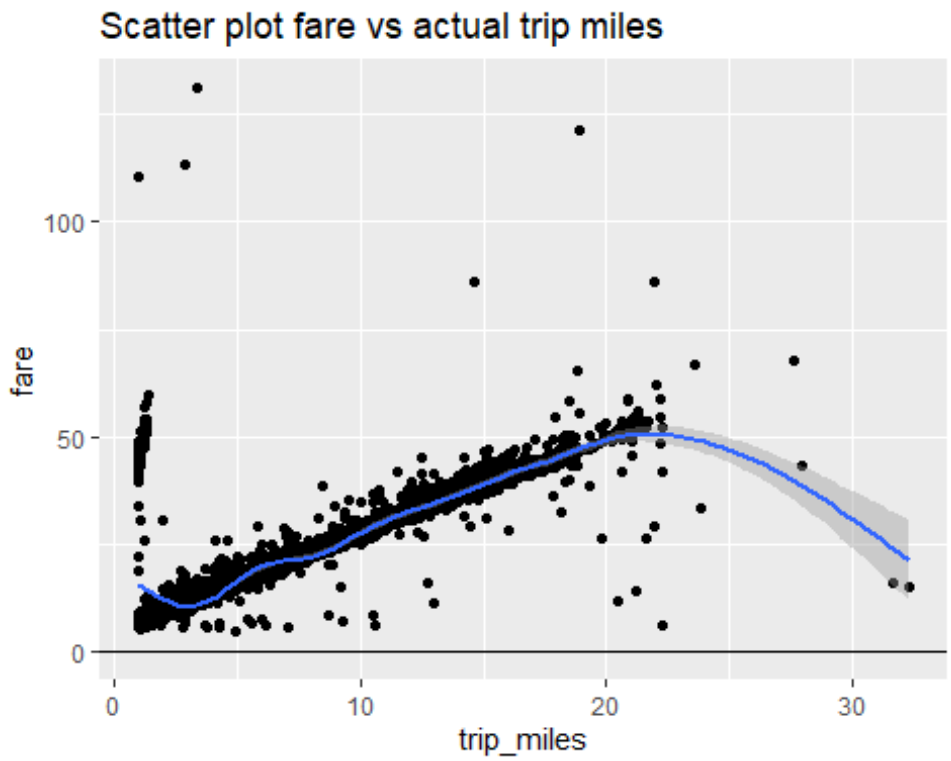
Data 603- Project

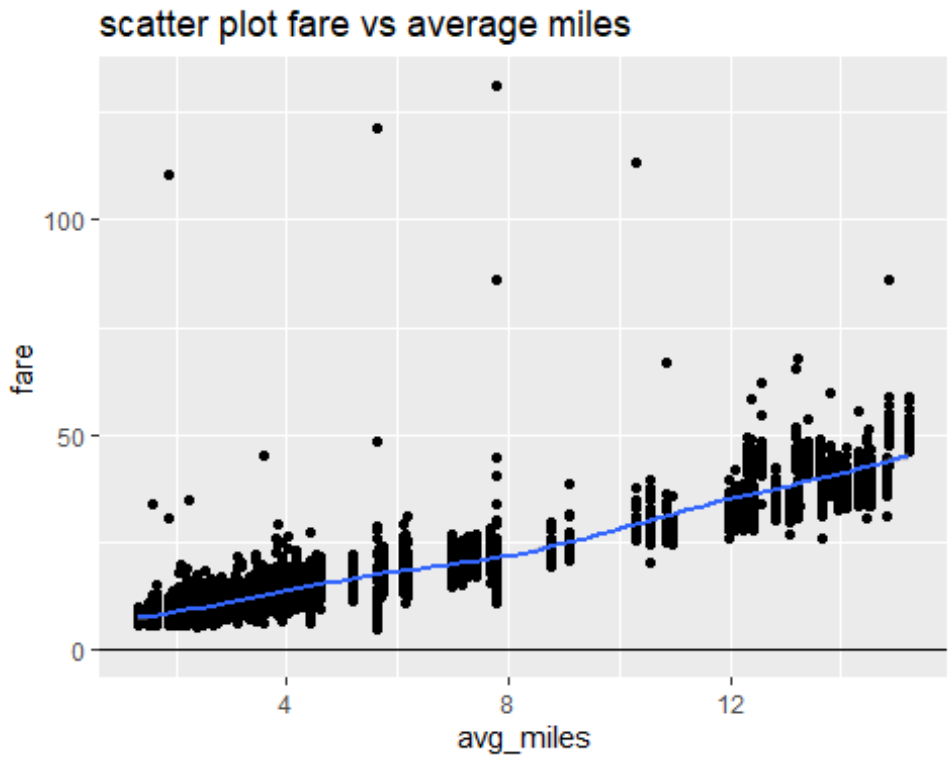
Shora Dehkordi, Ryan leeson, Guarav Kumar, Maryam Sarafraz

03/12/2019

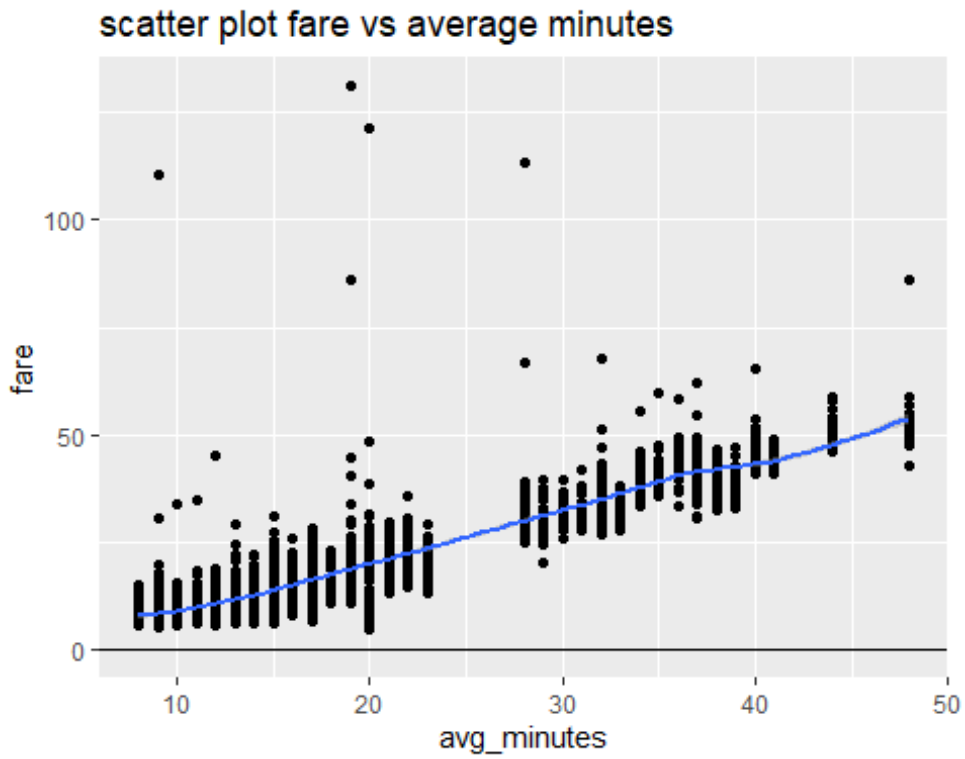
```
taxi_data = read.csv ("./taxitrip_sample_df_final.csv")
#head (taxi_data, 4)
#tail (taxi_data, 7)

#   convert day_of_week to a numerical value
transform (taxi_data, day_of_week = as.numeric (day_of_week))

#   Filter for weekend
#   Sunday = 1
#   Saturday = 7
taxi_data$weekend = 1
taxi_data$weekend[ taxi_data$day_of_week > 1 & taxi_data$day_of_week < 6] = 0

#   convert months to a numerical value
transform (taxi_data, months = as.numeric (months))

#   Filtering for season
taxi_data$season = "Winter"   #   Winter
taxi_data$season[taxi_data$months > 2 & taxi_data$months < 6] = "Spring"   #
Spring
taxi_data$season[taxi_data$months > 5 & taxi_data$months < 9] = "Summer"   #
Summer
taxi_data$season[taxi_data$months > 8 & taxi_data$months < 12] = "Fall"   #
Fall

transform (taxi_data, hours = as.numeric (hours))

taxi_data$time_of_day = "Night"   #   Night
taxi_data$time_of_day[taxi_data$hours >= 6 & taxi_data$hours < 12] =
"Morning"   #   Morning
taxi_data$time_of_day[taxi_data$hours >= 12 & taxi_data$hours < 18] =
"Afternoon"   #   Afternoon
taxi_data$time_of_day[taxi_data$hours >= 18 & taxi_data$hours < 24] =
"Evening"   #   Evening

transform (taxi_data, season = factor (season), weekend = factor (weekend),
time_of_day = factor (time_of_day))

ggplot(taxi_data, aes(x=trip_miles, y=fare)) +
  geom_point() + geom_smooth()+
  geom_hline(yintercept = 0) + ggtitle("Scatter plot fare vs actual trip
miles")

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



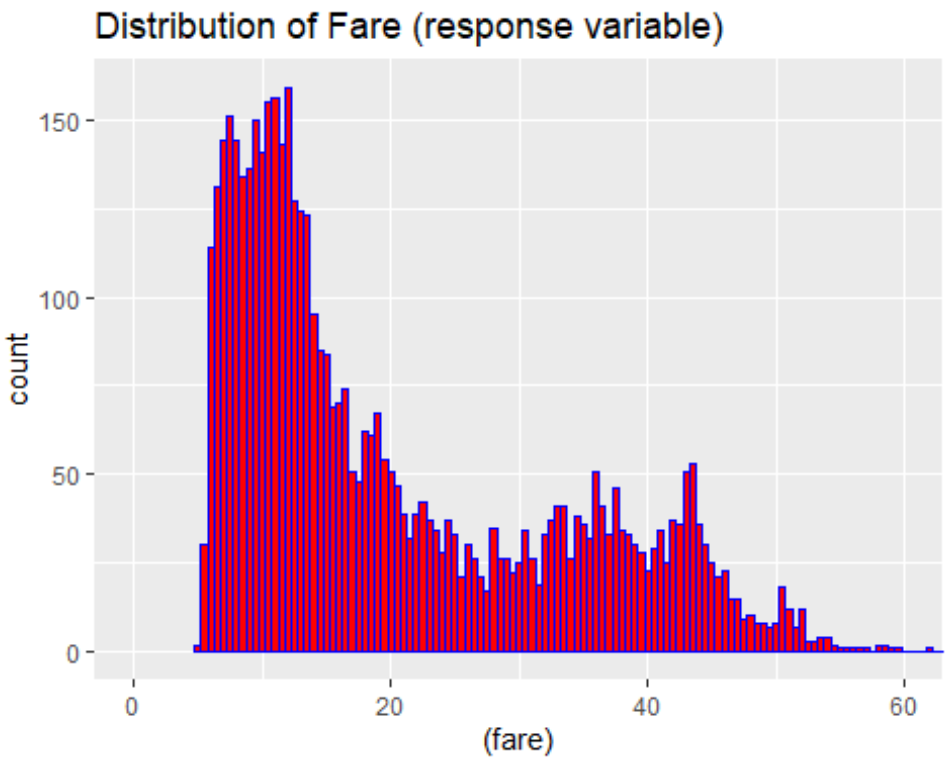

```
# we can see some outliers but in general we have a good correlation

ggplot(taxi_data, aes(x=avg_miles, y=fare)) +
  geom_point() + geom_smooth()+
  geom_hline(yintercept = 0) + ggtitle("scatter plot fare vs average miles")

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



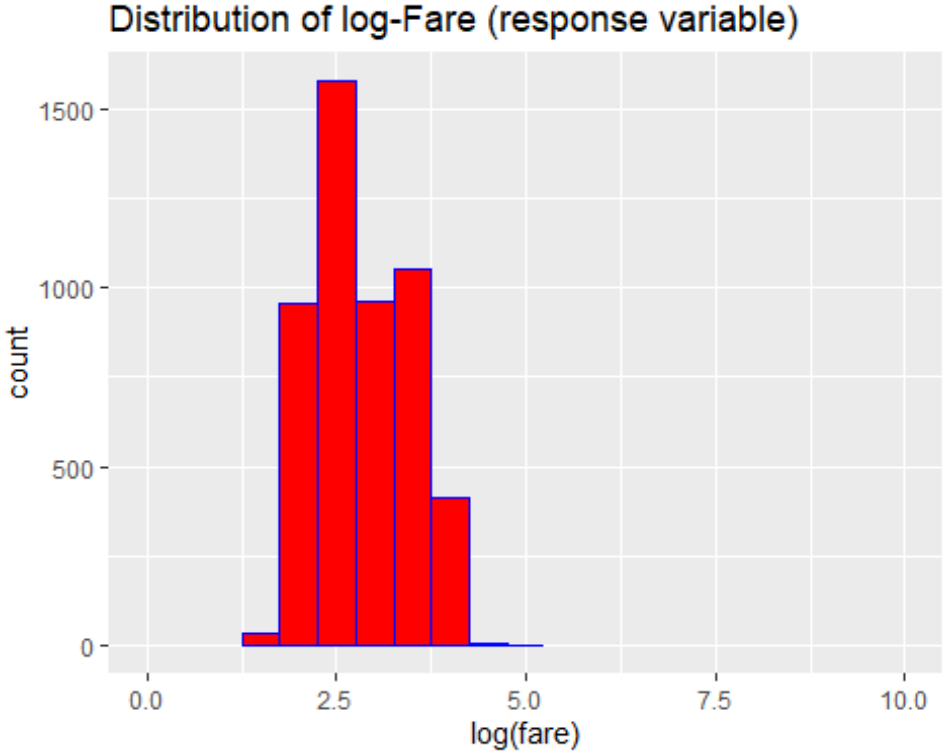

```
# we can see some outliers but in general we have a good correlation

ggplot(taxi_data, aes(x=avg_minutes, y=fare)) +
  geom_point() + geom_smooth()+
  geom_hline(yintercept = 0) + ggtitle("scatter plot fare vs average
minutes")

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
# we can see some outliers but in general we have a good correlation

ggplot(data= taxi_data , aes(x= (fare) ) )+ geom_histogram(col= 'blue' ,
fill='red',binwidth=0.5)+ coord_cartesian(xlim = c(0, 60))
+ggtitle("Distribution of Fare (response variable)")
```




```
ggplot(data= taxi_data , aes(x= log(fare) ) )+ geom_histogram(col= 'blue' ,
fill='red',binwidth=0.5)+ coord_cartesian(xlim = c(0, 10))
+ggtitle("Distribution of log-Fare (response variable)")
```



