_longitude                      1984.5                                -3.1
dropoff_latitude                     -5854.17 -5484.35 -4925.88   -13.54    -8.92
distance_miles                         154142   160949               392      263
pickup_neighborhoodGarment District     4.874            44.494    0.102    0.073
pickup_neighborhoodMidtown           -16.3071          -54.3136   0.1210   0.0897
pickup_neighborhoodUpper East Side    75.8879          -11.7016   0.0111   0.0979
pickup_neighborhoodUpper West Side   -162.070          -254.838   -0.227   -0.156
dropoff_neighborhoodGarment District  229.594  218.504  268.257    0.365    0.291
dropoff_neighborhoodMidtown           137.883  163.420  146.204    0.474    0.369
dropoff_neighborhoodUpper East Side   149.437  228.339  125.154    0.387    0.463
dropoff_neighborhoodUpper West Side   113.876  119.220   72.052    0.447    0.306
same_neighborhood1                    -228.88  -227.06                      -0.48
log(distance_miles)                                         262
```

Partial F-test - Transformed Target Variable

- Anova of g6 and g7
- As P- Value is less than the alpha (0.05), we conclude that the model g6 is better



```
> anova(g6,g7)
Analysis of Variance Table

Model 1: log(trip_duration) ~ pickup_latitude + dropoff_latitude + pickup_neighborhood +
    dropoff_neighborhood + distance_miles
Model 2: log(trip_duration) ~ passenger_count + pickup_longitude + pickup_latitude +
    dropoff_longitude + dropoff_latitude + distance_miles + pickup_neighborhood +
    dropoff_neighborhood + same_neighborhood
  Res.Df   RSS Df Sum of Sq   F  Pr(>F)
1    347 106.4
2    339  92.2  8      14.2 6.5 6.9e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Data Acquisition

