Flight Delay Analysis

By: Sena Hiraoka, Yi-Lin Liao

Flight Dataset

- The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers.

Data Source:

- Kaggle - https://www.kaggle.com/
- Dataset on Kaggle - https://www.kaggle.com/usdot/flight-delays

Basic Info of Dataset

```
> # Basic Info of Dataset
> print (format (object.size (flightData), units="Mb"))   # storage in megabytes
[1] "799.5 Mb"
> print (c ("Number of columns: ", ncol (flightData)))
[1] "Number of columns: " "31"
> print (c ("Number of rows: ", nrow (flightData)))
[1] "Number of rows: " "5819079"
> print (colnames (flightData))
 [1] "YEAR"                "MONTH"               "DAY"                 "DAY_OF_WEEK"         "AIRLINE"             "FLIGHT_NUMBER"       "TAIL_NUMBER"
 [8] "ORIGIN_AIRPORT"      "DESTINATION_AIRPORT" "SCHEDULED_DEPARTURE" "DEPARTURE_TIME"      "DEPARTURE_DELAY"     "TAXI_OUT"            "WHEELS_OFF"
[15] "SCHEDULED_TIME"      "ELAPSED_TIME"        "AIR_TIME"            "DISTANCE"            "WHEELS_ON"           "TAXI_IN"             "SCHEDULED_ARRIVAL"
[22] "ARRIVAL_TIME"        "ARRIVAL_DELAY"       "DIVERTED"            "CANCELLED"           "CANCELLATION_REASON" "AIR_SYSTEM_DELAY"    "SECURITY_DELAY"
[29] "AIRLINE_DELAY"       "LATE_AIRCRAFT_DELAY" "WEATHER_DELAY"
>
```

Cleaning and Transformations

Clean

- Remove cancelled and diverted flight - subset
- Remove unused columns – select
- Remove rows with NAs – na.omit

Transform

- Create a column with values night, daytime, and evening based on the scheduled departure time – mutate
- Change character columns to factors
- (Optional) Take only around 100,000 rows