###Full Linear model

```
taxi_fulllm = lm ( fare ~ factor(payment_type) + factor(company) + avg_miles
+ avg_minutes + factor(time_of_day) + factor(season) + factor(weekend) +
factor(hour_type), data = taxi_data)
```

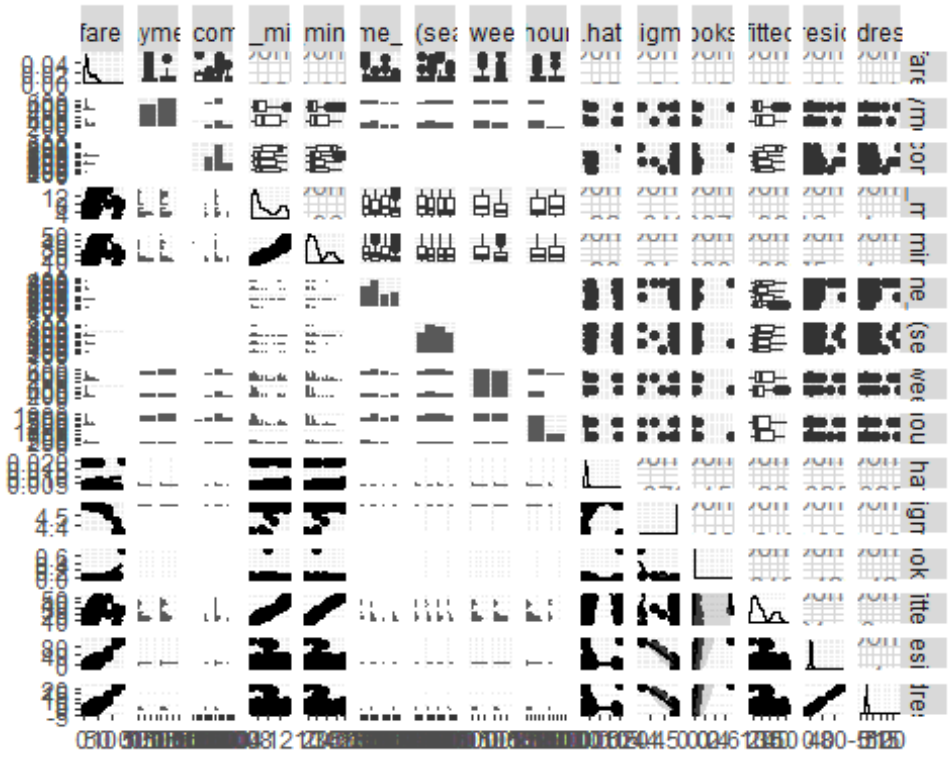
#####Check multi colinearity

```
vif (taxi_fulllm)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## factor(payment_type)   1.065684  1        1.032320
## factor(company)        1.033313  3        1.005477
## avg_miles             13.801971  1        3.715100
## avg_minutes           13.862480  1        3.723235
## factor(time_of_day)    1.221832  3        1.033956
## factor(season)         1.017771  3        1.002940
## factor(weekend)        1.081302  1        1.039857
## factor(hour_type)      1.113469  1        1.055211
```

avg_miles and avg_minutes have colinearity, so, avg_minutes will be removed from the model

```
ggpairs (taxi_fulllm, lower = list ( continuous = "smooth_loess", combo =
"facethist", discrete = "facetbar", na = "na"), cardinality_threshold = 25)
```



Model varaible testing

```
taxi_fulllm_new = lm ( fare ~ factor(payment_type) + factor(company) +
avg_miles + factor(time_of_day) + factor(season) + factor(weekend) +
factor(hour_type), data = taxi_data)
```

stepwise regression

```
taxi_stepw = ols_step_both_p ( taxi_fulllm_new, pent = 0.05, prem = 0.1,
details = FALSE)
```

```
## Stepwise Selection Method
## --------------------------
##
## Candidate Terms:
##
## 1. factor(payment_type)
## 2. factor(company)
## 3. avg_miles
## 4. factor(time_of_day)
## 5. factor(season)
## 6. factor(weekend)
## 7. factor(hour_type)
##
## We are selecting variables based on p value...
##
## Variables Entered/Removed:
##
## - avg_miles added
## - factor(company) added
## - factor(time_of_day) added
## - factor(hour_type) added
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.926       RMSE                 4.954
## R-Squared               0.858       Coef. Var           24.193
## Adj. R-Squared          0.858       MSE                 24.540
## Pred R-Squared          0.857       MAE                  2.981
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## --------------------------------------------------------------------------
```

```
##                    Sum of
##                   Squares          DF    Mean Square       F          Sig.
## ---------------------------------------------------------------------------
## Regression    739081.451           8      92385.181    3764.658    0.0000
## Residual      122479.776        4991         24.540
## Total         861561.227        4999
## ---------------------------------------------------------------------------
##
##                                          Parameter Estimates
## ---------------------------------------------------------------------------
------------------------------
##                        model      Beta    Std. Error    Std. Beta       t
Sig      lower     upper
## ---------------------------------------------------------------------------
------------------------------
##                  (Intercept)     8.601         0.689                 12.477
0.000     7.250     9.953
##                    avg_miles     2.644         0.016        0.914    167.345
0.000     2.613     2.675
##          factor(company)101    -5.561         0.673       -0.189     -8.264
0.000    -6.880    -4.242
##          factor(company)107    -4.857         0.666       -0.185     -7.292
0.000    -6.163    -3.551
##          factor(company)109    -5.292         0.678       -0.157     -7.803
0.000    -6.621    -3.962
## factor(time_of_day)Evening    -0.966         0.181       -0.035     -5.334
0.000    -1.321    -0.611
## factor(time_of_day)Morning    -0.570         0.214       -0.017     -2.663
0.008    -0.990    -0.150
##   factor(time_of_day)Night    -0.928         0.226       -0.027     -4.111
0.000    -1.371    -0.485
## factor(hour_type)rush_hour     0.803         0.179        0.025      4.480
0.000     0.452     1.155
## ---------------------------------------------------------------------------
------------------------------
```

avg_miles, company, hour_type, and time_of_day are suggested for the model

```
taxi_formodel = ols_step_forward_p ( taxi_fulllm_new, pent = 0.05, details =
FALSE)
```

```
## Forward Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. factor(payment_type)
## 2. factor(company)
## 3. avg_miles
## 4. factor(time_of_day)
```

```
## 5. factor(season)
## 6. factor(weekend)
## 7. factor(hour_type)
##
## We are selecting variables based on p value...
##
## Variables Entered:
##
## - avg_miles
## - factor(company)
## - factor(time_of_day)
## - factor(hour_type)
##
## No more variables to be added.
##
## Final Model Output
## ------------------
##
##                          Model Summary
## ----------------------------------------------------------------
## R                        0.926       RMSE                 4.954
## R-Squared                0.858       Coef. Var           24.193
## Adj. R-Squared           0.858       MSE                 24.540
## Pred R-Squared           0.857       MAE                  2.981
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## -------------------------------------------------------------------------
##                   Sum of
##                  Squares          DF    Mean Square       F          Sig.
## -------------------------------------------------------------------------
## Regression    739081.451           8      92385.181    3764.658    0.0000
## Residual      122479.776        4991         24.540
## Total         861561.227        4999
## -------------------------------------------------------------------------
##
##                                          Parameter Estimates
## ----------------------------------------------------------------------------
------------------------------
##                       model      Beta    Std. Error    Std. Beta       t
Sig      lower     upper
## ----------------------------------------------------------------------------
------------------------------
##                 (Intercept)     8.601         0.689                 12.477
0.000     7.250     9.953
##                   avg_miles     2.644         0.016        0.914    167.345
0.000     2.613     2.675
```

```
##         factor(company)101    -5.561         0.673       -0.189     -8.264
0.000    -6.880    -4.242
##         factor(company)107    -4.857         0.666       -0.185     -7.292
0.000    -6.163    -3.551
##         factor(company)109    -5.292         0.678       -0.157     -7.803
0.000    -6.621    -3.962
## factor(time_of_day)Evening    -0.966         0.181       -0.035     -5.334
0.000    -1.321    -0.611
## factor(time_of_day)Morning    -0.570         0.214       -0.017     -2.663
0.008    -0.990    -0.150
##   factor(time_of_day)Night    -0.928         0.226       -0.027     -4.111
0.000    -1.371    -0.485
## factor(hour_type)rush_hour     0.803         0.179        0.025      4.480
0.000     0.452     1.155
## ---------------------------------------------------------------------------
-----------------------------
```

avg_miles, company, hour_type, and time_of_day are suggested for the model

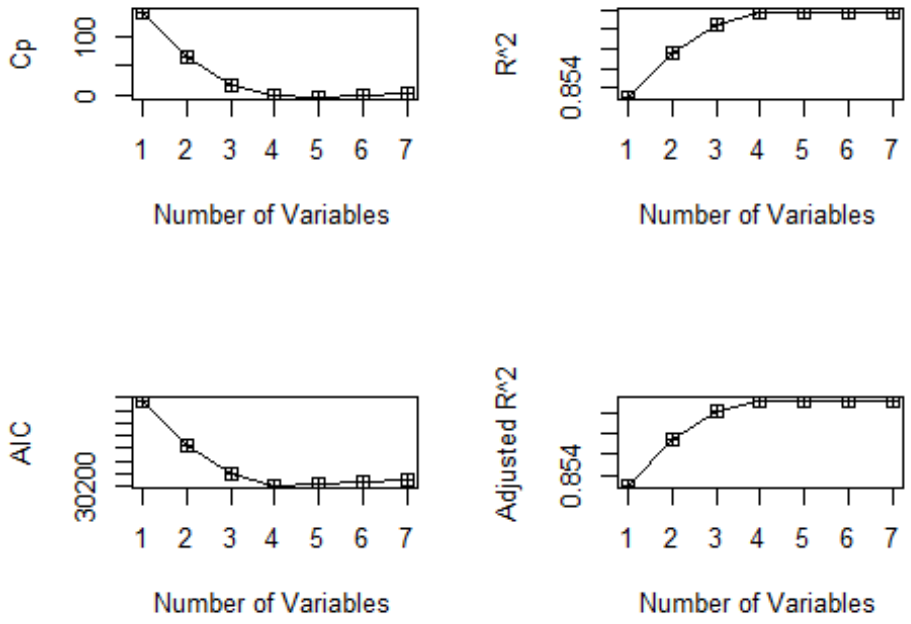
```
taxi_backmodel = ols_step_backward_p ( taxi_fulllm_new, prem = 0.05, details
= FALSE)

## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . factor(payment_type)
## 2 . factor(company)
## 3 . avg_miles
## 4 . factor(time_of_day)
## 5 . factor(season)
## 6 . factor(weekend)
## 7 . factor(hour_type)
##
## We are eliminating variables based on p value...
##
## Variables Removed:
##
## - factor(weekend)
## - factor(payment_type)
## - factor(season)
##
## No more variables satisfy the condition of p value = 0.05
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
```

```
## ----------------------------------------------------------------
## R                       0.926       RMSE                 4.954
## R-Squared               0.858       Coef. Var           24.193
## Adj. R-Squared          0.858       MSE                 24.540
## Pred R-Squared          0.857       MAE                  2.981
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## --------------------------------------------------------------------------
##                   Sum of
##                  Squares          DF    Mean Square       F          Sig.
## --------------------------------------------------------------------------
## Regression    739081.451           8      92385.181    3764.658    0.0000
## Residual      122479.776        4991         24.540
## Total         861561.227        4999
## --------------------------------------------------------------------------
##
##                                          Parameter Estimates
## ----------------------------------------------------------------------------
-----------------------------
##                       model      Beta    Std. Error    Std. Beta       t
Sig      lower     upper
## ----------------------------------------------------------------------------
-----------------------------
##                (Intercept)     8.601         0.689                 12.477
0.000     7.250     9.953
##         factor(company)101    -5.561         0.673       -0.189     -8.264
0.000    -6.880    -4.242
##         factor(company)107    -4.857         0.666       -0.185     -7.292
0.000    -6.163    -3.551
##         factor(company)109    -5.292         0.678       -0.157     -7.803
0.000    -6.621    -3.962
##                  avg_miles     2.644         0.016        0.914    167.345
0.000     2.613     2.675
## factor(time_of_day)Evening    -0.966         0.181       -0.035     -5.334
0.000    -1.321    -0.611
## factor(time_of_day)Morning    -0.570         0.214       -0.017     -2.663
0.008    -0.990    -0.150
##   factor(time_of_day)Night    -0.928         0.226       -0.027     -4.111
0.000    -1.371    -0.485
## factor(hour_type)rush_hour     0.803         0.179        0.025      4.480
0.000     0.452     1.155
## ----------------------------------------------------------------------------
-----------------------------
```

avg_miles, company, time_of_day and hour_type are suggested for the model.

```
##                                                          Best Subsets
Regression
## ------------------------------------------------------------------------
-------------------------------------------------------
## Model Index    Predictors
## ------------------------------------------------------------------------
-------------------------------------------------------
##      1         avg_miles
##      2         factor(company) avg_miles
##      3         factor(company) avg_miles factor(time_of_day)
##      4         factor(company) avg_miles factor(time_of_day)
factor(hour_type)
##      5         factor(company) avg_miles factor(time_of_day)
factor(season) factor(hour_type)
##      6         factor(payment_type) factor(company) avg_miles
factor(time_of_day) factor(season) factor(hour_type)
##      7         factor(payment_type) factor(company) avg_miles
factor(time_of_day) factor(season) factor(weekend) factor(hour_type)
## ------------------------------------------------------------------------
-------------------------------------------------------
##
##                                                        Subsets Regression
Summary
## ------------------------------------------------------------------------
-------------------------------------------------------------
##                         Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)         AIC
SBIC          SBC         MSEP        FPE       HSP       APC
## ------------------------------------------------------------------------
-------------------------------------------------------------
##   1        0.8536      0.8535      0.8534    142.9352    30336.0410
16146.5461    30355.5926    25.2527    25.2527    0.0051    0.1466
##   2        0.8558      0.8557      0.8551     64.8715    30263.5288
16070.0733    30302.6319    24.8791    24.8791    0.0050    0.1443
##   3        0.8573      0.8571      0.8564     16.9994    30219.9924
16022.5927    30278.6472    24.6536    24.6536    0.0049    0.1430
##   4        0.8578      0.8576      0.8568     -1.0635    30201.9254
16004.5622    30267.0973    24.5647    24.5647    0.0049    0.1424
##   5        0.8579      0.8576      0.8568     -1.8501    30205.1318
16003.7798    30289.8553    24.5707    24.5706    0.0049    0.1424
##   6        0.8579      0.8576      0.8567      0.0349    30207.0164
16005.6703    30298.2571    24.5799    24.5799    0.0049    0.1425
##   7        0.8579      0.8576      0.8567      2.0000    30208.9815
16007.6411    30306.7394    24.5896    24.5896    0.0049    0.1425
## ------------------------------------------------------------------------
-------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
```

```
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```




