Analysis of 2015 Airline Data

I. Introduction

For many, including myself, airline travel is a necessary part of life. The United States Department of Transportation (DOT) collects data on passenger airlines in the US and has made this data publicly available. The data dates back to 1987 and can hopefully be used to help better understand the airline industry and help consumers and airlines make smarter decisions. All code for this analysis can be found on my GitHub. https://github.com/rtcastellano/AirlineAnalysis

Initial questions

One of the aspects of the data that most interested me was the on-time information, which has been recorded for over a decade. My initial research tasks were the following:

1) Describe the overall state of the airline industry. What are the characteristics of the major airlines? What are they good at? What can they improve? Give the basic information that someone studying the airline industry should know.
2) Everyone who flies needs to deal with delays and cancellations. I was particularly interested in studying delays. What can I learn about delays and their causes?
3) I wanted to build a model to predict whether or not a flight is delayed. What factors correlate with flights being delayed? Such a tool would be useful to both consumers, in order to better plan their travel, and for airlines, in order to identify room for improvement and increase customer satisfaction.

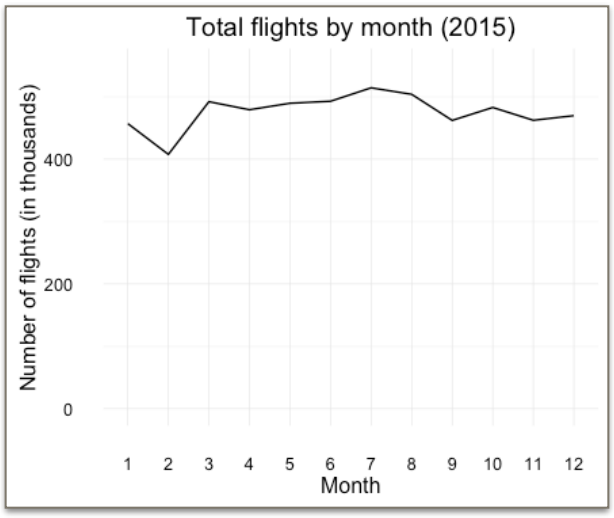
I decided to focus my research on the last year of full data, 2015. Thus, **all data considered here is from the year 2015**. This helped in reducing the amount of data so as to make my analysis more focused. Additionally, while I did not analyze other years, 2015 would be my starting point for analyzing data in 2016 and the future. Sections II and III, which address goals (1) and (2), are nontechnical, while section IV is technical.

II. State of the airline industry

Overall number of flights

The DOT data consists of scheduled flights by U.S. carriers that account for at least 1% of domestic passenger revenue. For these carriers, there were a total of **over 5.7 million** flights in the year 2015. From the data, I removed flights without arrival delay information. This included all flights that were cancelled; the amount of flights that were not cancelled, but still had missing arrival delay data was less than 1% of the data, so I believed I could remove them without biasing my data. I also joined this with the airline data provided to me in order to interpret my findings.

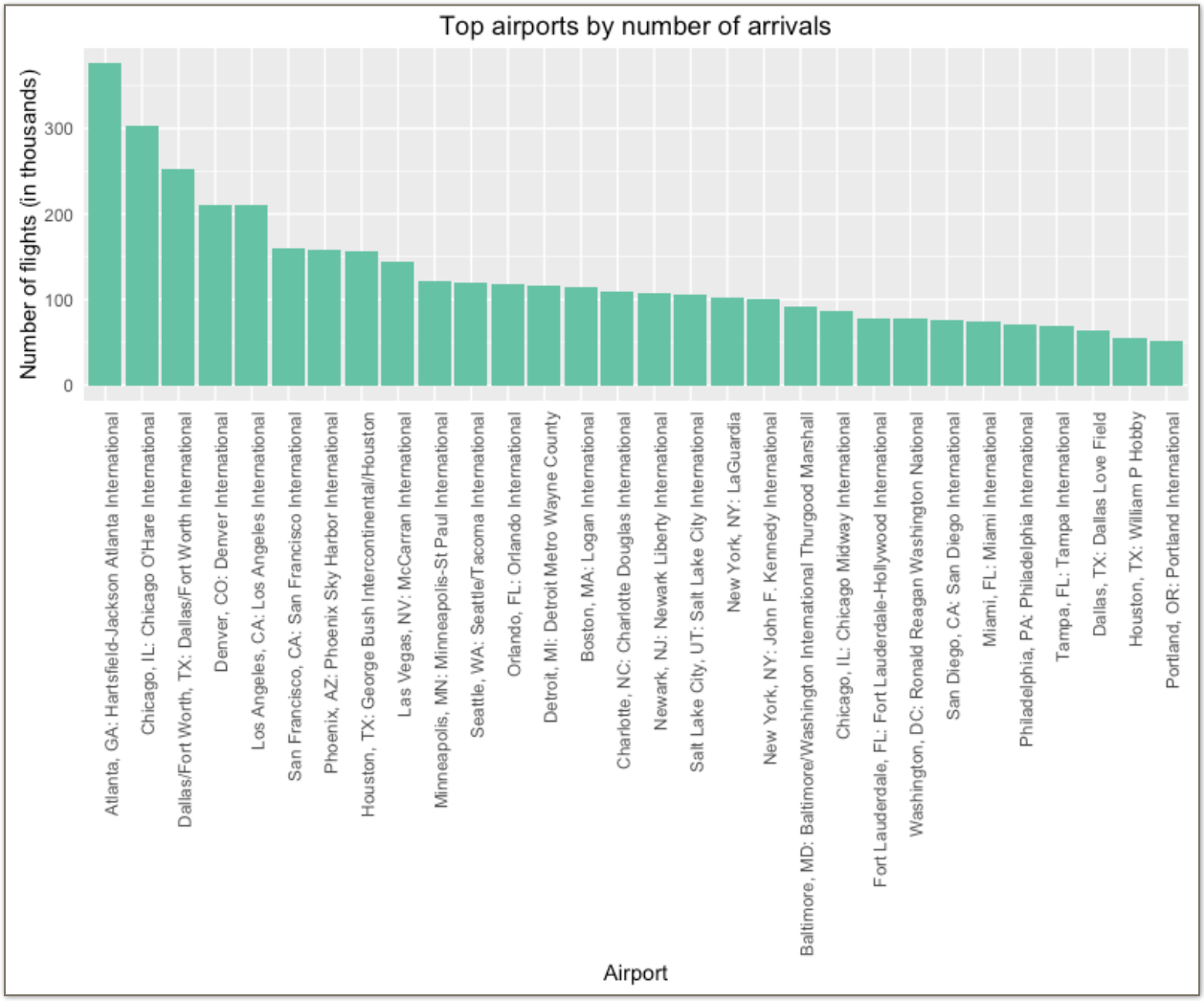
Each month of 2015 had between 400,000 and 520,000 flights with traffic at its lowest in February with just over 400,000 flights. (February also had the lowest flights

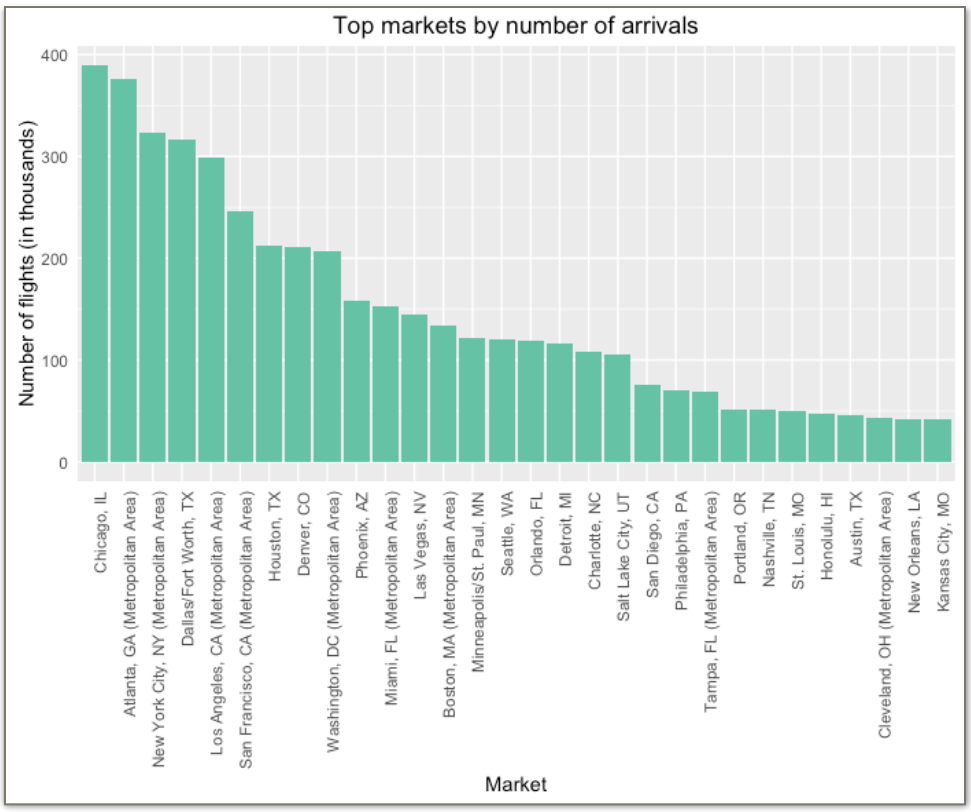


per day.) Travel was highest in the summer months (June, July, and August), reaching a maximum in July with over 510,000 flights.

