

## Capstone Project 1: Data Wrangling

The first piece of wrangling I did was the most important and was motivated by two problems. The first was that the initial datasets I downloaded were too large to load into the memory of my aging Macbook pro. The second was that the file that contained the target values for a model had much fewer observations than the dataset. So there were many rows in the dataset for which there were no targets to use for training or testing a model.

I was able to find one solution which solved both problems. I did this by using SQLAlchemy to create a local database chunk by chunk (something my computer was able to handle) and then using SQL commands to filter and create a new csv that filtered out all rows in the dataset which had no target values. The result was small enough that I could load it into memory without needed to lose useful data. I did this in the notebook entitled SQL-dataset.

The next steps were pretty normal. I dropped columns and rows that were mostly empty. Converted categorical variables to integers and datetime variables to datetime.

I also noticed that some of the columns I dropped for having mostly empty values were not actually empty, but rather had null values representing 0 or FALSE entries. So I brought them back in and cleaned them up.