```
ks_stat2 = data.frame ( c(1, 2, 3, 4, 5, 6, 7), ks$cp, ks$aic, ks$adjr,
ks$rsq)
names (ks_stat2) = c( "Predictors", "CP", "AIC", "Adjusted R^2", "R^2")
ks_stat2
```

```
##   Predictors           CP      AIC Adjusted R^2       R^2
## 1          1 142.93518208 30336.04    0.8535362 0.8535655
## 2          2  64.87151821 30263.53    0.8557315 0.8558469
## 3          3  16.99937537 30219.99    0.8570679 0.8572680
## 4          4  -1.06347616 30201.93    0.8576119 0.8578397
## 5          5  -1.85009439 30205.13    0.8576058 0.8579191
## 6          6   0.03485776 30207.02    0.8575805 0.8579224
## 7          7   2.00000000 30208.98    0.8575530 0.8579234
```

Cp four variable model is best AIC four variable is the best ajd.rsq four variables is the best but five variables is very close

taxi_fulllm_new = lm (fare ~ factor(payment_type) + factor(company) + avg_miles + factor(time_of_day) + factor(season) + factor(weekend) + factor(hour_type), data = taxi_data)

```
best.subset = regsubsets ( fare ~ factor(payment_type) + factor(company) +
avg_miles + factor(time_of_day) + factor(season) + factor(weekend) +
```

```
factor(hour_type), data = taxi_data, nv = 10)
summary ( best.subset)

## Subset selection object
## Call: regsubsets.formula(fare ~ factor(payment_type) + factor(company) +
##     avg_miles + factor(time_of_day) + factor(season) + factor(weekend) +
##     factor(hour_type), data = taxi_data, nv = 10)
## 13 Variables  (and intercept)
##                                Forced in Forced out
## factor(payment_type)Credit Card     FALSE      FALSE
## factor(company)101                  FALSE      FALSE
## factor(company)107                  FALSE      FALSE
## factor(company)109                  FALSE      FALSE
## avg_miles                           FALSE      FALSE
## factor(time_of_day)Evening          FALSE      FALSE
## factor(time_of_day)Morning          FALSE      FALSE
## factor(time_of_day)Night            FALSE      FALSE
## factor(season)Spring                FALSE      FALSE
## factor(season)Summer                FALSE      FALSE
## factor(season)Winter                FALSE      FALSE
## factor(weekend)1                    FALSE      FALSE
## factor(hour_type)rush_hour          FALSE      FALSE
## 1 subsets of each size up to 10
## Selection Algorithm: exhaustive
##          factor(payment_type)Credit Card factor(company)101
## 1  ( 1 )  " "                             " "
## 2  ( 1 )  " "                             " "
## 3  ( 1 )  " "                             "*"
## 4  ( 1 )  " "                             "*"
## 5  ( 1 )  " "                             "*"
## 6  ( 1 )  " "                             "*"
## 7  ( 1 )  " "                             "*"
## 8  ( 1 )  " "                             "*"
## 9  ( 1 )  " "                             "*"
## 10  ( 1 ) " "                             "*"
##          factor(company)107 factor(company)109 avg_miles
## 1  ( 1 )  " "                " "                "*"
## 2  ( 1 )  " "                " "                "*"
## 3  ( 1 )  " "                " "                "*"
## 4  ( 1 )  "*"                "*"                "*"
## 5  ( 1 )  "*"                "*"                "*"
## 6  ( 1 )  "*"                "*"                "*"
## 7  ( 1 )  "*"                "*"                "*"
## 8  ( 1 )  "*"                "*"                "*"
## 9  ( 1 )  "*"                "*"                "*"
## 10  ( 1 ) "*"                "*"                "*"
##          factor(time_of_day)Evening factor(time_of_day)Morning
## 1  ( 1 )  " "                        " "
## 2  ( 1 )  " "                        " "
## 3  ( 1 )  " "                        " "
```

```
## 4  ( 1 )  " "                          " "
## 5  ( 1 )  " "                          " "
## 6  ( 1 )  "*"                          " "
## 7  ( 1 )  "*"                          " "
## 8  ( 1 )  "*"                          "*"
## 9  ( 1 )  "*"                          "*"
## 10  ( 1 ) "*"                          "*"
##           factor(time_of_day)Night factor(season)Spring
## 1  ( 1 )  " "                        " "
## 2  ( 1 )  " "                        " "
## 3  ( 1 )  " "                        " "
## 4  ( 1 )  " "                        " "
## 5  ( 1 )  " "                        " "
## 6  ( 1 )  " "                        " "
## 7  ( 1 )  "*"                        " "
## 8  ( 1 )  "*"                        " "
## 9  ( 1 )  "*"                        " "
## 10  ( 1 ) "*"                        "*"
##           factor(season)Summer factor(season)Winter factor(weekend)1
## 1  ( 1 )  " "                  " "                  " "
## 2  ( 1 )  " "                  " "                  " "
## 3  ( 1 )  " "                  " "                  " "
## 4  ( 1 )  " "                  " "                  " "
## 5  ( 1 )  " "                  " "                  " "
## 6  ( 1 )  " "                  " "                  " "
## 7  ( 1 )  " "                  " "                  " "
## 8  ( 1 )  " "                  " "                  " "
## 9  ( 1 )  " "                  "*"                  " "
## 10  ( 1 ) " "                  "*"                  " "
##           factor(hour_type)rush_hour
## 1  ( 1 )  " "
## 2  ( 1 )  "*"
## 3  ( 1 )  "*"
## 4  ( 1 )  " "
## 5  ( 1 )  "*"
## 6  ( 1 )  "*"
## 7  ( 1 )  "*"
## 8  ( 1 )  "*"
## 9  ( 1 )  "*"
## 10  ( 1 ) "*"
```

```
reg.summary = summary ( best.subset)
```

five variables : avg_miles, company, hour_type Four variables : company, avg_miles,

```
summary (taxi_fulllm_new)
```

```
##
## Call:
## lm(formula = fare ~ factor(payment_type) + factor(company) +
##     avg_miles + factor(time_of_day) + factor(season) + factor(weekend) +
```

```
##     factor(hour_type), data = taxi_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.685  -2.478  -0.478   1.712 102.243
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     8.663657   0.710390  12.196  < 2e-16 ***
## factor(payment_type)Credit Card  0.050547   0.145697   0.347  0.72866
## factor(company)101              -5.542683   0.673320  -8.232 2.33e-16 ***
## factor(company)107              -4.832322   0.666413  -7.251 4.76e-13 ***
## factor(company)109              -5.271958   0.678582  -7.769 9.53e-15 ***
## avg_miles                        2.642412   0.016353 161.584  < 2e-16 ***
## factor(time_of_day)Evening      -0.957462   0.181408  -5.278 1.36e-07 ***
## factor(time_of_day)Morning      -0.565612   0.214576  -2.636  0.00842 **
## factor(time_of_day)Night        -0.929296   0.229076  -4.057 5.05e-05 ***
## factor(season)Spring            -0.207040   0.201237  -1.029  0.30361
## factor(season)Summer             0.002421   0.206214   0.012  0.99063
## factor(season)Winter            -0.263646   0.216219  -1.219  0.22277
## factor(weekend)1                 0.027239   0.145897   0.187  0.85190
## factor(hour_type)rush_hour       0.805608   0.179598   4.486 7.43e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.955 on 4986 degrees of freedom
## Multiple R-squared:  0.8579, Adjusted R-squared:  0.8576
## F-statistic:  2316 on 13 and 4986 DF,  p-value: < 2.2e-16
```

Ajd.rsq = 0.8576 payment_type is insignificant (t = 0.347 p-value > 0.05) company is significant (all p-values < 0.05) avg_miles is significant (t = 161.584 p-value < 0.05) time_of_day is significant (all p-values < 0.05) season is insignificant (all p-values > 0.05) weekend is insignificant (t = 0.187 p-value > 0.05) hour_type is significant (t = 4.486 p-value < 0.05)

The results of the individual t-tests indicates company, avg_miles, time_of_day, and hour_type should be kept in the model.

Models

```
head (taxi_data, 4)
```

```
##   X pickup_area dropoff_area trip_miles trip_seconds fare
## 1 1           8            8        1.1          540 7.00
## 2 2          32           32        1.3          600 7.75
## 3 3          32           32        1.0          540 7.00
## 4 4          32           32        1.1          540 7.00
##   trip_start_timestamp tips tolls trip_total payment_type company extras
## 1  2016-12-24 13:15:00  0.0     0       8.00         Cash     107      1
## 2   2016-9-20 17:45:00  1.0     0       8.75  Credit Card     107      0
## 3    2016-9-7 12:30:00  1.5     0       9.50  Credit Card     109      1
```

```
## 4    2016-6-7 18:00:00  1.0     0       8.00  Credit Card     107      0
##   pickup_dropoff avg_miles avg_minutes hours months day_of_week
## 1          8 _ 8  1.326054           8    13     12           7
## 2         32 _ 32  1.326054           8    17      9           3
## 3         32 _ 32  1.326054           8    12      9           4
## 4         32 _ 32  1.326054           8    18      6           3
##       hour_type tip_pct tip_type pickup_dropoff_dummy weekend season
## 1 not_rush_hour    0.00  unknown                -1_-1       1 Winter
## 2     rush_hour    0.13  regular                -1_-1       0   Fall
## 3 not_rush_hour    0.21     high                -1_-1       0   Fall
## 4     rush_hour    0.14  regular                -1_-1       0 Summer
##   time_of_day
## 1   Afternoon
## 2   Afternoon
## 3   Afternoon
## 4     Evening

#   From indiv t-tests, same as SW, BW, FW regression
taxi_lm_red = lm ( fare ~ factor(company) + avg_miles + factor(time_of_day) +
factor(hour_type), data = taxi_data)

#   Without time of day
taxi_lm_3 = lm ( fare ~ factor(company) + avg_miles + factor(hour_type), data
= taxi_data)

#   Without hour_type
taxi_lm_2 = lm ( fare ~ factor(company) + avg_miles, data = taxi_data)
```

Partial F-test

```
#   Comparison between full and reduced model
anova (taxi_fulllm_new, taxi_lm_red)

## Analysis of Variance Table
##
## Model 1: fare ~ factor(payment_type) + factor(company) + avg_miles +
factor(time_of_day) +
##     factor(season) + factor(weekend) + factor(hour_type)
## Model 2: fare ~ factor(company) + avg_miles + factor(time_of_day) +
factor(hour_type)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4986 122408
## 2   4991 122480 -5   -72.092 0.5873 0.7098

summary (taxi_lm_red)

##
## Call:
## lm(formula = fare ~ factor(company) + avg_miles + factor(time_of_day) +
##     factor(hour_type), data = taxi_data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -17.784  -2.468  -0.475   1.736 102.335
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   8.6012     0.6894  12.477  < 2e-16 ***
## factor(company)101           -5.5607     0.6728  -8.264  < 2e-16 ***
## factor(company)107           -4.8569     0.6660  -7.292 3.53e-13 ***
## factor(company)109           -5.2915     0.6781  -7.803 7.31e-15 ***
## avg_miles                     2.6438     0.0158 167.345  < 2e-16 ***
## factor(time_of_day)Evening   -0.9657     0.1810  -5.334 1.00e-07 ***
## factor(time_of_day)Morning   -0.5700     0.2140  -2.663  0.00776 **
## factor(time_of_day)Night     -0.9280     0.2258  -4.111 4.01e-05 ***
## factor(hour_type)rush_hour    0.8033     0.1793   4.480 7.63e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.954 on 4991 degrees of freedom
## Multiple R-squared:  0.8578, Adjusted R-squared:  0.8576
## F-statistic:  3765 on 8 and 4991 DF,  p-value: < 2.2e-16
```

Ajd.rsq = 0.8576. The value has not changed from the full model.

Just checking a few things.

