Capstone Project 1: Texas Oil Wells Produced Water Dataset Analysis before the Permian Oil Shale Revolution (1920 - 2010)

Problem Statement

The current report describes 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 questions that I am trying to answer are: i) whether water characteristics were significantly different before the Permian oil shale revolution in 2012, indicating that water contamination has increased due to the introduction of hydraulic fracturing; ii) whether TDS values of wells can be regressed using water quality parameters and origin basins. 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.

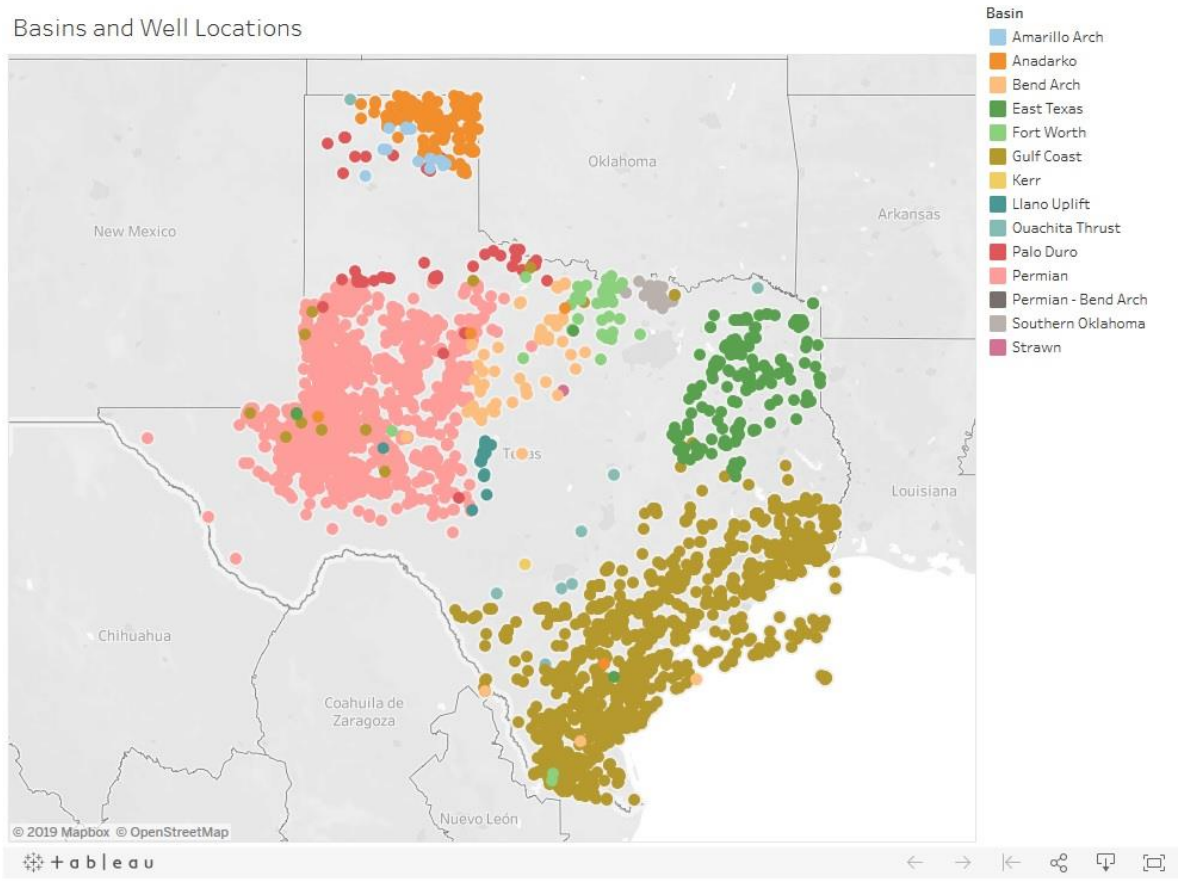


Figure 1: Basins and Well Locations

Produced water emerges from the earth along with oil and gas, as wells are drilled and cased (completed). 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 increased. 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

Project Proposal and Data Wrangling

The current report describes the analysis of produced water characteristics of onshore Texas oil wells between 1920 and 2010. The dataset was 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 with the data set.

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.