Data Cleaning

EDA

Building Data Models

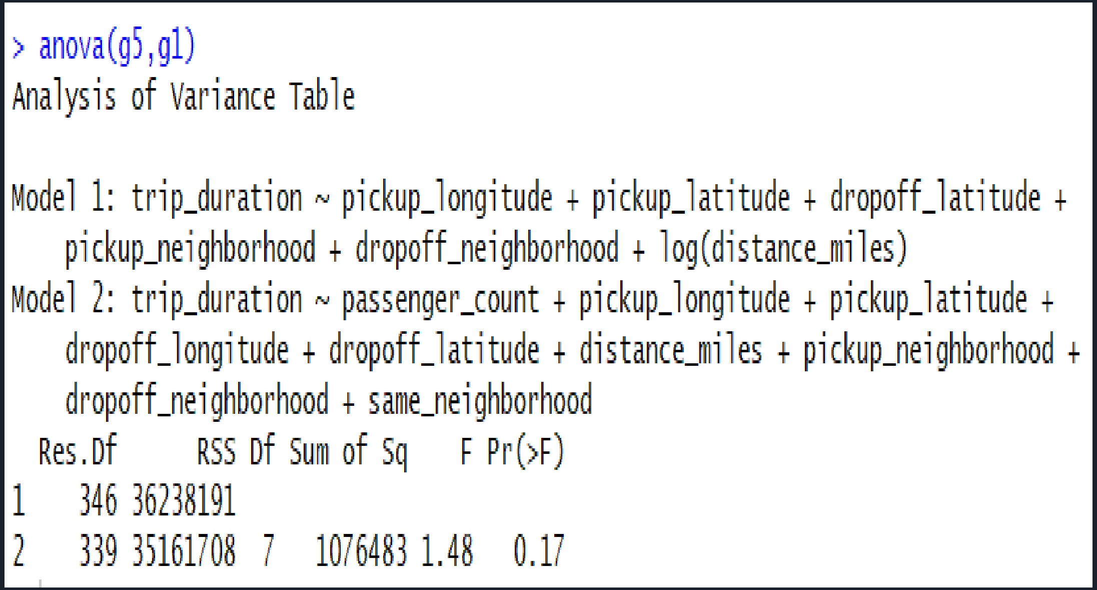
Challenges to OLS

Model Validation

Prediction

Partial F-test - Transformed Predictor Variable

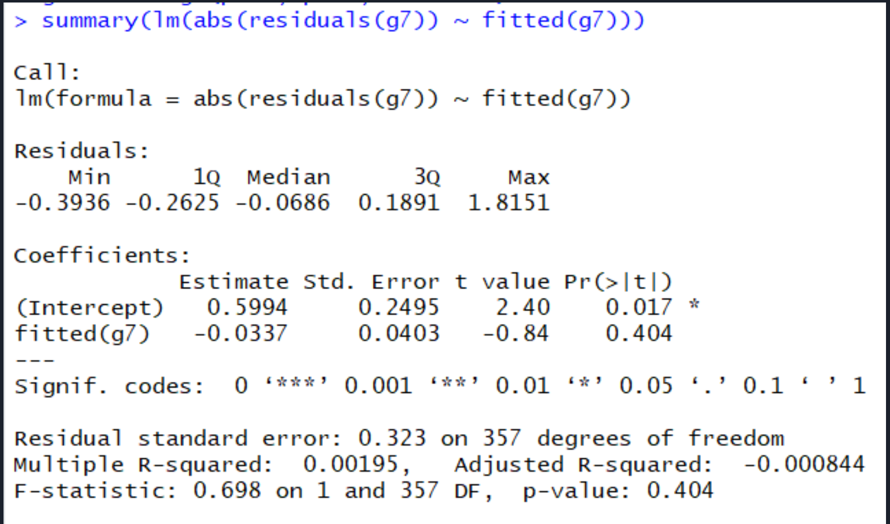
- Anova of g5 and g1
- As P- Value is greater than the alpha (0.05), we conclude that the smaller model, with one predictor as log transformed, is better i.e. g5 is better



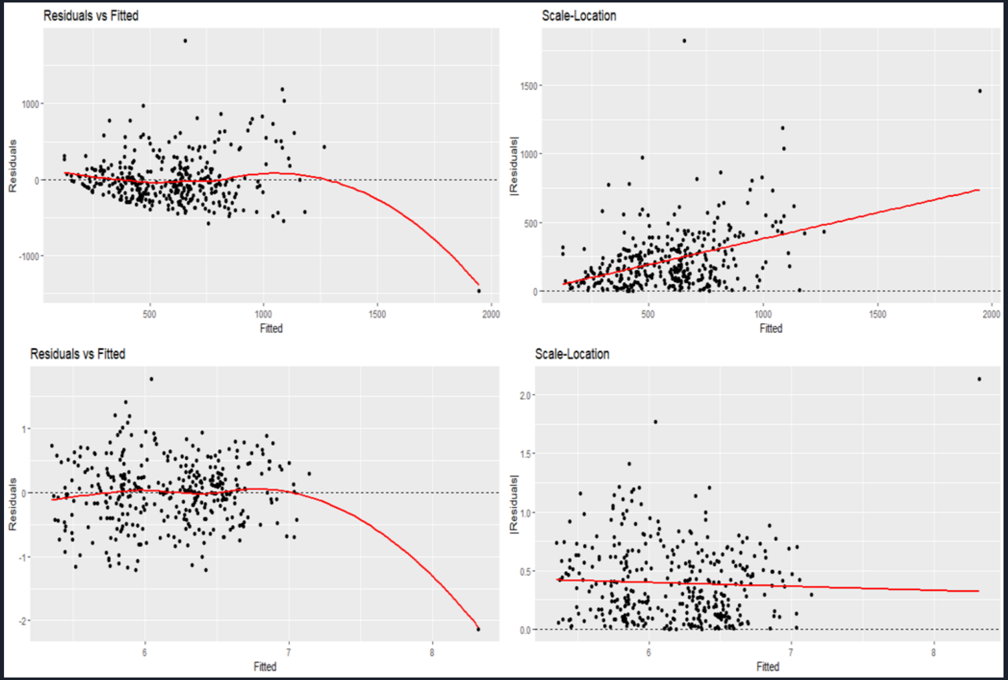
```
> anova(g5,g1)
Analysis of Variance Table

Model 1: trip_duration ~ pickup_longitude + pickup_latitude + dropoff_latitude +
    pickup_neighborhood + dropoff_neighborhood + log(distance_miles)
Model 2: trip_duration ~ passenger_count + pickup_longitude + pickup_latitude +
    dropoff_longitude + dropoff_latitude + distance_miles + pickup_neighborhood +
    dropoff_neighborhood + same_neighborhood
  Res.Df      RSS Df Sum of Sq    F Pr(>F)
1    346 36238191
2    339 35161708  7   1076483 1.48   0.17
```

Non-Constant Variance

The t-test does not reject constant error variance with a level of significance 5%, since the p-value, 0.404, is greater than 0.05

