Group Number – 128: Airline Delay

First Name	Last Name	Monday or Tuesday class
Ayush	Goyal	Online(Tuesday)
Puneet	Ojha	Online (Tuesday)

Table of Contents

1. Introduction

The aim of this project is to perform Multiple linear regression analysis on the data of airline delay. Multiple linear regression is considered as one of the perfect technique in data analytics, and for our future prediction we are driven towards multiple linear regression. Analysis on flight delay can be done based on the features such as Origin Airport, Destination Airport, weather, days (like – holidays, weekdays), distance, and elapsed time. With the given data and statistical technique, we will predict the future outcomes.

2. Data

Our data set is regarding the airline delay with the following attributes which are listed below. We have obtained are data from United States department of Transportation. Data contain information about the flights in United States.
The dataset has 13524 records.
There are 16 attributes in total

Dataset Attributes:

1- year- year for which data is recorded
2- month- month for which data is recorded
3- Carrier (code of the airlines)
> There are Total 12 airline carrier in the data such as American Airline , Alaska Airline, Southwest Airline with code name AA, AS, SW respectively.
4- carrier_name (full name of the airlines)
5- Airport (code of the airport)
> Code of all the airport name available in the data such as, DFW, DTW, JFK, ORD etc.
6- airport_name (full name of the airport)
7- arr_flights (number of flights arriving)
> Number of flights arrived on a respective airport of a respective airline in a specific month.
8- arr_canceled (number of flights cancelled)
> Number of flights cancelled on a respective airport of a respective airline in a specific month.
9- arr_diverted (number of flights diverted)
> Number of flights diverted on a respective airport of a respective airline in a specific month.
10- arr_delay (total number of minutes flights delayed)
> Number of flights delayed due to whatsoever reason on a respective airport of a respective airline in a specific month.
11- carrier_delay (number of minutes flights delayed because of factors involving the carrier)
12- weather delay (number of minutes flights delayed because of weather)
> Number of flights got delayed due to bad weather condition.
13- Nas_delay (number of minutes flights delayed because of nas conditions)
14- security_delay (number of minutes flights delayed because of security issues)
> Number of flights delayed due to the security issues on the airport.
15- late_aircraft_delay (number of minutes delayed flights which are further delayed)
> Number of flights which are already late but again get delayed due so some other reasons.
16- arr_delay- number of flights delayed at specific airport.

URL Link for the data

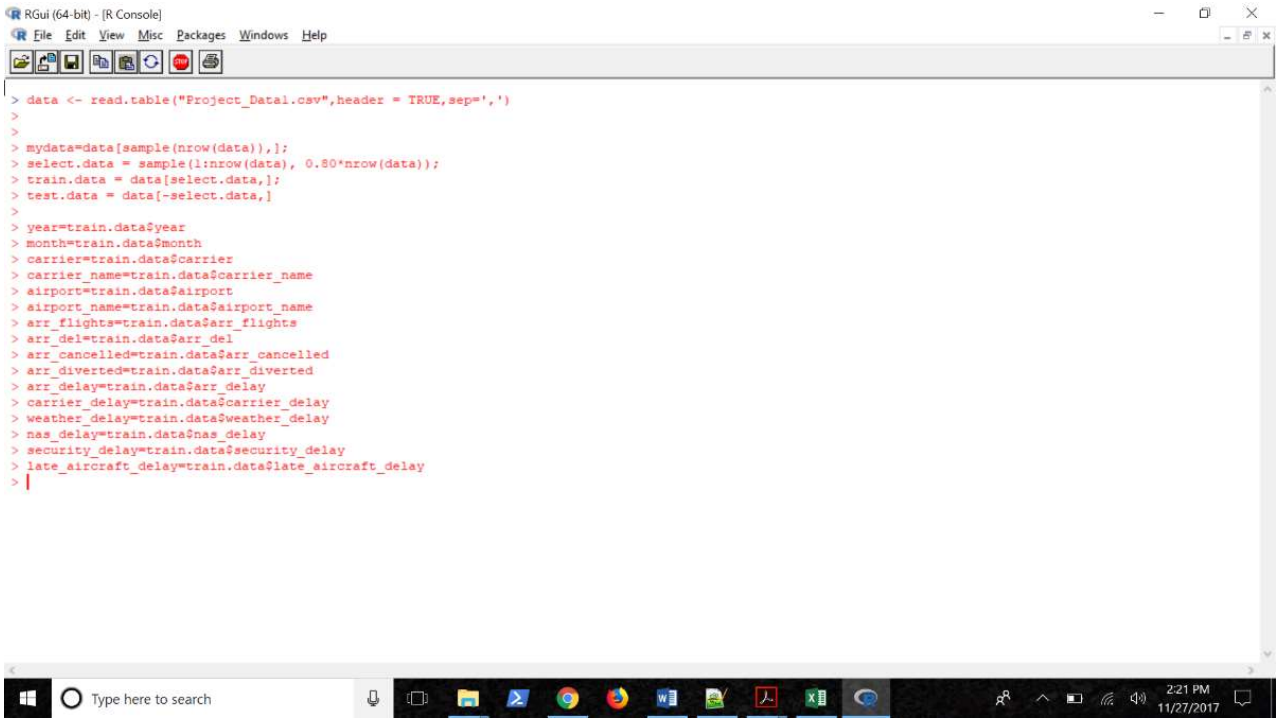
https://www.transtats.bts.gov/OT_Delay/ot_delaycause1.asp?display=data&pn=1

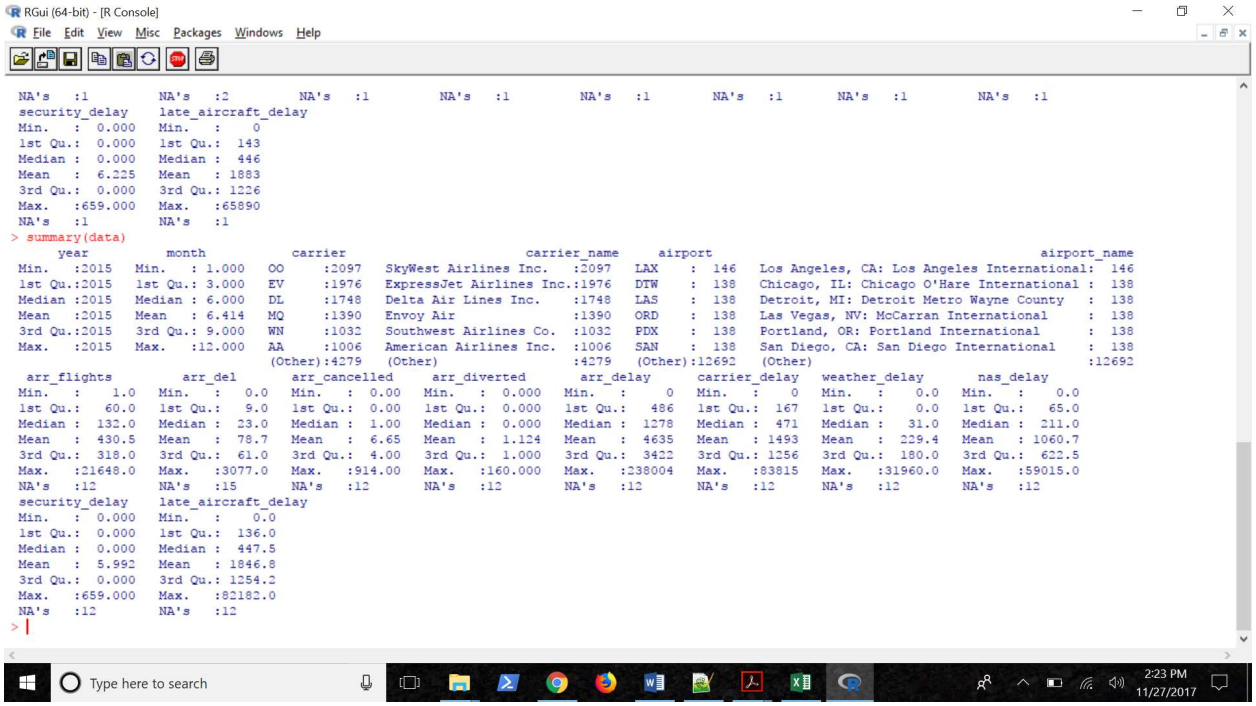
3. Problems to be solved

Some of the problem which we are trying to solve are :
Best fit model
Interpreting different plots to understand the relationship between the attributes.
What is the correlation between the attributes.
Airlines which has the maximum and minimum delay for a specific month.
On a particular airport what is the average delay for a airline.
Solutions:
Multiple linear regressions is being used to find the best fit model
Ggplots and the scaterplots were used.
DPLYR- function is used for the deep analysis of the data. From this function we have calculated min, max average delay of the flights. Also we have used it for some further analysis.

4. Data Processing

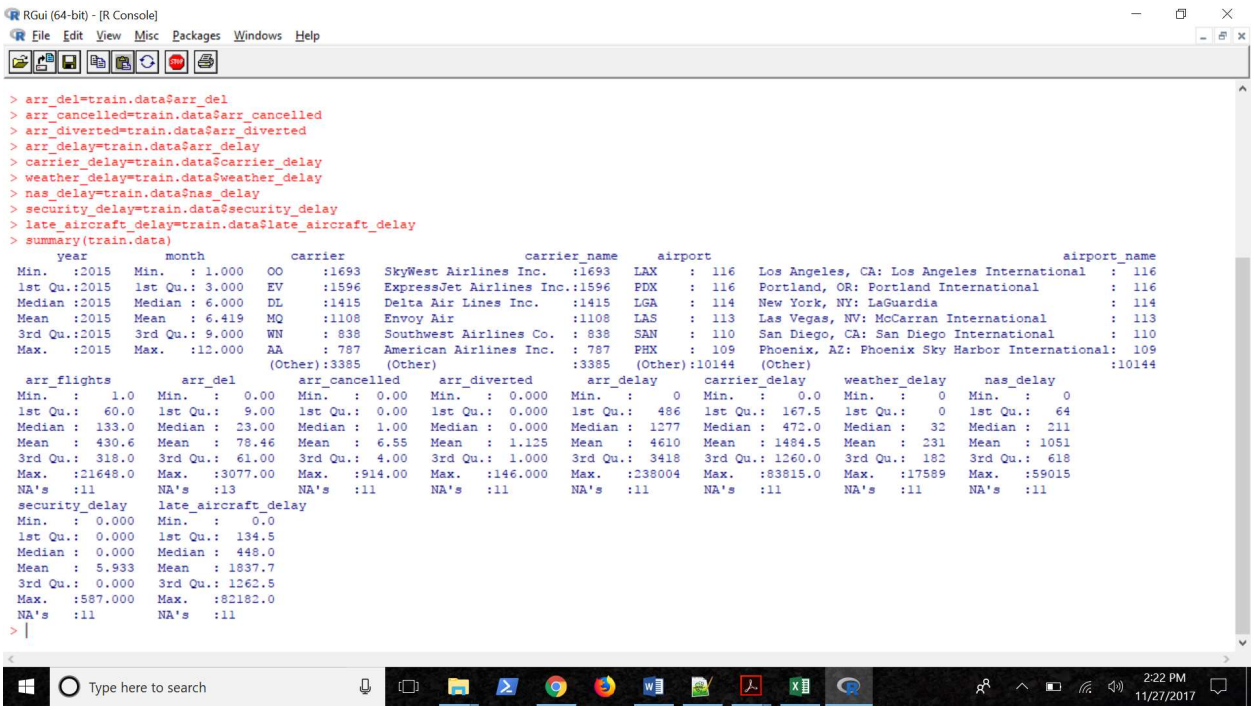
We first, import data into R. then defining th attributes and then we clean the data by using na.omit function. It removes the na values from the data.





