Keenan Flynn

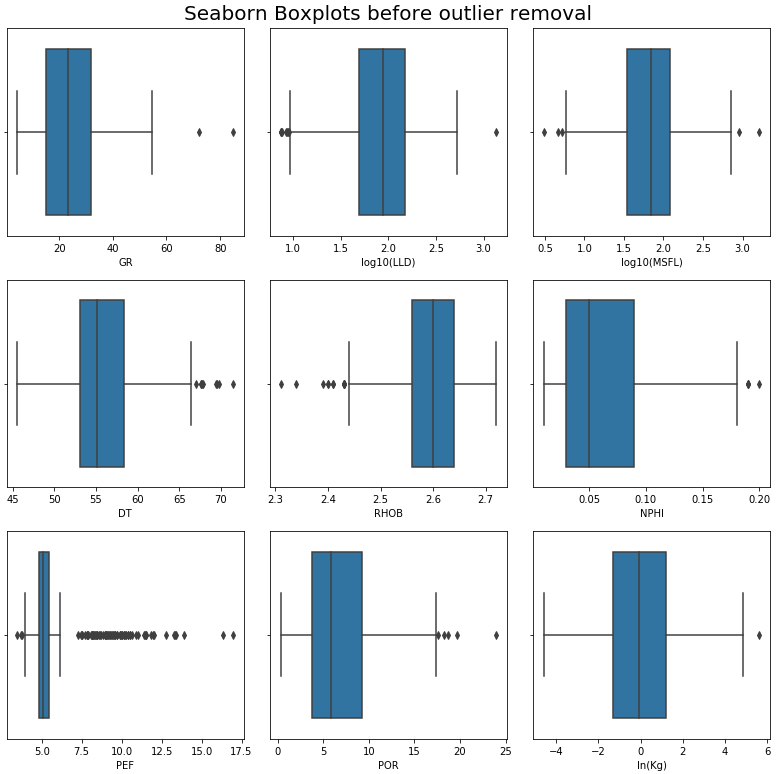
PETE 4161

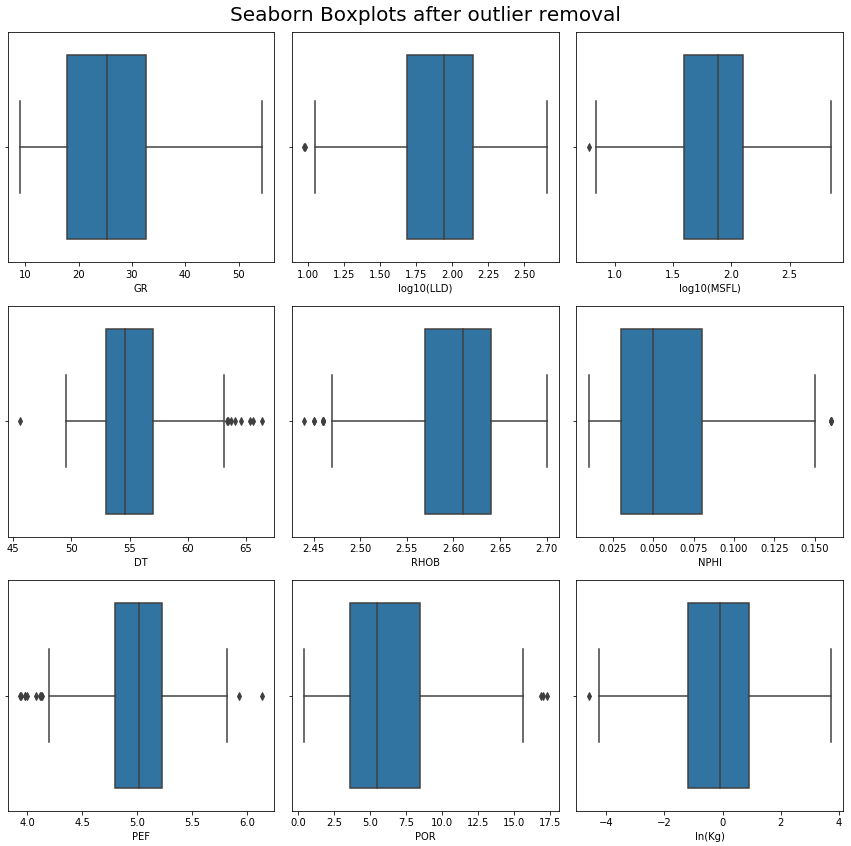
Salt Creek Data Exploration & Reduction, Classification, and Regression