Where are people traveling?

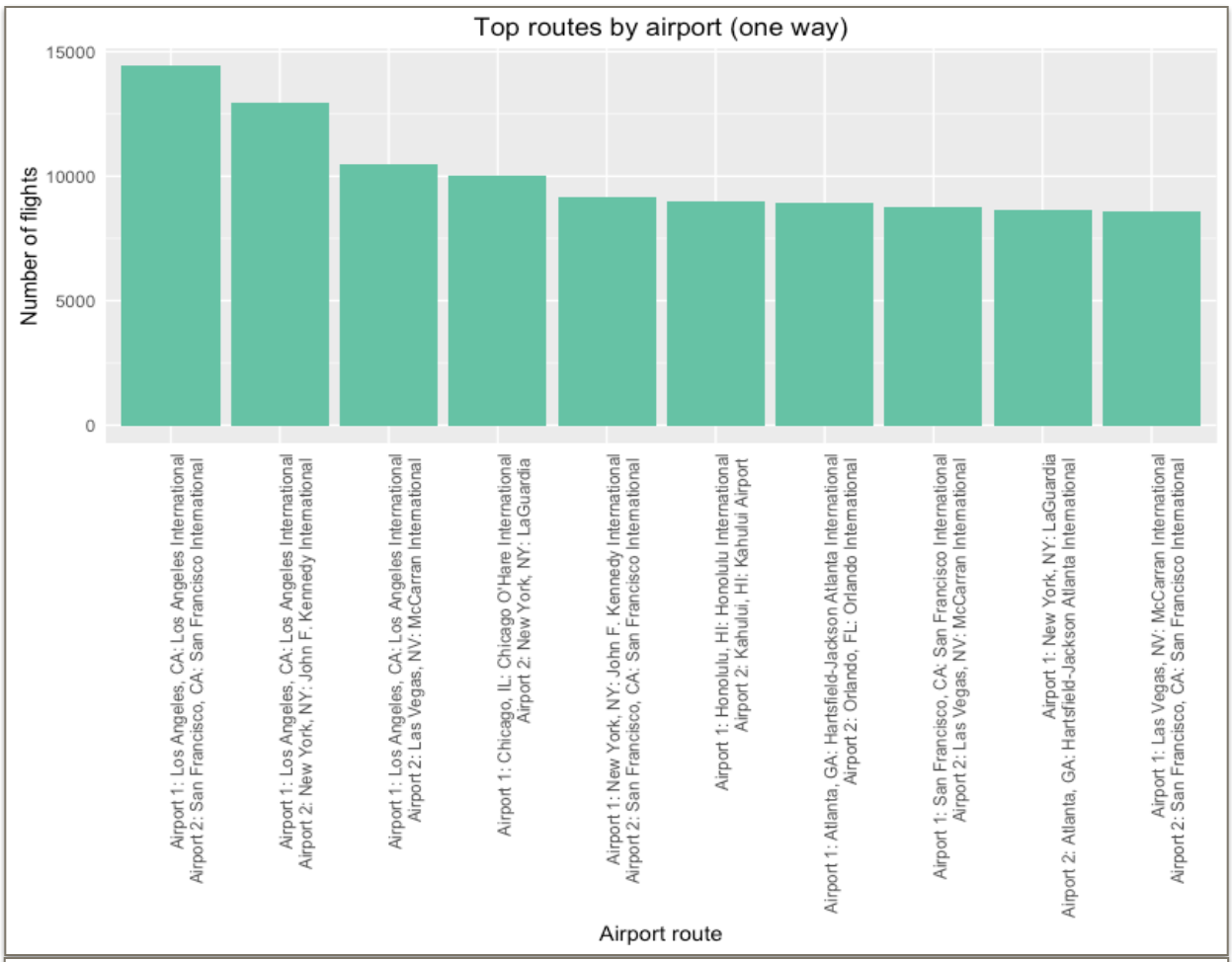
I next looked at the highest volume airports. Above are the top 30 airports by number of arrivals. Since the number of arrivals and departures from an airport are almost identical, this can also be used to describe the top airports by total volume (arrivals and departures). The highest volume airport is Hartsfield-Jackson Atlanta. As some cities contain multiple airports, we can also examine volume by metropolitan area.

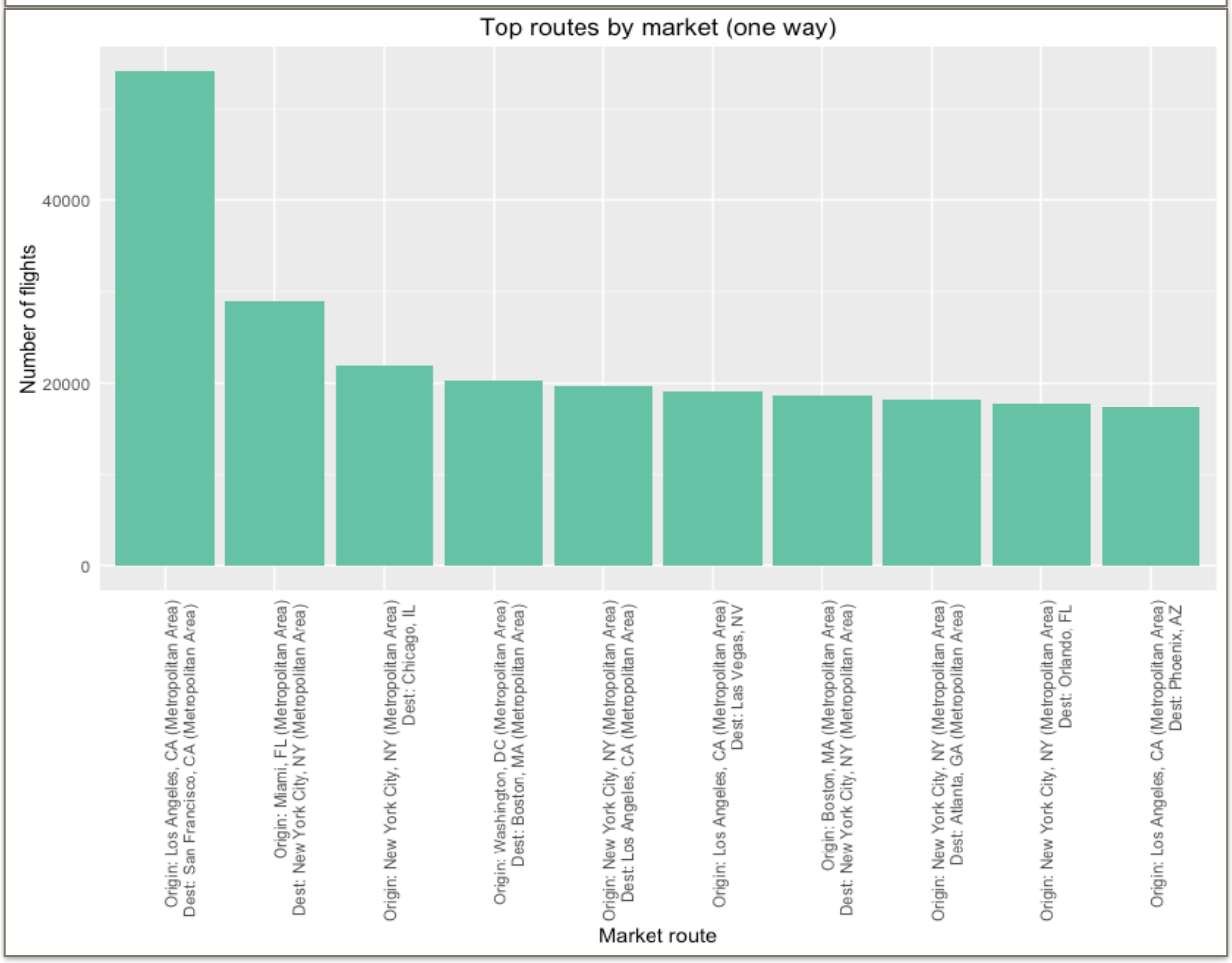




By this measure, Chicago is the busiest air market. We can see that it has two airports in the top 30 (although New York has three).

