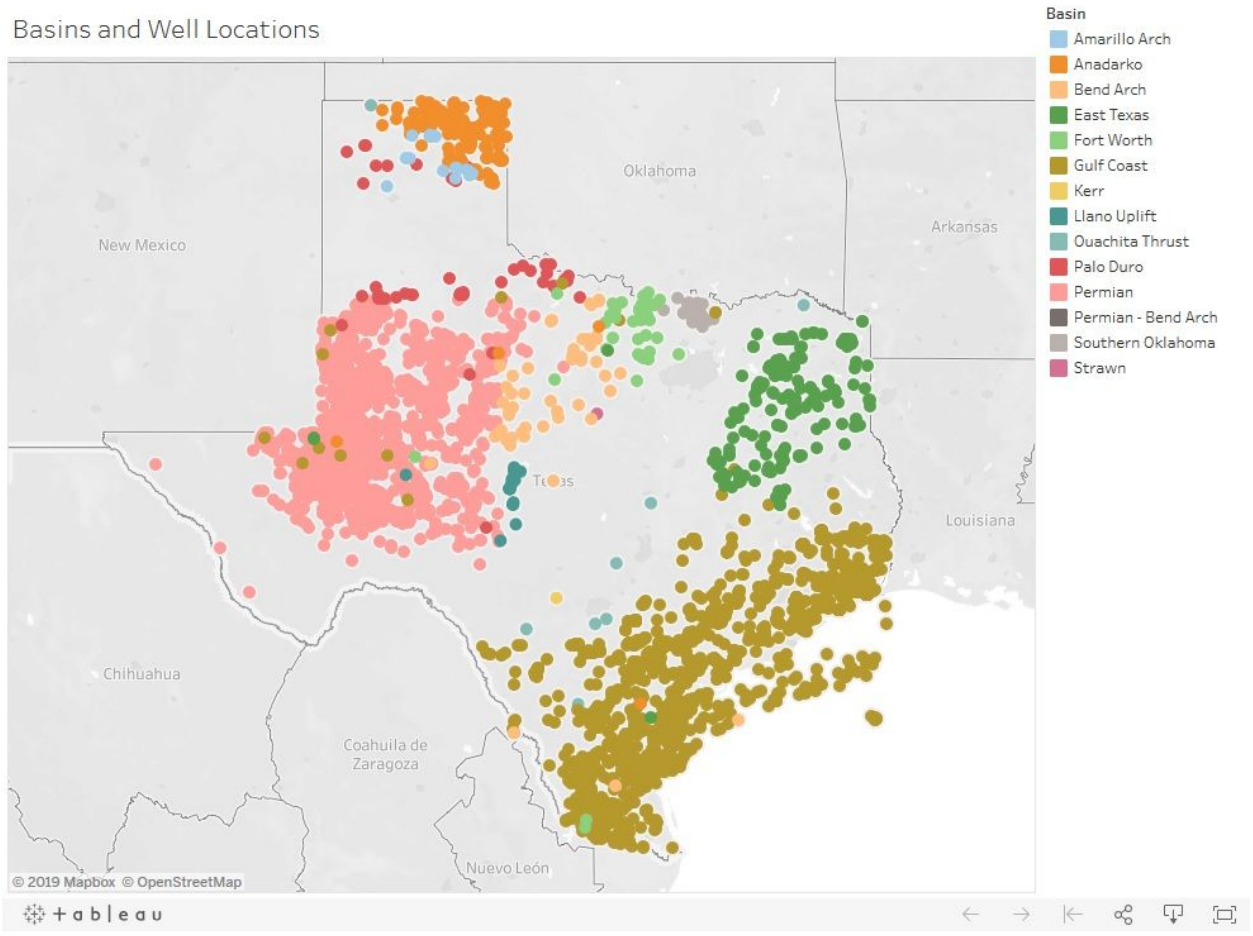
Milestone Report 1

Problem Statement

The dataset that I am working on describes the produced water characteristics from onshore wells drilled in Texas between 1920 and 2010. It has been obtained from the United States Geological Survey (USGS) National Produced Waters Database v2.3. The question that I am trying to answer are how do the water characteristics differ before the Permian oil shale revolution 2012 onwards. Hydraulic fracturing was introduced after 2012 and with advances in horizontal and automated drilling, long horizontal wells started being drilled to fracture shale rock formations under the surface of the earth. The largest oil shale formation exists in the Permian basin region of Texas, Oklahoma and New Mexico. I have decided to look at Texas wells only as a proof-of-concept that can be applied throughout the USA and Canada with shale formations.