In the textbook Applied Statistical Modeling and Data Analytics, the dataset ‘Salt-Creek.DAT’ is used as a vehicle to convey various methods of data modeling. In this report, we will similarly be examining and exploring the well logs from Salt Creek field. The data file contains 403 rows and 10 columns of well log data. The columns in order are ‘GR, log10(LLD), log10(MSFL), DT, RHOB, NPHI, PEF, POR, Kg, ln(Kg)’. The file contains the typical suite of well logs: Gamma Ray, Induction, Density, Photoelectric Absorption, and Sonic. These are our independent variables. Dependent porosity and permeability information is also given. Well log analysis is a particularly vital aspect of petroleum engineering as it is one of the main tools used to evaluate economic zones in a well. Historically this examination has been done by specialized professionals who manually interpret the well logs for hydrocarbon bearing strata. In this report, we will use statistical analysis, classification, and regression to gain a computer driven insight into the Salt Creek dataset.

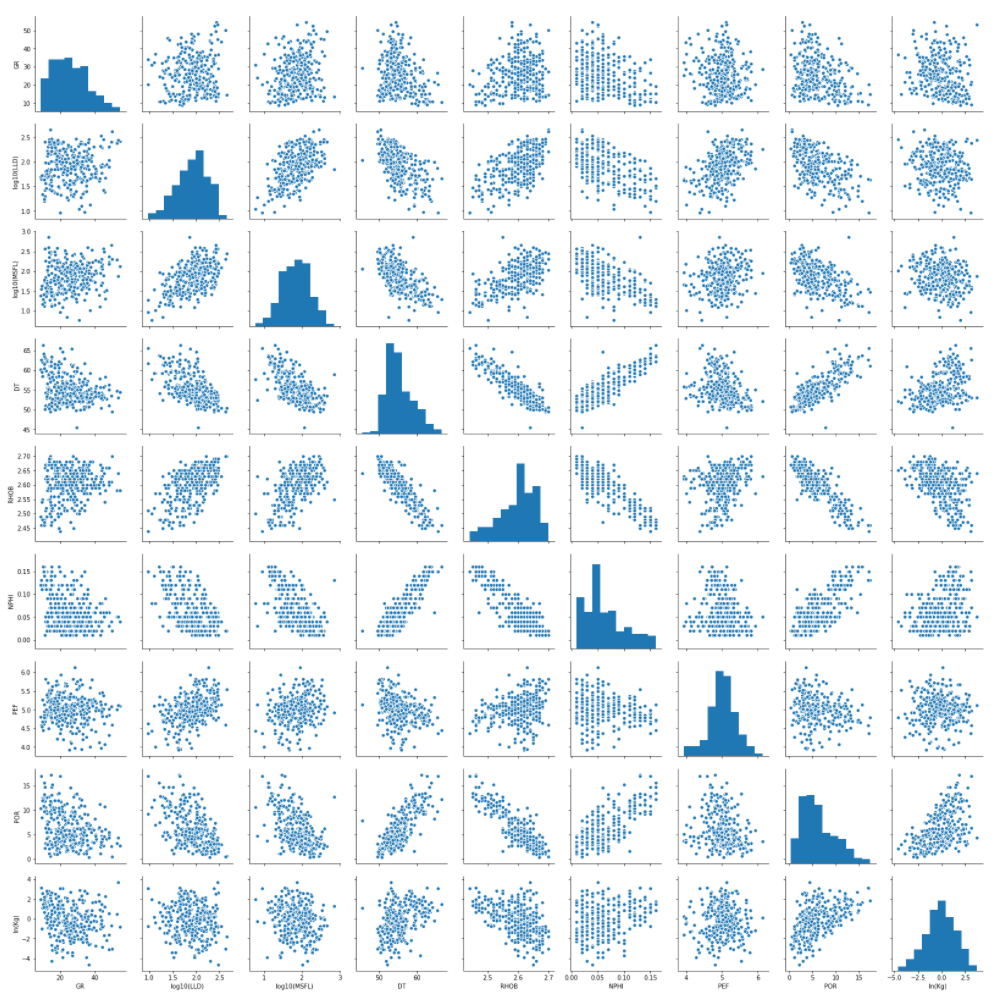
The first step in analyzing the data is to get our data into a usable format. We can do this by reading the DAT file into a Pandas DataFrame object. The datatype of each column is a 64 digit float, which is important to know when performing mathematical operations on the data later. With the datatype information, we can drop irrelevant columns which include the variable ‘Kg’ because we already have this information in the form of ‘ln(Kg)’. I checked to see if there were any duplicate rows, however there were none. There were also no null values.

