Predicting Flight Delay @ US Airports
Fahad Alsehami
2018-05-20
https://github.com/alfahadama/Airline_Delayed_in_US

1. Introduction

Every year, millions of passengers experience delays in flights, resulting in missing connections and spending more time away from home among others. The data is about the analysis of all flights that departed from New York City (e.g. EWR, JFK and LGA), the raw dataset contains 336,776 flights in total in 2013. In order to explain the causes of delays happen in 2013, variables also include a number of other datasets:

Dataset	Filename	Description
Flights	U.S Flight Dataset	Flight departures from US in 2013
Weather	U.S. Weather-Dataset	Hourly meteorological data for each airport
Planes	U.S. Planes_Dataset	Construction information about each plane
Airports	U.S. Airports_dataset	Airport names and locations
Airlines	U.S. Airlines_Dataset	Translation between two letter carrier codes and names

The following variables were recorded:

Variables	Description
year,month,day	Date of departure
dep_time,arr_time	Actual departure and arrival times, local tz.
sched_dep_time,sched_arr_time	Scheduled departure and arrival times, local tz.
dep_delay,arr_delay	Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.
carrier	Two letter carrier abbreviation. See airlines to get name
flight	Flight number

tailnum	Plane tail number
origin ,dest	Origin and destination. See airports for additional metadata.
air_time	Amount of time spent in the air, in minutes
distance	Distance between airports, in miles
hour,minute	Time of scheduled departure broken into hour and minutes.
time_hour	Scheduled date and hour of the flight as a date.
Airline	Full name
type	Type of plane
manufacturer,model	Manufacturer and model
engines ,seats	Number of engines and seats
speed	Average cruising speed in mph
engine	Type of engine
age	Age of plane
name.dest	Usual name of the airport
lat.dest ,lon.dest	Location of airport
alt.dest	Altitude, in feet
name.origin	Usual name of the airport
lat.origin ,lon.origin	Location of airport
alt.origin	Altitude, in feet
temp ,dewp	Temperature and dewpoint in F
humid	Relative humidity
wind_dir ,wind_speed ,wind_gust	Wind direction (in degrees), speed and gust speed (in mph)

precip	Precipitation, in inches
pressure	Sea level pressure in millibars
visib	Visibility in miles

This study intent to predict the total delay time for flights departing from NYC based on the hourly meteorological data for each airport, construction information about each plane, airport locations and the Flight characteristics. Thus the response variable is the total delay time (arr_delay + dep_delay), which denote the total of Departure and arrival delays, in minutes, all remaining variables are predictors.

Our specific objectives are as follows:

- To identify possible factors that may influence the delay times for flights departing from NYC.
- To provide recommendations for improving U.S flight.

The raw data have been preprocessed, a set of 12 different variables was obtained which affect delay of the U.S flights. In this variable set, while three of them are categorical variable, the rest are numeric variables, Among three categorical variables, two of them have four levels and the other has three levels. Moreover, in this data set, 24 variables are removed based on their lack of information content so they are omitted from the regression analysis.
First of all the data set divided into two groups which contain test and train parts. 70 percent (n = 200363 observations) of this data set will be used to train this regression model and the 30 percent (n = 85871 observations) of it will be used to test regression model obtained from train part.

Indicators: There are three categorical variables, and two of them have four levels and the other one has three levels. Therefore, in total there are 8 dummy variables in hand.
Standardization: In order to get rid of different units in the data set, it is needed to standardize all variables except for dummy variables. Thus, each variable have the same standard.
Multiple linear regressions will be performed to determine whether or not the variation that is observed in the response variable (which corresponds to ‘arr_delay’+ ‘dep_delay’ in this analysis) can be predicted by the Flight characteristics, airport location and the weather. Therefore the null hypothesis is that any of variables concerning flight characteristics, location of airports and weather does not have a significant effect on the total delay in departure and arrival time.

The hypothesis test follows;

```
H0 : There is no lack of fit in model.
Ha : There is lack of fit in model.
```

Furthermore, model selection methods are applied. Therefore stepwise regression is applied in order to obtain the best model in this study.This is a combination of backward elimination and forward selection. This addresses the situation where variables are added or removed early in the process and we want to change our mind about them later.

2. Results

2.1. Data Pre-processing

Data pre-processing or initial data analysis generally performed to prepare and understand the data. In this matter, univariate descriptive statistics were gathered:

- Numerical summaries - means, sds, five-number summaries, correlations,
- Graphical summaries - histograms and scatter plots were created.

Additionally, some predictors are removed based on their lack of information content, and some new variables are created.

We start by creating new variables “Quarter” and “TimeOfDay” in the dataframe. Such as “Q1”, “Q2”, “Q3”, “Q4”, referring to the four quarters of a calendar year. The TimeOfDay is the Splitting of hour variable into six hour segments:

- Midnight - 6am: Overnight
- 6am - Midday: Morning
- Midday - 6pm: Afternoon
- 6pm - Midnight: Evening

Below, the frequency distribution of each categorical variable. There were 286,234 flights taking off from New York, 34.97% of them (n = 100106) taking off from Newark Liberty International Airport, 34.23% (n = 98069) taking off from John F Kennedy International and 30.76% (n = 88059) that took off from La Guardia airport. The quarterly dispersion of 2013 was the same (around 25% of Total flights each, or 70,000 per quarter). However, the majority of flights (39.43%) took off in the morning, 37.73% took off in the afternoon, 22.23% left in the evening, while 0.6% of total 2013 took off during the night.

