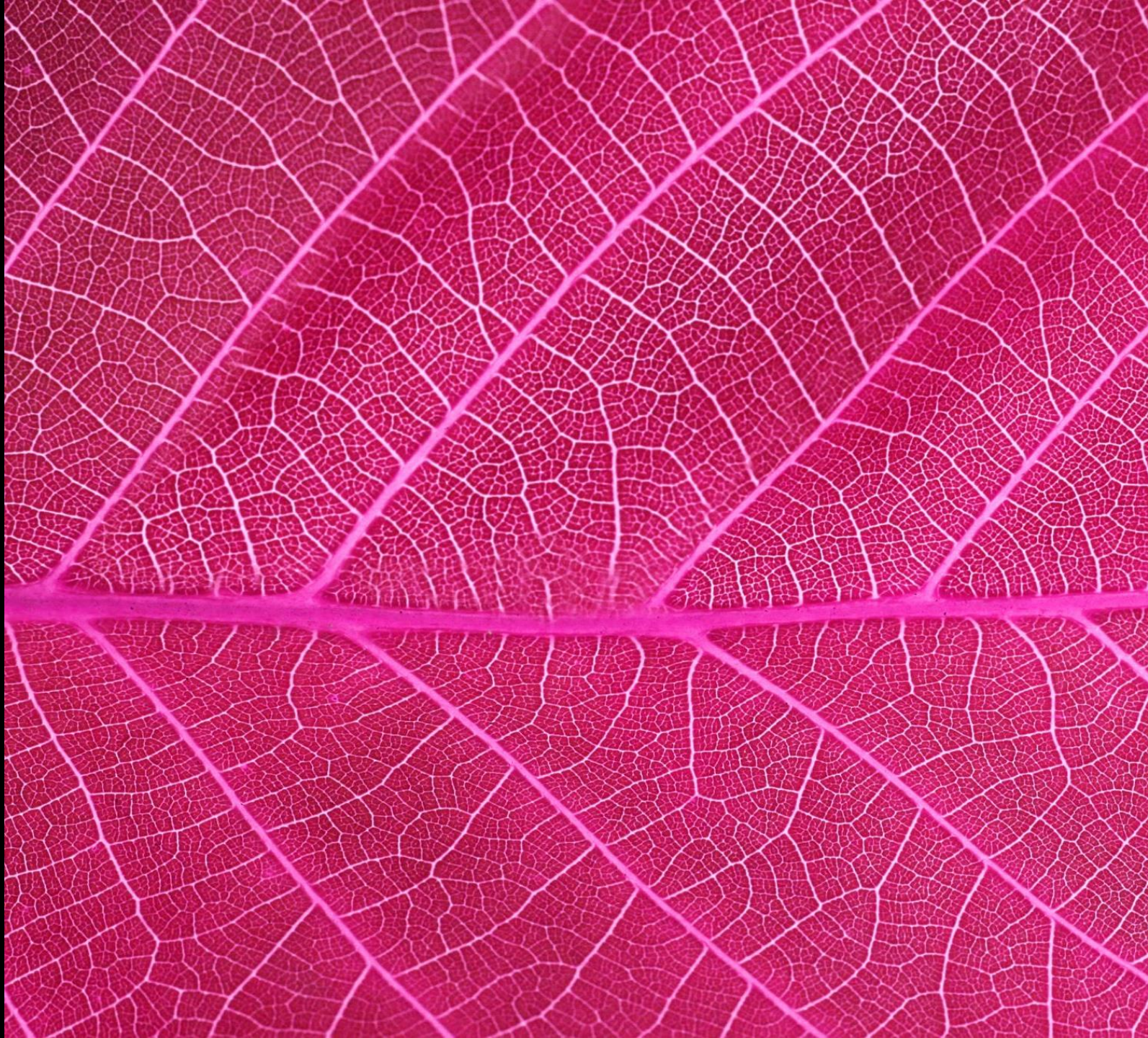


Texas oil wells produced water dataset analysis before the Permian oil shale revolution (1920 – 2010)

