Highlights

- Found interesting trends in airline cancellations
  - Variance in cancellations
  - Patterns by day of week
- Found proper data and columns to use from the Bureau of Transportation Statistics

Review Progress

~~Data Gathering - Pull Data from Website - Download a year worth of data files (1)~~

~~Data Gathering - Pull Data from Website - Append data to one another (1)~~

~~Data Gathering - Pull Data from Website - Upload into Python (1)~~

~~Data Gathering - Exploratory Data Analysis - Check data for missing values or outliers (1)~~

~~Gathering - Exploratory Data Analysis - Check for any multicollinearity in the data I'll be using (1)~~

~~Data Gathering - Exploratory Data Analysis - Look for any additional factors for possible inclusion into the model (1)~~

~~Data Modeling - Data Cleaning - Reformat any columns as necessary (1)~~

~~Data Modeling - Data Cleaning - Merge necessary information (1)~~

Data Modeling - Feature Engineering - Calculate new columns necessary for the app (2)

Data Modeling - Model Building - Run basic linear model for cancellation detection (1)

Data Modeling - Model Building - Look at more complex models for cancellation detection (2)

Data Modeling - Model Building - Run basic linear model for delay time prediction (1)

Data Modeling - Model Building - Look at more complex models for delay time prediction (2)

Data Modeling - Model Building - Compare misclassification rates, and pick the best cancellation detection modelone (1)

Data Modeling - Model Building - Compare delay time models, and pick the best one (1)

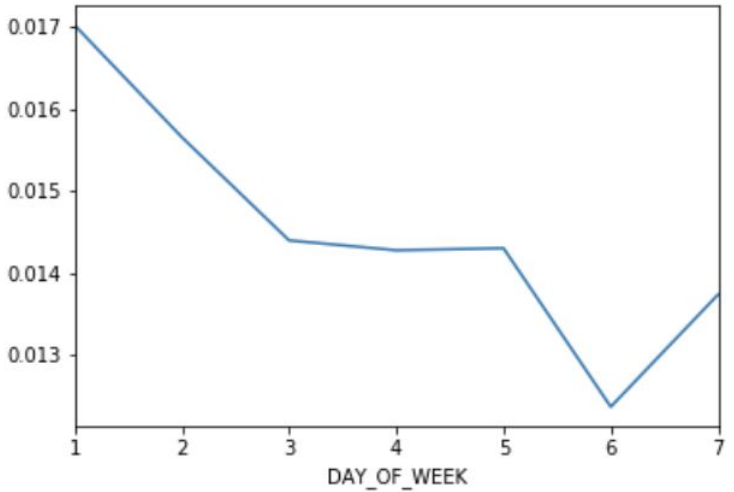
Building Web App - Build the Web App - Build an interface to allow people to input their flight information (8)

Building Web App - Build the Web App - Make App output delay/cancellation probability, predict delay time (8)

Demo/Analysis

Found that the worst states in terms of cancellation percentage seemed to be in new england

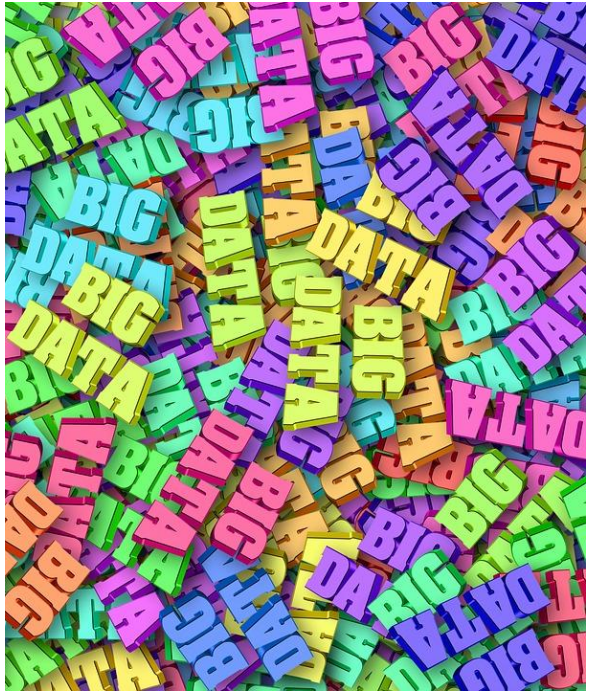
ORIGIN_STATE_ABR	
ME	0.041351
VT	0.039478
NH	0.034201
NY	0.033039
NJ	0.031770
SC	0.030994
RI	0.029520
WV	0.029301
NC	0.027501
VA	0.026408
MA	0.026077
CT	0.025931
PA	0.024964
OH	0.022120
IL	0.022024
MD	0.021707



Surprisingly, the weekends were the days with the lowest percentage of cancellations. Monday had the most.

Lessons Learned

- Only store as much data as necessary, and nothing more
  - I was too liberal with downloading columns from the BOTS
- This tool will be much more useful for large airports
  - Small airport data just too variable
- Normalization could be necessary for unusual weather years/months



Recommendations

In the following sprint, the following tasks should be done:

- Create linear and possibly more complex models for predicting cancellations
- Explore opportunities for feature creation to improve the model
- Create a user interface to input data