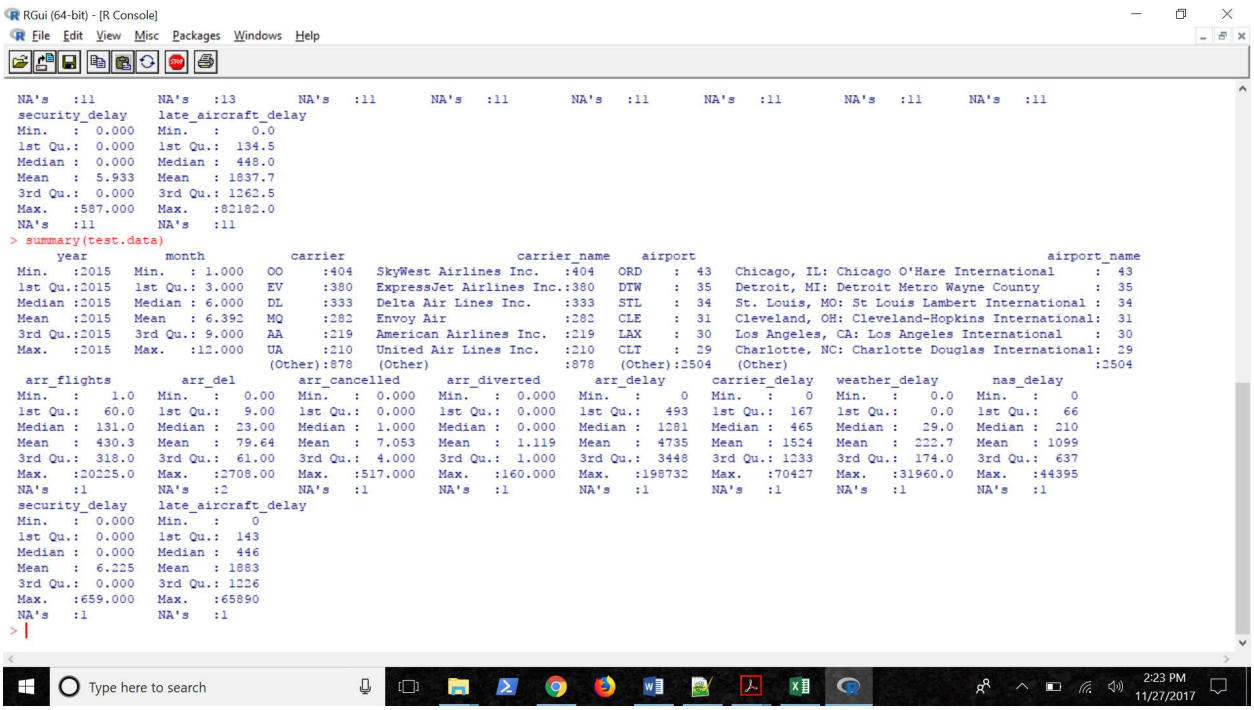
Removing the na values

Train Dataset



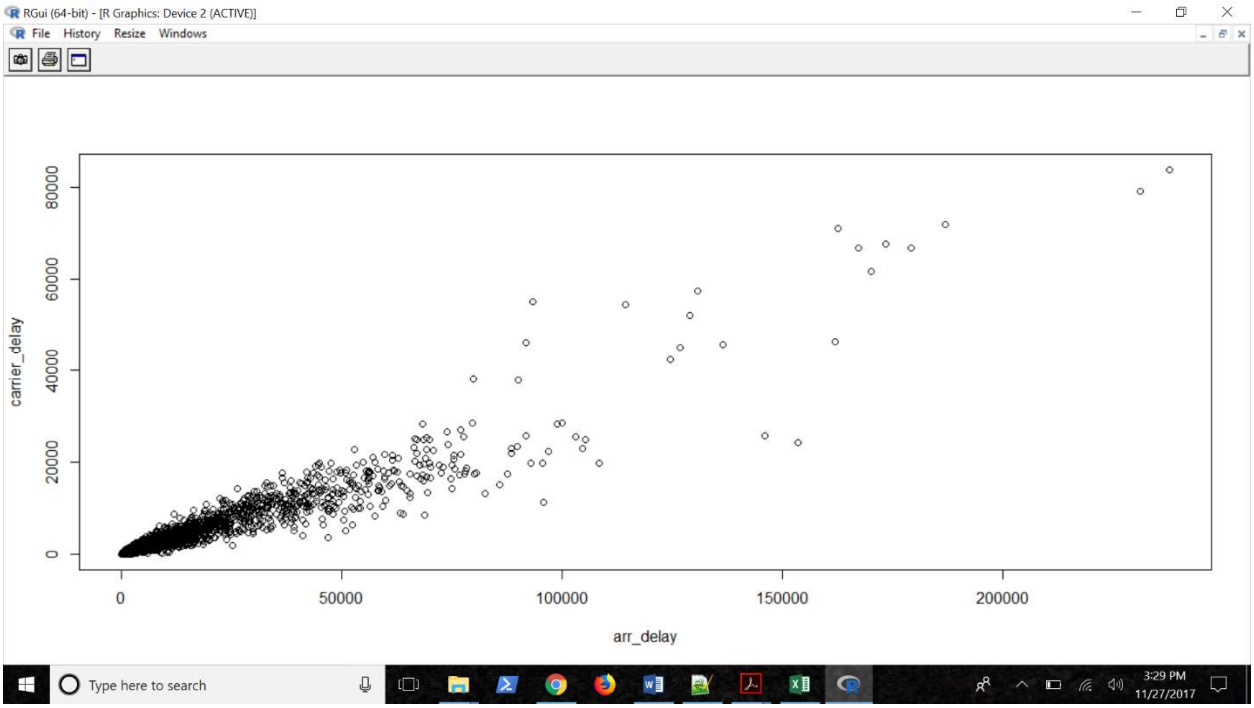
```
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

> arr_del=train.data$arr_del
> arr_cancelled=train.data$arr_cancelled
> arr_diverted=train.data$arr_diverted
> arr_delay=train.data$arr_delay
> carrier_delay=train.data$carrier_delay
> weather_delay=train.data$weather_delay
> nas_delay=train.data$nas_delay
> security_delay=train.data$security_delay
> late_aircraft_delay=train.data$late_aircraft_delay
> summary(train.data)
      year          month          carrier                     carrier_name     airport                                                airport_name
 Min.   :2015   Min.   : 1.000   OO     :1693   SkyWest Airlines Inc.   :1693   LAX    :  116   Los Angeles, CA: Los Angeles International   :  116
 1st Qu.:2015   1st Qu.: 3.000   EV     :1596   ExpressJet Airlines Inc.:1596   PDX    :  116   Portland, OR: Portland International         :  116
 Median :2015   Median : 6.000   DL     :1415   Delta Air Lines Inc.    :1415   LGA    :  114   New York, NY: LaGuardia                      :  114
 Mean   :2015   Mean   : 6.419   MQ     :1108   Envoy Air               :1108   LAS    :  113   Las Vegas, NV: McCarran International        :  113
 3rd Qu.:2015   3rd Qu.: 9.000   WN     : 838   Southwest Airlines Co.  : 838   SAN    :  110   San Diego, CA: San Diego International       :  110
 Max.   :2015   Max.   :12.000   AA     : 787   American Airlines Inc.  : 787   PHX    :  109   Phoenix, AZ: Phoenix Sky Harbor International:  109
                                 (Other):3385   (Other)                 :3385   (Other):10144   (Other)                                      :10144
  arr_flights         arr_del          arr_cancelled      arr_diverted        arr_delay        carrier_delay     weather_delay      nas_delay
 Min.   :    1.0   Min.   :   0.00   Min.   : 0.00   Min.   : 0.000   Min.   :     0   Min.   :    0.0   Min.   :    0   Min.   :    0
 1st Qu.:   60.0   1st Qu.:   9.00   1st Qu.: 0.00   1st Qu.: 0.000   1st Qu.:   486   1st Qu.:  167.5   1st Qu.:    0   1st Qu.:   64
 Median :  133.0   Median :  23.00   Median : 1.00   Median : 0.000   Median :  1277   Median :  472.0   Median :   32   Median :  211
 Mean   :  430.6   Mean   :  78.46   Mean   : 6.55   Mean   : 1.125   Mean   :  4610   Mean   : 1484.5   Mean   :  231   Mean   : 1051
 3rd Qu.:  318.0   3rd Qu.:  61.00   3rd Qu.: 4.00   3rd Qu.: 1.000   3rd Qu.:  3418   3rd Qu.: 1260.0   3rd Qu.:  182   3rd Qu.:  618
 Max.   :21648.0   Max.   :3077.00   Max.   :914.00   Max.   :146.000   Max.   :238004   Max.   :83815.0   Max.   :17589   Max.   :59015
 NA's   :11        NA's   :13        NA's   :11       NA's   :11        NA's   :11       NA's   :11        NA's   :11      NA's   :11
 security_delay    late_aircraft_delay
 Min.   :  0.000   Min.   :    0.0
 1st Qu.:  0.000   1st Qu.:  134.5
 Median :  0.000   Median :  448.0
 Mean   :  5.933   Mean   : 1837.7
 3rd Qu.:  0.000   3rd Qu.: 1262.5
 Max.   :587.000   Max.   :82182.0
 NA's   :11        NA's   :11
>
```

Test Dataset



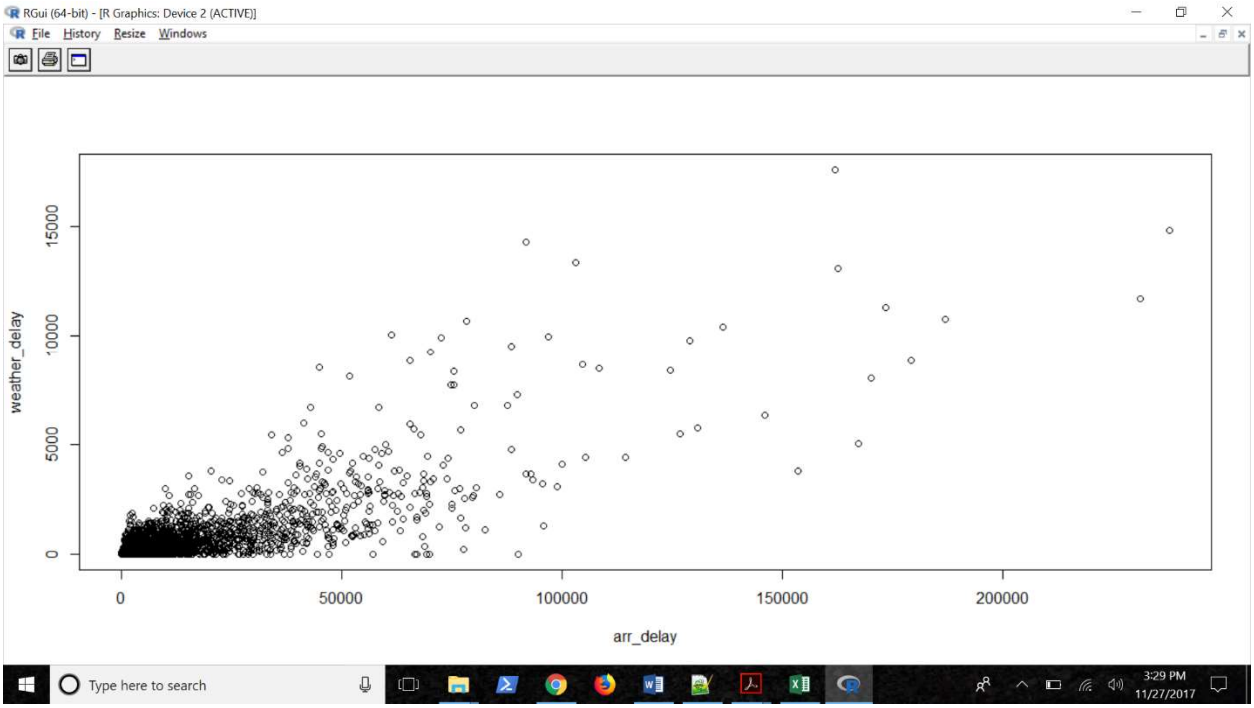
```
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

 NA's   :11        NA's   :13        NA's   :11       NA's   :11        NA's   :11       NA's   :11        NA's   :11      NA's   :11
 security_delay    late_aircraft_delay
 Min.   :  0.000   Min.   :    0.0
 1st Qu.:  0.000   1st Qu.:  134.5
 Median :  0.000   Median :  448.0
 Mean   :  5.933   Mean   : 1837.7
 3rd Qu.:  0.000   3rd Qu.: 1262.5
 Max.   :587.000   Max.   :82182.0
 NA's   :11        NA's   :11
> summary(test.data)
      year          month          carrier                     carrier_name     airport                                                airport_name
 Min.   :2015   Min.   : 1.000   OO     :404   SkyWest Airlines Inc.   :404   ORD    :  43   Chicago, IL: Chicago O'Hare International     :  43
 1st Qu.:2015   1st Qu.: 3.000   EV     :380   ExpressJet Airlines Inc.:380   DTW    :  35   Detroit, MI: Detroit Metro Wayne County       :  35
 Median :2015   Median : 6.000   DL     :333   Delta Air Lines Inc.    :333   STL    :  34   St. Louis, MO: St Louis Lambert International :  34
 Mean   :2015   Mean   : 6.392   MQ     :282   Envoy Air               :282   CLE    :  31   Cleveland, OH: Cleveland-Hopkins International:  31
 3rd Qu.:2015   3rd Qu.: 9.000   AA     :219   American Airlines Inc.  :219   LAX    :  30   Los Angeles, CA: Los Angeles International    :  30
 Max.   :2015   Max.   :12.000   UA     :210   United Air Lines Inc.   :210   CLT    :  29   Charlotte, NC: Charlotte Douglas International:  29
                                 (Other):878   (Other)                 :878   (Other):2504   (Other)                                      :2504
  arr_flights         arr_del          arr_cancelled      arr_diverted        arr_delay        carrier_delay     weather_delay      nas_delay
 Min.   :    1.0   Min.   :   0.00   Min.   : 0.000   Min.   : 0.000   Min.   :     0   Min.   :    0   Min.   :    0.0   Min.   :    0
 1st Qu.:   60.0   1st Qu.:   9.00   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:   493   1st Qu.:  167   1st Qu.:    0.0   1st Qu.:   66
 Median :  131.0   Median :  23.00   Median : 1.000   Median : 0.000   Median :  1281   Median :  465   Median :   29.0   Median :  210
 Mean   :  430.3   Mean   :  79.64   Mean   : 7.053   Mean   : 1.119   Mean   :  4735   Mean   : 1524   Mean   :  222.7   Mean   : 1099
 3rd Qu.:  318.0   3rd Qu.:  61.00   3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.:  3448   3rd Qu.: 1233   3rd Qu.:  174.0   3rd Qu.:  637
 Max.   :20225.0   Max.   :2708.00   Max.   :517.000   Max.   :160.000   Max.   :198732   Max.   :70427   Max.   :31960.0   Max.   :44395
 NA's   :1         NA's   :2         NA's   :1        NA's   :1         NA's   :1        NA's   :1       NA's   :1         NA's   :1
 security_delay    late_aircraft_delay
 Min.   :  0.000   Min.   :    0
 1st Qu.:  0.000   1st Qu.:  143
 Median :  0.000   Median :  446
 Mean   :  6.225   Mean   : 1883
 3rd Qu.:  0.000   3rd Qu.: 1226
 Max.   :659.000   Max.   :65890
 NA's   :1         NA's   :1
>
```

