Flights Analyzed

Anh Tran

10/15/2021

Introduction

Given a csv file contains information on all commercial flights departing the Washington, DC area and arriving New York during January 2004.

The given data set have the following variables:

- CRS_DEP_TIME – scheduled departure time
- CARRIER
- DEP_TIME – actual departure time
- DEST – destination airport
- DISTANCE of the flight
- FL_DATE – flight date
- ORIGIN – origin airport
- Weather – (0 = normal conditions, 1 = rain/snow)
- DAY_WEEK – (1 = Sunday, 2 = Monday, . . . , 7 = Saturday)
- DAY_OF_MONTH – (1 = January 1st, . . . 31 = January 31st)
- Flight.Status (on-time or delayed)

We are trying to predict whether or not the flight will be delayed based on the sample data set in 2004.

A flight is consider a delay flight is the one arrive at least 15 minutes after its scheduled.

Load library

```
## Loading required package: carData
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: lattice
```

Dimension of the data set

x
2201
12

Structure of the data set

```
## 'data.frame':    2201 obs. of  12 variables:
##  $ CRS_DEP_TIME : int  1455 1640 1245 1715 1039 840 1240 1645 1715 2120 ...
##  $ CARRIER      : chr  "OH" "DH" "DH" "DH" ...
##  $ DEP_TIME     : int  1455 1640 1245 1709 1035 839 1243 1644 1710 2129 ...
##  $ DEST         : chr  "JFK" "JFK" "LGA" "LGA" ...
##  $ DISTANCE     : int  184 213 229 229 229 228 228 228 228 228 ...
##  $ FL_DATE      : chr  "1/1/2004" "1/1/2004" "1/1/2004" "1/1/2004" ...
##  $ FL_NUM       : int  5935 6155 7208 7215 7792 7800 7806 7810 7812 7814 ...
##  $ ORIGIN       : chr  "BWI" "DCA" "IAD" "IAD" ...
##  $ Weather      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ DAY_WEEK     : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ DAY_OF_MONTH : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Flight.Status: chr  "ontime" "ontime" "ontime" "ontime" ...
```

|| || || ||

Example of the first 5 lines from given data set

CRS_DEP_TIME	CARRIER	DEP_TIME	DEST	DISTANCE	FL_DATE	ORIGIN	Weather	DAY_WEEK	DAY_OF_MONTH	Flight.Status
1455	OH	1455	JFK	184	1/1/2004	BWI	0	4	1	ontime
1640	DH	1640	JFK	213	1/1/2004	DCA	0	4	1	ontime
1245	DH	1245	LGA	229	1/1/2004	IAD	0	4	1	ontime
1715	DH	1709	LGA	229	1/1/2004	IAD	0	4	1	ontime
1039	DH	1035	LGA	229	1/1/2004	IAD	0	4	1	ontime

Summary of our given data from the csv file which included all lengths, types, and statistical such as mean, median, and mode of each particular variable

CRS_DEP_TIME	CARRIER	DEP_TIME	DEST	DISTANCE	FL_DATE	FL_NUM	ORIGIN	Weather	DAY_WEEK	DAY_OF_MONTH	Flight.Status
Min. : 600	Length:2201	Min. : 10	Length:2201	Min. :169.0	Length:2201	Min. : 746	Length:2201	Min. :0.00000	Min. :1.000	Min. : 1.00	Length:2201
1st Qu.:1000	Class :character	1st Qu.:1004	Class :character	1st Qu.:213.0	Class :character	1st Qu.:2156	Class :character	1st Qu.:0.00000	1st Qu.:2.000	1st Qu.: 8.00	Class :character
Median :1455	Mode :character	Median :1450	Mode :character	Median :214.0	Mode :character	Median :2385	Mode :character	Median :0.00000	Median :4.000	Median :16.00	Mode :character
Mean :1372	NA	Mean :1369	NA	Mean :211.9	NA	Mean :3815	NA	Mean :0.01454	Mean :3.905	Mean :16.02	NA
3rd Qu.:1710	NA	3rd Qu.:1709	NA	3rd Qu.:214.0	NA	3rd Qu.:6155	NA	3rd Qu.:0.00000	3rd Qu.:5.000	3rd Qu.:23.00	NA

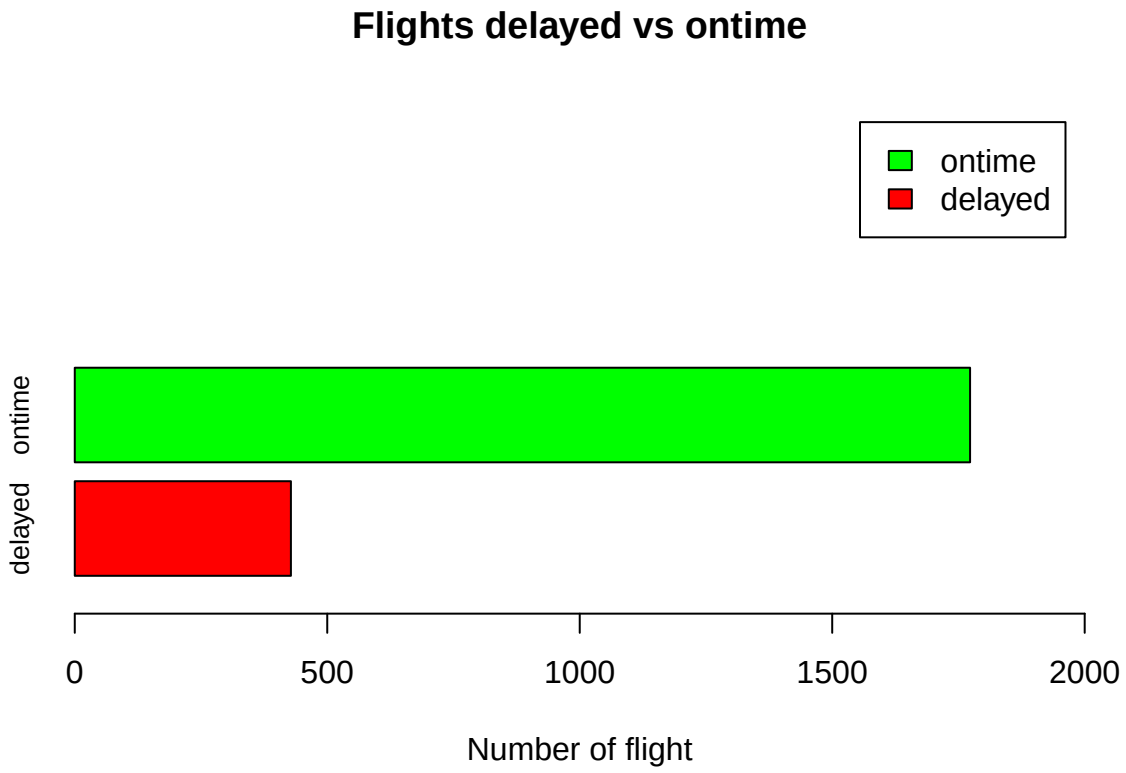
	CRS_DEP_TIME	CARRIER	DEP_TIME	DEST	DISTANCE	FL_DATE	FL_NUM	ORIGIN	Weather	DAY_WEEK	DAY_OF_MONTH	Flight.Status
	Max. :2130	NA	Max. :2330	NA	Max. :229.0	NA	Max. :7924	NA	Max. :1.00000	Max. :7.000	Max. :31.00	NA