```
summary (taxi_lm_3)
```

```
##
## Call:
## lm(formula = fare ~ factor(company) + avg_miles + factor(hour_type),
##     data = taxi_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.354  -2.489  -0.476   1.717 102.489
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  7.84646    0.67744  11.582  < 2e-16 ***
## factor(company)101          -5.57541    0.67437  -8.268  < 2e-16 ***
## factor(company)107          -4.84803    0.66761  -7.262 4.41e-13 ***
## factor(company)109          -5.28680    0.67988  -7.776 9.03e-15 ***
## avg_miles                    2.65542    0.01558 170.451  < 2e-16 ***
## factor(hour_type)rush_hour   1.06257    0.17101   6.213 5.61e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.968 on 4994 degrees of freedom
## Multiple R-squared:  0.857,  Adjusted R-squared:  0.8568
## F-statistic:  5984 on 5 and 4994 DF,  p-value: < 2.2e-16
```

```
#
anova (taxi_fulllm_new, taxi_lm_3)

## Analysis of Variance Table
##
## Model 1: fare ~ factor(payment_type) + factor(company) + avg_miles +
factor(time_of_day) +
##     factor(season) + factor(weekend) + factor(hour_type)
## Model 2: fare ~ factor(company) + avg_miles + factor(hour_type)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4986 122408
## 2   4994 123244 -8   -836.27 4.2579 4.117e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 summary (taxi_lm_3)

##
## Call:
## lm(formula = fare ~ factor(company) + avg_miles + factor(hour_type),
##     data = taxi_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.354  -2.489  -0.476   1.717 102.489
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  7.84646    0.67744  11.582  < 2e-16 ***
## factor(company)101          -5.57541    0.67437  -8.268  < 2e-16 ***
## factor(company)107          -4.84803    0.66761  -7.262 4.41e-13 ***
## factor(company)109          -5.28680    0.67988  -7.776 9.03e-15 ***
## avg_miles                    2.65542    0.01558 170.451  < 2e-16 ***
## factor(hour_type)rush_hour  1.06257    0.17101   6.213 5.61e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.968 on 4994 degrees of freedom
## Multiple R-squared:  0.857,  Adjusted R-squared:  0.8568
## F-statistic:  5984 on 5 and 4994 DF,  p-value: < 2.2e-16

anova (taxi_lm_red, taxi_lm_3)

## Analysis of Variance Table
##
## Model 1: fare ~ factor(company) + avg_miles + factor(time_of_day) +
factor(hour_type)
## Model 2: fare ~ factor(company) + avg_miles + factor(hour_type)
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1   4991 122480
## 2   4994 123244 -3   -764.18 10.38 8.312e-07 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova (taxi_fulllm_new, taxi_lm_3)
```

```
## Analysis of Variance Table
##
## Model 1: fare ~ factor(payment_type) + factor(company) + avg_miles +
factor(time_of_day) +
##     factor(season) + factor(weekend) + factor(hour_type)
## Model 2: fare ~ factor(company) + avg_miles + factor(hour_type)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4986 122408
## 2   4994 123244 -8   -836.27 4.2579 4.117e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova (taxi_lm_3, taxi_lm_2)
```

```
## Analysis of Variance Table
##
## Model 1: fare ~ factor(company) + avg_miles + factor(hour_type)
## Model 2: fare ~ factor(company) + avg_miles
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4994 123244
## 2   4995 124197 -1   -952.75 38.606 5.605e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova (taxi_fulllm_new, taxi_lm_2)
```

```
## Analysis of Variance Table
##
## Model 1: fare ~ factor(payment_type) + factor(company) + avg_miles +
factor(time_of_day) +
##     factor(season) + factor(weekend) + factor(hour_type)
## Model 2: fare ~ factor(company) + avg_miles
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4986 122408
## 2   4995 124197 -9     -1789 8.0968 5.155e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# So, the final model is taxi_lm_red with 4 variables. next we are going to
start adding the interactions and perform the required tests on the model
```

```
taxi_lm_red_int = lm ( fare ~ (factor(company) + avg_miles +
factor(time_of_day) + factor(hour_type))^2, data = taxi_data)
```

```
summary(taxi_lm_red_int)
```

```
##
## Call:
## lm(formula = fare ~ (factor(company) + avg_miles + factor(time_of_day) +
##     factor(hour_type))^2, data = taxi_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.431  -2.357  -0.461   1.702 102.406
##
## Coefficients: (1 not defined because of singularities)
##                                                        Estimate Std. Error
## (Intercept)                                            -2.16734    2.14110
## factor(company)101                                      5.12001    2.15021
## factor(company)107                                      5.12188    2.13552
## factor(company)109                                      4.58331    2.16624
## avg_miles                                               3.56515    0.15797
## factor(time_of_day)Evening                              0.44231    1.59753
## factor(time_of_day)Morning                              0.08704    2.17325
## factor(time_of_day)Night                                4.36152    2.89373
## factor(hour_type)rush_hour                              7.64661    1.80133
## factor(company)101:avg_miles                           -0.92865    0.15846
## factor(company)107:avg_miles                           -0.76484    0.15707
## factor(company)109:avg_miles                           -0.85408    0.15951
## factor(company)101:factor(time_of_day)Evening           0.09077    1.58593
## factor(company)107:factor(time_of_day)Evening          -0.48897    1.56777
## factor(company)109:factor(time_of_day)Evening           0.32443    1.59935
## factor(company)101:factor(time_of_day)Morning           0.76215    2.16226
## factor(company)107:factor(time_of_day)Morning          -0.05384    2.13816
## factor(company)109:factor(time_of_day)Morning           0.68035    2.17200
## factor(company)101:factor(time_of_day)Night            -4.13189    2.89146
## factor(company)107:factor(time_of_day)Night            -4.46426    2.87858
## factor(company)109:factor(time_of_day)Night            -3.71816    2.91783
## factor(company)101:factor(hour_type)rush_hour          -8.02143    1.78926
## factor(company)107:factor(hour_type)rush_hour          -7.70087    1.77142
## factor(company)109:factor(hour_type)rush_hour          -7.01456    1.79819
## avg_miles:factor(time_of_day)Evening                   -0.21708    0.03837
## avg_miles:factor(time_of_day)Morning                   -0.09550    0.04520
## avg_miles:factor(time_of_day)Night                     -0.18500    0.05462
## avg_miles:factor(hour_type)rush_hour                    0.06853    0.03785
## factor(time_of_day)Evening:factor(hour_type)rush_hour   1.40762    0.40684
## factor(time_of_day)Morning:factor(hour_type)rush_hour  -1.01008    0.46161
## factor(time_of_day)Night:factor(hour_type)rush_hour           NA         NA
##                                                        t value Pr(>|t|)
## (Intercept)                                             -1.012 0.311464
## factor(company)101                                       2.381 0.017295 *
## factor(company)107                                       2.398 0.016502 *
## factor(company)109                                       2.116 0.034412 *
## avg_miles                                               22.568  < 2e-16 ***
## factor(time_of_day)Evening                               0.277 0.781888
## factor(time_of_day)Morning                               0.040 0.968053
```

```
## factor(time_of_day)Night                                    1.507 0.131816
## factor(hour_type)rush_hour                                  4.245 2.23e-05 ***
## factor(company)101:avg_miles                               -5.860 4.91e-09 ***
## factor(company)107:avg_miles                               -4.869 1.15e-06 ***
## factor(company)109:avg_miles                               -5.354 8.97e-08 ***
## factor(company)101:factor(time_of_day)Evening               0.057 0.954360
## factor(company)107:factor(time_of_day)Evening              -0.312 0.755139
## factor(company)109:factor(time_of_day)Evening               0.203 0.839262
## factor(company)101:factor(time_of_day)Morning               0.352 0.724496
## factor(company)107:factor(time_of_day)Morning              -0.025 0.979911
## factor(company)109:factor(time_of_day)Morning               0.313 0.754116
## factor(company)101:factor(time_of_day)Night                -1.429 0.153068
## factor(company)107:factor(time_of_day)Night                -1.551 0.121000
## factor(company)109:factor(time_of_day)Night                -1.274 0.202621
## factor(company)101:factor(hour_type)rush_hour              -4.483 7.52e-06 ***
## factor(company)107:factor(hour_type)rush_hour              -4.347 1.41e-05 ***
## factor(company)109:factor(hour_type)rush_hour              -3.901 9.71e-05 ***
## avg_miles:factor(time_of_day)Evening                       -5.657 1.62e-08 ***
## avg_miles:factor(time_of_day)Morning                       -2.113 0.034661 *
## avg_miles:factor(time_of_day)Night                         -3.387 0.000712 ***
## avg_miles:factor(hour_type)rush_hour                        1.811 0.070255 .
## factor(time_of_day)Evening:factor(hour_type)rush_hour       3.460 0.000545 ***
## factor(time_of_day)Morning:factor(hour_type)rush_hour      -2.188 0.028705 *
## factor(time_of_day)Night:factor(hour_type)rush_hour            NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.893 on 4970 degrees of freedom
## Multiple R-squared:  0.8619, Adjusted R-squared:  0.8611
## F-statistic:  1069 on 29 and 4970 DF,  p-value: < 2.2e-16
```

*# The full interaction model has increased the adjusted R^2 from 0.8576 to 0.8611. But the individual t-Test indicates that ONLY avg_miles*company, hour_type*company, time_of_day*ave_miles, time_of_day*hour_type interactions are significant. So, the model to be reduced and all insignificant interactions to be removed.*

```
taxi_lm_red_int_red = lm ( fare ~ factor(company) + avg_miles +
factor(time_of_day) + factor(hour_type)+ avg_miles*factor(company)+

factor(company)*factor(hour_type)+avg_miles*factor(time_of_day)+
                            factor(time_of_day)*factor(hour_type) , data =
taxi_data)

summary(taxi_lm_red_int_red)
```

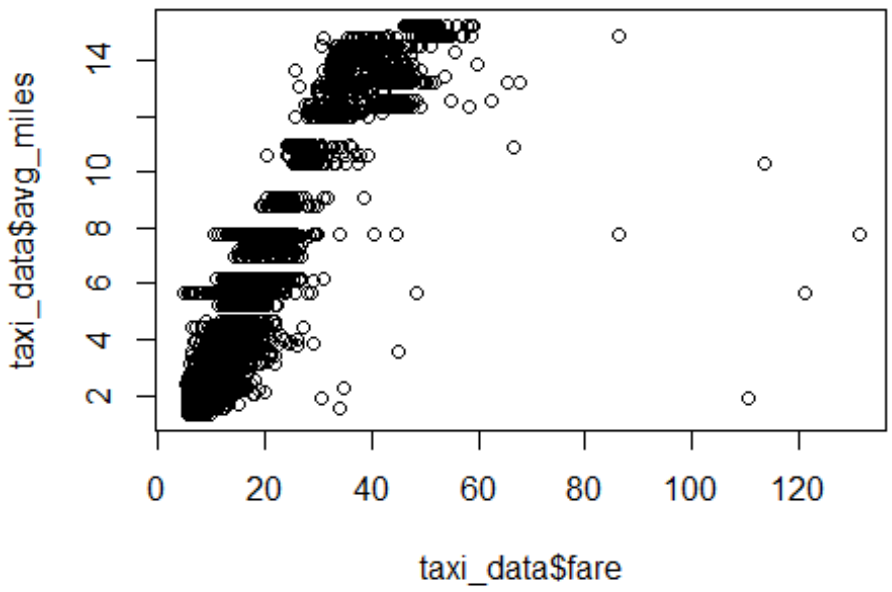
```
##
## Call:
## lm(formula = fare ~ factor(company) + avg_miles + factor(time_of_day) +
##     factor(hour_type) + avg_miles * factor(company) + factor(company) *
##     factor(hour_type) + avg_miles * factor(time_of_day) +
```

```
factor(time_of_day) *
##     factor(hour_type), data = taxi_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.913  -2.357  -0.443   1.712 102.383
##
## Coefficients: (1 not defined because of singularities)
##                                                        Estimate Std. Error
## (Intercept)                                            -1.24072    1.73131
## factor(company)101                                      4.21074    1.73079
## factor(company)107                                      3.77125    1.72325
## factor(company)109                                      3.88466    1.74166
## avg_miles                                               3.51612    0.15102
## factor(time_of_day)Evening                              0.35003    0.33801
## factor(time_of_day)Morning                              0.44600    0.41387
## factor(time_of_day)Night                                0.29885    0.38317
## factor(hour_type)rush_hour                              7.86401    1.66009
## factor(company)101:avg_miles                           -0.85648    0.15200
## factor(company)107:avg_miles                           -0.68965    0.15059
## factor(company)109:avg_miles                           -0.78778    0.15303
## factor(company)101:factor(hour_type)rush_hour          -7.74934    1.68444
## factor(company)107:factor(hour_type)rush_hour          -7.35395    1.66704
## factor(company)109:factor(hour_type)rush_hour          -6.86604    1.69271
## avg_miles:factor(time_of_day)Evening                   -0.22565    0.03754
## avg_miles:factor(time_of_day)Morning                   -0.10362    0.04479
## avg_miles:factor(time_of_day)Night                     -0.20600    0.05267
## factor(time_of_day)Evening:factor(hour_type)rush_hour  1.35763    0.40557
## factor(time_of_day)Morning:factor(hour_type)rush_hour -1.00604    0.46112
## factor(time_of_day)Night:factor(hour_type)rush_hour          NA         NA
##                                                        t value Pr(>|t|)
## (Intercept)                                             -0.717 0.473631
## factor(company)101                                       2.433 0.015016 *
## factor(company)107                                       2.188 0.028683 *
## factor(company)109                                       2.230 0.025763 *
## avg_miles                                               23.283  < 2e-16 ***
## factor(time_of_day)Evening                               1.036 0.300464
## factor(time_of_day)Morning                               1.078 0.281250
## factor(time_of_day)Night                                 0.780 0.435461
## factor(hour_type)rush_hour                               4.737 2.23e-06 ***
## factor(company)101:avg_miles                            -5.635 1.85e-08 ***
## factor(company)107:avg_miles                            -4.580 4.77e-06 ***
## factor(company)109:avg_miles                            -5.148 2.74e-07 ***
## factor(company)101:factor(hour_type)rush_hour           -4.601 4.32e-06 ***
## factor(company)107:factor(hour_type)rush_hour           -4.411 1.05e-05 ***
## factor(company)109:factor(hour_type)rush_hour           -4.056 5.06e-05 ***
## avg_miles:factor(time_of_day)Evening                    -6.012 1.97e-09 ***
## avg_miles:factor(time_of_day)Morning                    -2.314 0.020735 *
## avg_miles:factor(time_of_day)Night                      -3.911 9.31e-05 ***
## factor(time_of_day)Evening:factor(hour_type)rush_hour    3.347 0.000822 ***
```

```
## factor(time_of_day)Morning:factor(hour_type)rush_hour  -2.182 0.029175 *
## factor(time_of_day)Night:factor(hour_type)rush_hour        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.894 on 4980 degrees of freedom
## Multiple R-squared:  0.8616, Adjusted R-squared:  0.861
## F-statistic:  1631 on 19 and 4980 DF,  p-value: < 2.2e-16

# the reduced interaction model has increased the adjusted R^2 from 0.8576 to
0.8601. The R^2 is slightly less than the full interaction model but no
insignificant variables should be kept in the model.

