Capstone Project 1: In-Depth Analysis 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 machine learning methods applied during the in-depth analysis, as well as the conclusions drawn from the results. This in-depth analysis identifies the weather conditions that are associated with increased delay durations. This analysis also shows that the relationships between departure delay duration and each of sky level 1 altitude and wind speed are practically significant.

The jupyter notebook with the in-depth analysis code can be found here.

Sample Data Preparation

A sample set of data was obtained from the full pre-processed dataset. A series of filters and transformations were applied to select the data points of interest and prepare the sample dataset for analysis.

The sample data 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 data set.
    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 with a time less than 7 hours after the previous sample observation 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. Features that were not of interest were excluded from the sample dataset. See below for a list of features that were excluded.
5. Each missing value for the wind gust speed is replaced with a sentinel 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 are each replaced by a sentinel value of 42,000 feet. This sentinel 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 features were excluded from the sample dataset:

- **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 rigorous analysis would include the wind direction relative to the runway direction for each flight.
- **Sea level pressure:** Preliminary results indicated this feature was strongly correlated with the pressure altimeter.
- **Altitudes and coverages for sky levels 2 - 4:** This analysis only includes the lowest sky level for simplicity. A more rigorous analysis would include the remaining sky levels.

This prepared sample set of data was used in all portions of the in-depth analysis, although additional filters and transformations were applied for certain portions. This sample dataset will be referred to as **DS_0** for the remainder of the document.

Sky Level Variables

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 weather dataset had this value.

See the Federal Meteorological Handbook No. 1 (FMH1) for more details about the definitions and reporting standards of the sky condition parameters. This analysis only includes the altitude and coverage for sky level 1. If multiple layers are reported, this analysis only includes information about the lowest layer. A more rigorous analysis could include information about multiple layers by including sky levels 2-4.

Comparison Between Weather Conditions

The distributions of departure delays under various weather conditions were compared by using inferential statistics. The comparison between weather conditions was conducted using the dataset **DS_0**. The comparison serves to identify which weather conditions are associated with increased departure delay durations.

Comparisons to Clear Conditions

This portion of the analysis compared mean departure delays under various weather conditions to the mean departure delay under clear conditions. The weather conditions are given by the following categories: RAIN, FOG, SNOW/ICE, CLEAR, or OTHER. Refer to the jupyter notebook for more details about how the weather conditions were classified for the data points. The mean departure delay and number of sample data points for each weather condition is given in Table 1.

Table 1: The mean departure delay and count for each weather condition.

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

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 mean 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, OTHER, and SNOW/ICE 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 is 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.

Relationship Between Departure Delays and Weather Variables

This portion of the analysis examined 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. This stage's analysis indicated that wind speed and sky level 1 altitude 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.

Important Features

In this stage of the analysis, the relationship between departure delays and various weather variables was explored by performing Lasso regression. Data for departure delays and six of the weather variables was taken from the dataset **DS_0** for this portion 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.

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_0** 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 speed:** This variable had many missing values.
- **Sky level 1 coverage:** This portion of the analysis uses the altitude of sky level 1 without specifying the coverage.
- **Weather condition:** This variable would provide some redundant information with respect to other variables.

The value for the tuning parameter, $\boldsymbol{\alpha}$, was chosen by performing two grid searches in sequence. Each grid search used 10-fold cross-validation to select the best value for $\boldsymbol{\alpha}$ 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.

The coefficient of determination of the Lasso regression was 0.010, indicating 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 and their corresponding p-values are listed below in Table 2.

Table 2: 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	0.000
Relative Humidity	-0.3260	0.000
Wind Speed	0.4389	0.000
Pressure Altimeter	-0.1966	0.006
Visibility	-0.2766	0.000
Sky Level 1 Altitude	-0.4532	0.000

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 features 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. Further analysis could examine the relationships between departure delays and the remaining four features in more detail.

Conditional Regression Analysis

In this stage of the analysis, the two most important features identified by the Lasso regression are 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 portion of the analysis uses data from the dataset **DS_0**. For each linear regression, the data from **DS_0** is filtered to select the data points under the conditions of interest.

Ideal Conditions

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 sentinel value described by step 6 in the section **Sample Data Preparation**. This indicates that skies are clear, and that the sky level 1 altitude was originally a missing value before it was assigned the sentinel value.
- Weather condition value is CLEAR, indicating clear conditions.
- Visibility value is 10 miles. This is the highest reportable value for an automated visibility report.
- Gust wind speed value is equal to the sentinel value described by step 5 in the section **Sample Data Preparation**. This indicates that the gust wind speed was originally a missing value before it was assigned the sentinel value.
- 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.

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 corresponding sky level 1 altitudes originally had missing values that were replaced by sentinel values. There was some ambiguity about which sentinel 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 sentinel value. For simplicity, data points associated with clear skies were excluded from this portion of the analysis.