Our data set showed that there are 2201 lines of information in the csv file. For those variable using numeric to determine their meaning such as CRS_DEP_TIMEFL_DATE (scheduled depart time), DEP_TIME (actual depart time), DISTANCE (distance of each flight), Weather (weather at that time), DAY_WEEK (the day flight took place), DAY_OF_MONTH (the particular day of the month when the flight took place) , and Flight.Status (whether or not any particular flight delayed) are also provided min, max, mode, median, mean, the first and third Quartiles of the values of each column.

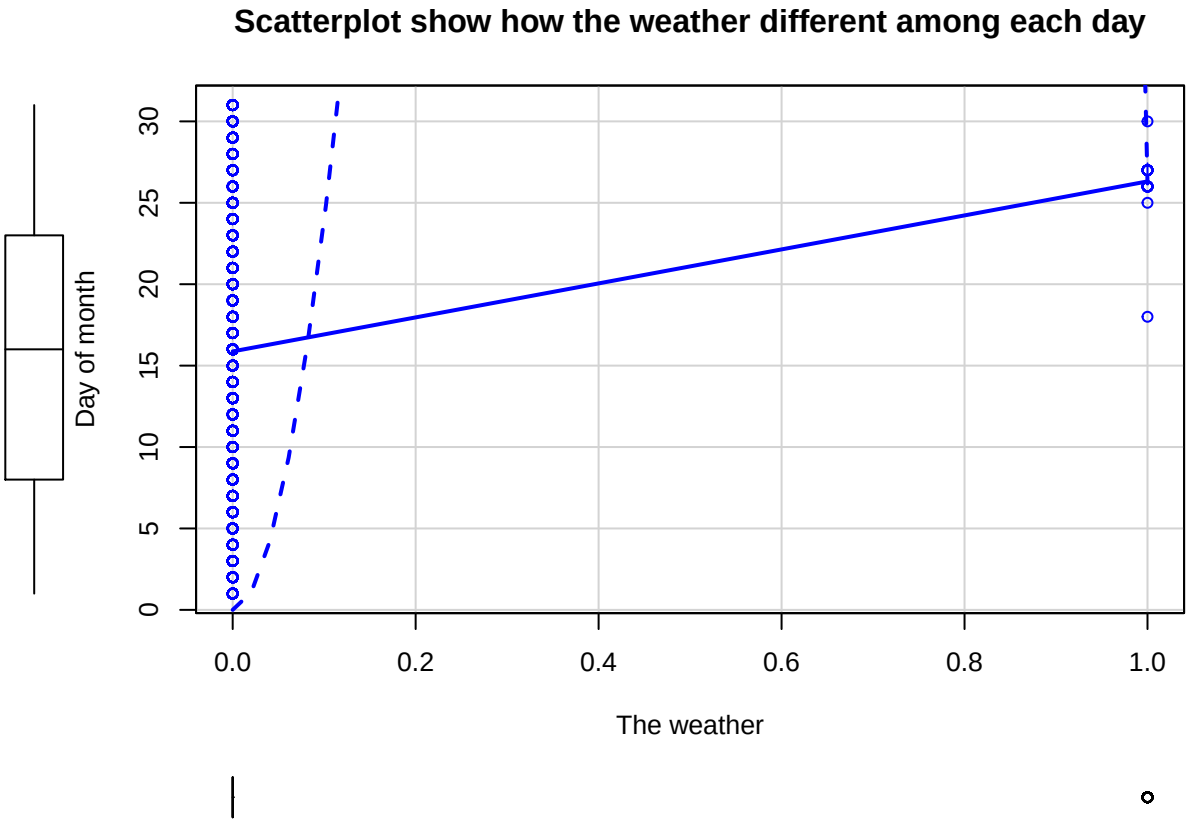
New data set after adding one column converted flight status into numeric "1" and "0" instead of "delayed" or "on-time"

CRS_DEP_TIME	CARRIER	DEP_TIME	DEST	DISTANCE	FL_DATE	FL_NUM	ORIGIN	Weather	DAY_WEEK	DAY_OF_MONTH	Flight.Status	flight.status
1455	OH	1455	JFK	184	1/1/2004	5935	BWI	0	4	1	ontime	0
1640	DH	1640	JFK	213	1/1/2004	6155	DCA	0	4	1	ontime	0
1245	DH	1245	LGA	229	1/1/2004	7208	IAD	0	4	1	ontime	0
1715	DH	1709	LGA	229	1/1/2004	7215	IAD	0	4	1	ontime	0
1039	DH	1035	LGA	229	1/1/2004	7792	IAD	0	4	1	ontime	0