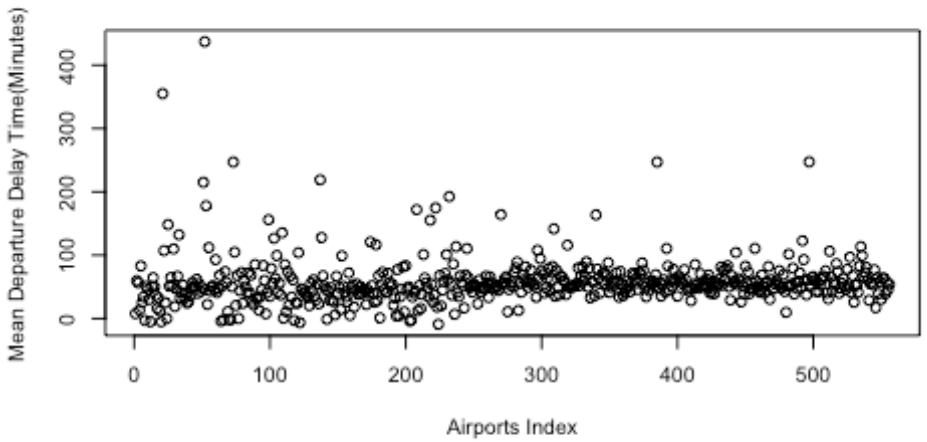
Structure of Dataset

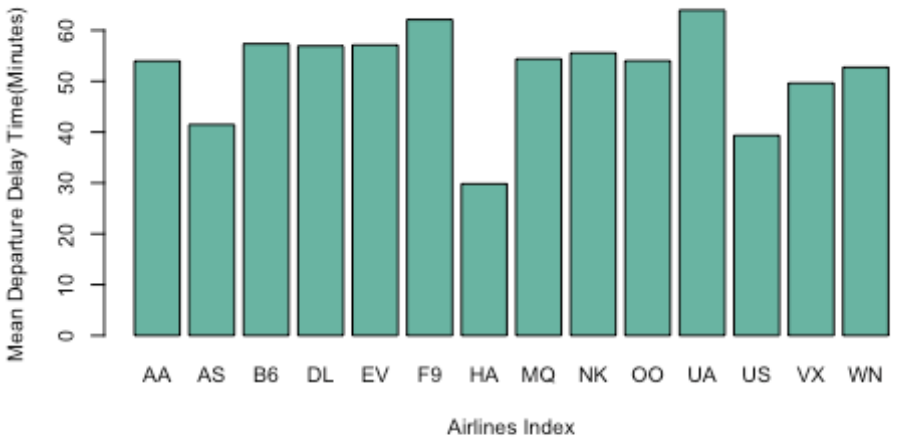
```
'data.frame':   106343 obs. of  22 variables:
 $ MONTH                          : int  1 1 1 1 1 1 1 1 1 1 ...
 $ DAY                            : int  1 1 1 1 1 1 1 1 1 1 ...
 $ DAY_OF_WEEK                    : int  4 4 4 4 4 4 4 4 4 4 ...
 $ AIRLINE                        : Factor w/ 14 levels "AA","AS","B6",..: 1 1 11 10 10 14 10 10 5 1 ...
 $ ORIGIN_AIRPORT                 : Factor w/ 625 levels "10135","10136",..: 455 389 591 389 531 306 342 367 357 453 ...
 $ DESTINATION_AIRPORT            : Factor w/ 625 levels "10135","10136",..: 507 390 389 480 389 447 480 480 455 390 ...
 $ DEPARTURE_TIME                 : int  618 623 634 658 758 607 639 654 621 622 ...
 $ DEPARTURE_DELAY                : int  58 53 56 73 119 7 39 54 1 -3 ...
 $ SCHEDULED_TIME                 : int  141 125 142 150 141 120 130 51 161 210 ...
 $ ELAPSED_TIME                   : int  137 138 164 133 138 135 121 85 184 230 ...
 $ AIR_TIME                       : int  111 96 129 112 123 103 89 24 154 200 ...
 $ DISTANCE                       : int  964 641 909 862 819 759 674 86 878 1172 ...
 $ WHEELS_ON                      : int  928 931 1009 801 1105 919 732 806 820 856 ...
 $ TAXI_IN                        : int  7 10 9 10 11 3 8 13 5 16 ...
 $ SCHEDULED_ARRIVAL              : int  841 835 900 715 920 900 710 651 801 855 ...
 $ ARRIVAL_DELAY                  : int  54 66 78 56 116 22 30 88 24 17 ...
 $ AIR_SYSTEM_DELAY               : int  0 13 22 0 0 15 0 34 23 17 ...
 $ SECURITY_DELAY                 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AIRLINE_DELAY                  : int  54 53 56 56 0 7 30 0 1 0 ...
 $ LATE_AIRCRAFT_DELAY            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ WEATHER_DELAY                  : int  0 0 0 0 116 0 0 54 0 0 ...
 $ SCHEDULED_DEPARTURE_TIME_GROUP: Factor w/ 3 levels "Daytime","Evening",..: 3 3 3 3 3 3 3 3 3 3 ...
 - attr(*, "na.action")= 'omit' Named int [1:4650569] 1 2 3 4 5 6 7 8 9 10 ...
  ..- attr(*, "names")= chr [1:4650569] "1" "2" "3" "4" ...
```

