Capstone Project 1: Milestone Report

Jonathon Poage
Springboard Data Science Career Track
2/3/2019

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, commercial flight delays are still a fairly common occurrence. 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. This project investigates variables that influence flight delays in the United States.

Goals of Analysis

This project analyzes trends in flight departure delays for non-stop domestic flights in the United States. This analysis focuses 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 combines on-time performance data for flights in 2016 through 2017 with weather data reported by METAR stations at the airports. Note that the actual scope of the final analysis will depend on a variety of factors which will become apparent while the data is analyzed.

The analysis begins by exploring the total distribution of delays for all airports and airline carriers combined. The analysis examines characteristics of the distribution, such as the central tendency and spread. The analysis also investigates trends in departure delays over time. After the total distribution of departure delays is better understood, the analysis incorporates the weather conditions, airports, and airline carriers.

The primary goal of this analysis is to identify the relationships between departure delays and weather variables. This analysis will not provide a comprehensive predictive model for delay duration. Such a task would likely exceed the scope of the project due to the amount of variables involved and the complexity of their relationships. Even so, the analysis results could potentially be used to anticipate increased delays under certain conditions.

The secondary goal of this analysis is to identify practically significant areas for improvement for the airports and airlines. The analysis will compare the distributions of departure delays between the airports and airlines. In addition, the analysis will identify features at specific airports that may be associated with increased delays. Future analysis could focus on these practically significant areas in order to investigate ways to reduce delay durations.

Benefit to Clients

This project serves the Airport Authorities and Major Airline Carriers. The analysis aims to identify variables associated with flight delays, such as weather conditions, airport location, and airline carrier. After the significant variables and their association with delays are identified, airport authorities could potentially better anticipate delays for future flights under certain conditions, which would help them plan airport operations. More accurate scheduling of departures and arrivals would directly improve efficiency of airport operations and management of resources, which may indirectly create other benefits such as improved passenger satisfaction and increased profits.

In addition, this analysis highlights areas of improvement for the airports and airline carriers. The airline carriers and airports that have the worst performance in terms of delays are identified. The analysis also closely examines features at certain airports that are associated with longer delays. The airport authorities and airline carriers involved would benefit by receiving a summary of specific areas they may wish to investigate and improve.

Data

The analysis for this project involves data from two different datasets. One dataset contains the On-Time performance data for flights, and the other dataset contains weather data. Data wrangling was performed separately for each dataset. The cleaned data was combined by merging flight information with weather conditions based on location and time.

The following jupyter notebook contains the code used to pre-process the data for this analysis: Data Wrangling Jupyter Notebook.

Flight Data

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

The full unfiltered dataset contains millions of observations. The data was filtered to reduce the number of observations in a way that highlights the most valuable information for a streamlined analysis. This project will focus 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 analysis will focus on departure delays from the origin airport.

The flight data pre-processing involved the following steps:

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 observation.
3. The DataFrames were concatenated into a single DataFrame. Unnecessary columns were dropped, and the index was reset to a default RangeIndex.

Missing values were removed in the columns related to departure dates, times, and durations. Missing values in other columns were not removed, in order to preserve the departure data.

Weather Data

The weather information for this analysis was obtained from Meteorological Terminal Aviation Routine Weather Reports (METARs) generated by the 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.

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 analysis. 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 cleaned, transformed, and combined using python.

The weather data was pre-processed 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.

    NOTE: A single observation from DEN for May 11, 2017 at 14:53 local time was removed. See the Outliers section for a discussion.
3. The DataFrames were concatenated into a single DataFrame. Unnecessary columns were dropped, and the index was reset to a default RangeIndex.

Missing observation timestamps were removed from the DataFrame. All other missing values were kept, in order to preserve the number of observations.

Merged Data

The pre-processed flight and weather DataFrames were merged into a single DataFrame. 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 DataFrame.

The final merged DataFrame contains the pre-processed dataset. This DataFrame was written to a csv file, to be used in later stages of the analysis. The pre-processed data file is too large to store on github. The file is created, stored, and loaded locally on my machine during my analysis. The input flight and weather data files are also stored and loaded locally on my machine.

