Midterm SFO

Jim Turman

November 18, 2016

Introduction

Below are the descriptive statistics, ANOVA and regression analysis of flight data for the SFO airport with regards to on time departure and arrival of flights. For the ANOVA testing a sample of 1000 observations is used where the departure time negatively effects the flight time meaning that there was a delay greater than zero. Attached with this file are plots describing the linear regression model that was done to predict the arrival time of flights leaving SFO.

Descriptive Statistics

Descriptive Statistics for Airline Carriers

These are the descriptive statistics for the Airline Carriers and the delays associated with them. Here we can see that the airline carrier EV has the largest mean delay time (in minutes) as well as the largest median delay time which might suggest that they are the worst airline of the sample with regards to departure delay. This also shows that the Airline Carrier HA has the smallest average delay time of 15.058 minutes and a median of 6.5 minutes suggesting that they might have the the smallest average departure delay. Below those descriptive statistics are another set describing the Departure Delay as well as the types of delays that we have data for. The largest average delay time is due to late aircraft and the smallest average delay time is due to security delays. Following that block are the relative frequency of delays that have an effect on departure time (delay time is greater than zero) which would mean a late departure. From the given data set we can see that late aircraft delays represent the largest amount of delays with 2387 and security delays the least amount with 4.

```
## Loading required package: tcltk
```

```
## $AA
##                sum               mean             median
##            "46177"            "46.177"               "23"
##               mode                var                 sd
##          "numeric" "6288.36203303303" "79.2991931423834"
##                  n           kurtosis               skew
##             "1000" "59.3850564388431" "6.29135739153806"
##                max                min              range
##             "1095"                "1"             "1094"
##       quartiles.0%      quartiles.25%      quartiles.50%
##                "1"                "7"               "23"
##      quartiles.75%     quartiles.100%                iqr
##               "56"             "1095"               "49"
##
## $AS
##                sum               mean             median
##            "22678"            "22.678"               "11"
##               mode                var                 sd
##          "numeric" "1178.27859459459" "34.3260629055328"
##                  n           kurtosis               skew
```

```
##             "1000" "33.1298395383019" "4.49099624349342"
##                max               min             range
##              "429"               "1"             "428"
##       quartiles.0%     quartiles.25%     quartiles.50%
##                "1"               "4"              "11"
##      quartiles.75%    quartiles.100%               iqr
##               "28"             "429"              "24"
##
## $B6
##                sum              mean            median
##            "44929"          "44.929"              "22"
##               mode               var                sd
##          "numeric" "3581.1370960961" "59.842602684844"
##                  n          kurtosis              skew
##             "1000" "12.7568916207651" "2.86519281794865"
##                max               min             range
##              "600"               "1"             "599"
##       quartiles.0%     quartiles.25%     quartiles.50%
##                "1"               "7"              "22"
##      quartiles.75%    quartiles.100%               iqr
##               "57"             "600"              "50"
##
## $DL
##                sum              mean            median
##            "53862"          "53.862"              "17"
##               mode               var                sd
##          "numeric" "7887.1400960961" "88.8095720972469"
##                  n          kurtosis              skew
##             "1000" "16.9401625848085" "3.40976123778738"
##                max               min             range
##              "813"               "1"             "812"
##       quartiles.0%     quartiles.25%     quartiles.50%
##                "1"               "6"              "17"
##      quartiles.75%    quartiles.100%               iqr
##               "61"             "813"              "55"
##
## $EV
##                sum              mean            median
##            "55690"           "55.69"              "29"
##               mode               var                sd
##          "numeric" "6720.33223223223" "81.9776325117543"
##                  n          kurtosis              skew
##             "1000"  "41.328777557351"  "5.0502858581609"
##                max               min             range
##              "954"               "1"             "953"
##       quartiles.0%     quartiles.25%     quartiles.50%
##                "1"              "10"              "29"
##      quartiles.75%    quartiles.100%               iqr
##               "69"             "954"              "59"
##
## $F9
##                sum              mean            median
##            "45330"           "45.33"              "25"
##               mode               var                sd
```

```