Table 1: Frequency Distributions

Dummy Variables	n	percent
United Air Lines Inc.	49514	17.3%
JetBlue Airways	48105	16.81%
ExpressJet Airlines Inc.	44138	15.42%
Delta Air Lines Inc.	41856	14.62%
American Airlines Inc.	28109	9.82%
Envoy Air	22008	7.69%
US Airways Inc.	17301	6.04%
Endeavor Air Inc.	15561	5.44%
Southwest Airlines Co.	10298	3.6%
Virgin America	4523	1.58%
AirTran Airways Corporation	2791	0.98%
Alaska Airlines Inc.	630	0.22%
Frontier Airlines Inc.	587	0.21%
Mesa Airlines Inc.	482	0.17%
Hawaiian Airlines Inc.	304	0.11%
SkyWest Airlines Inc.	27	0.01%
Total	286234	100%
Newark Liberty International Airport	100106	34.97%
John F Kennedy International Airport	98069	34.26%
La Guardia Airport	88059	30.76%
Q3	74422	26%

Q4	71777	25.08%
Q2	70592	24.66%
Q1	69443	24.26%
Morning	112876	39.43%
Afternoon	108004	37.73%
Evening	63626	22.23%
Overnight	1728	0.6%

The US flight delay ranged from -100 to 2573 minutes with an average of 15.83 minutes. Figure 1 shows that the distribution of the outcome is right skewed, it has long tail in the high values.

Table 2: Descriptive statistics of the outcome

Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
-100	-21	-7	15.83	19	2573	77.65

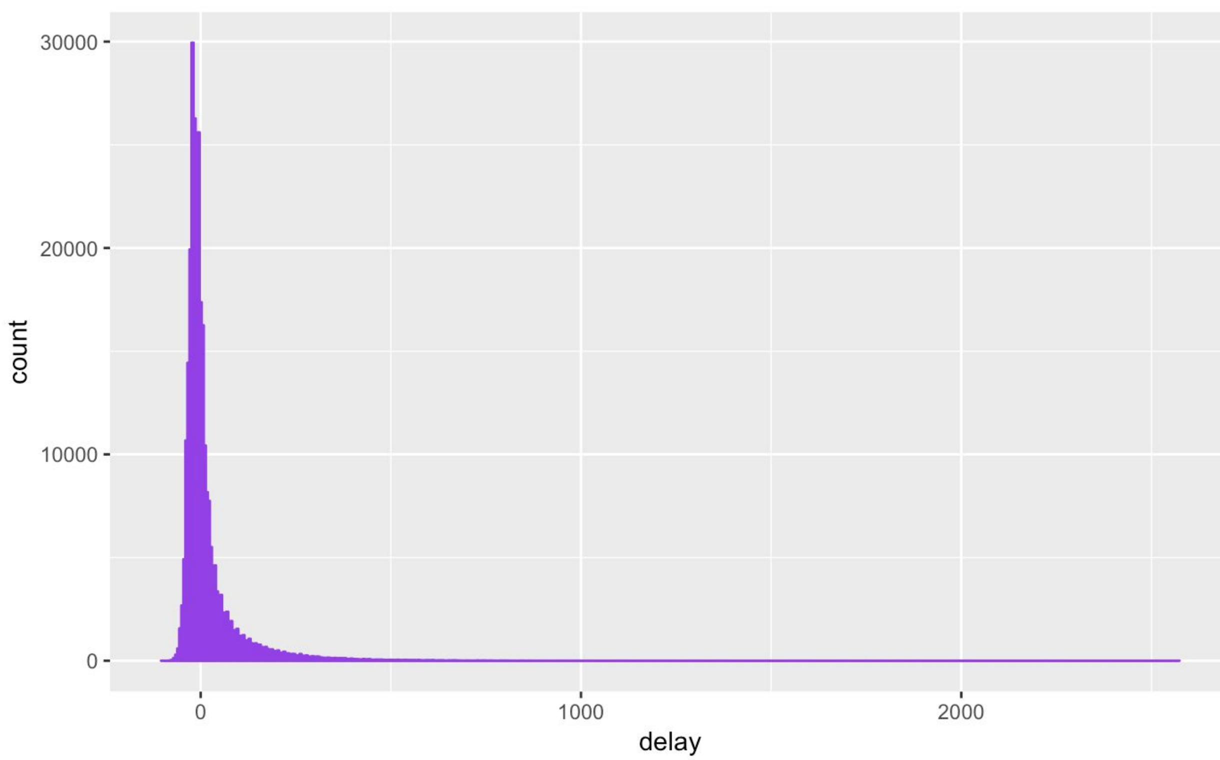


Figure 1: Distribution of the flight delay at NYC Airports

Here, some predictors are removed based on their lack of information content, the caret package function nearZeroVar is used in order to filter all predictors with near zero variance. In our data, there are three problematic predictors that should be removed from the data.
Similarly, ***findCorrelation*** function from the caret package is used in order to filter on high absolute pairwise between-predictor correlations:
The Figure below helps up to visually examine the between-predictor of the data:

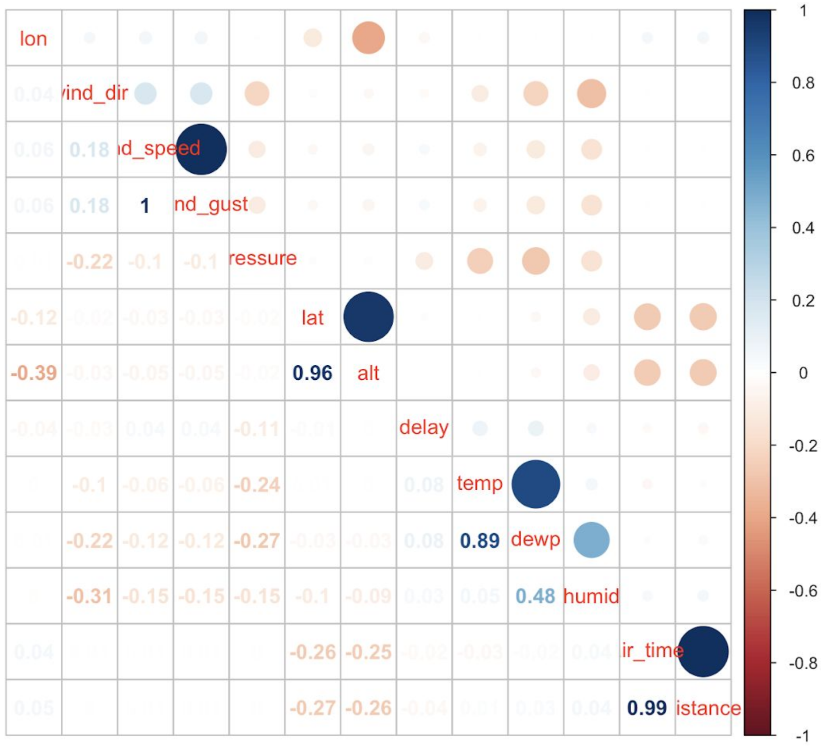


Figure 2: Correlation Matrix of the Flights NYC Airports
As shown in the Figure above, the pairs, (air_time and distance), (wind_speed and wind_gust) and (alt and lat) have a strong positive correlation, respectively.
Now, let us reduce the four predictors identified as collinear that have an absolute pairwise correlation above 0.75:
The variables *dewp*, alt, *wind_gust*, *distance* are highly correlated with others predictors.
Moving on, we can evaluate the continuous predictors for skewness. The skewness statistic ranges from a minimum of -0.5 to a maximum of 66.52, indicating that most of our predictors are right skewed. To correct for this skewness, a Box-Cox transformation was applied to all predictors.
Table 3: skewness statistic

air_time	temp	humid	wind_dir	wind_speed	pressure	lat	lon	delay
1.061	-0.0007674	0.1504	-0.5233	66.53	0.0885	0.3601	-0.4783	4.673

Figure 3 shows scatter plots of the predictors against the outcome along with a regression line from a flexible "smoother" model. According to these two figures, we can assume that the relationship between the predictors and the outcome is linear.