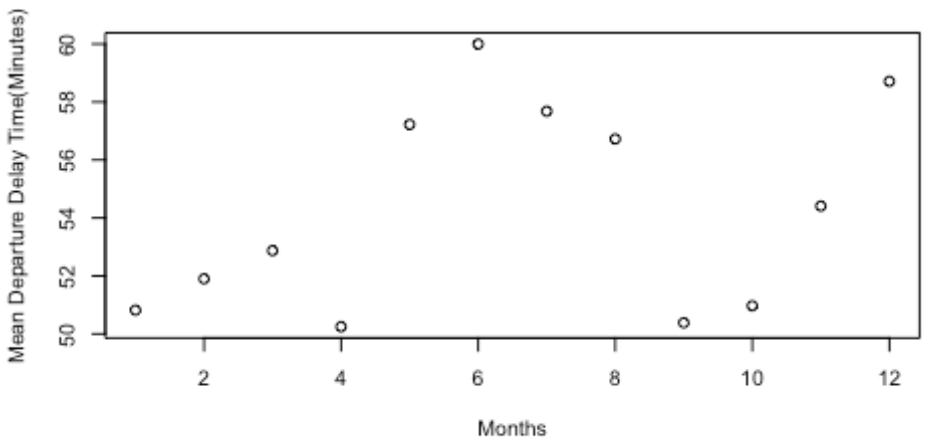
Summary of columns in dataset

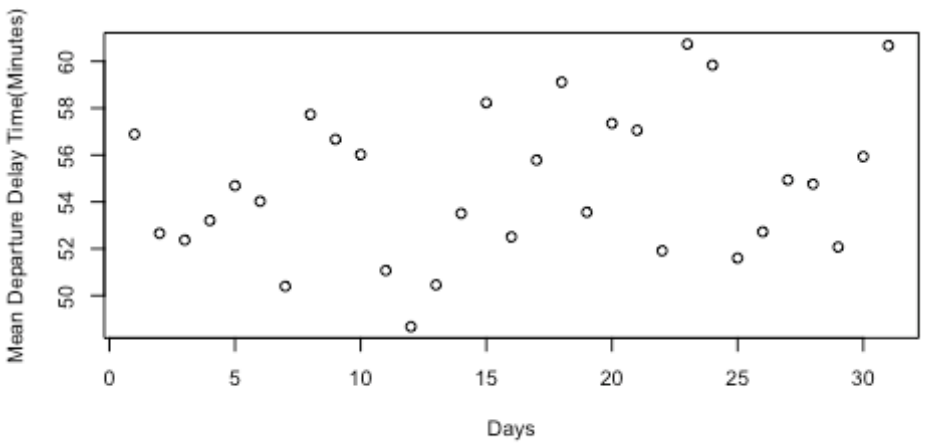
```
> summary(flightData)
     MONTH              DAY          DAY_OF_WEEK       AIRLINE      ORIGIN_AIRPORT  DESTINATION_AIRPORT DEPARTURE_TIME DEPARTURE_DELAY   SCHEDULED_TIME   ELAPSED_TIME
 Min.   : 1.000   Min.   : 1.00   Min.   :1.000   WN     :23807   ORD    : 6665   ORD    : 5890      Min.   :   1   Min.   : -28.00   Min.   : 20.0   Min.   : 16.0
 1st Qu.: 3.000   1st Qu.: 8.00   1st Qu.:2.000   AA     :13002   ATL    : 5643   ATL    : 5148      1st Qu.:1156   1st Qu.:  17.00   1st Qu.: 86.0   1st Qu.: 90.0
 Median : 6.000   Median :16.00   Median :4.000   DL     :11710   DFW    : 5046   DFW    : 4488      Median :1605   Median :  37.00   Median :125.0   Median :130.0
 Mean   : 6.232   Mean   :15.58   Mean   :3.879   EV     :10880   DEN    : 4261   LAX    : 4413      Mean   :1527   Mean   :  54.75   Mean   :143.3   Mean   :147.3
 3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:5.000   OO     :10743   LAX    : 4019   DEN    : 3789      3rd Qu.:1924   3rd Qu.:  70.00   3rd Qu.:175.0   3rd Qu.:182.0
 Max.   :12.000   Max.   :31.00   Max.   :7.000   UA     :10391   IAH    : 3107   SFO    : 3254      Max.   :2400   Max.   :1450.00   Max.   :705.0   Max.   :724.0
                                                  (Other):25810   (Other):77602   (Other):79361
    AIR_TIME        DISTANCE        WHEELS_ON        TAXI_IN      SCHEDULED_ARRIVAL ARRIVAL_DELAY     AIR_SYSTEM_DELAY SECURITY_DELAY      AIRLINE_DELAY    LATE_AIRCRAFT_DELAY
 Min.   :  9.0   Min.   :  31.0   Min.   :   1   Min.   :  1.000   Min.   :   1     Min.   :  15.00   Min.   :  0.00   Min.   :  0.00000   Min.   :   0.0   Min.   :   0.00
 1st Qu.: 62.0   1st Qu.: 383.0   1st Qu.:1241   1st Qu.:  4.000   1st Qu.:1300     1st Qu.:  23.00   1st Qu.:  0.00   1st Qu.:  0.00000   1st Qu.:   0.0   1st Qu.:   0.00
 Median : 99.0   Median : 677.0   Median :1717   Median :  6.000   Median :1715     Median :  37.00   Median :  2.00   Median :  0.00000   Median :   2.0   Median :   3.00
 Mean   :117.8   Mean   : 835.2   Mean   :1579   Mean   :  8.761   Mean   :1620     Mean   :  58.77   Mean   : 13.55   Mean   :  0.06384   Mean   :  18.9   Mean   :  23.33
 3rd Qu.:149.0   3rd Qu.:1074.0   3rd Qu.:2038   3rd Qu.:  9.000   3rd Qu.:2025     3rd Qu.:  70.00   3rd Qu.: 18.00   3rd Qu.:  0.00000   3rd Qu.:  19.0   3rd Qu.:  29.00
 Max.   :670.0   Max.   :4983.0   Max.   :2400   Max.   :176.000   Max.   :2359     Max.   :1444.00   Max.   :738.00   Max.   :221.00000   Max.   :1444.0   Max.   :1102.00

 WEATHER_DELAY     SCHEDULED_DEPARTURE_TIME_GROUP
 Min.   :  0.000   Daytime:49108
 1st Qu.:  0.000   Evening:47996
 Median :  0.000   NIGHT  : 9239
 Mean   :  2.926
 3rd Qu.:  0.000
 Max.   :995.000
```

