**Notes from the Meeting**

what policies govern how the delay times are assigned to  these categories?

·         arrival delay will have more variables affecting it than departure delay

·         arrival delay = departure delay + extended flight time

·         for categorical inputs, boxplot or violin plot

·         for clustering

o   just look at delayed flights

o   can’t use too many dimensions

§  use PCA

§  or just pick a few important variables

§  use sci-kit learning tutorials

·         how to categorize routes

o   use some other characteristic of the flight

§  flight capacity

§  flight price

§  flight direction

§  geographical direction

§  flight distance

·         how to do clustering with categorical variables?

o   look at histograms (in more than 2 dimensions?)

**\*\*Include date, time, your name and a detailed description of your progress.**

* Only use data from the last 3 years
* Split test and training sets randomly
  + all data ⇒ dataframe, subset ⇒ csv for training and test
* put into GitHub folder
  + ipython notebook
  + test and train data csvs

**Annie 11.23 ~2pm**

* I discovered that manually downloading data is very inefficient because you have to select a single month at a time.
* I’m trying to use the instructions from this website <http://www.tableausoftware.com/public/blog/2013/08/data-scraping-python-2098> to try to scrape the data.
* figured out how to download! now will work on assembling all the data into dataframe

Annie and Evan 11.28 ~9pm

* Manually gathered all the data from 9/2011 to 9/2014.
* Decided that it would be best to subset the data (over 5 GB!)
* Will decide very soon with Robert which popular routes to pick so that we can reduce our data set size, to hopefully under 100 MB to fit on github
* Next step after deciding routes
  + Download all the data from Drive, unzip it all,
  + Extract from each month the routes we have chosen
  + Save those to separate csv files (e.g. jfk-lax.csv, bos-jfk.csv…etc.)

Robert, 11/30

Separated out test data from training data. Test data: 10/2013 to 9/2014 (size 205961) Training data: 10/2011 to 9/2013 (size 426112). Saved these in the Github folder.

Doing analysis in iPython file called Exploratory1. Note the first line of code! Each person needs their own file path in order to use the data…

Table of Contents of analysis done so far: (Copy and pasted from Exploratory1 notebook)

'Things to do/Table of Contents:

(Will a lot of this depend on getting rid of weather/unpredictable long delays first and then studying remaining relationship?)

How does total flight time relate to arrival time delay?

How does total distance of flight relate to arrival time delay?

How does time of day for departure relate to arrival time delay?

How does time of the year relate to arrival time delay?

How does which day of the week relate to arrival time delay?

How does the particular airline carrier relate to arrival time delay?

Different cities and analyzing differences? Consider top 10 airports to and from Boston.

Top 10, taken from http://en.wikipedia.org/wiki/Logan\_International\_Airport:

ORD Chicago, DCA Washington, ATL Atlanta, PHL Philadelphia, SFO San Francisco, CLT Charlotte, JFK New York, LAX Los Angeles,

BWI Baltimore, EWR Newark

-Need to subset the data frame for this.

-Compare cities to one another

-Compare flight paths from Boston to the city to those from other city to Boston

-Consider specific flight paths with many of the variables considered earlier on (maybe later...)

-Generate some metric or rating system to give different city routes scores for how delay-free they are (maybe later...)

Build REGRESSION models (linear at first) or other form of quantiative interpretation. Once with just data probably and once with

hopefully better modification after getting rid of weather/unpredictable long delays.

Next: Clustering using clustering techniques. 2 variables at once. k-nearest neighbors?

SVM clustering too? Need to make categories for different amount of delay. Then one path to doing machine learning is using this

directly on the test data and see which cluster a flight ends up in.

regression with many vars to generate predictor? Just like baseball PSET.

Other ideas (copied from Proposal doc:

Other metrics? Such as Actual total flight in-air delay vs. delayed departure, or even taxi time.

types of delays? security, weather...is the labelling too rare/un-uniform and sort of unreliable/arbitrary perhaps?

Other countries.

Relationship of cardinal direction of travel, or airport geographic location, or airport size

Plane type/passenger capacity.

'''