##          "numeric" "4070.16726726727"  "63.797862560334"
##                  n           kurtosis               skew
##             "1000"  "22.801820086359" "3.93514054274555"
##                max                min              range
##              "683"                "1"              "682"
##       quartiles.0%      quartiles.25%      quartiles.50%
##                "1"               "10"               "25"
##      quartiles.75%     quartiles.100%                iqr
##            "55.25"              "683"            "45.25"
##
## $HA
##                sum               mean             median
##            "15058"           "15.058"              "6.5"
##               mode                var                 sd
##          "numeric" "1380.60323923924" "37.1564696821326"
##                  n           kurtosis               skew
##             "1000" "250.571100168815" "13.0648249765298"
##                max                min              range
##              "836"                "1"              "835"
##       quartiles.0%      quartiles.25%      quartiles.50%
##                "1"                "3"              "6.5"
##      quartiles.75%     quartiles.100%                iqr
##               "15"              "836"               "12"
##
## $NK
##                sum               mean             median
##            "51338"           "51.338"               "26"
##               mode                var                 sd
##          "numeric" "6197.48924524525" "78.7241338170529"
##                  n           kurtosis               skew
##             "1000" "45.9699398453598"  "5.3159704705498"
##                max                min              range
##             "1011"                "1"             "1010"
##       quartiles.0%      quartiles.25%      quartiles.50%
##                "1"                "9"               "26"
##      quartiles.75%     quartiles.100%                iqr
##               "63"             "1011"               "54"
##
## $OO
##                sum               mean             median
##            "41316"           "41.316"               "19"
##               mode                var                 sd
##          "numeric" "4199.57972372372" "64.8041644010917"
##                  n           kurtosis               skew
##             "1000" "50.9615555806469" "5.33522085058122"
##                max                min              range
##              "932"                "1"              "931"
##       quartiles.0%      quartiles.25%      quartiles.50%
##                "1"                "6"               "19"
##      quartiles.75%     quartiles.100%                iqr
##               "52"              "932"               "46"
##
## $UA
##                sum               mean             median
```

```
##            "37289"           "37.289"               "16"
##               mode                var                 sd
##          "numeric" "2925.51299199199" "54.0880115366797"
##                  n           kurtosis               skew
##             "1000" "13.1340821644088" "3.01865027817975"
##                max                min              range
##              "468"                "1"              "467"
##       quartiles.0%      quartiles.25%      quartiles.50%
##                "1"                "5"               "16"
##      quartiles.75%     quartiles.100%                iqr
##               "48"               "468"               "43"
##
## $VX
##                sum               mean             median
##            "28699"           "28.699"               "15"
##               mode                var                 sd
##          "numeric" "1449.68208108108"  "38.074690820558"
##                  n           kurtosis               skew
##             "1000" "13.1655551101411" "3.05585187417878"
##                max                min              range
##              "343"                "1"              "342"
##       quartiles.0%      quartiles.25%      quartiles.50%
##                "1"                "5"               "15"
##      quartiles.75%     quartiles.100%                iqr
##               "38"               "343"               "33"
##
## $WN
##                sum               mean             median
##            "24081"           "24.081"               "12.5"
##               mode                var                 sd
##          "numeric" "1015.11555455455"   "31.86087811964"
##                  n           kurtosis               skew
##             "1000" "12.5829632786527"  "2.9696306827761"
##                max                min              range
##              "307"                "1"              "306"
##       quartiles.0%      quartiles.25%      quartiles.50%
##                "1"                "5"              "12.5"
##      quartiles.75%     quartiles.100%                iqr
##               "29"               "307"               "24"
```

```
##                    vars    n    mean     sd median trimmed    mad min  max
## DepDelay              1 7061   36.23  52.15     19   25.89  22.24   1 1200
## DayOfWeek             2 7061    3.77   2.03      4    3.71   2.97   1    7
## DepTime               3 7061 1426.45 505.88   1408 1434.76 462.57   1 2400
## Cancelled             4 7061    0.00   0.03      0    0.00   0.00   0    1
## CarrierDelay          5 3609   18.76  44.93      3    9.18   4.45   0 1193
## WeatherDelay          6 3609    0.61   7.46      0    0.00   0.00   0  252
## NASDelay              7 3609    9.17  21.09      0    4.04   0.00   0  232
## SecurityDelay         8 3609    0.02   0.78      0    0.00   0.00   0   37
## LateAircraftDelay     9 3609   33.04  47.40     19   23.87  28.17   0  683
##                    range  skew kurtosis   se
## DepDelay            1199  4.96    55.90 0.62
## DayOfWeek              6  0.14    -1.23 0.02
## DepTime             2399 -0.24     0.09 6.02
```

```
## Cancelled             1 34.25  1171.50 0.00
## CarrierDelay       1193  8.79   157.84 0.75
## WeatherDelay        252 20.59   528.71 0.12
## NASDelay            232  4.26    24.60 0.35
## SecurityDelay        37 38.40  1616.72 0.01
## LateAircraftDelay   683  3.66    26.20 0.79
```

```
## Number of carrier delays:  1980
```

```
## Number of weather delays:  71
```