Basins and well locations



Produced water comes out of the earth along with oil and gas as wells are drilled, completed and start producing. Before the shale oil revolution, Texas was producing from different basins across the state. Due to the introduction of frac fluids, one of the hypotheses is

that the total dissolved solids (TDS) concentrations have gone up after the fracking revolution. The answer to this question may be important to study the effects of fracking on the environment, produced water quality in this case. Governments, NGOs, oil producers, water and wastewater companies, politicians, municipalities close to the Permian/Midland region should all find this problem significant. The Jupyter notebook for the project is available at https://github.com/shubacca/Produced-Waters/tree/master/Produced%20Waters

Data Wrangling

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.

Exploratory Data Analysis

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.

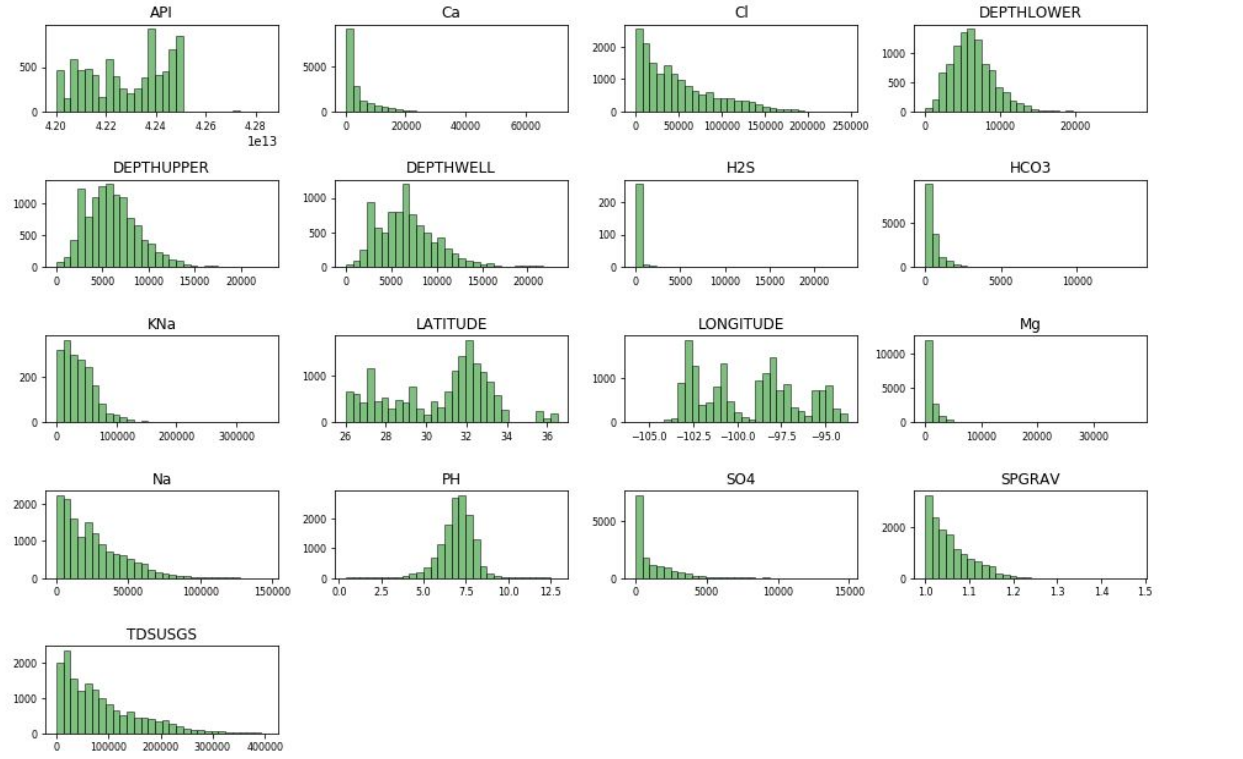
Accordingly, the major plays are Amarillo, Anadarko, Permian, East Texas and Gulf Coast. These basins were drilled and their produced water characteristics studied between 1920 and 2010. (It is important to point out that these are data points, and do not necessary mean the number of wells. One well 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.

