**Summary:**

We performed data analysis to determine the components driving average airport departure delay, finding that there are five principal components that explain our data. We derived this from Principal Component Analysis. Further, we found that when using the five principal components along with FAA regions and type of airport, we could do a principal component regression that had an R-squared of 96% on our test data, which is excellent. Finally, out of the FAA regions in the data, four of them were neutral with regard to departure delays.

|  |  |
| --- | --- |
| **Region** | Impact on Delays (minutes |
| **FAA REGION\_AAL** | 0 |
| **FAA REGION\_ACE** | 0.000000 |
| **FAA REGION\_AEA** | 0.145053 |
| **FAA REGION\_AGL** | -0.248698 |
| **FAA REGION\_ANE** | 0.000000 |
| **FAA REGION\_ANM** | -0.377643 |
| **FAA REGION\_ASO** | -0.000000 |
| **FAA REGION\_ASW** | -0.021190 |
| **FAA REGION\_AWP** | 0.066559 |

For further analysis, we would want to do the following:

1. Analysis components impacting arrival delays;
2. Create new features such as airport concentration in a region - i.e. we could hypothesis that more airports in a region or within a certain distance of one another as measured by Longitude/Latitude may mean that arrival delays are less of an issue because of multiple options for travellers and for diversions if necessary. For example, a person flying to John Wayne could be redirected to a nearby airport after 9 pm if their departure time was delayed enough for their arrival to violate curfew there.
3. Perform clustering on the principal components to see what groupings we discover, and then potentially link that to actual departure delay times.

**Problem Statement:**

Given the data the FAA has on arrivals and departures, and operational data for 74 major airports, we asked if there are features that are associated enough with delays to test changes to them to reduce delays. For the purpose of this study, we focused on departure delays as quantified by average airport departure delay, which is operational data for the year by airport of all delayed departures from it.

Our goal was to go from all the features the FAA data contains and reduce the dimensions of our data to the most descriptive components as measured by explained variance. We performed Principal Component Analysis to do this.

**Risks**

1) The airports with missing data for all periods do not impact the airports around them. I.e. The Gary/Chicago international airport, which is a joint civil/military public airport, does not impact delays at other Chicago airports. Or the Teterboro airport does not impact airports around it, or the Van Nuys airport does not impact airports around it (these are the airports for which we have operations data but not cancellation data).

2) The airports with missing data for certain years (2004 to 2006) also are not much of an issue (SMF - Sacramento) and PSP.

3) Given that this data is over multiple years, during which airlines have merged, gone bankrupt, or faced disruption from low-cost carriers (among other factors), it is possible this industry data has an impact on delays.

Other data that could impact delays includes population growth in the areas these airports serve. This is data that is not in our dataset. Yes, it is data that the FAA can not control, but what if it impacts delays? Then it would matter for describing what impacts delays and highlighting something the FAA can not directly control.

4) Technological factors such as the FAA having old equipment that could impact flight times and paths are not directly handled in this data.

In short, certain supply and demand factors do not appear to be in the data.

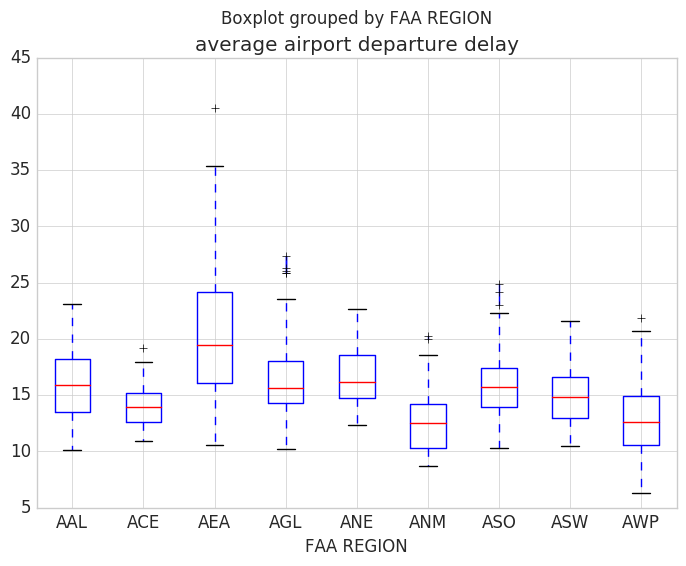
**Assumptions**

1) The Data collection and creation is good enough.

2) The features in the data are potentially explanatory.

**Exploratory Data Analysis**

*Summary Statistics*



**Principal Component Analysis**

After finding five principal components, we examined the new feature space. Below are some plots of the new feature space, which shows a high amount of potential clustering.





**Regression Output**

Since one of the purposes of PCA is to reduce dimensions and find the most explanatory data, we also ran a linear regression on the components and FAA regional data to see if it truly was explanatory of average airport departure delay. Our output is below, including model coefficients:

R2 score for Lasso CV Regression is: 0.960834652868  
Intercept is: 15.7351893293

|  |  |
| --- | --- |
| **Feature** | **Coefficient** |
| **FAA REGION\_ACE** | -0.000000 |
| **FAA REGION\_AEA** | 0.145053 |
| **FAA REGION\_AGL** | -0.248698 |
| **FAA REGION\_ANE** | 0.000000 |
| **FAA REGION\_ANM** | -0.377643 |
| **FAA REGION\_ASO** | -0.000000 |
| **FAA REGION\_ASW** | -0.021190 |
| **FAA REGION\_AWP** | 0.066559 |
| **AP Type\_Public Use** | 0.000000 |
| **PC\_1** | 1.696914 |
| **PC\_2** | 0.081837 |
| **PC\_3** | -0.067868 |
| **PC\_4** | 0.281914 |
| **PC\_5** | -0.853355 |

While interesting, we would caveat that a number of the operational data components probably ultimately rely on supply/demand factors that we could not directly observe (i.e. # of travelers, flights that day, etc.)

**Results**

Our model and principal component analysis suggest that the FAA should be less concerned about regions when it comes to departure delays, despite the difference in delays in the boxplot above. Instead, the FAA should look at the five principal components discovered via PCA.