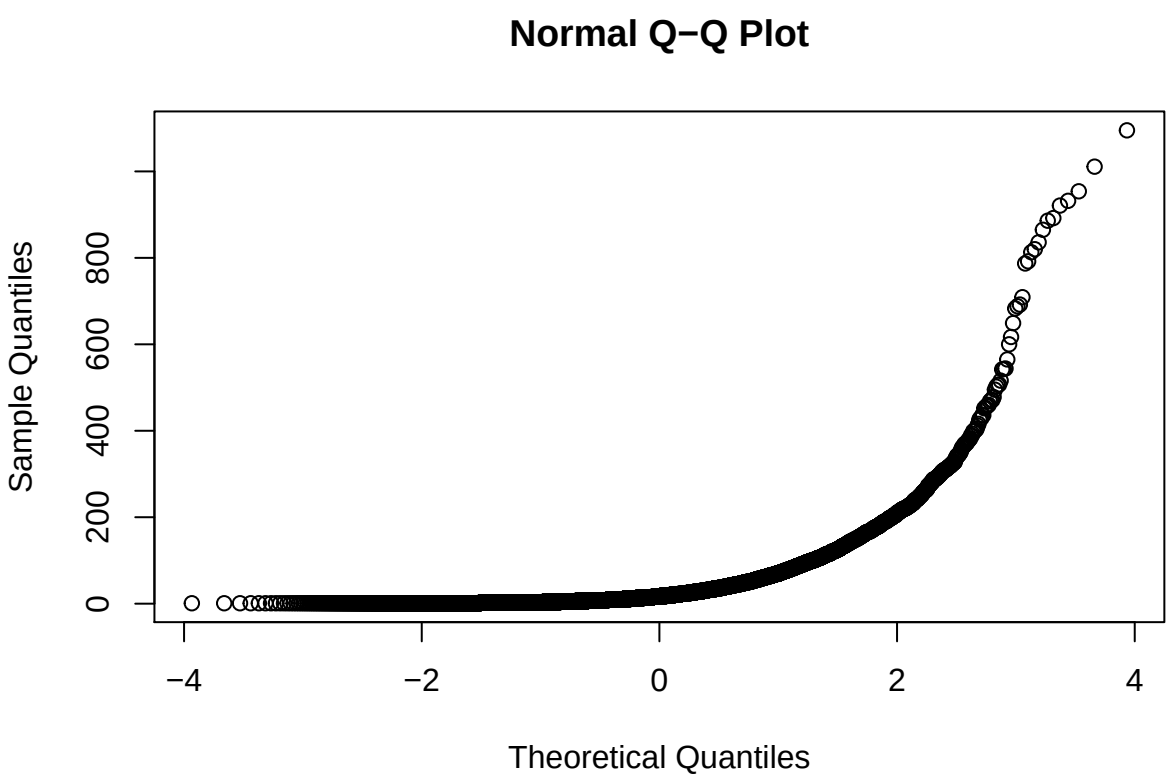
```
## Number of NAS delays:  1480
```

```
## Number of security delays:  4
```

```
## Number of aircraft delays:  2387
```