Springboard Capstone project I

Shubham Tiwari





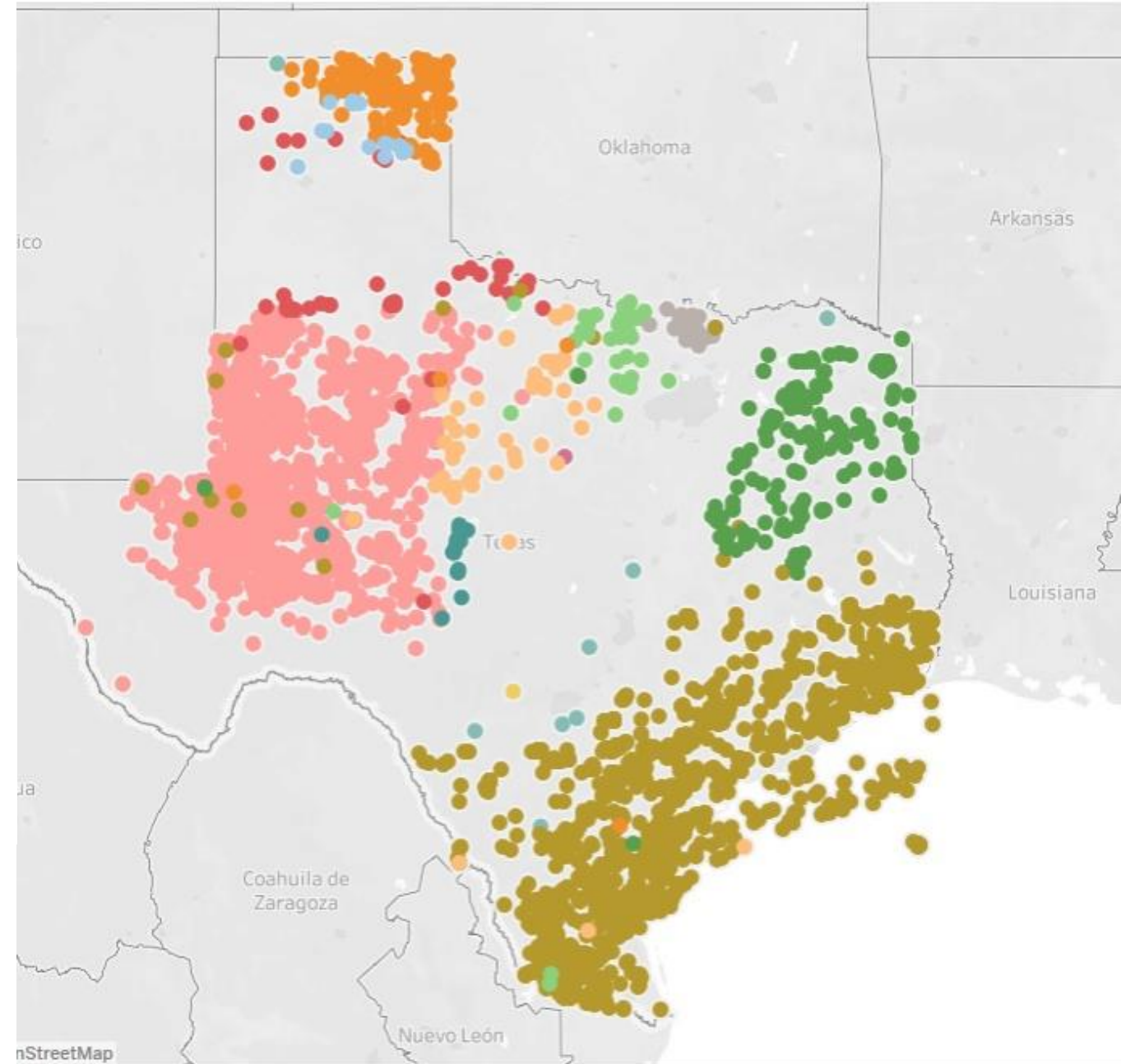
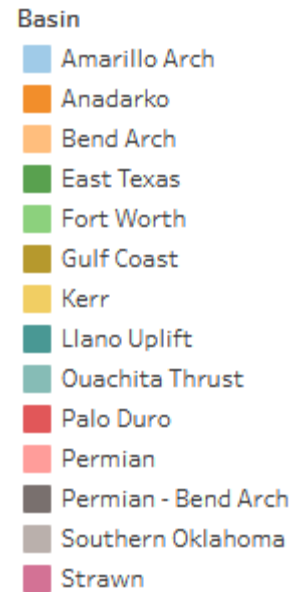
Introduction: Terms Definition


- Produced water: emerges from the earth with crude oil and gas at natural pressures, as wells are drilled and cased (completed)
- Permian shale oil revolution: increased oil output from the Permian basin in southern USA due to combination of techniques including:
 - Horizontal drilling
 - Multi-stage hydraulically fractured oil-rich shale rock
- Hydraulic Fracturing: well-stimulation process of pumping high quantities of frac fluid (primarily water, containing sand and other thickening agents) at high pressures, to fracture tight formations like shale

Problem Statement

1. Hypothesis: Water characteristics (total dissolved solids) were significantly different before the Permian oil shale revolution in 2012
2. Depths of wells can be found using water quality parameters and basins of origin.

