Capstone Project 1: Analysis of Departure Delays in the U.S.

Jonathon Poage
Springboard Data Science Career Track
3/7/2019

Table of Contents

Introduction

Air travel is an essential mode of transportation that many people rely on for personal and business purposes. U.S. airlines carried an estimated 849 million passengers in 2017 alone. Despite impressive advancements in air travel systems, departure delays for commercial flights are still a fairly common occurrence. Departure delays can negatively impact passengers, airlines, and airports in significant ways. They often arise unexpectedly, making it difficult to schedule actual departure and arrival times effectively in advance. Reducing the average departure delay duration by even one minute could be practically significant for airport operations.

This project analyzed departure delays for non-stop domestic flights in the United States. This project focused on the top 10 busiest U.S. airports, in terms of passenger traffic, and the 8 largest U.S. major airline carriers, in terms of passengers carried. The analysis combined on-time performance data for flights in 2016 and 2017 with weather data reported by METAR stations at the airports.

Goals of Analysis

This project had three primary goals. These goals are discussed below, along with a description of the analysis for this project.

The first goal was to examine departure delay distributions in general and to identify the airline carrier and airport with the worst performance in terms of departure delays. The analysis examined characteristics of the overall distribution of departure delays and investigated trends over time. Next, the departure delay distributions were compared between airline carriers using parallel box plots. Departure delay distributions were similarly compared between airports. The worst performers in terms of departure delays were identified.

The second goal was to examine departure delays under various weather conditions and to determine whether there is an association between airports and weather conditions. The analysis identified certain weather conditions with mean departure delays that were significantly higher than the mean departure delay under clear conditions. The analysis also compared weather conditions between airports, and the results indicated that there is an association between airports and weather conditions.

The third goal of this analysis was to identify the relationships between departure delay duration and various weather variables. A Lasso regression was performed to examine the relationship between departure delay duration and six weather variables. Next, conditional linear regressions were performed to examine the relationship between departure delay duration and sky level 1 altitude as well as the relationship between departure delay duration and wind

speed. The results of the regressions indicated that sky level 1 altitude and wind speed are of practical significance. Confidence intervals for the mean sky level 1 altitude and the mean wind speed were compared between airports to determine which airports had the worst conditions with respect to these two weather variables.

Benefit to Clients

This project serves the airport authorities and major airline carriers. Airport authorities could potentially use the results of this analysis to better anticipate departure delays for future flights under certain conditions, which would help them plan airport operations. More accurate scheduling of departures and arrivals could directly improve efficiency of airport operations, which may indirectly create other benefits such as improved passenger satisfaction and increased profits.

In addition, this analysis highlighted potential areas of improvement for certain airports and airline carriers. The airline carriers and airports with the worst performance in terms of departure delays were identified. The airports with the worst conditions for wind speed and sky level 1 altitude were also identified. The airport authorities and airline carriers involved could investigate ways to reduce departure delay durations.

Analysis Code

See below for links to the Jupyter Notebooks that contain the analysis code used for this project.

Data Wrangling

Data Story

Exploratory Data Analysis (EDA)

In-Depth Analysis

Data Preprocessing

This project involves data collected from two different sources. One data source provided On-Time performance data for flights, and the other data source provided weather data. Data was collected from each source and preprocessed. The preprocessed flight data was later combined with the preprocessed weather data by merging flight information with weather conditions based on location and time. The following sections describe the data collection and preprocessing methods.

Flight Data

The On-Time performance data for flights was collected by downloading csv files from the Bureau of Transportation Statistics (BTS) website. Each file contains data for one month during the years 2016 and 2017. The flight data includes information about departure times, departure delays, airport codes for origin and destination, and unique carrier codes. The data in the files was preprocessed using python.

The full unfiltered dataset contained millions of data points. The data was filtered during the preprocessing procedure to reduce the number of data points in a way that highlighted the most valuable information for a streamlined analysis. This project focused on data for the top 10 busiest U.S. airports (See **Appendix A**) and the top 8 largest U.S. airline carriers (See **Appendix B**). This project also focused on departure delays from the origin airport. See **Appendix C** for more details about the flight data preprocessing procedure.

Weather Data

The weather information for this analysis was obtained from Meteorological Terminal Aviation Routine Weather Reports (METARs) generated by weather stations at the top 10 busiest U.S. airports. METARs include observations about weather conditions, such as visibility, wind speed, temperature, and precipitation type. They are typically generated hourly, though at some locations they may be generated more frequently. See the Federal Meteorological Handbook No. 1 (FMH1) for more details about reporting standards for METARs.

Weather data was originally intended to be collected from the National Oceanic and Atmospheric Administration (NOAA). Unfortunately, there was a prolonged U.S. government shutdown during the data collection phase of this project. The NOAA website temporarily suspended services during the shutdown and was not providing data files. Weather data was obtained from an alternate source in order to proceed with this project.

Weather data was collected by downloading txt files with comma separated values from the following website: https://mesonet.agron.iastate.edu/request/download.phtml. One file was downloaded for each of the top 10 busiest U.S. airports. Each file contains data for all of 2016 through 2017, presented in local time for the METAR station. The data in the files was preprocessed using python. See **Appendix D** for more details about the weather data preprocessing procedure.

The sky level coverage and altitude variables are described here in more detail. The sky level altitude and coverage variables contain information about the appearance of the sky. If at least one layer of clouds, obscurations, or a combination of the two is reported, the sky level 1 coverage variable indicates the amount of sky cover at or below the lowest layer. The sky level

1 altitude variable indicates the height of the lowest layer. Sky levels 2-4 contain information about any layers above the lowest layer, in ascending order up to the first overcast layer.

A sky level 1 coverage value of VV identifies an indefinite ceiling, in which case the sky level 1 altitude indicates the vertical visibility into the indefinite ceiling. Clear skies are indicated by a sky level 1 coverage value of CLR with a missing value for the sky level 1 altitude. Clear skies can also be indicated by a sky level 1 coverage value of SKC, but none of the data points in the preprocessed weather dataset had this value.

Merged Data

The preprocessed flight and weather datasets were merged into a single dataset using python. Each scheduled departure time for a particular origin airport was matched with the nearest observation time for the corresponding weather station, as long as the observation time was either an exact match or was at most 1 hour earlier than the scheduled departure time. The rows without a successfully matched observation time were excluded from the merged dataset. The resulting preprocessed dataset was saved to a csv file, to be used in later stages of the analysis.

Treatment of Outliers

There were many outliers in the preprocessed dataset. A data point was considered an outlier if it deviated by more than 1.5 times the interquartile range (IQR) beyond the 25th or 75th percentile. One suspicious data point was identified during the initial pass of data preprocessing. This data point was investigated and excluded from the final version of the preprocessed dataset. See **Appendix D** for more details about this data point. All of the other outliers were kept for analysis.

Departure Delay Distributions

This section describes the analysis of the departure delay distributions. Departure delays are given as the difference in minutes between the actual departure time and the scheduled departure time. Negative values represent early departures.

The overall departure delay distribution was examined in detail, and trends were identified in the data. Next, the departure delay distributions were compared between airline carriers and between airports. The comparisons give an indication of the typical delays experienced by each airline carrier and each airport. These comparisons also serve to identify airline carriers and airports with the worst performance in terms of departure delays. The results of the analysis are discussed below.

Overall Departure Delay Distribution

Descriptive statistics are presented for the overall distribution of departure delays. The statistics include the mean, sample standard deviation, minimum and maximum values, as well as the 25th, 50th, and 75th percentiles.

Table 1: Descriptive statistics for the overall distribution of departure delays.