Anyone else using my data wrangling code should download a copy of the data files from my google drive. A backup copy of the pre-processed data file (version uploaded on 1/21/2019), as well as copies of the input weather and flight data files, can be downloaded here: https://drive.google.com/open?id=1PkbeFC4E2Vea67YqzfooeZacI9eKd9-P

Treatment of outliers

A suspicious data point was found during the initial pass of data cleaning. 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 jumps 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 observation was removed from the weather dataset for the final cleaned version.

Box plots were examined for the departure delay distributions at each origin airport. There were many outliers, based on the deviation of 1.5 times the interquartile range (IQR) beyond the 25th or 75th percentile, but none of the inspected outlier values looked unreasonable. All of the outliers, except for the one that was removed for having suspiciously high wind speed, are retained in the pre-processed dataset.

Analysis

The data analysis for this project is currently in progress. The first two out of three stages of the analysis have been completed. In the first stage, the data was explored using descriptive statistics and visualizations. The departure delay distribution was examined in detail, and trends were identified in the data. In the second stage, inferential statistics methods were applied to compare weather conditions between airports and draw inferences about population parameters. The results from these two stages of analysis are discussed below.

The final stage of analysis will be completed at a later time. This stage will include an in-depth analysis of the relationship between departure delays and various weather variables. The complete analysis will include regression models and Analysis of Variance (ANOVA).

Departure Delay Distribution

Descriptive statistics are presented for the 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. The departure delays are listed as the difference in minutes between the actual departure time and the scheduled departure time. Negative values represent early departures.

Table 1: Descriptive statistics for the distribution of departure delays. All values are displayed in minutes. Missing values are excluded from the statistics.

Statistic	Value (Minutes)
Mean	10.5
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 full 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 compared to the scheduled departure time. The plot displays departure delays between -30 minutes and 240 minutes, inclusive.