Locations of wells by basin (1920-2010)





Dataset and Motivations of Study

- Obtained from United States Geological Survey (USGS) [National Produced Waters Database v2.3](#)
- To study the effects of fracking on environment, water quality
- Significant stakeholders: governments, NGOs, oil producers, water and wastewater companies, politicians, municipalities close to the Permian/Midland region
- The Jupyter notebook for the project is available at <https://github.com/shubacca/Produced-Waters/tree/master/Produced%20Waters>

Data Wrangling and Missing Data Imputations

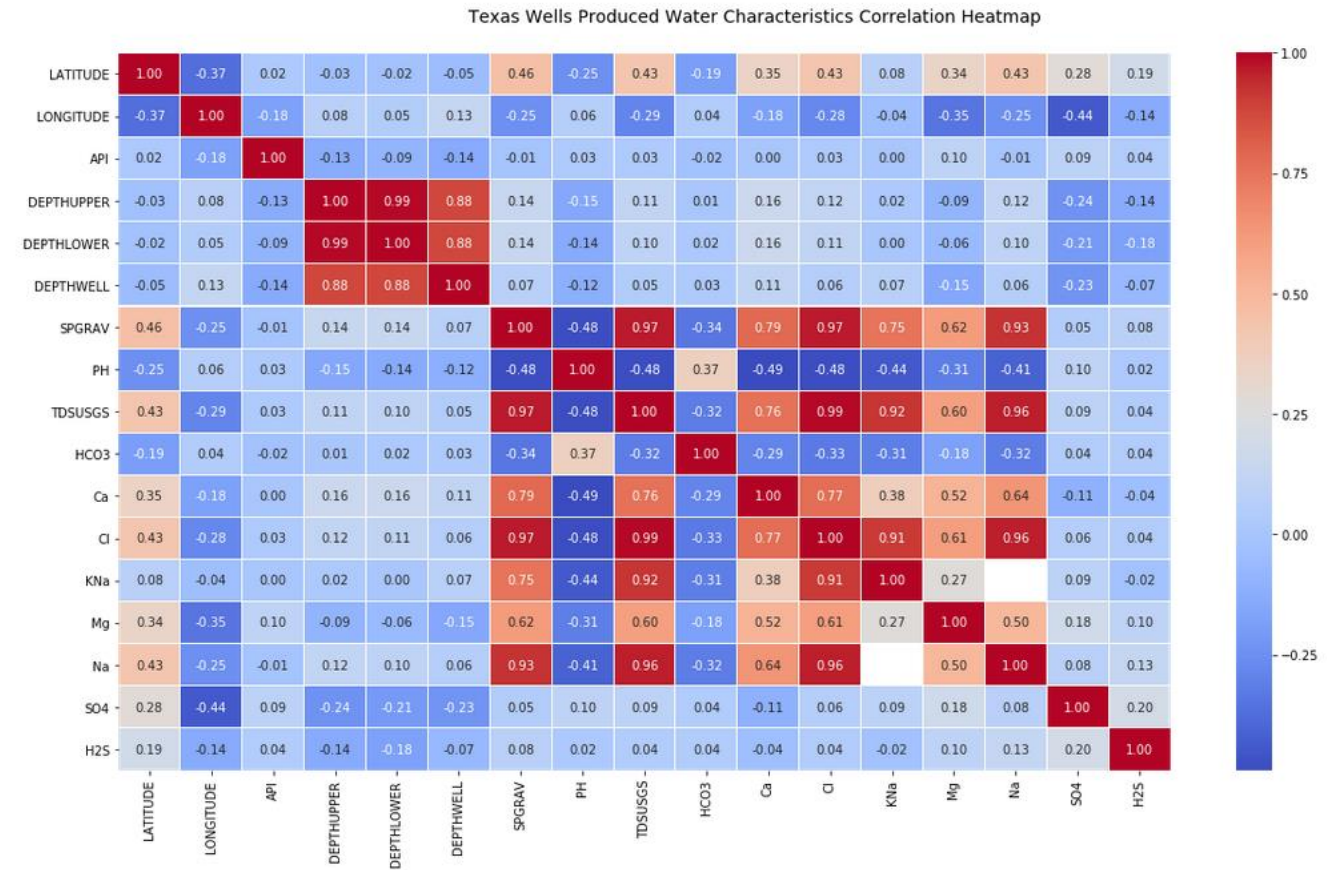
- Initial dataset: 190 features and 19388 entries
- Features reduced to 27:
 - *Positional*: 'LATITUDE', 'LONGITUDE'
 - *Basin characteristics*: 'API', 'BASIN', 'STATE', 'DATECOMP', 'DATESAMPLE', 'FORMATION', 'PERIOD', 'DEPTHUPPER', 'DEPTHLOWER', 'DEPTHWELL', 'LITHOLOGY',
 - *Water characteristics*: 'SG', 'SPGRAV', 'PH', 'TDSUSGS', 'TDS', 'HCO3', 'Ca', 'Cl', 'KNa', 'Mg', 'Na', 'SO4', 'H2S', 'cull_chargeb'
- Missing positional values imputed through checks with unique API values and well names
- Median values used for missing salts imputation like Ca, Cl, Mg.