Statistic	Value (Minutes)
Mean	10.5
Sample Standard Deviation	38.9
Minimum	-234
25th Percentile	-4
50th Percentile	-1
75th Percentile	8
Maximum	2,149

The 25th, 50th, and 75th percentiles indicate that many flights depart within several minutes before or after the scheduled departure time. The mean departure delay is 10.5 minutes, which is above the 75th percentile. The minimum departure delay indicates that the earliest flight departed almost 4 hours early. The maximum departure delay is close to 36 hours, which is a large but reasonable value. The data point with the maximum departure delay has a reported Carrier Delay of 35.7 hours. The deviation of the maximum departure delay from the mean is much larger than the deviation of the minimum from the mean.

A histogram of departure delays is plotted below. Although the overall distribution of departure delays ranges from roughly -4 hours to 36 hours, relatively few departures left more than 30 minutes early or more than 4 hours late. The histogram only includes departure delays between -30 minutes and 240 minutes, inclusive, so that this region can be visualized more clearly.

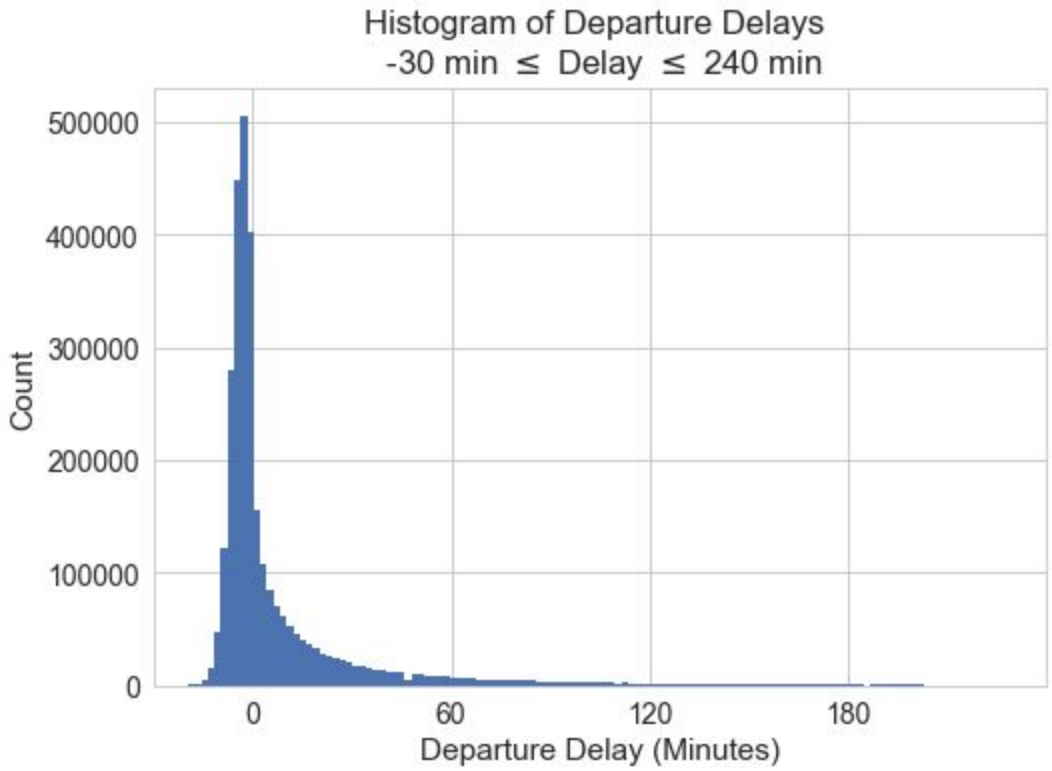

Figure 1: Histogram of departure delays. Departure delays less than -30 minutes and greater than 240 minutes are not shown in the plot. Out of the roughly 3 million total data points, 12,591 had departures more than 4 hours late, and 22 had departures more than 30 minutes early.

The histogram shows a peak at a negative value on the order of a few minutes. The distribution has a positive skew, with a tail extending towards larger delay values.

Trends in the departure delays over time are displayed in the time series plot below. The daily mean and daily median values of the overall distribution of departure delays are plotted over the entire timeframe for the years 2016 and 2017.

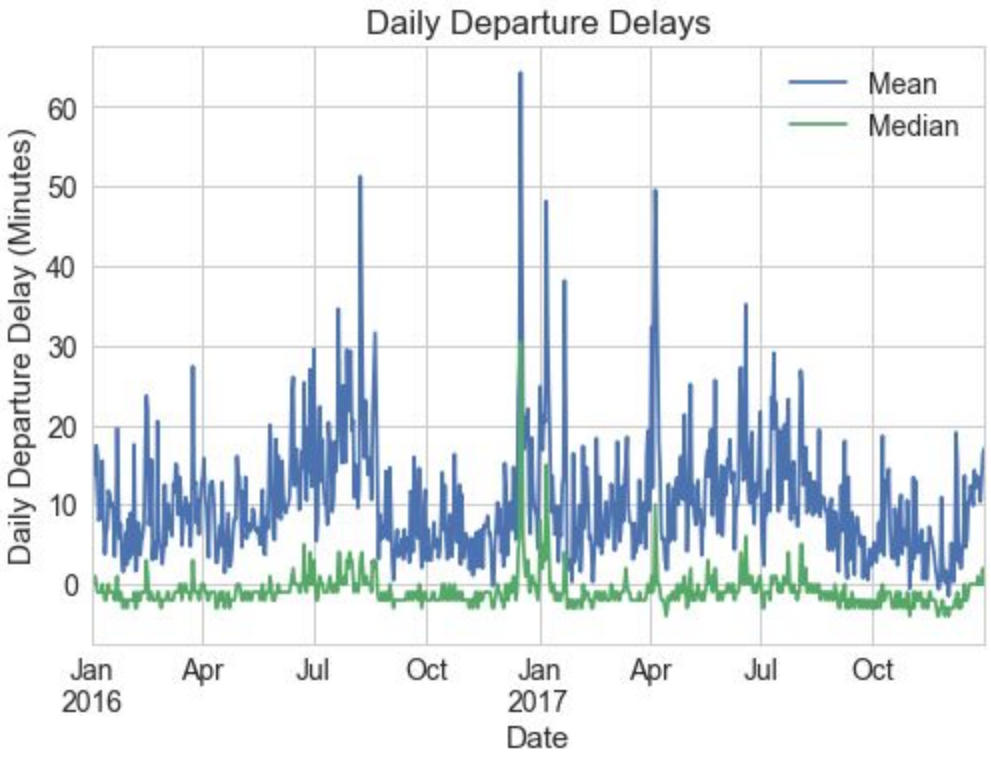

Figure 2: Time series plot with daily mean and daily median departure delays.

The time series plot indicates that daily mean departure delays are typically larger than daily median departure delays. This suggests that the daily distributions of departure delays were typically positively skewed. The plot also shows seasonal variations over long time scales as well as fluctuations over short time scales. The delays tend to be largest in the summer months and smallest in the fall months, with the exception of a few days in late 2016 and early 2017. The mean is more sensitive to extreme outliers than the median, and so the variations in the median over time are milder than for the mean.

Comparison of Departure Delays Between Airlines Carriers

The distributions of departure delays are visually compared between airline carriers using parallel box plots. The plots are shown below. A data point was considered an outlier if it deviated by more than 1.5 times the IQR above the third quartile (Q3) or below the first quartile (Q1).