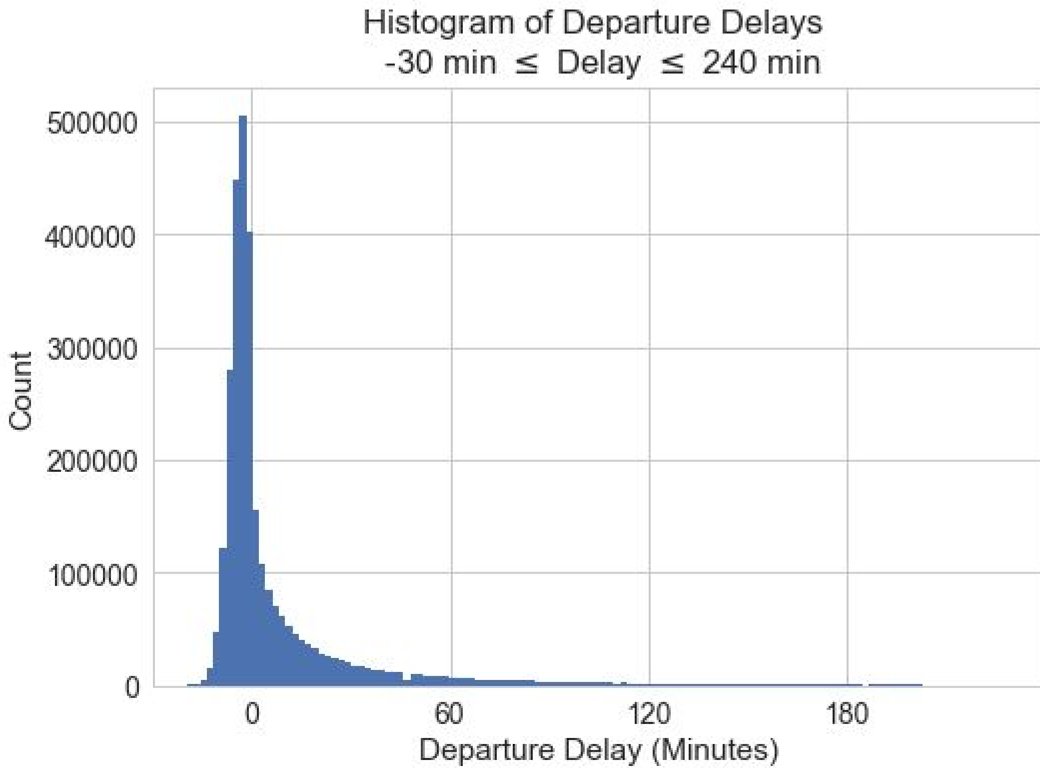

Figure 1: Histogram of the distribution of departure delays. Departure delays less than -30 minutes and greater than 240 minutes are not shown in the plot. Out of the roughly 3 million total data points, 12,591 had departures more than 4 hours late, and 22 had departures more than 30 minutes early.

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

The trends in departure delays over time are displayed in the time series plot below. The daily mean and daily median departure delays are plotted over the entire timeframe for the years 2016 and 2017. Departure delays for all airports and airline carriers are included in the daily means and medians.

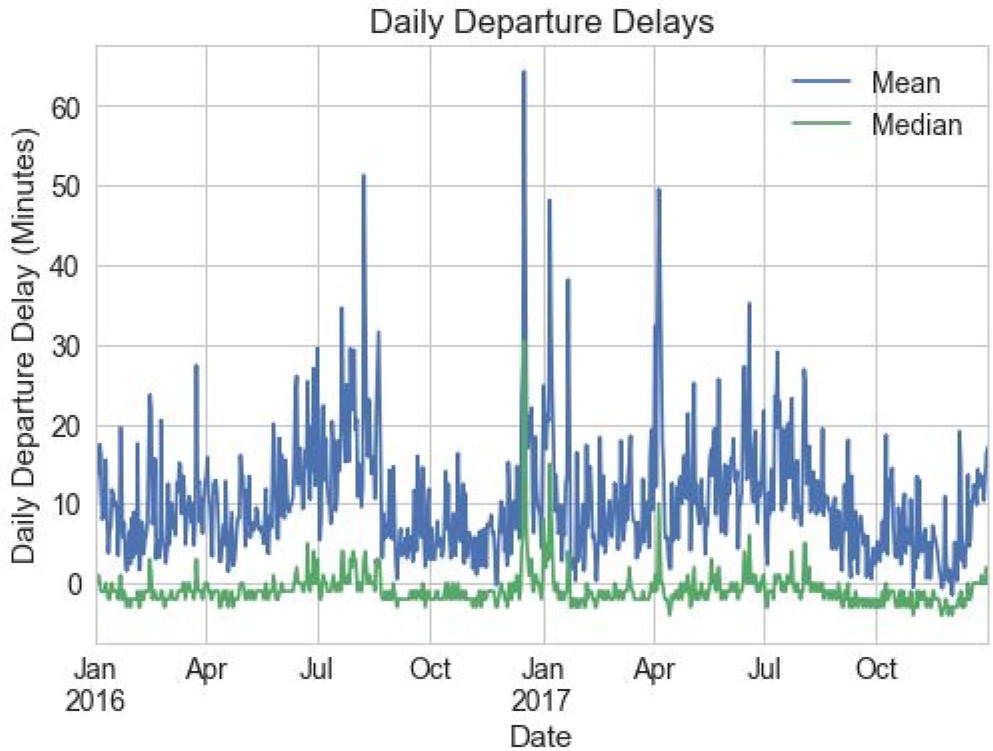

Figure 2: Time series plot with daily mean and daily median departure delays.

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 typically have positive skew. 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 and Airports

The distributions of departure delays are compared between airline carriers, and again 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 best and worst performance in terms of delays. Future analysis could focus on the worst performers and investigate ways to reduce delay durations.

Airline Carriers

The distributions of departure delays are visually compared between airline carriers using parallel box plots. The plots are shown below. Outliers are defined with respect to the interquartile range (IQR). A data point is considered an outlier if it deviates by more than 1.5 times the IQR above the third quartile (Q3) or below the first quartile (Q1).