ANOVA

Below are the results of the ANOVA showing the difference between airline carriers with regards to the average negative delay time of their flights out of SFO. This analysis is testing to see if there is a statistically signifcant difference between the mean delay time of each carrier. If there is a statistically significant difference between the mean delay time in minutes for each carrier we would expect to see a P value $< .05$. As we can see below the P value is much less than .05 so we can reject the null hypothesis that there is not a significant difference between the average delay times. There is also a plot showing how the data is distributed, because the data does not look like a normal bell curve we can assume that the sample of 1000 observations for each airline is not normally distributed.




```
##                Df   Sum Sq Mean Sq F value Pr(>F)
## group          11  1954797  177709   45.48 <2e-16 ***
## Residuals   11988 46846507    3908
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene's Test

Below is the Levene's test used to determine the homogeneity of variance between the airline carriers. My alpha level is .05 to determine the level of significance. If P is less than .05 I can conclude that I am 95% confident that the variance between the carriers is significantly different. The p value is far less than .05 so I can indeed conclude that there is a statistically significant difference between the variance of the carriers.

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group    11  37.803 < 2.2e-16 ***
##       11988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

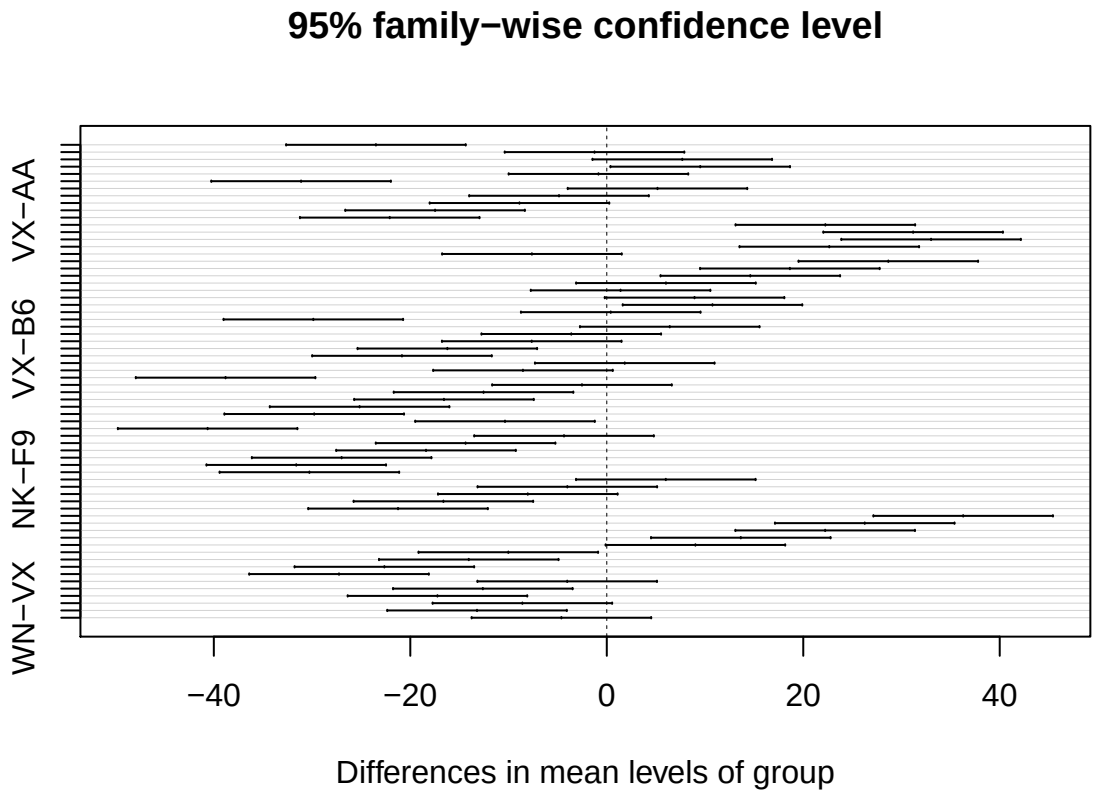
Tukey-Kramer Post Hoc Test

This test is to determine which means have a difference between each other. The left hand column shows the two carriers that are being compared and the right hand shows the p value. the lower the P value the more

significant the difference is between the average delay time of the carriers. My level of significance is .05. This suggests that the best performing airline in terms of average delay time is HA and the worst is EV.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = fit)
##
## $group
##          diff         lwr         upr     p adj
## AS-AA -23.499 -32.6369634 -14.3610366 0.0000000
## B6-AA  -1.248 -10.3859634   7.8899634 0.9999993
## DL-AA   7.685  -1.4529634  16.8229634 0.2025930
## EV-AA   9.513   0.3750366  18.6509634 0.0326643
## F9-AA  -0.847  -9.9849634   8.2909634 1.0000000
## HA-AA -31.119 -40.2569634 -21.9810366 0.0000000
## NK-AA   5.161  -3.9769634  14.2989634 0.7922810
## OO-AA  -4.861 -13.9989634   4.2769634 0.8502858
## UA-AA  -8.888 -18.0259634   0.2499634 0.0654739
## VX-AA -17.478 -26.6159634  -8.3400366 0.0000000
## WN-AA -22.096 -31.2339634 -12.9580366 0.0000000
