Capstone Project 1

Data Wrangling

The dataset that I am working on describes the produced water characteristics from onshore wells drilled in Texas. It has been obtained from the United States Geological Survey (USGS) National Produced Waters Database v2.3. There are about 190 parameters that each water sample yields, and just for the state of Texas, there are 19388 unique entries. Importing the files to my local directories, I started by choosing the parameters of interest, and identified the important ones to 27, namely: 'LATITUDE', 'LONGITUDE', 'API', 'BASIN', 'STATE', 'DATECOMP', 'DATESAMPLE', 'FORMATION', 'PERIOD', 'DEPTHUPPER', 'DEPTHLOWER', 'DEPTHWELL', 'LITHOLOGY', 'SG', 'SPGRAV', 'PH', 'TDSUSGS', 'TDS', 'HCO3', 'Ca', 'Cl', 'KNa', 'Mg', 'Na', 'SO4', 'H2S', 'cull_chargeb'.

The particular cleaning steps that I performed were to remove the missing values from columns (a lot of the columns were empty), using the .dropna() method on the Pandas dataframe along the columns axis: *df_drop = df.dropna(axis= 1, how='all')*. Next I filtered the rest of the columns by calling out the above columns and storing them in another dataframe. There were dates and times parameters that needed to be converted into datetime parameters, and so I implemented those using *pd.to_datetime(df.DATECOMP)*.

The method that I used for handling missing latitude and longitude values was by corroborating well names with their longitude and latitude tags. Even though there are 19388 entries, the number of unique latitude and longitudes are 5308 and 5255, respectively. This may be due to some wells with repeated time series data, or the fact that multiple wells exist in the same location, in the form of a separate wellhead, or the same wellhead but a divergent wellpath. The approach then was to go search for wells with similar latitude and longitude values in a ‘for’ loop, and create a mini-dataframe with these similar values. If there were missing values in the latitude and longitude columns, I would check the API values (unique well ID) and the well names in that mini-dataframe. If the two API values and well names were the same, then I would add in their lat/lon coordinates in the missing columns.

I also had missing formation upper depth and lower depth values for many data points. This gives the formation depths from where the well produces water and oil. I added those

missing points in using a similar technique as described above. Given a particular well name and well API number, I would create mini-dataframes corresponding to these values. For missing depths, I would copy the depths from the same well names and API numbers.

There were no outliers from the data set, and I am moving forward with the exploratory data analysis.