Uber and Lyft prices prediction

STATISTICAL LEARNING FINAL PROJECT

Prof. ALBERTO ROVERATO

MANOJ KUMAR NAGABANDI

2039097

Obtaining Data

Source: Kaggle

Data collection: Website of Massachusetts State, USA

Data set of size 693,071 rows and 57 columns

Outline of the project:

Tidy	Tidy the data
Individuate	Individuate some outcome variables to study
Explore	Explore the data visually
Build	Build significative models
Interpret and justify	Interpret and justify the results

Features in Data Set

- The rideshare dataset contain 56 features:
- "id" , "timestamp" , "hour" , "day" , "month", "datetime" , "timezone", "source", "destination", "cab_type", "product_id", "name", "price", "distance", "surge_multiplier", "latitude", "longitude,"sunriseTime", "uvIndexTime" ,"sunsetTime","short_summary", "long_summary, "windGustTime" ,"icon"
- "visibility", "dewpoint", "pressure", "windBearing", "cloudCover", "uvIndex", "ozone", "moonPhase","precipIntensityMax", "precipIntesity", "precipProbability, ", "humidity","windSpeed" , "windGust", , "visibility.1"
- "temperatureMin" , "temperatureMinTime", "temperatureMax" , "temperatureMaxTime", "apparentTemperatureMin" , "apparentTemperatureMinTime","apparentTemperatureMax" , "apparentTemperatureMaxTime" , "temperature" , "temperatureHigh", "temperatureHighTime","temperatureLow", "apparentTempertaure", "temperatureLowTime".

Climate Related features

Temperature related features

❖ Clean & Filter Data(Pre-processing)

- Initially, checks for NAN, infinite values, and missing values are done and 55,095 missing values are present in the data which are omitted.
- Both “visibility” and “visibility.1“ features have exactly the same data in their columns one of them is dropped.
- Checks for skewness are done on all numeric features and 15 features are found to be negatively skewed and 26 features are found to be positively skewed.
- Since skewness is greater than 3 for "surge_multiplier“ and "precipIntensity" features we apply cube root transformation is applied to normalize and reduce the skewness of the features.

Clean & Filter Data(Pre-processing)

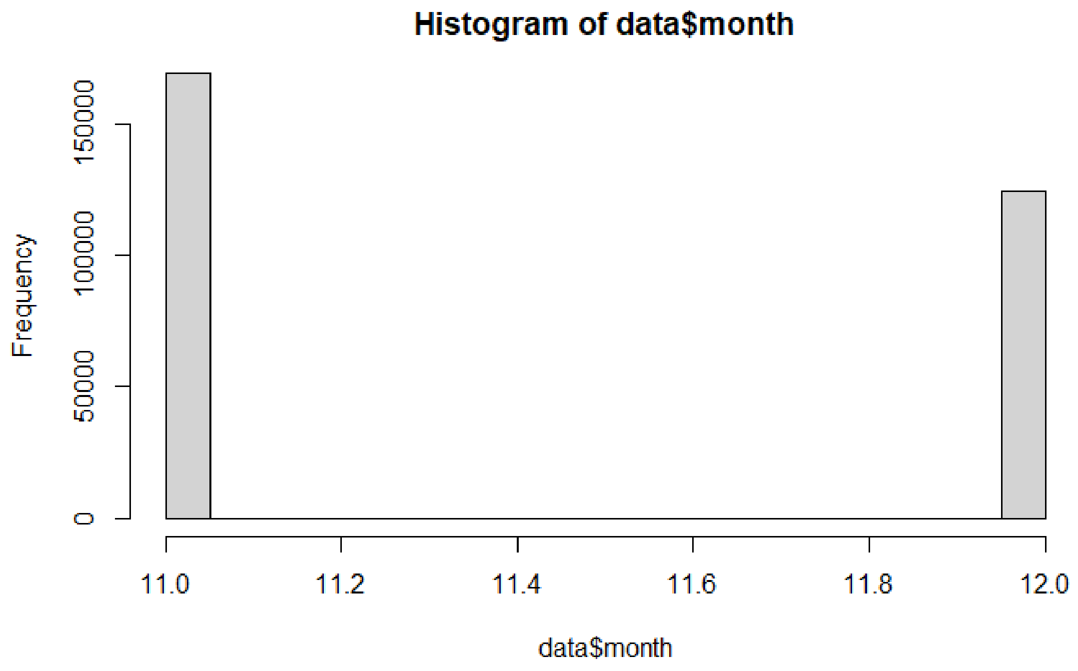
- Outliers are found by using the interquartile range.
- Features whose outliers(>10%) are:
- latitude ← **0.127 %**
- Visibility ← **0.197 %**
- temperatureHigh ← **0.236 %**
- apparentTemperatureHigh ← **0.103 %**
- apparentTemperatureLow ← **0.126 %**
- temperatureMax ← **0.197 %**
- apparentTemperatureMin ← **0.109 %**
- **After deleting rows with outlier values, the final dimension used is 293877 rows and 56 columns.**

Clean & Filter Data(Pre-processing)

- There are 11 features with character datatype are:
- Id, datetime, timezone, source, destination, cab_type, product_id, name, short_summary, long_summary, icon
- Since every row of the "id" feature values are unique and every row of the "timezone" feature has the same value.
- From the above 2 features model does not learn anything so they can be discarded.
- Finally, in the feature "product_id" we have unidentified information so this feature can be dropped as well.
- Therefore, there are 8 categorical features remaining to which one hot encoding is applied to convert them to binary vectors
- Therefore, the final size of our data frame is **293,877 rows and 102 columns**

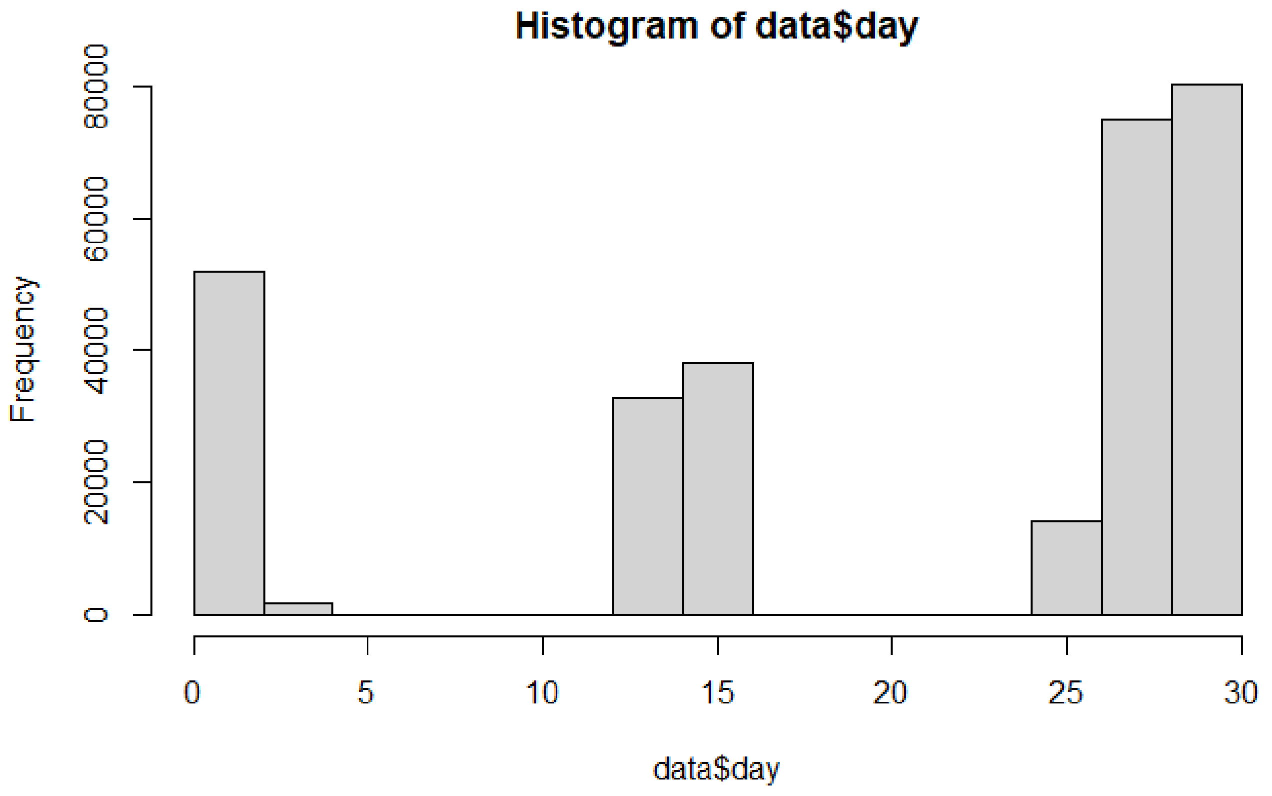
Exploratory Data Analysis (EDA)