```
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))
```

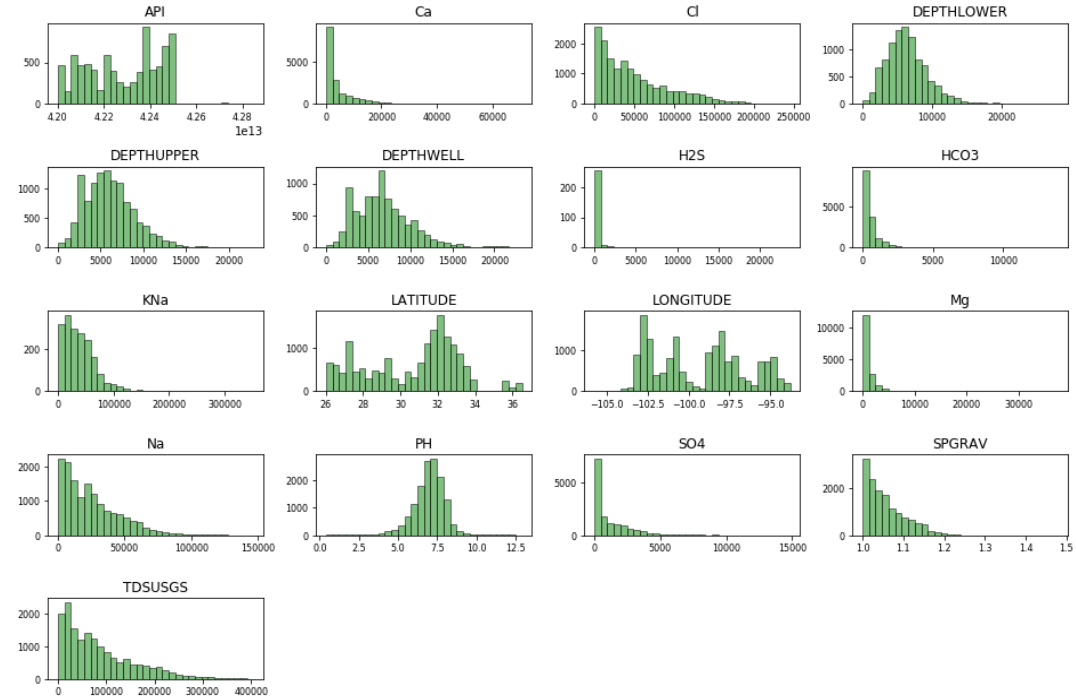


Figure 2: Individual Histograms for all Numeric Data

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 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)
```

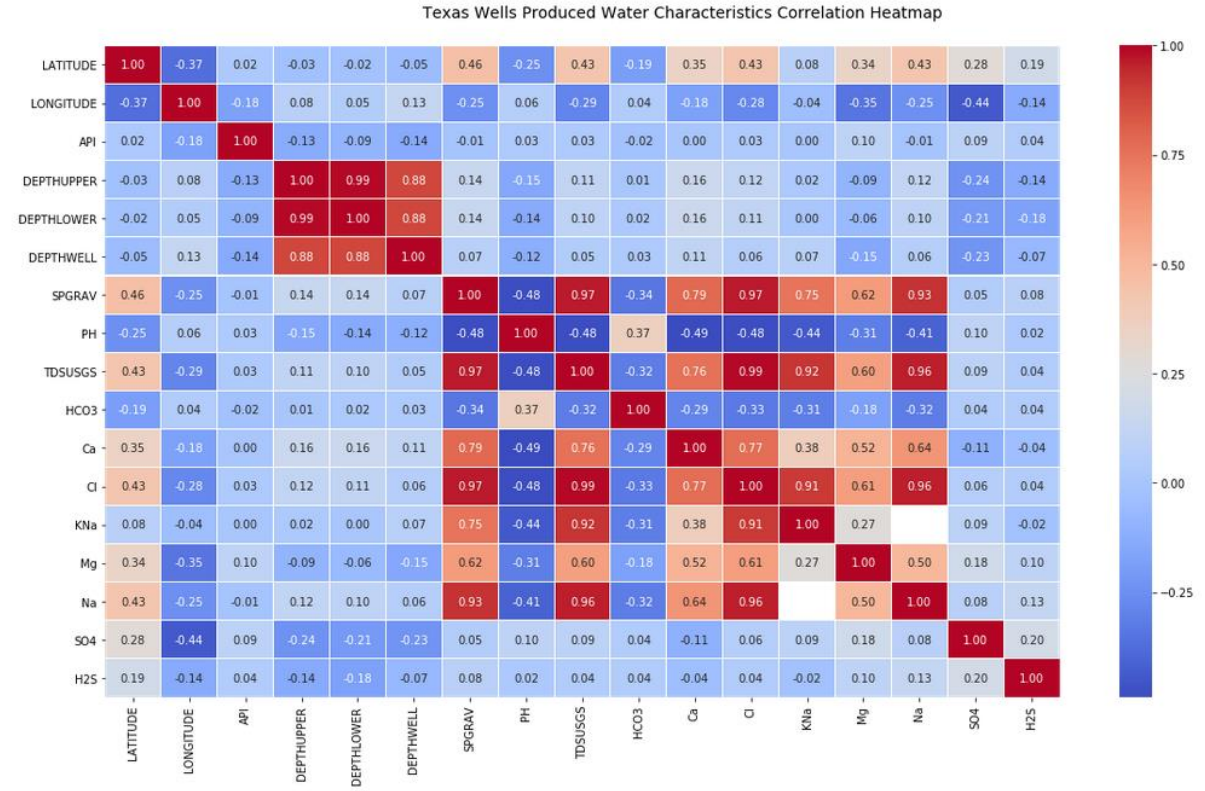


Figure 3: Correlation Heat Map for all Numeric Data

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.