Visualizations – Departure Delay

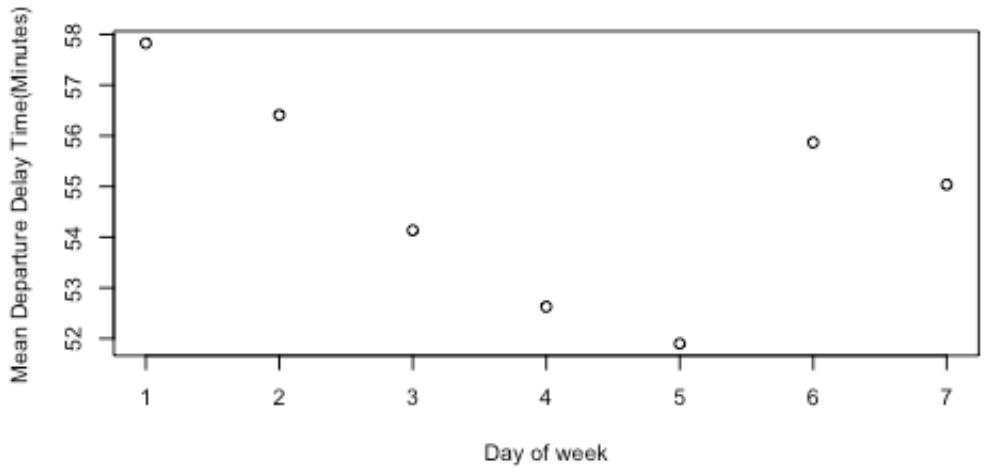


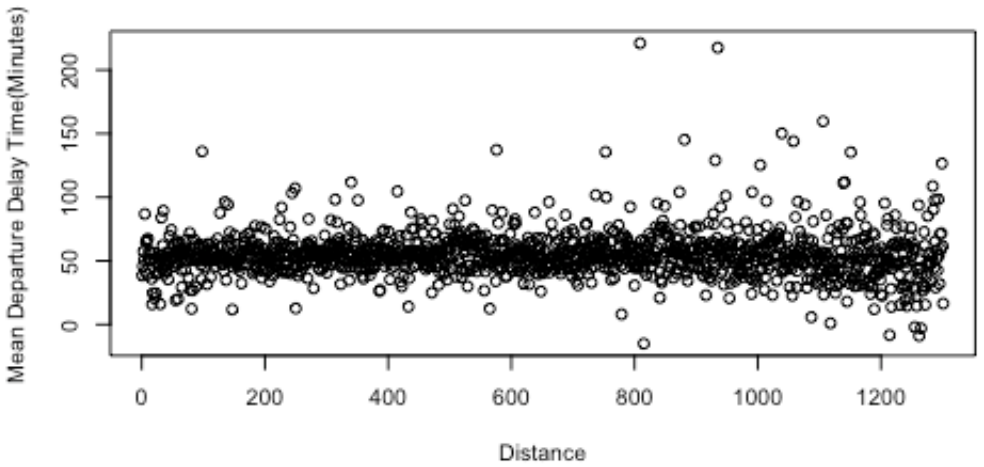


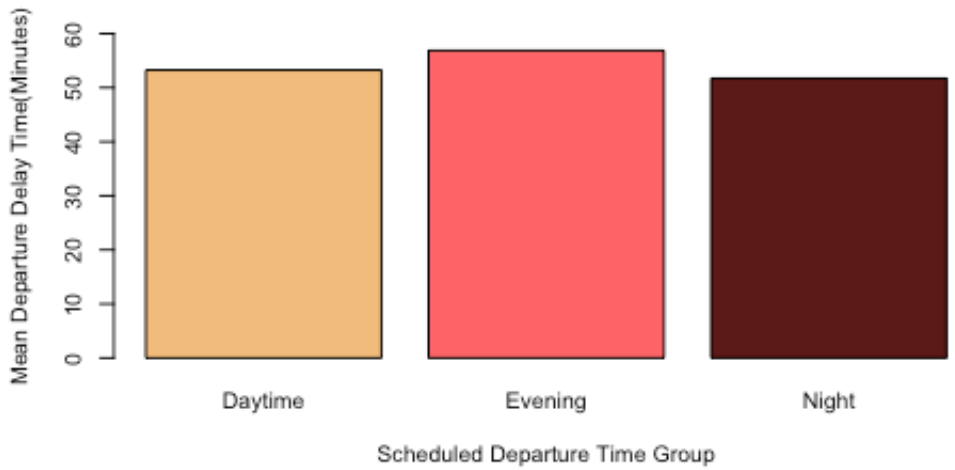




Visualizations







Prediction Model Preparation

- Training data to Testing data ratio is 7:3 (70000 to 30000 records)
- Conducted K-fold Cross validation on models using training dataset.

```
> print(model)
Linear Regression

70000 samples
   21 predictor

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 46667, 46666, 46667
Resampling results:

  RMSE          Rsquared  MAE
  4.725147e-12  1         4.503171e-12

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
> print(model2)
Linear Regression

70000 samples
    5 predictor

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 46667, 46668, 46665
Resampling results:

  RMSE          Rsquared  MAE
  5.98378e-12  1         5.415189e-12

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Create Prediction Model for Delay time