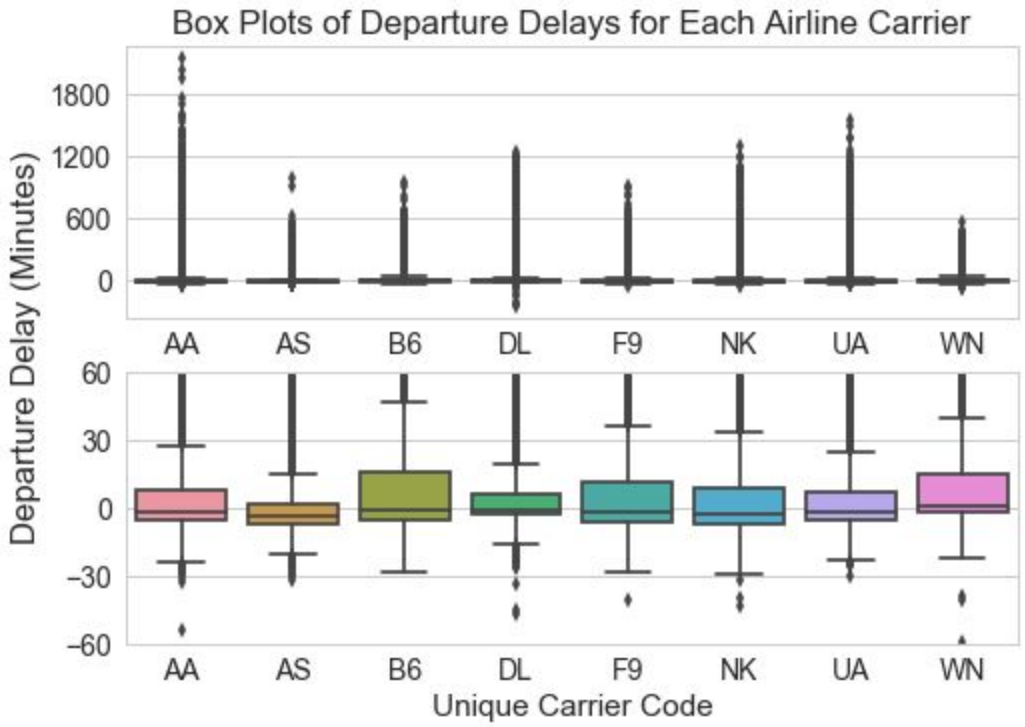

Figure 3: Box plots of departure delays for each airline carrier. The top subplot shows the full distributions for the departure delays. The values for the first quartiles, medians, and third quartiles are difficult to distinguish in the top plot, so the bottom plot provides a close up view of the distributions in the region with delays between -60 minutes and 60 minutes.

The analysis code supplemented the plots with descriptive statistics in order to obtain exact values. The plots and statistics indicate that the medians range from -4 minutes to 1 minute. The values for Q1 and Q3 are all between -7 minutes and 16 minutes, inclusive. The medians are closer to the values for Q1 than to the values for Q3. For each airline carrier, the IQR is relatively small compared to the full range of the distribution, and there are many outliers with large values. The distributions appear to be positively skewed.

The plots indicate that the distributions of departure delays are similar between the airline carriers, but there are sufficient differences for a comparison. For example, AS and DL have the smallest IQR, and B6 has the largest IQR. AA has the largest range, and WN has the smallest range. B6 has the highest mean departure delay out of all of the airline carriers, with a value of 16.3 minutes. B6 also has the highest sample standard deviation, with a value of 46.6 minutes. For these reasons, B6 is considered the worst performing airline carrier in terms of departure delays.

Comparison of Departure Delays Between Airports

The distributions of departure delays are visually compared between airports using parallel box plots. The plots are shown below.

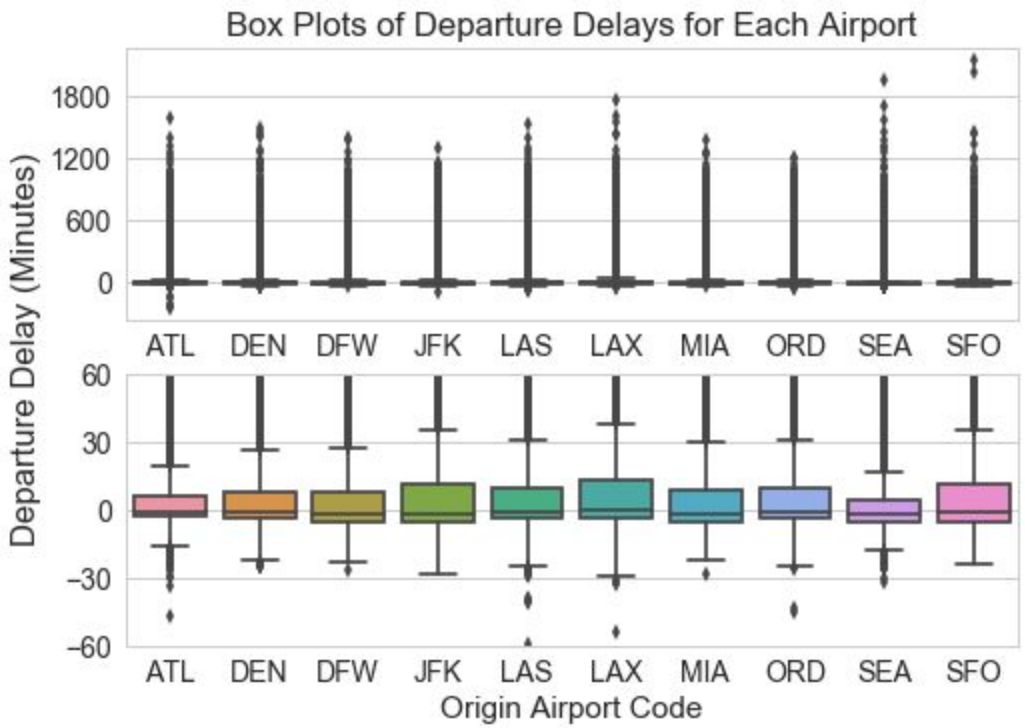

Figure 4: Box plots of departure delays for each airport. The top subplot shows the full distributions for the departure delays. The values for the first quartiles, medians, and third quartiles are difficult to distinguish in the top plot, so the bottom plot provides a close up view of the distributions in the region with delays between -60 minutes and 60 minutes.

The analysis code supplemented the plots with descriptive statistics in order to obtain exact values. The plots and statistics indicate that the medians range from -2 minutes to 0 minutes. The values for Q1 and Q3 are all between -5 minutes and 13 minutes, inclusive. The medians are closer to the values for Q1 than to the values for Q3. For each airport, the IQR is relatively small compared to the full range of the distribution, and there are many outliers with large values. The distributions appear to be positively skewed.

The plots indicate that the distributions of departure delays are similar between the airports, but there are sufficient differences for a comparison. For example, SEA and ATL have the smallest IQR, while LAX has the largest IQR. JFK has the highest mean departure delay out of all of the airports, with a value of 14.2 minutes. JFK also has the highest sample standard deviation, with a value of 49.3 minutes. For these reasons, JFK is considered the worst performing airport in terms of departure delays.

Sample Dataset Preparation

Many of the observations in the preprocessed weather dataset are not completely independent. Weather conditions tend to vary on longer timescales than the interval between METARs, and

subsequent observations from the same station are too close together to be considered independent. Similarly, many data points in the preprocessed dataset are too close in time to be considered independent.

Two sample datasets were obtained and prepared for the remainder of the analysis. For each of these sample datasets, data points from the same station are separated by 7 hours or more. Weather observations at least 7 hours apart are treated as independent for this analysis. See the following sections for more details about the two sample datasets.