Plot between arrival delay and carrier delay

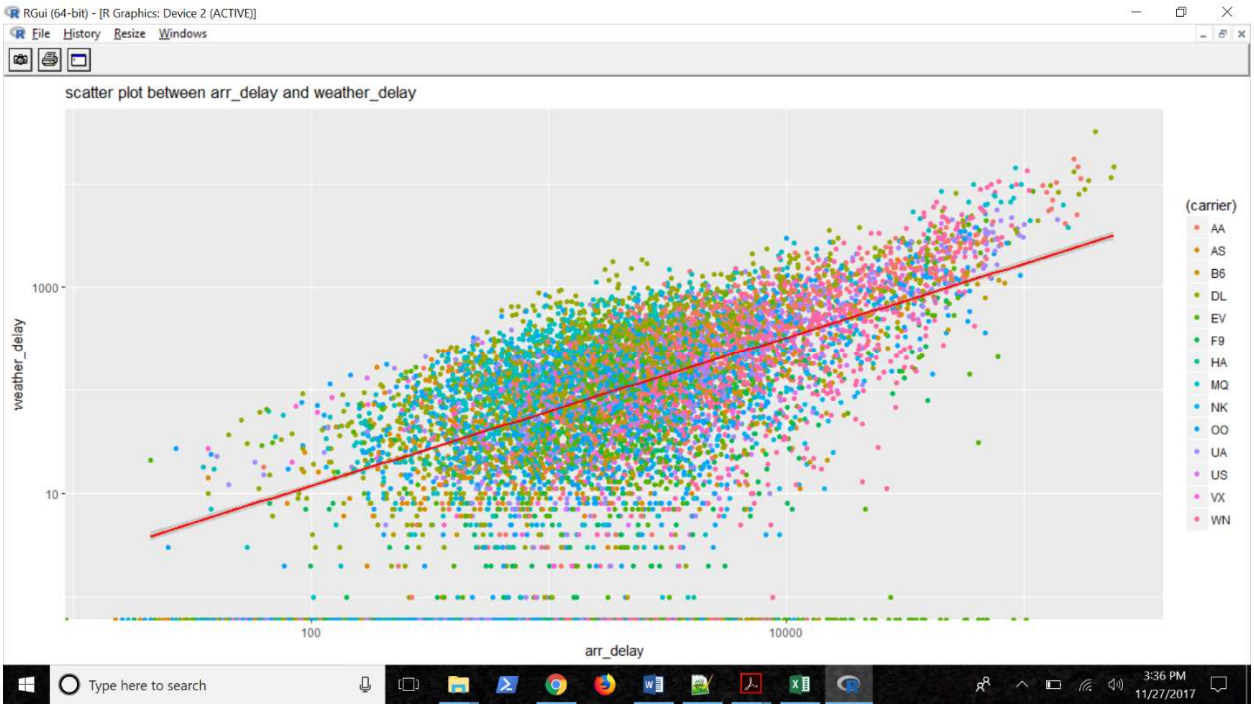


GGplot between arrival delay and Carrier delay

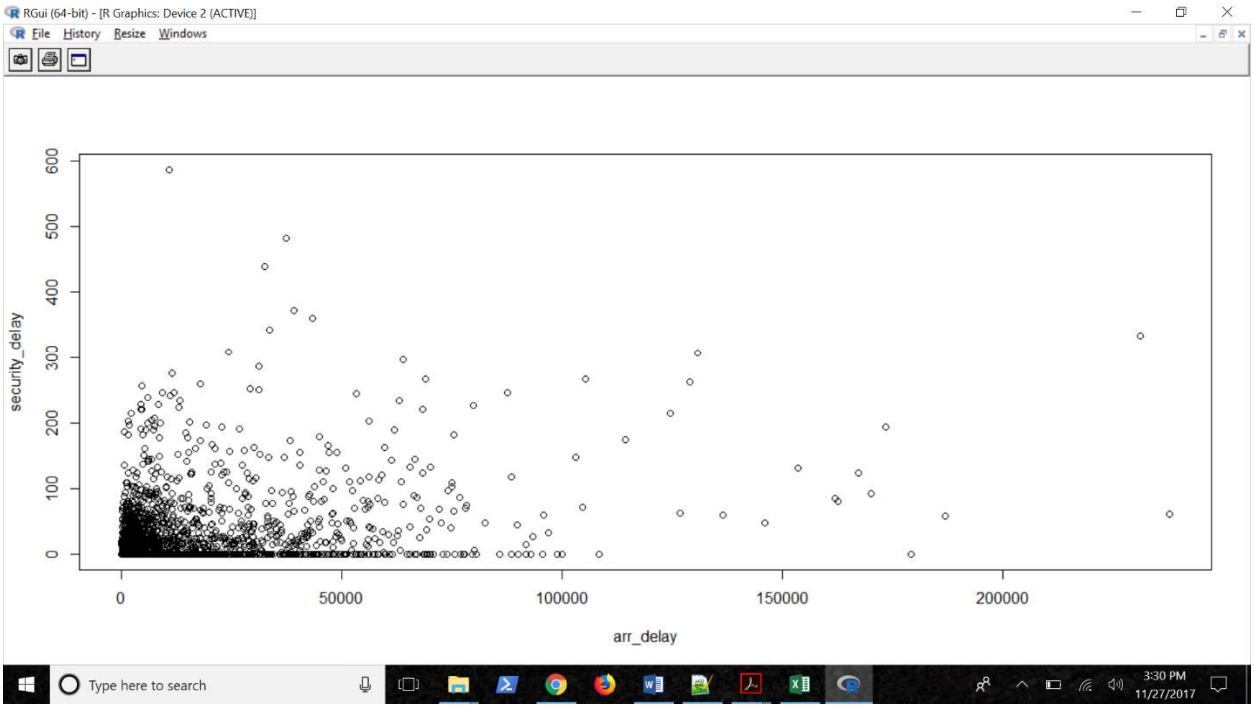
Plot between arrival delay and Weather delay



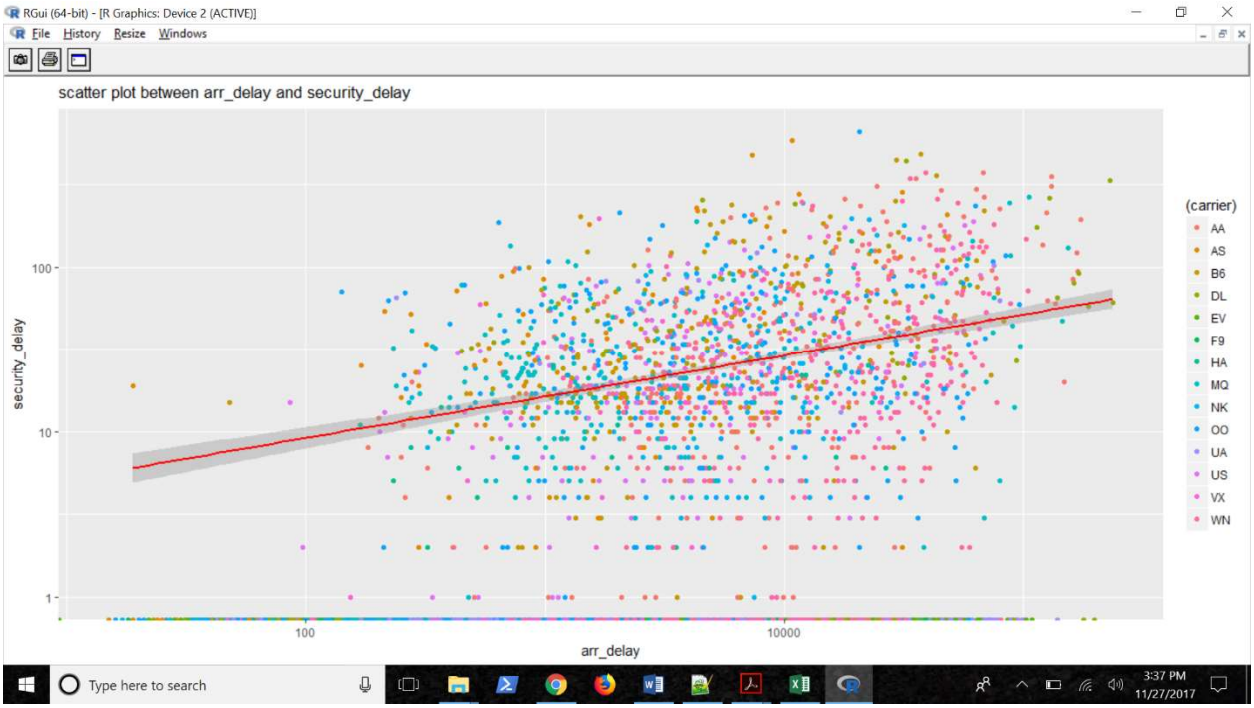
GGplot between arrival delay and weather delay



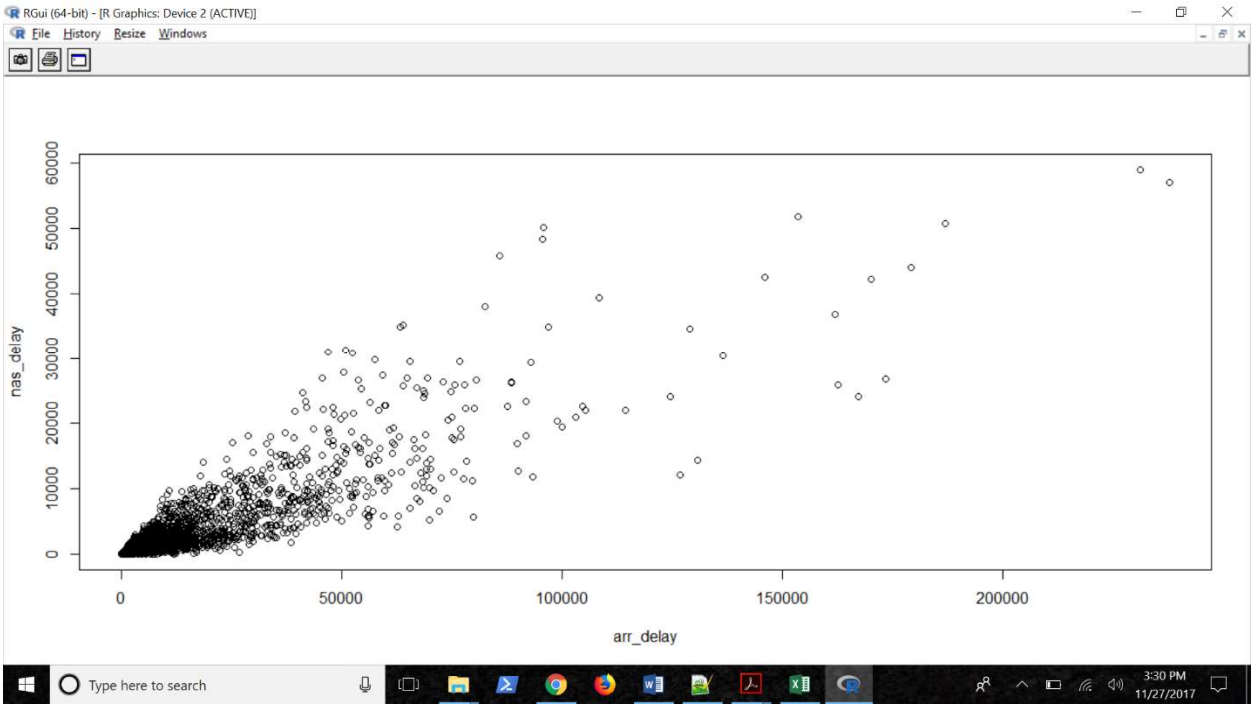
Plot between arrival delay and security delay