```
fig = plt.figure(num=2, figsize=(17, 8), dpi=80, facecolor='w', edgecolor='none')
ax = sns.boxplot(x='BASIN', y='TDSUSGS',data=df_drop, showfliers=False)
plt.xticks(rotation = 45)
```

```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13]),
 <a list of 14 Text xticklabel objects>)
```

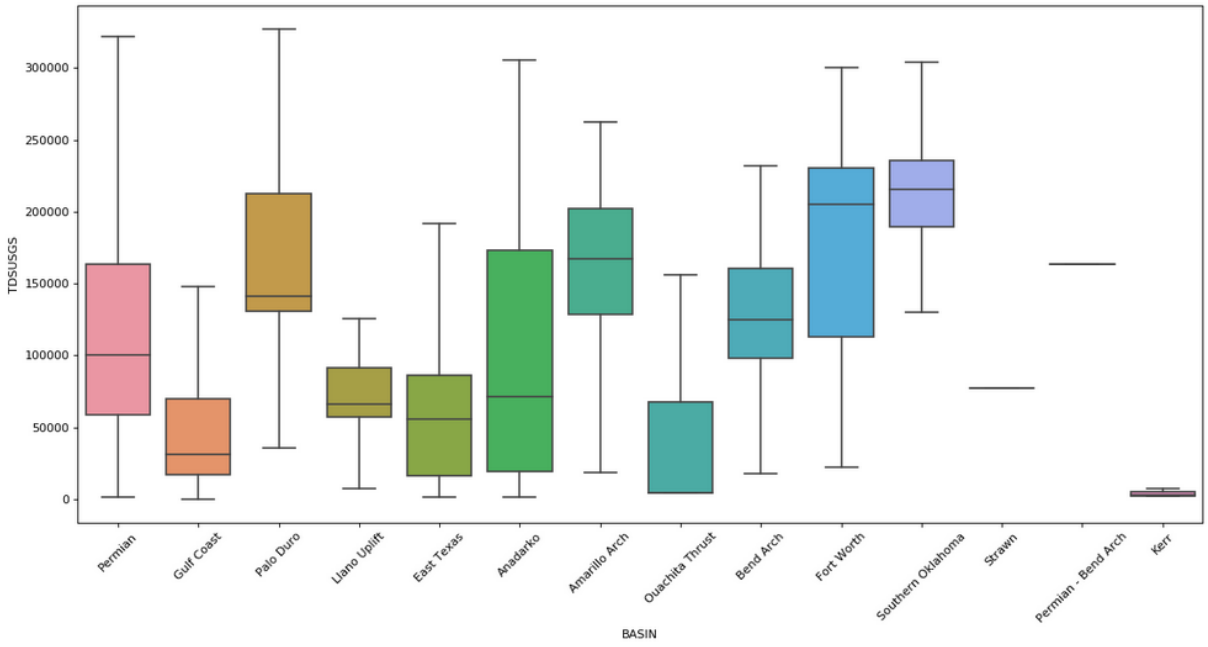


Figure 4: Boxplot for TDS Values by Basins

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.

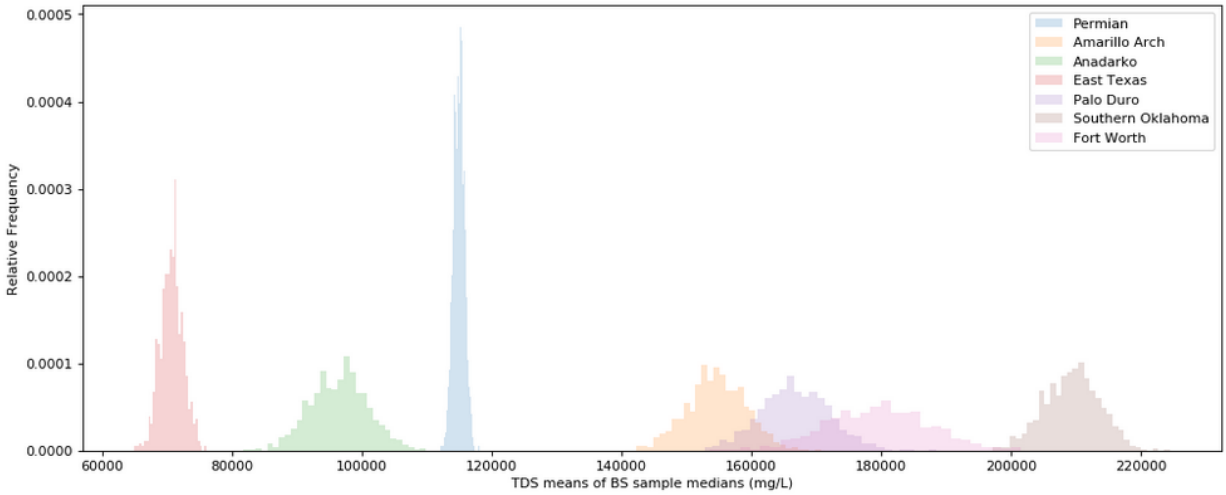


Figure 5: Bootstrapped Samples for TDS Values by Basins

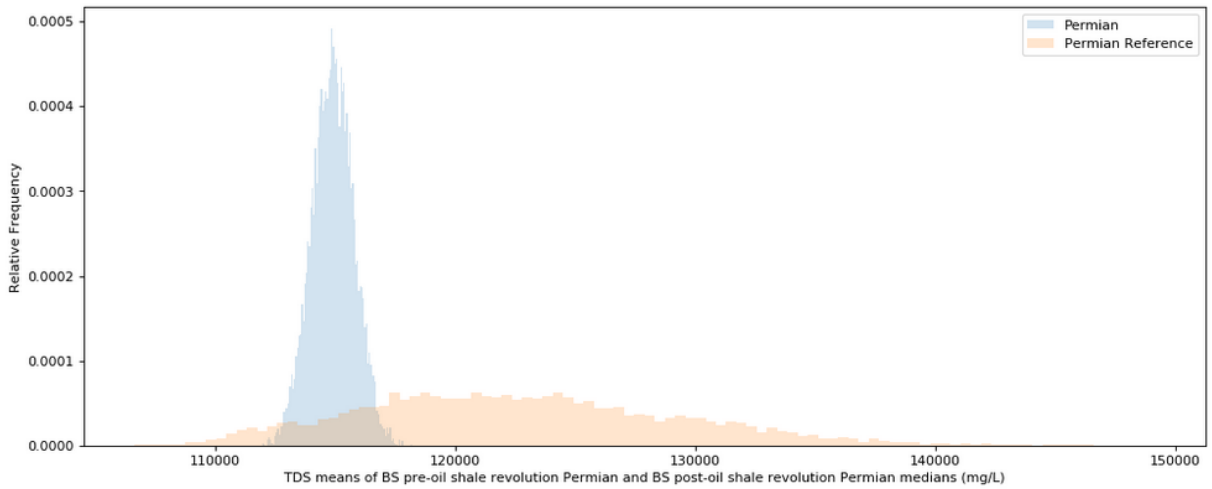


Figure 6: Bootstrapped TDS samples for pre-Oil Shale Revolution Permian Basin and post-Oil Shale Revolution Permian Basin

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.