```
> summary(model)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
       Min         1Q     Median         3Q        Max
-2.519e-09 -1.800e-13 -2.000e-14  1.300e-13  1.047e-10

Coefficients: (184 not defined because of singularities)
                                    Estimate Std. Error    t value Pr(>|t|)
(Intercept)                       -9.947e-12  9.695e-12 -1.026e+00  0.30490
MONTH                             -9.279e-15  1.155e-14 -8.040e-01  0.42163
DAY                               -1.908e-15  4.214e-15 -4.530e-01  0.65071
DAY_OF_WEEK                        2.328e-14  1.865e-14  1.249e+00  0.21177
AIRLINEAS                         -2.755e-13  3.663e-13 -7.520e-01  0.45195
AIRLINEB6                          6.297e-13  2.477e-13  2.543e+00  0.01101 *
AIRLINEDL                          8.095e-14  2.024e-13  4.000e-01  0.68914
AIRLINEEV                         -8.622e-14  2.280e-13 -3.780e-01  0.70534
AIRLINEF9                          2.211e-13  3.010e-13  7.350e-01  0.46248
AIRLINEHA                          1.684e-12  7.347e-13  2.292e+00  0.02190 *
AIRLINEMQ                          2.089e-13  2.365e-13  8.830e-01  0.37698
AIRLINENK                          1.217e-13  2.667e-13  4.560e-01  0.64809
AIRLINEOO                          3.246e-13  2.183e-13  1.487e+00  0.13698
AIRLINEUA                          4.841e-13  2.002e-13  2.418e+00  0.01559 *
AIRLINEUS                         -6.434e-15  2.583e-13 -2.500e-02  0.98013
AIRLINEVX                          1.772e-13  3.952e-13  4.480e-01  0.65389
AIRLINEWN                          5.043e-13  1.940e-13  2.599e+00  0.00934 **
ORIGIN_AIRPORT10136                       NA         NA         NA       NA
ORIGIN_AIRPORT10140                1.617e-12  7.241e-12  2.230e-01  0.82328
ORIGIN_AIRPORT10141                       NA         NA         NA       NA
ORIGIN_AIRPORT10146                1.152e-12  9.696e-12  1.190e-01  0.90541
ORIGIN_AIRPORT10154                       NA         NA         NA       NA
```

```
> summary(model2)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
       Min         1Q     Median         3Q        Max
-2.524e-09 -7.000e-14  0.000e+00  9.000e-14  5.207e-10

Coefficients:
                                             Estimate Std. Error    t value Pr(>|t|)
(Intercept)                                -1.000e-11  1.524e-13 -6.560e+01  < 2e-16 ***
AIRLINEAS                                   1.118e-12  2.756e-13  4.057e+00 4.98e-05 ***
AIRLINEB6                                   3.874e-14  1.896e-13  2.040e-01 0.838041
AIRLINEDL                                   4.865e-13  1.544e-13  3.151e+00 0.001629 **
AIRLINEEV                                  -4.939e-12  1.617e-13 -3.054e+01  < 2e-16 ***
AIRLINEF9                                   1.408e-12  2.690e-13  5.235e+00 1.66e-07 ***
AIRLINEHA                                   3.996e-13  4.176e-13  9.570e-01 0.338722
AIRLINEMQ                                   3.624e-12  1.917e-13  1.890e+01  < 2e-16 ***
AIRLINENK                                   4.496e-12  2.329e-13  1.931e+01  < 2e-16 ***
AIRLINEOO                                   2.269e-12  1.621e-13  1.400e+01  < 2e-16 ***
AIRLINEUA                                   2.356e-12  1.601e-13  1.472e+01  < 2e-16 ***
AIRLINEUS                                   8.918e-13  2.254e-13  3.958e+00 7.58e-05 ***
AIRLINEVX                                   1.254e-12  3.666e-13  3.422e+00 0.000623 ***
AIRLINEWN                                   1.212e-12  1.352e-13  8.965e+00  < 2e-16 ***
SCHEDULED_TIME                              1.000e+00  2.044e-15  4.893e+14  < 2e-16 ***
ELAPSED_TIME                               -1.000e+00  1.970e-15 -5.076e+14  < 2e-16 ***
ARRIVAL_DELAY                               1.000e+00  5.802e-16  1.723e+15  < 2e-16 ***
SCHEDULED_DEPARTURE_TIME_GROUPEvening      -1.018e-13  7.811e-14 -1.304e+00 0.192294
SCHEDULED_DEPARTURE_TIME_GROUPNIGHT        -4.907e-15  1.369e-13 -3.600e-02 0.971409
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