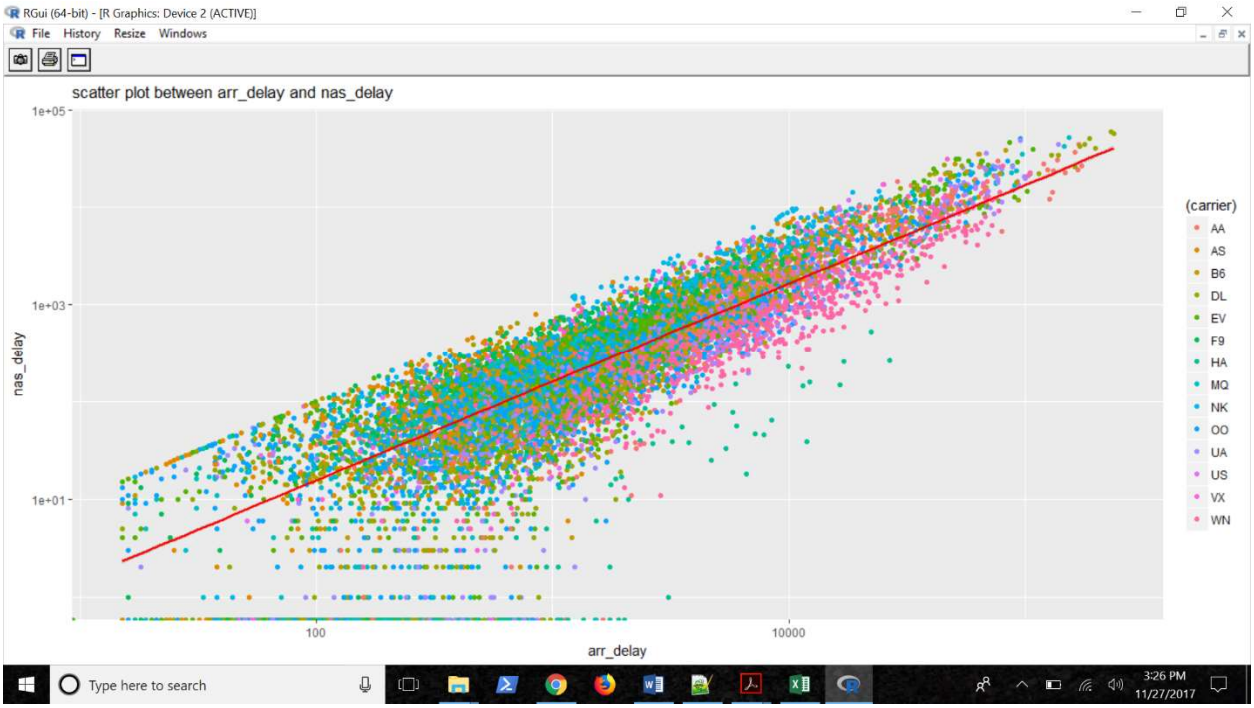
GGplot between arrival delay and security delay



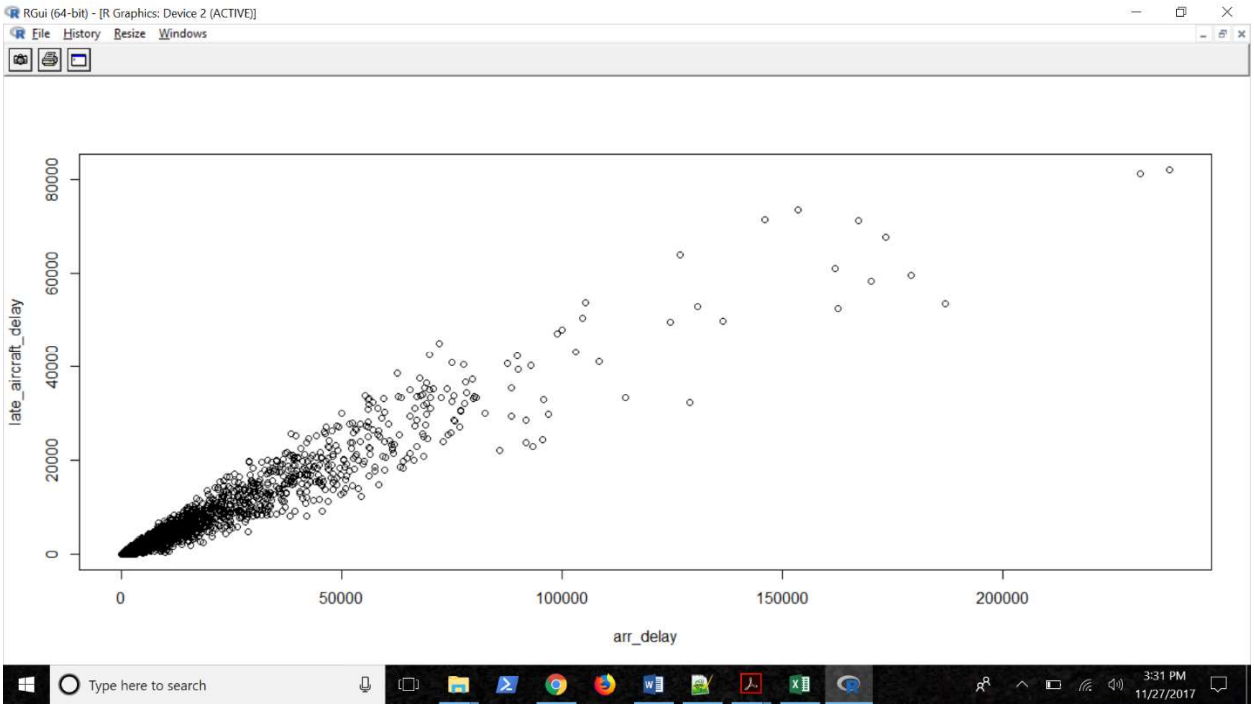
Plot between arrival delay and Nas delay



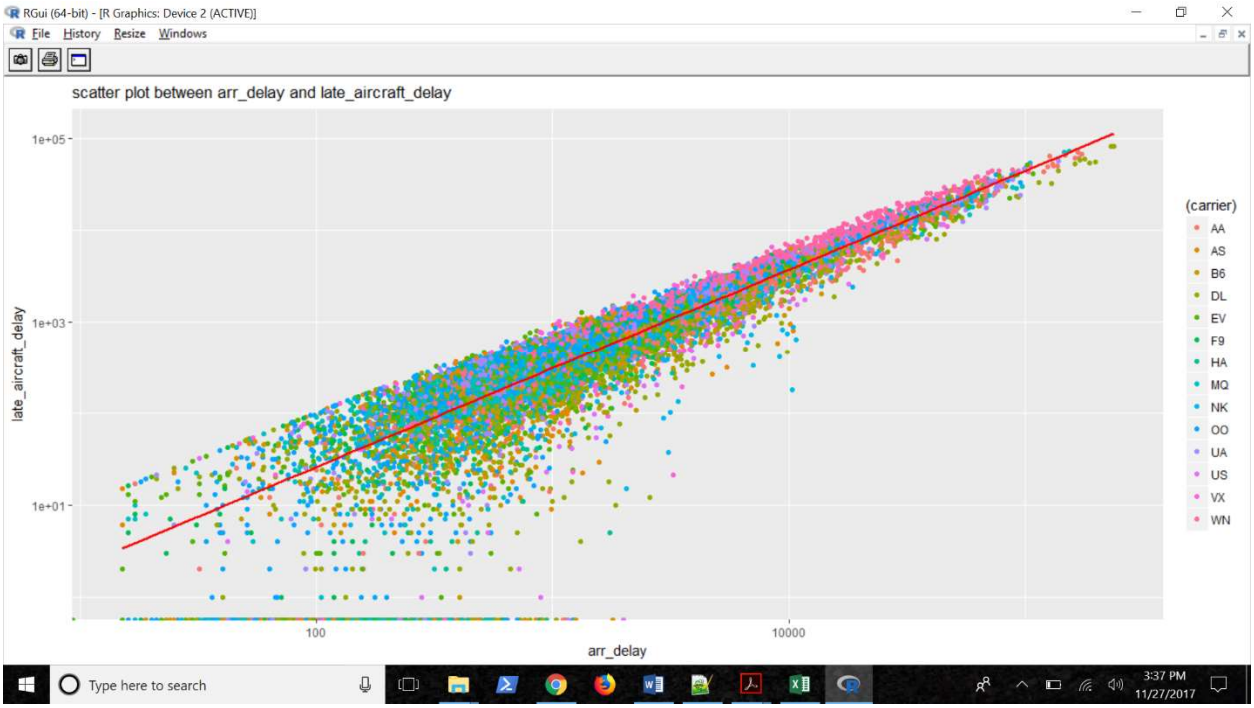
GGplot between arrival delay and nas delay



Plot between arrival delay and late aircraft delay



GGplot between arrival delay and late aircraft delay

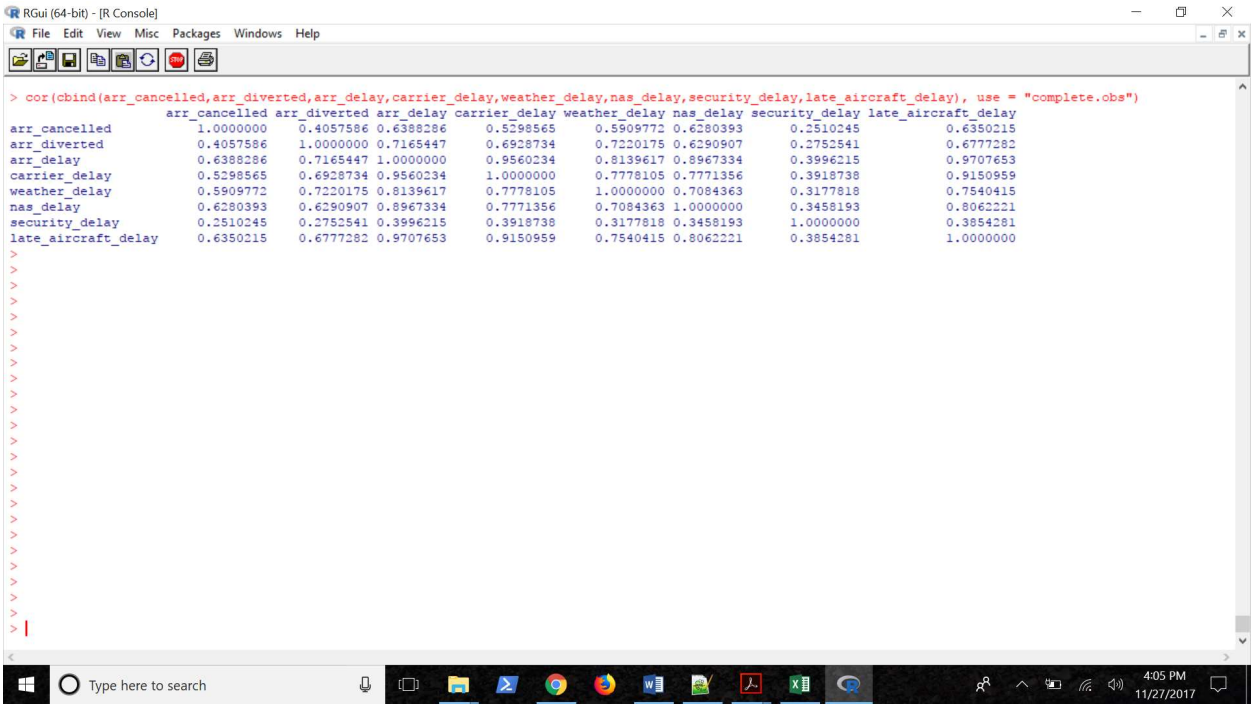


Interpretation of the plots

- From the above graphs we can say that there is linear relationship between the following attributes.
- arrival delay and carrier delay
- arrival delay and weather delay
- arrival delay and nas delay
- arrival delay and late aircraft delay
- From this we can say that with increase in carrier delay, weather delay, nas delay, and late aircraft delay, arrival delay also increases.
- However, in case of arrival delay and Security delay there is no linear relationship, as from the graph data is scattered all over.