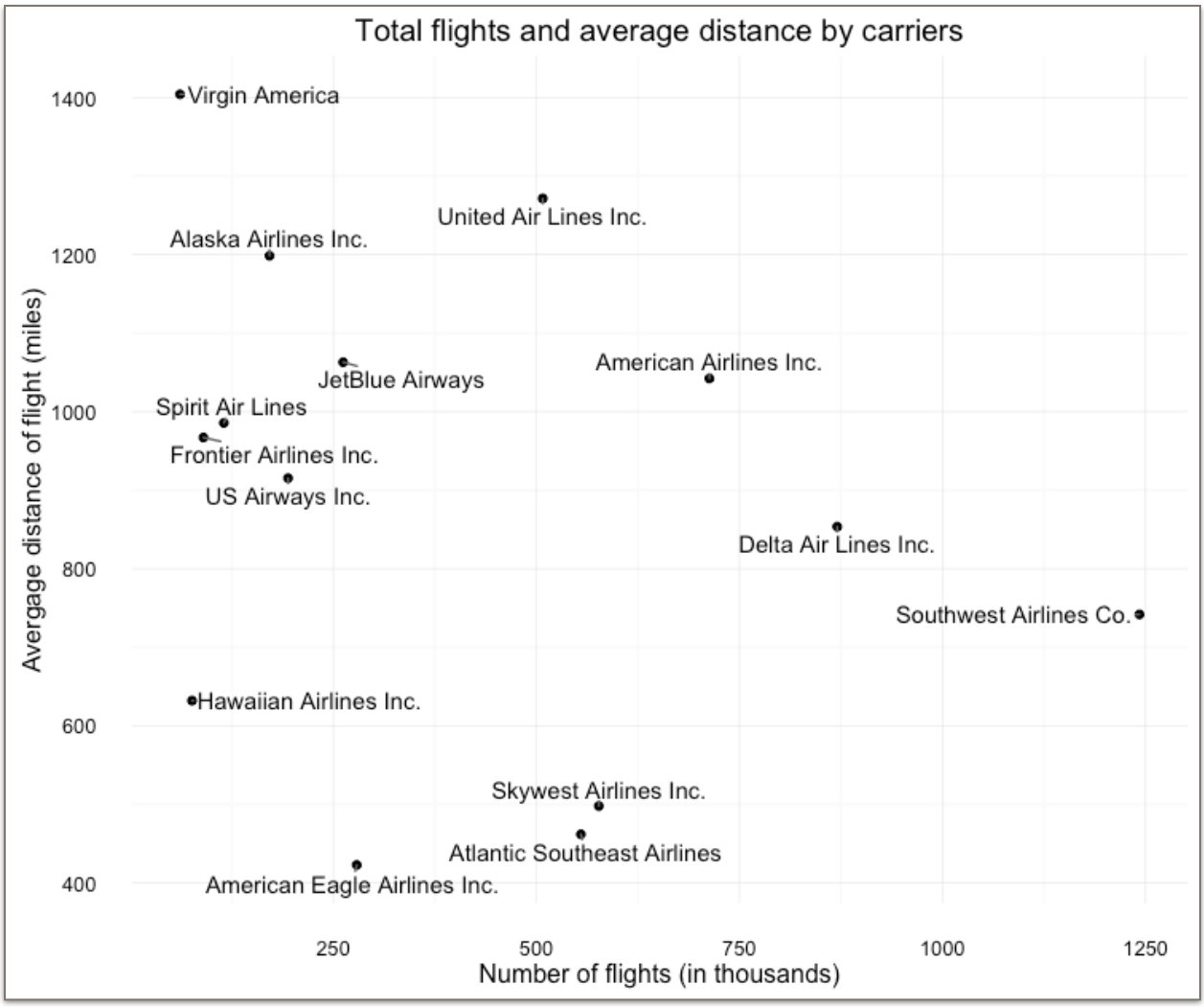
I also looked at the most popular travel routes. Below are the 10 most popular routes. The most flights occurred between Los Angeles International Airport and San Francisco International airport, followed by LAX and JFK. As with the airport data above, this is one-way with a nearly identical number of flights in the reverse direction. The top route by market is also between Los Angeles and San Francisco; the margin between this route and the next most popular route, New York and Miami, is now quite large.


Top routes by airport (one way)
Number of flights
15000
10000
5000
0
Airport 1: Los Angeles, CA: Los Angeles International
Airport 2: San Francisco, CA: San Francisco International
Airport 1: Los Angeles, CA: Los Angeles International
Airport 2: New York, NY: John F. Kennedy International
Airport 1: Los Angeles, CA: Los Angeles International
Airport 2: Las Vegas, NV: McCarran International
Airport 1: Chicago, IL: Chicago O'Hare International
Airport 2: New York, NY: LaGuardia
Airport 1: New York, NY: John F. Kennedy International
Airport 2: San Francisco, CA: San Francisco International
Airport 1: Honolulu, HI: Honolulu International
Airport 2: Kahului, HI: Kahului Airport
Airport 1: Atlanta, GA: Hartsfield-Jackson Atlanta International
Airport 2: Orlando, FL: Orlando International
Airport 1: San Francisco, CA: San Francisco International
Airport 2: Las Vegas, NV: McCarran International
Airport 1: New York, NY: LaGuardia
Airport 2: Atlanta, GA: Hartsfield-Jackson Atlanta International
Airport 1: Las Vegas, NV: McCarran International
Airport 2: San Francisco, CA: San Francisco International
Airport route


Top routes by market (one way)
Number of flights
40000
20000
0
Origin: Los Angeles, CA (Metropolitan Area)
Dest: San Francisco, CA (Metropolitan Area)
Origin: Miami, FL (Metropolitan Area)
Dest: New York City, NY (Metropolitan Area)
Origin: New York City, NY (Metropolitan Area)
Dest: Chicago, IL
Origin: Washington, DC (Metropolitan Area)
Dest: Boston, MA (Metropolitan Area)
Origin: New York City, NY (Metropolitan Area)
Dest: Los Angeles, CA (Metropolitan Area)
Origin: Los Angeles, CA (Metropolitan Area)
Dest: Las Vegas, NV
Origin: Boston, MA (Metropolitan Area)
Dest: New York City, NY (Metropolitan Area)
Origin: New York City, NY (Metropolitan Area)
Dest: Atlanta, GA (Metropolitan Area)
Origin: New York City, NY (Metropolitan Area)
Dest: Orlando, FL
Origin: Los Angeles, CA (Metropolitan Area)
Dest: Phoenix, AZ
Market route

Airlines

In 2015, American Airlines and US Airways merged to create the so-called “Big Four” airlines: Southwest, United, Delta, and American. In 2015, these four airlines plus US Airways (which began reporting as American in July 2015), accounted for over 60% of flights. Additionally, two other high volume airlines, Skywest Airlines and Atlantic Southeast Airlines, operate as contractors for the Big Four airlines, making the share of the airline industry they control even higher. Below we have a graph of number of flights and average flight distance of the 14 major carriers (those that account for at least 1%