# partial F test between full interaction model & the reduced interaction
model
anova (taxi_lm_red_int_red, taxi_lm_red_int)

## Analysis of Variance Table
##
## Model 1: fare ~ factor(company) + avg_miles + factor(time_of_day) +
factor(hour_type) +
##     avg_miles * factor(company) + factor(company) * factor(hour_type) +
##     avg_miles * factor(time_of_day) + factor(time_of_day) *
factor(hour_type)
## Model 2: fare ~ (factor(company) + avg_miles + factor(time_of_day) +
factor(hour_type))^2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4980 119278
## 2   4970 119009 10    268.08 1.1195 0.3428

# Hypothesis
# H0: Bi = 0 ,               i = all coefficient indexes that are in the full
interaction model but not in the reduced model
# Ha: at least one Bi != 0 , i = all coefficient indexes that are in the full
interaction model but not in the reduced model

# Partial F test returned a P-value of 0.3428 > 0.05 meaning the H0 cannot be
rejected. This confirms that the reduced interaction model works better than
the full interaction model.

# partial F test between reduced interaction model & the simple model
anova (taxi_lm_red_int_red, taxi_lm_red)

## Analysis of Variance Table
##
## Model 1: fare ~ factor(company) + avg_miles + factor(time_of_day) +
factor(hour_type) +
##     avg_miles * factor(company) + factor(company) * factor(hour_type) +
##     avg_miles * factor(time_of_day) + factor(time_of_day) *
factor(hour_type)
## Model 2: fare ~ factor(company) + avg_miles + factor(time_of_day) +
```

```
factor(hour_type)
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1   4980 119278
## 2   4991 122480 -11   -3202.2 12.154 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Hypothesis
# H0: Bi = 0 ,                i = all coefficient indexes for the interactions
# Ha: at least one Bi != 0 ,  i = all coefficient indexes for the interactions

# Partial F test returned a small P-value  < 0.05 meaning the H0 can be
rejected in favor of the alternative. This means that the reduced interaction
model works better than the simple model.

# Next we will be checking whether a higher order relation exists between
avg_miles and  fare.

#pairs (~fare+ avg_miles ,data = taxi_data)
plot(taxi_data$fare,taxi_data$avg_miles)
```




```
taxi_lm_red_int_red_high = lm ( fare ~ factor(company) + factor(time_of_day)
+factor(hour_type)+
                                avg_miles*factor(company)+
factor(company)*factor(hour_type)+
                                avg_miles*factor(time_of_day)+
factor(time_of_day)*factor(hour_type)+
```

```
                                poly(avg_miles,degree= 12, raw =TRUE), data
= taxi_data)

summary(taxi_lm_red_int_red_high)

##
## Call:
## lm(formula = fare ~ factor(company) + factor(time_of_day) +
factor(hour_type) +
##     avg_miles * factor(company) + factor(company) * factor(hour_type) +
##     avg_miles * factor(time_of_day) + factor(time_of_day) *
factor(hour_type) +
##     poly(avg_miles, degree = 12, raw = TRUE), data = taxi_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.006  -2.096  -0.415   1.676 102.633
##
## Coefficients: (2 not defined because of singularities)
##                                                      Estimate
## (Intercept)                                        -3.668e+02
## factor(company)101                                  3.019e+00
## factor(company)107                                  2.604e+00
## factor(company)109                                  2.853e+00
## factor(time_of_day)Evening                          3.293e-01
## factor(time_of_day)Morning                          3.615e-01
## factor(time_of_day)Night                            4.875e-02
## factor(hour_type)rush_hour                          8.526e+00
## avg_miles                                           9.945e+02
## poly(avg_miles, degree = 12, raw = TRUE)1                  NA
## poly(avg_miles, degree = 12, raw = TRUE)2          -1.139e+03
## poly(avg_miles, degree = 12, raw = TRUE)3           7.393e+02
## poly(avg_miles, degree = 12, raw = TRUE)4          -3.033e+02
## poly(avg_miles, degree = 12, raw = TRUE)5           8.320e+01
## poly(avg_miles, degree = 12, raw = TRUE)6          -1.572e+01
## poly(avg_miles, degree = 12, raw = TRUE)7           2.072e+00
## poly(avg_miles, degree = 12, raw = TRUE)8          -1.897e-01
## poly(avg_miles, degree = 12, raw = TRUE)9           1.181e-02
## poly(avg_miles, degree = 12, raw = TRUE)10         -4.765e-04
## poly(avg_miles, degree = 12, raw = TRUE)11          1.122e-05
## poly(avg_miles, degree = 12, raw = TRUE)12         -1.170e-07
## factor(company)101:avg_miles                       -6.311e-01
## factor(company)107:avg_miles                       -4.685e-01
## factor(company)109:avg_miles                       -5.840e-01
## factor(company)101:factor(hour_type)rush_hour      -8.122e+00
## factor(company)107:factor(hour_type)rush_hour      -7.857e+00
## factor(company)109:factor(hour_type)rush_hour      -7.499e+00
## factor(time_of_day)Evening:avg_miles               -1.810e-01
## factor(time_of_day)Morning:avg_miles               -6.339e-02
## factor(time_of_day)Night:avg_miles                 -1.349e-01
```

```
## factor(time_of_day)Evening:factor(hour_type)rush_hour  8.613e-01
## factor(time_of_day)Morning:factor(hour_type)rush_hour -1.058e+00
## factor(time_of_day)Night:factor(hour_type)rush_hour           NA
##                                                     Std. Error t value
## (Intercept)                                          9.247e+01  -3.967
## factor(company)101                                   1.659e+00   1.819
## factor(company)107                                   1.652e+00   1.576
## factor(company)109                                   1.669e+00   1.709
## factor(time_of_day)Evening                           3.246e-01   1.015
## factor(time_of_day)Morning                           3.963e-01   0.912
## factor(time_of_day)Night                             3.697e-01   0.132
## factor(hour_type)rush_hour                           1.589e+00   5.365
## avg_miles                                            2.363e+02   4.209
## poly(avg_miles, degree = 12, raw = TRUE)1                   NA      NA
## poly(avg_miles, degree = 12, raw = TRUE)2            2.579e+02  -4.416
## poly(avg_miles, degree = 12, raw = TRUE)3            1.594e+02   4.639
## poly(avg_miles, degree = 12, raw = TRUE)4            6.233e+01  -4.865
## poly(avg_miles, degree = 12, raw = TRUE)5            1.633e+01   5.095
## poly(avg_miles, degree = 12, raw = TRUE)6            2.952e+00  -5.326
## poly(avg_miles, degree = 12, raw = TRUE)7            3.729e-01   5.556
## poly(avg_miles, degree = 12, raw = TRUE)8            3.281e-02  -5.780
## poly(avg_miles, degree = 12, raw = TRUE)9            1.970e-03   5.995
## poly(avg_miles, degree = 12, raw = TRUE)10           7.687e-05  -6.199
## poly(avg_miles, degree = 12, raw = TRUE)11           1.757e-06   6.388
## poly(avg_miles, degree = 12, raw = TRUE)12           1.783e-08  -6.562
## factor(company)101:avg_miles                         1.460e-01  -4.323
## factor(company)107:avg_miles                         1.446e-01  -3.239
## factor(company)109:avg_miles                         1.469e-01  -3.976
## factor(company)101:factor(hour_type)rush_hour        1.612e+00  -5.037
## factor(company)107:factor(hour_type)rush_hour        1.596e+00  -4.924
## factor(company)109:factor(hour_type)rush_hour        1.621e+00  -4.627
## factor(time_of_day)Evening:avg_miles                 3.619e-02  -5.002
## factor(time_of_day)Morning:avg_miles                 4.295e-02  -1.476
## factor(time_of_day)Night:avg_miles                   5.086e-02  -2.653
## factor(time_of_day)Evening:factor(hour_type)rush_hour  3.894e-01   2.212
## factor(time_of_day)Morning:factor(hour_type)rush_hour  4.415e-01  -2.396
## factor(time_of_day)Night:factor(hour_type)rush_hour           NA      NA
##                                                     Pr(>|t|)
## (Intercept)                                         7.37e-05 ***
## factor(company)101                                   0.06890 .
## factor(company)107                                   0.11505
## factor(company)109                                   0.08754 .
## factor(time_of_day)Evening                           0.31038
## factor(time_of_day)Morning                           0.36171
## factor(time_of_day)Night                             0.89509
## factor(hour_type)rush_hour                          8.46e-08 ***
## avg_miles                                           2.61e-05 ***
## poly(avg_miles, degree = 12, raw = TRUE)1                 NA
## poly(avg_miles, degree = 12, raw = TRUE)2           1.03e-05 ***
## poly(avg_miles, degree = 12, raw = TRUE)3           3.59e-06 ***
```

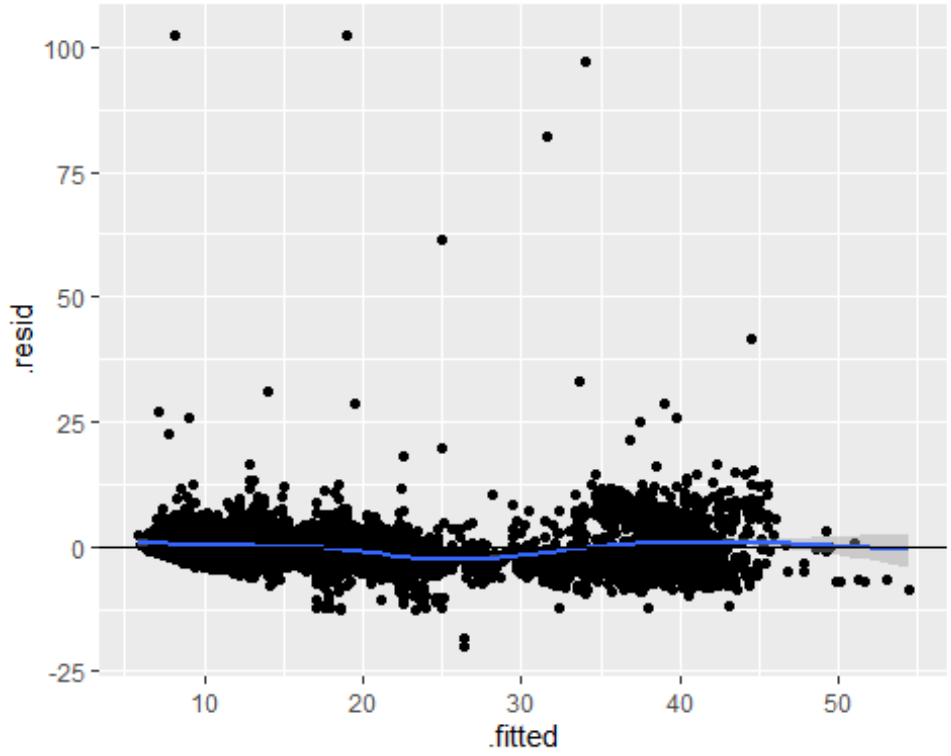
```
## poly(avg_miles, degree = 12, raw = TRUE)4              1.18e-06 ***
## poly(avg_miles, degree = 12, raw = TRUE)5              3.61e-07 ***
## poly(avg_miles, degree = 12, raw = TRUE)6              1.05e-07 ***
## poly(avg_miles, degree = 12, raw = TRUE)7              2.90e-08 ***
## poly(avg_miles, degree = 12, raw = TRUE)8              7.92e-09 ***
## poly(avg_miles, degree = 12, raw = TRUE)9              2.17e-09 ***
## poly(avg_miles, degree = 12, raw = TRUE)10             6.15e-10 ***
## poly(avg_miles, degree = 12, raw = TRUE)11             1.83e-10 ***
## poly(avg_miles, degree = 12, raw = TRUE)12             5.84e-11 ***
## factor(company)101:avg_miles                           1.57e-05 ***
## factor(company)107:avg_miles                            0.00121 **
## factor(company)109:avg_miles                           7.10e-05 ***
## factor(company)101:factor(hour_type)rush_hour          4.90e-07 ***
## factor(company)107:factor(hour_type)rush_hour          8.76e-07 ***
## factor(company)109:factor(hour_type)rush_hour          3.80e-06 ***
## factor(time_of_day)Evening:avg_miles                   5.87e-07 ***
## factor(time_of_day)Morning:avg_miles                    0.14005
## factor(time_of_day)Night:avg_miles                      0.00801 **
## factor(time_of_day)Evening:factor(hour_type)rush_hour   0.02702 *
## factor(time_of_day)Morning:factor(hour_type)rush_hour   0.01662 *
## factor(time_of_day)Night:factor(hour_type)rush_hour           NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.682 on 4969 degrees of freedom
## Multiple R-squared:  0.8736, Adjusted R-squared:  0.8728
## F-statistic:  1145 on 30 and 4969 DF,  p-value: < 2.2e-16

# All the higher order variables seem to be significant. Also, the higher
order model increased the adjusted R2 from 0.8601  to 0.8719

# partial F test between reduced interaction model & the higher order model
anova (taxi_lm_red_int_red, taxi_lm_red_int_red_high)