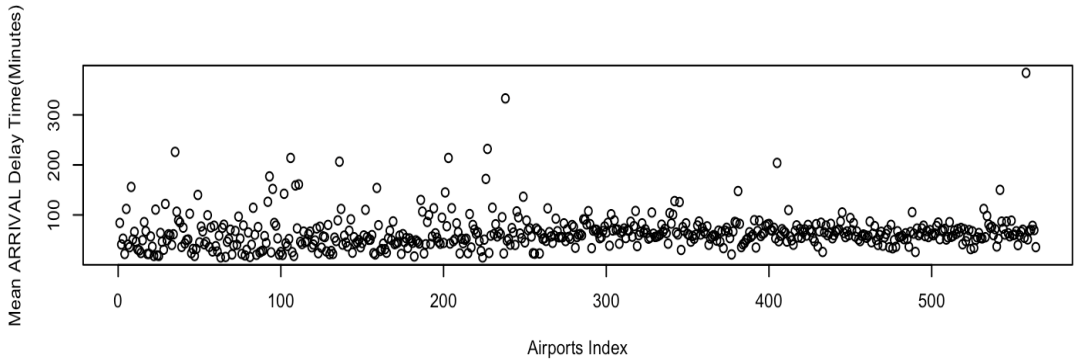
Prediction Model Validation

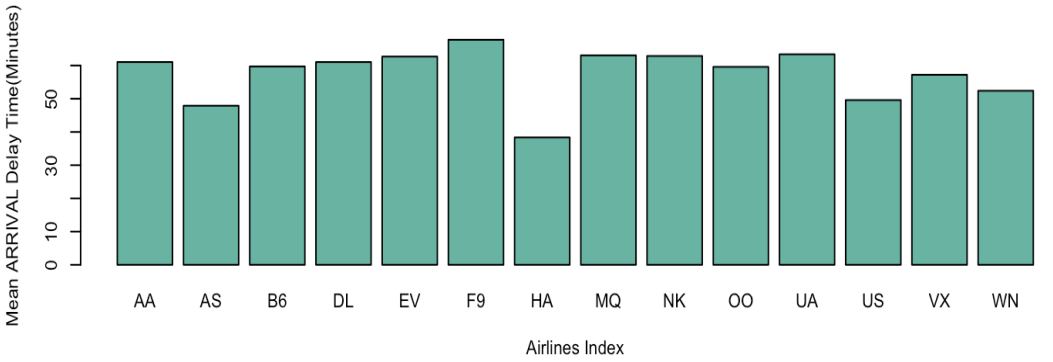
```
> rmse(predictions,test_df$DEPARTURE_DELAY)
[1] 9.615932e-12
```

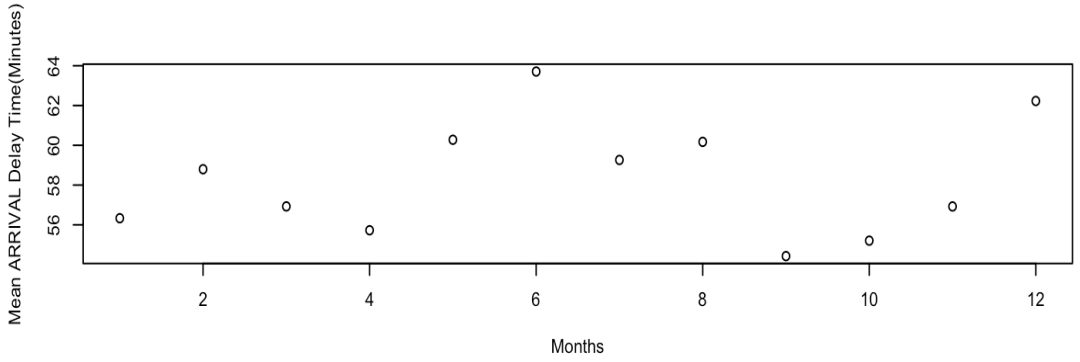
```
> rmse(predictions2,test_df$DEPARTURE_DELAY)
[1] 9.82185e-12
```