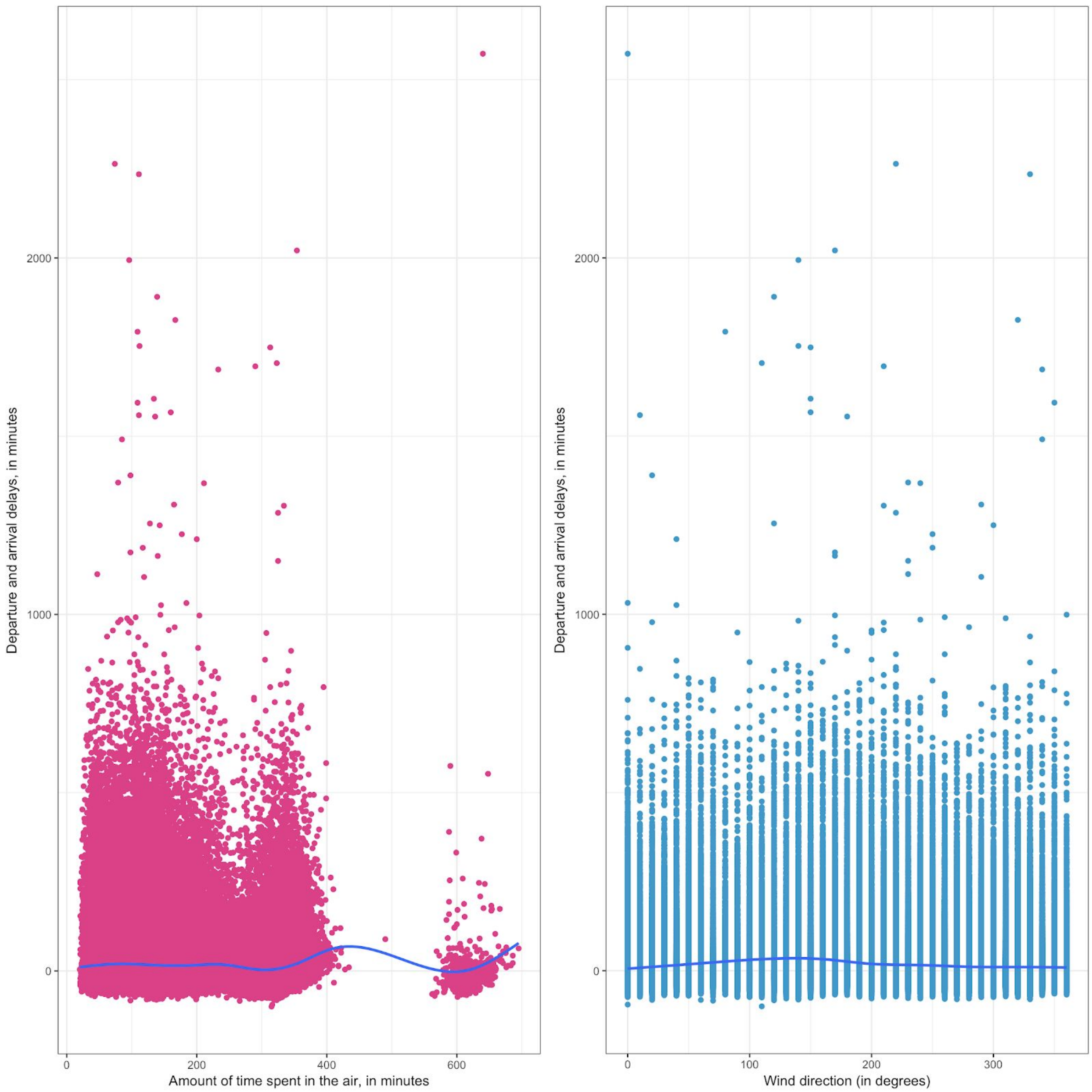


Figure 3: Graphical Representation of flights.2013 dataset

2.2. Data Splitting

To allocate data to model building and evaluating performance: let's split the *flights.2013* dataset into two parts: training and test data. 70% of the 2013 U.S flights will belong to the training set, and the other 30% to the test set:
Allocate *training* dataset to model building, which contains 200363 cases, and *test* dataset to evaluating performance, which contains 200363 cases.

2.3. Model Building

The regression line can be written in the form:
Where:

- : mean of the dependent variable when all (Center)
- Binary X = "dummy variable" for group
  - :i=1,.., total groups -1: mean difference in outcome between groups
- Continuous X
  - difference in mean outcome corresponding to a 1-unit increase in X

delay of fights is the response, and the remaining variables are the predictors. we have 9 continuous variables, eight dummy variables and no missing data.
The null hypothesis is as follows:

- : all new 's are zero
- Assess using F-test

The Table below displays model summary statistics, the parameter estimates, their standard errors, and p-values for testing whether each individual coefficient is different than 0:
Table 4: Fitting linear model: delay ~ air_time + temp + humid + wind_dir + wind_speed + pressure + name + Qtr + TimeOfDay

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	901	27.41	32.87	2.252e-236
air_time	-0.01944	0.001863	-10.44	1.727e-25
temp	-0.1022	0.01714	-5.964	2.466e-09
humid	0.4757	0.01141	41.67	0
wind_dir	-0.01871	0.001743	-10.73	7.121e-27
wind_speed	0.1318	0.01246	10.58	3.79e-26

pressure	-0.919	0.02632	-34.92	2.238e-266
La Guardia Airport	1.903	0.4418	4.307	1.654e-05
Newark Liberty International Airport	8.057	0.4109	19.61	1.523e-85
QtrQ2	10.3	0.659	15.64	4.518e-55
QtrQ3	4.729	0.8011	5.902	3.589e-09
QtrQ4	-1.016	0.5355	-1.897	0.05786
TimeOfDayMorning	6.44	2.187	2.944	0.003239
TimeOfDayAfternoon	35.43	2.199	16.11	2.223e-58
TimeOfDayEvening	47.38	2.214	21.4	1.77e-101

Observations	Residual Std. Error		Adjusted
200363	75.32	0.05568	0.05561

The simple estimates of the RMSE and were 75.32 and 0.05568, respectively.
So when name=LGA , the prediction is , 2.10 minutes more than for , and when name=EWR , the prediction is = 8.4, 8.4 minutes more than for . And from the extremely small *p*-value, this is a significant finding. So we are quite sure that flights from JFK made a significantly lower delay LGA or EWR .

The Location of airport (Lat and lon) are not significant at all (*p*-value > 0.05), while the Temperature and dewpoint in F (*temp*), Relative humidity (*humid*), Wind direction (in degrees) (*wind_dir*), speed (in mph)(*wind_dir*), Sea level pressure in millibars *pressure* are highly significant for U.S flight delay (*p*-value <0.05).