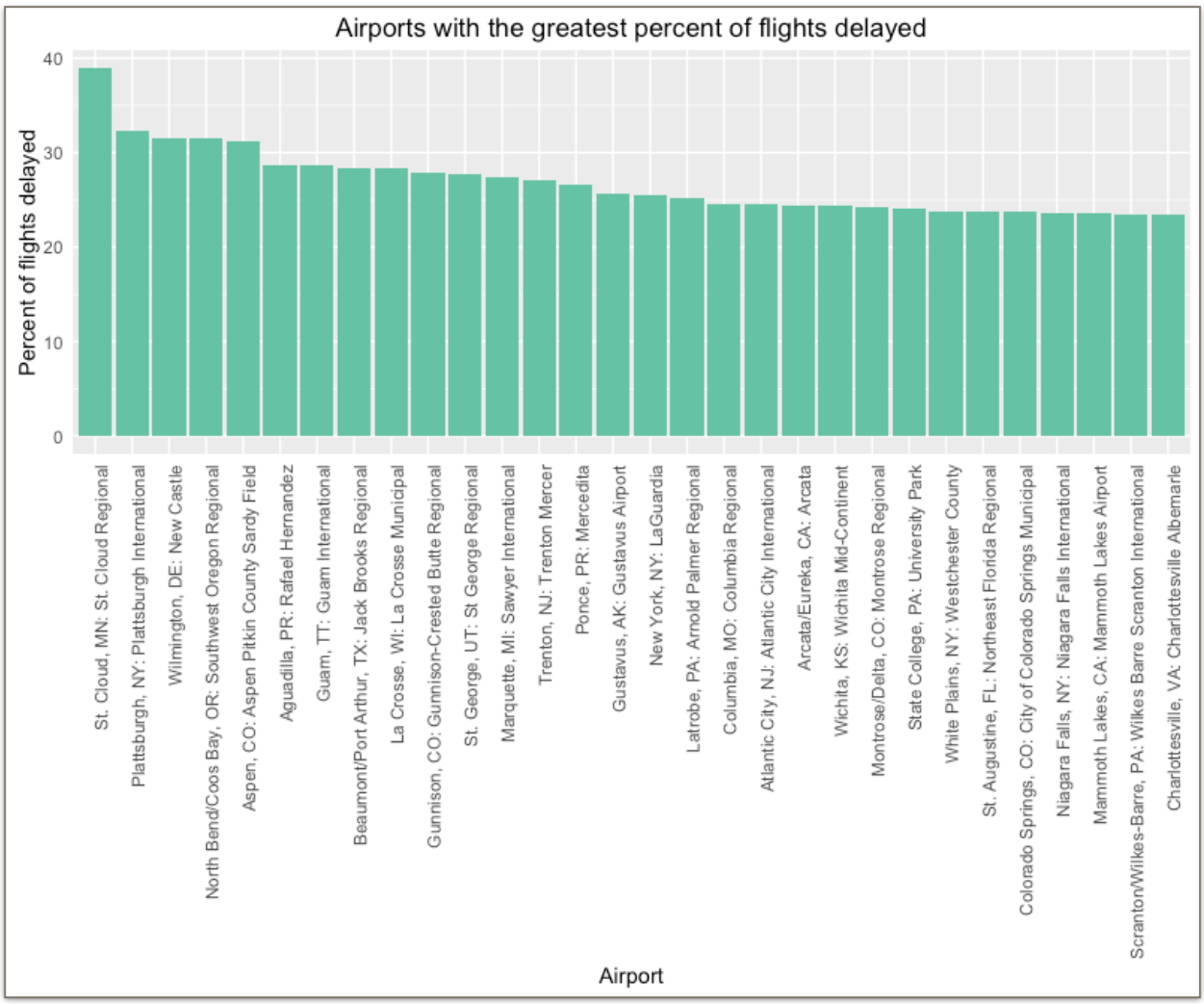
of domestic passenger revenue). From this we can see that the business strategy of airlines differs. Southwest operates a very large number of flights with a low average distance. Virgin America, on the other hand, focuses on fewer, but longer flights.

III. Delays

Most people who fly have experienced the following scenario: Leave hours before your flight, and wait in ever-growing security lines, only to find out that your flight has been delayed. Once your flight is delayed you don't know if it will be for minutes or hours. Is there any knowledge we can gain for the DOT data to help passengers and airlines better manage delays?

The DOT data contains scheduled departure and arrival and actual departure and arrival data that comes from carriers' Computer Reservation System (CRS). Additionally, since 2003, the cause of delay has been recorded (more on this below). I defined a delayed flight as one which arrives greater than 15 minutes past its scheduled arrival time (this is also the FAA definition of delayed). I also used data in which early flights are designated as being 0 minutes delayed (as opposed to negative). We also remark that according to the DOT, 29 airports report on-time data, but the data used contains more than just those airports.

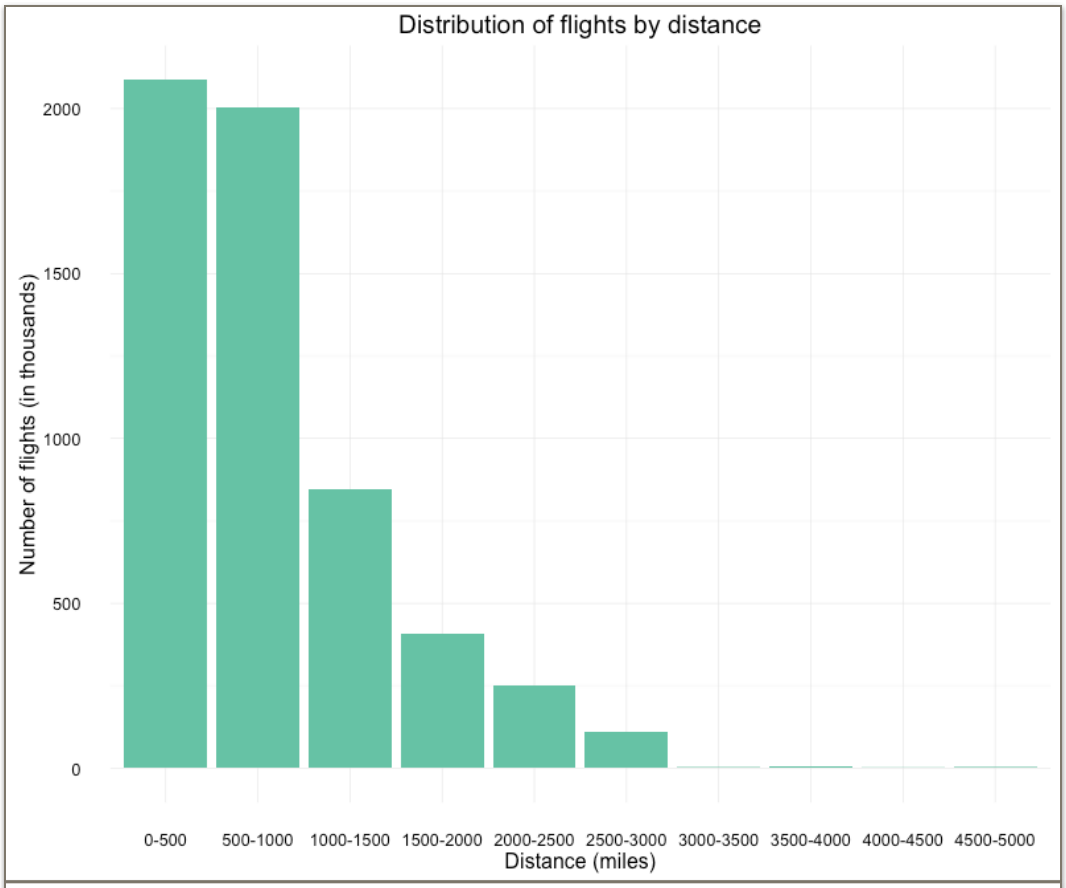
I focused on arrival delay (as opposed to departure delay) because I believe this is of greatest interest to customers. In 2015, **18.6 % of all flights were delayed**. Among all flights, the **average flight was delayed 12.1 minutes** (recall this includes on-time and early flights which are recorded as having 0 delay). Among delayed flights, the average delay was just under an hour with a **median delay of 37 minutes.**