The choice of the minimum separation threshold between weather observations was limited based on the amount of data available. The threshold of 7 hours was chosen to balance the length of the separation between observations with the size of the selected sample. Although a separation of 7 hours between observations may not be enough to consider them truly independent, the analysis can still draw meaningful conclusions as long as their limitations are acknowledged.

Sample Weather Dataset

A sample set of data was obtained from the preprocessed weather dataset, and a series of transformations were applied to prepare the sample dataset for analysis This sample dataset was used in portions of the analysis that compared weather conditions between airports. This sample dataset will be referred to as **DS_W** for the remainder of the document. See **Appendix E** for more details about procedure used to prepare the dataset **DS_W**.

Sample Departure Delay and Weather Variable Dataset

A sample set of data was obtained from the preprocessed dataset, and a series of filters and transformations were applied to select the data points of interest and prepare the sample dataset for analysis. This sample dataset was used in portions of the analysis that examined the relationships between departure delays and weather variables. This sample dataset will be referred to as **DS_DW** for the remainder of the document. See **Appendix F** for more details about the procedure used to prepare the dataset **DS_DW**.

Comparisons Between Weather Conditions

This section describes the analysis between weather conditions. Mean departure delays under various weather conditions were compared to the mean departure delay under clear conditions. These comparisons served to identify any weather conditions that had higher mean departure delays than clear conditions. Next, the weather conditions at each airport were compared to determine whether there is an association between airports and weather conditions. The results of the analysis are discussed below.

Weather Conditions

The weather conditions were defined for this analysis as RAIN, FOG, SNOW/ICE, CLEAR, and OTHER. The data points in the sample datasets were each assigned a single weather condition based on the value of the wxcodes variable. See FMH1 for more details about METAR present weather reporting standards.

Data points with fog, shallow fog, or mist were assigned the weather condition FOG. Data points with rain or drizzle were assigned the weather condition RAIN. Data points with snow, hail, snow pellets, ice pellets, snow grains, or ice crystals were assigned the weather condition SNOW/ICE. Data points without a reported present weather group were assigned the weather condition CLEAR. All other data points, including data points with more than one reported present weather group, were assigned the weather condition OTHER. For example, a data point with both fog and rain was assigned the weather condition OTHER. Refer to the jupyter notebooks for more details about how the weather conditions were assigned to the data points.

Comparisons of Mean Departure Delays

This portion of the analysis compared mean departure delays under the weather conditions RAIN, FOG, SNOW/ICE, and OTHER to the mean departure delay under clear conditions. These comparisons were conducted using the dataset **DS_DW**. The mean departure delay and number of sample data points for each weather condition are given in Table 2.

Table 2: The mean departure delay and count for each weather condition.

Weather Condition	Mean Departure Delay (Minutes)	Count
RAIN	-0.29	519
FOG	-1.13	349
SNOW/ICE	1.91	58
CLEAR	-0.93	17,714
OTHER	1.80	525

For each weather condition category other than CLEAR, a two-sample t-test was performed to determine whether the mean departure delay under the chosen weather condition category is

significantly greater than the mean departure delay under the weather condition CLEAR. The significance level of α = 0.05 was used for each test. The null and alternative hypotheses for each test are:

$$H_0 : \mu_{CLEAR} = \mu_{WC}$$
$$H_a : \mu_{CLEAR} < \mu_{WC}$$

Where μ_{CLEAR} is the mean departure delay under the weather condition CLEAR and μ_{WC} is the mean departure delay for the chosen weather condition category. The resulting p-values for RAIN, SNOW/ICE, and OTHER were below the significance level of 0.05. For each of these weather condition categories, the null hypothesis is rejected and the alternative hypothesis is accepted. The p-value for SNOW was not below the significance level of 0.05, so in this case the null hypothesis is not rejected.

The results of the two-sample t-tests indicate that the mean departure delay under weather conditions of RAIN, SNOW/ICE, or OTHER is significantly greater than the mean departure delay under the weather condition CLEAR. The two-sample t-test for FOG did not indicate that the mean departure delay under the weather condition FOG is significantly greater than the mean departure delay under the weather condition CLEAR.

A corresponding two-tailed test was also performed to determine whether the mean departure delay under the weather condition FOG is significantly different from the mean departure delay under the weather condition CLEAR. This test used a significance level of α = 0.05. The null and alternative hypotheses are:

$$H_0 : \mu_{CLEAR} = \mu_{FOG}$$
$$H_a : \mu_{CLEAR} \neq \mu_{FOG}$$

Where μ_{FOG} is the mean departure delay under the weather condition FOG. The results of the two-tailed test do not indicate that the mean departure delay under the weather condition FOG is significantly different from the mean departure delay under the weather condition CLEAR.

Comparison of Weather Conditions Between Airports

This portion of the analysis compared weather conditions between airports. This comparison was conducted using the dataset **DS_W**. A chi-square test of independence was performed to determine whether there was a significant association between the weather conditions and the airports. The contingency table below shows the frequency count for each weather condition at each airport.

Table 3: Contingency table with observed frequency counts for weather conditions at airports. Note that while some cells contain frequency counts of 0, the expected frequency count for each cell is greater than 5. LAX has the highest frequency counts for FOG and OTHER. SEA has the highest frequency count for RAIN. ORD has the highest frequency count for SNOW/ICE. LAS has the highest frequency count for CLEAR.

wcond station	FOG	RAIN	SNOW/ICE	CLEAR	OTHER
ATL	30	50	1	2144	58
DEN	19	13	18	2205	43
DFW	15	33	0	2168	35
JFK	35	69	17	2122	82
LAS	0	19	0	2406	17
LAX	102	17	0	2062	93
MIA	13	49	0	2205	25
ORD	65	52	30	2024	48
SEA	16	170	1	1923	88
SFO	12	57	0	2235	27

The chi-square test of independence was performed using this contingency table. A significance level of α = 0.05 was used for the test. The null and alternative hypotheses are:

$$H_0: \textit{There is no relationship between airports and weather conditions.}$$
$$H_a: \textit{There is a relationship between airports and weather conditions.}$$

The resulting p-value was well below the significance level of 0.05. The null hypothesis is rejected, and the alternative hypothesis is accepted. The hypothesis test indicates that there is a significant association between airports and weather conditions.

Weather Conditions and Airports of Interest

The comparisons between weather conditions indicated that the mean departure delays for conditions of RAIN, SNOW/ICE, or OTHER were higher than the mean departure delay under clear conditions. This analysis suggests that increased departure delays might be anticipated under conditions of RAIN, SNOW/ICE, or OTHER.

The comparison of weather conditions between airports indicated that there is a significant association between the airports and the weather conditions. This analysis suggests that some airports may experience certain weather conditions more frequently than other airports.

Relationships Between Departure Delays and Weather Variables

This section describes the relationships between departure delays and various weather variables. This portion of the analysis was conducted in two stages. In the first stage, Lasso regression was performed to identify the most important features. The results indicated that sky level 1 altitude and wind speed were the two most important features. In the second stage, the relationship between departure delay duration and sky level 1 altitude was examined in more detail by performing two conditional linear regressions. Similarly, a separate conditional linear regression was performed to examine the relationship between departure delay duration and wind speed. The results of these two stages of the analysis are discussed below.

Lasso Regression

The relationship between departure delays and various weather variables was examined by performing Lasso regression. Data for departure delays and six of the weather variables was obtained from the dataset **DS_DW** for this stage of the analysis. The selected weather variables were then standardized, and a Lasso regression was performed using the departure delays and the standardized weather variables. The resulting regression coefficients were compared in order to identify the most important features.