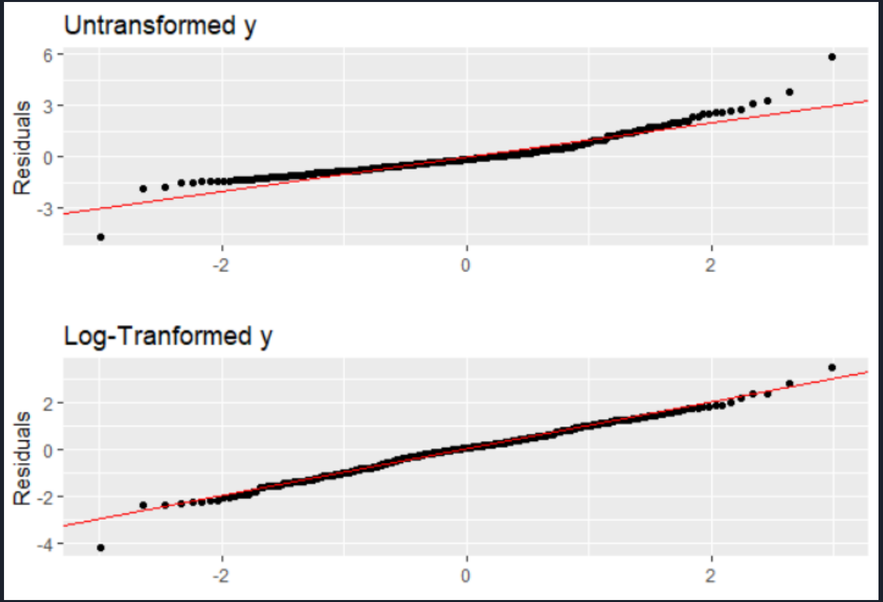


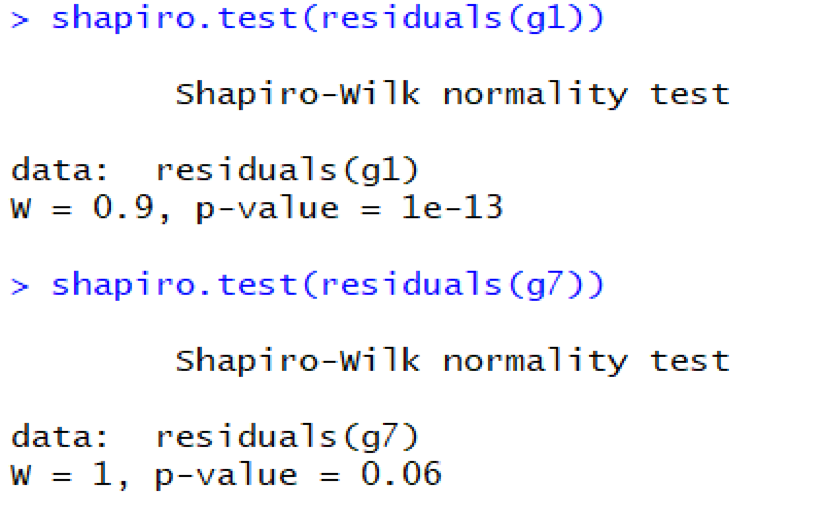
```
> summary(lm(abs(residuals(g7)) ~ fitted(g7)))

Call:
lm(formula = abs(residuals(g7)) ~ fitted(g7))

Residuals:
    Min      1Q  Median      3Q     Max 
-0.3936 -0.2625 -0.0686  0.1891  1.8151 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   0.5994     0.2495    2.40    0.017 *
fitted(g7)   -0.0337     0.0403   -0.84    0.404  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.323 on 357 degrees of freedom
Multiple R-squared:  0.00195,   Adjusted R-squared:  -0.000844 
F-statistic: 0.698 on 1 and 357 DF,  p-value: 0.404
```