- In which months did most of the rides occur?
- It appears that we only have November and December data in our monthly data
- November ←169512 values
  December ←124365 values



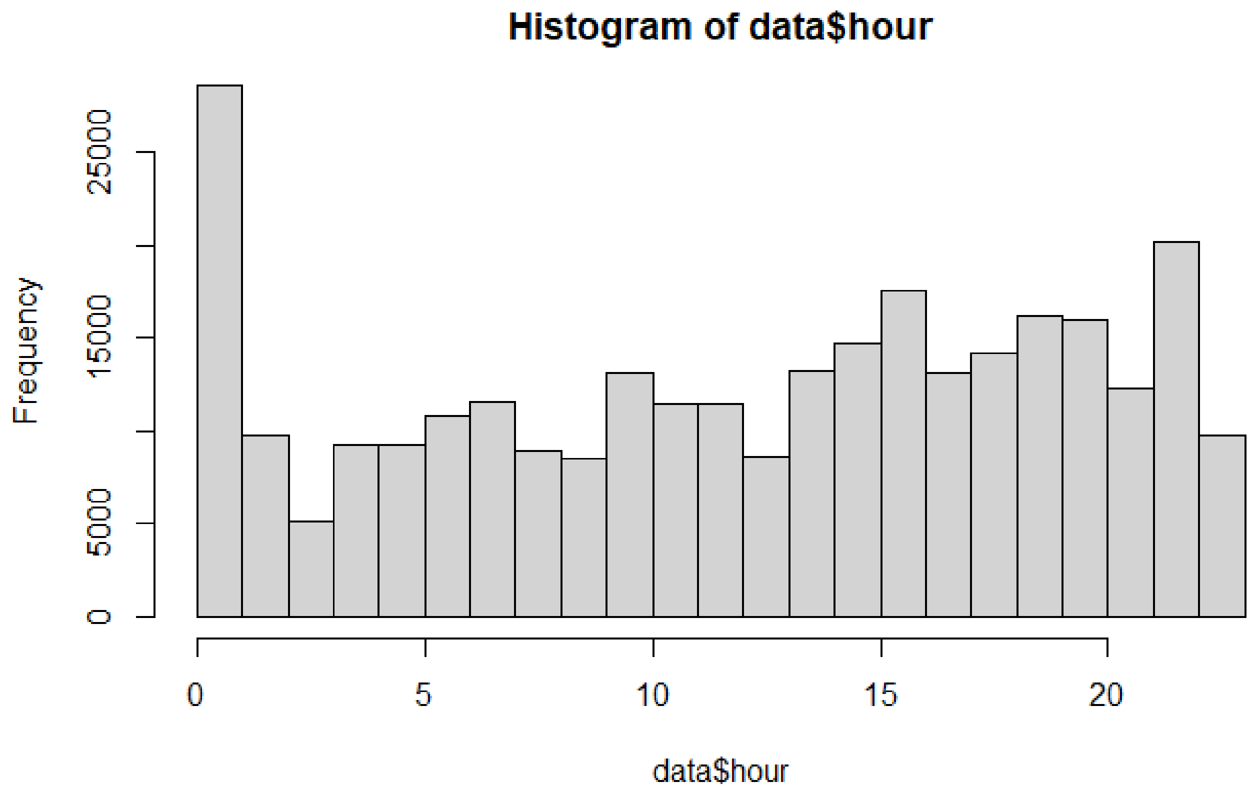
❖ Exploratory Data Analysis (EDA)

- On what dates most rides have not taken place?
- We have many gaps in data between 4th - 12th days and from 17th - 25th days data are not present in each month



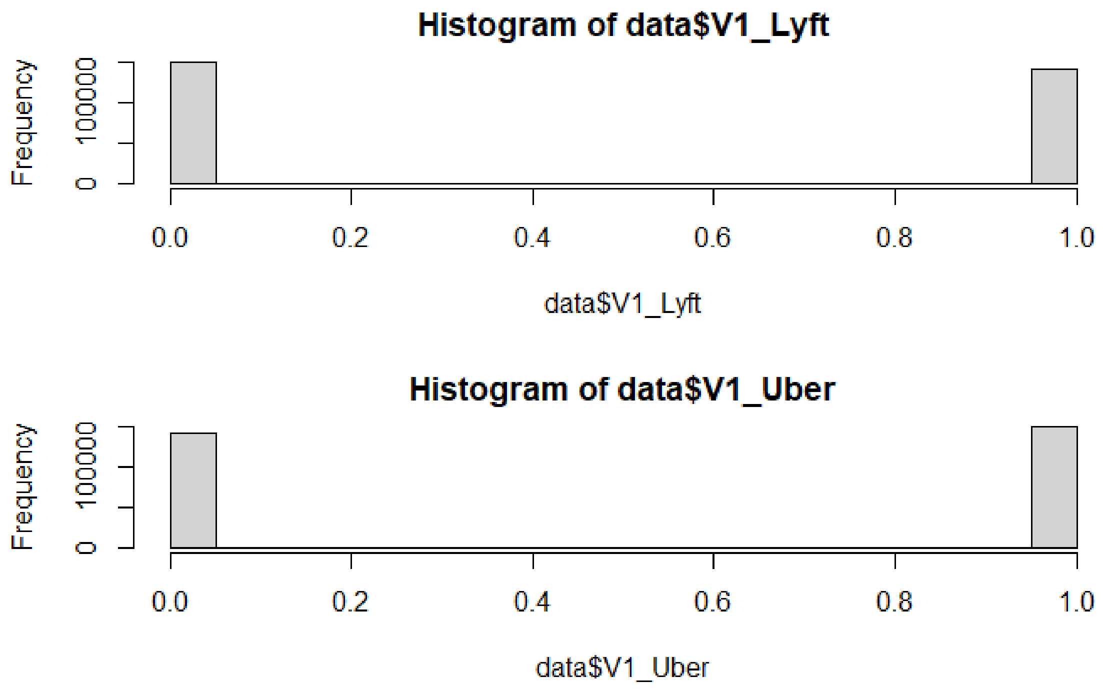
Exploratory Data Analysis (EDA)