## Analysis of Variance Table
##
## Model 1: fare ~ factor(company) + avg_miles + factor(time_of_day) +
factor(hour_type) +
##     avg_miles * factor(company) + factor(company) * factor(hour_type) +
##     avg_miles * factor(time_of_day) + factor(time_of_day) *
factor(hour_type)
## Model 2: fare ~ factor(company) + factor(time_of_day) + factor(hour_type)
+
##     avg_miles * factor(company) + factor(company) * factor(hour_type) +
##     avg_miles * factor(time_of_day) + factor(time_of_day) *
factor(hour_type) +
##     poly(avg_miles, degree = 12, raw = TRUE)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4980 119278
## 2   4969 108904 11     10374 43.031 < 2.2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Partial F test provides a small P-value <0.05 which suggests rejecting H0
and keeping the higher order model.
# Next we are going to test all the model conditions
```

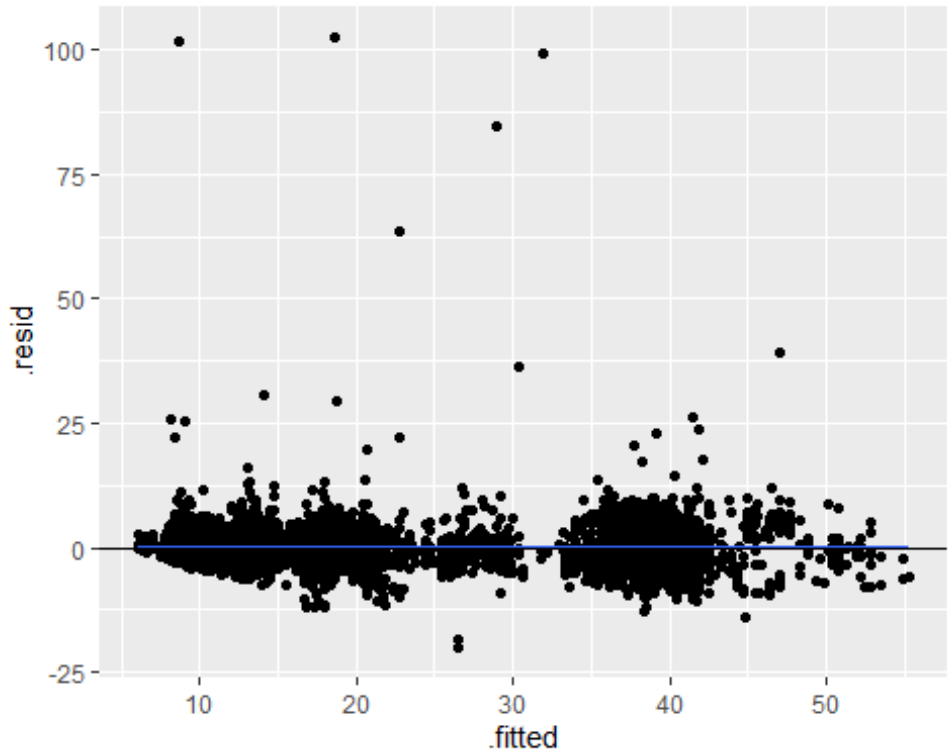
Assumption Test

#1. Linearity Assumption

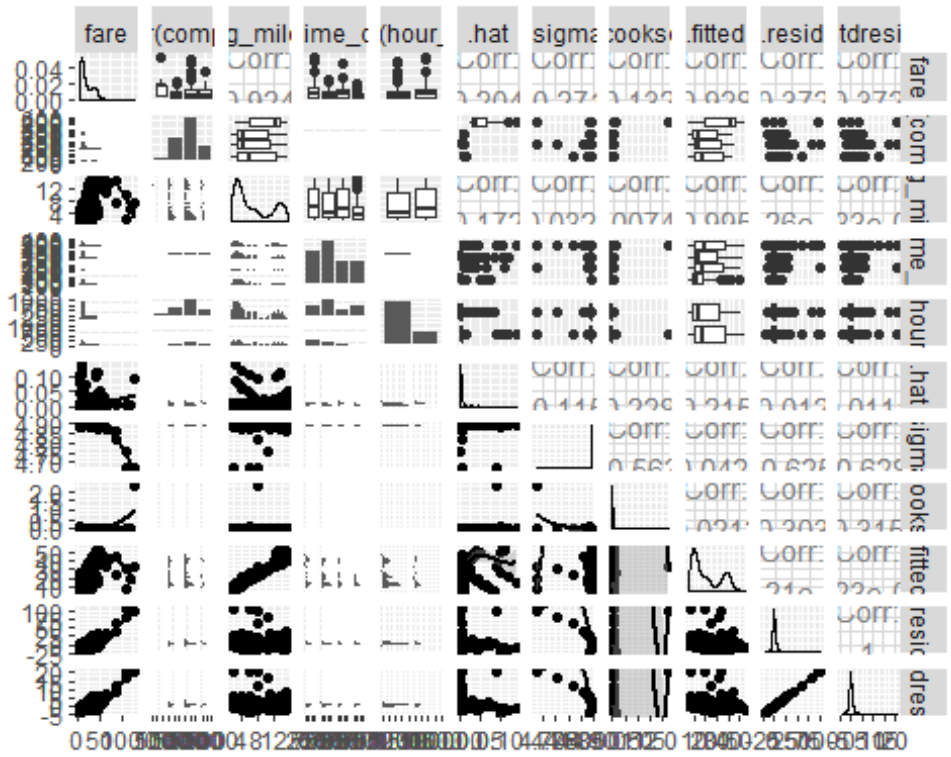
```
ggplot(taxi_lm_red_int_red, aes(x=.fitted, y=.resid)) + geom_point() +
geom_smooth()+ geom_hline(yintercept = 0)

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(taxi_lm_red_int_red_high, aes(x=.fitted, y=.resid)) + geom_point() +
geom_smooth()+ geom_hline(yintercept = 0)

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
# From the above plot we can observe that there is slight pattern in the
residual plot but there is no pattern for our higher order model. so we can
say that higher order model holds the Linearity assumption.

ggpairs(taxi_lm_red_int_red,lower = list(continuous = "smooth_loess", combo
="facethist", discrete = "facetbar", na = "na"))
```



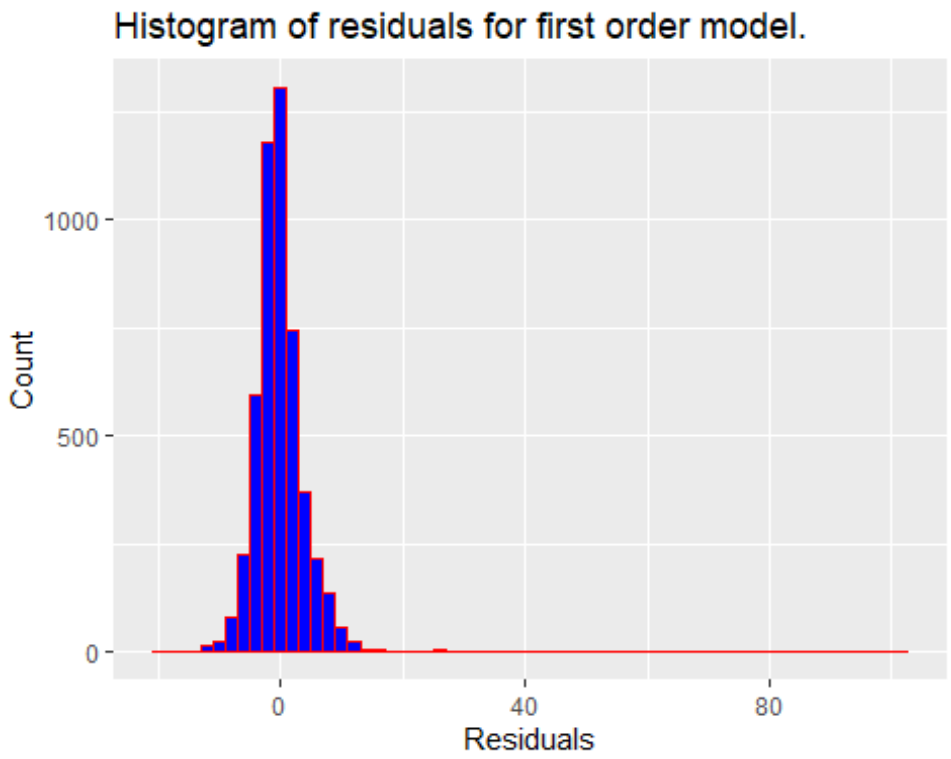
#3. Equal Variance Assumption

```
#Breusch-Pagan test

#Ho: homoscedasticity
#Ha: heteroscedasticity

bptest(taxi_lm_red_int_red)

##
##  studentized Breusch-Pagan test
##
## data:  taxi_lm_red_int_red
## BP = 156.15, df = 19, p-value < 2.2e-16

bptest(taxi_lm_red_int_red_high)

##
##  studentized Breusch-Pagan test
##
## data:  taxi_lm_red_int_red_high
## BP = 169.1, df = 30, p-value < 2.2e-16

# From the above output of Breusch-Pagan test p-value is 0.00000000000000022
less than alpha=0.05 so we reject the null hypothesis and conclude that both
the model have heteroscedasticity.
```

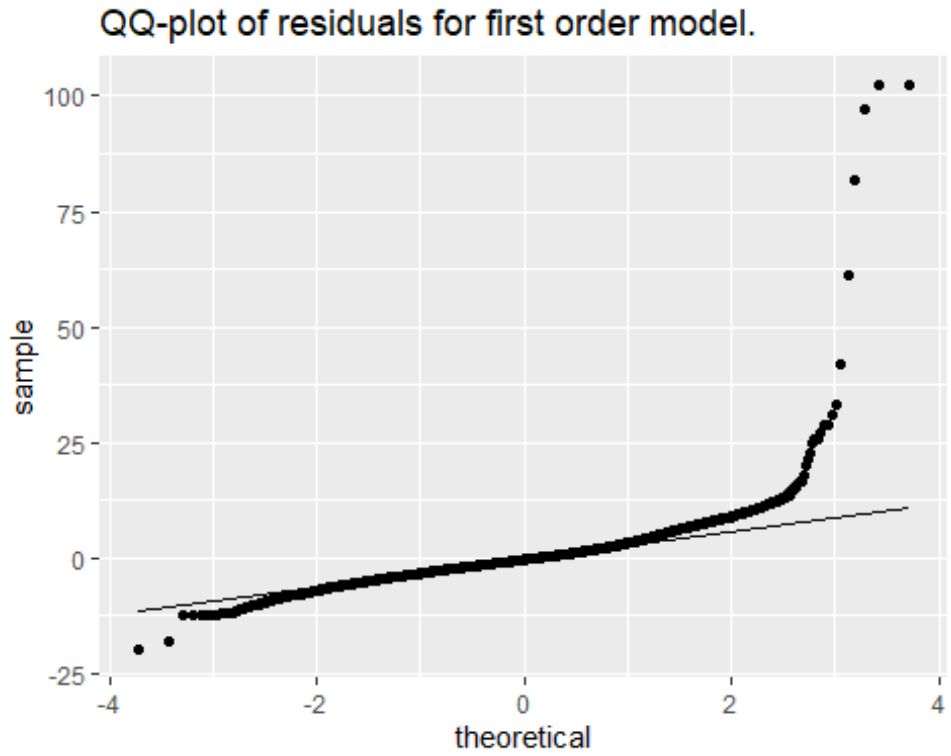
#4. Normality Assumption

```r
#Ho: the sample data are significantly normally distributed
#Ha: the sample data are not significantly normally distributed

ggplot(data=taxi_data, aes(residuals(taxi_lm_red_int_red_high))) +
geom_histogram(col="red", fill="blue", binwidth=2) + labs(title="Histogram of
residuals for twelve order model") + labs(x="Residuals", y="Count")
```

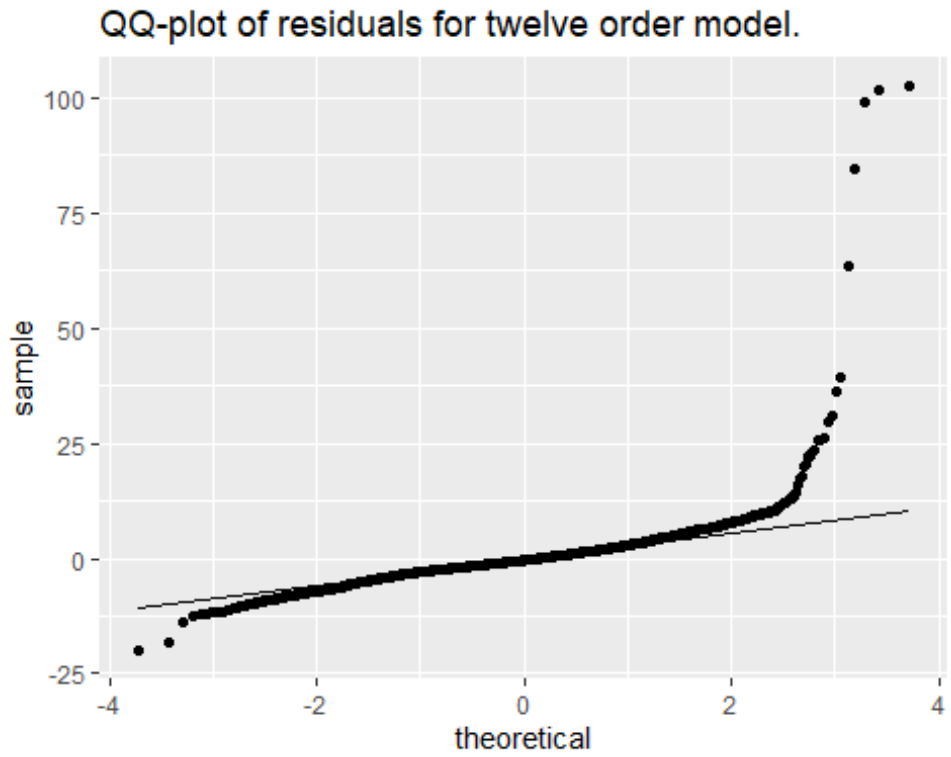

```r
ggplot(data=taxi_data, aes(residuals(taxi_lm_red_int_red))) +
geom_histogram(col="red", fill="blue", binwidth=2) + labs(title="Histogram of
residuals for first order model.") + labs(x="Residuals", y="Count")
```