5. Methods and Process

Correlation between the attributes



Model

RGui (64-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> model1=lm(arr_cancelled~arr_delay+carrier_delay+weather_delay+nas_delay+late_aircraft_delay)
> summary(model1)

Call:
lm(formula = arr_cancelled ~ arr_delay + carrier_delay + weather_delay +
    nas_delay + late_aircraft_delay)

Residuals:
    Min      1Q  Median      3Q     Max
-220.53   -1.95   -0.56    1.11  717.30

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.5633499  0.1918286   2.937  0.00332 **
arr_delay         0.0030261  0.0075936   0.399  0.69026
carrier_delay    -0.0062830  0.0076073  -0.826  0.40887
weather_delay     0.0079535  0.0075901   1.048  0.29471
nas_delay        -0.0008921  0.0075986  -0.117  0.90655
late_aircraft_delay  0.0003079  0.0075993   0.041  0.96768
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.52 on 10804 degrees of freedom
Multiple R-squared:  0.4932,    Adjusted R-squared:  0.4929
F-statistic:  2103 on 5 and 10804 DF,  p-value: < 2.2e-16
```

RGui (64-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> model2=lm(arr_cancelled~arr_delay+carrier_delay+weather_delay+nas_delay)
> summary(model2)

Call:
lm(formula = arr_cancelled ~ arr_delay + carrier_delay + weather_delay +
    nas_delay)

Residuals:
    Min      1Q  Median      3Q     Max
-220.49   -1.95   -0.56    1.11  717.32

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.627e-01  1.912e-01   2.943  0.00326 **
arr_delay      3.334e-03  9.753e-05  34.184  < 2e-16 ***
carrier_delay -6.591e-03  2.089e-04 -31.555  < 2e-16 ***
weather_delay  7.646e-03  3.924e-04  19.488  < 2e-16 ***
nas_delay     -1.200e-03  1.641e-04  -7.311 2.85e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.52 on 10805 degrees of freedom
Multiple R-squared:  0.4932,    Adjusted R-squared:  0.493
F-statistic:  2629 on 4 and 10805 DF,  p-value: < 2.2e-16
```

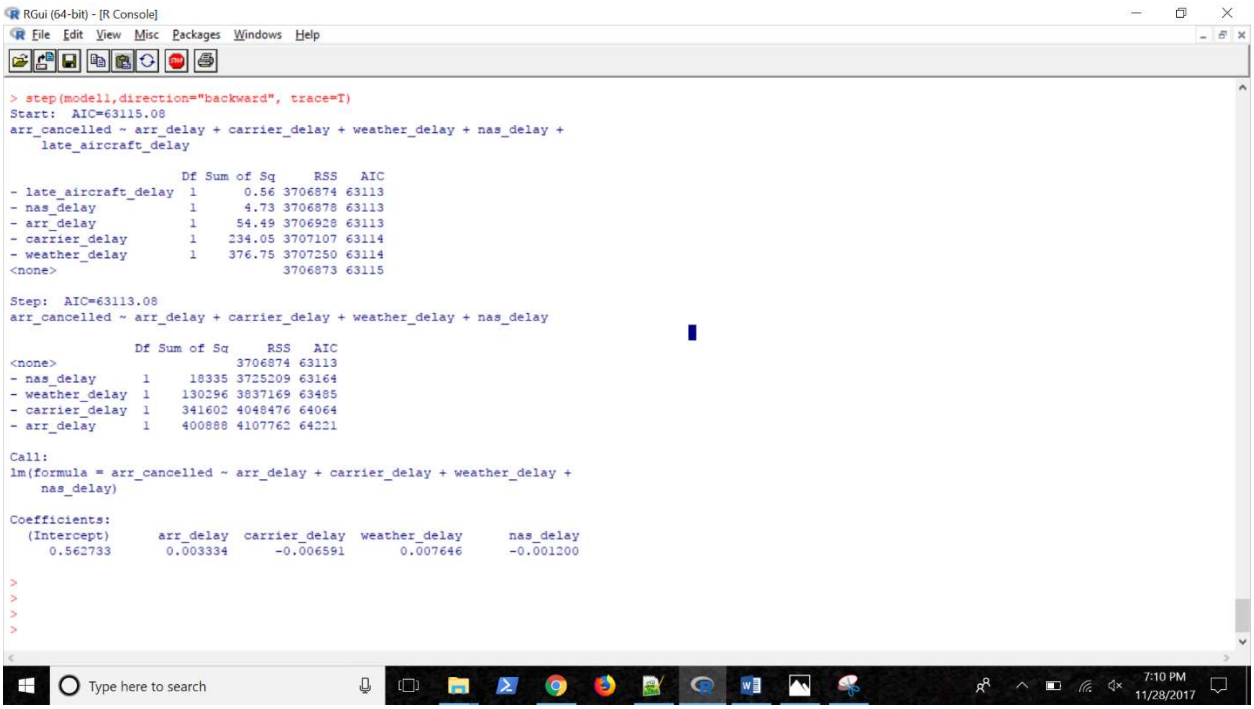
Forward model

RGui (64-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> base=lm(arr_cancelled~arr_delay)
> step(base,scope=list(upper=ml, lower=~1),direction="forward", trace=F)

Call:
lm(formula = arr_cancelled ~ arr_delay + carrier_delay + weather_delay +
    nas_delay)

Coefficients:
   (Intercept)       arr_delay   carrier_delay   weather_delay       nas_delay
      0.562733        0.003334       -0.006591        0.007646       -0.001200

>
>
>
>
>
>
```

Backward model



Both

RGui (64-bit) - [R Console]

File Edit View Misc Packages Windows Help

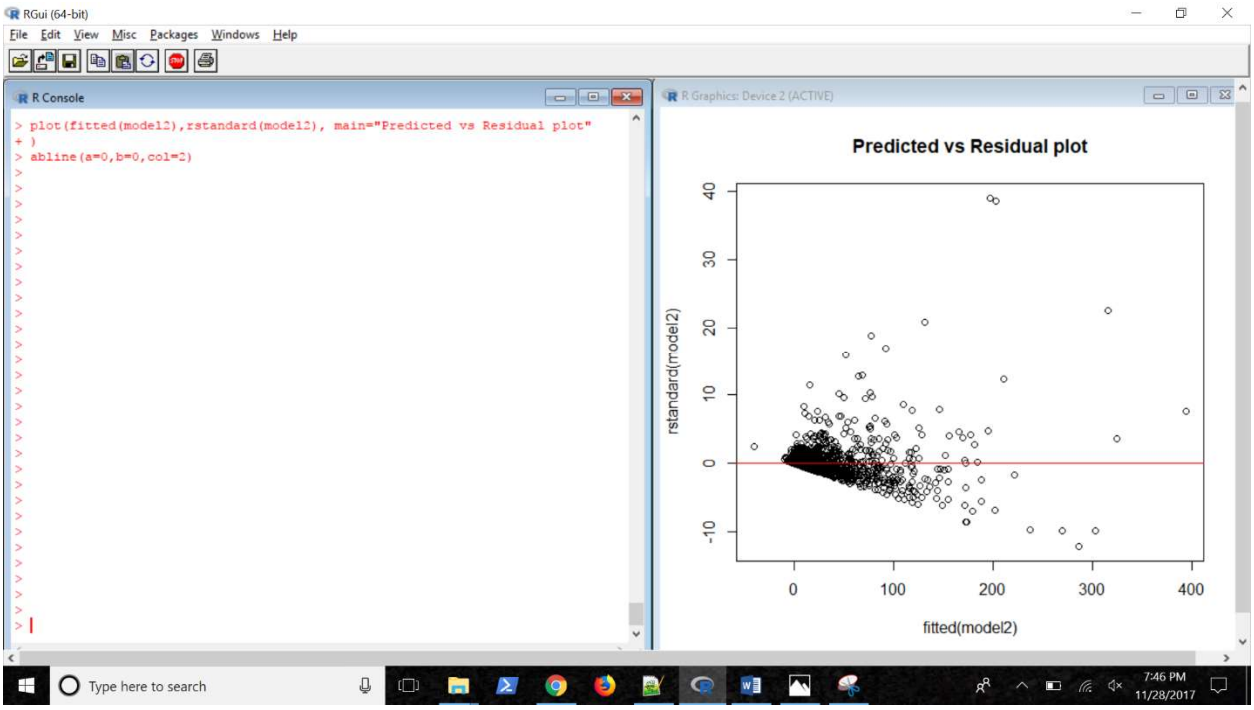
```
> step(base,scope=list(upper=ml, lower=~1),direction="both", trace=F)

Call:
lm(formula = arr_cancelled ~ arr_delay + carrier_delay + weather_delay +
    nas_delay)

Coefficients:
   (Intercept)       arr_delay   carrier_delay   weather_delay       nas_delay
      0.562733        0.003334       -0.006591        0.007646       -0.001200

>
>
>
>
```

From all the above model we get model2 as the best model.

6. Evaluations and Results

6.1. Evaluation Methods

Residual Plots



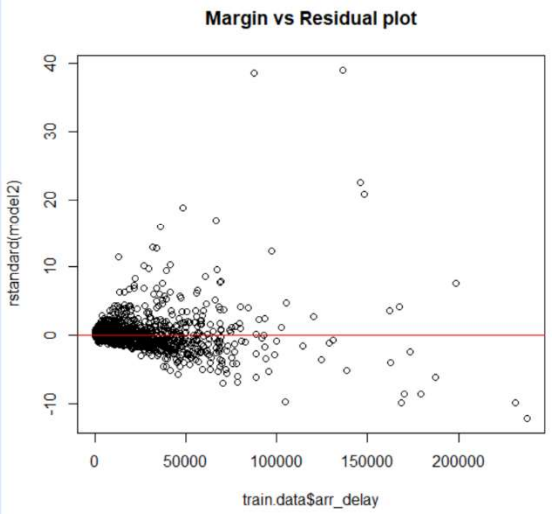
RGui (64-bit)

File Edit View Misc Packages Windows Help

R Console