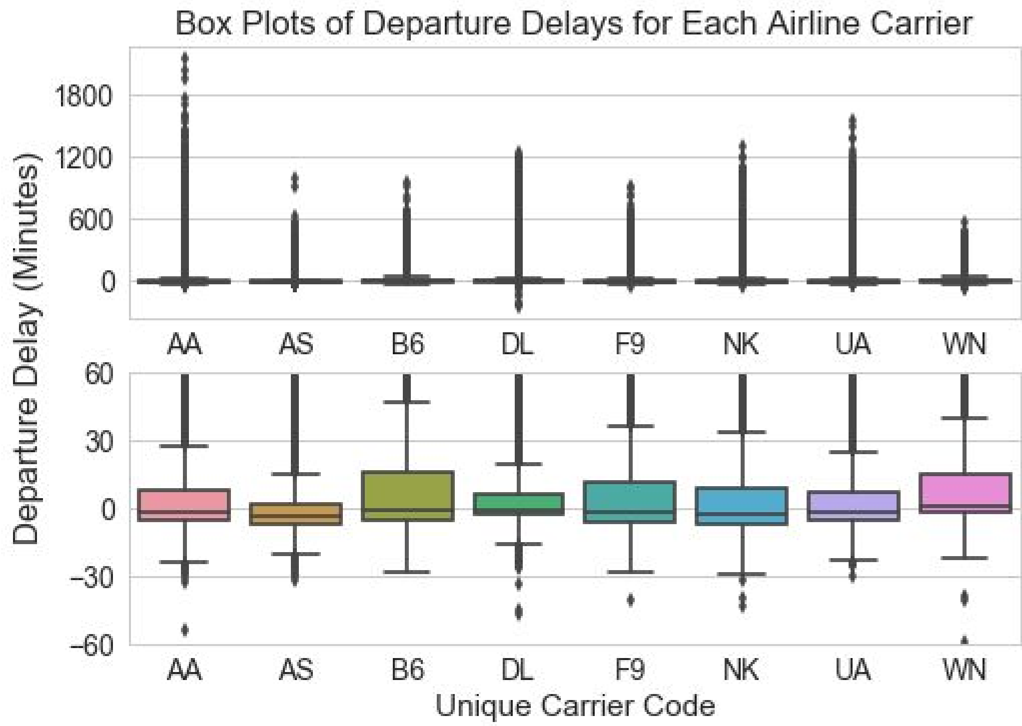

Figure 3: Box plots of the departure delays for each airline carrier. The top subplot shows the full distributions for the departure delays. The values for the first quartiles, medians, and third quartiles are difficult to distinguish in the top plot, so the bottom plot provides a close up view of the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

Several trends can be observed in the plots above. In addition, this analysis supplemented the plots with descriptive statistics, in order to obtain exact values. The plots and statistics show 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 have positive skew.

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 standard deviation, with a value of 46.6 minutes. For these reasons, B6 is considered the worst performing airline carrier in terms of delays.