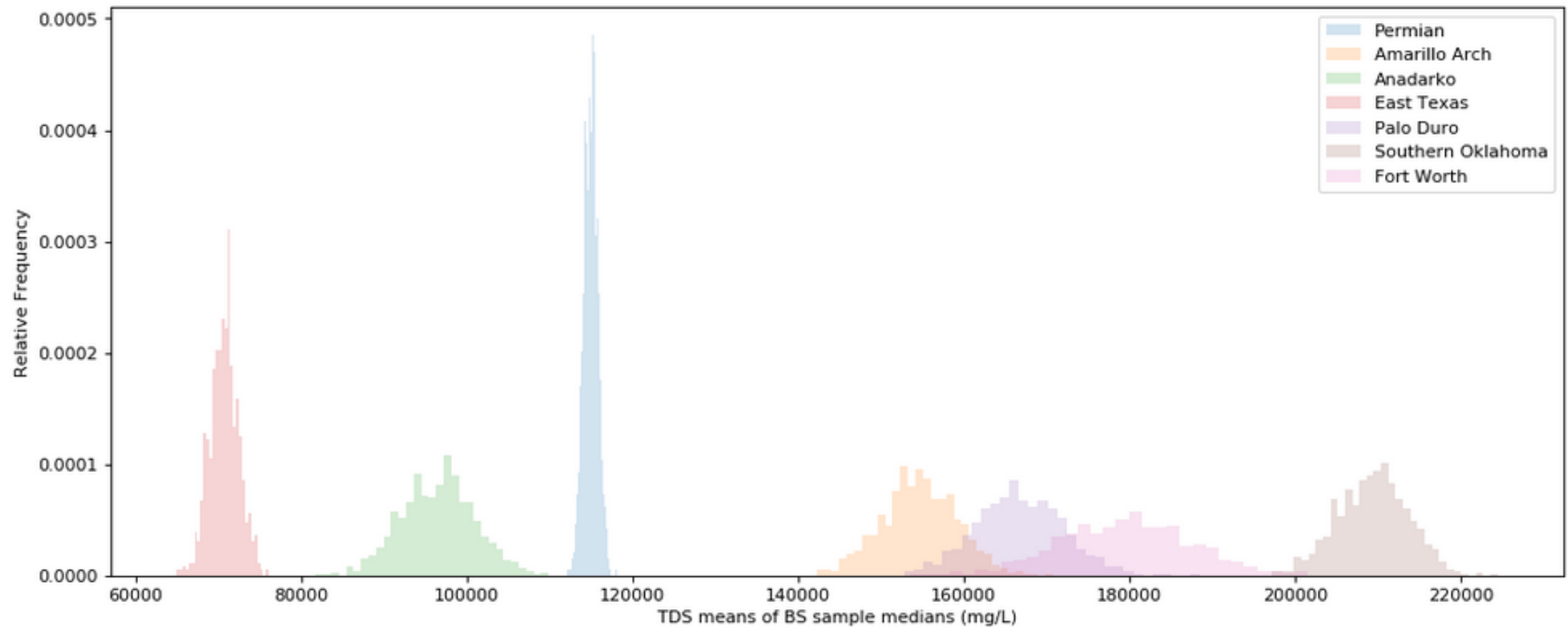
Exploratory Data Analysis

Correlation Heatmap Analysis

Magnesium and calcium don't contribute to TDS as much, suggesting these elements are found in insoluble phases

```
# 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)
```