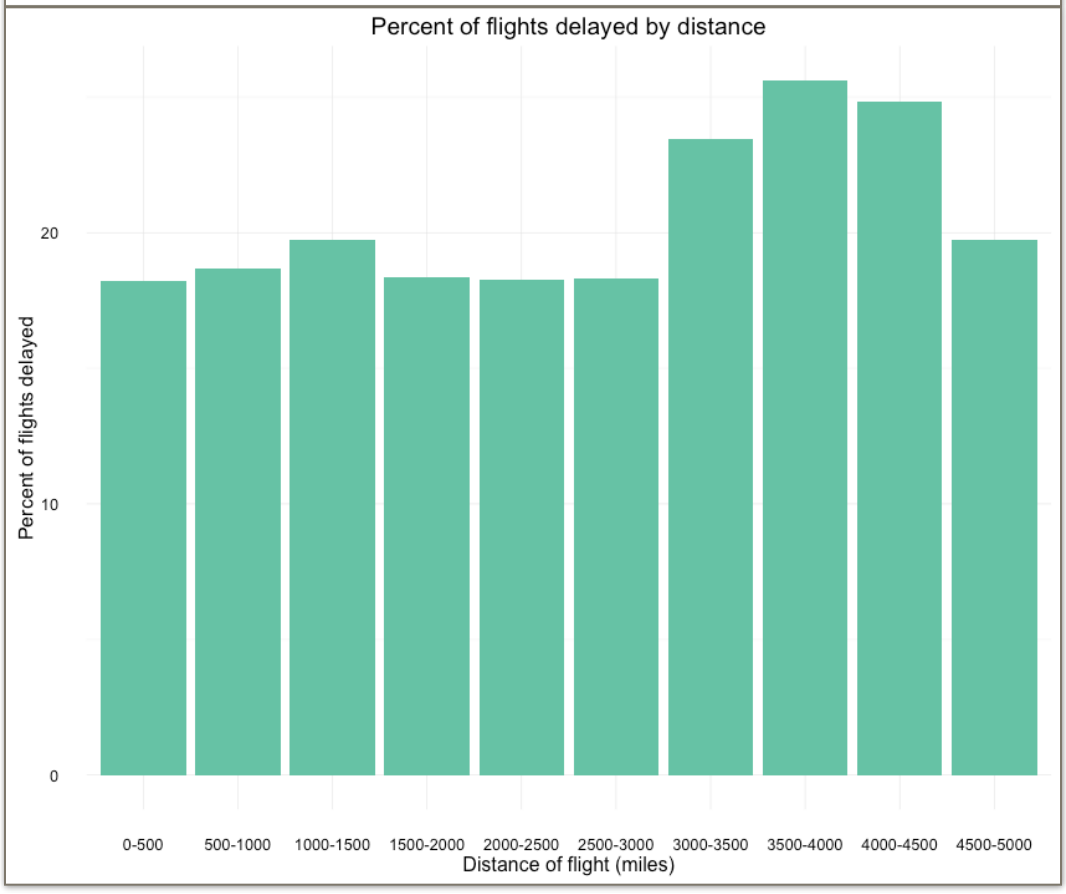


Above we see the 30 airports with largest percent of flights delayed. St. Cloud Regional Airport has the worst delay percentage.

Delays and distance

We saw earlier that different airlines had a wide range of average flight distance. Next, I looked into the relationship between flight distance and delays. Below is the distribution of flights by distance—we see that most flights are between 0 and 1000 miles. However, examining the percent of flights delayed by distance, we see that longer flights are more delayed.


Distribution of flights by distance
Number of flights (in thousands)
2000
1500
1000
500
0
0-500
500-1000
1000-1500
1500-2000
2000-2500
2500-3000
3000-3500
3500-4000
4000-4500
4500-5000
Distance (miles)

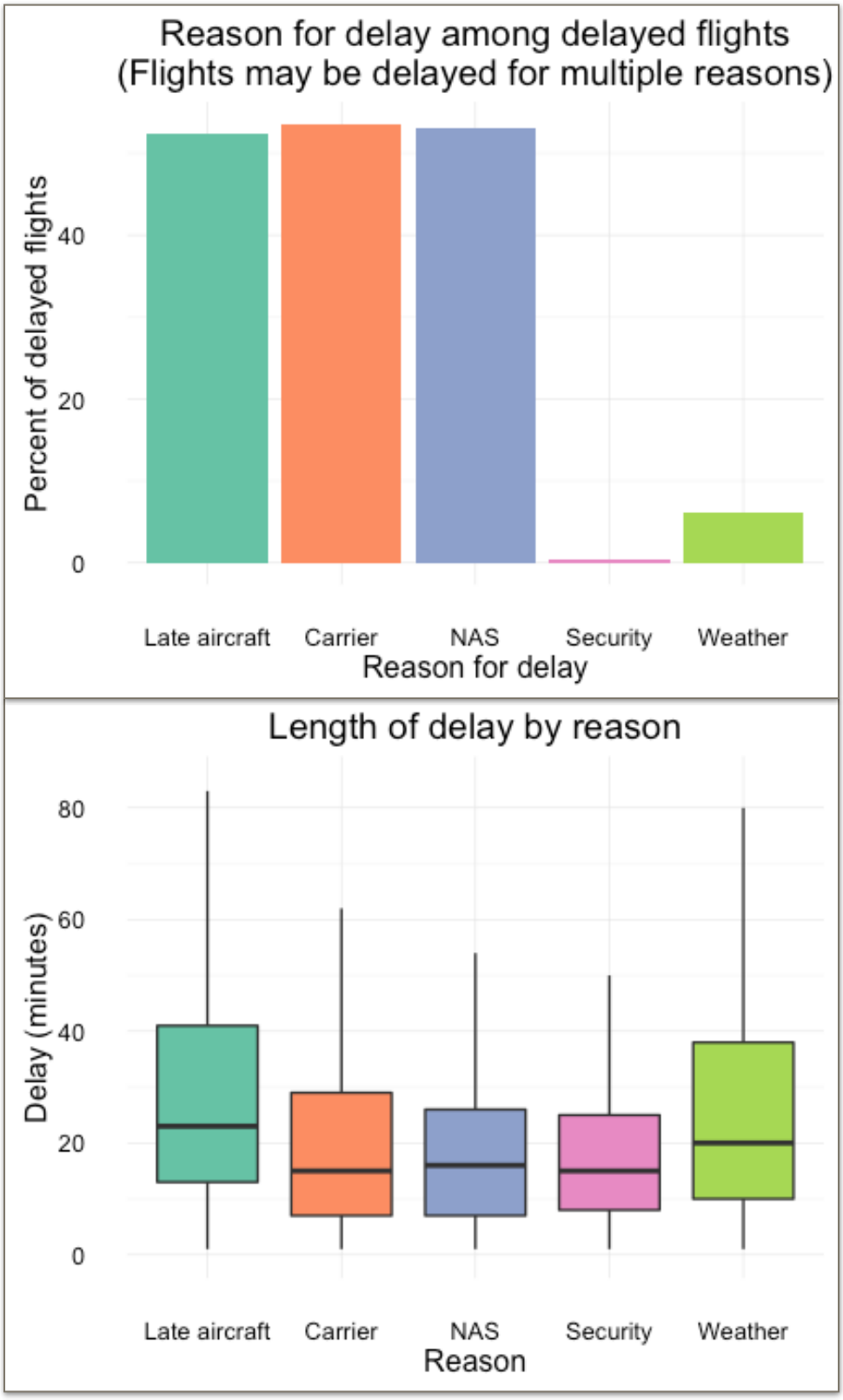

Percent of flights delayed by distance
Percent of flights delayed
20
10
0
0-500
500-1000
1000-1500
1500-2000
2000-2500
2500-3000
3000-3500
3500-4000
4000-4500
4500-5000
Distance of flight (miles)

Types of delays

The DOT data breaks down the causes of delays into five categories:

- Carrier: The cause of delay are within the airline's control (e.g. maintenance, baggage loading, fueling, etc.).
- Weather: Significant weather events.
- NAS (National Aviation System): Delays attributed to the national aviation system (e.g. non-extreme weather, airport operations, and air traffic control).
- Late aircraft: A previous flight arrived late.
- Security: Delays caused by the evacuation of a terminal due to a security breach.