Model

The relationship between departure delays and the standardized weather variables is modeled by the following equation in matrix notation:

$$\mathbf{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$$

where **y** denotes a vector with the observed departure delay values, $\boldsymbol{X}$ denotes the design matrix, β denotes the parameter vector, and $\boldsymbol{\varepsilon}$ denotes the error term.

The design matrix had seven columns total. One column was populated with ones. The other six columns contained the observed values for the following standardized features:

- **Air temperature**
- **Relative humidity**
- **Wind speed**
- **Pressure altimeter**
- **Visibility**
- **Sky level 1 altitude**

The following weather variables in the dataset **DS_DW** were excluded from the Lasso regression:

- **One hour precipitation:** This variable did not have a sufficient amount of data points with positive values.
- **Wind gust:** This variable had many missing values.
- **Sky level 1 coverage:** This stage of the analysis used the sky level 1 altitude without specifying the coverage.
- **Weather condition:** This variable would provide some redundant information with respect to other variables.

The value for the tuning parameter, α, was chosen by performing two grid searches in sequence. Each grid search used 10-fold cross-validation to select the best value for α from an array of values. The first grid search covered a large region of parameter space, and the second grid search covered a smaller region of parameter space around the best value from the first grid search. The best value from the second grid search was used in the Lasso regression.

Results

The coefficient of determination of the Lasso regression was 0.010. The slope coefficients and their corresponding p-values are listed below in Table 4.

Table 4: Slope coefficients and their corresponding p-values. All of the features were standardized before performing the Lasso regression. The target variable was not standardized.

Standardized Feature	Coefficient (Minutes)	p-value
Air Temperature	0.2667	< 0.001
Relative Humidity	-0.3260	< 0.001
Wind Speed	0.4389	< 0.001
Pressure Altimeter	-0.1966	0.006
Visibility	-0.2766	< 0.001
Sky Level 1 Altitude	-0.4532	< 0.001

Discussion

The coefficient of determination of the Lasso regression indicated that 1.0% of the variance in the departure delays can be explained by the model. A more complex model could potentially result in a higher coefficient of determination, but creating a comprehensive predictive model for departure delays is beyond the scope of this analysis. The goal of the Lasso regression was to identify the most important features. The resulting p-values for the coefficients were all below even a conservative significance level of 0.01, indicating that all of the coefficients were significantly different from 0.

The slope coefficients indicate that sky level 1 altitude and wind speed were the two most important features. The next stage of this analysis examines these two weather variables in more detail. It is worth noting that none of the slope coefficients were extremely close to 0. The absolute values of the slope coefficients range from 0.1966 to 0.4532.

Conditional Linear Regressions

In this stage of the analysis, the weather variables sky level 1 altitude and wind speed were examined in more detail. Two conditional linear regressions were performed to determine the relationship between departure delay duration and sky level 1 altitude when weather variables other than sky level 1 altitude and coverage have ideal conditions. Similarly, a separate conditional linear regression was performed to determine the relationship between departure delay duration and wind speed when all other weather variables have ideal conditions.

This stage of the analysis used data from the dataset **DS_DW**. For each linear regression, the data from **DS_DW** was filtered to select the data points under the conditions of interest. See **Appendix G** for the definitions of the ideal conditions used in this stage of the analysis.

Sky Level 1 Altitude

Two linear regressions were performed to determine the relationship between departure delays and sky level 1 altitudes when all weather variables other than sky level 1 altitude and coverage are under ideal conditions. These two regressions were performed in order to compare results with and without the inclusion of several outliers that have extremely large departure delays. One regression included all values for departure delays, and the other regression only included data points with departure delays less than 1 hour. These regressions used 29.72% and 29.69%, respectively, of the data points that were used in the Lasso regression.

Both regressions also excluded each data point that had a sky level 1 coverage value of CLR. These data points were excluded because the corresponding sky level 1 altitudes originally had missing values that were replaced by custom values. There was some ambiguity about which

value to assign to the sky level 1 altitudes for the data points associated with clear skies. The results of a regression may vary depending on the chosen value. For simplicity, data points associated with clear skies were excluded from the two regressions.

A scatter plot of departure delay vs sky level 1 altitude under ideal conditions is displayed in Figure 5. The plot also includes the best fit line for each of the linear regressions.

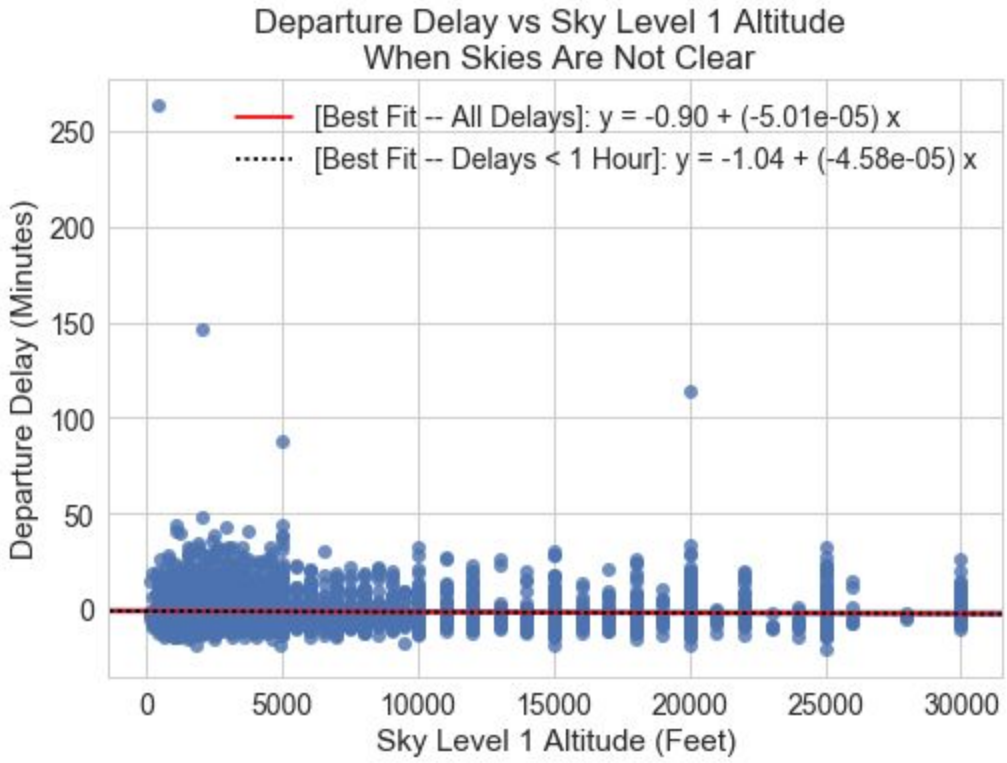

Figure 5: Scatter plot of departure delay vs sky level 1 altitude when skies are not clear. All weather variables other than sky level 1 altitude and coverage have ideal conditions. The red line indicates the best fit line for the linear regression that included all departure delays. The black dashed line indicates the best fit line for the linear regression that only included departure delays less than 1 hour.

The results of the linear regressions indicate that each of the slope coefficients is significantly different from 0. The models predict that a decrease of 30,000 feet in the sky level 1 altitude will increase departure delay durations by 1.38 to 1.50 minutes, under the appropriate conditions for the other weather variables. These results indicate that sky level 1 altitude is of practical significance. A more rigorous analysis could incorporate the altitude and coverage of all four sky levels.