Airports

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

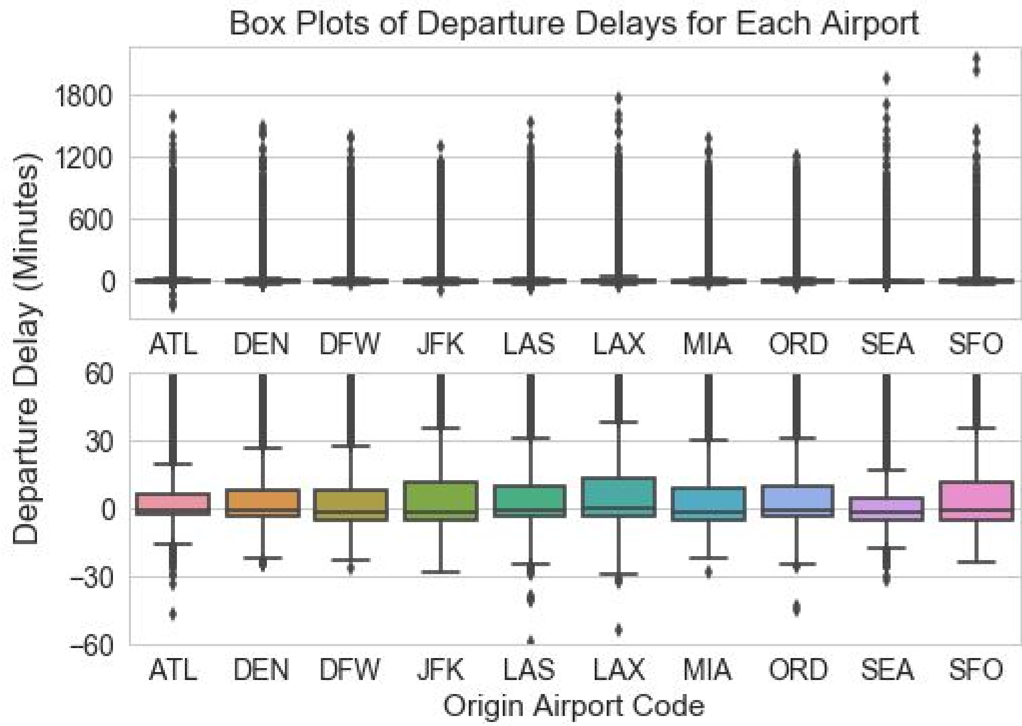

Figure 4: Box plots of the departure delays for each airport. The top subplot shows the full distributions for the departure delays. The values for the first quartiles, medians, and third quartiles are difficult to distinguish in the top plot, so the bottom plot provides a close up view of the distributions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

As with the comparison between airline carriers, this analysis supplemented the plots above with descriptive statistics. 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 have positive skew.

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 standard deviation, with a value of 49.3 minutes. For these reasons, JFK is considered the worst performing airport in terms of delays.

Comparison of Weather Conditions Between Airports

The weather conditions at each airport were compared by using inferential statistics. The comparison between airports was conducted using a sample set of data drawn from the full pre-processed weather dataset for 2016 and 2017. The methods used to draw the sample data are described below. The comparison serves to identify features at specific airports that may be associated with increased delays.

The following jupyter notebook contains the code used for this analysis to compare weather conditions between airports: [EDA Jupyter Notebook]

Sample Data

A sample set of data was drawn from the full weather dataset for this portion of the analysis. Weather conditions tend to vary on longer timescales than the interval between METARs, so the observations in the full weather dataset are too close together to be considered independent. The sample set of data was obtained by selecting observations from the full weather dataset that are separated by 7 hours or more.

The sample data was obtained for analysis using the following procedure:

1. Weather observations 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 data.
   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.

The choice of the minimum separation threshold between 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 the limitations are acknowledged. Observations at least 7 hours apart are treated as independent for this analysis.

Test of Independence

A chi-square test of independence was performed on the sample to determine whether there is a significant association between weather conditions and the airports. The weather conditions are classified as FOG, RAIN, SNOW/ICE, CLEAR, and OTHER. Each weather observation is assigned a single weather condition based on the value in the wxcodes column of the DataFrame. See the [Federal Meteorological Handbook No. 1 (FMH1)] for more details about METAR present weather reporting standards.

Observations with fog, shallow fog, or mist are classified as FOG. Observations with rain or drizzle are classified as RAIN. Observations with snow, hail, snow pellets, ice pellets, snow grains, or ice crystals are classified as SNOW/ICE. Observations without a reported present weather group are classified as CLEAR. All other observations, including observations with more than one reported present weather group, are classified as OTHER. For example, an observation with both fog and rain is classified as OTHER. Refer to the EDA jupyter notebook for more details about how the weather conditions were classified for the observations in the analysis.

The contingency table below shows the frequency count for each weather condition at each airport.

Table 2: 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.

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 $\alpha = 0.05$ 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 is 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.

The contingency table indicates that SEA has the highest frequency count for RAIN and the lowest frequency count for CLEAR out of the sample. LAX has the highest frequency count for

FOG out of the sample. ORD has the highest frequency count for SNOW/ICE out of the sample.

Confidence Intervals

This analysis compared the mean wind speed and mean visibility between airports. This comparison was conducted using the sample data set. A 95 % confidence interval for the mean wind speed was obtained separately for each airport. Similarly, a 95 % confidence interval for the mean visibility was obtained separately for each airport. The confidence intervals and sample means were compared between airports in order to determine which airports had the best and worst conditions in terms of wind speed and visibility.