Non-Normal Errors

Normal QQ-plots for detecting



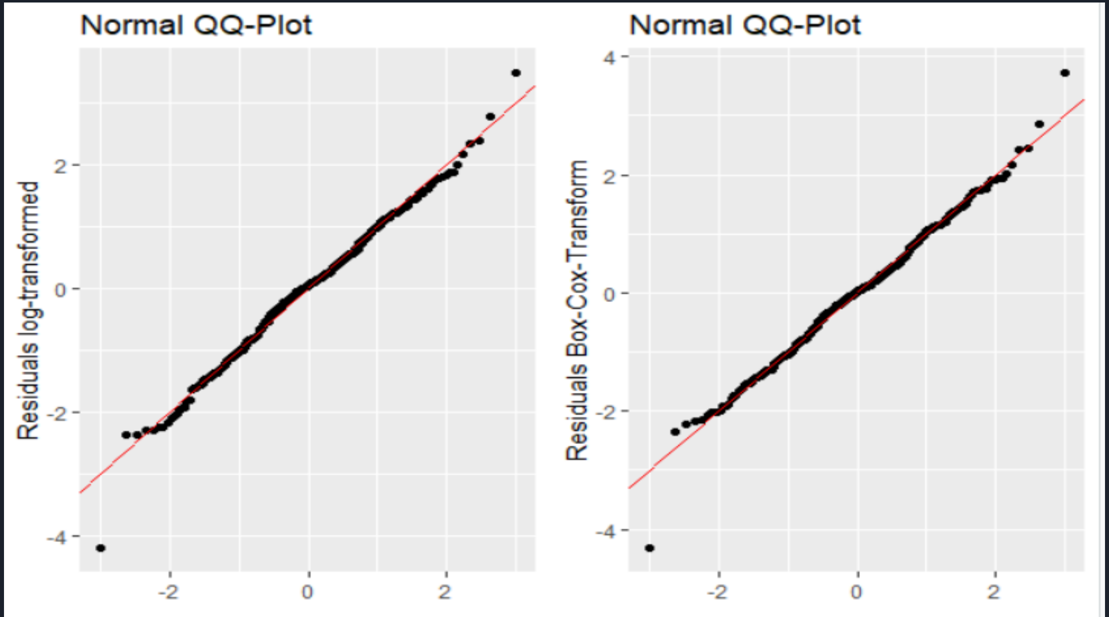


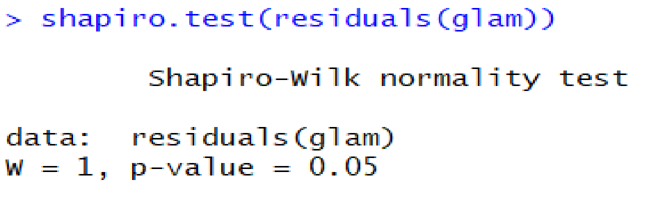
```
> shapiro.test(residuals(g1))

        Shapiro-Wilk normality test

data:  residuals(g1)
W = 0.9, p-value = 1e-13

> shapiro.test(residuals(g7))

        Shapiro-Wilk normality test

data:  residuals(g7)
W = 1, p-value = 0.06
```

We fail to reject the null hypothesis of normality for the residuals of log-transformed model with level of significance 5% since the p-value is greater than 0.05.

Non-Normal Errors

Box-Cox Power Transformation



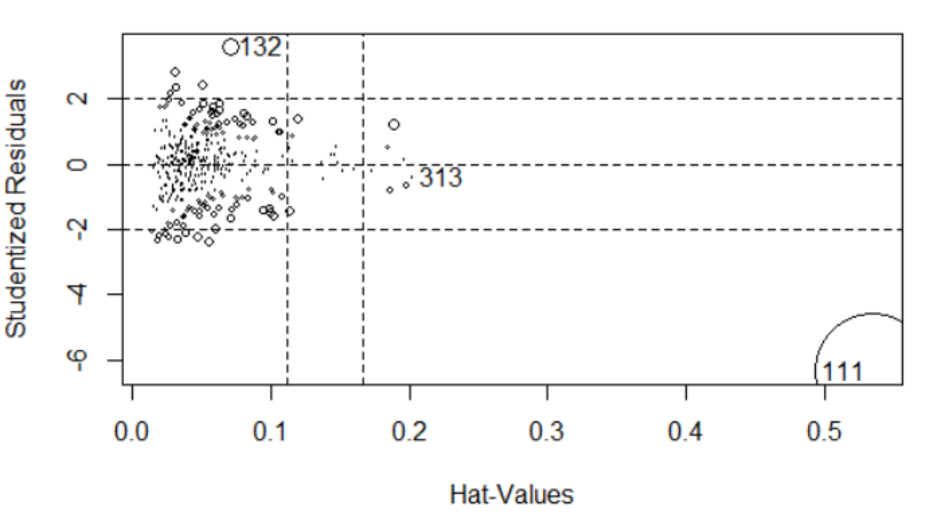


```
> shapiro.test(residuals(glam))

        Shapiro-Wilk normality test

data:  residuals(glam)
W = 1, p-value = 0.05
```

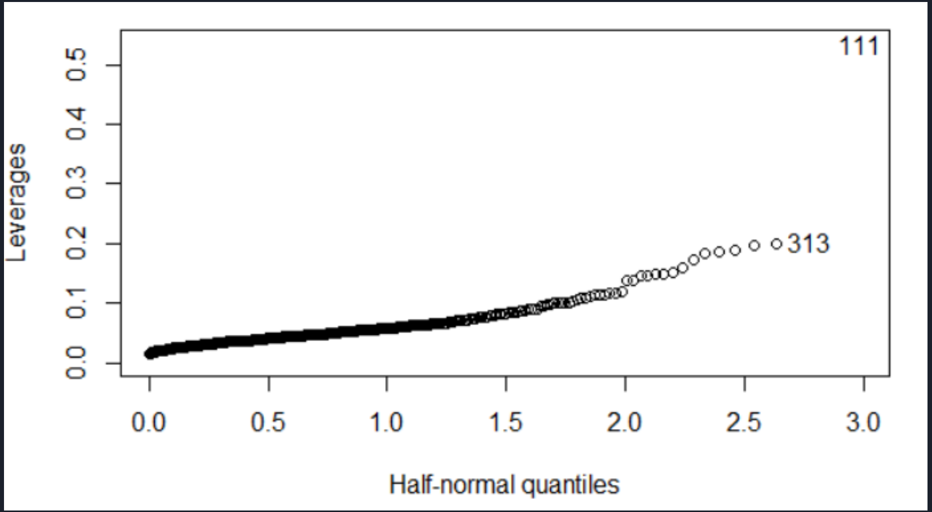
The Shapiro-Wilk test concludes that the errors are normal for the Box-Cox Transformed model with level of significance 5% since the p-value is approximately equal to 0.05.

