Project 7: Airport Delays + Cluster Analysis

Description

This week, we learned about clustering and how to prepare your data for clustering, as well as useful ways for storing and accessing your data. Now, we're going to apply each of the skills, as well as skills you've learned in previous courses, to successfully store, understand, prepare, and model your data using unsupervised learning methods.

You've been hired by the FAA as a consultant to analyze the operations of major airports around the country. The FAA wants to cut down on delays nationwide, and the most important part of this task is understanding the characteristics and groupings of airports based on a dataset of departure and operational delays.

- A certain degree of delay is expected in airport operations, however the FAA is noticing significant delays with certain airports
- When a flight takes off, it's departure delay is recorded in minutes, as well as operational data relating to this delay
- At the end of the year, this data is averaged out for each airport. Your datasets have these averaged for a 10 year range between 2004 and 2014
- Over this 10 year range, some delay times have not improved or have worsened.

Point: Your task is to understand the distribution, characteristics, and components of individual airports operations that are leading to these delays.

Project Summary

In this project, we're going to be using three different datasets related to airport operations. These include a dataset detailing the arrival and departure delays/diversions by airport, a dataset that provides metrics related to arrivals and departures for each airport, and a dataset that details names and characteristics for each airport code.

You will help the FAA:

- Organize and store their data so that they can easily understand it after your consulting work is done
- Mine and refine the data to uncover its basic attributes and characteristics
- Use your skills with PCA to uncover the core components of operations related to delays.

When you've finished your analysis, the FAA would like a report detailing your findings, with recommendations as to which airports and operational characteristics they should target to decrease delays.

Here are some questions to keep in mind:

- What operational factors are most directly correlated to delays?
- Take a look at airports groupings - are there any relationships by region? Size?

Requirements

- Complete all of the tasks below:
- Write a problem statement & describe the goals of your study to be included in the final report
- Create a database to store your data; the FAA has dictated that you use PostgreSQL
- Conduct Exploratory Data Analysis to understand the attributes of our data; include your EDA findings in your final report to the FAA
- Mine & refine your data

- Conduct a PCA to discover the principal components behind departure delays
- Present the results of your findings in a formal report to the FAA, including the problem statement, summary statistics of the takeoff delays and operational delays, your PCA analysis detailing the principal components related to relays, and a case study on one specific airport that best illustrates your findings to FAA officials.
- Plot your PCA analysis on a 3-dimensional graph
- Create a blog post from your notebook of at least 500 words (and 1-2 graphics!) that describes your project and includes your analysis, findings, and recommendations. Link to it in your Jupyter notebook.