- How many hours of data are logged?
- We have logged data of 24hrs



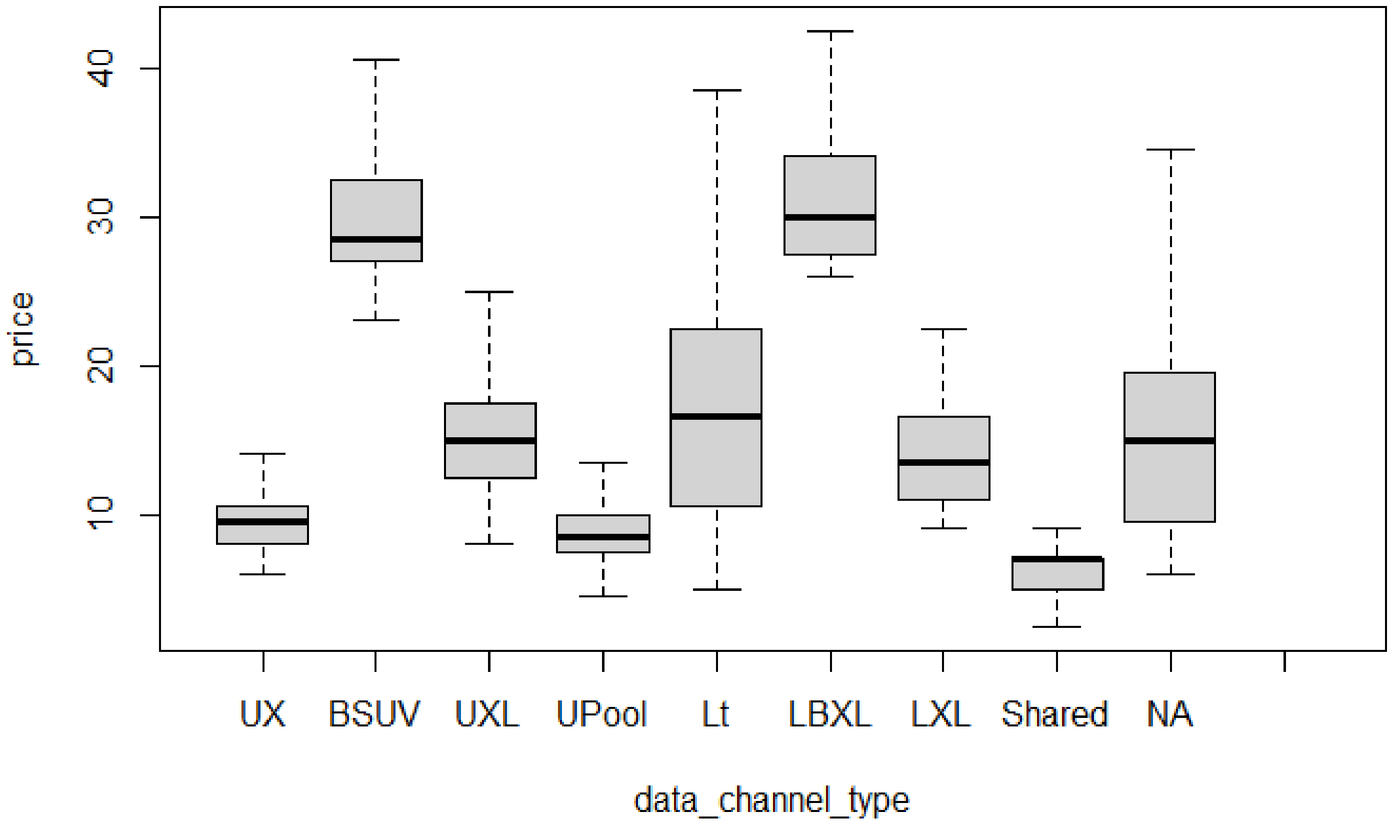
Exploratory Data Analysis (EDA)

- What company cabs are used more?
- Uber and Lyft categories have almost same size
- Uber ← 151,560 values
- Lyft ← 142,317 values



Exploratory Data Analysis (EDA)

- Which cab has best price per mile?
- Firstly, Lyft XL has a slightly lower fare per mile than UberXL.
- Uber Black SUV shows a lower rate than Lyft Black XL.
- Lyft's ordinary ride when compared to UberX has higher fares per mile.



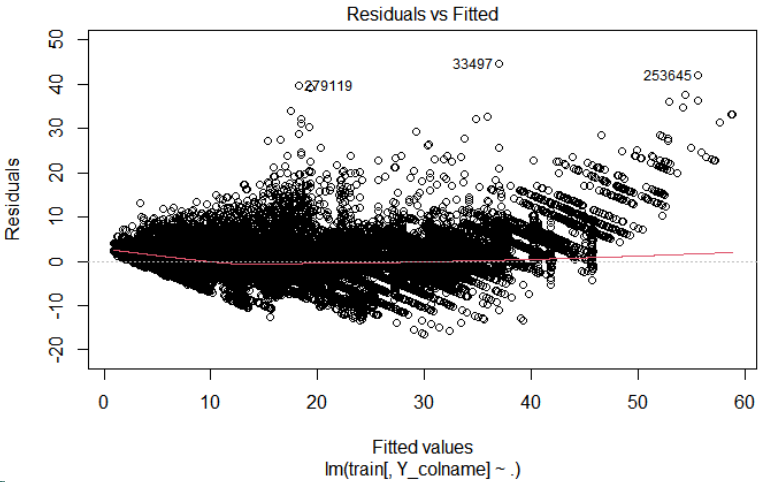
❖ Model Building

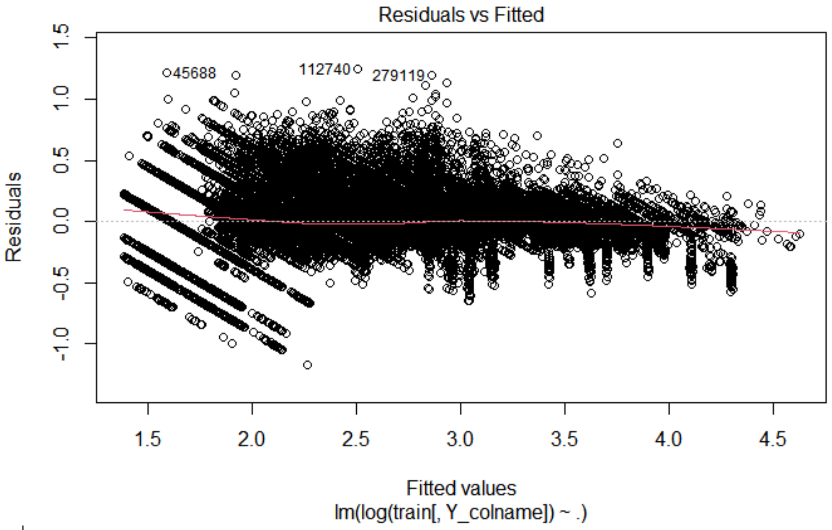
- 293,877 rows are taken and randomly split into training(60%), test(20%), and validation(20%) sets which are used to train, validate and test the model.
- Before building a linear model 4 assumptions are needed to be checked.
- The checking of assumptions are done by training 3 models below:
- 1) linear model
- 2) linear model with log applied
- 3) linear model with sqrt appled

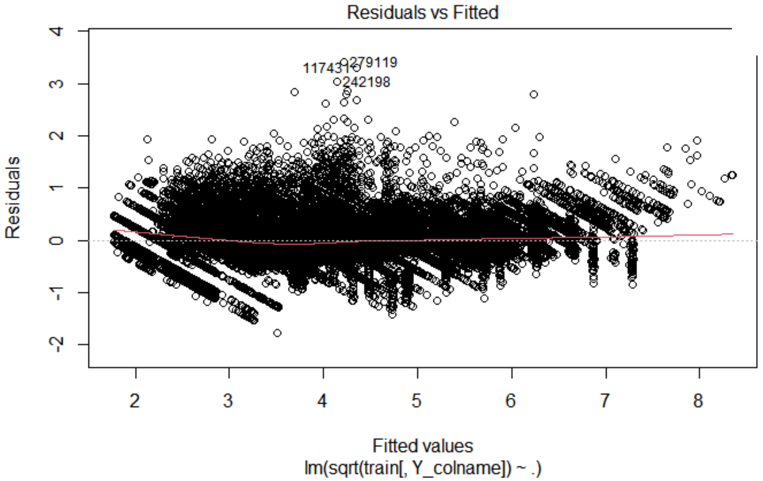
There are 4 assumptions considered for linear regression are :

1) There is a linear relationship between the predictors (x) and the outcome (y).