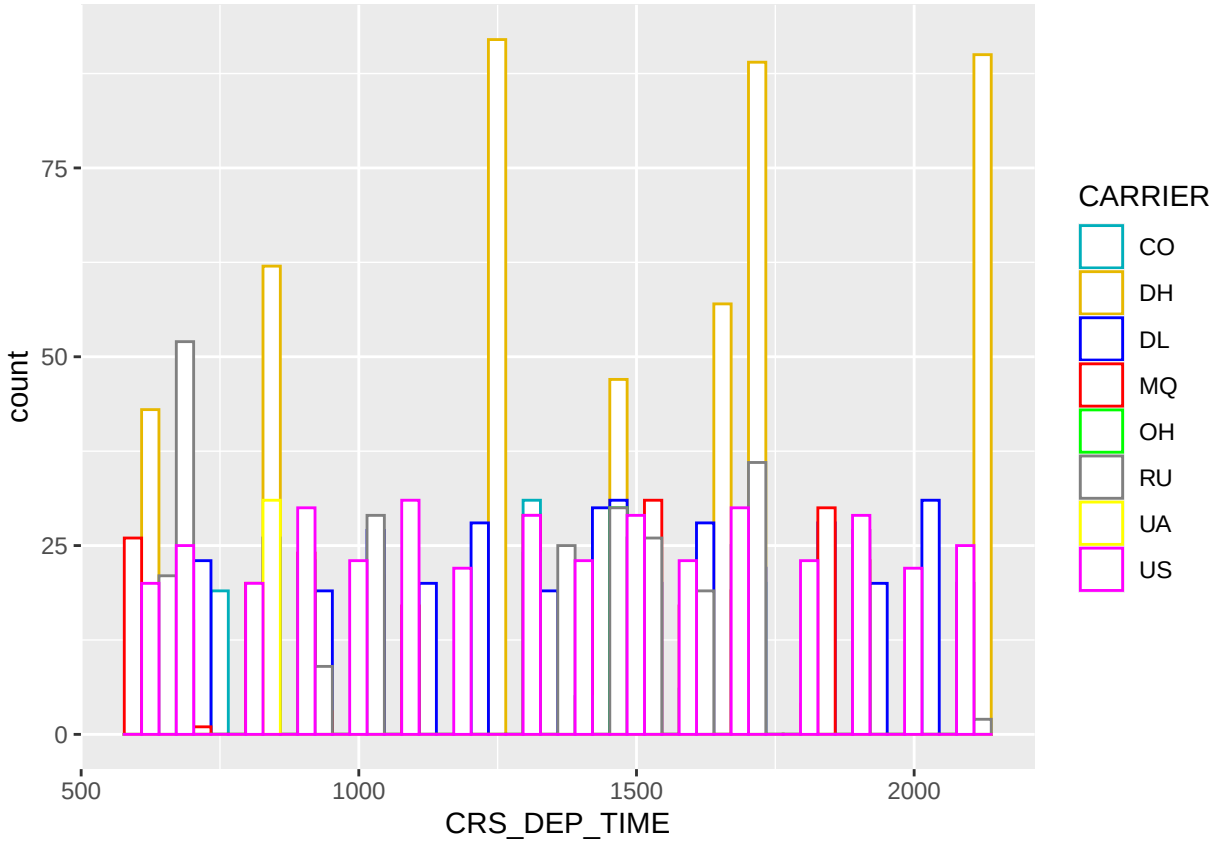
Bar plot showing number of flight delayed vs on time



Scatter plot shows the weather of each day of month



Histogram shows scheduled depart time group by Carriers



Calculate percentage of flights delay vs on time

Var1	Freq
delayed	0.1944571
ontime	0.8055429

Flight delayed = 19.445%

Flight on time = 80.554%

To determine whether or not any variable in the given data set affect flights arrive on time or delay, we will create cross table between the following four variable: CRS_DEP_TIME, DEST, ORIGIN, Weather and Flight.Status

Cross table between ORIGIN (where flight departed) and Flight.Status (delayed or not)

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
```

```
## |-------------------------|
##
##
## Total Observations in Table:  2201
##
##
##                      | new_df$ORIGIN
## new_df$Flight.Status |       BWI |       DCA |       IAD | Row Total |
## ---------------------|-----------|-----------|-----------|-----------|
##              delayed |        37 |       221 |       170 |       428 |
##                      |     2.749 |     7.739 |    10.043 |           |
##                      |     0.086 |     0.516 |     0.397 |     0.194 |
##                      |     0.255 |     0.161 |     0.248 |           |
##                      |     0.017 |     0.100 |     0.077 |           |
## ---------------------|-----------|-----------|-----------|-----------|
##               ontime |       108 |      1149 |       516 |      1773 |
##                      |     0.664 |     1.868 |     2.424 |           |
##                      |     0.061 |     0.648 |     0.291 |     0.806 |
##                      |     0.745 |     0.839 |     0.752 |           |
##                      |     0.049 |     0.522 |     0.234 |           |
## ---------------------|-----------|-----------|-----------|-----------|
##         Column Total |       145 |      1370 |       686 |      2201 |
##                      |     0.066 |     0.622 |     0.312 |           |
## ---------------------|-----------|-----------|-----------|-----------|
##
##
```

Based on the results of the column total from the cross table above, we can clearly see that all the flights from given data set which were departed from:

- Flights from "BWI" delayed by 25.5% and 74.55% were on time.
- Flights from "DCA" delayed by 16.1% and 83.9% were on time.
- Flights from "IAD" delayed by 24.8% and 75.2% were on time.

The percentage of delayed flights among three different origin, "BWI", "DCA", and "IAD", demonstrate that the origin of each flight does affect whether or not any particular flight will arrive on time. There were only 16.15% of flights from given data set departed from "DCA" delayed while the other two origin, "BWI" and "IAD", were much higher 25.5% and 24.8%.

Cross table between DAY_WEEK (flight on a particular day of the week) and Flight.Status (delayed or not)

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  2201
```

```
##
##
##                  | new_df$Flight.Status
## new_df$DAY_WEEK  |   delayed |    ontime | Row Total |
## -----------------|-----------|-----------|-----------|
##                1 |        84 |       224 |       308 |
##                  |     9.703 |     2.342 |           |
##                  |     0.273 |     0.727 |     0.140 |
##                  |     0.196 |     0.126 |           |
##                  |     0.038 |     0.102 |           |
## -----------------|-----------|-----------|-----------|
##                2 |        63 |       244 |       307 |
##                  |     0.183 |     0.044 |           |
##                  |     0.205 |     0.795 |     0.139 |
##                  |     0.147 |     0.138 |           |
##                  |     0.029 |     0.111 |           |
## -----------------|-----------|-----------|-----------|
##                3 |        57 |       263 |       320 |
##                  |     0.439 |     0.106 |           |
##                  |     0.178 |     0.822 |     0.145 |
##                  |     0.133 |     0.148 |           |
##                  |     0.026 |     0.119 |           |
## -----------------|-----------|-----------|-----------|
##                4 |        57 |       315 |       372 |
##                  |     3.252 |     0.785 |           |
##                  |     0.153 |     0.847 |     0.169 |
##                  |     0.133 |     0.178 |           |
##                  |     0.026 |     0.143 |           |
## -----------------|-----------|-----------|-----------|
##                5 |        75 |       316 |       391 |
##                  |     0.014 |     0.003 |           |
##                  |     0.192 |     0.808 |     0.178 |
##                  |     0.175 |     0.178 |           |
##                  |     0.034 |     0.144 |           |
## -----------------|-----------|-----------|-----------|
##                6 |        24 |       226 |       250 |
##                  |    12.463 |     3.008 |           |
##                  |     0.096 |     0.904 |     0.114 |
##                  |     0.056 |     0.127 |           |
##                  |     0.011 |     0.103 |           |
## -----------------|-----------|-----------|-----------|
##                7 |        68 |       185 |       253 |
##                  |     7.186 |     1.735 |           |
##                  |     0.269 |     0.731 |     0.115 |
##                  |     0.159 |     0.104 |           |
##                  |     0.031 |     0.084 |           |
## -----------------|-----------|-----------|-----------|
##    Column Total  |       428 |      1773 |      2201 |
##                  |     0.194 |     0.806 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

There were three days that have the most significant flights delayed showed in the cross table above which

were Monday, Thursday, and Saturday. The percentages of flights delayed departed on these days are need to be consider as one of the secondary caused flights delayed.

Cross table between DEST (where flight arrived) and Flight.Status (delayed or not)

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  2201
##
##
##                      | new_df$DEST
## new_df$Flight.Status |       EWR |       JFK |       LGA | Row Total |
## ---------------------|-----------|-----------|-----------|-----------|
##              delayed |       161 |        84 |       183 |       428 |
##                      |     7.764 |     1.065 |     7.380 |           |
##                      |     0.376 |     0.196 |     0.428 |     0.194 |
##                      |     0.242 |     0.218 |     0.159 |           |
##                      |     0.073 |     0.038 |     0.083 |           |
## ---------------------|-----------|-----------|-----------|-----------|
##               ontime |       504 |       302 |       967 |      1773 |
##                      |     1.874 |     0.257 |     1.782 |           |
##                      |     0.284 |     0.170 |     0.545 |     0.806 |
##                      |     0.758 |     0.782 |     0.841 |           |
##                      |     0.229 |     0.137 |     0.439 |           |
## ---------------------|-----------|-----------|-----------|-----------|
##         Column Total |       665 |       386 |      1150 |      2201 |
##                      |     0.302 |     0.175 |     0.522 |           |
## ---------------------|-----------|-----------|-----------|-----------|
##
##
```