Wind Speed

A linear regression was performed to determine the relationship between departure delays and wind speeds when all other weather variables have ideal conditions. This regression used

5.60% of the data points that were used in the Lasso regression. A scatter plot of departure delay vs wind speed is displayed in Figure 6. The plot also includes the best fit line for the regression.

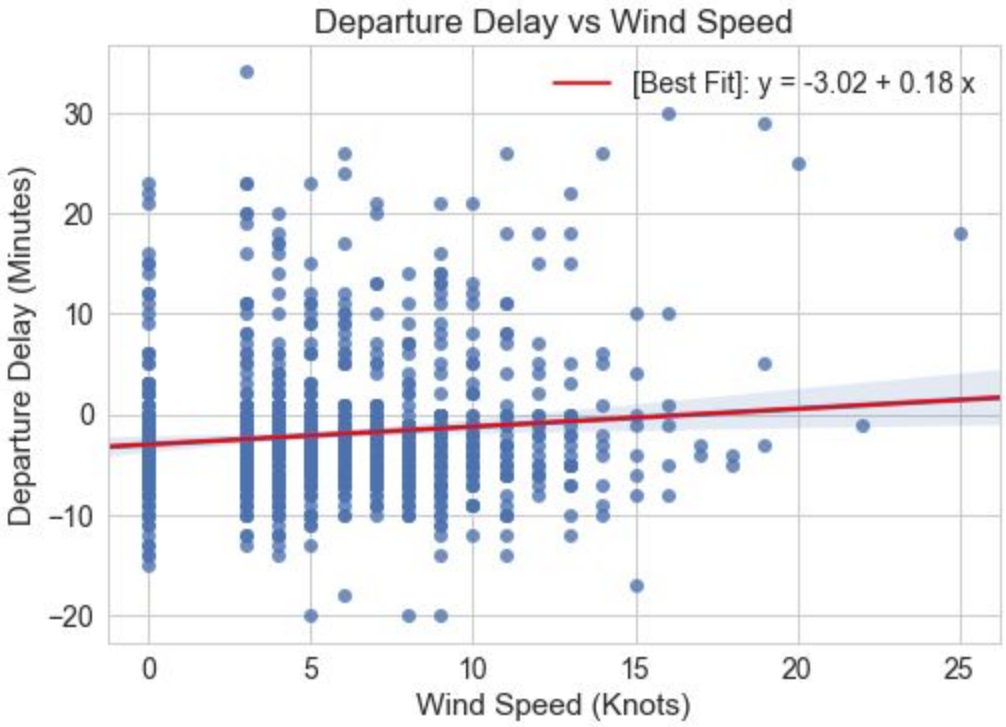

Figure 6: Scatter plot of departure delay vs wind speed. All weather variables other than wind speed have ideal conditions. The red line indicates the best fit line for the linear regression. The blue band around the red line indicates the 95% confidence interval.

The results of the linear regression indicate that the slope coefficient is significantly different from 0. The model predicts that an increase of 25 knots in the wind speed will increase departure delays by 4.4 minutes, under the appropriate conditions for the other weather variables. These results indicate that wind speed is of practical significance.

Comparison of Weather Variables Between Airports

This section describes the analysis of the sky level 1 altitudes and wind speeds at each of the airports. This portion of the analysis used data from the dataset **DS_W**. A 95% confidence interval for the mean sky level 1 altitude was obtained separately for each airport. Similarly, a 95% confidence interval for the mean wind speed was obtained separately for each airport. These confidence intervals were compared between airports in order to determine which airports had the best and worst mean values for sky level 1 altitude and wind speed.

Sky Level 1 Altitude Confidence Intervals

Figure 7 shows the 95% confidence intervals for the mean sky level 1 altitude. There was some ambiguity about which custom value to assign to the sky level 1 altitudes for the data points associated with clear skies. Different choices for the value may result in different confidence intervals. Even so, the mean sky level 1 altitude evaluated with the chosen value is a useful metric for a comparison between airports.

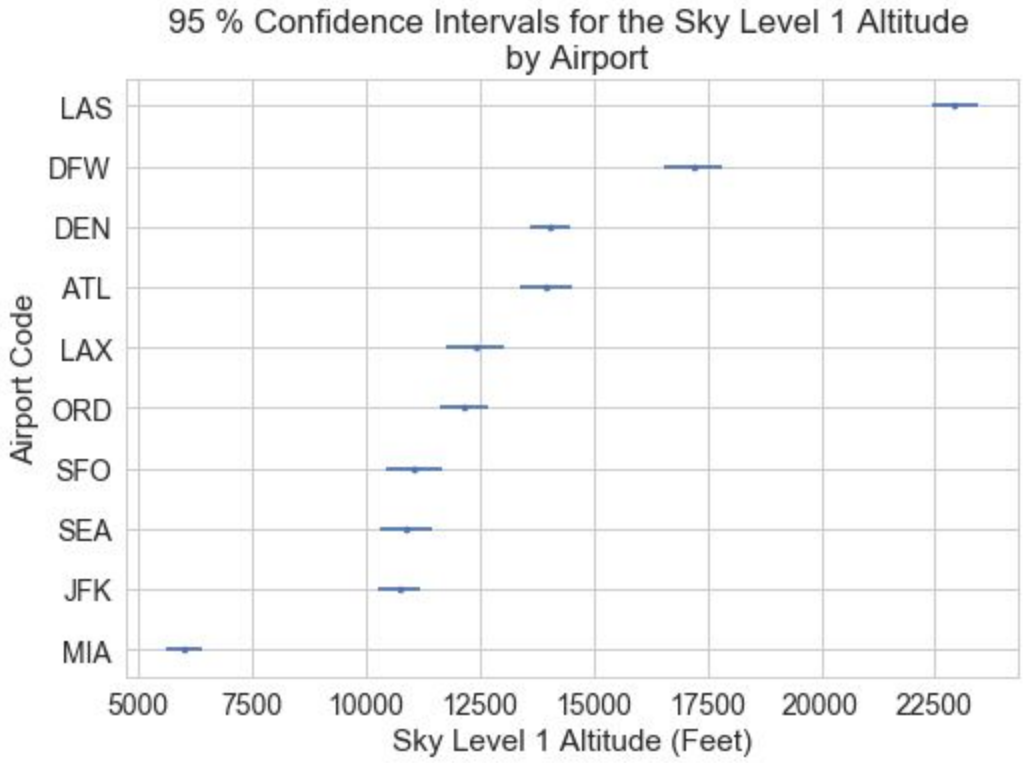

Figure 7: 95% confidence intervals for the sky level 1 altitude. The points represent the sample means. The lines represent the confidence intervals for the mean sky level 1 altitudes.

The plot shows some spread in the mean sky level 1 altitudes between airports. The confidence intervals suggest that the mean sky level 1 altitudes are all between 5,000 feet and 25,000 feet. The plot indicates that MIA has the lowest mean sky level 1 altitude of the sample dataset. The plot also indicates that LAS has the highest mean sky level 1 altitude of the sample dataset.

Wind Speed Confidence Intervals

Figure 8 shows the 95% confidence intervals for the mean wind speed. Note that wind speeds are reported as integer values in knots, and that the wind speed sensors typically have a starting speed of 3 knots.