2) Residual Errors have a mean value of zero.







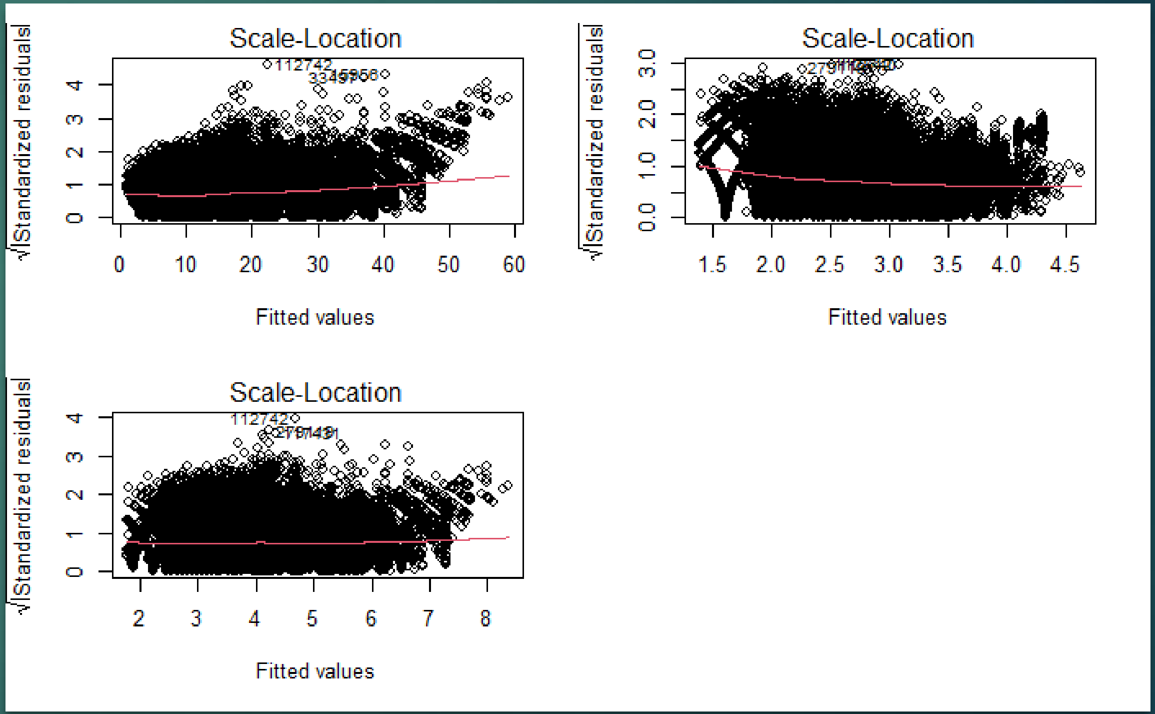
3) Predictors (x) are independent and observed with negligible error

- We use Durbin Watson test in which null hypothesis of the test states that there is no auto-correlation of residuals.
- Implicitly, our target has enough evidence and fail to reject H0 hypotheses.
- Therefore all 3 models which we assumed have no auto correlation among their predicted variables.

```
lag Autocorrelation D-W Statistic p-value
  1     -0.00428263      2.008562   0.082
Alternative hypothesis: rho != 0
lag Autocorrelation D-W Statistic p-value
  1    -0.002703031      2.005401   0.282
Alternative hypothesis: rho != 0
lag Autocorrelation D-W Statistic p-value
  1    -0.004471079      2.008937    0.07
Alternative hypothesis: rho != 0
```

4) Residual errors have constant variance

- The red line is roughly horizontal across the plot.
- But, to check homoscedasticity we use Breusch-Pagan Test since it is not clear from the red line that we have constant variance.



Breusch-Pagan Test

- From the output we can see that the p-value of the test is less than 0.05, we reject the null hypothesis.
- We have sufficient evidence to say that heteroscedasticity is present in the regression model which means there may be some non constant variance which is not desirable.

```
        studentized Breusch-Pagan test

data:  assumption_test_model
BP = 21247, df = 84, p-value < 2.2e-16


        studentized Breusch-Pagan test

data:  assumption_test_model_log
BP = 23192, df = 84, p-value < 2.2e-16


        studentized Breusch-Pagan test

data:  assumption_test_model_sqrt
BP = 10260, df = 84, p-value < 2.2e-16
```

- Several assumptions are satisfied by 3 linear models.
- Now, we experiment with the following models by training them:

  - Full linear model

  - Poisson GLM (log transform of target variable)

  - Backward and forward coefficients selection is done and best models, based on Bayesian Information Criterion and Mallow's Cp coefficient are trained.
- After training the models, we compare and select the best in terms of:

  - Adjuster R squared

  - Akaike Information Criterion

  - Bayesian Information Criterion

  - Number of parameters

  - Validation R squared

1)Full linear model

- The adjusted R squared for the full linear model is 0.9247 with p-value less than 0.05
- With other techniques, we will try to improve these metrics while decreasing the number of parameters considered.

```
Call:
lm(formula = train[, Y_colname] ~ ., data = train[, X_colnames])

Residuals:
    Min      1Q  Median      3Q     Max
-16.542  -1.425  -0.154   1.263  53.568

Coefficients: (17 not defined because of singularities)
                              Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 -2.323e+04  3.250e+04   -0.715 0.474705
timestamp                   -6.150e-06  5.787e-06   -1.063 0.287957
hour                         2.433e-02  2.097e-02    1.160 0.245893
day                          5.604e-01  5.015e-01    1.117 0.263853
month                        1.685e+01  1.504e+01    1.120 0.262759
distance                     2.901e+00  6.873e-03  422.095  < 2e-16 ***
surge_multiplier             6.832e+01  2.340e-01  291.914  < 2e-16 ***
V1_Back.Bay.1               -7.017e-02  2.928e-02   -2.396 0.016555 *
V1_Beacon.Hill.1            -3.747e-01  2.928e-02  -12.795  < 2e-16 ***
V1_Boston.University.1      -5.070e-01  3.051e-02  -16.617  < 2e-16 ***
V1_Fenway.1                 -2.971e-01  2.989e-02   -9.943  < 2e-16 ***
V1_Financial.District.1      3.079e-01  2.946e-02   10.449  < 2e-16 ***
V1_Haymarket.Square.1        1.949e-01  2.970e-02    6.564 5.26e-11 ***
V1_North.End.1               3.581e-01  2.936e-02   12.198  < 2e-16 ***
V1_North.Station.1           1.799e-04  2.917e-02    0.006 0.995079
V1_Northeastern.University.1 -5.122e-01  2.984e-02  -17.163  < 2e-16 ***
V1_South.Station.1                  NA         NA       NA       NA
V1_Theatre.District.1        4.572e-01  2.934e-02   15.581  < 2e-16 ***
V1_West.End.1                       NA         NA       NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.506 on 176241 degrees of freedom
Multiple R-squared:  0.9281,    Adjusted R-squared:  0.9281
F-statistic: 2.71e+04 on 84 and 176241 DF,  p-value: < 2.2e-16
```