```
> plot(train.data$arr_delay,rstandard(model2), main="Margin vs Residual plot")
> abline(a=0,b=0,col=2)
>
```

R Graphics: Device 2 (ACTIVE)



Type here to search

7:59 PM
11/28/2017

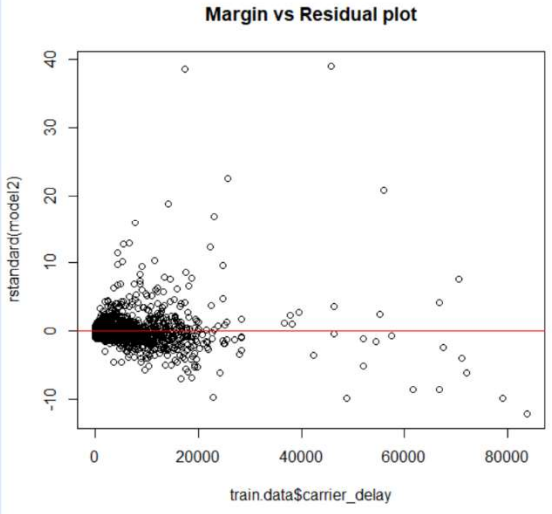
RGui (64-bit)

File Edit View Misc Packages Windows Help

R Console

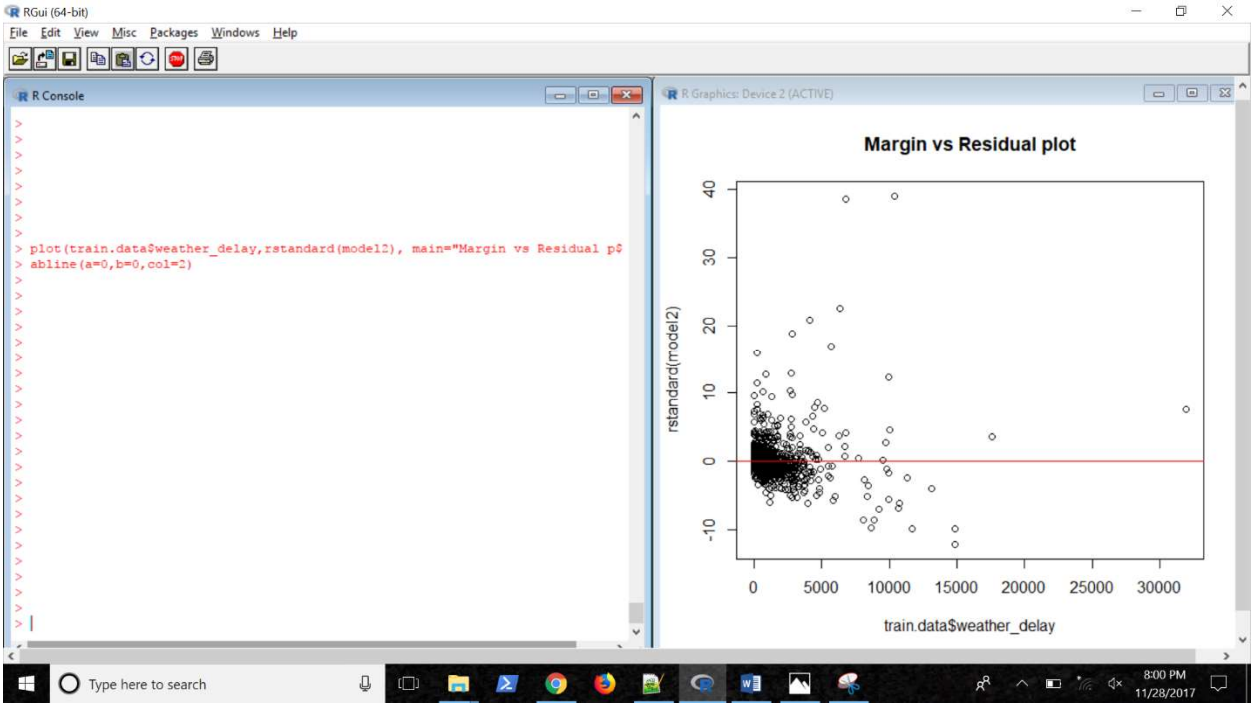
```
> plot(train.data$carrier_delay,rstandard(model2), main="Margin vs Residual p$
> abline(a=0,b=0,col=2)
>
```

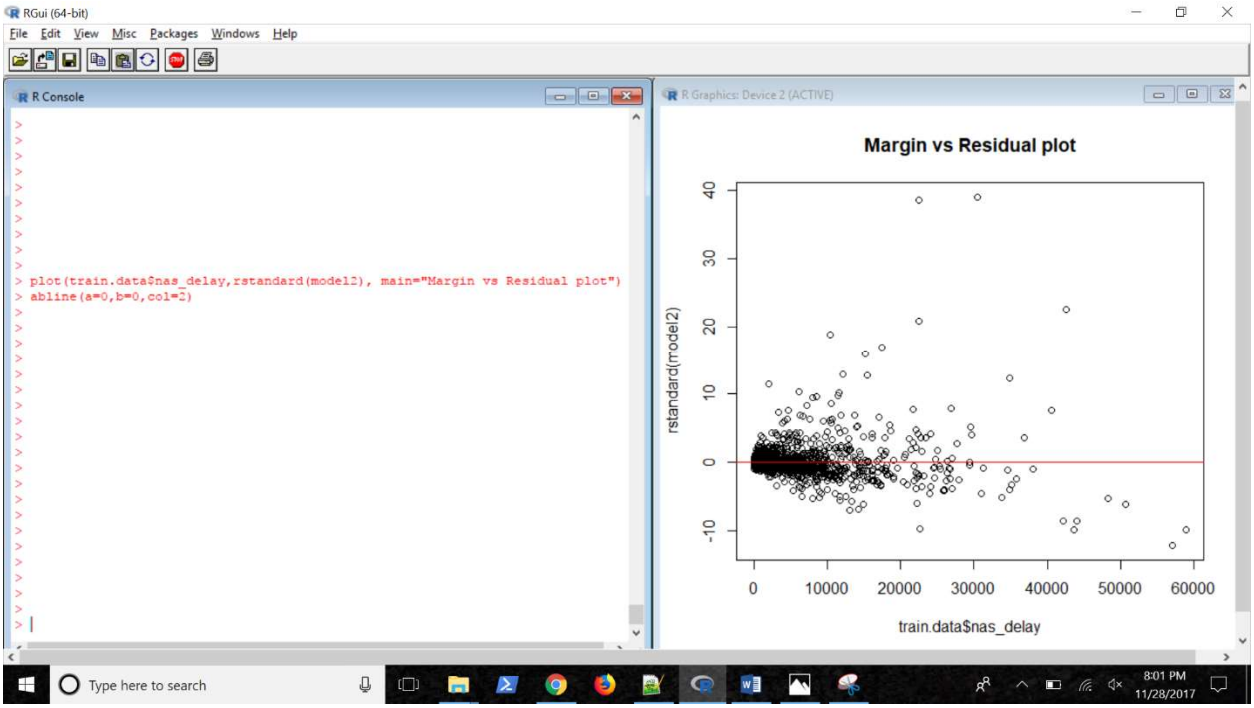
R Graphics: Device 2 (ACTIVE)



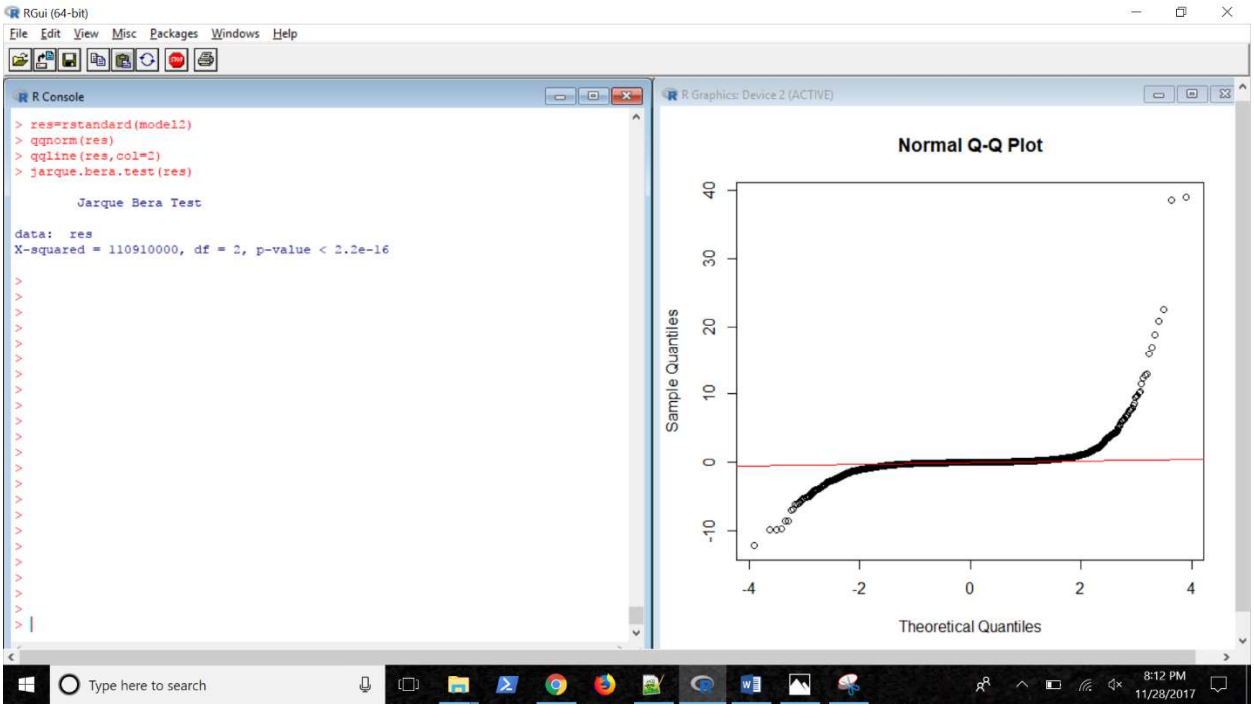
Type here to search

8:00 PM
11/28/2017


RGui (64-bit)
File Edit View Misc Packages Windows Help
R Console
> plot(train.data$weather_delay,rstandard(model2), main="Margin vs Residual p$
> abline(a=0,b=0,col=2)
R Graphics: Device 2 (ACTIVE)
Margin vs Residual plot
rstandard(model2)
train.data$weather_delay
0
5000
10000
15000
20000
25000
30000
40
30
20
10
0
-10
Type here to search
8:00 PM
11/28/2017


RGui (64-bit)
File Edit View Misc Packages Windows Help
R Console
> plot(train.data$nas_delay,rstandard(model2), main="Margin vs Residual plot")
> abline(a=0,b=0,col=2)
R Graphics: Device 2 (ACTIVE)
Margin vs Residual plot
rstandard(model2)
train.data$nas_delay
0
10000
20000
30000
40000
50000
60000
40
30
20
10
0
-10
Type here to search
8:01 PM
11/28/2017

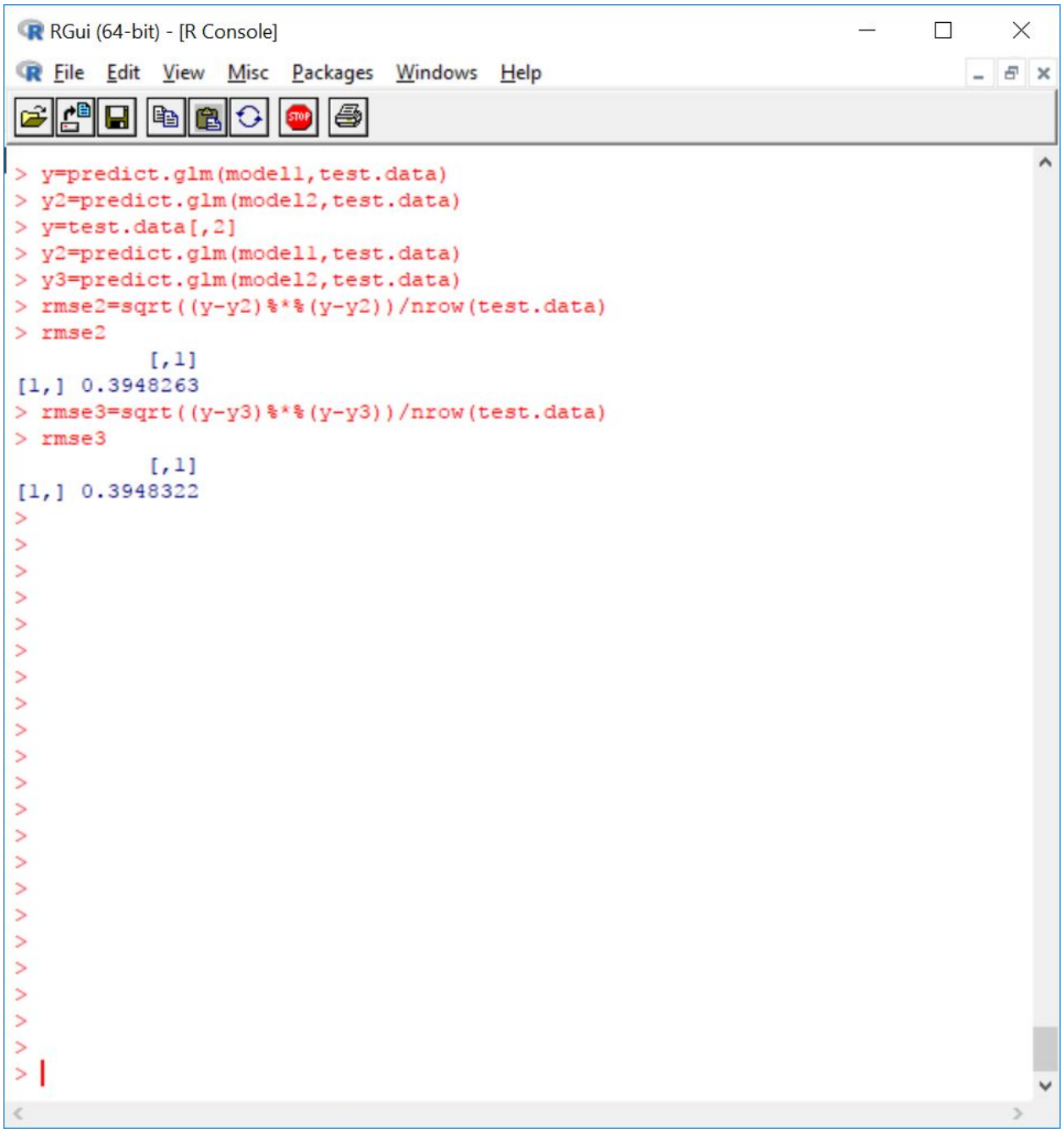
Normality test



Multicollinearity