## B6-AS  22.251  13.1130366  31.3889634 0.0000000
## DL-AS  31.184  22.0460366  40.3219634 0.0000000
## EV-AS  33.012  23.8740366  42.1499634 0.0000000
## F9-AS  22.652  13.5140366  31.7899634 0.0000000
## HA-AS  -7.620 -16.7579634   1.5179634 0.2135552
## NK-AS  28.660  19.5220366  37.7979634 0.0000000
## OO-AS  18.638   9.5000366  27.7759634 0.0000000
## UA-AS  14.611   5.4730366  23.7489634 0.0000114
## VX-AS   6.021  -3.1169634  15.1589634 0.5836353
## WN-AS   1.403  -7.7349634  10.5409634 0.9999975
## DL-B6   8.933  -0.2049634  18.0709634 0.0624247
## EV-B6  10.761   1.6230366  19.8989634 0.0066642
## F9-B6   0.401  -8.7369634   9.5389634 1.0000000
## HA-B6 -29.871 -39.0089634 -20.7330366 0.0000000
## NK-B6   6.409  -2.7289634  15.5469634 0.4819513
## OO-B6  -3.613 -12.7509634   5.5249634 0.9802376
## UA-B6  -7.640 -16.7779634   1.4979634 0.2101403
## VX-B6 -16.230 -25.3679634  -7.0920366 0.0000004
## WN-B6 -20.848 -29.9859634 -11.7100366 0.0000000
## EV-DL   1.828  -7.3099634  10.9659634 0.9999619
## F9-DL  -8.532 -17.6699634   0.6059634 0.0942079
## HA-DL -38.804 -47.9419634 -29.6660366 0.0000000
## NK-DL  -2.524 -11.6619634   6.6139634 0.9991113
## OO-DL -12.546 -21.6839634  -3.4080366 0.0004501
## UA-DL -16.573 -25.7109634  -7.4350366 0.0000002
## VX-DL -25.163 -34.3009634 -16.0250366 0.0000000
## WN-DL -29.781 -38.9189634 -20.6430366 0.0000000
## F9-EV -10.360 -19.4979634  -1.2220366 0.0114207
## HA-EV -40.632 -49.7699634 -31.4940366 0.0000000
## NK-EV  -4.352 -13.4899634   4.7859634 0.9242802
## OO-EV -14.374 -23.5119634  -5.2360366 0.0000179
## UA-EV -18.401 -27.5389634  -9.2630366 0.0000000
## VX-EV -26.991 -36.1289634 -17.8530366 0.0000000
```

```
## WN-EV -31.609 -40.7469634 -22.4710366 0.0000000
## HA-F9 -30.272 -39.4099634 -21.1340366 0.0000000
## NK-F9   6.008  -3.1299634  15.1459634 0.5870457
## OO-F9  -4.014 -13.1519634   5.1239634 0.9565638
## UA-F9  -8.041 -17.1789634   1.0969634 0.1494576
## VX-F9 -16.631 -25.7689634  -7.4930366 0.0000002
## WN-F9 -21.249 -30.3869634 -12.1110366 0.0000000
## NK-HA  36.280  27.1420366  45.4179634 0.0000000
## OO-HA  26.258  17.1200366  35.3959634 0.0000000
## UA-HA  22.231  13.0930366  31.3689634 0.0000000
## VX-HA  13.641   4.5030366  22.7789634 0.0000688
## WN-HA   9.023  -0.1149634  18.1609634 0.0566820
## OO-NK -10.022 -19.1599634  -0.8840366 0.0176234
## UA-NK -14.049 -23.1869634  -4.9110366 0.0000328
## VX-NK -22.639 -31.7769634 -13.5010366 0.0000000
## WN-NK -27.257 -36.3949634 -18.1190366 0.0000000
## UA-OO  -4.027 -13.1649634   5.1109634 0.9555493
## VX-OO -12.617 -21.7549634  -3.4790366 0.0004005
## WN-OO -17.235 -26.3729634  -8.0970366 0.0000000
## VX-UA  -8.590 -17.7279634   0.5479634 0.0889318
## WN-UA -13.208 -22.3459634  -4.0700366 0.0001475
## WN-VX  -4.618 -13.7559634   4.5199634 0.8895458
```