2)Poisson GLM

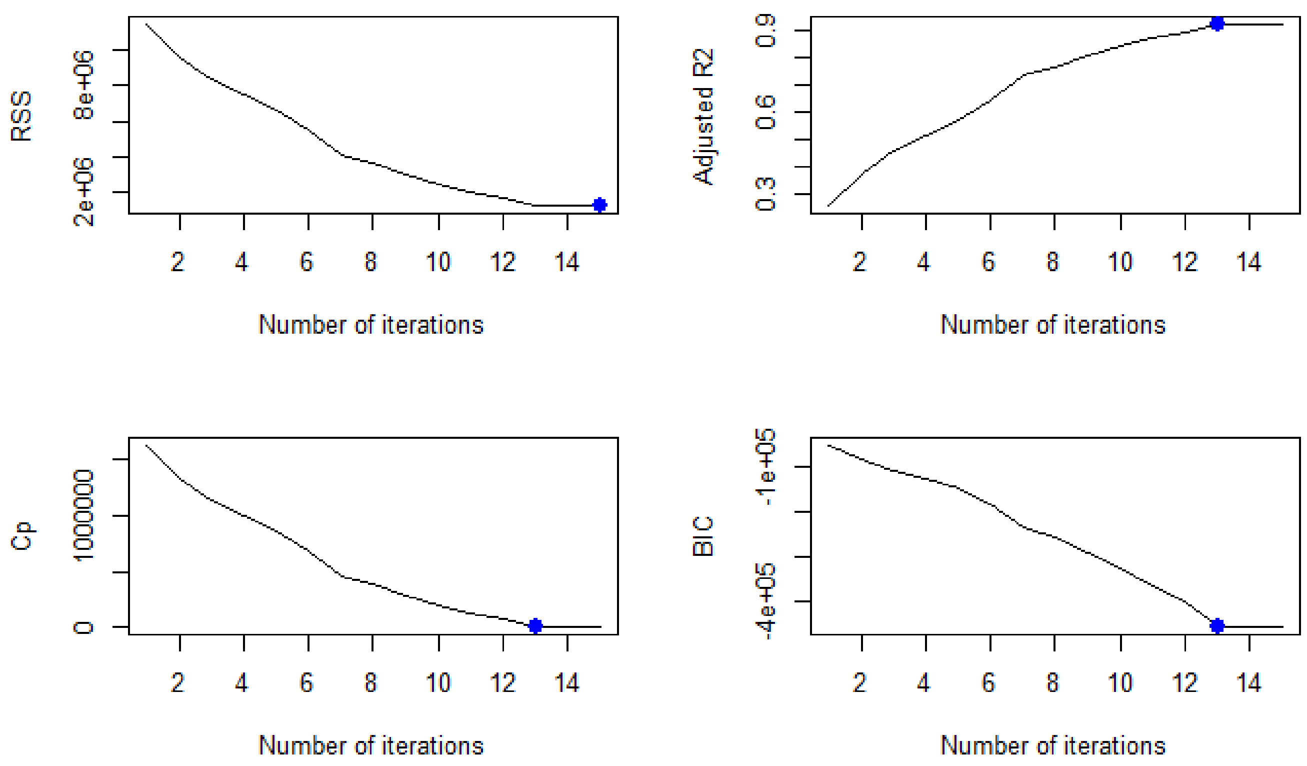
- Log transformation is used to increase R squared up to 0.9386 from 0.9247.
- To improve these metrics while decreasing the number of parameters considered forward and backward selection are used.

```
Call:
lm(formula = log(train[, Y_colname]) ~ ., data = train[, X_colnames])

Residuals:
     Min       1Q   Median       3Q      Max 
-1.16268 -0.07493 -0.00433  0.06844  1.24677 

Coefficients: (17 not defined because of singularities)
                              Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                 -2.897e+03  1.827e+03   -1.586 0.112799    
timestamp                   -3.978e-07  3.253e-07   -1.223 0.221367    
hour                         1.421e-03  1.179e-03    1.206 0.227964    
day                          3.479e-02  2.819e-02    1.234 0.217062    
month                        1.046e+00  8.456e-01    1.237 0.216184    
distance                     1.754e-01  3.863e-04  454.001  < 2e-16 ***
surge_multiplier             2.597e+00  1.315e-02  197.415  < 2e-16 ***
latitude                     4.609e-01  2.609e-01    1.766 0.077330 .  
longitude                   -6.897e-01  3.658e-01   -1.886 0.059343 .  

V1_.Possible.Drizzle.               NA         NA       NA       NA    
V1_Back.Bay.1               -2.518e-03  1.646e-03   -1.530 0.125977    
V1_Beacon.Hill.1            -5.183e-03  1.646e-03   -3.149 0.001639 ** 
V1_Boston.University.1      -4.110e-02  1.715e-03  -23.967  < 2e-16 ***
V1_Fenway.1                 -2.122e-02  1.680e-03  -12.635  < 2e-16 ***
V1_Financial.District.1     -3.426e-02  1.656e-03  -20.689  < 2e-16 ***
V1_Haymarket.Square.1       -1.514e-02  1.669e-03   -9.072  < 2e-16 ***
V1_North.End.1               1.751e-02  1.650e-03   10.611  < 2e-16 ***
V1_North.Station.1          -7.108e-03  1.639e-03   -4.336 1.45e-05 ***
V1_Northeastern.University.1 -2.905e-02 1.677e-03  -17.317  < 2e-16 ***
V1_South.Station.1                  NA         NA       NA       NA    
V1_Theatre.District.1        2.784e-02  1.649e-03   16.878  < 2e-16 ***
V1_West.End.1                       NA         NA       NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1408 on 176241 degrees of freedom
Multiple R-squared:  0.9387,	Adjusted R-squared:  0.9386 
F-statistic: 3.211e+04 on 84 and 176241 DF,  p-value: < 2.2e-16
```

3) Linear model (forward / backward model selection)

- Many correlated indices were removed by trial & error method and nvmax is chosen 15 for both forward and backward subset selection.

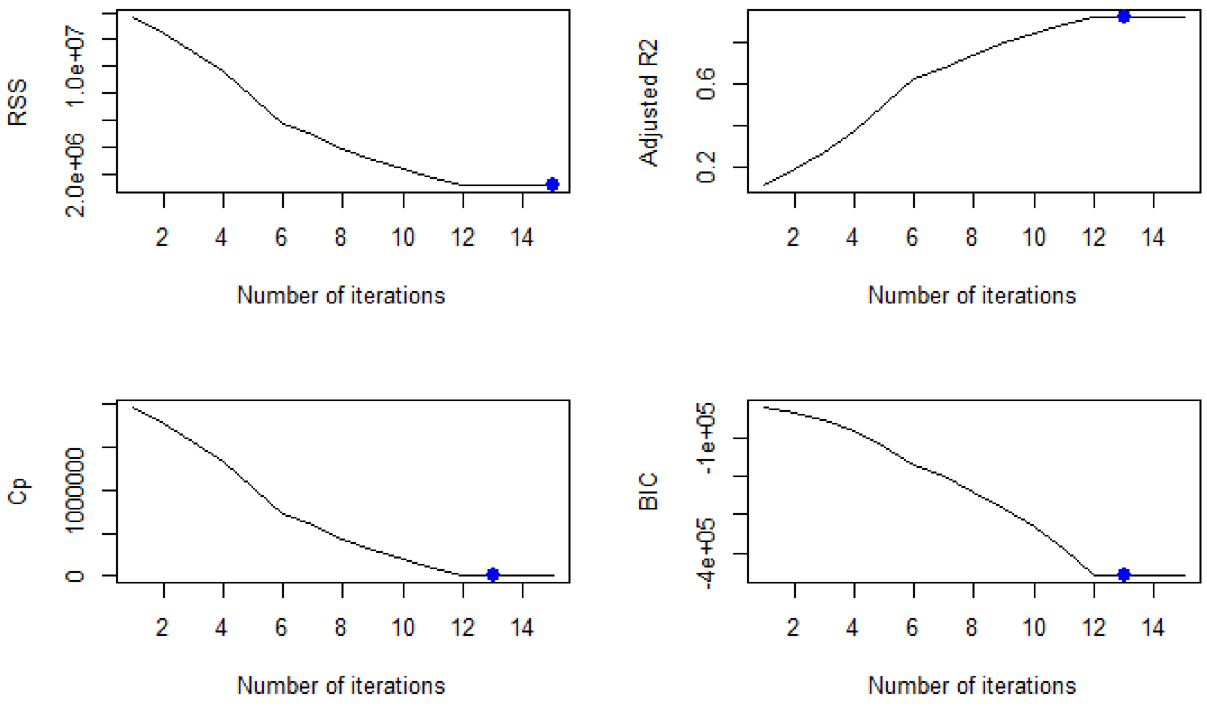
Forward subsets



3) Linear model (forward / backward model selection)