Individual histograms for all numeric data.

```
df_drop.hist(bins=30, color='green', alpha=0.5,edgecolor='black', linewidth=1.0,
             xlabelsize=8, ylabelsize=8, grid=False)
plt.tight_layout(rect=(0, 0, 2.2, 2.2))
```



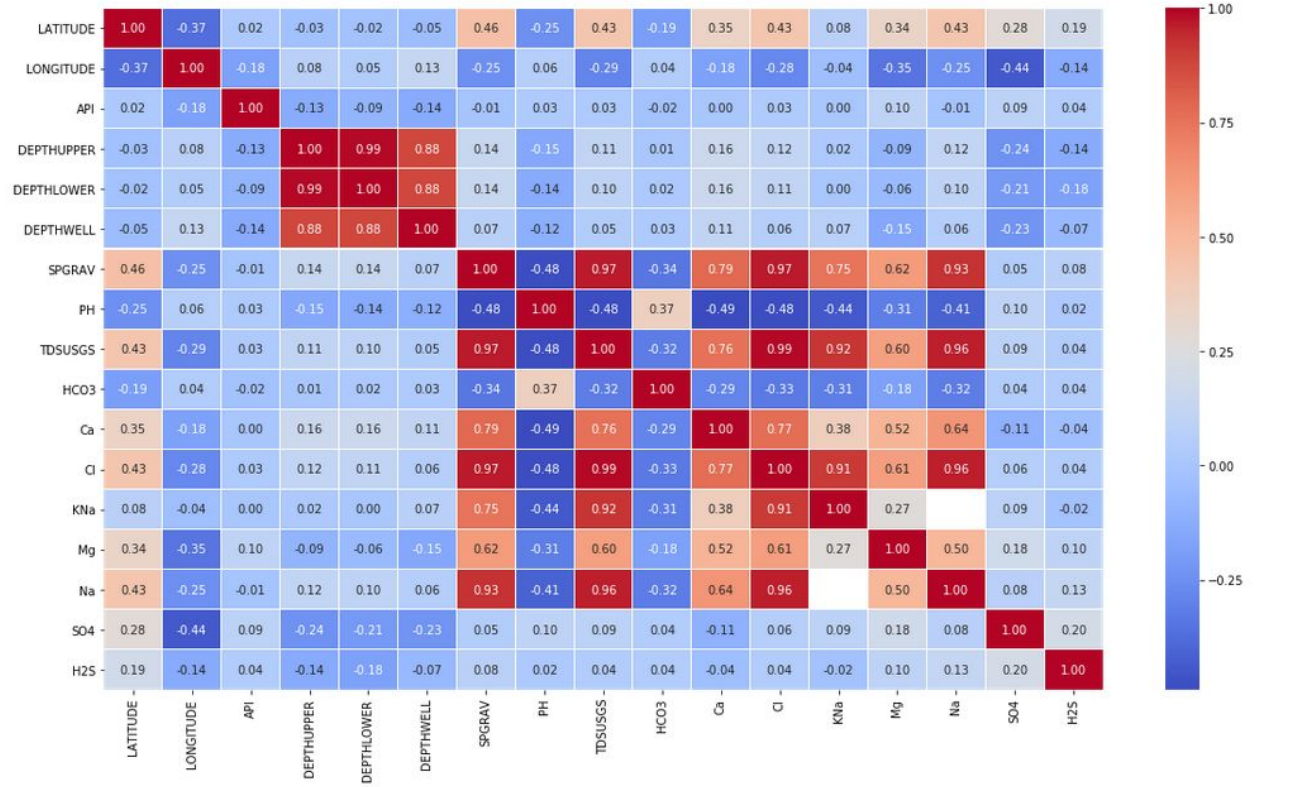
3. Heat-map and pair-wise scatter plot generation describing the correlations between these numerical variables:

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.