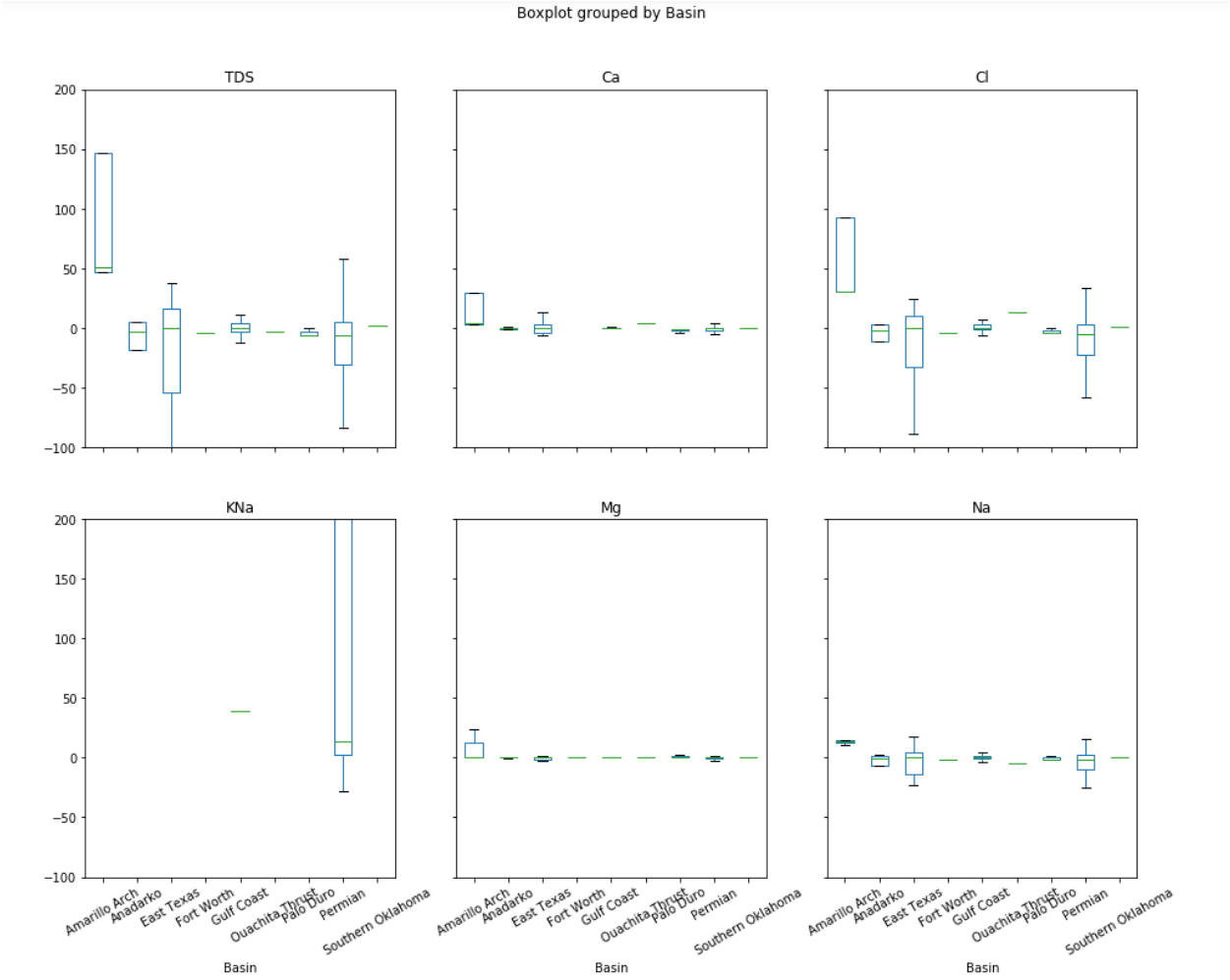


Figure 7: Boxplots displaying the Rates of Change for Different Parameters (TDS, Ca, Cl, KNa, Mg and Na; units of mg/L/day), Separated by Different Basins.

In-depth Analysis (Machine Learning)

The current project derived its machine learning question with an aim to imputing missing values for the depths. The hypothesis is that given some of the major variables that characterize produced waters and the basins they are present in, one can reasonably well predict the depth at which they originate. The problem is framed as a regression problem given the predicted variable (DEPTHUPPER) is continuous. In order to do that, the data needed to be prepared as follows:

1. Solve for Ca and Na using median imputation.
2. Drop null TDS values (given that there are only 12 missing).
3. Separate missing depths data, and create a test/train dataset with remaining variables.
4. Test and evaluate different linear regression models to solve for DEPTHUPPER using relevant variables.

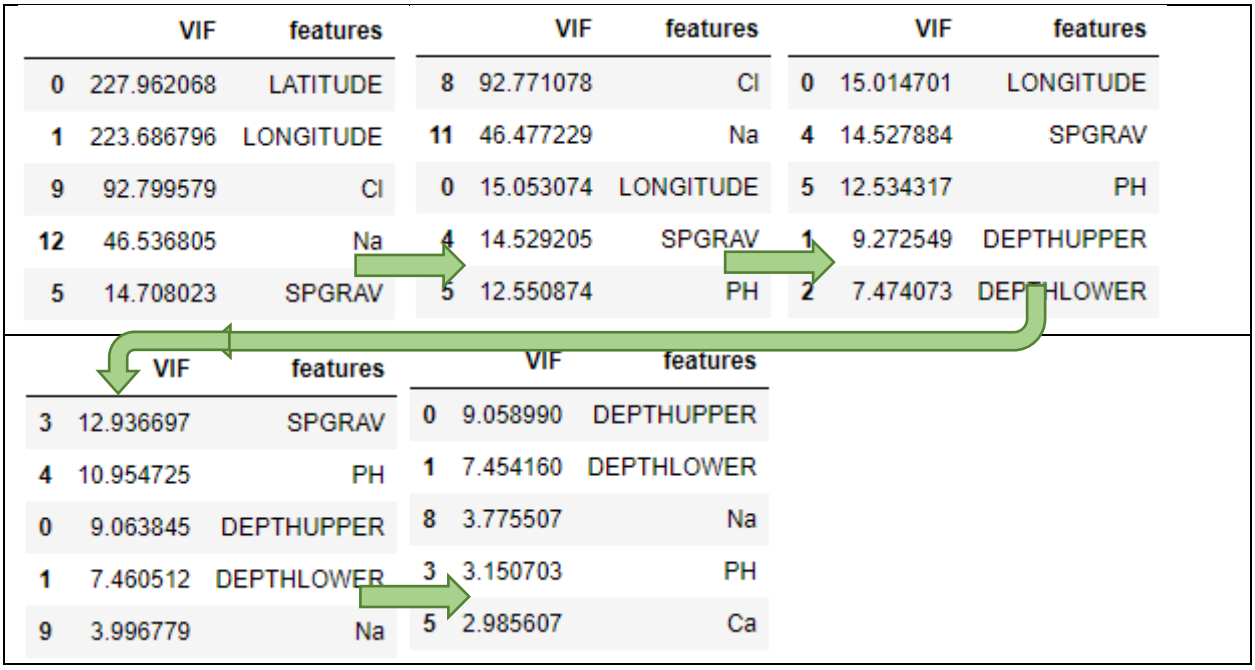
The most important part of selecting features comes from background knowledge. Produced waters have characteristics determined by their origin basins, formations and absolute depths below sea level. Conversely, depths should be related to the water quality and the basins of origin. It should be noted that these waters are naturally pressurized in the reservoirs, and not the result of hydraulic fracture stimulations. As a result the major predictor variables are latitude, longitude, basin, well type, chlorine (Cl), calcium (Ca), sodium (Na) and total dissolved solids (TDS) concentrations. A few more variables like magnesium and potassium-sodium concentrations could be chosen, but given the heat map generated between the continuous variables, multicollinearity development seemed apparent.

Hence a multicollinearity analysis was performed and the variance inflation factor was used on the numerical variables. The various iterations of the VIF table are given in the figure below. In order, LATITUDE, Cl, LONGITUDE and SPGRAV were dropped and the model was reduced to the remaining variables.

Steps 1 and 2 were easily accomplished, using *fillna()* and filtering for *notnull()* values from the TDSUSGS column. The data was filtered to include the variables mentioned above and stored in *df_prep*. For the third step, the missing depths were easily filtered out (4946 values out of total 16619 values). Using the *sklearn.metrics.train_test_split()* function, 20% of the data was separated then as test data. The rest of the training data was fitted with different regression models.

One of the challenges of handling the data was the presence of categorical types like basin (15 types) and well type (3 kinds). Formations were initially meant to be used as a predictor variable, but there are 2029 unique formations available, and so the idea was not pursued. To handle the categorical data, both variables were converted to dummy variables, with 1 less variable each (to avoid multicollinearity and redundancy). Then they were joined with the train data. The other variables were then standardized within (0, 1) using the *sklearn.preprocessing.MinMaxScaler()* function.

Table 1: Variable inflation factors in their iterations, until they all drop to below 10. Start from top left corner, after which LATITUDE is dropped, and then VIF run again to reveal Cl with a high VIF value. Cl is then dropped, and the process is repeated until all VIF values are below 10. These variables do not have multicollinearity.



	VIF	features
0	227.962068	LATITUDE
1	223.686796	LONGITUDE
9	92.799579	Cl
12	46.536805	Na