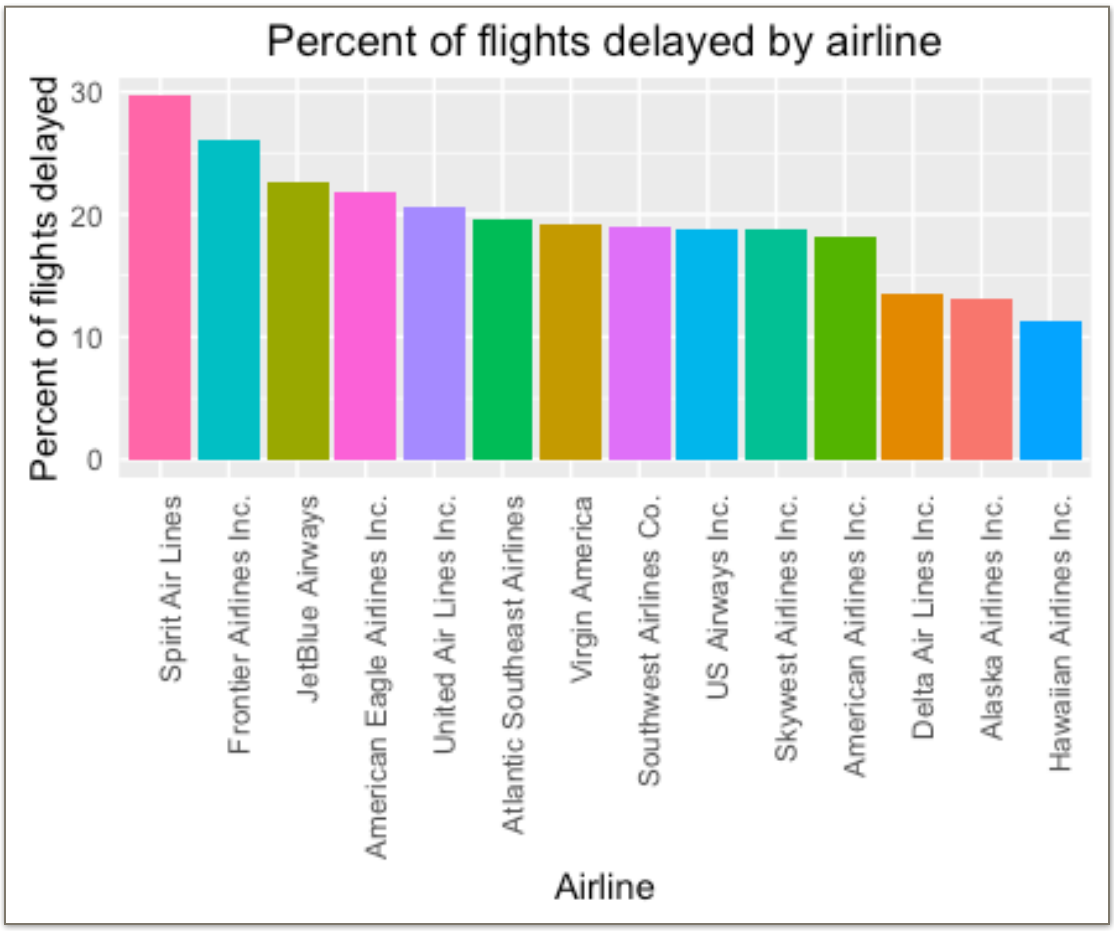
The DOT gives minutes of delay for each of these five reason reason. Below is a a description of the reasons for delay among delayed flights (defined as that reason causing a nonzero delay). Among delayed flights, approximately 50% involve a late aircraft delay, and the same for carrier and NAS delay. Surprising to me is that weather accounts for such a small percentage of delays. However, below we see that the median weather related delay is longer than security, NAS, and carrier delays. The spread of weather and late aircraft delays are also the widest.

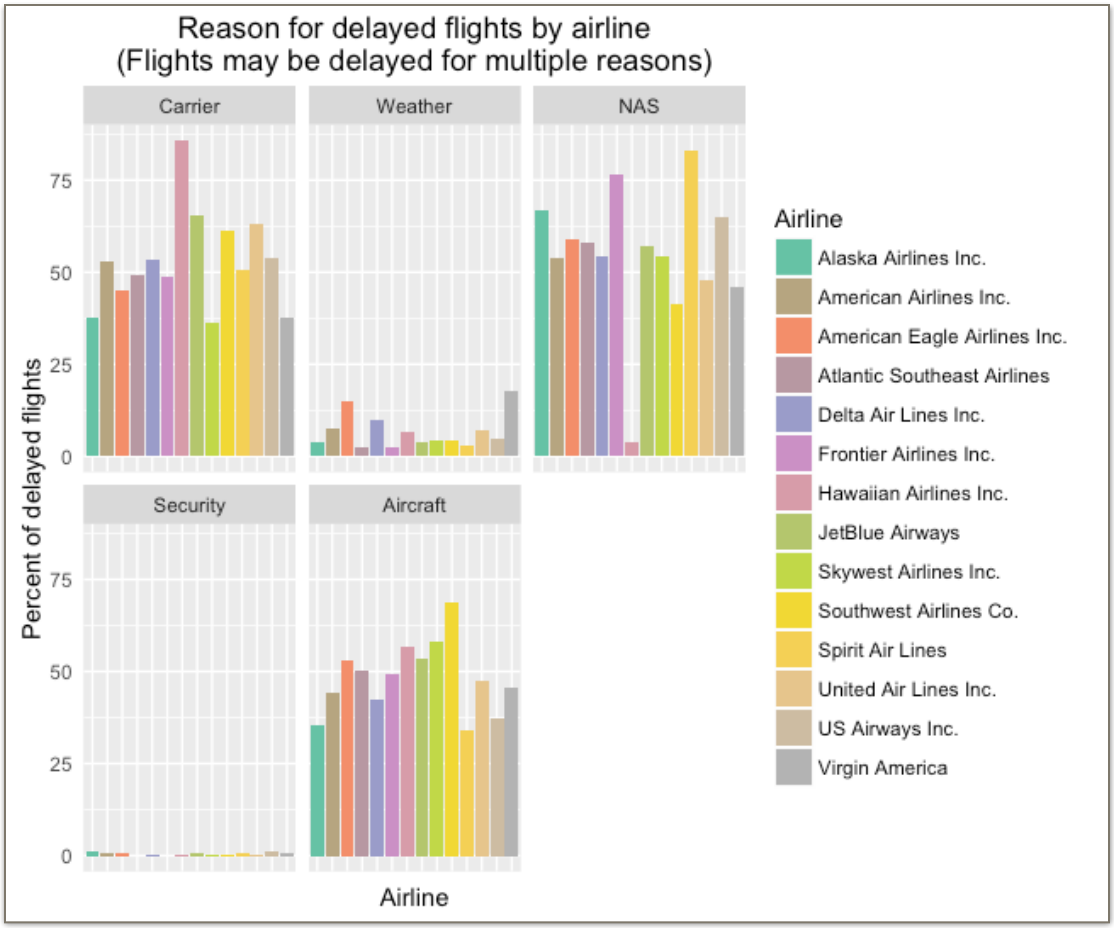

Reason for delay among delayed flights
(Flights may be delayed for multiple reasons)
Percent of delayed flights
0
20
40
Late aircraft
Carrier
NAS
Security
Weather
Reason for delay
Length of delay by reason
Delay (minutes)
0
20
40
60
80
Late aircraft
Carrier
NAS
Security
Weather
Reason

Delays and airlines

Next, I examined delays among airlines. Below we see airlines ranked by percent of flights that are delayed. Note that the lower volume airlines show up at both of the top and bottom of this ranking.

Finally, as above, I examined the reason for delay, now divided by airline. This could be useful for a customer deciding between airlines as well as for airlines to see how they are performing against their competition. For example, Southwest has a high number of late aircraft delays—possibly related to the largest volume of flights they operate. Of particular interest might be carrier delays: those delays that are directly within the airline's control. Frontier airlines might want to look into improving on this front.





IV. Predictive model

My goal was to build a model to predict whether or not a flight will be delayed. The variables that I downloaded from the DOT data set and considered using as predictors for my model were:

- Month
- Day of week
- Hour of departure
- Hour of arrival
- Distance covered

- Airline
- Origin airport
- Destination airport

With such a large data set, I decided to build and test my model on a sample of size 200,000. I then made an 80-20 training-test split on this sample. This was done to focus my analysis while still ensuring a random sample. I also limited the origin airport and destination airport to the 30 airports with the most traffic. This was also done to focus my analysis as well as the fact that these were almost the same as the 29 airports that the DOT says reports arrival data. We remark that Hawaiian Airlines does not have any flights between these airports.

Preliminary analysis

Before building any model, my exploratory data analysis suggested that delays are not independent of the origin and destination airport. To confirm this, I performed a chi square test for the response of being delayed or not and the destination airport. This returned a significant p-value ($p < 2.2e\text{-}16$) implying that these variables are not independent.

The model

I chose a logistic regression for my model due to its interpretability in order to give advice to consumers and airlines. I performed this analysis in R.

My final model uses the variables above with the following changes:

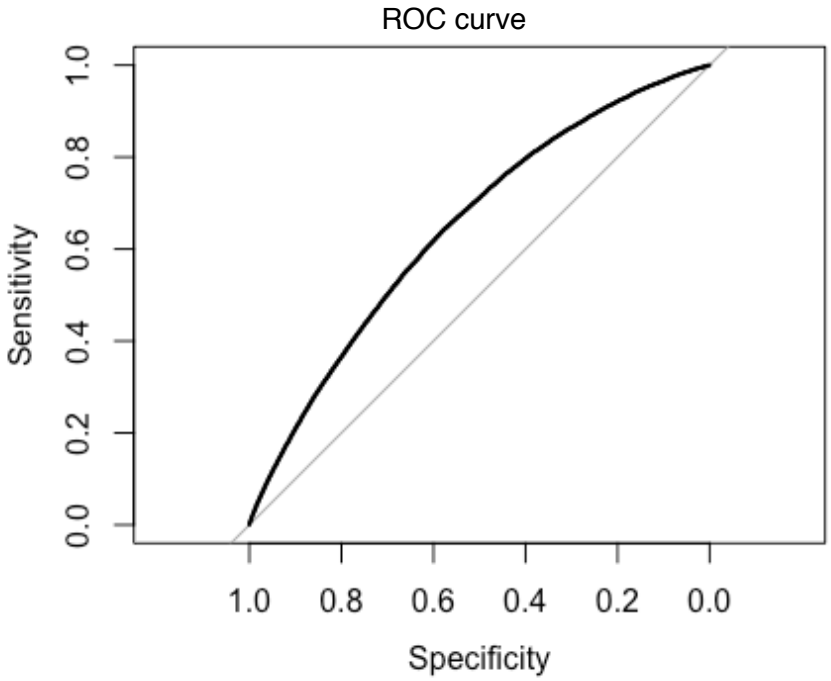
- I binned ‘Month’ into 4 seasons.
- I removed hour of arrival.