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

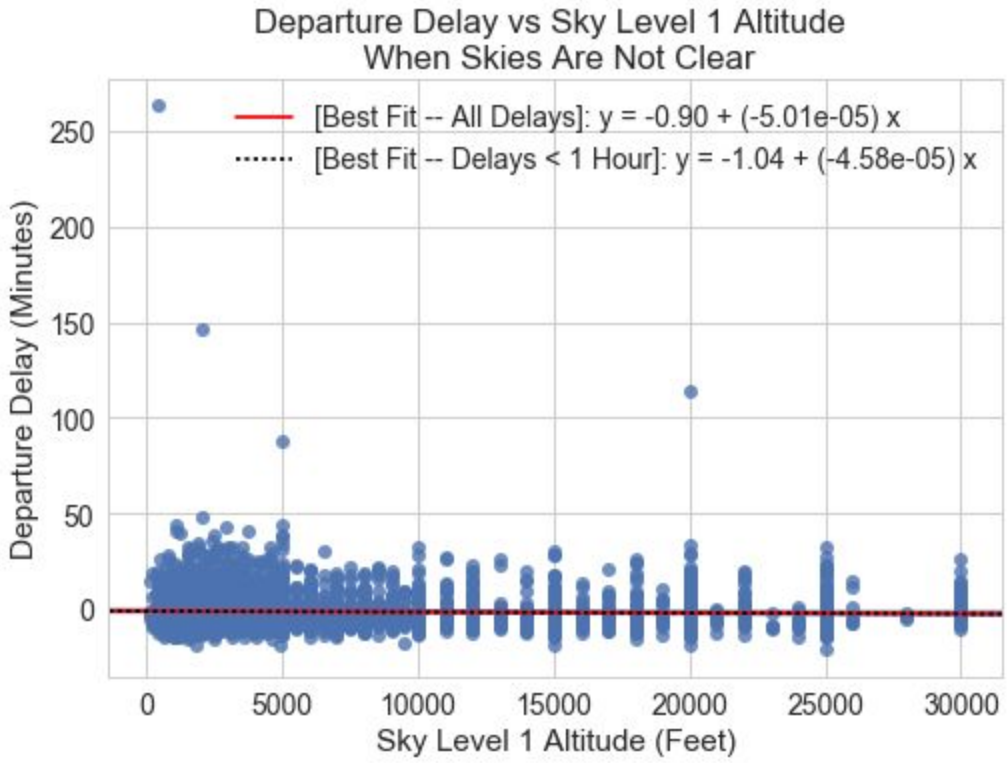

Figure 1: Scatter plot of departure delay vs sky level 1 altitude when skies are not clear. All weather variables other than sky level 1 altitude and coverage have ideal conditions. The red line indicates the best fit line for the linear regression that included all departure delays. The black dashed 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 2. The plot also includes the best fit line for the regression.

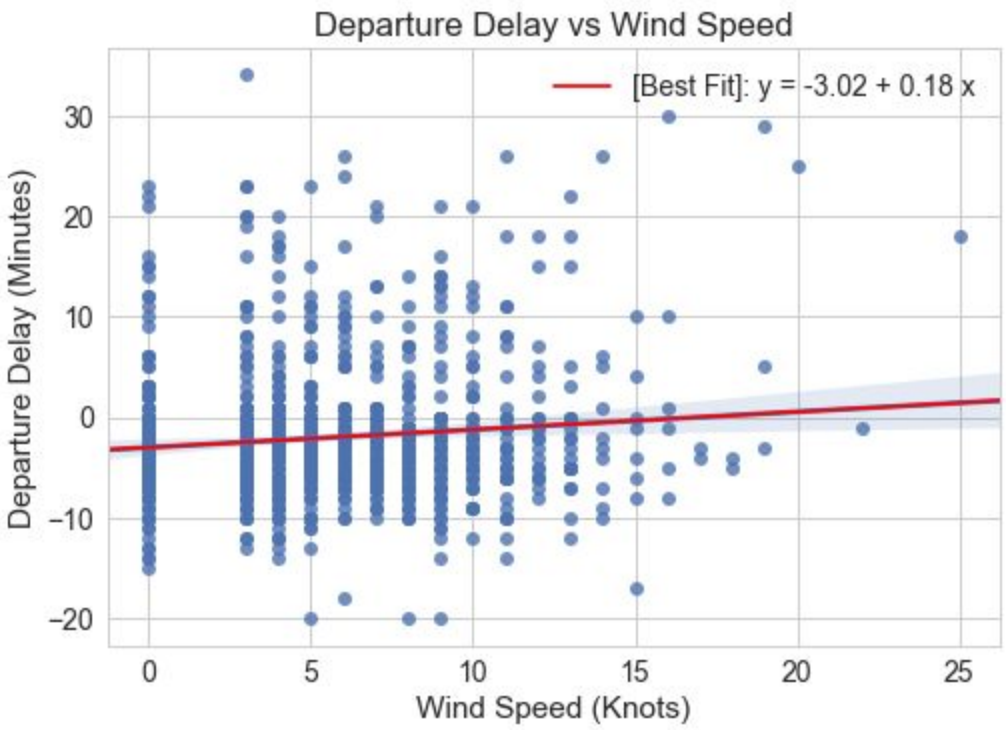

Figure 2: Scatter plot of departure delay vs wind speed. All weather variables other than wind speed have ideal conditions. The red line indicates the best fit line for the linear regression. The blue band around the red line indicates the 95 % confidence interval.

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.

Conclusion

The in-depth analysis examined the relationships between departure delays and various weather variables. The comparison between weather conditions indicated that weather conditions of RAIN, SNOW/ICE, and OTHER are associated with higher departure delays than CLEAR conditions. Sky level 1 altitude and wind speed were identified as the two most important features for the Lasso regression. The conditional linear regressions indicated that wind speed and sky level 1 altitude are of practical significance.

A full predictive model of departure delays is beyond the scope of this analysis. Future analysis could explore the relationship between departure delays and weather variables in more detail.