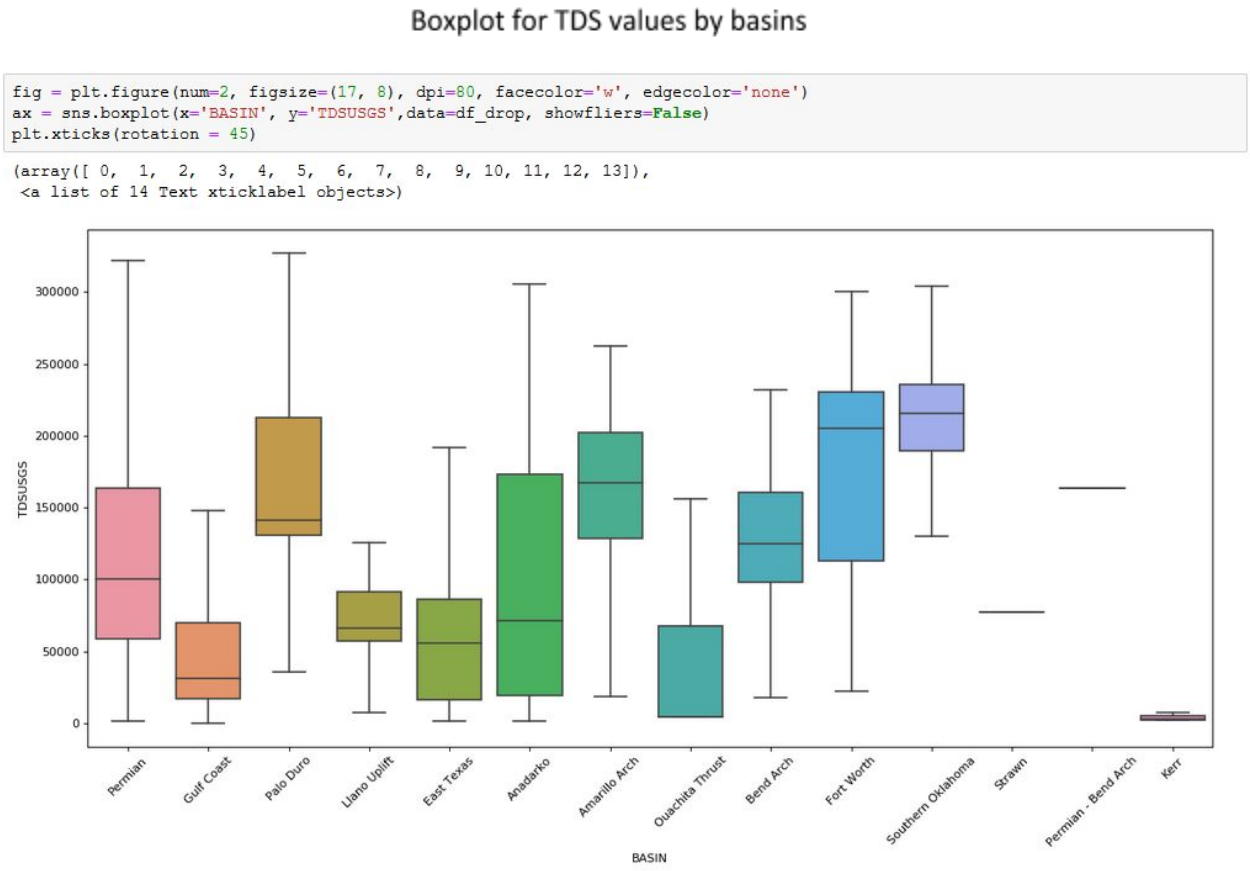
Correlation heatmap for all numeric data.

```
# Correlation Matrix Heatmap Comparisons
f, ax = plt.subplots(figsize=(19, 10))
corr = df_drop.corr()
hm = sns.heatmap(round(corr,2), annot=True, ax=ax, cmap="coolwarm",fmt='.2f',
                 linewidths=.05)
f.subplots_adjust(top=0.93)
t= f.suptitle('Texas Wells Produced Water Characteristics Correlation Heatmap', fontsize=14)
```



4. 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.