Best model for the stepwise selection is the following;
Model characteristics: , F-statistic: 828.7 on 14 and 200348 DF, p-value: < 2.2e-16
To compute the model flight delay values for new samples, the predict method is used:

```
lmPred1 <- predict(lm.delay, test)
 head(lmPred1)
##    1     2     3     4     5     6
##  2.96 -8.80  3.07 -4.58 -9.21 -5.49
```

The caret function defaultSummary is used to estimate the test set performance:

```
lmValues1 <- data.frame(obs = test$delay, pred = lmPred1)
 defaultSummary(lmValues1)
##     RMSE Rsquared     MAE
##  75.8810   0.0532  43.4417
```

Based on the test set, the summaries produced by the summary function for lm were pessimist.

3. Conclusion

The aim of this paper was to construct a linear model that predicts well the relationships in all flights that departed from New York City data, with this reliable analysis, it's not easy at to predict the U.S flight delay of an unknown data because our output prediction did not worked at all in the above mentioned experiment. Further modeling assumptions had failed in this experiment. Thus, multiple linear regression failed to predict the delay of US flights. A future analysis using non-parametric methods may be conducted to carry out the estimation of delays flights departing from NYC, for instance decision trees, random forests can be used in this matter.

Appendix

 # Load the packsge into memory

```
library(pander)
library(tidyverse)
library(janitor)
library(caret)
```

```
#Prepare the dataset
#Extract records and Add the Total Delay field "TotalDelay"
flights.2013 <-
  flights %>%
       #Hourly meterological data for LGA, JFK and EWR.
  left_join(weather %>% select(origin,temp,dewp,humid, wind_dir,
wind_speed,wind_gust,precip,pressure, visib,time_hour), by = c("origin", "time_hour")) %>%
  #Add airline names for carriers targeted for study
  left_join(airlines, by = c("carrier"="Code")) %>%
  rename(Airline = Description) %>%
  # Add the location of the airport
  left_join(airports %>% select(faa,name,lat,lon,alt), by = c("origin" = "faa"))  %>%
```

```
  na.omit()
```

```
#Creating new variables
flights.2013 <-
  flights.2013 %>% mutate(delay = dep_delay+arr_delay, #Create the outcome delay
                      Qtr = factor(quarters(time_hour)), #Create a new column "quarter"
                      TimeOfDay = cut(hour, c(0, 6, 12, 18, 24),
                                  labels = c("Overnight", "Morning", "Afternoon",
"Evening"),
                                  right = FALSE)
                      ) %>%
  droplevels()
```

```
#Data investigation
data.fac <- Filter(is.factor,flights.2013 )
data.fac %>% tabyl(Airline) %>% arrange(desc(n)) %>% adorn_totals("row") %>%
mutate(percent=paste0(round(100*percent,2),"%")) %>% pander()
data.fac %>% tabyl(name) %>% arrange(desc(n)) %>% adorn_totals("row") %>%
mutate(percent=paste0(round(100*percent,2),"%")) %>% pander()
data.fac %>% tabyl(Qtr) %>% arrange(desc(n)) %>% adorn_totals("row") %>%
mutate(percent=paste0(round(100*percent,2),"%")) %>% pander()
data.fac %>% tabyl(TimeOfDay) %>% arrange(desc(n)) %>% adorn_totals("row") %>%
mutate(percent=paste0(round(100*percent,2),"%")) %>% pander()
flights.2013 %>%
  summarise(Min = min(delay),`1st Qu.` = quantile(delay, 0.25),Median = median(delay),Mean =
mean(delay),`3rd Qu.` = quantile(delay, 0.75),Max = max(delay), SD = sd(delay)) %>%
  pander(caption ="Descriptive statistics of the outcome")
ggplot(data = flights.2013, aes(delay)) +
  geom_histogram(color="blue", bins = 500)
```

```
## A vector of three (1,26,28) integers is returned that indicates which columns should be
removed.
nearZeroVar(flights.2013)
flights.2013 <- flights.2013[, -c(1:14,17:19,26,28,29)]
```

```
data.num <- Filter(is.numeric, flights.2013)
correlations <- cor(data.num)
 dim(correlations)
## [1] 13 13
```

```
library(corrplot)
corrplot.mixed(correlations, order="hclust")
highCorr <- findCorrelation(correlations, cutoff = .75)
 length(highCorr)
```

```
## [1] 4
```

```
flights.2013 <- flights.2013 %>% select(-dewp,-alt,-wind_gust,-distance)

rm(list = c( "airlines" ,"airports","correlations",
"data.fac","data.num","flights","flights13" ,"highCorr","planes","trainingRows","weather"
))#ls()) #Clear workspace
library(e1071)

skewValues <- apply(Filter(is.numeric,flights.2013), 2, skewness)

skewValues %>% pander(caption="Skewness across columns")
library(cowplot) #Arranging plots in a grid
fig1 <- ggplot(data=flights.2013, aes(x = air_time, y = delay)) +
  geom_point(color ="blue")+
  geom_smooth(se=F)+
  labs( x="Amount of time spent in the air, in minutes", y="Departure and arrival delays, in
minutes")+
  theme_bw()
```

```
fig2 <- ggplot(data=flights.2013, aes(x = wind_dir, y = delay)) +
  geom_point(color ="blue")+
  geom_smooth(se=F)+
  labs( x="Wind direction (in degrees)", y="Departure and arrival delays, in minutes")+
  theme_bw()
```

```
plot_grid(fig1, fig2)
```

```
# Create Training and Test data
set.seed(100) # setting seed to reproduce results of random sampling
n <- nrow(flights.2013)
trainingRows <- sample(n, 0.7*n)# row indices for training data
training <- flights.2013[trainingRows, ] # model training data

test <- flights.2013[-trainingRows, ]   # test data

lm.delay <- step(lm(delay ~. , data = training), direction ="both")
## Start:  AIC=1731850
## delay ~ air_time + temp + humid + wind_dir + wind_speed + pressure +
##     name + lat + lon + Qtr + TimeOfDay
##
##
## Step:  AIC=1731850
## delay ~ air_time + temp + humid + wind_dir + wind_speed + pressure +
##     name + lat + Qtr + TimeOfDay
##
```

```
##
## Step:  AIC=1731850
## delay ~ air_time + temp + humid + wind_dir + wind_speed + pressure +
##     name + Qtr + TimeOfDay
##
##              Df Sum of Sq   RSS     AIC
## <none>                    1.14e+09 1731850
## - temp       1      201785 1.14e+09 1731883
## - air_time   1      617836 1.14e+09 1731957
## - wind_speed  1     634900 1.14e+09 1731960
## - wind_dir   1      653717 1.14e+09 1731963
## - name       2   2427332 1.14e+09 1732273
## - Qtr        3   2829314 1.14e+09 1732342
## - pressure   1   6918349 1.14e+09 1733064
## - humid      1   9852280 1.15e+09 1733577
## - TimeOfDay   3  39826206 1.18e+09 1738744
summary(lm.delay) %>% pander()
```