Wind Speed

The plot below shows the 95 % confidence interval for the mean wind speed at each airport. 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.

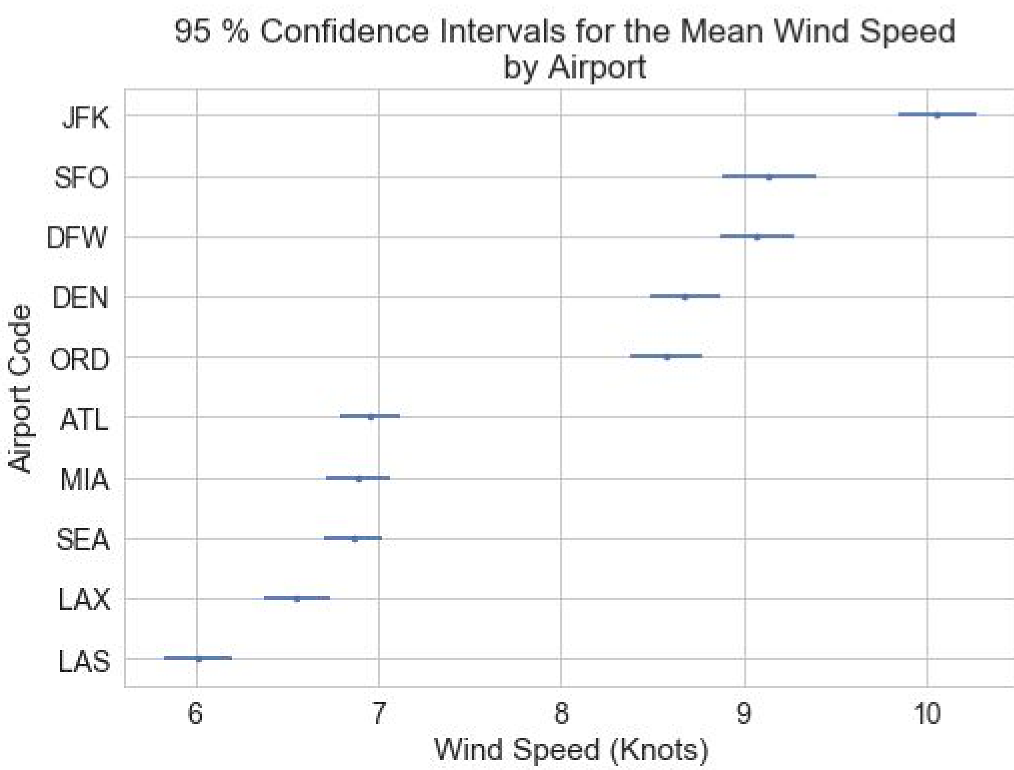

Figure 5: 95 % confidence intervals for the mean wind speed. The points represent the sample means. The lines represent the confidence intervals for the mean wind speeds.

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. The plot also indicates that JFK has the largest mean wind speed of the sample.

Visibility

The plot below shows the 95 % confidence interval for the mean visibility at each airport. Note that the reportable visibility values are not uniformly spaced. There is greater resolution for lower visibilities than for higher visibilities. Even so, the mean visibility is a useful metric for a comparison between airports. See FMH1 for more details about the visibility reporting standards.

