**Revised Capstone Proposals**

1.  2015 Traffic Fatalities provided by NHTSA.  Similar to my earlier proposal regarding predicting traffic fatalities.  This could be used to predict fatalities by time of day (binary early or late, for example) and time of year (Winter, Spring, Summer or Fall—factors could be created based on month of year given), and other correlative factors. (This dataset, unlike the ones I was looking at before, appears to be in better shape.)

<https://www.kaggle.com/nhtsa/2015-traffic-fatalities>

2.  Movies Dataset + Academy Awards dataset.  These two datasets properly combined could be used to predict Academy Award winners/nominees. (Combination may be made easier since there are apparently numerical identifiers associated with each movie).  (Possible predictors are budget of a typical movie, ratings, length, star quality of the actors/directors, etc.)  I would probably have to do more data cleaning here and remove a lot of the movies in the movie data set to just get the ones that were nominated for/won Oscars and to get rid of unnecessary/extraneous columns.

<https://www.kaggle.com/rounakbanik/the-movies-dataset/data>

<https://www.kaggle.com/theacademy/academy-awards>

3.  Credit Card Fraud prediction.  I found this idea of interest, but am not too sure about this particular dataset. The description notes that the data is imbalanced. The description also says that a confusion matrix won't work and recommends Area Under the Precision-Recall Curve. Although this is not necessarily my first choice of topic, I would be curious if this dataset could still be made workable.

<https://www.kaggle.com/dalpozz/creditcardfraud>