Influential Outliers

Influence Plot

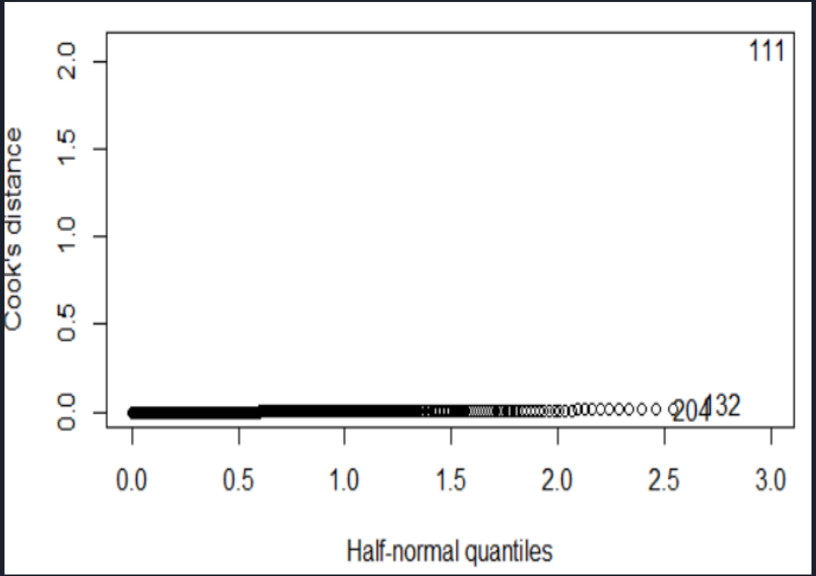


Half Normal Plot

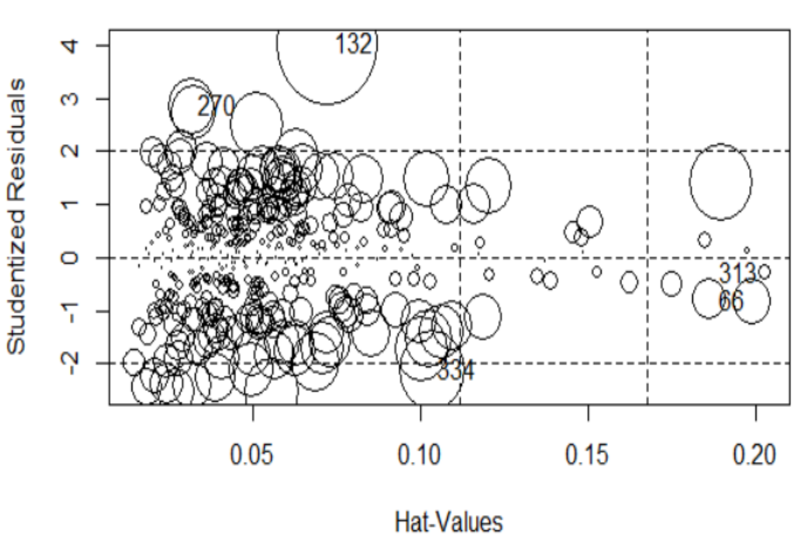


Influential Outlier

Half normal plot of Cook’s Distance

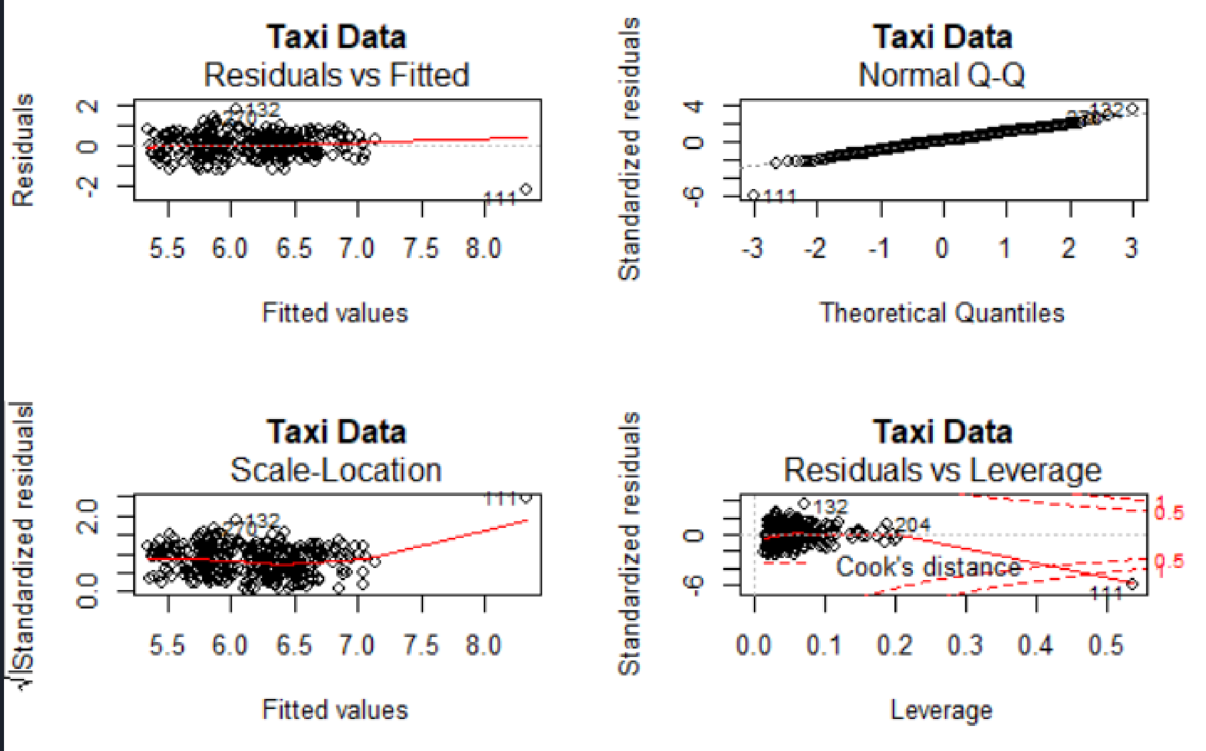


Influence Plot (without 111 data point)