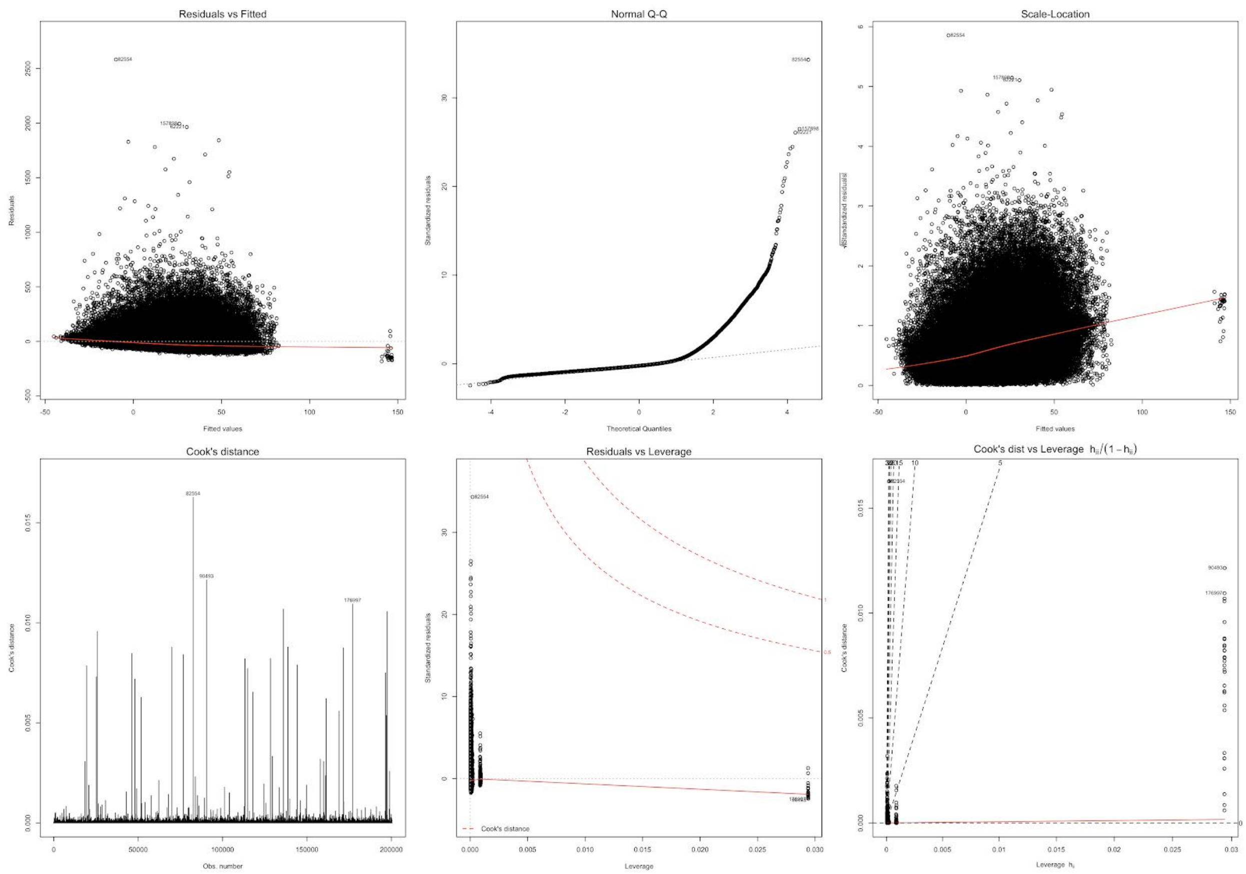
```
# backward elimination
summary(step(lm(delay ~. , data = training), direction ="backward"))
## Start:  AIC=1731850
## delay ~ air_time + temp + humid + wind_dir + wind_speed + pressure +
##     name + lat + lon + Qtr + TimeOfDay
##
##
## Step:  AIC=1731850
## delay ~ air_time + temp + humid + wind_dir + wind_speed + pressure +
##     name + lat + Qtr + TimeOfDay
##
##
## Step:  AIC=1731850
## delay ~ air_time + temp + humid + wind_dir + wind_speed + pressure +
##     name + Qtr + TimeOfDay
##
##              Df Sum of Sq   RSS     AIC
## <none>                   1.14e+09 1731850
## - temp       1      201785 1.14e+09 1731883
## - air_time   1      617836 1.14e+09 1731957
## - wind_speed  1     634900 1.14e+09 1731960
## - wind_dir   1      653717 1.14e+09 1731963
## - name       2   2427332 1.14e+09 1732273
## - Qtr        3   2829314 1.14e+09 1732342
## - pressure   1   6918349 1.14e+09 1733064
## - humid      1   9852280 1.15e+09 1733577
## - TimeOfDay   3  39826206 1.18e+09 1738744
##
## Call:
## lm(formula = delay ~ air_time + temp + humid + wind_dir + wind_speed +
```

```
##     pressure + name + Qtr + TimeOfDay, data = training)
##
## Residuals:
##     Min     1Q Median      3Q     Max
## -181.9  -37.2  -17.7    8.3 2582.9
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             900.95328   27.40668   32.87
## air_time                                 -0.01944    0.00186  -10.44
## temp                                     -0.10223    0.01714   -5.96
## humid                                     0.47569    0.01141   41.67
## wind_dir                                 -0.01871    0.00174  -10.73
## wind_speed                                0.13179    0.01246   10.58
## pressure                                 -0.91896    0.02632  -34.92
## nameLa Guardia Airport                    1.90306    0.44183    4.31
## nameNewark Liberty International Airport   8.05746    0.41088   19.61
## QtrQ2                                    10.30422    0.65904   15.64
## QtrQ3                                     4.72868    0.80115    5.90
## QtrQ4                                    -1.01578    0.53554   -1.90
## TimeOfDayMorning                          6.44013    2.18748    2.94
## TimeOfDayAfternoon                       35.42676    2.19854   16.11
## TimeOfDayEvening                         47.37881    2.21405   21.40
##                                          Pr(>|t|)
## (Intercept)                              < 2e-16 ***
## air_time                                 < 2e-16 ***
## temp                                     2.5e-09 ***
## humid                                    < 2e-16 ***
## wind_dir                                 < 2e-16 ***
## wind_speed                               < 2e-16 ***
## pressure                                 < 2e-16 ***
## nameLa Guardia Airport                   1.7e-05 ***
## nameNewark Liberty International Airport  < 2e-16 ***
## QtrQ2                                    < 2e-16 ***
## QtrQ3                                    3.6e-09 ***
## QtrQ4                                    0.0579 .
## TimeOfDayMorning                         0.0032 **
## TimeOfDayAfternoon                       < 2e-16 ***
## TimeOfDayEvening                         < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75.3 on 200348 degrees of freedom
## Multiple R-squared:  0.0557, Adjusted R-squared:  0.0556
## F-statistic:  844 on 14 and 2e+05 DF,  p-value: <2e-16
# forward elimination
summary(step(lm(delay ~. , data = training), direction ="forward"))