The results from the column total of the cross table above showed that destination of flights can affect the ability of being on time.

- Table showed that there were 24.2% of flights arrived at "EWR" delayed and 75.8% were on time
- Table showed that there were 21.8% of flights arrived at "JFK" delayed and 78.2% were on time
- Table showed that there were 15.9% of flights arrived at "EWR" delayed and 84.1% were on time

In conclusion, we can say that flights destination were affect the percentage of flight delayed. There were only 15.9% of flight arrived at "EWR" destination delayed compared to 24.2% and 21.8% delayed at "EWR" and "JFK" destinations.

Cross table between Weather (weather condition at that time) and Flight.Status (delayed or not)

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  2201
##
##
##                      | new_df$Weather
## new_df$Flight.Status |         0 |         1 | Row Total |
## ---------------------|-----------|-----------|-----------|
##              delayed |       396 |        32 |       428 |
##                      |     1.575 |   106.783 |           |
##                      |     0.925 |     0.075 |     0.194 |
##                      |     0.183 |     1.000 |           |
##                      |     0.180 |     0.015 |           |
## ---------------------|-----------|-----------|-----------|
##               ontime |      1773 |         0 |      1773 |
##                      |     0.380 |    25.777 |           |
##                      |     1.000 |     0.000 |     0.806 |
##                      |     0.817 |     0.000 |           |
##                      |     0.806 |     0.000 |           |
## ---------------------|-----------|-----------|-----------|
##         Column Total |      2169 |        32 |      2201 |
##                      |     0.985 |     0.015 |           |
## ---------------------|-----------|-----------|-----------|
##
##
```

The results from columns total of the table above showed the possibility of weather can affect flights arrive on time or delay.

- Table showed that 18.3% flights delayed and 81.7% on time if the weather is under normal condition
- Table showed that 100% flights delayed and 0% on time if the weather is identified as snow or rain

We can clearly see that the weather affected how flights arrived. Based on the result we have from the cross table above, there were 100% flights delayed due to the bad weather condition.

Cross table between CRS_DEP_TIME (time flight scheduled to depart) and Flight.Status (delayed or not)

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
```

```
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  2201
##
##
##                     | new_df$Flight.Status
## new_df$CRS_DEP_TIME |   delayed |    ontime | Row Total |
## --------------------|-----------|-----------|-----------|
##                 600 |         2 |        24 |        26 |
##                     |     1.847 |     0.446 |           |
##                     |     0.077 |     0.923 |     0.012 |
##                     |     0.005 |     0.014 |           |
##                     |     0.001 |     0.011 |           |
## --------------------|-----------|-----------|-----------|
##                 630 |         4 |        53 |        57 |
##                     |     4.528 |     1.093 |           |
##                     |     0.070 |     0.930 |     0.026 |
##                     |     0.009 |     0.030 |           |
##                     |     0.002 |     0.024 |           |
## --------------------|-----------|-----------|-----------|
##                 640 |         9 |        13 |        22 |
##                     |     5.212 |     1.258 |           |
##                     |     0.409 |     0.591 |     0.010 |
##                     |     0.021 |     0.007 |           |
##                     |     0.004 |     0.006 |           |
## --------------------|-----------|-----------|-----------|
##                 645 |         1 |        20 |        21 |
##                     |     2.328 |     0.562 |           |
##                     |     0.048 |     0.952 |     0.010 |
##                     |     0.002 |     0.011 |           |
##                     |     0.000 |     0.009 |           |
## --------------------|-----------|-----------|-----------|
##                 700 |        18 |        74 |        92 |
##                     |     0.001 |     0.000 |           |
##                     |     0.196 |     0.804 |     0.042 |
##                     |     0.042 |     0.042 |           |
##                     |     0.008 |     0.034 |           |
## --------------------|-----------|-----------|-----------|
##                 730 |         3 |        21 |        24 |
##                     |     0.595 |     0.144 |           |
##                     |     0.125 |     0.875 |     0.011 |
##                     |     0.007 |     0.012 |           |
##                     |     0.001 |     0.010 |           |
## --------------------|-----------|-----------|-----------|
##                 735 |         2 |        15 |        17 |
##                     |     0.516 |     0.125 |           |
##                     |     0.118 |     0.882 |     0.008 |
##                     |     0.005 |     0.008 |           |
##                     |     0.001 |     0.007 |           |
```

```
## --------------------|-----------|-----------|-----------|
##                 759 |         0 |         2 |         2 |
##                     |     0.389 |     0.094 |           |
##                     |     0.000 |     1.000 |     0.001 |
##                     |     0.000 |     0.001 |           |
##                     |     0.000 |     0.001 |           |
## --------------------|-----------|-----------|-----------|
##                 800 |         8 |        32 |        40 |
##                     |     0.006 |     0.002 |           |
##                     |     0.200 |     0.800 |     0.018 |
##                     |     0.019 |     0.018 |           |
##                     |     0.004 |     0.015 |           |
## --------------------|-----------|-----------|-----------|
##                 830 |         4 |        22 |        26 |
##                     |     0.221 |     0.053 |           |
##                     |     0.154 |     0.846 |     0.012 |
##                     |     0.009 |     0.012 |           |
##                     |     0.002 |     0.010 |           |
## --------------------|-----------|-----------|-----------|
##                 840 |         9 |        53 |        62 |
##                     |     0.775 |     0.187 |           |
##                     |     0.145 |     0.855 |     0.028 |
##                     |     0.021 |     0.030 |           |
##                     |     0.004 |     0.024 |           |
## --------------------|-----------|-----------|-----------|
##                 845 |         0 |         3 |         3 |
##                     |     0.583 |     0.141 |           |
##                     |     0.000 |     1.000 |     0.001 |
##                     |     0.000 |     0.002 |           |
##                     |     0.000 |     0.001 |           |
## --------------------|-----------|-----------|-----------|
##                 850 |         5 |        26 |        31 |
##                     |     0.175 |     0.042 |           |
##                     |     0.161 |     0.839 |     0.014 |
##                     |     0.012 |     0.015 |           |
##                     |     0.002 |     0.012 |           |
## --------------------|-----------|-----------|-----------|
##                 900 |        10 |        67 |        77 |
##                     |     1.652 |     0.399 |           |
##                     |     0.130 |     0.870 |     0.035 |
##                     |     0.023 |     0.038 |           |
##                     |     0.005 |     0.030 |           |
## --------------------|-----------|-----------|-----------|
##                 925 |         0 |         3 |         3 |
##                     |     0.583 |     0.141 |           |
##                     |     0.000 |     1.000 |     0.001 |
##                     |     0.000 |     0.002 |           |
##                     |     0.000 |     0.001 |           |
## --------------------|-----------|-----------|-----------|
##                 930 |         1 |        27 |        28 |
##                     |     3.628 |     0.876 |           |
##                     |     0.036 |     0.964 |     0.013 |
##                     |     0.002 |     0.015 |           |
##                     |     0.000 |     0.012 |           |
```