Data Acquisition — Data Cleaning — EDA — Building Data Models — **Challenges to OLS** — Model Validation — Prediction

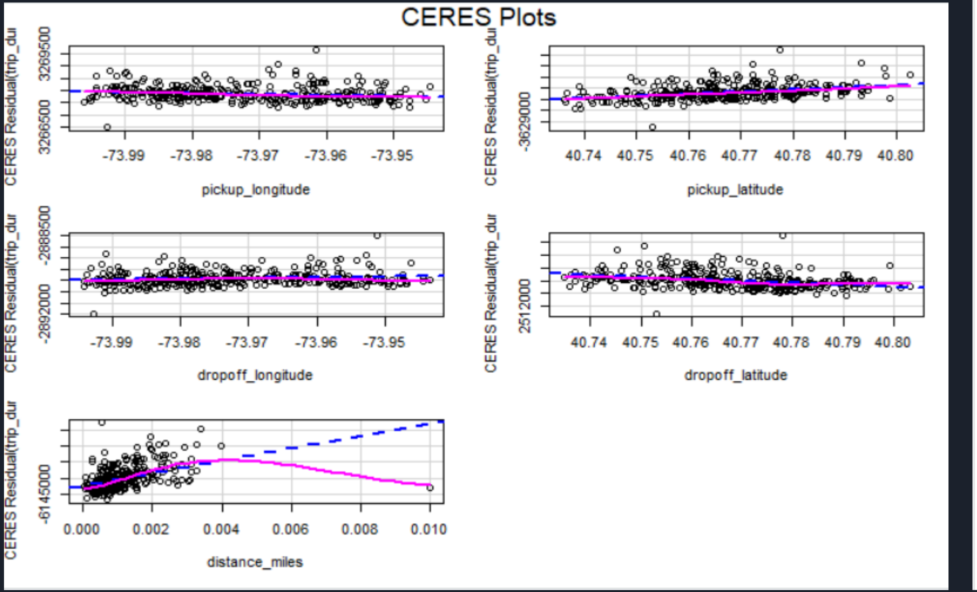
Omnibus diagnostic plot function



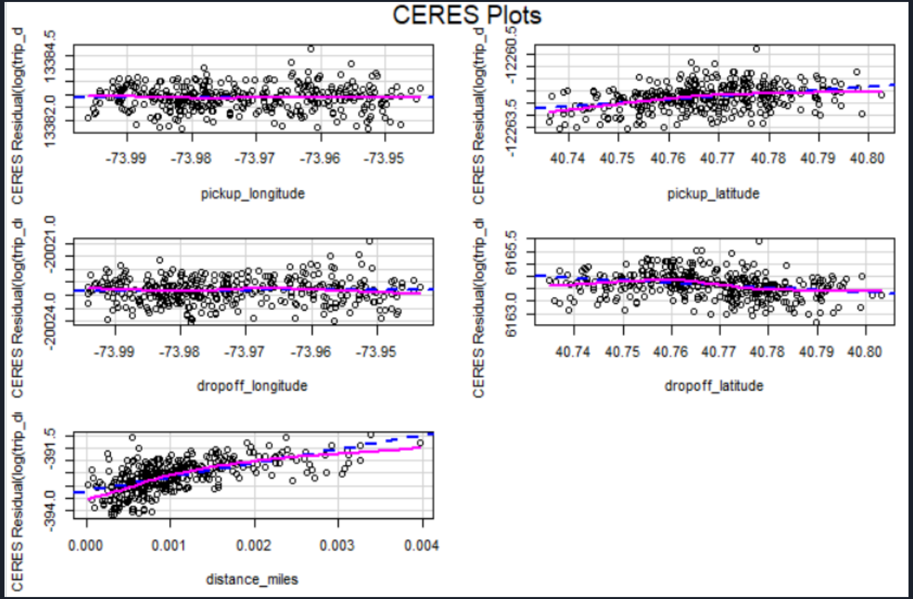
Data Acquisition — Data Cleaning — EDA — Building Data Models — **Challenges to OLS** — Model Validation — Prediction

