**Capstone 1: Exploratory Data Analysis**

The dataset obtained from the USGS website on produced water characteristics of wells between 1920 and 2010, was first cleaned and wrangled for exploratory data analysis. The Jupyter notebook for the project is available at <https://github.com/shubacca/Produced-Waters/tree/master/Produced%20Waters>

The particular techniques used to explore the data were as follows:

1. Visualization of locations of the wells using the latitude/longitude data and Tableau: This data was color-coded according to the basins.Accordingly, the major plays are Amarillo, Anadarko, Permian, East Texas and Gulf Coast. These basins were drilled and their produced water characteristics studied between 1920 and 2010. (It is important to point out that these are data points, and do not necessary mean the number of wells. One well can have multiple data points due to data being collected over time.)
2. Individual histograms creation for each of the numerical variables: It is interesting to note that out of all the basins studied, the upper depths and pH form near-normal distribution profiles, with population means of 6120 ft and 6.95, respectively.
3. Heat-map and pair-wise scatter plot generation describing the correlations between these numerical variables: The positive correlations with respect to the upper depths, lower depths (of the formations) and depths of the wells are expected. High positive correlations were also found between TDS and calcium, chlorine, potassium plus sodium, and just sodium, mostly because inclusion of these elements creates the TDS values. Magnesium and calcium don't contribute to TDS as much, which suggest that these elements are found in the insoluble phases coming out of the wells.

It is interesting to note that the bicarbonate has a weak negative correlation with TDS, suggesting that it was found in an insoluble form. Also interesting to note is that the specific gravity of the water increases with an increase in TDS content. Chlorine was mostly found to be high in positive correlation with both potassium and sodium, suggesting it was found in aqueous soluble form as KCl or NaCl. Both sodium and chlorine have strong positive correlations with TDS and specific gravity of the produced waters. Also interesting to note is that salt content does not necessarily correlate with the depth of the wells drilled.

1. TDS histogram distribution and boxplot by basin: The distributions show that Permian Basin has a higher TDS content and spread than the other basins. The Gulf coast basin has the least amount of TDS and the least spread. It will be interesting to note if the TDS values decrease over time in each of these basins.
2. Aggregate stats (mean and standard deviation) for TDS, depths, calcium, chlorine, sodium, potassium-sodium, magnesium: The basin with the highest TDS content is Southern Oklahoma, followed by Fort Worth and then Palo Duro. The standard deviation of TDS in the Permian basin is about 65% of the mean, and that the highest TDS spreads are noticeable for Anadarko, Permian and Fort Worth. The most hard waters (higher Ca and Mg content) are found in the Southern Oklahoma, Fort Worth and Palo Duro basins, and these waters respectively have the highest Na and Cl concentrations as well. It is also interesting to note that Permian has the highest variability in terms of depths of wells drilled, followed by Gulf coast, East Texas, Anadarko and then Amarillo.
3. Depths histograms distribution by basin: lots of variability was observed.
4. Bootstrapped tests on TDS data per basin: Results were plotted, and one particular basin (Permian) was taken for a hypothesis test, whether the pre-oil shale revolution Permian TDS values were significantly different from post-oil shale revolution Permian TDS values. Both z-score and p-value were calculated, with values of 116.77 and 0.0001, respectively. At this z-score, the p-value is very low, and hence the null hypothesis can be rejected safely, and said that the TDS did in fact increase after the shale oil revolution. This can be attributed to the drilling of more horizontal wells with more minerals and salts seeping in. Addition of frac fluids can also be a major cause of an increase in the TDS and salt content of the produced waters.
5. Time-series analyses looking at rates of change in TDS, and other elemental concentrations: Box plots and aggregate tables were created by basin for each of rate of change of TDS, Ca, Cl, KNa, Mg and Na. Based on these plots and table, many inferences can be made. The TDS seems to increase for Amarillo Arch and Southern Oklahoma basins at about 50.8 mg/L/day and 1.5 mg/L/day, respectively. Other basins including Permian, Palo Duro, Ouachita Thrust and Anadarko all decrease over time in TDS concentrations. Fort Worth has a very rapid decline in calcium concentrations, but that could be an outlier. The Ouachita Thrust basin saw an increase in calcium and chlorine concentrations, while an overall decrease in TDS concentrations. This may be due to presence of insoluble calcium chloride, or other insoluble forms of both calcium and chlorine. Sodium and chlorine otherwise mimicked TDS patterns quite closely across basins.