R-lab Final

Nina Bucar

March 16, 2016

Data Analysis For Departure Delays Using nycflights13 Data

1) Weather

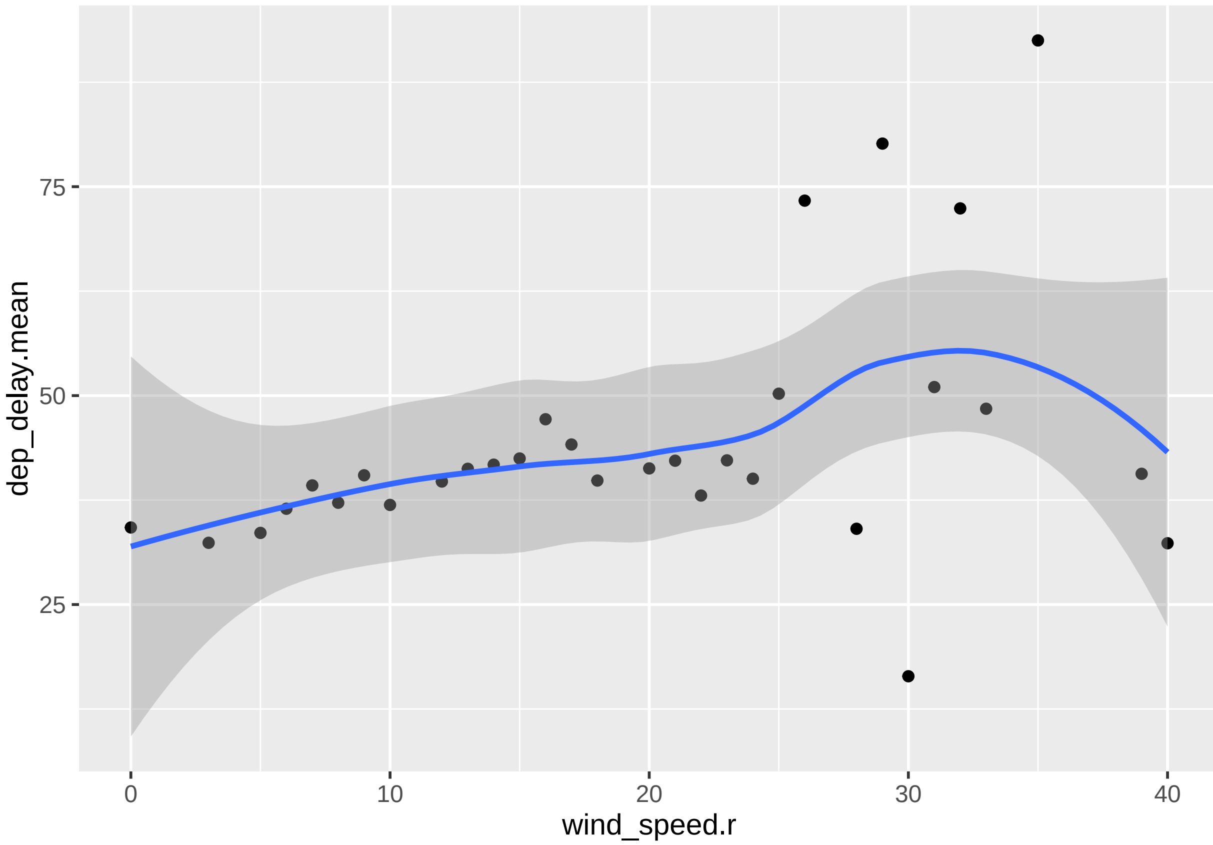
The weather table in the nycflights13 database has several columns of data that seem to have the potential for influencing flight delays. I chose to examine and analyze the wind speed and wind gusts data. The approach I took is to do a left-join of the flights data with the weather data.

I visually inspected the weather data and determined that almost all of the weather data is from the EWR (Newark, New Jersey) airport. Since there is not enough data for the JFK and LGA airports in the weather table to do a reasonable analysis, I used the dplyr and sql functions to filter out everything except the weather data for EWR. I did the same thing when querying the flights table.

I also noticed that the weather table has data only at the level of hours and nothing more granular. So, I created a new variable, "hour", for each flight in the flights table. Then I did the left-join using year, month, day, hour to extend the flight row (observations) with the weather columns.

There are a limited number of wind speeds in the weather table, so I was able to use the group_by and summarize functions to calculate the mean value of departure delays for the various wind speeds in the weather table.