5	14.708023	SPGRAV

	VIF	features
8	92.771078	Cl
11	46.477229	Na
0	15.053074	LONGITUDE
4	14.529205	SPGRAV
5	12.550874	PH

	VIF	features
0	15.014701	LONGITUDE
4	14.527884	SPGRAV
5	12.534317	PH
1	9.272549	DEPTHUPPER
2	7.474073	DEPTHLOWER

	VIF	features
3	12.936697	SPGRAV
4	10.954725	PH
0	9.063845	DEPTHUPPER
1	7.460512	DEPTHLOWER
9	3.996779	Na

	VIF	features
0	9.058990	DEPTHUPPER
1	7.454160	DEPTHLOWER
8	3.775507	Na
3	3.150703	PH
5	2.985607	Ca

Regression evaluation takes the form of accuracy scores (R2 values between the predicted and real target variables), or root mean squared errors (RMSE). The results for the various regressions are denoted as follows:

Table 2: Algorithms for Regression for Scaled Parameters and their Scores

Algorithm	Parameters	Training R^2 value	Test R^2 value	RMSE
Ordinary Least Squares	*Default*	0.975	0.957	14944.3
Lasso Regression	*alpha = 0.1, max_iter = 10000*	0.975	0.957	14943.0
Ridge Regression	alpha = 0.1, max_iter = 10000	0.975	0.958	14891.8
Stochastic Gradient Descent Regression	*max_iter=1000, tol=1e-4, eta0=0.1*	0.974	0.960	14420.6
Linear Support Vector Machines Regression	*epsilon=.01*	-0.879	-0.880	99098.3
Polynomial Support Vector Machines Regression	*kernel='poly', degree=2, C=100, epsilon=1*	0.069	0.084	69161.4
Decision Tree Regression	*max_depth=15*	0.999	0.963	13856.4
Random Forest Regression	*max_depth=20, n_estimators=100*	0.997	0.984	9034.7

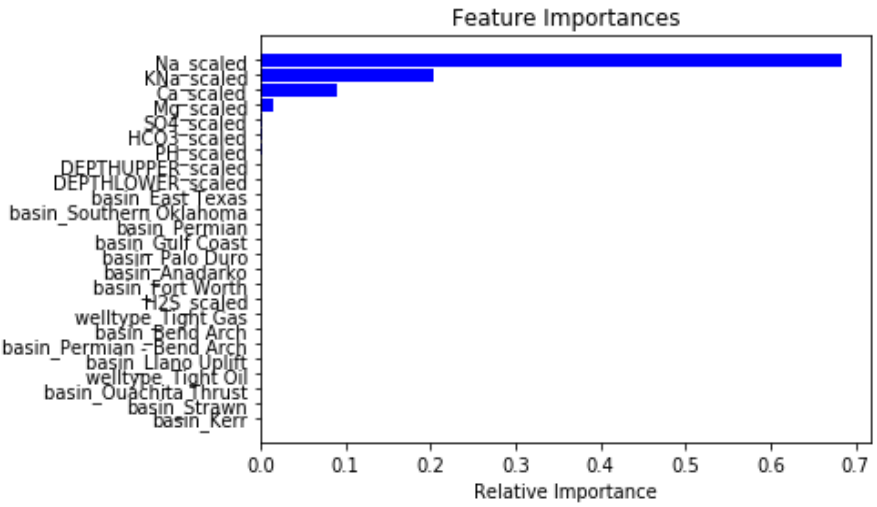


Figure 8: Feature Importances

Running the feature importances we can see that the TDS values were most influenced by Na, and rightfully so. This is because a majority of the total dissolved solids is usually sodium or chlorine. Only about 4 features were revealed to have any major effects on the target variable, TDSUSGS.

We could have run a 5-fold cross-validation on the best model, but our fit was so close to 100% that the model was accepted as is. Based on this data, a prediction was performed for the missing

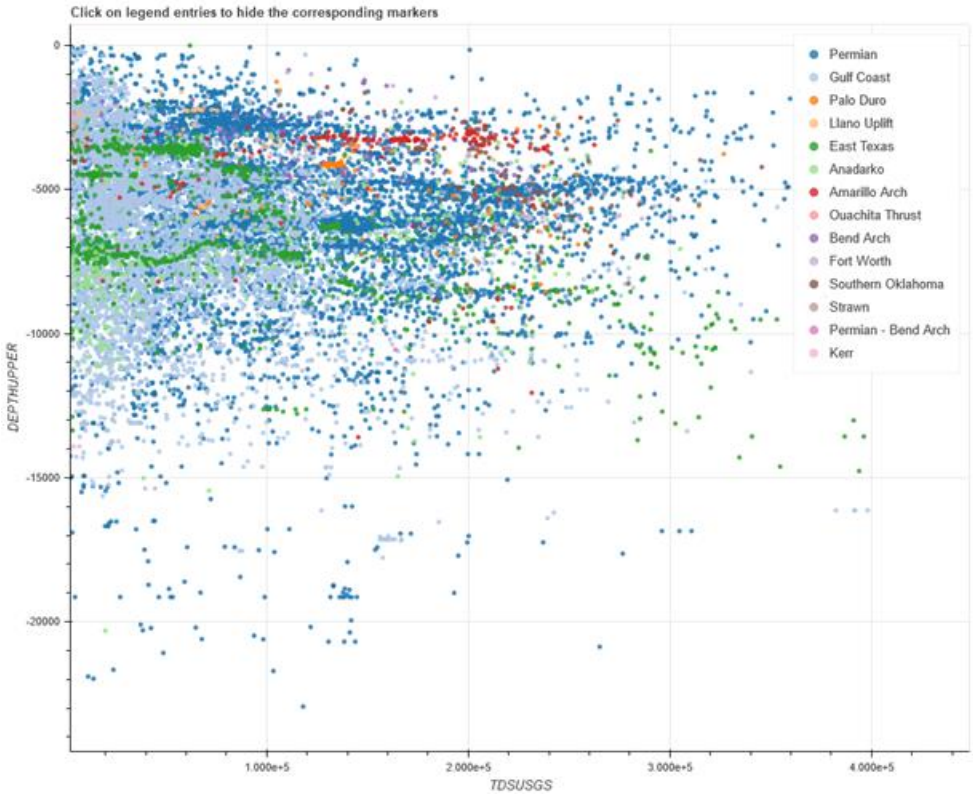


Figure 9: Distribution of Depths and their corresponding TDS Values Separated by Basins

depths and the missing values replaced. An interactive Bokeh plot was then constructed as follows, outlining the depths versus TDS relationships.

Conclusions

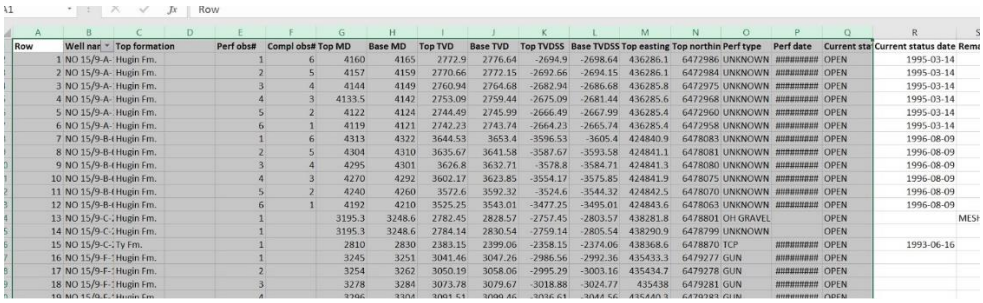
The two questions posed at the beginning of the study were satisfactorily answered by the analyses: i) the post-Oil Shale Revolution produced waters were significantly higher in TDS concentrations from the pre-revolution waters, found using bootstrapping techniques. This result can possibly be attributed to the drilling of more horizontal wells with more minerals and salts seeping in, or the addition of frac fluids; ii) given a R^2 score of 0.98 on the test data, the Random Forest model was best able to determine the TDS values given other datapoints. This was trained on a combination of categorical basin and well type values, and continuous water variables.