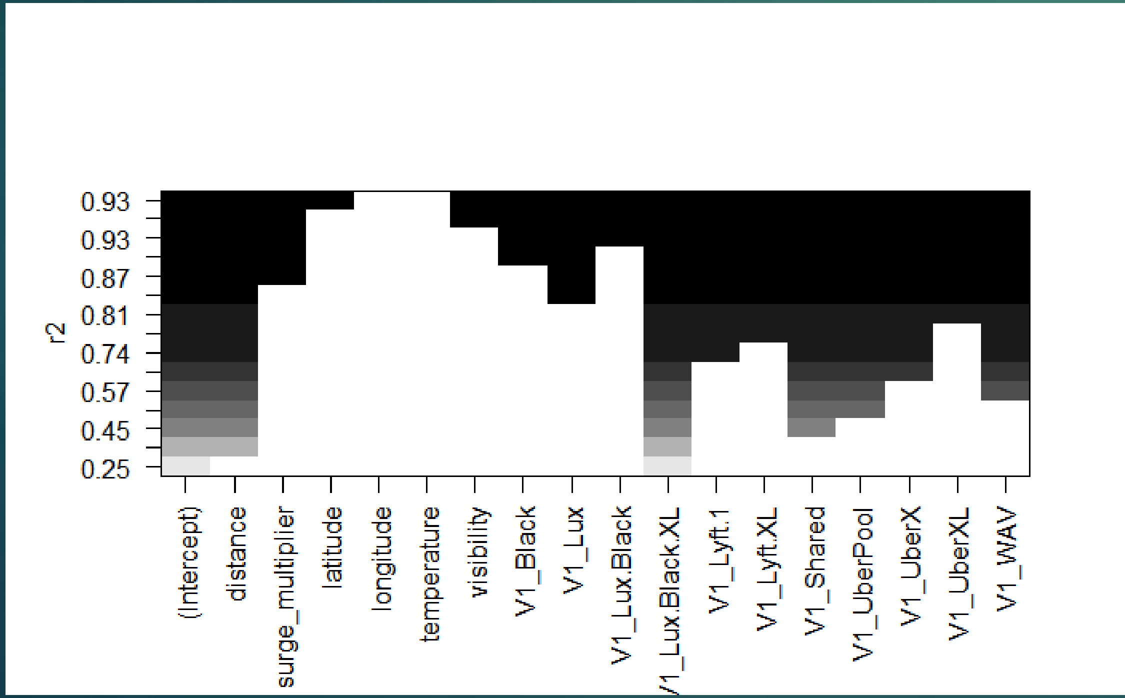
- Best number of parameters in Backward selection and Forward selection are same with
- BIC →13
- Cp →13

Backward subsets

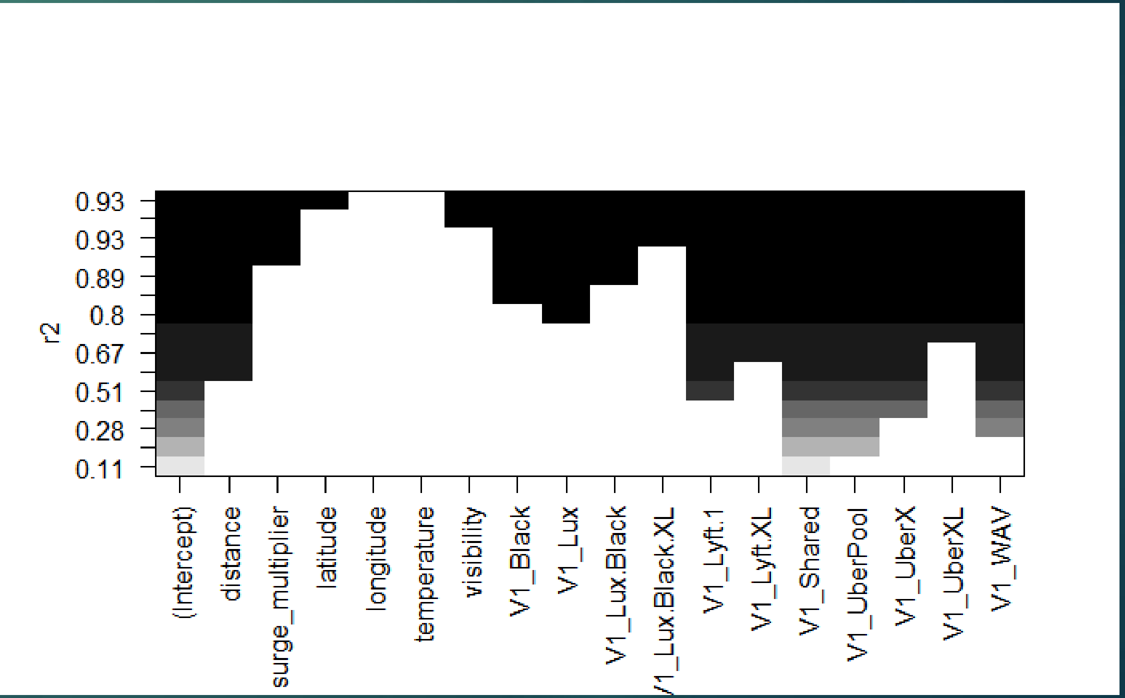


- To represent the most significant variables, we used variable elimination plots based on r^2 value.
- After elimination of variables, best BIC and Cp model are retrained.