##Regression Analysis Below are the steps taken to develop a model to predict the arrival delay of a flight leaving SFO including flights that left both earlier and later than expected. The first step was to select fields that could possibly predict the arrival delay of a flight. I wanted to get as close to a 95% confidence level as possible without making the model too flexible by adding too many fields. Below is an exhaustive search of

the possible model for predicting the arrival delay. The higher the R-sq (adj) value the better the model will be at predicting arrival times. As we can see below the 10th iteration of the model produces the highest R-sq (adj) value and leaves us with greater than a 95% confidence level. That model includes the carrier, departure delay, taxi out time, distance traveled, air time, carrier delay, weather delay, NAS delay, security delay and late aircraft delay.

```
##    (Intercept) Carrier DepDelay TaxiOut Distance AirTime CarrierDelay
## 1            1       0        1       0        0       0            0
## 2            1       0        1       0        0       0            0
## 3            1       0        0       0        0       0            1
## 4            1       0        1       0        0       0            1
## 5            1       0        1       1        0       0            1
## 6            1       0        1       1        0       0            1
## 7            1       0        1       1        1       0            1
## 8            1       1        1       1        1       0            1
## 9            1       1        1       1        1       1            1
## 10           1       1        1       1        1       1            1
##    WeatherDelay NASDelay SecurityDelay LateAircraftDelay      R-Sq
## 1             0        0             0                 0 0.8993179
## 2             0        1             0                 0 0.9279504
## 3             0        1             0                 1 0.9412610
## 4             0        1             0                 1 0.9514045
## 5             0        1             0                 1 0.9577018
## 6             1        1             0                 1 0.9611432
## 7             1        1             0                 1 0.9617787
## 8             1        1             0                 1 0.9619832
## 9             1        1             0                 1 0.9620710
## 10            1        1             1                 1 0.9621319
##    R-Sq (adj)          Cp
## 1   0.8993116 26451.85436
## 2   0.9279413 14393.12692
## 3   0.9412500  8788.34519
## 4   0.9513923  4517.65351
## 5   0.9576886  1867.06601
## 6   0.9611286   419.45866
## 7   0.9617619   153.77439
## 8   0.9619641    69.63832
## 9   0.9620496    34.66376
## 10  0.9621082    11.00000
```

Linear Regression

I have chosen to do a linear regression because the data is not categorical and is predicting a numerical value in terms of minutes. I am testing to see that at least one of the variables chosen contributes significantly to the model. If the P value is less than .05 then the variable is a significant contributor to the model. As shown below all of the variables contribute significantly and the overall P value of the model is 2.2e-16. The Adjusted R-squared is 0.9621 suggesting that with this model we can accurately predict arrival delay of flights leaving SFO with 96.21% confidence.

```
##
## Call:
## lm(formula = ArrDelay ~ ., data = model.dat)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.423  -5.022   0.212   5.134  40.409
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.467e+01  2.919e-01 -50.240  < 2e-16 ***
## Carrier            3.154e-01  3.469e-02   9.091  < 2e-16 ***
## DepDelay           3.880e-01  6.521e-03  59.504  < 2e-16 ***
## TaxiOut            4.247e-01  8.763e-03  48.470  < 2e-16 ***
## Distance          -2.973e-03  3.327e-04  -8.934  < 2e-16 ***
## AirTime            1.810e-02  2.988e-03   6.058 1.41e-09 ***
## CarrierDelay       6.331e-01  7.336e-03  86.296  < 2e-16 ***
## WeatherDelay       6.991e-01  1.875e-02  37.280  < 2e-16 ***
## NASDelay           8.218e-01  6.991e-03 117.555  < 2e-16 ***
## SecurityDelay      8.807e-01  1.738e-01   5.066 4.11e-07 ***
## LateAircraftDelay  6.374e-01  7.334e-03  86.903  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.145 on 15951 degrees of freedom
## Multiple R-squared:  0.9621, Adjusted R-squared:  0.9621
## F-statistic: 4.053e+04 on 10 and 15951 DF,  p-value: < 2.2e-16
```