- I removed day of the week.

Although removing these feature increased my deviance, this increase was very small, so I opted to remove the features for simplicity of the model. Additionally, I performed a drop in deviance test, which returned an insignificant p-value (< 2.2e-16) implying that the reduced model is sufficient, i.e. the removed variables are not needed anyway.

Results of the model

My sample data set had 80% of flights not delayed. Due to the imbalance of the classification problem, I performed an ROC analysis. My model had an AUC of .65: this means my model has a 65% chance of assigning a higher score to a positive sample than to a negative sample (50% would be a random model). This indicates that the model is not a very good fit and I will comment on this in the next section. I've included the ROC curve below.



I used the ROC analysis to chose the optimal threshold (.19). This optimal threshold does not maximize the accuracy, but rather sensitivity + specificity, where sensitivity is the true positive rate and specificity is the true negative rate. With this optimal threshold the test accuracy (59%) is lower than the baseline, but has a test sensitivity of 63% and specificity of 58%. With this model, although not as accurate as the baseline model on the whole, correctly identifies approximately 60% of both delayed and not delayed flights. Depending on the application, one may want a model to increases sensitivity (such as the baseline guess) or specificity.

I performed an overall goodness of fit test, which returned an insignificant p-value (1) indicating that the difference between my model and the null model (predicting only non-delayed flights) is statistically significant. However, as described above, that statistical significance does not necessarily translate to practical accuracy.

Interpretation of the model

I've included at the end of this report a screenshot from R of all of the coefficients of my model. However, in this section I will walk though some of the insights one can gain from examining the model.

All of my predictors, except for distance, are categorical and hence dummified. One level of each factor is taken as the "base" factor—it does not have a coefficient and is the baseline level that other coefficients are taken with respect to. These base levels are:

- Departure time: 12am-1am.
- Airline: Alaska airlines
- Origin and departure airport: Atlanta

- Season: Fall

With this is the baseline we can make the following observations about our model:

- All of the airline coefficients are statistically significant. The coefficients tell us whether that airline is more likely/less likely to be delayed and the degree to which it affects the model (compared to Alaska Airlines). For example, Spirit has the highest positive coefficient meaning flying Spirit will increase the probability of a delayed flight most in the model. Frontier airlines also does not perform well in this model (agreeing with the earlier EDA).
- All of the departure hour coefficients are statistically significant. We see that the only hour of departure resulting in a lower prediction of delay than 12am-1am is 6am-7am.
- The longer the distance of the flight, the less likely it will predict the flight is delayed—contrary to our preliminary EDA.
- Several of the origin and destination airport coefficients are significant. For example, having a destination of Portland decreases the probability of a delay in this model.