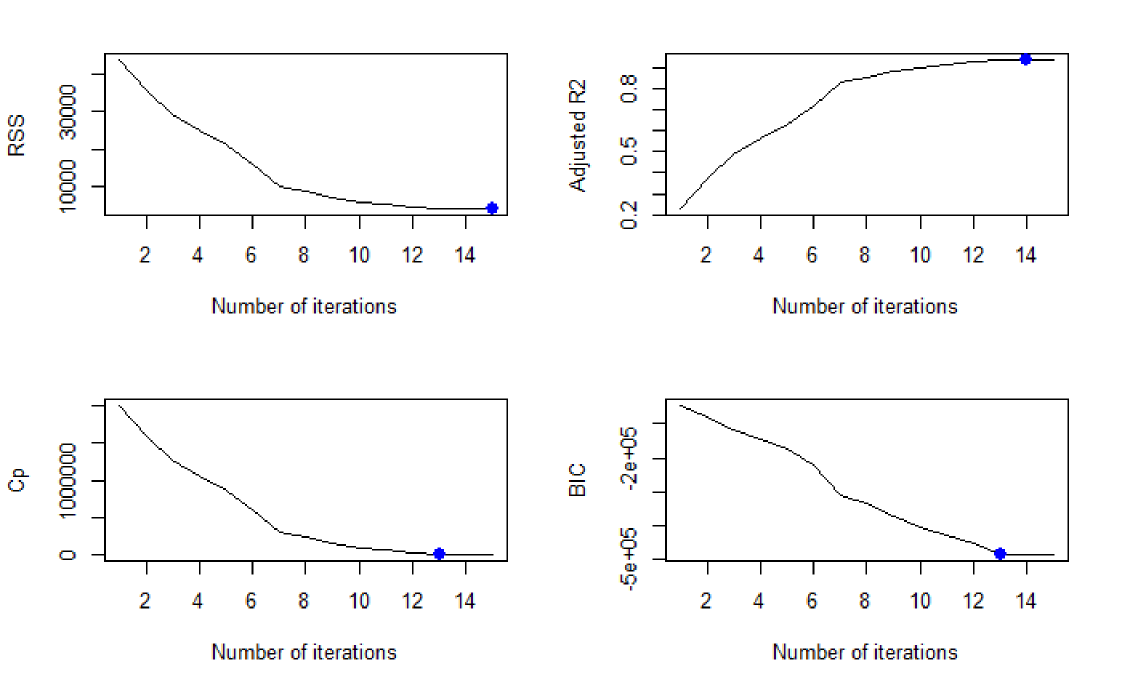
forward



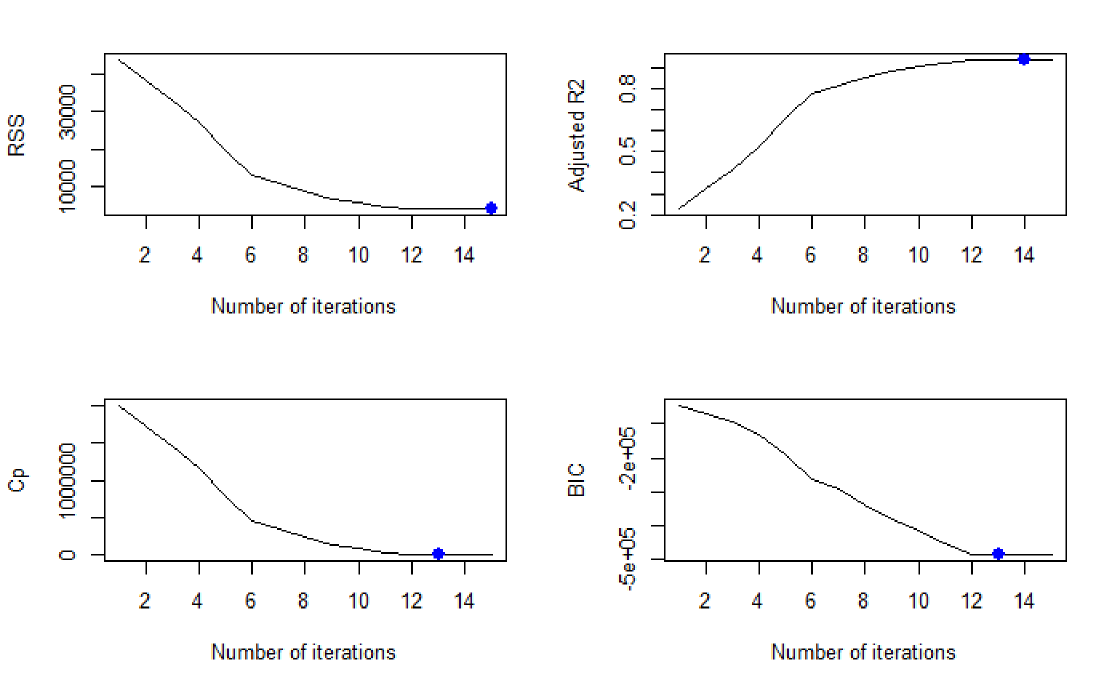
backward



4) Forward / backward selection for log(Y) transform



Forward subsets

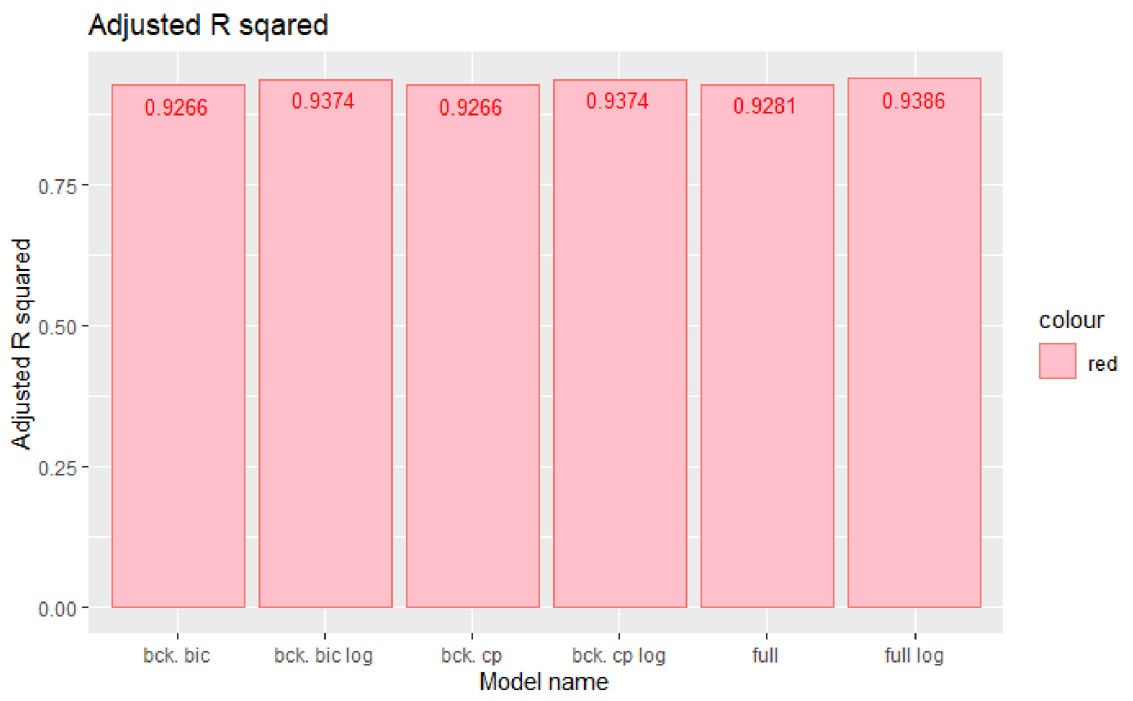


Backward subsets

Number of best coefficients for both selections are BIC→13 & Cp→13

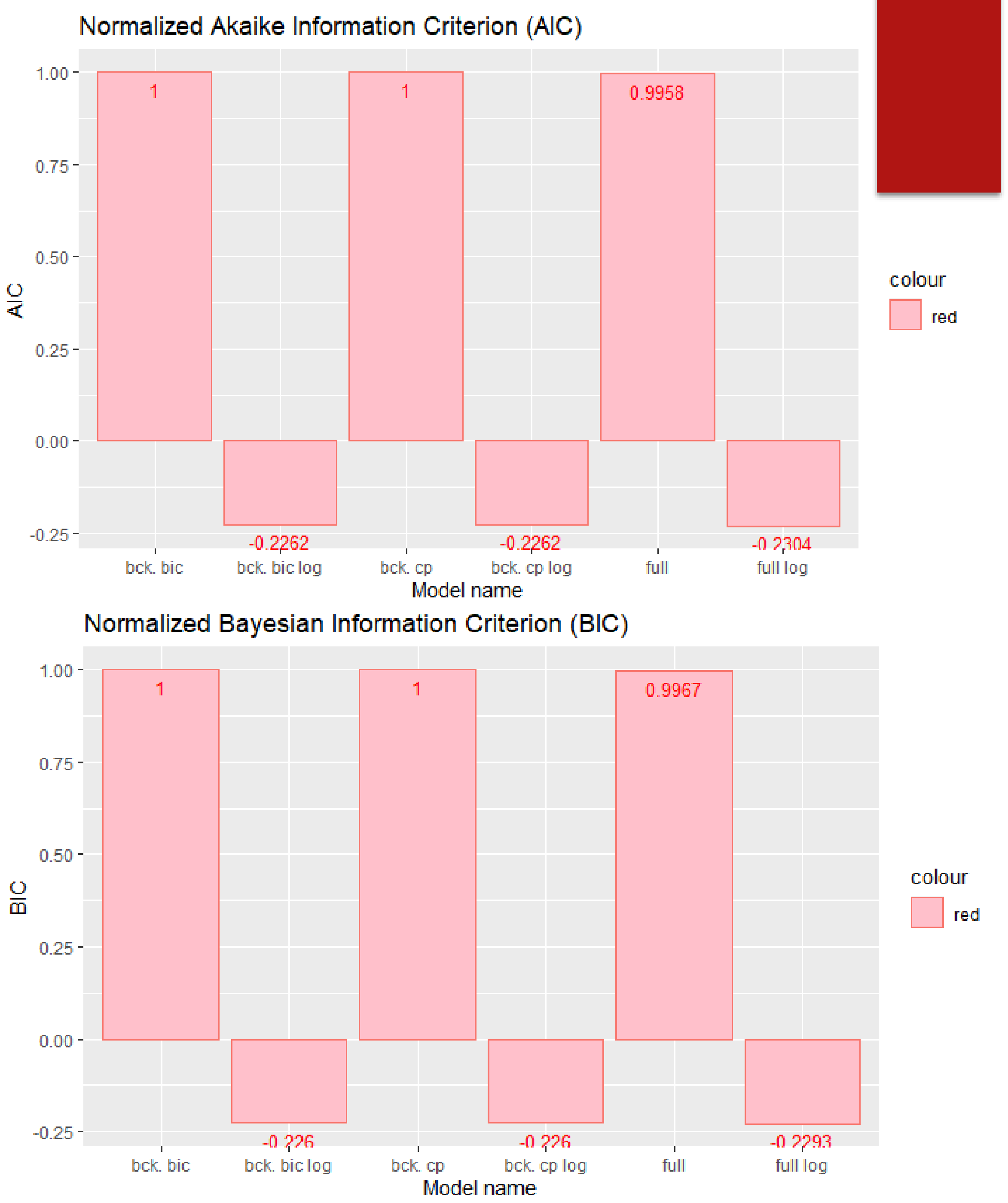
❖ Comparison of models

- **Comparison of adjusted R^2**
- From the plot we can say that independent of parameter selection technique generalized linear model was capable of achieving better R^2 results.



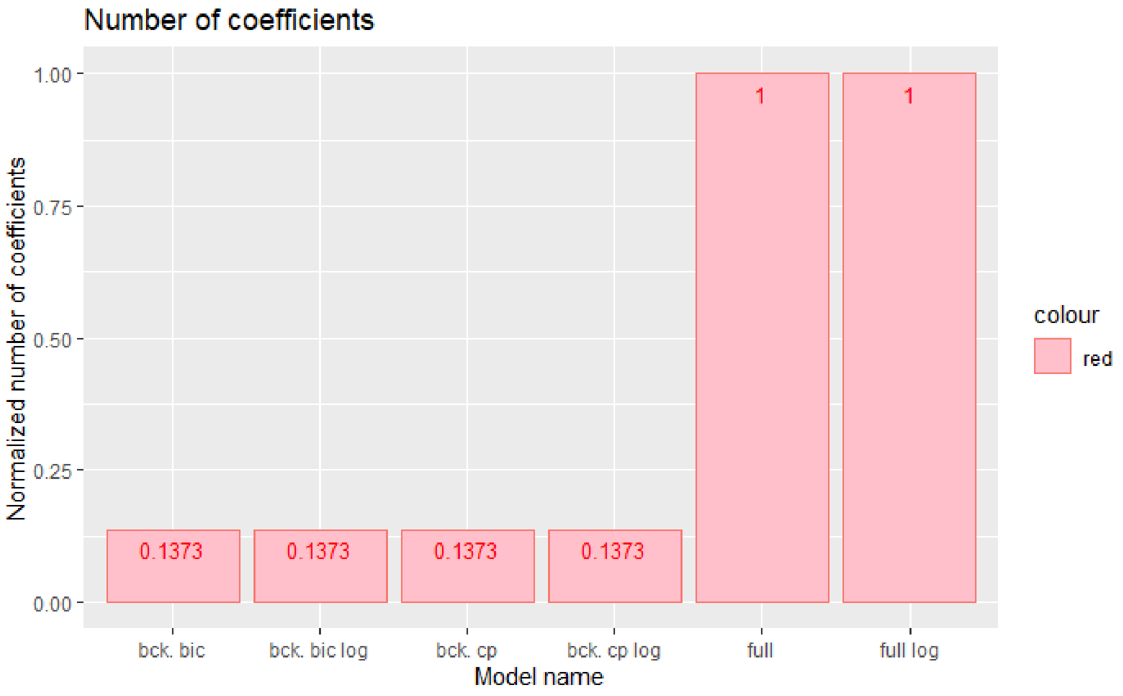
2)Comparison of AIC and BIC

- A remarkable difference is also observed in terms of Akaike and Bayesian information criteria
