Capstone Project 1: EDA Summary

Introduction

This project analyzes trends in flight departure delays for the 10 busiest U.S. airports and the 8 largest U.S. major airline carriers. The analysis combines on-time performance data for flights in 2016 through 2017 with weather data reported by METAR stations at the airports. The primary goal of the analysis is to describe the relationship between departure delays and various weather variables. The analysis also compares trends in the departure delay distributions between airports and airline carriers.

This document summarizes the statistical methods applied during the exploratory data analysis (EDA) as well as the conclusions drawn from the results.

The jupyter notebook with the EDA code can be found here.

Comparison 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 cleaned 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.

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 dataset are too close together to be considered independent. The sample set of data was obtained by selecting observations from the full dataset that are separated by a minimum threshold.

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 between observations was limited based on the amount of data available. The threshold of 7 hours was chosen to balance 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 METAR WX codes. 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 jupyter notebook for more details about how the weather conditions were classified for the observations.

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

Table 1: 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 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 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 Seattle-Tacoma International Airport has the highest frequency count for RAIN and the lowest frequency count for CLEAR out of the sample. Los Angeles International Airport has the highest frequency count for FOG out of the sample. O'Hare International Airport 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.

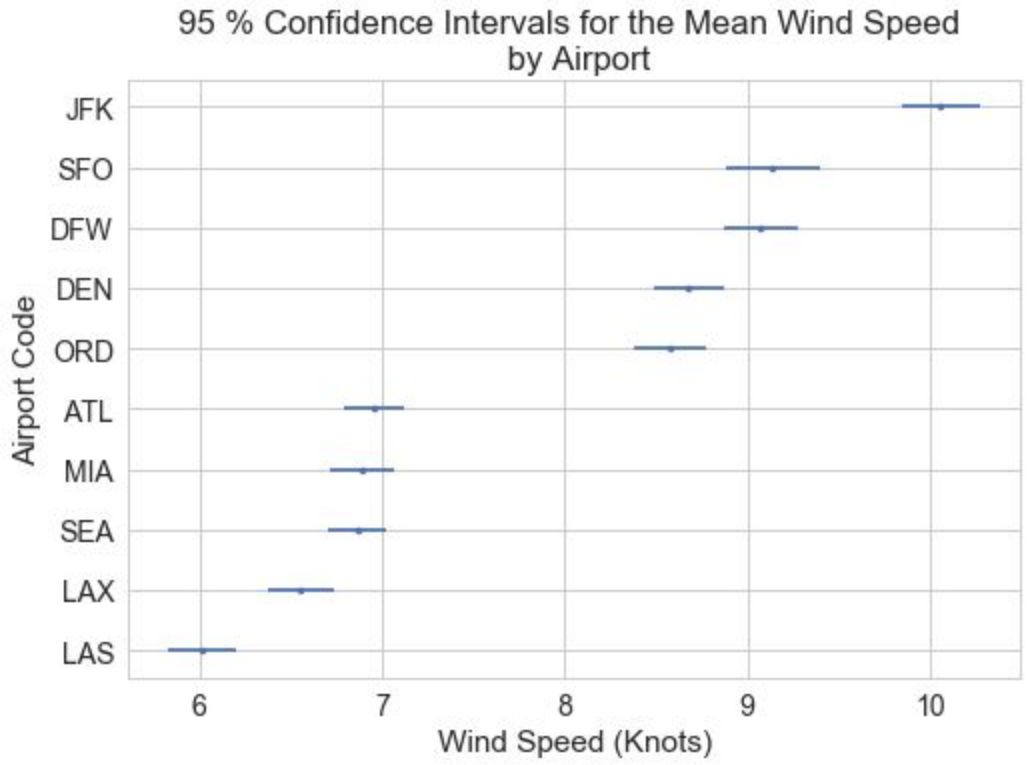

Figure 1: 95 % confidence intervals for the mean wind speed. The points represent the sample means. The lines represent the confidence intervals for the mean wind speeds.