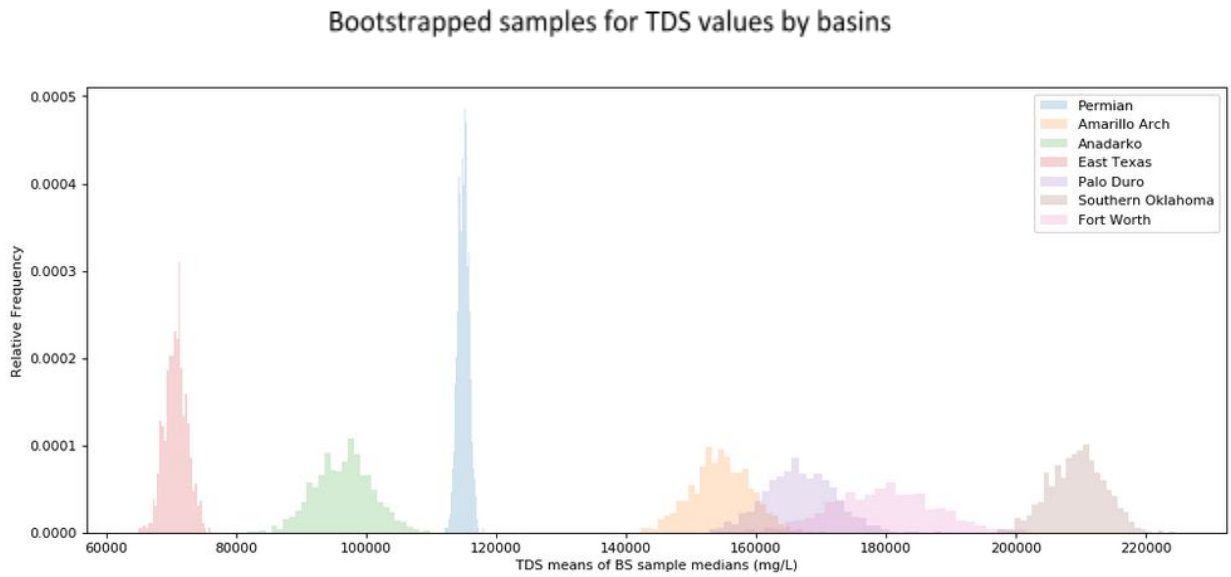
5. Aggregate stats (mean and standard deviation) for TDS, depths, calcium, chlorine, sodium, potassium-sodium, magnesium:

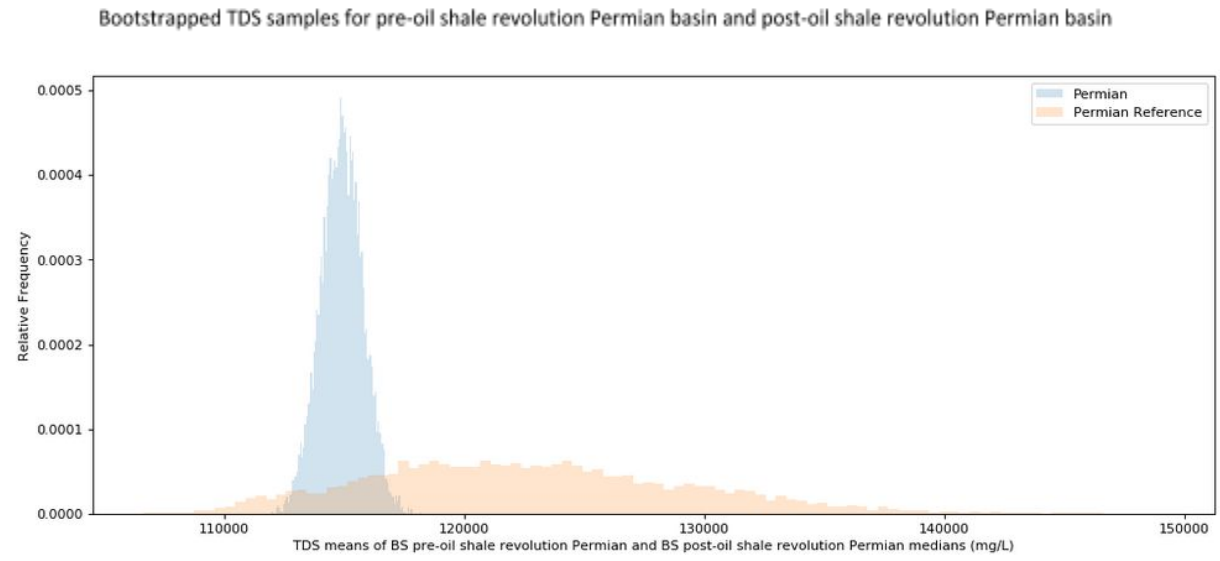
The basin with the highest TDS content is Southern Oklahoma, followed by Fort Worth and then Palo Duro. The standard deviation of TDS in the Permian basin is about 65% of the mean, and that the highest TDS spreads are noticeable for Anadarko, Permian and Fort Worth. The most hard waters (higher Ca and 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.

6. Depths histograms distribution by basin: lots of variability was observed.

7. 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.





8. 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.

Boxplots displaying the rates of change for different parameters (TDS, Ca, Cl, KNa, Mg and Na; units of mg/L/day), separated by different basins.