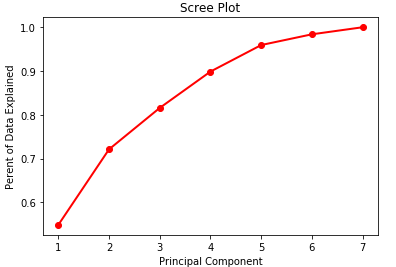
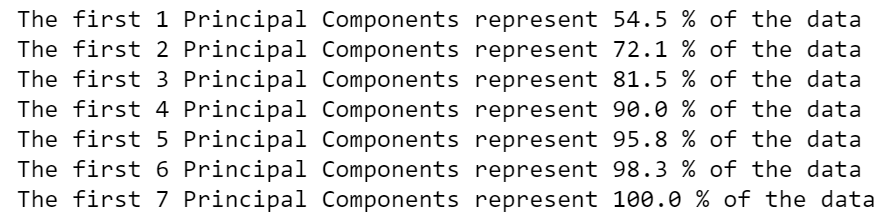
The next step in exploratory data analysis was to remove outliers. We can visualize what kind of outliers we have with boxplots and we can determine which data to remove using the Interquartile range method (IQR). The first and third quartile was taken of the aggregate data, and IQR calculated as third quartile minus first quartile. If a sample of data was not within +/- 1.5 times the IQR, it was removed. 1.5 is a common number to use in this method, but if we want fewer or more outliers removed, we can tune this number. In this case, 403 rows of data were initially in the DataFrame, and after removal of outliers there were 286.





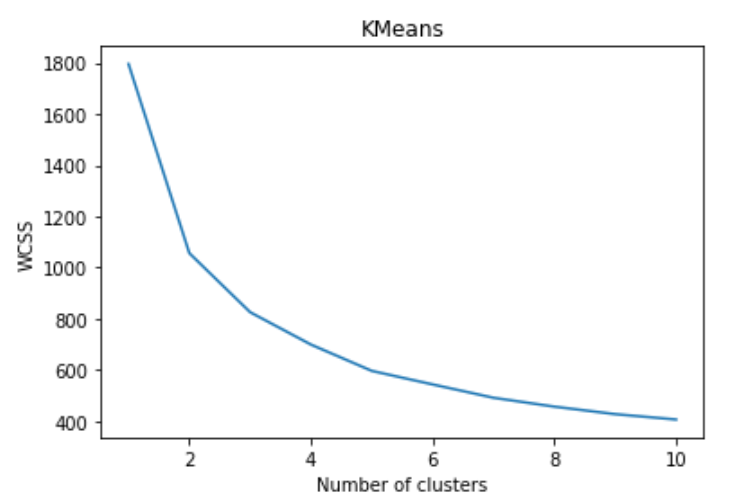
The final step in data exploration is visualizing correlations between the data. With Seaborn’s pairplot we can see relationships between variables. In the case of salt creek, when examining the diagonal of the pair plot, all distributions are either gaussian or log normal.



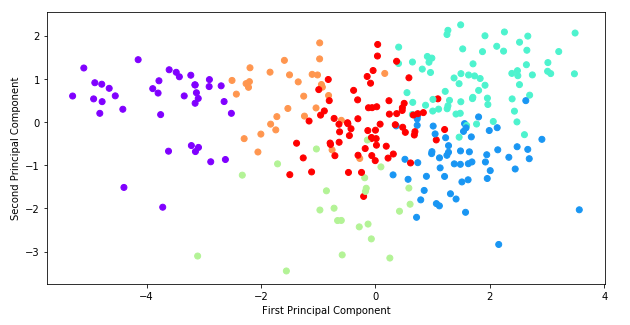
The salt creek data is now ready for dimensional reduction using the PCA method. This method is used rather than dropping columns because of the minimal variance lost. After standardizing the data, a covariance matrix was found. This eigenvalues and eigenvectors of this “C” matrix were then found and sorted so that we can examine the total variance explained of each eigenvector/eigenvalue pair or principal component. The results of this examination are as follows, in the form of text and in the graphical form of a scree plot.

