11/11/16 Documentation

What I did: (Lauren)

* Subset events table by year in MySQL so that queries are faster and don’t use too much disk space (before: 34 mill events per year)
* Got random sample of each subset containing 5% of events from each year (About 1.2 million observations per year)
* Made large CSV file combining all of these with unix cat command
* For future reference:

Macintosh HD:Users:Lauren:Desktop:Screen Shot 2016-11-13 at 3.03.29 PM.png

Goal: Want to plot feature relationships and distributions of features to see if event and inquiry tables can be modeled similarly.

11/13 Documentation

What I did: (Lauren)

* Tried same query for inquiry table
* Realized that there are many fewer observations – not necessary to get random sample for these (10% of data for 2013-2014 is only 43k rows)
* Get data from each year separately and merge CSV files
* Go into R project and clean the data
* For events table: in image ID feature, there are a lot of null values and the ones that aren’t null are not meaningful, they just increment with time. Because of this, I replaced null values with 0 and any other value with 1.
* Replace “NULL” entries of inquiry table search queries as “NONE” 🡪 this is what null means for this column
* Turn data from BOTH tables into character data so that factors can be replaced easily (i.e. replace “NULLS” with 0’s, otherwise 1’s).

11/14 Documentation

What I did: (Lauren)

* Created vector from search queries in inquiry table
* Got top 100 search queries with their frequencies
* Created vector with corresponding terms and frequencies
* Made word cloud with top 100 frequent words to see how frequencies compare to each other

11/17 Documentation

What I did: (Lauren)

* Converted character data from inquiry table to date format so that I can create a new column for duration (still in progress, issues with data format conversion)
* Replace null controlIds in events table with 0’s 🡪 assuming that is what null implies, easier to plot this way
* Convert events table image ID to numeric data, diagnosis ID, and control ID
* Plot imageID vs. diagnosis ID and find it not very useful
* Mosaic plot of these is more useful
* 2-ways tables of active view ID and image ID, made proportion tables to see how to predict these.

11/21 Documentation

What I did: (Lauren)

* Reformatted data from inquiry table for analysis c
  + changed null values to none for ddx module id column
  + inquiry type id from character data to numeric
  + diagnosis lookup ids null 🡪 none
  + convert application answer ID, outcome answer ID to numeric data
  + Remove NA values (2) from coercing inquiry type ID
  + Start time and end time to POSIXct date format
* Split duration for inquiry into sub intervals from min to max (0 to 24000) and get frequency for each sub interval of size 250
* Plot histogram and barplot of this distribution
* Make pairwise plot between other features to examine relationships
* Found 🡪 correlation of 1 between application answer ID and diagnosis lookup ID
  + Remove one of them from the data frame: doesn’t matter which (choose application answer ID) 🡪 nevermind this was a silly mistake
* However: found correlation between application answer ID and outcome answer ID is 0.74 so that’s pretty significant too

11/24 Documentation

What I did (Lauren):

* Plot density of start time by day to see if there are clusters/ spikes
* Plot start time by year to see if some years are more dense than others (have more usage, sessions started) because of trend noticed from day plot
* Make boxplot of starttimes, axis labels aren’t very useful I will try to change that.
* Look at table when different conditions are true, i.e. ddx module id and diagnosis lookup id are “NONE”
  + Discovery: when application answer ID is not 0 outcome answer ID is also not 0
  + Only 2877 rows of data where this is true as compared to total 3410777 total rows
  + Something else interesting: one of ddx module ID or diagnosis lookup id is always NONE and only one row where those are both true
* Used ggplot library to try to get a sense for distributions of one variable against another… some more effective than others