Future work in this topic would focus on gathering more water quality data after the shale oil revolution, and connecting previously drilled wells with newer frac projects on the same well paths. This way an increase in oil production quantities and a change in water characteristics can be studied. One can examine and impute the missing depths of the wells using all the other information provided. Predictions can be made given newer horizontal drilling technologies, their completion methods and the contribution of different frac stages to the quality of water flowing back. Given water constituents, and trends in TDS and other minerals over time, one can gain concrete evidence of the steps needed to characterize and recycle these produced waters, resulting in significant savings in environmental costs.

Appendix A: Initial Project Ideas

The project began with an exploration of possible ideas, as given below:

1. Oil, gas and water production from Volve DataSet in North Sea: Correlations between geophysical, reservoir, drilling and completions, and production data (2007-2016)
    a. DataSet used from the Volve database by Equinor: https://data.equinor.com/dataset/Volve



Row	Well nar	Top formation	Perf obs#	Compl obs#	Top MD	Base MD	Top TVD	Base TVD	Top TVDSS	Base TVDSS	Top easting	Top northin	Perf type	Perf date	Current sta	Current status date	Rem
1	NO 15/9-A-	Hugin Fm.	1	6	4160	4165	2772.9	2776.64	-2694.9	-2698.64	436286.1	6472986	UNKNOWN	########	OPEN	1995-03-14	
2	NO 15/9-A-	Hugin Fm.	2	5	4157	4159	2770.66	2772.15	-2692.66	-2694.15	436286.1	6472984	UNKNOWN	########	OPEN	1995-03-14	
3	NO 15/9-A-	Hugin Fm.	3	4	4144	4149	2760.94	2764.68	-2682.94	-2686.68	436285.8	6472975	UNKNOWN	########	OPEN	1995-03-14	
4	NO 15/9-A-	Hugin Fm.	4	3	4133.5	4142	2753.09	2759.44	-2675.09	-2681.44	436285.6	6472968	UNKNOWN	########	OPEN	1995-03-14	
5	NO 15/9-A-	Hugin Fm.	5	2	4122	4124	2744.49	2745.99	-2666.49	-2667.99	436285.4	6472960	UNKNOWN	########	OPEN	1995-03-14	
6	NO 15/9-A-	Hugin Fm.	6	1	4119	4121	2742.23	2743.74	-2664.23	-2665.74	436285.4	6472958	UNKNOWN	########	OPEN	1995-03-14	
7	NO 15/9-B-(	Hugin Fm.	1	6	4313	4322	3644.53	3653.4	-3596.53	-3605.4	424840.9	6478083	UNKNOWN	########	OPEN	1996-08-09	
8	NO 15/9-B-(	Hugin Fm.	2	5	4304	4310	3635.67	3641.58	-3587.67	-3593.58	424841.1	6478081	UNKNOWN	########	OPEN	1996-08-09	
9	NO 15/9-B-(	Hugin Fm.	3	4	4295	4301	3626.8	3632.71	-3578.8	-3584.71	424841.3	6478080	UNKNOWN	########	OPEN	1996-08-09	
10	NO 15/9-B-(	Hugin Fm.	4	3	4270	4292	3602.17	3623.85	-3554.17	-3575.85	424841.9	6478075	UNKNOWN	########	OPEN	1996-08-09	
11	NO 15/9-B-(	Hugin Fm.	5	2	4240	4260	3572.6	3592.32	-3524.6	-3544.32	424842.5	6478070	UNKNOWN	########	OPEN	1996-08-09	
12	NO 15/9-B-(	Hugin Fm.	6	1	4192	4210	3525.25	3543.01	-3477.25	-3495.01	424843.6	6478063	UNKNOWN	########	OPEN	1996-08-09	
13	NO 15/9-C-(	Hugin Fm.	1		3195.3	3248.6	2782.45	2828.57	-2757.45	-2803.57	438281.8	6478801	OH GRAVEL		OPEN		MES
14	NO 15/9-C-(	Hugin Fm.	1		3195.3	3248.6	2784.14	2830.54	-2759.14	-2805.54	438290.9	6478799	UNKNOWN		OPEN		
15	NO 15/9-C-	Ty Fm.	1		2810	2830	2383.15	2399.06	-2358.15	-2374.06	438368.6	6478870	TCP	########	OPEN	1993-06-16	
16	NO 15/9-F-	Hugin Fm.	1		3245	3251	3041.46	3047.26	-2986.56	-2992.36	435433.3	6479277	GUN	########	OPEN		
17	NO 15/9-F-	Hugin Fm.	2		3254	3262	3050.19	3058.06	-2995.29	-3003.16	435434.7	6479278	GUN	########	OPEN		
18	NO 15/9-F-	Hugin Fm.	3		3278	3284	3073.78	3079.67	-3018.88	-3024.77	435438	6479281	GUN	########	OPEN		

	A	B	C	D	E	F	G	H	I	J	K	L
1	Wellbore name	NPDCode	Year	Month	On Stream	Oil	Gas	Water	GI	WI		
2					hrs	Sm3	Sm3	Sm3	Sm3	Sm3		
3	15/9-F-1 C	7405	2014	4	228	11,142	1,597,937	0	NULL	NULL	1	
4	15/9-F-1 C	7405	2014	5	734	24,902	3,496,230	783	NULL	NULL	2	
5	15/9-F-1 C	7405	2014	6	706	19,618	2,886,662	2,068	NULL	NULL	3	
6	15/9-F-1 C	7405	2014	7	742	15,086	2,249,366	6,244	NULL	NULL	4	
7	15/9-F-1 C	7405	2014	8	433	6,970	1,048,191	4,530	NULL	NULL	5	
8	15/9-F-1 C	7405	2014	9	630	9,168	1,414,100	8,318	NULL	NULL	6	
9	15/9-F-1 C	7405	2014	10	745	9,468	1,462,064	10,365	NULL	NULL	7	
10	15/9-F-1 C	7405	2014	11	580	6,710	1,044,188	7,234	NULL	NULL	8	
11	15/9-F-1 C	7405	2014	12	28	120	25,857	183	NULL	NULL	9	
12	15/9-F-1 C	7405	2015	1	480	10,876	1,604,935	6,851	NULL	NULL	10	
13	15/9-F-1 C	7405	2015	2	437	9,587	1,439,454	10,745	NULL	NULL	11	

b. My clients are exploration and production oil and gas companies that can get smarter predictions for reservoir and oil production capacity given certain geophysical characteristics.