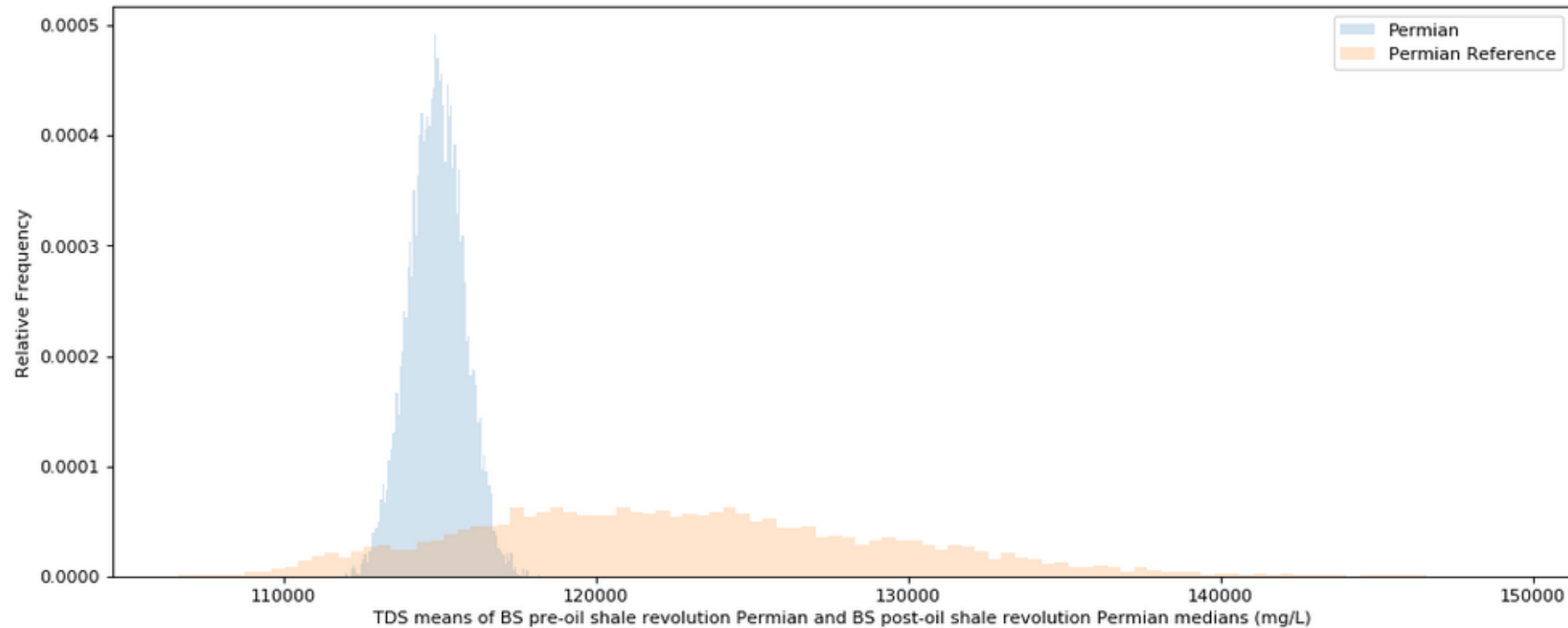


Exploratory Data Analysis

Bootstrapped Tests on TDS Data by Basin

1000 simulations on current dataset aggregated over TDS median values

Results showed Permian basin had the least spread in TDS values as compared to Fort Worth basin



Pre-oil shale revolution TDS values vs. post-revolution: Hypothesis Testing

1000 bootstrapped simulations on current dataset for Permian TDS medians compared with modern values¹

Z-score = 116.77, p-value = 0.0001 => reject null hypothesis => TDS did in fact increase after the revolution

Increase attributed to drilling of more horizontal wells with more minerals and salts seeping in and due to addition of frac fluids

1. Khan, N. A., Engle, M., Dungan, B., Holguin, F. O., Xu, P., & Carroll, K. C. (2016). Volatile-organic molecular characterization of shale-oil produced water from the Permian Basin. *Chemosphere*, 148, 126–136. doi:10.1016/j.chemosphere.2015.12.116

In-depth Analysis Methodology for Depth Imputation



Solve for Cl, Ca and Na using median imputation.



Drop null TDS values (given that there are only 12 missing).