To determine how many principal components to keep, I used a combination of the elbow method in the case of the scree plot and an abstract minumum value of 90% variance explained. This led me to choose the first four principal components as the components to keep, with the first 2 components explaining about two-thirds of all data. The data can now be classified.

K-means clustering was used to classify all 4 principal component data into groups. This algroithm was provided through scikit learn’s cluster library. The number of clusters chosen was 6. This was chosen using WCSS method. WCSS plotted against the number of clusters.

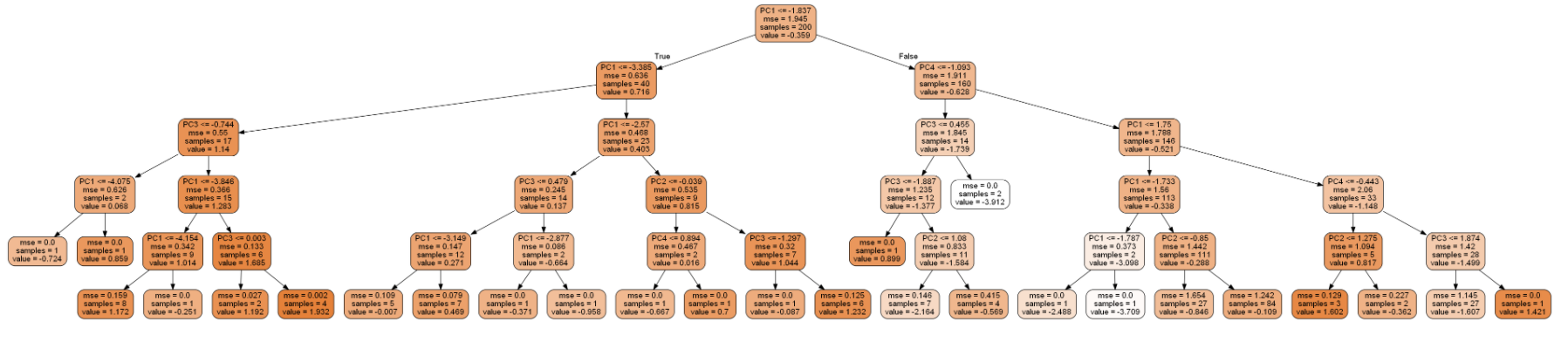


The evaluation of how many clusters to choose was done using the elbow method and 6 clusters were chosen. The 6 clusters were visualized along the first and second principal components.

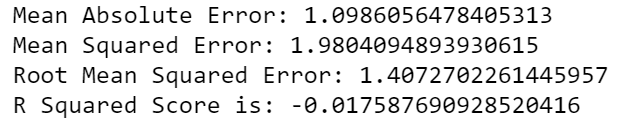


The principal components of the Salt Creek set were finally used to predict permeability (or ‘ln(Kg)’) by means of a regression tree. The tree was created using scikit learn’s ‘DecisionTreeRegressor’ function. A regression tree was chosen because our predictive values are continuous and quantitative rather than qualitative. To achieve qualitative predictions the ‘DecisionTreeClassifier’ function could replace the regression tree function and the code would work without error.

The data was split into 30% test and 70% training data. Our ‘x’ data is the input from our well logs and the ‘y’ is the variable we want to predict, being either porosity or permeability. Our criterion for the split is mean square error and our hyperparameter is a max depth of 5. The result is the tree below.



We can use the tree to get an array of predicted ‘y’ values. These predicted values were then evaluated using 4 different metrics: Mean absolute error, Mean Squared error, Root mean square error and R squared.



In conclusion, the unsupervised (Kmeans clustering) and supervised (regression tree) methods described above have created a model that can both classify data into groups and predict values given inputs that have been rigorously cleaned and transformed. The next step in this process would be to further tune hyperparameters by use of ensemble methods and further refine the clustering, whether improving the Kmeans algorithm or testing Kmeans against other types of clustering.