```
## --------------------|-----------|-----------|-----------|
##                1000 |         0 |        23 |        23 |
##                     |     4.473 |     1.080 |           |
##                     |     0.000 |     1.000 |     0.010 |
##                     |     0.000 |     0.013 |           |
##                     |     0.000 |     0.010 |           |
## --------------------|-----------|-----------|-----------|
##                1030 |         9 |        47 |        56 |
##                     |     0.328 |     0.079 |           |
##                     |     0.161 |     0.839 |     0.025 |
##                     |     0.021 |     0.027 |           |
##                     |     0.004 |     0.021 |           |
## --------------------|-----------|-----------|-----------|
##                1039 |         1 |         5 |         6 |
##                     |     0.024 |     0.006 |           |
##                     |     0.167 |     0.833 |     0.003 |
##                     |     0.002 |     0.003 |           |
##                     |     0.000 |     0.002 |           |
## --------------------|-----------|-----------|-----------|
##                1040 |         1 |        14 |        15 |
##                     |     1.260 |     0.304 |           |
##                     |     0.067 |     0.933 |     0.007 |
##                     |     0.002 |     0.008 |           |
##                     |     0.000 |     0.006 |           |
## --------------------|-----------|-----------|-----------|
##                1100 |         5 |        43 |        48 |
##                     |     2.012 |     0.486 |           |
##                     |     0.104 |     0.896 |     0.022 |
##                     |     0.012 |     0.024 |           |
##                     |     0.002 |     0.020 |           |
## --------------------|-----------|-----------|-----------|
##                1130 |         1 |        19 |        20 |
##                     |     2.146 |     0.518 |           |
##                     |     0.050 |     0.950 |     0.009 |
##                     |     0.002 |     0.011 |           |
##                     |     0.000 |     0.009 |           |
## --------------------|-----------|-----------|-----------|
##                1200 |         0 |        22 |        22 |
##                     |     4.278 |     1.033 |           |
##                     |     0.000 |     1.000 |     0.010 |
##                     |     0.000 |     0.012 |           |
##                     |     0.000 |     0.010 |           |
## --------------------|-----------|-----------|-----------|
##                1230 |         1 |        27 |        28 |
##                     |     3.628 |     0.876 |           |
##                     |     0.036 |     0.964 |     0.013 |
##                     |     0.002 |     0.015 |           |
##                     |     0.000 |     0.012 |           |
## --------------------|-----------|-----------|-----------|
##                1240 |         6 |        25 |        31 |
##                     |     0.000 |     0.000 |           |
##                     |     0.194 |     0.806 |     0.014 |
##                     |     0.014 |     0.014 |           |
##                     |     0.003 |     0.011 |           |
```

```
## --------------------|-----------|-----------|-----------|
##                1245 |        16 |        45 |        61 |
##                     |     1.444 |     0.348 |           |
##                     |     0.262 |     0.738 |     0.028 |
##                     |     0.037 |     0.025 |           |
##                     |     0.007 |     0.020 |           |
## --------------------|-----------|-----------|-----------|
##                1300 |        14 |        95 |       109 |
##                     |     2.443 |     0.590 |           |
##                     |     0.128 |     0.872 |     0.050 |
##                     |     0.033 |     0.054 |           |
##                     |     0.006 |     0.043 |           |
## --------------------|-----------|-----------|-----------|
##                1315 |         2 |         2 |         4 |
##                     |     1.920 |     0.464 |           |
##                     |     0.500 |     0.500 |     0.002 |
##                     |     0.005 |     0.001 |           |
##                     |     0.001 |     0.001 |           |
## --------------------|-----------|-----------|-----------|
##                1330 |         0 |        19 |        19 |
##                     |     3.695 |     0.892 |           |
##                     |     0.000 |     1.000 |     0.009 |
##                     |     0.000 |     0.011 |           |
##                     |     0.000 |     0.009 |           |
## --------------------|-----------|-----------|-----------|
##                1359 |         4 |        21 |        25 |
##                     |     0.153 |     0.037 |           |
##                     |     0.160 |     0.840 |     0.011 |
##                     |     0.009 |     0.012 |           |
##                     |     0.002 |     0.010 |           |
## --------------------|-----------|-----------|-----------|
##                1400 |         6 |        40 |        46 |
##                     |     0.970 |     0.234 |           |
##                     |     0.130 |     0.870 |     0.021 |
##                     |     0.014 |     0.023 |           |
##                     |     0.003 |     0.018 |           |
## --------------------|-----------|-----------|-----------|
##                1430 |        11 |        41 |        52 |
##                     |     0.078 |     0.019 |           |
##                     |     0.212 |     0.788 |     0.024 |
##                     |     0.026 |     0.023 |           |
##                     |     0.005 |     0.019 |           |
## --------------------|-----------|-----------|-----------|
##                1455 |        46 |        92 |       138 |
##                     |    13.687 |     3.304 |           |
##                     |     0.333 |     0.667 |     0.063 |
##                     |     0.107 |     0.052 |           |
##                     |     0.021 |     0.042 |           |
## --------------------|-----------|-----------|-----------|
##                1500 |        16 |        61 |        77 |
##                     |     0.070 |     0.017 |           |
##                     |     0.208 |     0.792 |     0.035 |
##                     |     0.037 |     0.034 |           |
##                     |     0.007 |     0.028 |           |
```

```
