VisualDX capstone project 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 copy command

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