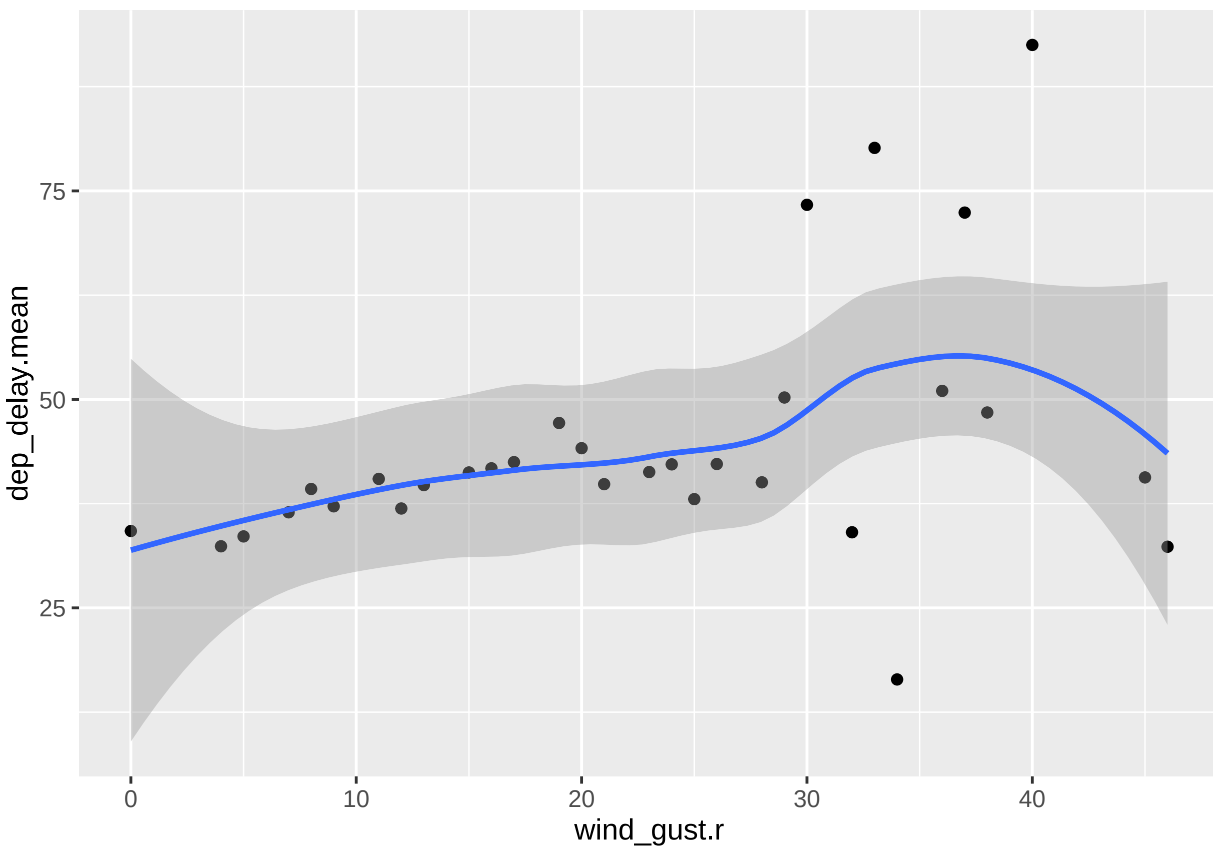
The plot below shows how the mean of the departure delays varies with wind speed. Generally, the mean departure delay for flights increases with increased wind speed for the departure hour for those flights – except for wind speeds above about 35 mph (which had limited data points). As the plot shows, the mean of departure delays increases significantly when the wind speed increases by 10 or 20 mph at departure time.



I also ran a regression of wind_speed on departure delay mean, which resulted in a statistically significant coefficient, as shown in the output, below.

```
##
## Call:
## lm(formula = dep_delay.mean ~ wind_speed.r, data = flights.windspeed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.973  -3.721  -0.516   1.468  39.285
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.4645     5.3563   6.248 8.09e-07 ***
## wind_speed.r   0.5643     0.2384   2.367   0.0248 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.31 on 29 degrees of freedom
## Multiple R-squared:  0.1619, Adjusted R-squared:  0.133
## F-statistic: 5.601 on 1 and 29 DF,  p-value: 0.02484
```

I did a similar analysis for the wind gust data and got similar results: mean departure delays increase with increasing wind gusts.



Regression of mean departure delays on wind gusts showing statistically significant coefficients.

```
##
## Call:
## lm(formula = dep_delay.mean ~ wind_gust.r, data = flights.windgust)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.750  -3.818  -0.393   1.263  39.370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.4031     5.3600   6.232 8.45e-07 ***
## wind_gust.r   0.4932     0.2075   2.377   0.0243 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.3 on 29 degrees of freedom
## Multiple R-squared:  0.1631, Adjusted R-squared:  0.1342
## F-statistic:  5.65 on 1 and 29 DF,  p-value: 0.02427
```

2) Time

To analyze the relationship between departure delays and time, I rounded the departure time to the nearest hour. Then, I used group_by, summarize, filter, arrange, etc. on the data in the flights table to calculate mean delays and cancellation percentages to show how departure delays vary with time.