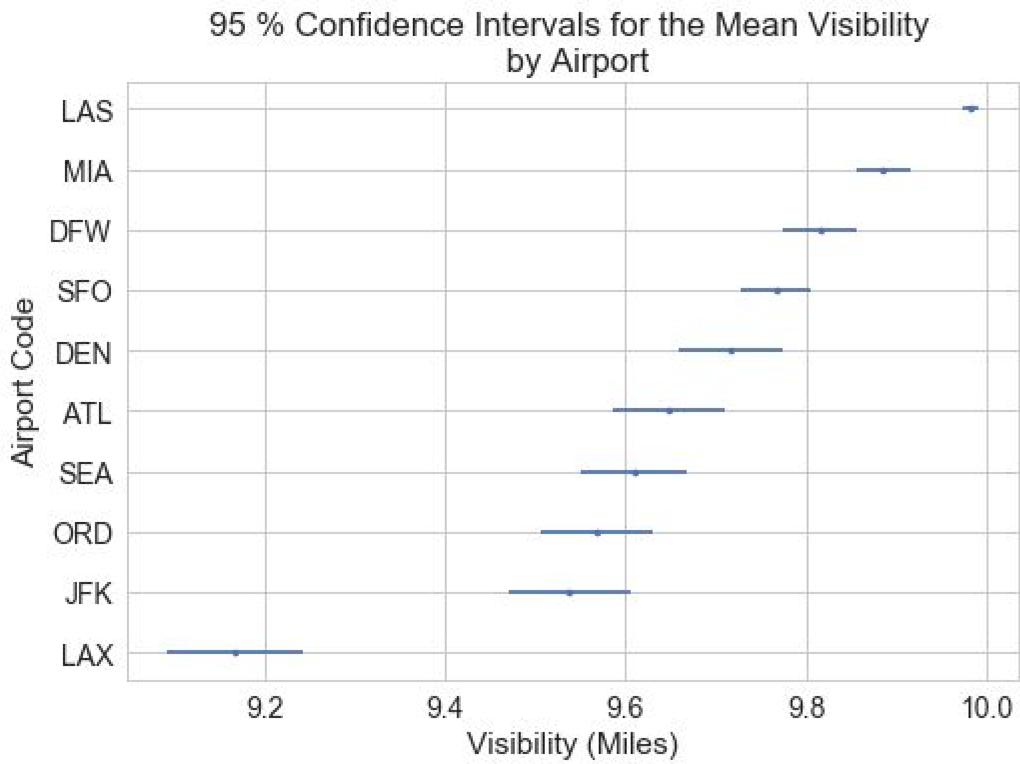

Figure 6: 95 % confidence intervals for the mean visibility. The points represent the sample means. The lines represent the confidence intervals for the mean visibilities.

The plot shows some spread in the mean visibility between airports. The confidence intervals suggest that the mean visibilities are all between 9 miles and 10 miles. The plot indicates that LAX has the lowest mean visibility out of the sample. The plot also indicates that LAS has the highest mean visibility out of the sample.

Summary

The distribution of departure delays indicates that many flights depart within several minutes before or after the scheduled departure time. The distribution shows positive skew, with several delays lasting hours. The daily mean and daily median departure delays show seasonal variations over long time scales as well as fluctuations over short time scales. The delays tend 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 show similarities between airline carriers and between airports, but there are sufficient differences for comparisons to be made. JetBlue Airways Corporation is considered the worst performing airline carrier, with a mean departure delay of 16.3 minutes. John F. Kennedy International Airport is considered the worst performing airport in terms of delays, with a mean departure delay of 14.2 minutes. These worst performers are of practical significance. They may benefit most from an investigation about ways to reduce departure delay durations.

The analysis indicates that there is a significant association between airports and weather conditions. Seattle–Tacoma International Airport has the most instances of rainy conditions and the fewest instances of clear conditions in the comparison of weather conditions between airports. Los Angeles International Airport has the most instances of foggy conditions in the comparison. O'Hare International Airport has the most instances of snowy/icy conditions in the comparison. The weather conditions at these airports may be associated with an increase in delays. These airports may benefit from further investigation about the effect of weather conditions on departure delays.

This analysis is still in progress. The next steps for the analysis are described in the following section.

Next Steps

The final stage of analysis will examine the relationship between departure delays and various numeric weather variables. The final analysis will likely include regression models and F-tests. The analysis may involve performing a linear regression between departure delays and each individual weather variable sequentially. In each case, the other weather variables will be held fixed under certain conditions in order to best isolate the variation in delays with respect to the chosen weather variable. The analysis may also involve modeling departure delay duration as a function of several weather variables, and performing a multiple regression. Although this analysis will not include a comprehensive predictive model of departure delay durations, airport authorities could potentially better anticipate delays for future flights under certain weather conditions.

Appendix A - Busiest U.S. Airports

The top 10 busiest U.S. airports by total passenger traffic in 2016 are displayed in the table below. 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 3: 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 below. 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 4: 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