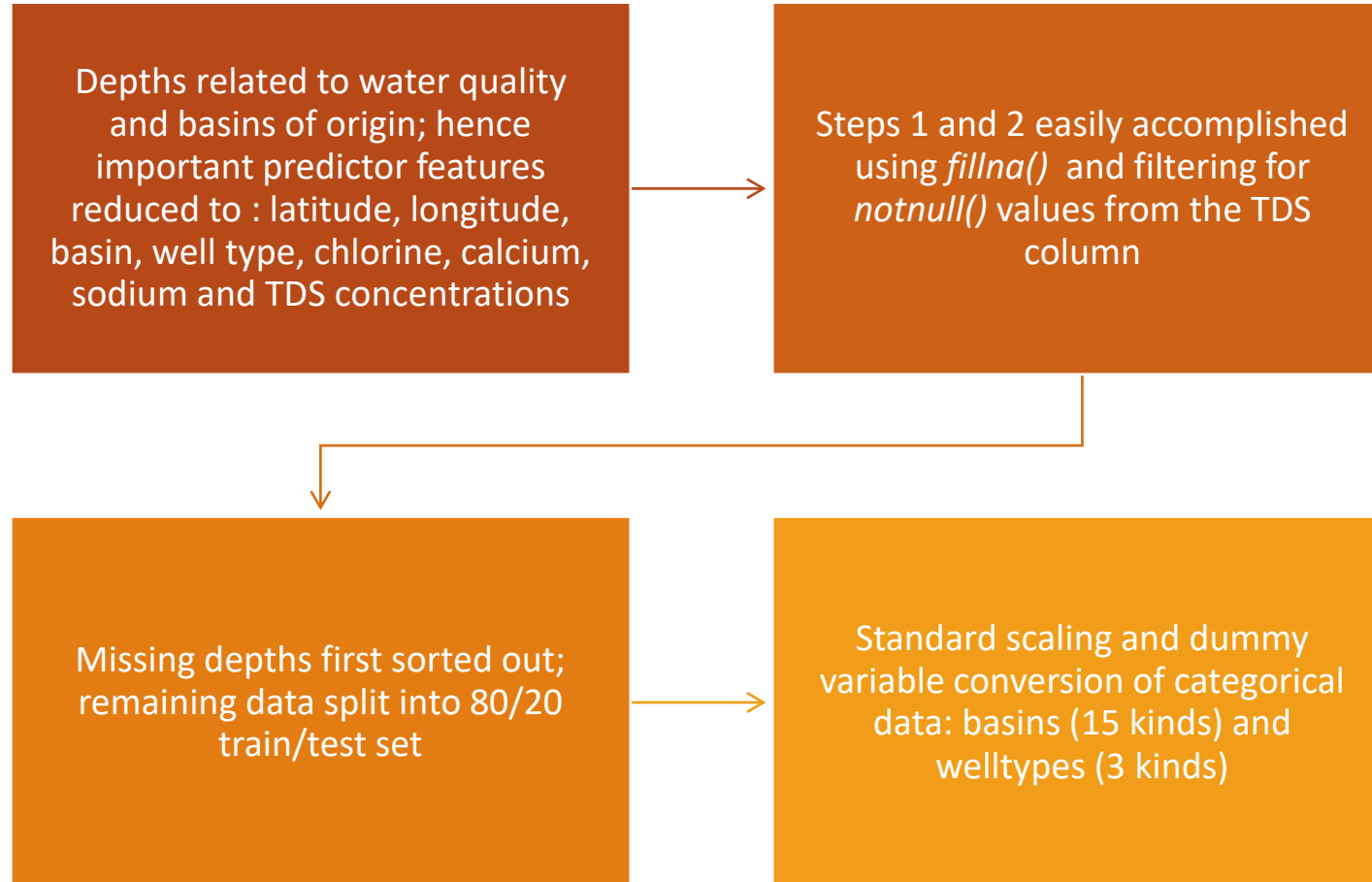


Separate missing depths data out, and create a test/train dataset with remaining variables.



Test, evaluate and tune different linear regression models to solve for DEPTHUPPER using relevant variables.