## Start:  AIC=1731850
## delay ~ air_time + temp + humid + wind_dir + wind_speed + pressure +
##     name + lat + lon + Qtr + TimeOfDay
```

```
##
## Call:
## lm(formula = delay ~ air_time + temp + humid + wind_dir + wind_speed +
##     pressure + name + lat + lon + Qtr + TimeOfDay, data = training)
##
## Residuals:
##     Min     1Q Median      3Q     Max
## -181.9  -37.2  -17.7    8.3 2582.9
##
## Coefficients: (2 not defined because of singularities)
##                                         Estimate Std. Error t value
## (Intercept)                            900.95328   27.40668   32.87
## air_time                               -0.01944       0.00186  -10.44
## temp                                    -0.10223      0.01714   -5.96
## humid                                    0.47569      0.01141   41.67
## wind_dir                               -0.01871       0.00174  -10.73
## wind_speed                                  0.13179   0.01246   10.58
## pressure                               -0.91896       0.02632  -34.92
## nameLa Guardia Airport                 1.90306 0.44183    4.31
## nameNewark Liberty International Airport   8.05746    0.41088   19.61
## lat                                          NA      NA      NA
## lon                                          NA      NA      NA
## QtrQ2                                   10.30422      0.65904   15.64
## QtrQ3                                      4.72868    0.80115    5.90
## QtrQ4                                   -1.01578      0.53554   -1.90
## TimeOfDayMorning                       6.44013 2.18748    2.94
## TimeOfDayAfternoon                     35.42676       2.19854   16.11
## TimeOfDayEvening                       47.37881       2.21405   21.40
##                                        Pr(>|t|)
## (Intercept)                            < 2e-16 ***
## air_time                               < 2e-16 ***
## temp                                   2.5e-09 ***
## humid                                  < 2e-16 ***
## wind_dir                               < 2e-16 ***
## wind_speed                             < 2e-16 ***
## pressure                              < 2e-16 ***
## nameLa Guardia Airport                 1.7e-05 ***
## nameNewark Liberty International Airport  < 2e-16 ***
## lat                                          NA
## lon                                          NA
## QtrQ2                                  < 2e-16 ***
## QtrQ3                                  3.6e-09 ***
## QtrQ4                                  0.0579 .
## TimeOfDayMorning                       0.0032 **
## TimeOfDayAfternoon                      < 2e-16 ***
## TimeOfDayEvening                       < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75.3 on 200348 degrees of freedom
```

```
## Multiple R-squared:  0.0557, Adjusted R-squared:  0.0556
## F-statistic:  844 on 14 and 2e+05 DF,  p-value: <2e-16
# stepwise regression
summary(step(lm(delay ~. , data = training), direction ="both"))
