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).

Jean:

* queried the join between the ClientSession table and Contract tables to find the contractTypeId (1-8). The contractTypeId equal to 1 and 2 were large, so I limited them to 20%. The other values were not an issue.
  + haven’t done anything with this data yet
* queried the entire ClientSession table. This table has about 5 million rows.
  + in R, columns we removed columns that weren’t necessary. We now have columns for productId, agent, sessionStart, sessionEnd, currentLicenseCount, imageCount, diagnosisCount
  + separated data based on mobile users and PC users using a string comparison
  + added a “duration” column by subtracting start time from end time
  + plotting duration is proving to be difficult due to the extreme variation in session length

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.

11/19/16

Jean:

* loaded 8 csv files into new R project (one for each contractTypeId 1-8)

11/20/16

Jean:

* created a barplot of the number of rows in each contract type
* created a histogram of the logarithm of image count for contract type 2 (may not be necessary for determining differences between types)
* experimented with plotting duration by month, but wasn’t able to incorporate our data effectively