The output, below, shows which departure hours result in the longest mean departure delays. The largest departure delay means are for the late night and after midnight hours, as would probably be expected.

```
## Source: local data frame [23 x 2]
##
##    dep_time.r dep_delay.mean
##         (dbl)          (dbl)
## 1           3      291.50000
## 2           2      237.65169
## 3           1      192.34615
## 4           0      126.13341
## 5          23      102.95066
## 6          22       63.96912
## 7          24       47.42014
## 8          21       36.57486
## 9          20       27.04789
## 10         19       19.85063
## ..        ...            ...
```

The output, below, shows how the mean departure delay corresponds to the month of departure. The data indicates the the largest depart delay means occur during the beginning-of-summer months and in December, presumably due to travel for the holidays.

```
## Source: local data frame [12 x 2]
##
##    month dep_delay.mean
##    (int)          (dbl)
## 1      7      21.727787
## 2      6      20.846332
## 3     12      16.576688
## 4      4      13.938038
## 5      3      13.227076
## 6      5      12.986859
## 7      8      12.611040
## 8      2      10.816843
## 9      1      10.036665
## 10     9       6.722476
## 11    10       6.243988
## 12    11       5.435362
```

3) Airport Destination

To analyze the relationship between departure delays and airport destinations, I used group_by, summarize, filter, arrange, etc. on the data in the flights table to calculate mean delays and cancellation percentages to show how departure delays vary with flight destinations.

```
## Source: local data frame [102 x 2]
##
##     dest dep_delay.mean
##    (chr)          (dbl)
## 1    CAE       35.57009
## 2    TUL       34.90635
```

```
## 3    OKC       30.56881
## 4    BHM       29.69485
## 5    TYS       28.49396
## 6    JAC       26.54545
## 7    DSM       26.23295
## 8    RIC       23.63985
## 9    ALB       23.62053
## 10   MSN       23.58007
## ..   ...            ...
```

This output shows the destination airports with the worst departure delays.

```
## Source: local data frame [102 x 2]
##
##     dest dep_delay.max
##    (chr)         (dbl)
## 1    HNL          1301
## 2    CMH          1137
## 3    ORD          1126
## 4    SFO          1014
## 5    CVG          1005
## 6    TPA           960
## 7    MSP           911
## 8    PDX           899
## 9    ATL           898
## 10   MIA           896
## ..   ...           ...
```

This output shows the destinations airports with the worst cancellation percentages.

```
## Source: local data frame [102 x 2]
##
##     dest cancelled.pct
##    (chr)         (dbl)
## 1    JAC     12.000000
## 2    CHO     11.538462
## 3    BHM      9.427609
## 4    CAE      8.620690
## 5    TYS      8.399366
## 6    DAY      8.196721
## 7    DSM      7.732865
## 8    MHT      7.631318
## 9    OKC      7.225434
## 10   BDL      6.997743
## ..   ...           ...
```

4) Characteristics of the Plane

To analyze the relationship between departure delays and planes, I used group_by, summarize, filter, arrange, etc. on the data in the flights table to calulate mean delays and cancelation percentages and then did a left_join with the planes table to show how departure delays vary with planes. With further analysis, it would be possible to see how the departure delays relate to specific characteristics of each plane type.

The output, below, shows which planes result in the longest mean departure delays.

```
## Source: local data frame [4,037 x 6]
##
##    dep_delay.mean tailnum      manufacturer       model engines seats
##             (dbl)   (chr)             (chr)       (chr)   (int) (int)
## 1             297  N844MH            BOEING   767-432ER       2   300
## 2             274  N922EV    BOMBARDIER INC CL-600-2B19       2    55
## 3             272  N587NW            BOEING     757-351       2   275
## 4             268  N911DA MCDONNELL DOUGLAS    MD-90-30       2   142
## 5             233  N851NW  AIRBUS INDUSTRIE    A330-223       2   379
## 6             227  N654UA            BOEING     767-322       2   330
## 7             203  N928DN            BOEING    MD-90-30       2   142
## 8             186  N7715E            BOEING     737-7BD       2   149
## 9             177  N665MQ                NA          NA      NA    NA
## 10            165  N136DL            BOEING     767-332       2   330
## ..            ...     ...               ...         ...     ...   ...
```

The output, below, shows the significant mean departure delays for planes manufactured by BOEING.

```
## Source: local data frame [1,627 x 6]
##
##    dep_delay.mean tailnum manufacturer     model engines seats
##             (dbl)   (chr)        (chr)     (chr)   (int) (int)
## 1           297.0  N844MH       BOEING 767-432ER       2   300
## 2           272.0  N587NW       BOEING   757-351       2   275
## 3           227.0  N654UA       BOEING   767-322       2   330
## 4           203.0  N928DN       BOEING  MD-90-30       2   142
## 5           186.0  N7715E       BOEING   737-7BD       2   149
## 6           165.0  N136DL       BOEING   767-332       2   330
## 7           132.0  N670US       BOEING   747-451       4   450
## 8           112.5  N305AS       BOEING   737-990       2   149
## 9           111.0  N78003       BOEING   777-224       2   400
## 10           91.0  N657UA       BOEING   767-322       2   330
## ..            ...     ...          ...       ...     ...   ...
```

My analysis of this data shows that you can see significant delays from the NYC airports in 2013 due to weather conditions such as wind speeds and wind gusts, departure times, with late night hours being particularly bad, departure times of year, with summer months and December being particularly bad, destination airports, such as CAE, TUL, OKC and others listed in my output, and particular planes.

The data in nycflights13 is interesting and allows for interesting analysis using RSQLite, dplyr, ggplot and other R language libraries and functions.