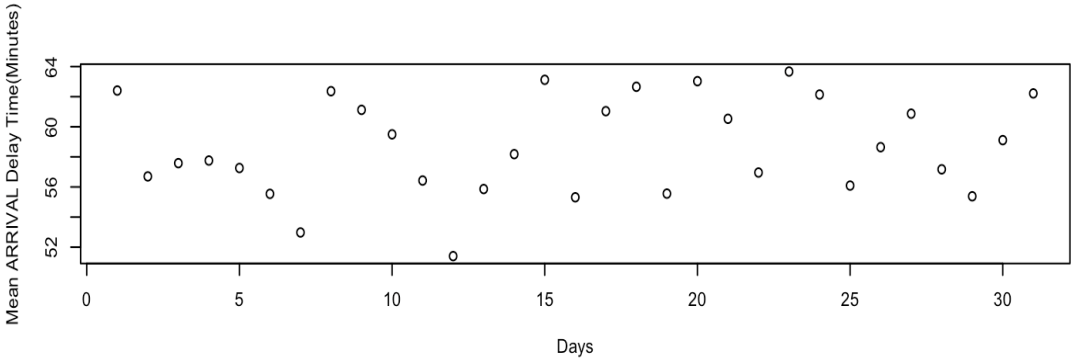
Now let’s look at Arrival Delay

Arrival Delay Visualizations

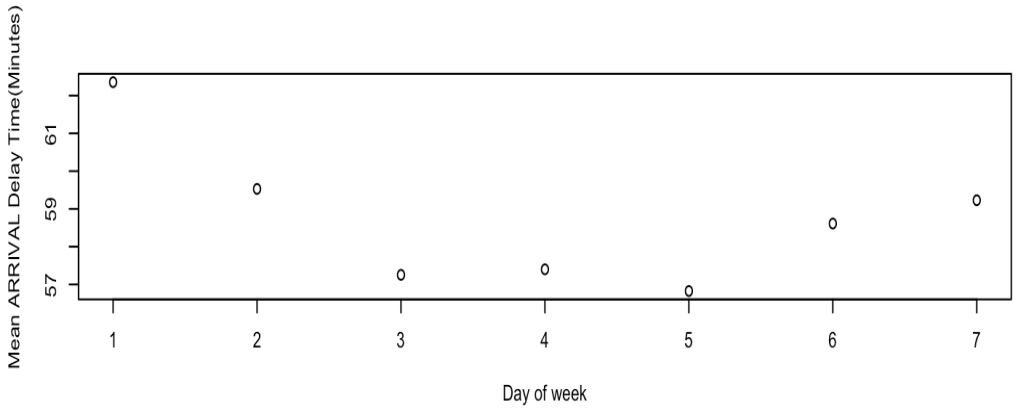


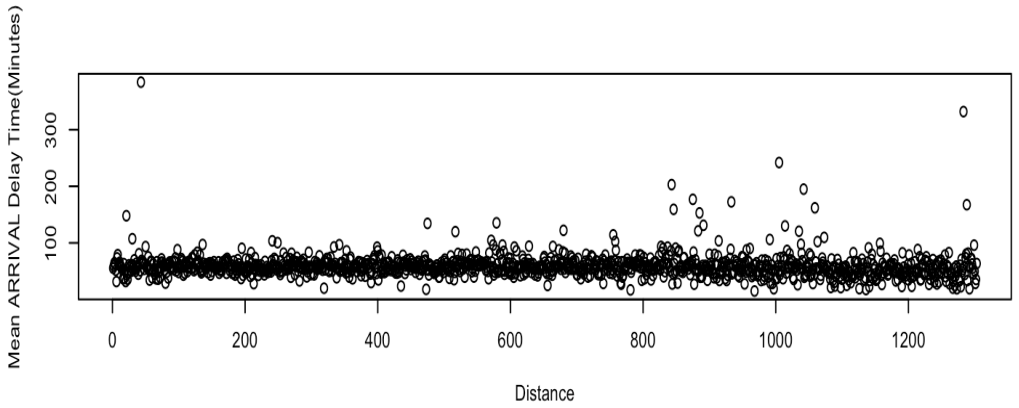


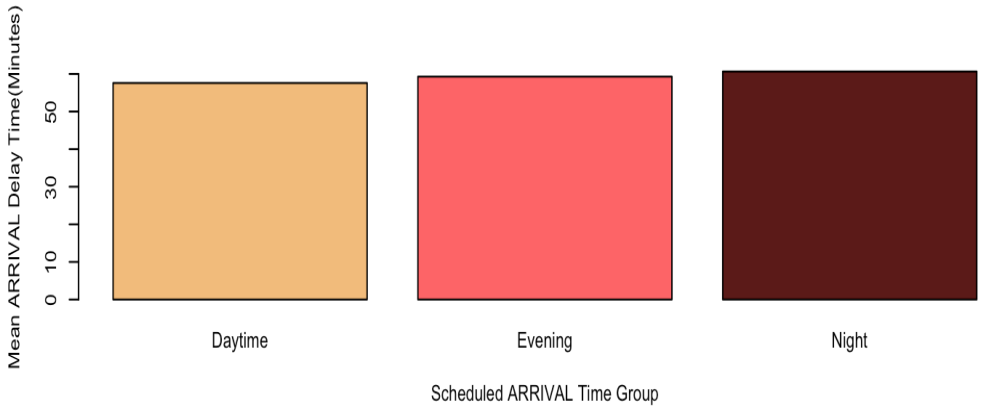




Arrival Delay Visualizations







Arrival Delay

```
> print(model)
Linear Regression

70000 samples
   24 predictor

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 46667, 46666, 46667
Resampling results:

  RMSE          Rsquared  MAE
  1.739963e-12  1         1.60676e-12

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
> print(model2)
Linear Regression

70000 samples
    5 predictor

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 46667, 46666, 46667
Resampling results:

  RMSE          Rsquared  MAE
  2.630467e-12  1         2.185625e-12

Tuning parameter 'intercept' was held constant at a value of TRUE
>
```

Arrival Delay

```
> summary(model)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
       Min         1Q     Median         3Q        Max
-2.601e-10 -1.400e-13  4.000e-14  2.200e-13  2.466e-09

Coefficients: (159 not defined because of singularities)
                                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)                      6.897e-12  6.200e-12  1.210e-01  0.90395
MONTH                            1.375e-15  1.139e-14 -1.470e-01  0.88343
DAY                             -6.083e-16  4.149e-15  2.230e-01  0.82338
DAY_OF_WEEK                      4.095e-15  1.835e-14  2.664e+00  0.00773 **
AIRLINEAS                        9.739e-13  3.656e-13 -2.178e+00  0.02938 *
AIRLINEB6                       -5.311e-13  2.438e-13  1.335e+00  0.18186
AIRLINEDL                        2.654e-13  1.988e-13  1.862e+00  0.06263 .
AIRLINEEV                        4.168e-13  2.238e-13  7.699e+00 1.39e-14 ***