```
> vif(model2)
    arr_delay carrier_delay weather_delay     nas_delay
    43.204896     20.513355      2.896123      8.669979
> sqrt(vif(model2))
    arr_delay carrier_delay weather_delay     nas_delay
     6.573043      4.529167      1.701800      2.944483
>
>
```

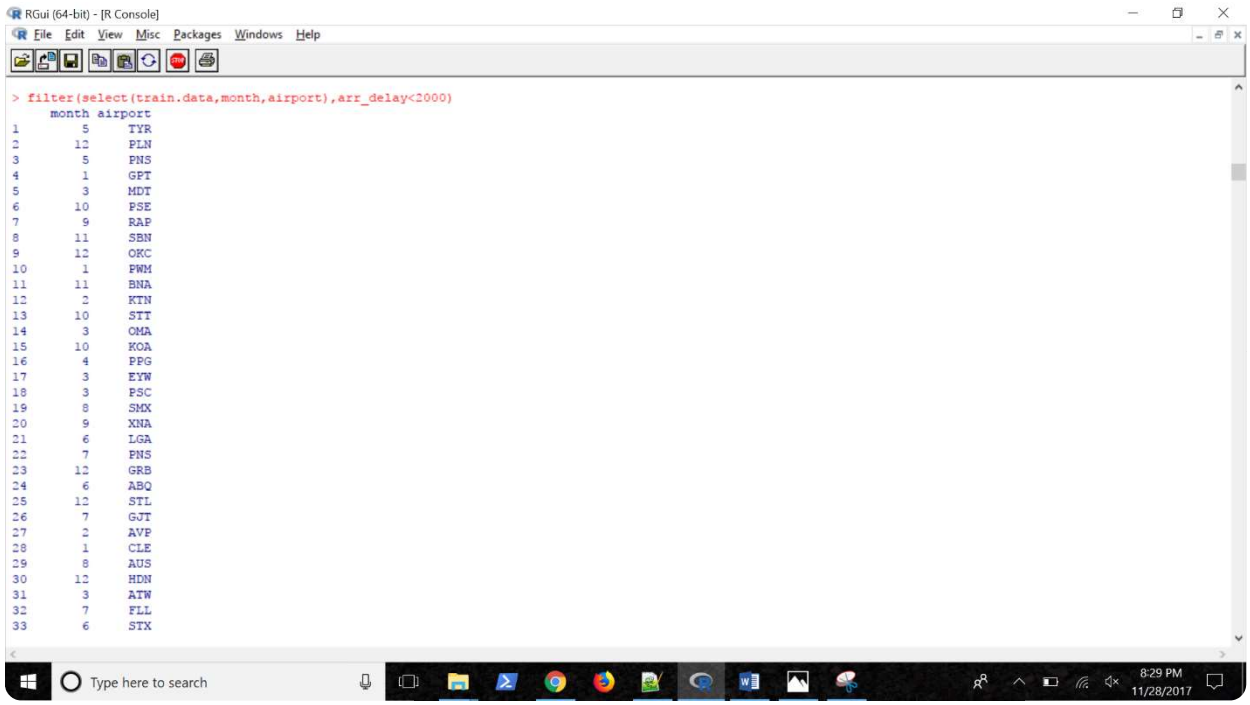
Rmse technique to get the best mpdel using r2 adj value.



6.2. Results and Findings

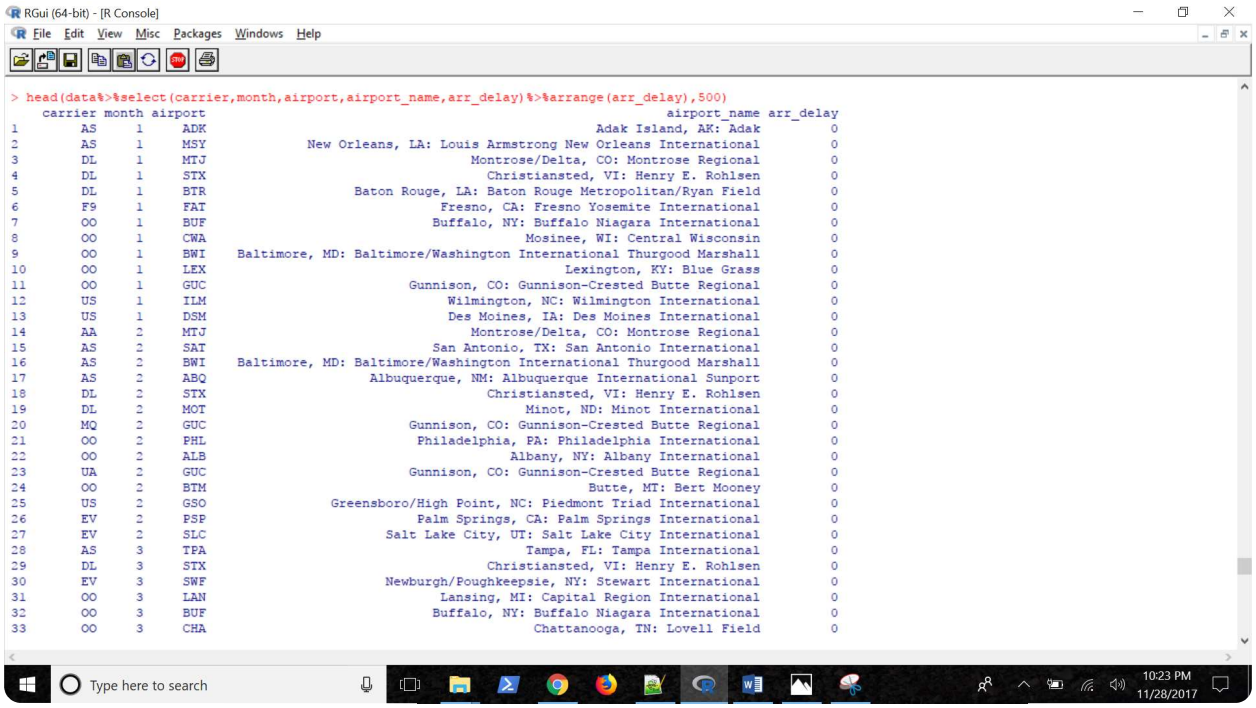
Using **dplyr** function for advanced analysis

Filter



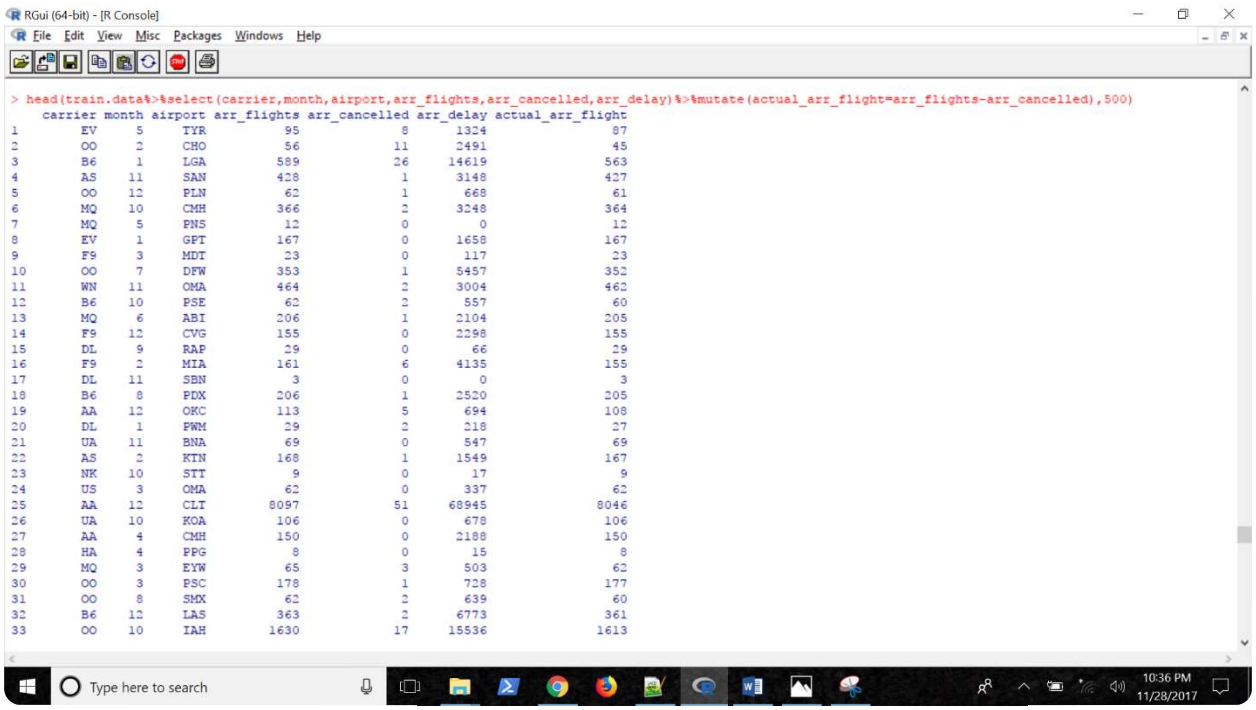
Maximum arrival delay.

Minimum arrival Delay



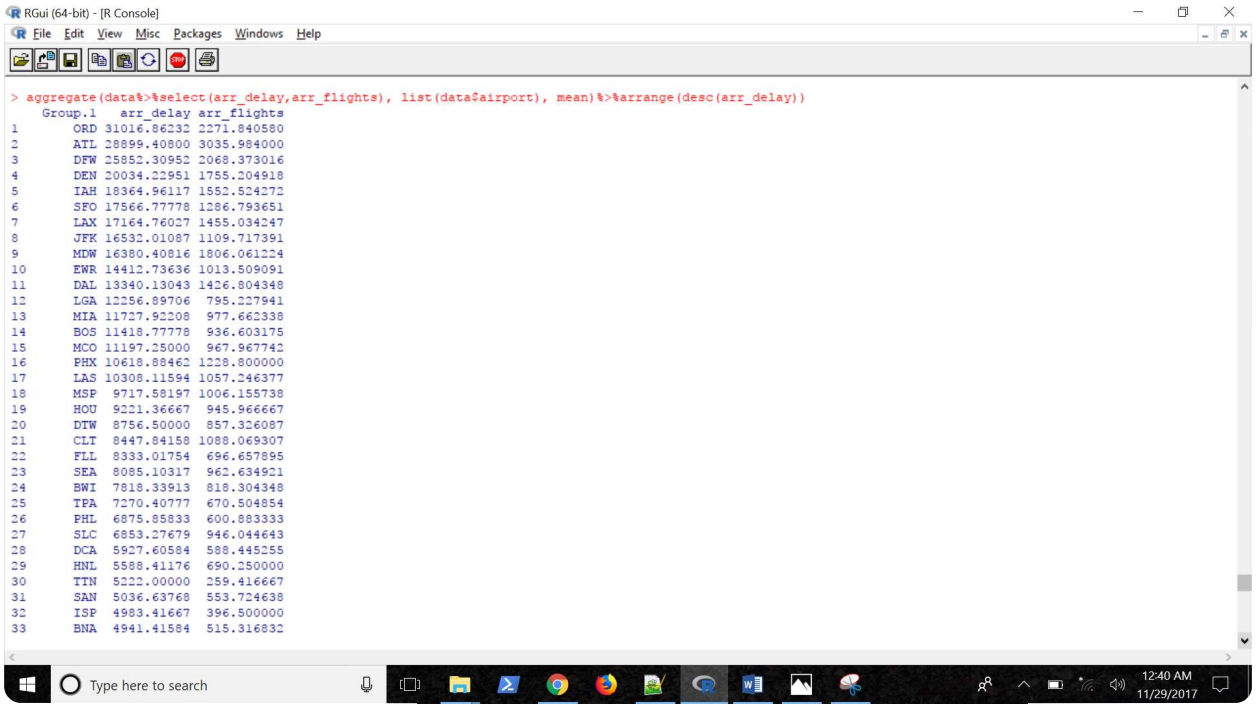
```
> head(data%>%select(carrier,month,airport,airport_name,arr_delay)%>%arrange(arr_delay),500)
    carrier month airport                                                airport_name arr_delay
1        AS     1     ADK                                                Adak Island, AK: Adak         0
2        AS     1     MSY            New Orleans, LA: Louis Armstrong New Orleans International         0
3        DL     1     MTJ                                Montrose/Delta, CO: Montrose Regional         0
4        DL     1     STX                                  Christiansted, VI: Henry E. Rohlsen         0
5        DL     1     BTR                  Baton Rouge, LA: Baton Rouge Metropolitan/Ryan Field         0
6        F9     1     FAT                              Fresno, CA: Fresno Yosemite International         0
7        OO     1     BUF                            Buffalo, NY: Buffalo Niagara International         0
8        OO     1     CWA                                    Mosinee, WI: Central Wisconsin         0
9        OO     1     BWI    Baltimore, MD: Baltimore/Washington International Thurgood Marshall         0
10       OO     1     LEX                                         Lexington, KY: Blue Grass         0
11       OO     1     GUC                          Gunnison, CO: Gunnison-Crested Butte Regional         0
12       US     1     ILM                               Wilmington, NC: Wilmington International         0
13       US     1     DSM                                Des Moines, IA: Des Moines International         0
14       AA     2     MTJ                                Montrose/Delta, CO: Montrose Regional         0
15       AS     2     SAT                              San Antonio, TX: San Antonio International         0
16       AS     2     BWI    Baltimore, MD: Baltimore/Washington International Thurgood Marshall         0
17       AS     2     ABQ                     Albuquerque, NM: Albuquerque International Sunport         0
18       DL     2     STX                                  Christiansted, VI: Henry E. Rohlsen         0
19       DL     2     MOT                                      Minot, ND: Minot International         0
20       MQ     2     GUC                          Gunnison, CO: Gunnison-Crested Butte Regional         0
21       OO     2     PHL                           Philadelphia, PA: Philadelphia International         0
22       OO     2     ALB                                  Albany, NY: Albany International         0
23       UA     2     GUC                          Gunnison, CO: Gunnison-Crested Butte Regional         0
24       OO     2     BTM                                            Butte, MT: Bert Mooney         0
25       US     2     GSO                Greensboro/High Point, NC: Piedmont Triad International         0
26       EV     2     PSP                          Palm Springs, CA: Palm Springs International         0
27       EV     2     SLC                      Salt Lake City, UT: Salt Lake City International         0
28       AS     3     TPA                                    Tampa, FL: Tampa International         0
29       DL     3     STX                                  Christiansted, VI: Henry E. Rohlsen         0
30       EV     3     SWF                      Newburgh/Poughkeepsie, NY: Stewart International         0
31       OO     3     LAN                              Lansing, MI: Capital Region International         0
32       OO     3     BUF                            Buffalo, NY: Buffalo Niagara International         0
33       OO     3     CHA                                  Chattanooga, TN: Lovell Field         0
```

Aircrafts that flew



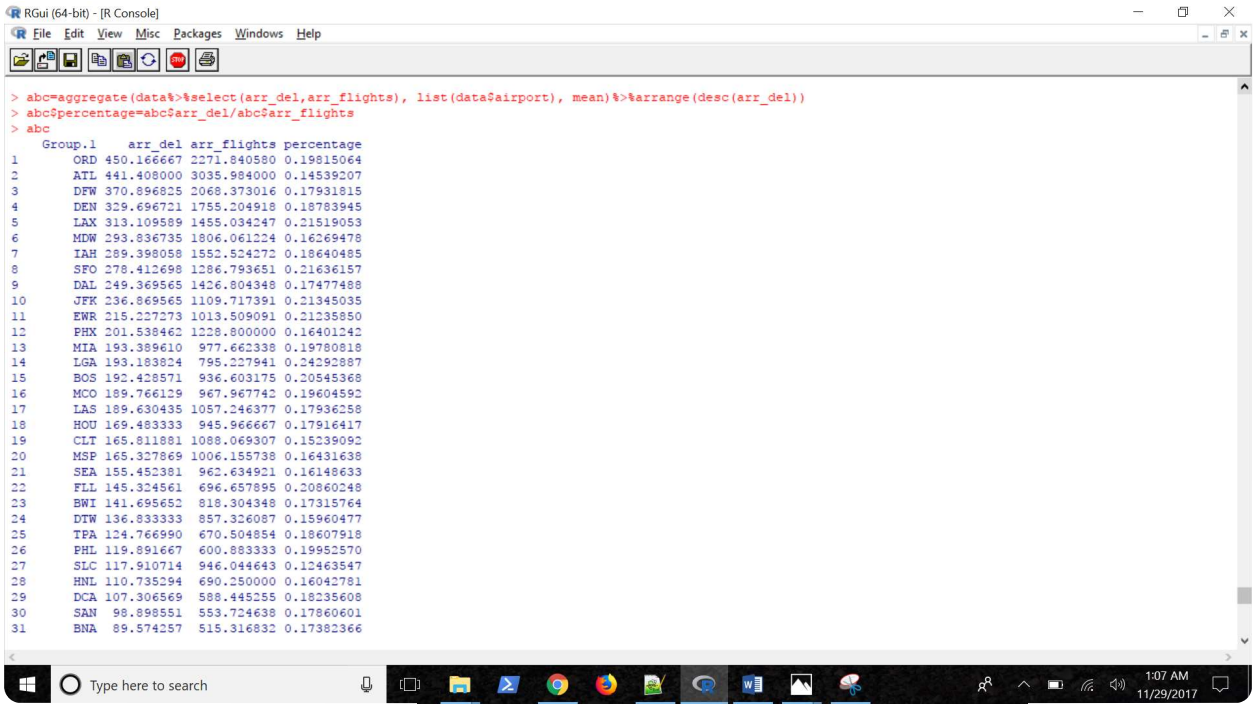
```
> head(train.data%>%select(carrier,month,airport,arr_flights,arr_cancelled,arr_delay)%>%mutate(actual_arr_flight=arr_flights-arr_cancelled),500)
    carrier month airport arr_flights arr_cancelled arr_delay actual_arr_flight
1        EV     5     TYR         95             8      1324              87
2        OO     2     CHO         56            11      2491              45
3        B6     1     LGA        589            26     14619             563
4        AS    11     SAN        428             1      3148             427
5        OO    12     PLN         62             1       668              61
6        MQ    10     CMH        366             2      3248             364
7        MQ     5     PNS         12             0         0              12
8        EV     1     GPT        167             0      1658             167
9        F9     3     MDT         23             0       117              23
10       OO     7     DFW        353             1      5457             352
11       WN    11     OMA        464             2      3004             462
12       B6    10     PSE         62             2       557              60
13       MQ     6     ABI        206             1      2104             205
14       F9    12     CVG        155             0      2298             155
15       DL     9     RAP         29             0        66              29
16       F9     2     MIA        161             6      4135             155
17       DL    11     SBN          3             0         0               3
18       B6     8     PDX        206             1      2520             205
19       AA    12     OKC        113             5       694             108
20       DL     1     PWM         29             2       218              27
21       UA    11     BNA         69             0       547              69
22       AS     2     KTN        168             1      1549             167
23       NK    10     STT          9             0        17               9
24       US     3     OMA         62             0       337              62
25       AA    12     CLT       8097            51     68945            8046
26       UA    10     KOA        106             0       678             106
27       AA     4     CMH        150             0      2188             150
28       HA     4     PPG          8             0        15               8
29       MQ     3     EYW         65             3       503              62
30       OO     3     PSC        178             1       728             177
31       OO     8     SMX         62             2       639              60
32       B6    12     LAS        363             2      6773             361
33       OO    10     IAH       1630            17     15536            1613
```

Mean of arrival delays and arrival flights



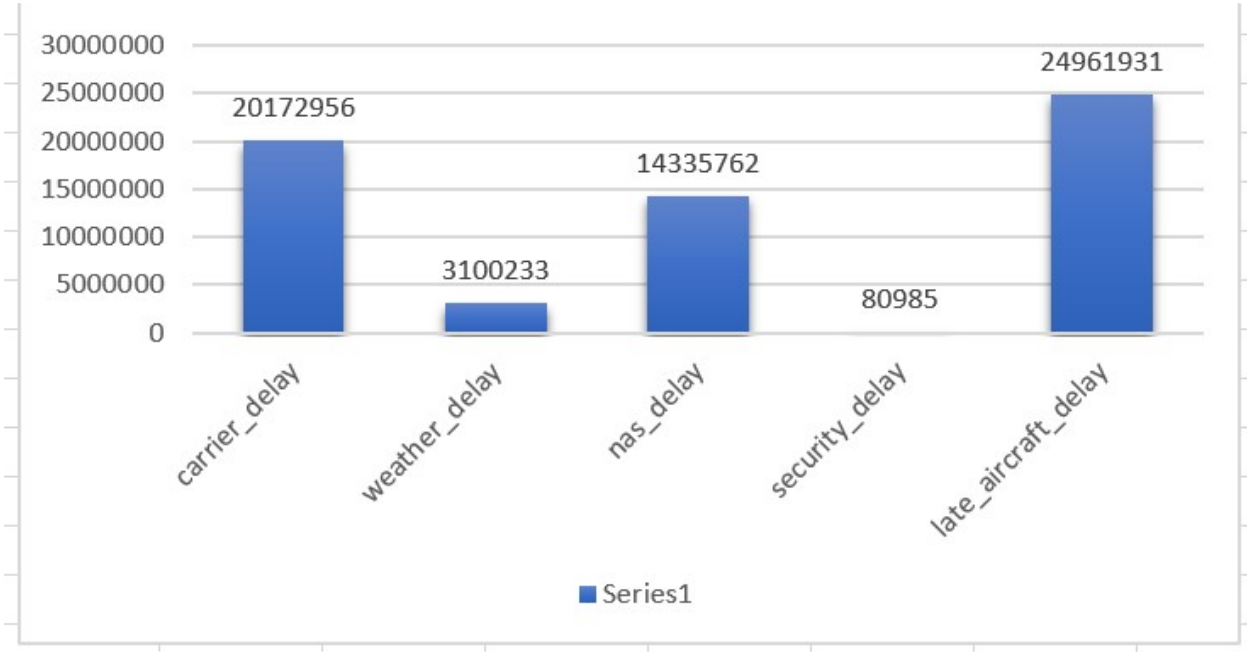
```
> aggregate(data%>%select(arr_delay,arr_flights), list(data$airport), mean)%>%arrange(desc(arr_delay))
   Group.1   arr_delay arr_flights
1      ORD 31016.86232 2271.840580
2      ATL 28899.40800 3035.984000
3      DFW 25852.30952 2068.373016
4      DEN 20034.22951 1755.204918
5      IAH 18364.96117 1552.524272
6      SFO 17566.77778 1286.793651
7      LAX 17164.76027 1455.034247
8      JFK 16532.01087 1109.717391
9      MDW 16380.40816 1806.061224
10     EWR 14412.73636 1013.509091
11     DAL 13340.13043 1426.804348
12     LGA 12256.89706  795.227941
13     MIA 11727.92208  977.662338
14     BOS 11418.77778  936.603175
15     MCO 11197.25000  967.967742
16     PHX 10618.88462 1228.800000
17     LAS 10308.11594 1057.246377
18     MSP  9717.58197 1006.155738
19     HOU  9221.36667  945.966667
20     DTW  8756.50000  857.326087
21     CLT  8447.84158 1088.069307
22     FLL  8333.01754  696.657895
23     SEA  8085.10317  962.634921
24     BWI  7818.33913  818.304348
25     TPA  7270.40777  670.504854
26     PHL  6875.85833  600.883333
27     SLC  6853.27679  946.044643
28     DCA  5927.60584  588.445255
29     HNL  5588.41176  690.250000
30     TTN  5222.00000  259.416667
31     SAN  5036.63768  553.724638
32     ISP  4983.41667  396.500000
33     BNA  4941.41584  515.316832
```


Delay percentage