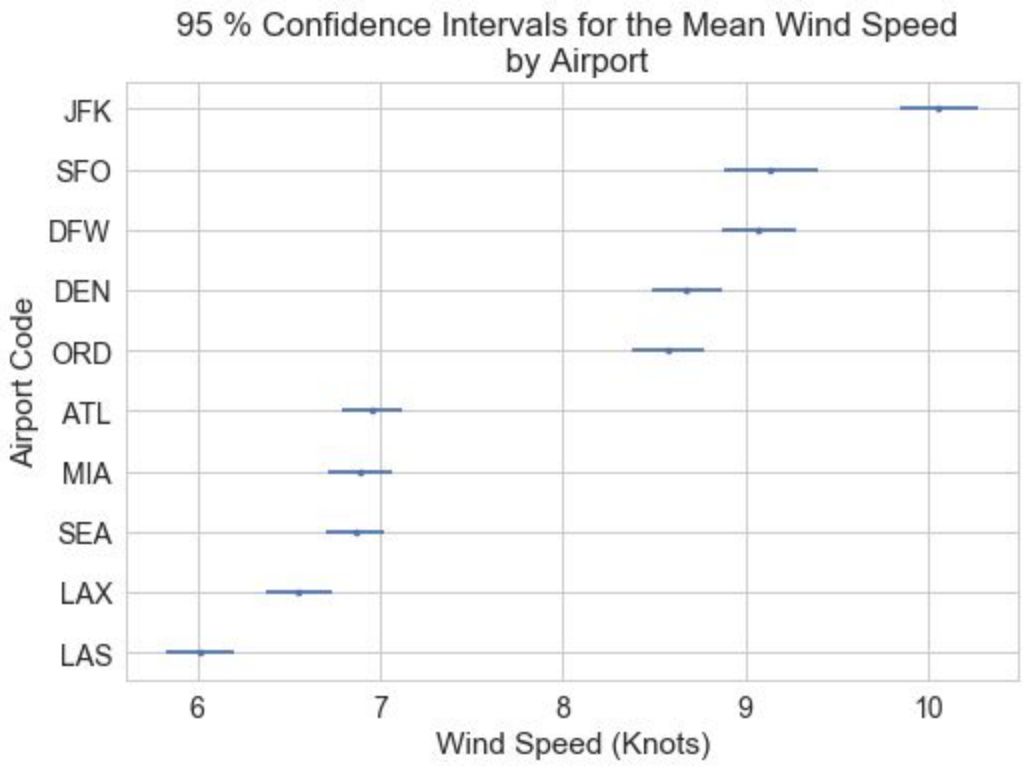

Figure 8: 95% confidence intervals for the mean wind speed. The points represent the sample means. The lines represent the confidence intervals for the mean wind speeds.

The plot shows considerable spread in the mean wind speeds between airports. The confidence intervals suggest that the mean wind speeds are all between 5 knots and 11 knots. The plot indicates that LAS has the lowest mean wind speed of the sample dataset. The plot also indicates that JFK has the highest mean wind speed of the sample dataset.

Summary

The overall distribution of departure delays indicated that many flights depart within several minutes before or after the scheduled departure time. The distribution has a positive skew, with several delays lasting hours. The daily mean and daily median departure delays showed variations over long time scales as well as fluctuations over short time scales. The departure delays tended to be longest in the summer months and shortest in the fall months, with the exception of a few days in late 2016 and early 2017.

The departure delay distributions showed similarities between airline carriers and between airports, but there were sufficient differences for comparisons to be made. JetBlue Airways Corporation was considered the worst performing airline carrier, with a mean departure delay of 16.3 minutes. John F. Kennedy International Airport was considered the worst performing airport, with a mean departure delay of 14.2 minutes. These worst performers may benefit most from investigations about ways to reduce departure delay durations.

The analysis indicated that the mean departure delays under the weather conditions RAIN, SNOW/ICE, and OTHER were significantly higher than the mean departure delay under the weather condition CLEAR. The analysis also indicated that there is a significant association between airports and weather conditions.

Sky level 1 altitude and wind speed were identified as the two most important features in the Lasso regression. The conditional linear regressions indicated that sky level 1 altitude and wind speed are of practical significance. The comparison between airports indicated that Miami International Airport had the lowest mean sky level 1 altitude and John F. Kennedy International Airport had the highest mean wind speed. These two airports could benefit from investigations about ways to reduce departure delay durations during conditions with strong winds or low sky level 1 altitudes.

Recommendations

This project identified JetBlue Airways Corporation as the worst performing airline carrier and John F. Kennedy International Airport as the worst performing airport, in terms of departure delays. It is recommended that these worst performers investigate ways to reduce departure delay durations. Other airline carriers and airports may benefit from similar investigations as well, though it is recommended that the worst performers are prioritized.

The results of this project indicated that departure delays tended to be longest in the summer months. It is recommended that airport authorities anticipate increased departure delays when scheduling departures in the summer months. It is also recommended that summer months are prioritized in investigations about the causes of the seasonal trends in departure delay duration. Additional variables could be considered, such as passenger demand and airport capacity. Further analysis could investigate ways to reduce departure delay durations in the summer months.

The results of this project indicated that the mean departure delays under the weather conditions RAIN, SNOW/ICE, and OTHER were significantly higher than the mean departure delay under the weather condition CLEAR. Further analysis could examine the broad OTHER category in more detail. At present, it is recommended that airport authorities anticipate increased departure delays when scheduling departures under the weather conditions RAIN and SNOW/ICE. Further analysis could investigate ways to reduce departure delay durations during the weather conditions RAIN and SNOW/ICE.

The results of this project indicated that sky level 1 altitude and wind speed are of practical significance. It is recommended that airport authorities anticipate increased departure delays when scheduling departures under conditions with low sky level 1 altitudes or high wind speeds. The results indicated that Miami International Airport had the lowest mean sky level 1 altitude

and John F. Kennedy International Airport had the highest mean wind speed. Further analysis could prioritize these airports in investigations about ways to reduce departure delay durations during poor conditions for sky level 1 altitude or wind speed.

The results of this project can be expanded upon. Further analysis could be performed with the goal of creating a comprehensive predictive model for departure delay durations.

Appendix A - Busiest U.S. Airports

The top 10 busiest U.S. airports by total passenger traffic in 2016 are displayed in Table 5. The list in the table is based on data compiled by Airports Council International - North America (ACI-NA). The data can be found in the ACI-NA 2016 North American Airport Traffic Summary (Passenger).

Table 5: Top 10 busiest U.S. airports by total passenger traffic in 2016

Airport Name	IATA Code	Total Passengers
Hartsfield–Jackson Atlanta International Airport	ATL	104,171,935
Los Angeles International Airport	LAX	80,921,527
O'Hare International Airport	ORD	77,960,588
Dallas/Fort Worth International Airport	DFW	65,670,697
John F. Kennedy International Airport	JFK	59,105,513
Denver International Airport	DEN	58,266,515
San Francisco International Airport	SFO	53,099,282
McCarran International Airport	LAS	47,496,614
Seattle–Tacoma International Airport	SEA	45,736,700
Miami International Airport	MIA	44,584,603

Appendix B - Largest U.S. Airline Carriers

The top 8 largest U.S. airline carriers based on passengers carried in 2017 are listed in Table 6. The list is based on data from the Bureau of Transportation Statistics tables T-100 Domestic Market (U.S. Carriers) and T-100 Market (US Carriers Only). Note that SkyWest Airlines and Republic Airline are excluded from the list because they are regional airlines that operate service for other airlines.