```
ggplot(taxi_data, aes(sample=taxi_lm_red_int_red$residuals)) +stat_qq() +
stat_qq_line()  + labs(title="QQ-plot of residuals for first order model.")
```



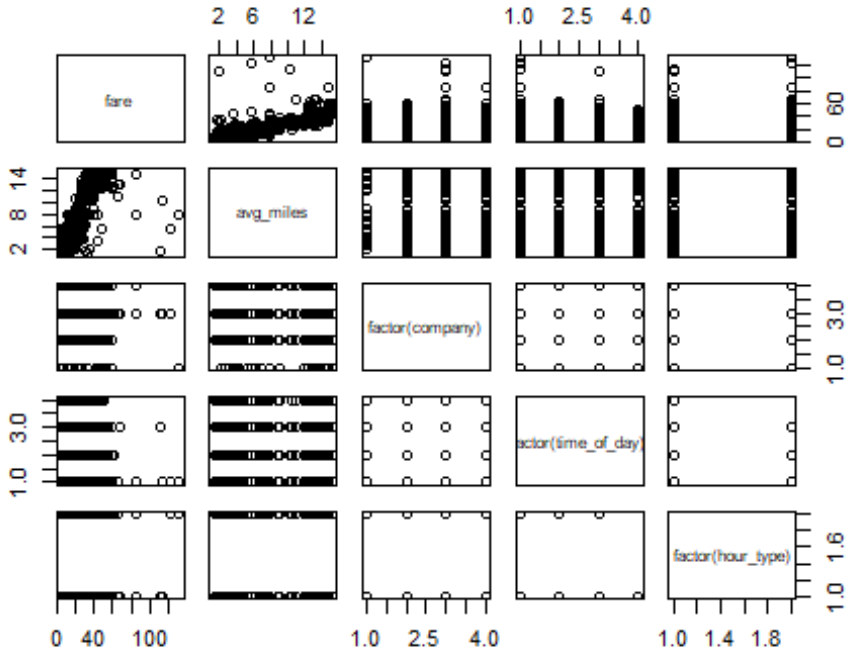
```
ggplot(taxi_data, aes(sample=taxi_lm_red_int_red_high$residuals)) +stat_qq()
+ stat_qq_line()  + labs(title="QQ-plot of residuals for twelve order
model.")
```




```
shapiro.test(residuals(taxi_lm_red_int_red_high))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(taxi_lm_red_int_red_high)
## W = 0.66, p-value < 2.2e-16

shapiro.test(residuals(taxi_lm_red_int_red))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(taxi_lm_red_int_red)
## W = 0.70125, p-value < 2.2e-16
```

5. Multicollinearity

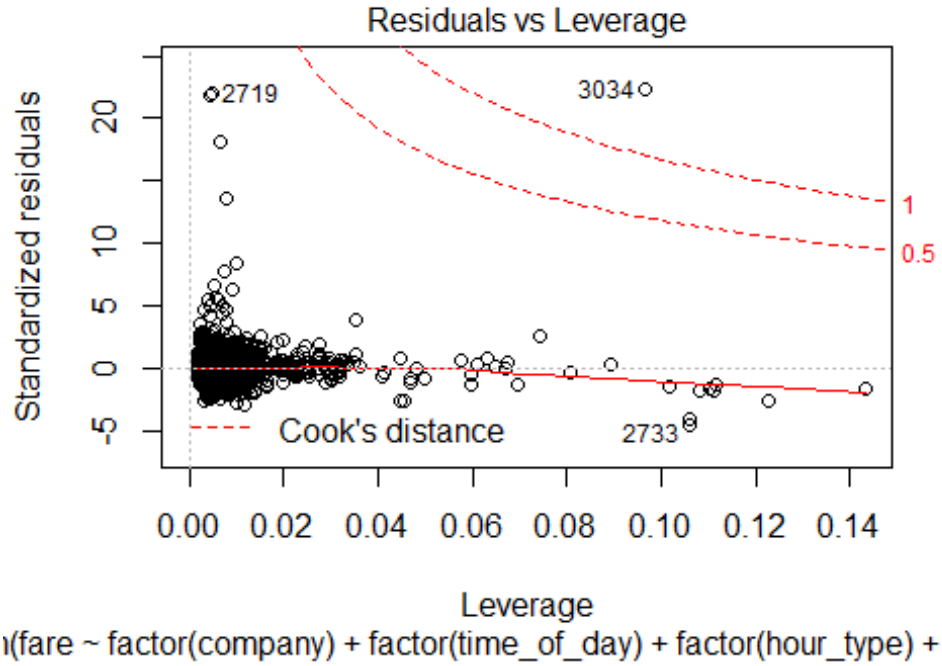
```
pairs(~fare+ avg_miles+factor(company) + factor(time_of_day)
+factor(hour_type),data=taxi_data)
```




```
X1<-cbind( taxi_data$avg_miles, factor(taxi_data$company),
factor(taxi_data$time_of_day), factor(taxi_data$hour_type))
imcdiag(X1,taxi_data$fare, method="VIF")

##
## Call:
## imcdiag(x = X1, y = taxi_data$fare, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##       VIF detection
## V1 1.0255         0
## V2 1.0029         0
## V3 1.1002         0
## V4 1.0761         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================

# From the below plot and VIF test we can state that there is no
multicollinearity in our variable.
```

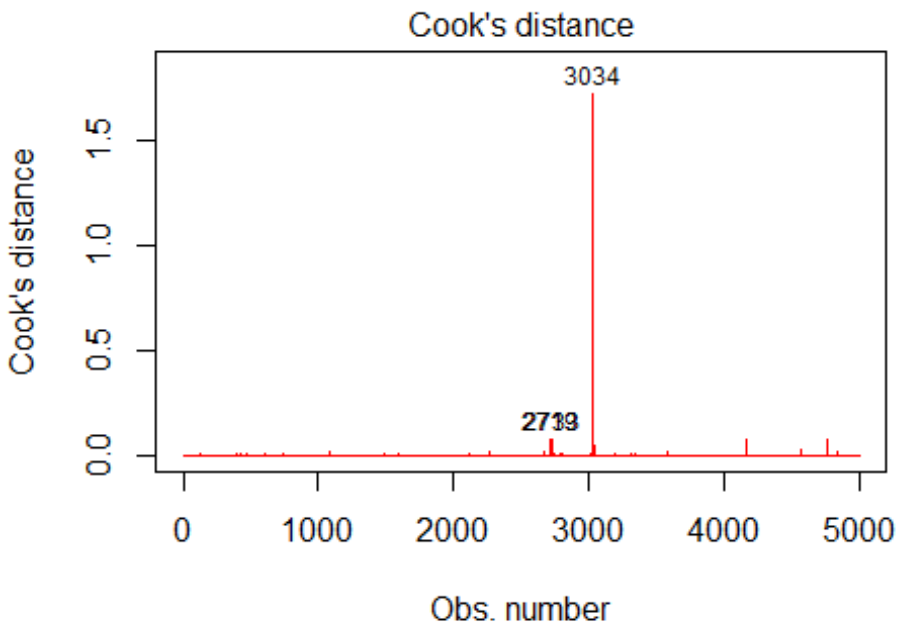
6. Outlier

```
#Residuals vs Leverage plot
plot(taxi_lm_red_int_red_high,which=5)
```



ı(fare ~ factor(company) + factor(time_of_day) + factor(hour_type) + a\

```
#Cook's Distance
taxi_data[cooks.distance(taxi_lm_red_int_red_high)>0.5,]

##         X pickup_area dropoff_area trip_miles trip_seconds   fare
## 3034 3034          76           76        3.4         5280 131.25
##      trip_start_timestamp tips tolls trip_total payment_type company
## 3034   2016-7-30 16:30:00    0     0     137.75         Cash       8
##      extras pickup_dropoff avg_miles avg_minutes hours months day_of_week
## 3034    6.5        76 _ 76   7.79902          19    16      7           7
##      hour_type tip_pct tip_type pickup_dropoff_dummy weekend season
## 3034 rush_hour       0  unknown               76 _ 76       1 Summer
##      time_of_day
## 3034   Afternoon

plot(taxi_lm_red_int_red_high,pch=18,col="red",which=c(4))
```



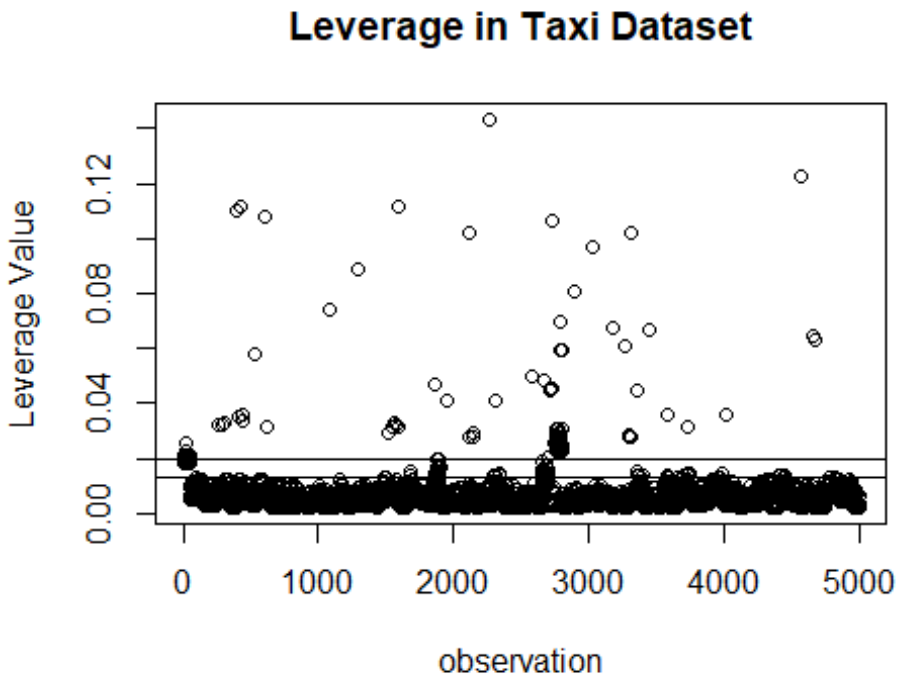

```
ı(fare ~ factor(company) + factor(time_of_day) + factor(hour_type) + a\
```

```
lev=hatvalues(taxi_lm_red_int_red_high)
p = length(coef(taxi_lm_red_int_red_high))
n = nrow(taxi_data)
outlier = lev[lev>(3*p/n)]
print(outlier)
```

```
##           2           3           4           9          12          13
## 0.01980660 0.02043110 0.02091317 0.02073254 0.01980658 0.01980662
##          17          21          24          27          28          29
## 0.02520953 0.02073215 0.02005443 0.02073215 0.02249231 0.02091324
##          30          31          35          45          47          50
## 0.02005443 0.02073215 0.01980661 0.01987216 0.02073215 0.02091324
##         263         288         300         401         408         419
## 0.03194823 0.03194823 0.03284992 0.11024201 0.03475214 0.11149194
##         436         437         528         609         613         615
## 0.03599579 0.03343886 0.05782265 0.10799810 0.03136067 0.03136067
##        1081        1293        1525        1555        1562        1571
## 0.07420049 0.08913463 0.02887110 0.03113004 0.03322079 0.03180329
##        1578        1586        1599        1862        1867        1954
## 0.03180329 0.03113004 0.11138208 0.04672691 0.04672691 0.04075505
##        2117        2122        2142        2143        2149        2268
## 0.02777517 0.10177129 0.02958916 0.02958916 0.02777517 0.14325982
##        2314        2577        2665        2692        2707        2711
## 0.04132194 0.05003701 0.04817071 0.02004096 0.04486946 0.04573375
##        2722        2725        2733        2751        2752        2753
## 0.04573375 0.10613737 0.10613737 0.02529740 0.02455999 0.02529740
##        2754        2755        2756        2757        2758        2759
```

```
## 0.02529740 0.02446378 0.02281584 0.02515108 0.02446378 0.02281584
##       2760       2761       2762       2763       2764       2765
## 0.02515108 0.02281584 0.02446378 0.02446378 0.02515108 0.02851662
##       2766       2767       2768       2769       2770       2771
## 0.03037770 0.02446378 0.02918939 0.02567763 0.02755828 0.03037770
##       2772       2773       2774       2775       2776       2777
## 0.02479757 0.02263362 0.02479757 0.02281584 0.02567763 0.02918939
##       2778       2779       2780       2781       2782       2783
## 0.02281584 0.02263362 0.02769595 0.02413846 0.02263362 0.02263362
##       2784       2785       2786       2787       2788       2789
## 0.02918939 0.06966335 0.02446378 0.02515108 0.05971495 0.02479757
##       2790       2791       2792       2793       2794       2795
## 0.02263362 0.02413846 0.02281584 0.02446378 0.02755828 0.02529740
##       2796       2797       2798       2799       2800       2901
## 0.02413846 0.02263362 0.03037770 0.02479757 0.05971495 0.08063101
##       3034       3182       3269       3301       3305       3306
## 0.09665883 0.06748451 0.06096966 0.02838702 0.02780825 0.10188035
##       3320       3361       3445       3578       3727       4012
## 0.02780825 0.04474765 0.06702912 0.03549112 0.03177241 0.03549843
##       4573       4652       4673
## 0.12263517 0.06482868 0.06309915
```

```r
plot(rownames(taxi_data),lev, main = "Leverage in Taxi Dataset", xlab=
"observation",ylab = "Leverage Value")
abline(h = 2 *p/n, lty = 1)
abline(h = 3 *p/n, lty = 1)
```



```r
taxi_data_wo = taxi_data[-as.numeric(rownames(data.frame(outlier))),]

taxi_lm_red_int_red_high_wo = lm ( fare ~ factor(company) +
factor(time_of_day) +factor(hour_type)+
                                   avg_miles*factor(company)+
factor(company)*factor(hour_type)+
                                   avg_miles*factor(time_of_day)+
factor(time_of_day)*factor(hour_type)+
                                   poly(avg_miles,degree= 12, raw =TRUE), data
= taxi_data_wo)
summary(taxi_lm_red_int_red_high_wo)