## --------------------|-----------|-----------|-----------|
##                1515 |         3 |         2 |         5 |
##                     |     4.229 |     1.021 |           |
##                     |     0.600 |     0.400 |     0.002 |
##                     |     0.007 |     0.001 |           |
##                     |     0.001 |     0.001 |           |
## --------------------|-----------|-----------|-----------|
##                1520 |         0 |         1 |         1 |
##                     |     0.194 |     0.047 |           |
##                     |     0.000 |     1.000 |     0.000 |
##                     |     0.000 |     0.001 |           |
##                     |     0.000 |     0.000 |           |
## --------------------|-----------|-----------|-----------|
##                1525 |         9 |        12 |        21 |
##                     |     5.919 |     1.429 |           |
##                     |     0.429 |     0.571 |     0.010 |
##                     |     0.021 |     0.007 |           |
##                     |     0.004 |     0.005 |           |
## --------------------|-----------|-----------|-----------|
##                1530 |        10 |        40 |        50 |
##                     |     0.008 |     0.002 |           |
##                     |     0.200 |     0.800 |     0.023 |
##                     |     0.023 |     0.023 |           |
##                     |     0.005 |     0.018 |           |
## --------------------|-----------|-----------|-----------|
##                1600 |        12 |        33 |        45 |
##                     |     1.207 |     0.291 |           |
##                     |     0.267 |     0.733 |     0.020 |
##                     |     0.028 |     0.019 |           |
##                     |     0.005 |     0.015 |           |
## --------------------|-----------|-----------|-----------|
##                1605 |         1 |         0 |         1 |
##                     |     3.337 |     0.806 |           |
##                     |     1.000 |     0.000 |     0.000 |
##                     |     0.002 |     0.000 |           |
##                     |     0.000 |     0.000 |           |
## --------------------|-----------|-----------|-----------|
##                1610 |         3 |        21 |        24 |
##                     |     0.595 |     0.144 |           |
##                     |     0.125 |     0.875 |     0.011 |
##                     |     0.007 |     0.012 |           |
##                     |     0.001 |     0.010 |           |
## --------------------|-----------|-----------|-----------|
##                1630 |        10 |        41 |        51 |
##                     |     0.001 |     0.000 |           |
##                     |     0.196 |     0.804 |     0.023 |
##                     |     0.023 |     0.023 |           |
##                     |     0.005 |     0.019 |           |
## --------------------|-----------|-----------|-----------|
##                1640 |         5 |        22 |        27 |
##                     |     0.012 |     0.003 |           |
##                     |     0.185 |     0.815 |     0.012 |
##                     |     0.012 |     0.012 |           |
##                     |     0.002 |     0.010 |           |
```

```
## --------------------|-----------|-----------|-----------|
##                1645 |         1 |        29 |        30 |
##                     |     4.005 |     0.967 |           |
##                     |     0.033 |     0.967 |     0.014 |
##                     |     0.002 |     0.016 |           |
##                     |     0.000 |     0.013 |           |
## --------------------|-----------|-----------|-----------|
##                1700 |        11 |        63 |        74 |
##                     |     0.799 |     0.193 |           |
##                     |     0.149 |     0.851 |     0.034 |
##                     |     0.026 |     0.036 |           |
##                     |     0.005 |     0.029 |           |
## --------------------|-----------|-----------|-----------|
##                1710 |         7 |        21 |        28 |
##                     |     0.444 |     0.107 |           |
##                     |     0.250 |     0.750 |     0.013 |
##                     |     0.016 |     0.012 |           |
##                     |     0.003 |     0.010 |           |
## --------------------|-----------|-----------|-----------|
##                1715 |        21 |        40 |        61 |
##                     |     7.040 |     1.699 |           |
##                     |     0.344 |     0.656 |     0.028 |
##                     |     0.049 |     0.023 |           |
##                     |     0.010 |     0.018 |           |
## --------------------|-----------|-----------|-----------|
##                1720 |        11 |        16 |        27 |
##                     |     6.296 |     1.520 |           |
##                     |     0.407 |     0.593 |     0.012 |
##                     |     0.026 |     0.009 |           |
##                     |     0.005 |     0.007 |           |
## --------------------|-----------|-----------|-----------|
##                1725 |         0 |         1 |         1 |
##                     |     0.194 |     0.047 |           |
##                     |     0.000 |     1.000 |     0.000 |
##                     |     0.000 |     0.001 |           |
##                     |     0.000 |     0.000 |           |
## --------------------|-----------|-----------|-----------|
##                1730 |        13 |        37 |        50 |
##                     |     1.105 |     0.267 |           |
##                     |     0.260 |     0.740 |     0.023 |
##                     |     0.030 |     0.021 |           |
##                     |     0.006 |     0.017 |           |
## --------------------|-----------|-----------|-----------|
##                1800 |         1 |        26 |        27 |
##                     |     3.441 |     0.831 |           |
##                     |     0.037 |     0.963 |     0.012 |
##                     |     0.002 |     0.015 |           |
##                     |     0.000 |     0.012 |           |
## --------------------|-----------|-----------|-----------|
##                1830 |        12 |        46 |        58 |
##                     |     0.046 |     0.011 |           |
##                     |     0.207 |     0.793 |     0.026 |
##                     |     0.028 |     0.026 |           |
##                     |     0.005 |     0.021 |           |
```

```
## --------------------|-----------|-----------|-----------|
##                1900 |        35 |        64 |        99 |
##                     |    12.883 |     3.110 |           |
##                     |     0.354 |     0.646 |     0.045 |
##                     |     0.082 |     0.036 |           |
##                     |     0.016 |     0.029 |           |
## --------------------|-----------|-----------|-----------|
##                1930 |         3 |        17 |        20 |
##                     |     0.203 |     0.049 |           |
##                     |     0.150 |     0.850 |     0.009 |
##                     |     0.007 |     0.010 |           |
##                     |     0.001 |     0.008 |           |
## --------------------|-----------|-----------|-----------|
##                2000 |         4 |        18 |        22 |
##                     |     0.018 |     0.004 |           |
##                     |     0.182 |     0.818 |     0.010 |
##                     |     0.009 |     0.010 |           |
##                     |     0.002 |     0.008 |           |
## --------------------|-----------|-----------|-----------|
##                2030 |         5 |        26 |        31 |
##                     |     0.175 |     0.042 |           |
##                     |     0.161 |     0.839 |     0.014 |
##                     |     0.012 |     0.015 |           |
##                     |     0.002 |     0.012 |           |
## --------------------|-----------|-----------|-----------|
##                2100 |         7 |        38 |        45 |
##                     |     0.350 |     0.085 |           |
##                     |     0.156 |     0.844 |     0.020 |
##                     |     0.016 |     0.021 |           |
##                     |     0.003 |     0.017 |           |
## --------------------|-----------|-----------|-----------|
##                2120 |        28 |        62 |        90 |
##                     |     6.298 |     1.520 |           |
##                     |     0.311 |     0.689 |     0.041 |
##                     |     0.065 |     0.035 |           |
##                     |     0.013 |     0.028 |           |
## --------------------|-----------|-----------|-----------|
##                2130 |         1 |         1 |         2 |
##                     |     0.960 |     0.232 |           |
##                     |     0.500 |     0.500 |     0.001 |
##                     |     0.002 |     0.001 |           |
##                     |     0.000 |     0.000 |           |
## --------------------|-----------|-----------|-----------|
##        Column Total |       428 |      1773 |      2201 |
##                     |     0.194 |     0.806 |           |
## --------------------|-----------|-----------|-----------|
##
##
```