c. Some questions that can be asked are:
    i. Exploratory Data Analysis per well drilled and produced
    ii. Oil and gas Production profile analysis (rate of production, total amount produced == estimated amount)
    iii. GIS analysis based on depth, x and y coordinates, and figuring out if the entire formation is tapped out through fracturing
    iv. Calculate rate of penetration using maintenance parameters (check paper Rate of penetration (ROP) modeling using hybrid models: deterministic and machine learning)
    v. Perform NLP on daily drilling reports to understand when a trip (an interruption) would occur, and if that can be prevented before it happens
    vi. Perform NLP on geophysical formation description to understand which layers have oil production capacities
    vii. Deliverables: Code, Report, Blog Post (also available is salt challenge from kaggle)

2. Song analysis of covers of popular songs on Spotify
    a. DataSets used: Spotify API call for top pop, rock, r&b, jazz songs, and their top 10-20 covers. The categories would include the following parameters:


```
  "duration_ms" : 255349,
  "key" : 5,
  "mode" : 0,
  "time_signature" : 4,
  "acousticness" : 0.514,
  "danceability" : 0.735,
  "energy" : 0.578,
  "instrumentalness" : 0.0902,
  "liveness" : 0.159,
  "loudness" : -11.840,
  "speechiness" : 0.0461,
  "valence" : 0.624,
  "tempo" : 98.002,
  "id" : "06AKEBrKUckW0KREUWRnvT",
  "uri" : "spotify:track:06AKEBrKUckW0KREUWRnvT"
,
  "track_href" : "https://api.spotify.com/v1/tra
cks/06AKEBrKUckW0KREUWRnvT",
  "analysis_url" : "https://api.spotify.com/v1/a
udio-analysis/06AKEBrKUckW0KREUWRnvT",
  "type" : "audio_features"
```

Track Features: Popularity, Duration, Key, Mode, Time signature, Acousticness, Danceability, Energy, Instrumentalness, Liveness, Loudness, Speechiness, Valence, Tempo

Rate Limits: 50 results per query (more information on this https://tgel0.github.io/blog/spotify-data-project-part-1-from-data-retrieval-to-first-insights/)

b. The clients are Spotify or any other streaming music platforms, trying to study what makes a song popular (how to create good covers based on Spotify parameters?)
c. Some questions that can be asked are:
    i. Initial exploratory data analysis
    ii. How is popularity affected by any of the original song's parameters? Or is the cover more popular than the original, why?
    iii. Frequency of cover songs, years they were published
    iv. How is valence (emotions conveyed by the song) affected by the other parameters?
    v. Top artists who cover frequently

Dataset: Spotify API and this conference paper: Applying Data Mining for Sentiment Analysis in Music (rate limits for Spotify API, run it for outputs) (https://www.researchgate.net/publication/318510880_Applying_Data_Mining_for_Sentiment_Analysis_in_Music)

3. Things to do recommendation system for Banff National Park, based on TripAdvisor API and weather reports (NLP on reviews)
    a. DataSets used: TripAdvisor API reference call. Parameters associated:

    Location ID, Name, Rating, Number of reviews, ranking data, category and subcategory, awards, latitude, longitude, price level, cuisine type, attraction type

    b. Rate limits: 50 calls/second, 1000 calls per day
    c. The clients are anyone interested in NLP
    d. Some questions to ask are:
        i. Initial exploratory data analysis
        ii. Rating based on other parameters
        iii. Rating based on Reviews written and reviews based on ratings
        iv. Depending on the season, one can decide to go hiking or decide to go snowshoeing or skiing

    Paper: Sentiment Analysis in TripAdvisor (https://ieeexplore.ieee.org/abstract/document/8012330)

    Paper: Sentiment Analysis in tourism: capitalizing on big data https://www.researchgate.net/profile/Alireza_Alaei/publication/321812828_Se

ntiment_Analysis_in_Tourism_Capitalizing_on_Big_Data/links/5a7bc1eeaca27233575b2184/Sentiment-Analysis-in-Tourism-Capitalizing-on-Big-Data.pdf

4. Other Ideas:
    a. FracFocus Registry has data on frac water characteristics and quantities injected into the ground
    b. US Data - Royalties because of oil and gas to federal government according to counties
    c. Data Mining the water table - competition by DataDriven
    d. Unearth competition on Explorer Challenge