By examine the coefficients further, there are even more insights to be gained.

```
Coefficients:
                                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                     -3.266e+00  8.809e-02 -37.073  < 2e-16 ***
DEP_TIME_BLK0600-0659                                           -1.647e-01  6.379e-02  -2.582 0.009814 **
DEP_TIME_BLK0700-0759                                            1.912e-01  6.091e-02   3.139 0.001695 **
DEP_TIME_BLK0800-0859                                            3.452e-01  6.089e-02   5.669 1.43e-08 ***
DEP_TIME_BLK0900-0959                                            4.490e-01  6.144e-02   7.308 2.72e-13 ***
DEP_TIME_BLK1000-1059                                            5.712e-01  6.063e-02   9.421  < 2e-16 ***
DEP_TIME_BLK1100-1159                                            6.595e-01  5.995e-02  11.002  < 2e-16 ***
DEP_TIME_BLK1200-1259                                            7.699e-01  6.044e-02  12.740  < 2e-16 ***
DEP_TIME_BLK1300-1359                                            8.667e-01  5.955e-02  14.554  < 2e-16 ***
DEP_TIME_BLK1400-1459                                            9.018e-01  6.008e-02  15.010  < 2e-16 ***
DEP_TIME_BLK1500-1559                                            1.059e+00  5.899e-02  17.946  < 2e-16 ***
DEP_TIME_BLK1600-1659                                            1.035e+00  5.962e-02  17.361  < 2e-16 ***
DEP_TIME_BLK1700-1759                                            1.129e+00  5.869e-02  19.229  < 2e-16 ***
DEP_TIME_BLK1800-1859                                            1.239e+00  5.926e-02  20.914  < 2e-16 ***
DEP_TIME_BLK1900-1959                                            1.286e+00  5.929e-02  21.691  < 2e-16 ***
DEP_TIME_BLK2000-2059                                            1.221e+00  6.051e-02  20.184  < 2e-16 ***
DEP_TIME_BLK2100-2159                                            1.186e+00  6.351e-02  18.676  < 2e-16 ***
DEP_TIME_BLK2200-2259                                            1.056e+00  6.795e-02  15.534  < 2e-16 ***
DEP_TIME_BLK2300-2359                                            7.612e-01  8.087e-02   9.413  < 2e-16 ***
DISTANCE                                                        -4.383e-05  1.162e-05  -3.773 0.000161 ***
AIRLINEAmerican Airlines Inc.                                    3.490e-01  5.215e-02   6.692 2.20e-11 ***
AIRLINEAmerican Eagle Airlines Inc.                              7.151e-01  8.928e-02   8.009 1.15e-15 ***
AIRLINEAtlantic Southeast Airlines                               6.299e-01  7.626e-02   8.260  < 2e-16 ***
AIRLINEDelta Air Lines Inc.                                      1.624e-01  5.224e-02   3.108 0.001881 **
AIRLINEFrontier Airlines Inc.                                    8.111e-01  6.381e-02  12.712  < 2e-16 ***
AIRLINEJetBlue Airways                                           6.568e-01  5.778e-02  11.367  < 2e-16 ***
AIRLINESkywest Airlines Inc.                                     5.515e-01  5.712e-02   9.656  < 2e-16 ***
AIRLINESouthwest Airlines Co.                                    5.893e-01  5.216e-02  11.299  < 2e-16 ***
AIRLINESpirit Air Lines                                          1.180e+00  5.822e-02  20.270  < 2e-16 ***
AIRLINEUnited Air Lines Inc.                                     4.351e-01  5.209e-02   8.353  < 2e-16 ***
AIRLINEUS Airways Inc.                                           3.602e-01  5.813e-02   6.197 5.75e-10 ***
AIRLINEVirgin America                                            2.965e-01  6.388e-02   4.642 3.45e-06 ***
ORIGIN_AIRPORTBaltimore, MD: Baltimore/Washington International Thurgood Marshall  1.275e-01  5.293e-02   2.409 0.016006 *
ORIGIN_AIRPORTBoston, MA: Logan International                    1.684e-01  4.462e-02   3.775 0.000160 ***
ORIGIN_AIRPORTCharlotte, NC: Charlotte Douglas International     1.763e-01  4.824e-02   3.655 0.000257 ***
ORIGIN_AIRPORTChicago, IL: Chicago Midway International          1.006e-01  5.382e-02   1.869 0.061570 .
ORIGIN_AIRPORTChicago, IL: Chicago O'Hare International          3.662e-01  4.007e-02   9.137  < 2e-16 ***
ORIGIN_AIRPORTDallas, TX: Dallas Love Field                      2.227e-01  6.211e-02   3.586 0.000335 ***
ORIGIN_AIRPORTDallas/Fort Worth, TX: Dallas/Fort Worth International  1.743e-01  4.435e-02   3.929 8.52e-05 ***
ORIGIN_AIRPORTDenver, CO: Denver International                   1.902e-01  4.190e-02   4.540 5.62e-06 ***
ORIGIN_AIRPORTDetroit, MI: Detroit Metro Wayne County            4.397e-02  4.756e-02   0.924 0.355255
ORIGIN_AIRPORTFort Lauderdale, FL: Fort Lauderdale-Hollywood International  -1.308e-01  5.131e-02  -2.549 0.010808 *
ORIGIN_AIRPORTHouston, TX: George Bush Intercontinental/Houston  2.351e-01  4.780e-02   4.917 8.77e-07 ***
ORIGIN_AIRPORTHouston, TX: William P Hobby                       7.736e-02  6.271e-02   1.234 0.217366
ORIGIN_AIRPORTLas Vegas, NV: McCarran International              1.113e-01  4.351e-02   2.557 0.010555 *
```

```
ORIGIN_AIRPORTLos Angeles, CA: Los Angeles International                  2.635e-01  4.001e-02   6.585 4.55e-11 ***
ORIGIN_AIRPORTMiami, FL: Miami International                             2.920e-01  4.960e-02   5.888 3.91e-09 ***
ORIGIN_AIRPORTMinneapolis, MN: Minneapolis-St Paul International         1.086e-01  4.747e-02   2.287 0.022181 *
ORIGIN_AIRPORTNew York, NY: John F. Kennedy International               1.596e-01  4.884e-02   3.268 0.001082 **
ORIGIN_AIRPORTNew York, NY: LaGuardia                                   3.211e-01  4.461e-02   7.198 6.12e-13 ***
ORIGIN_AIRPORTNewark, NJ: Newark Liberty International                  2.344e-01  4.905e-02   4.778 1.77e-06 ***
ORIGIN_AIRPORTOrlando, FL: Orlando International                        1.565e-02  4.619e-02   0.339 0.734670
ORIGIN_AIRPORTPhiladelphia, PA: Philadelphia International              2.497e-01  4.979e-02   5.014 5.32e-07 ***
ORIGIN_AIRPORTPhoenix, AZ: Phoenix Sky Harbor International             1.101e-01  4.471e-02   2.463 0.013759 *
ORIGIN_AIRPORTPortland, OR: Portland International                      -5.018e-02  6.643e-02  -0.755 0.450020
ORIGIN_AIRPORTSalt Lake City, UT: Salt Lake City International          8.245e-03  5.355e-02   0.154 0.877644
ORIGIN_AIRPORTSan Diego, CA: San Diego International                    -8.067e-02  5.355e-02  -1.506 0.131973
ORIGIN_AIRPORTSan Francisco, CA: San Francisco International            2.441e-01  4.316e-02   5.656 1.55e-08 ***
ORIGIN_AIRPORTSeattle, WA: Seattle/Tacoma International                 2.713e-01  5.091e-02   5.330 9.83e-08 ***
ORIGIN_AIRPORTTampa, FL: Tampa International                            -7.763e-02  5.470e-02  -1.419 0.155850
ORIGIN_AIRPORTWashington, DC: Ronald Reagan Washington National         3.147e-02  4.890e-02   0.644 0.519824
DEST_AIRPORTBaltimore, MD: Baltimore/Washington International Thurgood Marshall  5.245e-03  5.377e-02   0.098 0.922295
DEST_AIRPORTBoston, MA: Logan International                             1.035e-01  4.428e-02   2.338 0.019364 *
DEST_AIRPORTCharlotte, NC: Charlotte Douglas International              9.352e-03  5.050e-02   0.185 0.853089
DEST_AIRPORTChicago, IL: Chicago Midway International                   -8.002e-02  5.619e-02  -1.424 0.154396
DEST_AIRPORTChicago, IL: Chicago O'Hare International                   2.292e-01  4.106e-02   5.582 2.38e-08 ***
DEST_AIRPORTDallas, TX: Dallas Love Field                               -5.670e-02  6.415e-02  -0.884 0.376768
DEST_AIRPORTDallas/Fort Worth, TX: Dallas/Fort Worth International      1.934e-01  4.437e-02   4.358 1.31e-05 ***
DEST_AIRPORTDenver, CO: Denver International                            4.711e-02  4.318e-02   1.091 0.275243
DEST_AIRPORTDetroit, MI: Detroit Metro Wayne County                     -4.988e-02  4.872e-02  -1.024 0.305912
DEST_AIRPORTFort Lauderdale, FL: Fort Lauderdale-Hollywood International  1.185e-01  4.906e-02   2.416 0.015689 *
DEST_AIRPORTHouston, TX: George Bush Intercontinental/Houston           2.213e-01  4.818e-02   4.593 4.36e-06 ***
DEST_AIRPORTHouston, TX: William P Hobby                                -1.370e-02  6.364e-02  -0.215 0.829550
DEST_AIRPORTLas Vegas, NV: McCarran International                       -4.103e-02  4.408e-02  -0.931 0.351916
DEST_AIRPORTLos Angeles, CA: Los Angeles International                  2.597e-01  3.995e-02   6.500 8.02e-11 ***
DEST_AIRPORTMiami, FL: Miami International                              3.382e-01  5.040e-02   6.709 1.96e-11 ***
DEST_AIRPORTMinneapolis, MN: Minneapolis-St Paul International          -5.765e-02  4.947e-02  -1.165 0.243901
DEST_AIRPORTNew York, NY: John F. Kennedy International                 3.673e-01  4.799e-02   7.655 1.94e-14 ***
DEST_AIRPORTNew York, NY: LaGuardia                                     4.745e-01  4.394e-02  10.798  < 2e-16 ***
DEST_AIRPORTNewark, NJ: Newark Liberty International                    2.434e-01  4.887e-02   4.981 6.33e-07 ***
DEST_AIRPORTOrlando, FL: Orlando International                          2.473e-01  4.470e-02   5.532 3.16e-08 ***
DEST_AIRPORTPhiladelphia, PA: Philadelphia International                1.407e-01  5.073e-02   2.774 0.005546 **
DEST_AIRPORTPhoenix, AZ: Phoenix Sky Harbor International               -4.320e-02  4.512e-02  -0.957 0.338337
DEST_AIRPORTPortland, OR: Portland International                        -1.871e-01  6.545e-02  -2.859 0.004250 **
DEST_AIRPORTSalt Lake City, UT: Salt Lake City International            -1.407e-01  5.313e-02  -2.649 0.008076 **
DEST_AIRPORTSan Diego, CA: San Diego International                      -1.218e-01  5.278e-02  -2.309 0.020961 *
DEST_AIRPORTSan Francisco, CA: San Francisco International              3.169e-01  4.244e-02   7.467 8.23e-14 ***
DEST_AIRPORTSeattle, WA: Seattle/Tacoma International                   9.968e-02  5.104e-02   1.953 0.050832 .
DEST_AIRPORTTampa, FL: Tampa International                              9.398e-02  5.271e-02   1.783 0.074569 .
DEST_AIRPORTWashington, DC: Ronald Reagan Washington National           8.557e-03  4.925e-02   0.174 0.862068
MONTH_BUCKETSpring                                                      3.330e-01  2.416e-02  13.783  < 2e-16 ***
MONTH_BUCKETSummer                                                      4.827e-01  2.073e-02  23.280  < 2e-16 ***
MONTH_BUCKETWinter                                                      5.230e-01  2.112e-02  24.767  < 2e-16 ***
```

Comments on the model

I arrived at the model described above after trying several other logistic regression models. This model combined accuracy and interpretability. I tried binning time variables, removing certain predictors, and I chose this model as the best model. Al-

though it is not very accurate (as described above), I believe that having a model that can predict delays and non-delays 60% of the time is still useful to consumers and airlines. Additionally, the model is simple enough the coefficients can easily be interpreted (as I did above). If greater accuracy is desired I believe other predictors should be incorporated—I will discuss this in the next section.

V. Conclusions

The DOT on-time airline data has a lot to offer consumers and airlines. Using this data can help identify problems that can be solved in the airline industry and better educate fliers. There are several aspects of this data that I am interested in exploring as it pertains to delays and improving my delay prediction model.

- Join weather data and investigate if weather is a good predictor of delay. Weather as a predictor is not as useful for a consumer who would like to know how likely his/her flight is to be delayed 2 weeks in the future (when weather predictions are inaccurate), but would still be interesting to look at. It could also help airlines better prepare for inclement weather.
- Examine the association between delay and the region and route of a flight. I mentioned earlier that I experimented in binning some of the categorical data to use in my model. I would like to attempt to do this for region of origin/destination. For example, are flights to the southeast more likely to be delayed than flights to the northeast? Similarly examining delays by route (as opposed to origin and destination separate, as I did).

- Incorporate time series analysis. As I mentioned at the beginning, I only analyzed data from 2015. I am interested to know how delays how changed over time. I would also like to know if this information had any predictive value, and if so, add it to my model.