## Start:  AIC=1731850
## delay ~ air_time + temp + humid + wind_dir + wind_speed + pressure +
##     name + lat + lon + Qtr + TimeOfDay
##
##
## Step:  AIC=1731850
## delay ~ air_time + temp + humid + wind_dir + wind_speed + pressure +
##     name + lat + Qtr + TimeOfDay
##
##
## Step:  AIC=1731850
## delay ~ air_time + temp + humid + wind_dir + wind_speed + pressure +
##     name + Qtr + TimeOfDay
##
##              Df Sum of Sq   RSS     AIC
## <none>                     1.14e+09 1731850
## - temp       1     201785 1.14e+09 1731883
## - air_time   1     617836 1.14e+09 1731957
## - wind_speed  1    634900 1.14e+09 1731960
## - wind_dir   1     653717 1.14e+09 1731963
## - name       2    2427332 1.14e+09 1732273
## - Qtr        3    2829314 1.14e+09 1732342
## - pressure   1    6918349 1.14e+09 1733064
## - humid      1    9852280 1.15e+09 1733577
## - TimeOfDay   3  39826206 1.18e+09 1738744
##
## Call:
## lm(formula = delay ~ air_time + temp + humid + wind_dir + wind_speed +
##     pressure + name + Qtr + TimeOfDay, data = training)
##
## Residuals:
##     Min     1Q Median      3Q     Max
## -181.9  -37.2  -17.7    8.3 2582.9
##
## Coefficients:
##                                      Estimate Std. Error t value
## (Intercept)                         900.95328   27.40668   32.87
## air_time                             -0.01944    0.00186  -10.44
## temp                                 -0.10223    0.01714   -5.96
## humid                                 0.47569    0.01141   41.67
## wind_dir                             -0.01871    0.00174  -10.73
## wind_speed                            0.13179    0.01246   10.58
## pressure                             -0.91896    0.02632  -34.92
## nameLa Guardia Airport                1.90306    0.44183    4.31
## nameNewark Liberty International Airport   8.05746    0.41088   19.61
## QtrQ2                                10.30422    0.65904   15.64
```

```
## QtrQ3                                      4.72868      0.80115 5.90
## QtrQ4                                     -1.01578      0.53554   -1.90
## TimeOfDayMorning                          6.44013 2.18748    2.94
## TimeOfDayAfternoon                        35.42676       2.19854   16.11
## TimeOfDayEvening                          47.37881       2.21405   21.40
##                                           Pr(>|t|)
## (Intercept)                               < 2e-16 ***
## air_time                                  < 2e-16 ***
## temp                                      2.5e-09 ***
## humid                                       < 2e-16 ***
## wind_dir                                  < 2e-16 ***
## wind_speed                                < 2e-16 ***
## pressure                                  < 2e-16 ***
## nameLa Guardia Airport                    1.7e-05 ***
## nameNewark Liberty International Airport  < 2e-16 ***
## QtrQ2                                     < 2e-16 ***
## QtrQ3                                     3.6e-09 ***
## QtrQ4                                     0.0579 .
## TimeOfDayMorning                              0.0032 **
## TimeOfDayAfternoon                        < 2e-16 ***
## TimeOfDayEvening                          < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75.3 on 200348 degrees of freedom
## Multiple R-squared:  0.0557, Adjusted R-squared:  0.0556
## F-statistic:  844 on 14 and 2e+05 DF,  p-value: <2e-16
# 6-Plot of Fit
par(mfrow= c(2,3))# creates six panels for plotting
plot(lm.delay, which = 1:6)
```


Residuals vs Fitted
Fitted values
Residuals
Normal Q-Q
Theoretical Quantiles
Standardized residuals
Scale-Location
Fitted values
Cook's distance
Obs. number
Cook's distance
Residuals vs Leverage
Leverage
Standardized residuals
Cook's distance
Cook's dist vs Leverage $h_{ii}/(1-h_{ii})$
Leverage h_{ii}
Cook's distance