```
> abc=aggregate(data%>%select(arr_del,arr_flights), list(data$airport), mean)%>%arrange(desc(arr_del))
> abc$percentage=abc$arr_del/abc$arr_flights
> abc
   Group.1    arr_del arr_flights percentage
1      ORD 450.166667 2271.840580 0.19815064
2      ATL 441.408000 3035.984000 0.14539207
3      DFW 370.896825 2068.373016 0.17931815
4      DEN 329.696721 1755.204918 0.18783945
5      LAX 313.109589 1455.034247 0.21519053
6      MDW 293.836735 1806.061224 0.16269478
7      IAH 289.398058 1552.524272 0.18640485
8      SFO 278.412698 1286.793651 0.21636157
9      DAL 249.369565 1426.804348 0.17477488
10     JFK 236.869565 1109.717391 0.21345035
11     EWR 215.227273 1013.509091 0.21235850
12     PHX 201.538462 1228.800000 0.16401242
13     MIA 193.389610  977.662338 0.19780818
14     LGA 193.183824  795.227941 0.24292887
15     BOS 192.428571  936.603175 0.20545368
16     MCO 189.766129  967.967742 0.19604592
17     LAS 189.630435 1057.246377 0.17936258
18     HOU 169.483333  945.966667 0.17916417
19     CLT 165.811881 1088.069307 0.15239092
20     MSP 165.327869 1006.155738 0.16431638
21     SEA 155.452381  962.634921 0.16148633
22     FLL 145.324561  696.657895 0.20860248
23     BWI 141.695652  818.304348 0.17315764
24     DTW 136.833333  857.326087 0.15960477
25     TPA 124.766990  670.504854 0.18607918
26     PHL 119.891667  600.883333 0.19952570
27     SLC 117.910714  946.044643 0.12463547
28     HNL 110.735294  690.250000 0.16042781
29     DCA 107.306569  588.445255 0.18235608
30     SAN  98.898551  553.724638 0.17860601
31     BNA  89.574257  515.316832 0.17382366
```

Delay in minutes for each variable



7. Conclusions and Future Work

7.1. Conclusions

- Main reason of cancellation - The common attributes that make up the reason of cancellation are: arr_delay, nas_delay, weather_delay and carrier_delay.
- As adj-r2 is low, from the above we can say arr_cancelled is not well supported by the properties (attributes) chosen for analysis

7.2. Limitations

- In the output in R we are not able to see whole output because the data set is a large and we are not able to do the complete analysis.
- The timing and the numbers present in the details not being verified, we don't know whether they are recorder on time or there is some error in that.
- Not only airline delay, other important details can also be found out from this data.

7.3. Potential Improvements or Future Work

- Using PCR or ridge regression we can improve the accuracy of the model in the future.
- Removing the multicolinearity problem from the model.
- Some factors that affect the variable can be identified using the decision tree.
- Precision of the prediction can be increased by the random forest method.