Table 6: Top 8 largest U.S. airline carriers based on passengers carried in 2017

Airline Carrier Name	IATA Code	Enplaned Passengers (Domestic)	Enplaned Passengers (Domestic & International)
Southwest Airlines Co.	WN	153,859,080	157,727,005
Delta Air Lines, Inc.	DL	120,928,953	145,436,827
American Airlines, Inc.	AA	116,528,317	144,919,764
United Airlines, Inc.	UA	80,554,287	107,161,566
JetBlue Airways Corporation	B6	32,395,833	40,013,934
Alaska Airlines, Inc.	AS	24,089,158	26,110,618
Spirit Airlines, Inc.	NK	21,971,273	23,812,748
Frontier Airlines, Inc.	F9	15,970,347	16,799,968

Appendix C - Flight Data Preprocessing

The flight data was preprocessed using the following procedure:

1. The input csv files were read into a list of pandas DataFrames with the following settings:
    - Several fields in the csv files contain four digit numbers that represent time values, using the format HHMM. The values in these fields were read and stored as strings, in order to preserve the format and to store each missing value as NaN.
    - The csv files indicate whether each flight was diverted using a field populated by zeroes and ones. There is a similar field for cancellations. The values in these fields were read and stored as booleans rather than integers.
2. The following steps were performed on each DataFrame individually:
    a. The DataFrame was filtered. DataFrame rows were kept for analysis if they met the following conditions:
        - The airline carrier is among the top 8 largest U.S. airline carriers.
        - The origin airport is among the top 10 busiest U.S. airports.
        - The departure delay is not null.
        - The flight date is not null.
        - The flight was not cancelled.
    b. Three new columns were created with the following departure information:
        - Scheduled departure time, represented as a Timestamp object.
        - Actual departure time, represented as a Timestamp object.
        - Wheels off time, represented as a Timestamp object.

       The DataFrame was then filtered to remove all rows with missing values in any of these columns.
    c. All of the dates and times associated with departures are displayed in local time for the origin airport. These values are kept at local time, rather than converted to UTC, in order to preserve information about the time of day. A new column was created that contains the timezone of the origin airport as a string for each data point.
3. The DataFrames were concatenated into a single DataFrame. Unnecessary columns were dropped, and the index was reset to a default RangeIndex.

Data points with missing values in any of the columns related to departure dates, times, or durations were removed from the DataFrame. Data points with missing values in the other columns were not removed, in order to preserve the departure data.

Appendix D - Weather Data Preprocessing

The weather data was preprocessed using the following procedure:

1. The input txt files were read into a list of pandas DataFrames with the following settings:
    - The observation timestamp values were read and stored as Timestamp objects.
    - The txt files indicate missing data with ‘M’. All instances of ‘M’ in the input files were converted to the appropriate missing values (‘NaN’, ‘NaT’, …).
2. The following steps were performed on each DataFrame individually:
    a. The DataFrame was filtered. DataFrame rows were kept for analysis if they met the following conditions:
        - The station is located at one of the top 10 busiest U.S. airports.
        - The observation timestamp is not null.
    b. The wind directions with values of 360 degrees were converted to 0 degrees.
    c. All of the observation timestamps are displayed in local time for the station. These values are kept at local time, rather than converted to UTC, in order to match readily with the flight times. A new column was created that contains the timezone of the station as a string for each observation.
3. A single observation from DEN for May 11, 2017 at 14:53 local time was removed. See below for a discussion about this data point.
4. The DataFrames were concatenated into a single DataFrame. An unnecessary column was dropped, and the index was reset to a default RangeIndex.

Data points with a missing value for the observation timestamp were removed from the DataFrame. Data points with missing values in the other columns were not removed, in order to preserve the number of observations.

A suspicious data point was found during the initial pass of data preprocessing. A wind speed of 70 knots occurred at DEN on May 11, 2017 at 14:53 local time. The weather data for Denver International Airport was examined around that time, and the wind speed value of 70 knots seemed suspicious. The wind speed abruptly jumped to 70 knots for a single hourly report, with smaller wind speeds before and after. There are no special notes in the raw METAR suggesting severe weather conditions, such as a hurricane. This data point was excluded from the preprocessed weather dataset.

Appendix E - Dataset **DS_W** Preparation Procedure

The sample dataset **DS_W** was obtained and prepared for analysis using the following procedure:

1. Weather observations from the preprocessed weather dataset were grouped by weather station.
   The following steps were taken for each group:
   a. The observations were separated into 7 hour time intervals.
   b. The first observation of each 7 hour interval was selected for the sample dataset.
   c. Some selected observations were more or less than 7 hours apart, depending on how many data points were missing in the 7 hour intervals. Any sample observation that occurred less than 7 hours after the previous sample observation was removed from the sample dataset.
2. A new DataFrame column was created with information about the weather condition for each data point. This column contains one of the following strings for each data point: RAIN, FOG, SNOW/ICE, CLEAR, or OTHER.
3. For data points with clear skies, as indicated by a sky level 1 coverage value of CLR, the missing values for sky level 1 altitude were each replaced by a custom value of 42,000 feet. This value was chosen after considering typical service ceilings for commercial aircraft.

Appendix F - Dataset **DS_DW** Preparation Procedure

The sample dataset **DS_DW** was obtained and prepared for analysis using the following procedure:

1. The data points with missing values or values of 0 for all of the reported delay categories were kept for analysis. All other data points were excluded.
2. Data points were grouped by weather station. The following steps were taken for each group:
    a. The data points were separated into 7 hour time intervals based on the timestamps of the weather observations.
    b. The first data point of each 7 hour interval was selected for the sample dataset.
    c. Some selected data points had observation timestamps that were more or less than 7 hours apart, depending on how many data points were missing in the 7 hour intervals. Any sample data point with an observation time less than 7 hours after the previous observation time was removed from the sample dataset.
3. A new DataFrame column was created with information about the weather condition for each data point. This column contains one of the following strings for each data point: RAIN, FOG, SNOW/ICE, CLEAR, or OTHER.
4. Weather variables that were not of interest were excluded from the sample dataset. See below for a list of weather variables that were excluded.
5. Each missing value for the wind gust is replaced with a custom value of -999 knots.
6. For data points with clear skies, as indicated by a sky level 1 coverage value of CLR, the missing values for sky level 1 altitude were each replaced by a custom value of 42,000 feet. This value was chosen after considering typical service ceilings for commercial aircraft.
7. All data points with any remaining missing values were removed from the sample dataset.

The following weather variables were excluded from the dataset **DS_DW**:

- **Dew point:** The dew point can be derived from the air temperature and the relative humidity.
- **Wind direction:** Wind direction alone is not meaningful. A more rigorous analysis could include the wind direction relative to the runway direction for each flight.
- **Sea level pressure:** Preliminary results indicated that this variable was strongly correlated with the pressure altimeter.
- **Altitudes and coverages for sky levels 2 - 4:** This analysis only includes sky level 1 for simplicity. A more rigorous analysis could include sky levels 2 through 4.

Appendix G - Ideal Conditions for Weather Variables

The ideal conditions for the weather variables were defined as follows:

- **Sky level 1 coverage** value is CLR, indicating clear skies.
- **Sky level 1 altitude** value is equal to the value described by step 6 in **Appendix F**. This indicates that skies are clear, and that the sky level 1 altitude was originally a missing value before it was assigned the value of 42,000 feet.
- **Weather condition** value is CLEAR, indicating clear conditions.
- **Visibility** value is 10 miles. This is the highest reportable value for an automated visibility report.
- **Wind Gust** value is equal to the value described by step 5 in **Appendix F**. This indicates that the wind gust was originally a missing value before it was assigned the value of -999 knots.
- **Pressure altimeter** value is between 29.7 inches and 30.3 inches, inclusive.
- **One hour precipitation** value is 0 inches.
- **Wind speed** value is less than or equal to 10 knots. According to the Beaufort scale, wind speeds in this range correspond to gentle breezes and lighter.
- **Relative humidity** value is greater than or equal to 50%.
- **Air temperature** value is between 40 degrees F and 90 degrees F, inclusive.