Correct Model Specification

Target Variable: Trip Duration



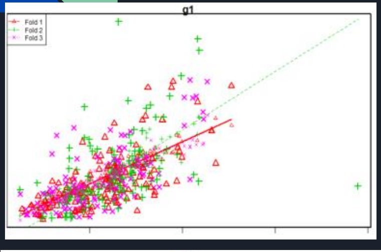
Target Variable: log(Trip Duration)

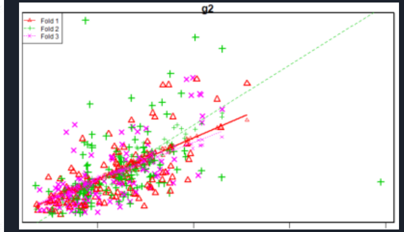


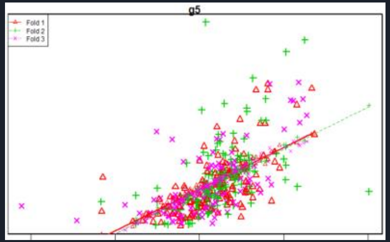
After log transformation, the model looks normal

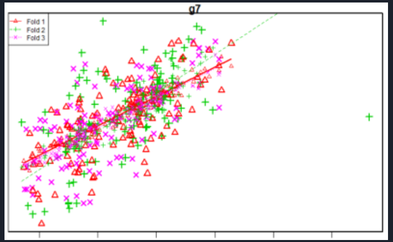
Data Acquisition — Data Cleaning — EDA — Building Data Models — **Challenges to OLS** — Model Validation — Prediction

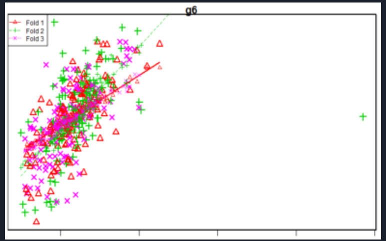
Cross Validation of linear models

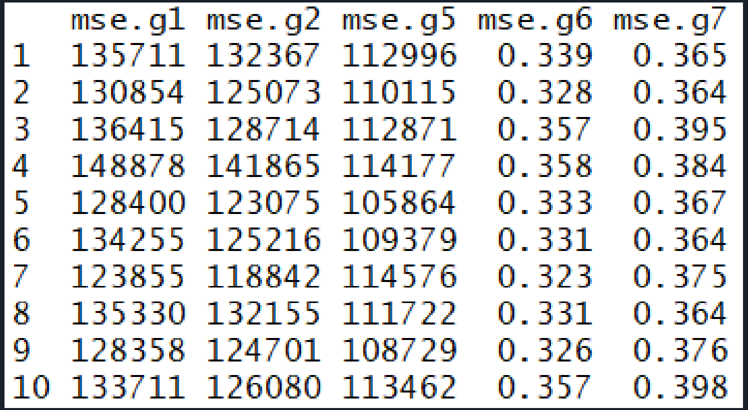












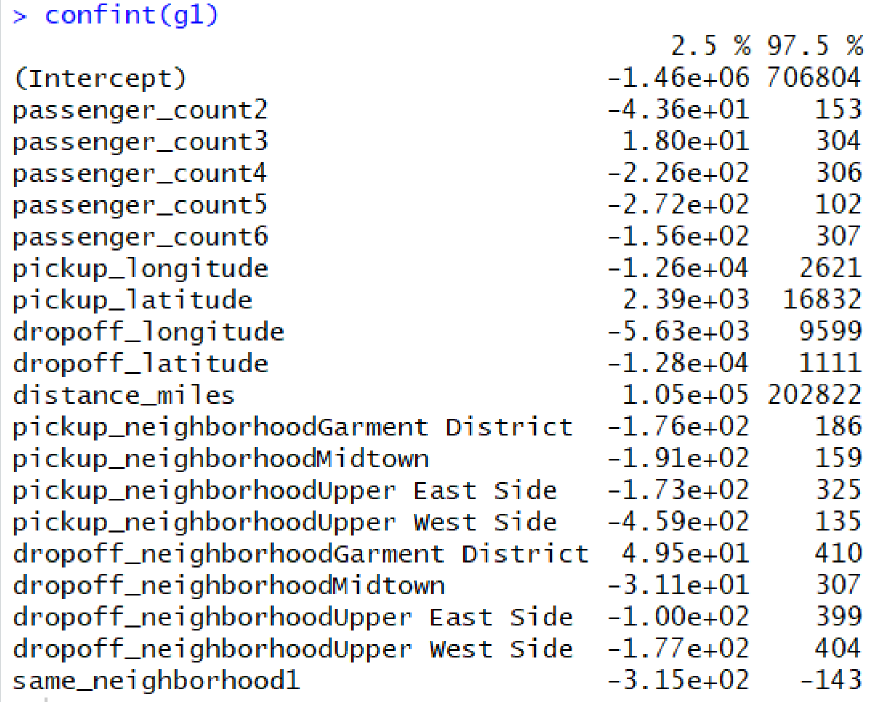
```
   mse.g1 mse.g2 mse.g5 mse.g6 mse.g7
1  135711 132367 112996  0.339  0.365
2  130854 125073 110115  0.328  0.364
3  136415 128714 112871  0.357  0.395
4  148878 141865 114177  0.358  0.384
5  128400 123075 105864  0.333  0.367
6  134255 125216 109379  0.331  0.364
7  123855 118842 114576  0.323  0.375
8  135330 132155 111722  0.331  0.364
9  128358 124701 108729  0.326  0.376
10 133711 126080 113462  0.357  0.398
```