Based on the results showed from the cross table between two variables, CRS_DEP_TIME and Flight.Status from the given data set, there was no strong evidence support that scheduled departure time may affect the possibility of flights delay or on-time. There were some significant amount of flights delayed appeared during certain time such as 46 flights delayed at 14:55, 35 flights delayed at 19:00, and 21 flights delayed at 17:21. The highest percentage of flights delayed was approximately 10.6% which not really high compared to other

variables from the data set. I most likely, scheduled departure time does not affect flights arrival time at all.

By the time we perform prediction, there is no variable DEP_TIME so we need to create new data frame without DEP_TIME variable.

Generate a random seed to randomly split a data set into training and validation set

The total observation in table is 2201 = 100% of data

70% of data = 2201 * 0.7 = 1540.7 rows

Generate random sample using 70% of data

Extract training set using random_sample

Create validation data set using the remaining data of 2201 - 1540.7 = 660.3 rows = 30% of data

Create classification tree model using training data set to predict whether or not any given flight on-time or delayed

Generate prediction for data training set

Create confusion matrix using predicted values and flight_delayed_training_set$Flight.Status

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1540
##
##
##                                      | flight_delayed_training_set$Flight.Status
## flight_delayed_training_set_prediction |   delayed |    ontime | Row Total |
## ---------------------------------------|-----------|-----------|-----------|
##                               delayed |       104 |        28 |       132 |
##                                       |   220.783 |    56.546 |           |
##                                       |     0.788 |     0.212 |     0.086 |
##                                       |     0.331 |     0.023 |           |
##                                       |     0.068 |     0.018 |           |
## ---------------------------------------|-----------|-----------|-----------|
##                                ontime |       210 |      1198 |      1408 |
##                                       |    20.698 |     5.301 |           |
##                                       |     0.149 |     0.851 |     0.914 |
##                                       |     0.669 |     0.977 |           |
##                                       |     0.136 |     0.778 |           |
## ---------------------------------------|-----------|-----------|-----------|
##                          Column Total |       314 |      1226 |      1540 |
##                                       |     0.204 |     0.796 |           |
## ---------------------------------------|-----------|-----------|-----------|
##
##
```

The accuracy of the tree model is percentage of all cases classified correctly

Accuracy = (true negative + true positive) / N = (94 + 1208) / 1540 = 0.061 + 0.784 = 0.845

It indicates that the models classified correctly 84.5% of the cases.

Misclassification rate = (false negative + false positive) / N = (22 + 216) / 1540 = 0.014 + 1.14 = 1.154. The model incorrectly classified about 11.5% of all training cases. This is a poor model performance

Classification tree tend to show higher accuracy on training data set used to develop tree model, we need to evaluate the model accuracy and misclassification rate on the unused validation set

Generate predictions for validation set using flight_delayed_models

Create confusion matrix for validation data set

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  661
##
##
##                                          | flight_delayed_validation_set$Flight.Status
## flight_delayed_validation_set_prediction |   delayed |    ontime | Row Total |
## -----------------------------------------|-----------|-----------|-----------|
##                                  delayed |        34 |        16 |        50 |
##                                          |    74.679 |    15.564 |           |
##                                          |     0.680 |     0.320 |     0.076 |
##                                          |     0.298 |     0.029 |           |
##                                          |     0.051 |     0.024 |           |
## -----------------------------------------|-----------|-----------|-----------|
##                                   ontime |        80 |       531 |       611 |
##                                          |     6.111 |     1.274 |           |
##                                          |     0.131 |     0.869 |     0.924 |
##                                          |     0.702 |     0.971 |           |
##                                          |     0.121 |     0.803 |           |
## -----------------------------------------|-----------|-----------|-----------|
##                             Column Total |       114 |       547 |       661 |
##                                          |     0.172 |     0.828 |           |
## -----------------------------------------|-----------|-----------|-----------|
##
##
```

