**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.)