AIRLINEF9                        2.305e-12  2.994e-13  4.905e+00 9.36e-07 ***
AIRLINEHA                        3.633e-12  7.406e-13  4.108e+00 4.00e-05 ***
AIRLINEMQ                        9.598e-13  2.336e-13 -9.150e-01  0.36037
AIRLINENK                       -2.428e-13  2.654e-13  5.286e+00 1.26e-07 ***
```

```
> summary(model2)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
       Min         1Q     Median         3Q        Max
-4.308e-10 -5.000e-14 -2.000e-14  1.000e-14  2.466e-09

Coefficients:
                 Estimate Std. Error    t value Pr(>|t|)
(Intercept)      1.511e-11  1.612e-13  9.370e+01  < 2e-16 ***
AIRLINEAS        8.099e-12  2.823e-13  2.869e+01  < 2e-16 ***
AIRLINEB6       -3.779e-12  1.924e-13 -1.964e+01  < 2e-16 ***
AIRLINEDL       -3.559e-12  1.567e-13 -2.272e+01  < 2e-16 ***
AIRLINEEV       -9.569e-12  1.638e-13 -5.841e+01  < 2e-16 ***
AIRLINEF9       -3.194e-12  2.740e-13 -1.166e+01  < 2e-16 ***
AIRLINEHA       -2.952e-12  4.280e-13 -6.898e+00 5.32e-12 ***
AIRLINEMQ       -4.039e-12  1.932e-13 -2.090e+01  < 2e-16 ***
AIRLINENK       -2.119e-12  2.386e-13 -8.880e+00  < 2e-16 ***
AIRLINEOO       -2.766e-12  1.641e-13 -1.686e+01  < 2e-16 ***
AIRLINEUA       -2.723e-12  1.616e-13 -1.685e+01  < 2e-16 ***
AIRLINEUS       -3.980e-12  2.273e-13 -1.751e+01  < 2e-16 ***
```

Arrival Delay

```
> rmse(predictions,test_df$ARRIVAL_DELAY)
[1] 9.439005e-12
```

```
> rmse(predictions2,test_df$ARRIVAL_DELAY)
[1] 9.957019e-12
```

Model 1 (using all attributes) has lower RMSE for both departure and arrival delay prediction.

Conclusion

1. Linear Regression maybe not suitable for predicting delay time.
2. Model 2 may have too less columns and be a bit overfitting.
3. K fold only 3 folds so both models aren’t trained enough to show more accurate results.