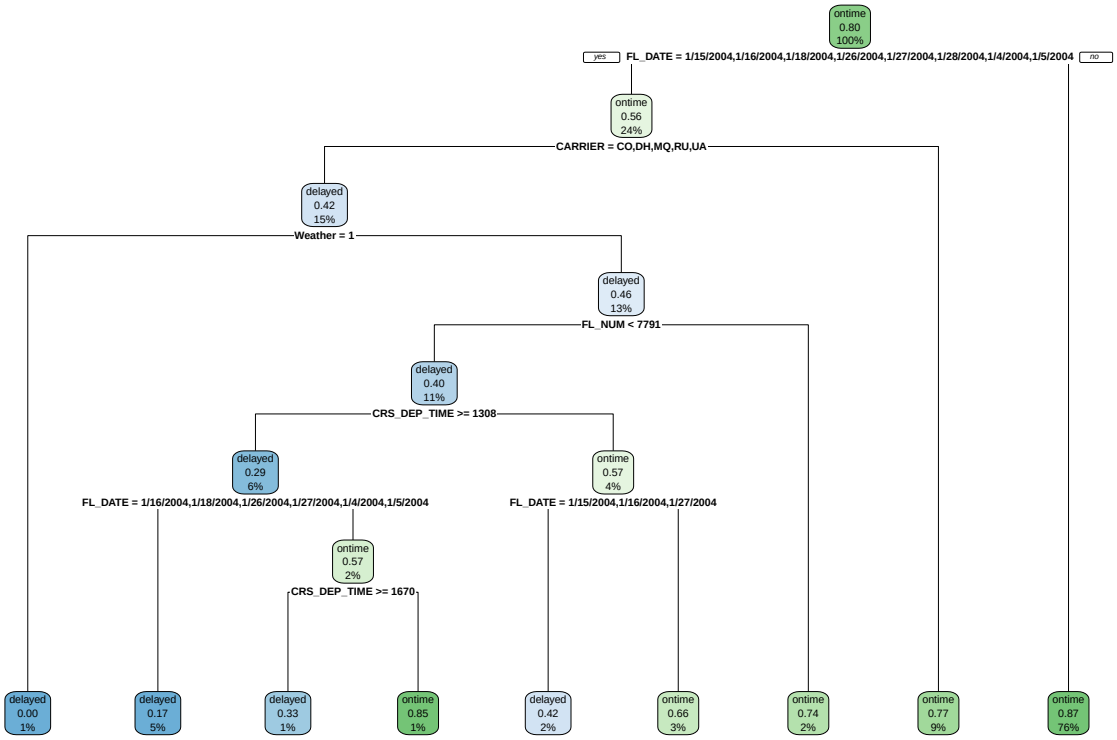
The accuracy of the tree model is percentage of all cases classified correctly

Accuracy = (true negative + true positive) / N = (33 + 525) / 661 = 0.05 + 0.794 = 0.844

It indicates that the models classified correctly 84.4% of the cases and it very close to the accuracy on the training set

Misclassification rate = (false negative + false positive) / N = (18 + 85) / 661 = 0.027 + 0.129 = 0.156. The model incorrectly classified about 15.6% of all training cases. This is a poor model performance

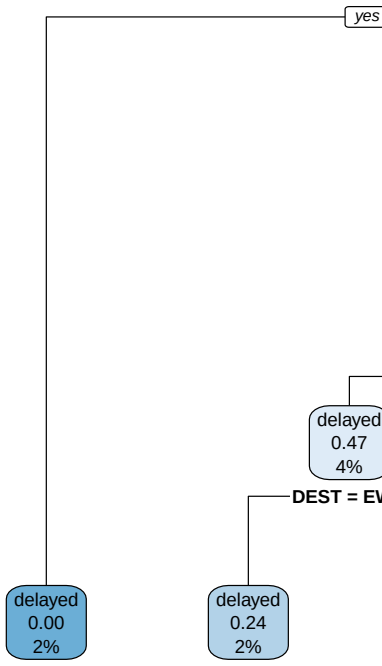
To understand the rules which lead to flight delayed vs on-time, we need to draw a decision tree model using the flight_delayed_model



Five of the terminal nodes correspond to cases with P yes > 0.5.Observation of these nodes are classified as condition which responded to flights arrive "on-time". These nodes followed the rules:

- Flights on-time = FL_DATE(flights departed on 4,5,15,16,18,26,27,28 of January, 2004), not in list of carrier (CO,DH,MQ,RU,UA)
- Flights on-time = FL_DATE(flights departed on 4,5,15,16,18,26,27,28 of January, 2004), in list of carrier (CO,DH,MQ,RU,UA), not departed on 26 and 27 of January, 2004, have scheduled departure time later than 5:13pm, and have flight number < 7303
- Flights on-time = FL_DATE(flights departed on 4,5,15,16,18,26,27,28 of January, 2004), in list of carrier (CO,DH,MQ,RU,UA), not departed on 26 and 27 of January, 2004, have scheduled departure time earlier than 5:13pm, departure date on 5,16,18 of January, 2004, and have scheduled departure time later than 1:08pm
- Flights on-time = FL_DATE(flights departed on 4,5,15,16,18,26,27,28 of January, 2004), in list of carrier (CO,DH,MQ,RU,UA), not departed on 26 and 27 of January, 2004, have scheduled departure time earlier than 5:13pm, departure date not on 5,16,18 of January, 2004
- Flights on-time = not departed on on 4,5,15,16,18,26,27,28 of January, 2004

To have better classification tree I would remove the CRS_DEP_TIME, FL_DATE, and FL_NUM variables



Create new decision tree model using the flight_delayed_model_after_dropped

This tree model is much better fit compared to the previous one. For instance, flights arrive on time:

- If flights under good weather condition and does not have carrier such as CO,DH,MQ,RU
- If flights under good weather condition, have carrier such as CO,DH,MQ,RU, flight between the 4th and 6th of January, 2004, and the flight destination is not EWR.
- If flights under good weather condition, have carrier such as CO,DH,MQ,RU, and flight before the 4th of January, 2004
- If flights under good weather condition, have carrier such as CO,DH,MQ,RU, have destination as LGA, and flight before 15th of January, 2004
- If flights under good weather condition, have carrier such as CO,DH,MQ,RU, have destination as LGA, flight after 26th of January, 2004, and flight between Wednesday and Satutday.
- If flights under good weather condition, have carrier such as CO,DH,MQ,RU, have destination as LGA, flight after 17th of January, 2004
- If flights under good weather condition, have carrier such as CO,DH,MQ,RU, have destination as LGA, and flight before 26th of January, 2004

Conclusion

My conclusion based on analyzed data from given data set of information on all commercial flights departing the Washington, DC area and arriving at New York during January 2004 are:

- Scheduled departure time (CRS_DEP_TIME), actual departure time (DEP_TIME), and flight date (FL_DATE) does not affect the possibility of flights delay or not. There were some cases appeared in certain time but have low percentage compared to the other variables. There were not clear connection between these three variables and the rest from the data set caused significant number of flights delayed
- The most significant variable causing flights delayed is the weather (WEATHER). The weather caused 100% of flights delayed despite any other affected of other variables.

- There were very number of flights delayed which departed between the 4th and 6th of January, 2004. This information not strong enough to support that flights delayed caused by date.
- There was approximately 19.6% of flights delayed which departed on Monday, 17.5% on Thursday, and 15.9% on Saturday. These numbers are not significant but still need to consider as a minor issue caused flights delayed.
- The variable DEST (destination airport) and ORIGIN (origin airport) does affect flights delayed may depend on how busy they are by the time flight depart or arrive. The percentages were vary from above 15% to below 26% approximately base on different airports. Time and date may be two more factors caused how busy each airport.
- There were no evidence showed DISTANCE (distance of flights) affect flights delayed.
- Both tree models model above showed that CARRIER variable does affected the possibility of flights delayed. The following carriers were show in the tree model: “CO”,“DH”,“MQ”, and “RU”.