Separating Train/Test Data and Model Fitting

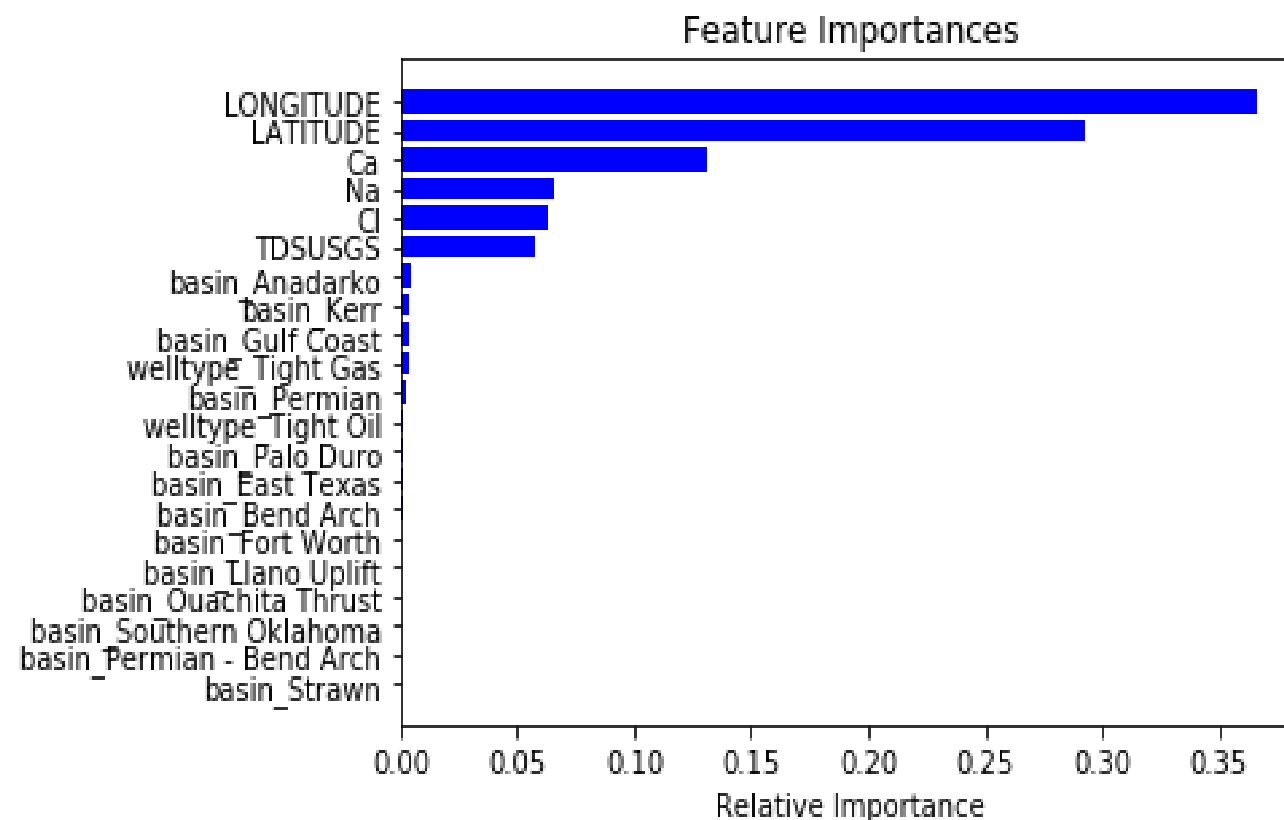


Algorithm	Parameters	Training R^2 value	Test R^2 value	RMSE
Ordinary Least Squares	Default	0.098	0.098	2789.5
Lasso Regression	alpha = 0.001, max_iter = 10000	0.098	0.098	2789.5
Ridge Regression	alpha = 0.001, max_iter = 10000	0.098	0.098	2789.5
Stochastic Gradient Descent Regression	max_iter=1000, tol=1e-4, eta0=0.1	0.091	0.089	2803.0
Linear Support Vector Machines Regression	epsilon=.01	-0.236	-0.237	3266.5
Polynomial Support Vector Machines Regression	kernel='poly', degree=2, C=100, epsilon=1	0.060	0.064	2842.2
Decision Tree Regression	max_depth=15	0.861	-0.139	3134.8
Random Forest Regression	max_depth=20, n_estimators=100	0.947	0.428	2221.3
Gradient Boosting Regression	max_depth=20, n_estimators=50, learning_rate=1	1.000	-0.213	3235.7
k-Nearest Neighbors Regression	n_neighbors=5	0.680	0.468	2142.2

Algorithms for Regression for Scaled Parameters and their Scores

Feature Selection

- Feature selection performed using the random forest regressor's in-built feature_importances_ attribute
- Only about 6 features have any major effects on the target depth variable
- Running RF regression with only these 6 variables gave similar results, i.e. train score = 0.948, test score = 0.426, RMSE = 2224.8



Algorithms for
Regression for
Unscaled Parameters
and their Scores

Standardization of the latitudes and longitudes doesn't make much sense, hence the unscaled datasets were run through the top regressors of scaled data, i.e. Random Forest, k-NN and gradient boosting regressors.