From the cross validation predicted values, we can see that the model g6 holds better mse overall.

*Rest of the graphs is in Rmd File.

Individual Confidence Interval

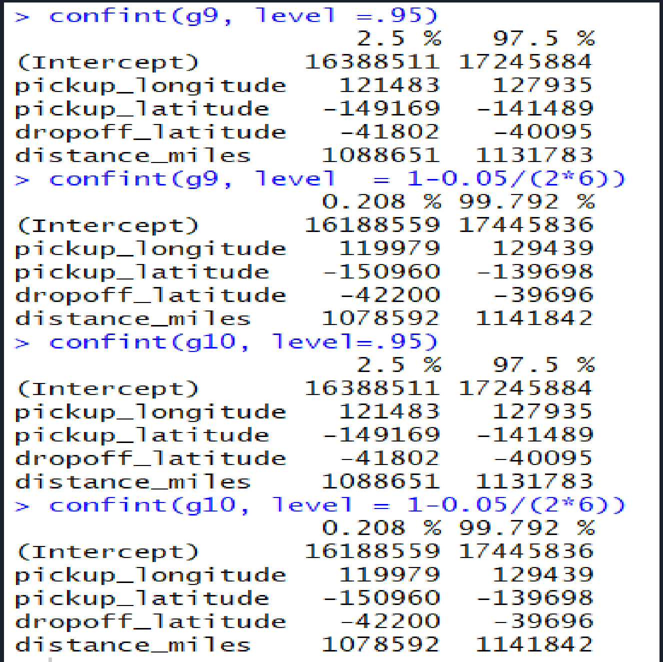
- We are focusing in the 95% CI on the model g1 .



```
> confint(g1)
                                          2.5 % 97.5 %
(Intercept)                           -1.46e+06 706804
passenger_count2                      -4.36e+01    153
passenger_count3                       1.80e+01    304
passenger_count4                      -2.26e+02    306
passenger_count5                      -2.72e+02    102
passenger_count6                      -1.56e+02    307
pickup_longitude                      -1.26e+04   2621
pickup_latitude                        2.39e+03  16832
dropoff_longitude                     -5.63e+03   9599
dropoff_latitude                      -1.28e+04   1111
distance_miles                         1.05e+05 202822
pickup_neighborhoodGarment District   -1.76e+02    186
pickup_neighborhoodMidtown            -1.91e+02    159
pickup_neighborhoodUpper East Side    -1.73e+02    325
pickup_neighborhoodUpper West Side    -4.59e+02    135
dropoff_neighborhoodGarment District   4.95e+01    410
dropoff_neighborhoodMidtown           -3.11e+01    307
dropoff_neighborhoodUpper East Side   -1.00e+02    399
dropoff_neighborhoodUpper West Side   -1.77e+02    404
same_neighborhood1                    -3.15e+02   -143
```

Comparing CI of Coefficients with and without Bonferroni Models

- Comparing the model g and model g where Trip duration is converted to factor.



```
> confint(g9, level =.95)
                     2.5 %   97.5 %
(Intercept)       16388511 17245884
pickup_longitude    121483   127935
pickup_latitude    -149169  -141489
dropoff_latitude    -41802   -40095
distance_miles     1088651  1131783
> confint(g9, level  = 1-0.05/(2*6))
                   0.208 % 99.792 %
(Intercept)       16188559 17445836
pickup_longitude    119979   129439
pickup_latitude    -150960  -139698
dropoff_latitude    -42200   -39696
distance_miles     1078592  1141842
> confint(g10, level=.95)
                     2.5 %   97.5 %
(Intercept)       16388511 17245884
pickup_longitude    121483   127935
pickup_latitude    -149169  -141489
dropoff_latitude    -41802   -40095
distance_miles     1088651  1131783
> confint(g10, level = 1-0.05/(2*6))
                   0.208 % 99.792 %
(Intercept)       16188559 17445836
pickup_longitude    119979   129439
pickup_latitude    -150960  -139698
dropoff_latitude    -42200   -39696
distance_miles     1078592  1141842
```

Thanks

Any Questions?