##
## Call:
## lm(formula = fare ~ factor(company) + factor(time_of_day) +
factor(hour_type) +
##     avg_miles * factor(company) + factor(company) * factor(hour_type) +
##     avg_miles * factor(time_of_day) + factor(time_of_day) *
factor(hour_type) +
##     poly(avg_miles, degree = 12, raw = TRUE), data = taxi_data_wo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.465  -2.086  -0.402   1.692 102.600
##
## Coefficients: (2 not defined because of singularities)
##                                                      Estimate
## (Intercept)                                        -2.000e+02
## factor(company)107                                 -4.624e-01
## factor(company)109                                 -1.690e-01
## factor(time_of_day)Evening                          1.864e-01
## factor(time_of_day)Morning                          2.720e-01
## factor(time_of_day)Night                           -1.952e-01
## factor(hour_type)rush_hour                          3.563e-01
## avg_miles                                           5.495e+02
## poly(avg_miles, degree = 12, raw = TRUE)1                  NA
## poly(avg_miles, degree = 12, raw = TRUE)2          -6.246e+02
## poly(avg_miles, degree = 12, raw = TRUE)3           4.029e+02
## poly(avg_miles, degree = 12, raw = TRUE)4          -1.643e+02
## poly(avg_miles, degree = 12, raw = TRUE)5           4.478e+01
## poly(avg_miles, degree = 12, raw = TRUE)6          -8.403e+00
## poly(avg_miles, degree = 12, raw = TRUE)7           1.098e+00
## poly(avg_miles, degree = 12, raw = TRUE)8          -9.954e-02
## poly(avg_miles, degree = 12, raw = TRUE)9           6.126e-03
## poly(avg_miles, degree = 12, raw = TRUE)10         -2.438e-04
## poly(avg_miles, degree = 12, raw = TRUE)11          5.650e-06
## poly(avg_miles, degree = 12, raw = TRUE)12         -5.785e-08
## factor(company)107:avg_miles                        1.772e-01
## factor(company)109:avg_miles                        5.336e-02
## factor(company)107:factor(hour_type)rush_hour       2.117e-01
```

```
## factor(company)109:factor(hour_type)rush_hour          5.680e-01
## factor(time_of_day)Evening:avg_miles                  -1.608e-01
## factor(time_of_day)Morning:avg_miles                  -5.655e-02
## factor(time_of_day)Night:avg_miles                    -9.415e-02
## factor(time_of_day)Evening:factor(hour_type)rush_hour  1.051e+00
## factor(time_of_day)Morning:factor(hour_type)rush_hour -8.930e-01
## factor(time_of_day)Night:factor(hour_type)rush_hour           NA
##                                                      Std. Error t value
## (Intercept)                                           1.053e+02  -1.900
## factor(company)107                                    2.671e-01  -1.731
## factor(company)109                                    3.561e-01  -0.475
## factor(time_of_day)Evening                            3.115e-01   0.599
## factor(time_of_day)Morning                            3.810e-01   0.714
## factor(time_of_day)Night                              3.548e-01  -0.550
## factor(hour_type)rush_hour                            3.667e-01   0.972
## avg_miles                                             2.705e+02   2.031
## poly(avg_miles, degree = 12, raw = TRUE)1                    NA      NA
## poly(avg_miles, degree = 12, raw = TRUE)2             2.977e+02  -2.098
## poly(avg_miles, degree = 12, raw = TRUE)3             1.860e+02   2.167
## poly(avg_miles, degree = 12, raw = TRUE)4             7.369e+01  -2.229
## poly(avg_miles, degree = 12, raw = TRUE)5             1.959e+01   2.285
## poly(avg_miles, degree = 12, raw = TRUE)6             3.601e+00  -2.334
## poly(avg_miles, degree = 12, raw = TRUE)7             4.629e-01   2.372
## poly(avg_miles, degree = 12, raw = TRUE)8             4.148e-02  -2.399
## poly(avg_miles, degree = 12, raw = TRUE)9             2.538e-03   2.414
## poly(avg_miles, degree = 12, raw = TRUE)10            1.009e-04  -2.415
## poly(avg_miles, degree = 12, raw = TRUE)11            2.352e-06   2.402
## poly(avg_miles, degree = 12, raw = TRUE)12            2.435e-08  -2.376
## factor(company)107:avg_miles                          3.403e-02   5.208
## factor(company)109:avg_miles                          4.261e-02   1.252
## factor(company)107:factor(hour_type)rush_hour         3.709e-01   0.571
## factor(company)109:factor(hour_type)rush_hour         4.589e-01   1.238
## factor(time_of_day)Evening:avg_miles                  3.561e-02  -4.516
## factor(time_of_day)Morning:avg_miles                  4.197e-02  -1.347
## factor(time_of_day)Night:avg_miles                    4.950e-02  -1.902
## factor(time_of_day)Evening:factor(hour_type)rush_hour  3.755e-01   2.799
## factor(time_of_day)Morning:factor(hour_type)rush_hour  4.237e-01  -2.108
## factor(time_of_day)Night:factor(hour_type)rush_hour           NA      NA
##                                                      Pr(>|t|)
## (Intercept)                                           0.05747 .
## factor(company)107                                    0.08346 .
## factor(company)109                                    0.63508
## factor(time_of_day)Evening                            0.54952
## factor(time_of_day)Morning                            0.47536
## factor(time_of_day)Night                              0.58222
## factor(hour_type)rush_hour                            0.33134
## avg_miles                                             0.04227 *
## poly(avg_miles, degree = 12, raw = TRUE)1                  NA
## poly(avg_miles, degree = 12, raw = TRUE)2             0.03593 *
## poly(avg_miles, degree = 12, raw = TRUE)3             0.03030 *
```

```
## poly(avg_miles, degree = 12, raw = TRUE)4               0.02585 *
## poly(avg_miles, degree = 12, raw = TRUE)5               0.02234 *
## poly(avg_miles, degree = 12, raw = TRUE)6               0.01966 *
## poly(avg_miles, degree = 12, raw = TRUE)7               0.01772 *
## poly(avg_miles, degree = 12, raw = TRUE)8               0.01646 *
## poly(avg_miles, degree = 12, raw = TRUE)9               0.01582 *
## poly(avg_miles, degree = 12, raw = TRUE)10              0.01577 *
## poly(avg_miles, degree = 12, raw = TRUE)11              0.01633 *
## poly(avg_miles, degree = 12, raw = TRUE)12              0.01755 *
## factor(company)107:avg_miles                           1.99e-07 ***
## factor(company)109:avg_miles                            0.21054
## factor(company)107:factor(hour_type)rush_hour           0.56825
## factor(company)109:factor(hour_type)rush_hour           0.21585
## factor(time_of_day)Evening:avg_miles                   6.46e-06 ***
## factor(time_of_day)Morning:avg_miles                    0.17793
## factor(time_of_day)Night:avg_miles                      0.05724 .
## factor(time_of_day)Evening:factor(hour_type)rush_hour   0.00514 **
## factor(time_of_day)Morning:factor(hour_type)rush_hour   0.03511 *
## factor(time_of_day)Night:factor(hour_type)rush_hour          NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.443 on 4849 degrees of freedom
## Multiple R-squared:  0.8766, Adjusted R-squared:  0.8759
## F-statistic:  1276 on 27 and 4849 DF,  p-value: < 2.2e-16
```

```
taxi_data[cooks.distance(taxi_lm_red_int_red_high_wo)>0.5,]
```

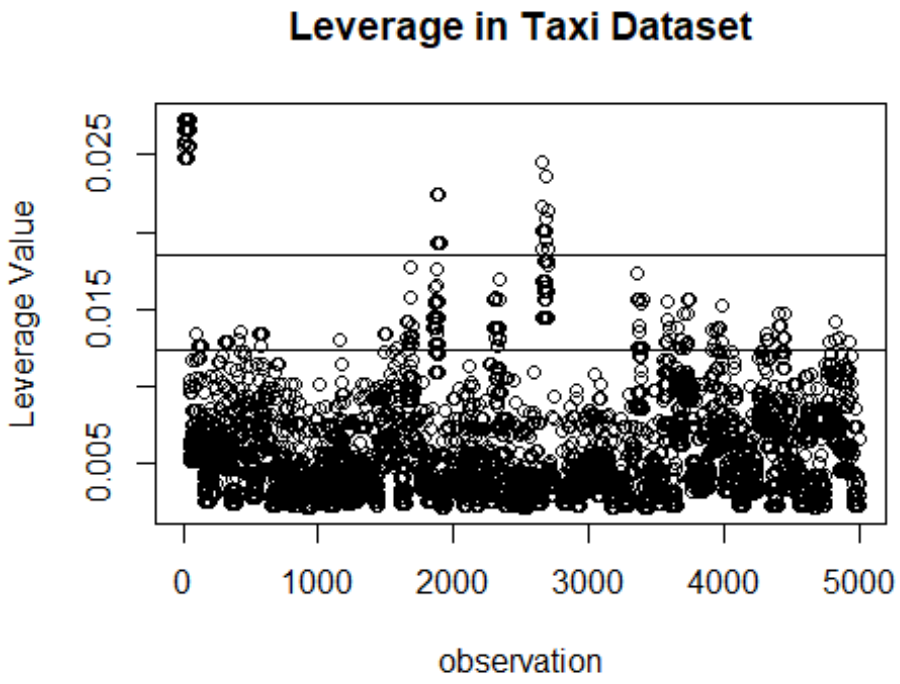
```
##  [1] X                    pickup_area          dropoff_area
##  [4] trip_miles           trip_seconds         fare
##  [7] trip_start_timestamp tips                 tolls
## [10] trip_total           payment_type         company
## [13] extras               pickup_dropoff       avg_miles
## [16] avg_minutes          hours                months
## [19] day_of_week          hour_type            tip_pct
## [22] tip_type             pickup_dropoff_dummy weekend
## [25] season               time_of_day
## <0 rows> (or 0-length row.names)
```

```
lev_wo=hatvalues(taxi_lm_red_int_red_high_wo)
p = length(coef(taxi_lm_red_int_red_high_wo))
n = nrow(taxi_data_wo)
outlier = lev[lev>(3*p/n)]
print(outlier)
```

```
##         2         3         4         6         7         8
## 0.01980660 0.02043110 0.02091317 0.01967040 0.01888571 0.01933299
##         9        11        12        13        16        17
## 0.02073254 0.01872634 0.01980658 0.01980662 0.01967039 0.02520953
##        18        20        21        22        23        24
## 0.01934181 0.01888578 0.02073215 0.01872636 0.01934181 0.02005443
```

```
##          25          27          28          29          30          31
## 0.01934181 0.02073215 0.02249231 0.02091324 0.02005443 0.02073215
##          32          33          35          36          38          39
## 0.01934181 0.01888578 0.01980661 0.01967039 0.01934181 0.01933299
##          40          45          46          47          49          50
## 0.01888578 0.01987216 0.01967039 0.02073215 0.01933299 0.02091324
##         263         288         300         401         408         419
## 0.03194823 0.03194823 0.03284992 0.11024201 0.03475214 0.11149194
##         436         437         528         609         613         615
## 0.03599579 0.03343886 0.05782265 0.10799810 0.03136067 0.03136067
##        1081        1293        1525        1555        1562        1571
## 0.07420049 0.08913463 0.02887110 0.03113004 0.03322079 0.03180329
##        1578        1586        1599        1862        1867        1881
## 0.03180329 0.03113004 0.11138208 0.04672691 0.04672691 0.01977205
##        1898        1954        2117        2122        2142        2143
## 0.01977205 0.04075505 0.02777517 0.10177129 0.02958916 0.02958916
##        2149        2268        2314        2577        2655        2665
## 0.02777517 0.14325982 0.04132194 0.05003701 0.01876402 0.04817071
##        2692        2707        2711        2722        2725        2733
## 0.02004096 0.04486946 0.04573375 0.04573375 0.10613737 0.10613737
##        2751        2752        2753        2754        2755        2756
## 0.02529740 0.02455999 0.02529740 0.02529740 0.02446378 0.02281584
##        2757        2758        2759        2760        2761        2762
## 0.02515108 0.02446378 0.02281584 0.02515108 0.02281584 0.02446378
##        2763        2764        2765        2766        2767        2768
## 0.02446378 0.02515108 0.02851662 0.03037770 0.02446378 0.02918939
##        2769        2770        2771        2772        2773        2774
## 0.02567763 0.02755828 0.03037770 0.02479757 0.02263362 0.02479757
##        2775        2776        2777        2778        2779        2780
## 0.02281584 0.02567763 0.02918939 0.02281584 0.02263362 0.02769595
##        2781        2782        2783        2784        2785        2786
## 0.02413846 0.02263362 0.02263362 0.02918939 0.06966335 0.02446378
##        2787        2788        2789        2790        2791        2792
## 0.02515108 0.05971495 0.02479757 0.02263362 0.02413846 0.02281584
##        2793        2794        2795        2796        2797        2798
## 0.02446378 0.02755828 0.02529740 0.02413846 0.02263362 0.03037770
##        2799        2800        2901        3034        3182        3269
## 0.02479757 0.05971495 0.08063101 0.09665883 0.06748451 0.06096966
##        3301        3305        3306        3320        3361        3445
## 0.02838702 0.02780825 0.10188035 0.02780825 0.04474765 0.06702912
##        3578        3727        4012        4573        4652        4673
## 0.03549112 0.03177241 0.03549843 0.12263517 0.06482868 0.06309915
```

```
plot(rownames(taxi_data_wo),lev_wo, main = "Leverage in Taxi Dataset", xlab=
"observation",ylab = "Leverage Value")
abline(h = 2 *p/n, lty = 1)
abline(h = 3 *p/n, lty = 1)
```


Leverage in Taxi Dataset
Leverage Value
0.005
0.015
0.025
0
1000
2000
3000
4000
5000
observation