- The lower the AIC and BIC the better, and a negative AIC or BIC indicates a lower degree of information loss than a positive AIC or BIC



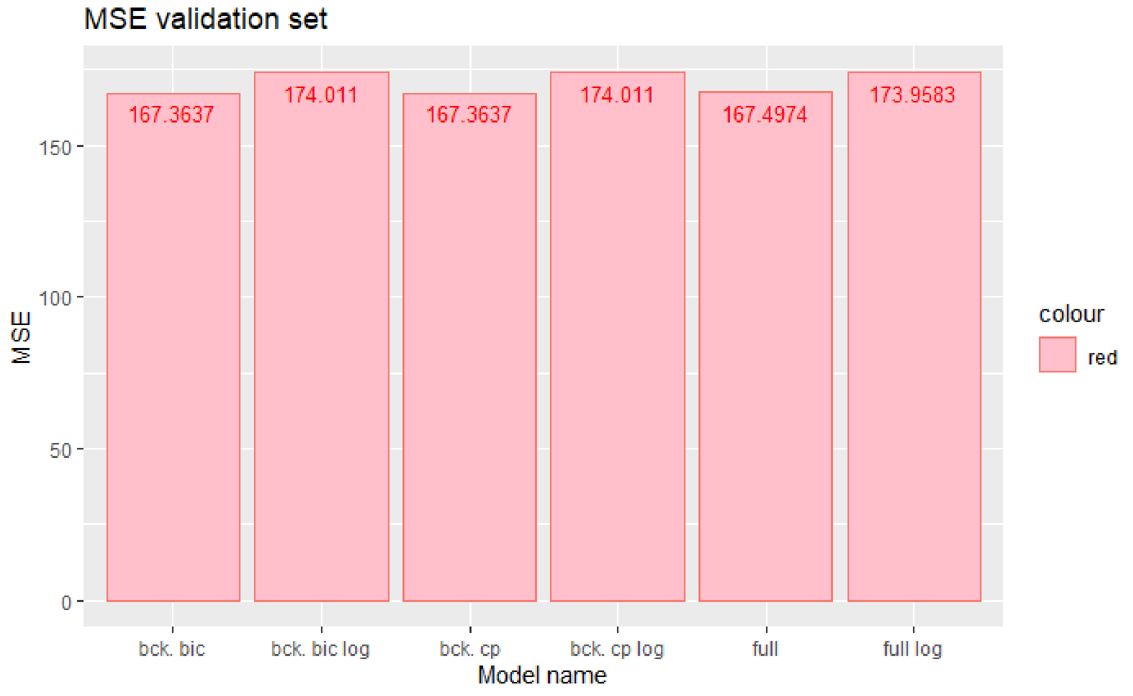
4)Comparison based on number of parameters

- A powerful result can be made here - having up to 87 percent less parameters, the lightweight models like GLM with backward selection and GLM log model with backward selection were able to achieve better R squared metrics with less parameters.



5)Comparison Based on Validation Set.

- On the validation set, all models achieved comparable results.
- **Conclusion:**
- The best model is a Poisson GLM, with backward coefficients selection.
- It has 87% less parameters than full model, allowing for around 0.9386 adjusted R squared.



Thank you