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 McCarran International Airport has the lowest mean wind speed of the sample. The plot also indicates that John F. Kennedy International Airport 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.

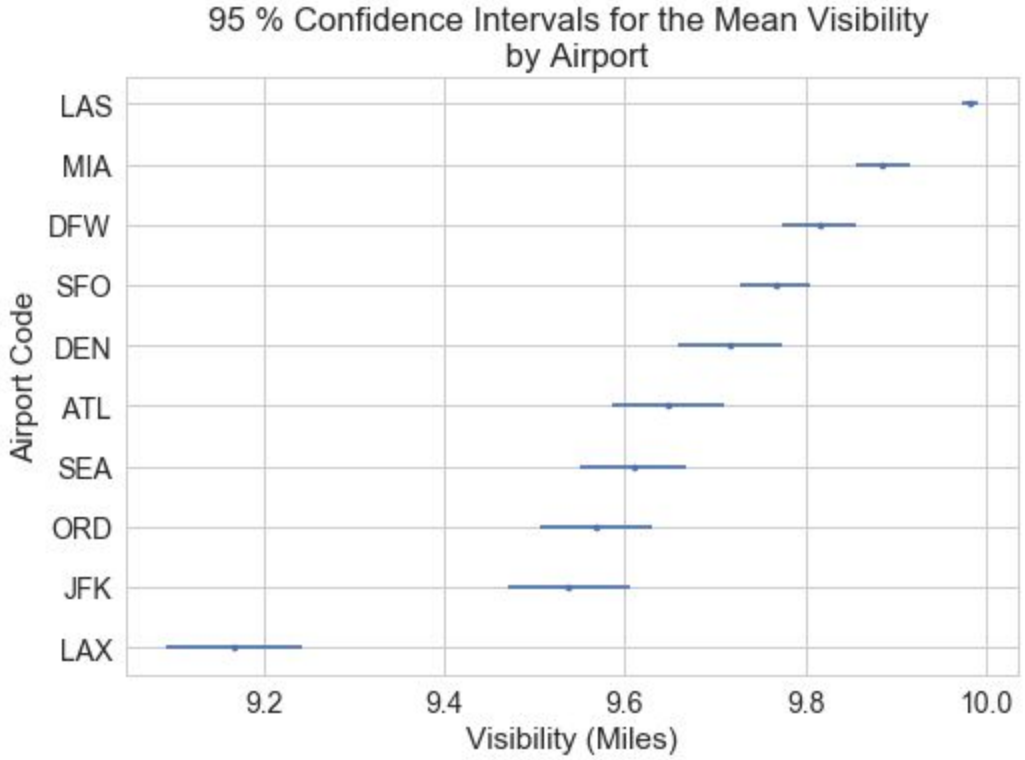

Figure 2: 95 % confidence intervals for the mean visibility. The points represent the sample means. The lines represent the confidence intervals for the mean visibilities.

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 Los Angeles International Airport has the lowest mean visibility out of the sample. The plot also indicates that McCarran International Airport has the highest mean visibility out of the sample.

Correlations Between Departure Delays and Weather Variables

The EDA analysis examined the relationship between departure delays and various numeric weather variables using the pre-processed dataset. Observations with positive values for any of the five reported delay categories, such as Security Delay or Carrier Delay, were excluded from this portion of the analysis in order to focus on the weather variables. In addition, wind direction was excluded from this analysis. A rigorous analysis would need to take into account the wind direction relative to the takeoff direction for each flight.

This analysis evaluated a correlation matrix to determine the strength of the linear relationships between departure delays and the weather variables. The Pearson correlation coefficients suggest that each weather variable is very weakly correlated with departure delays. The correlation matrix indicates that there are instances of moderate to very strong correlations

between some of the weather variables. Note that the significance of the correlation coefficients was not evaluated in this analysis.

The full analysis of the relationship between weather variables and departure delays is beyond the scope of the EDA. The final analysis will include regression models and F-tests.