Algorithm	Parameters	Training R^2 value	Test R^2 value	RMSE
Random Forest Regression	max_depth=25, n_estimators=100	0.958	0.745	1483.8
Gradient Boosting Regression	max_depth=20, n_estimators=50, learning_rate=1	1.000	0.497	2083.9
k-Nearest Neighbors Regression	n_neighbors=5	0.410	0.122	2752.7



Tuning Hyperparameters on Random Forest Model

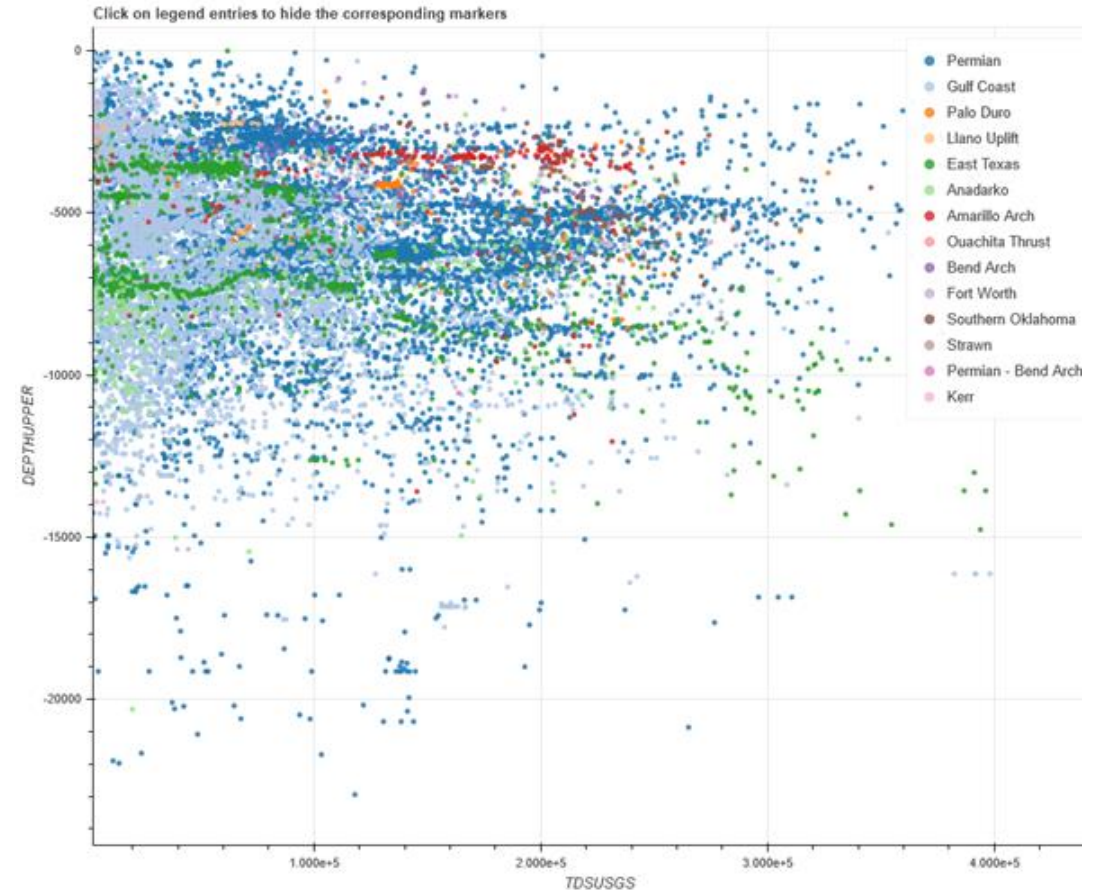
- 5-fold cross-validation run on best model (RF)
- Best Parameters Results:
 - Max_depth = 30
 - N_estimators = 150
- Best Score:
 - Train score = 0.960
 - Test score = 0.745
 - RMSE = 1478.4

Conclusions and Future Work

- i) Post-Oil Shale Revolution produced waters were significantly higher in TDS concentrations from the pre-revolution waters, found using bootstrapping techniques
- ii) Given a R^2 score of 0.75 on the test data, the missing depths were able to be filled in through a Random Forest Regression model

Future Work:

- Gather water quality data after the shale oil revolution, and connect previously drilled wells with newer frac projects on the same well paths
- Make predictions given newer horizontal drilling technologies, their completion methods and contribution of different frac stages to quality of flowback water



Interactive Bokeh plot (please refer to Jupyter notebook) of depths versus TDS concentration distribution by basins