INTRODUCTION

Production from unconventional petroleum reservoirs includes petroleum from shale, coal, tight-sand and oil-sand. These reservoirs contain enormous quantities of oil and natural gas but pose a technology challenge to both geoscientists and engineers to produce economically on a commercial scale. These reservoirs store large volumes and are widely distributed at different stratigraphic levels and basin types, offering long-term potential for energy supply. Most of these reservoirs are low permeability and porosity that need enhancement with hydraulic fracture stimulation to maximize fluid drainage. Production from these reservoirs is increasing with continued advancement in geological characterization techniques and technology for well drilling, logging, and completion with drainage enhancement. Currently, Australia, Argentina, Canada, Egypt, USA, and Venezuela are producing natural gas from low permeability reservoirs: tight-sand, shale, and coal (CBM). Canada, Russia, USA, and Venezuela are producing heavy oil from oil sands. USA is leading the development of techniques for exploring, and technology for exploiting unconventional gas resources, which can help to develop potential gas-bearing shales of Thailand.

In this project, I will investigate a dataset containing examples of the geological features of unconventional reservoirs such as 'Porosity (%)', 'Acoustic impedance (kg/m2s*10^6)', 'Brittleness Ratio', 'Vitrinite Reflectance (%)' and understand their relationship among each other, their significance and their relationship with the production of the reservoir $A\sqrt{K}$ which is simply the cross-sectional area multiplied by the square root of permeability.

DESCRIPTION OF THE DATASET

The Dataset has the following features 'Porosity (%)', 'Acoustic impedance (kg/m2s*10^6)', 'Brittleness Ratio', 'Vitrinite Reflectance (%)' and the target feature is the $A\sqrt{K}$ *Aroot(K)* as shown in Figure 1

	Porosity (%)	Matrix Perm (nd)	Acoustic impedance (kg/m2s*10^6)	Brittleness Ratio	TOC (%)	Vitrinite Reflectance (%)	Aroot(K)
0	8.456	292	3.080	97.680	4.64	1.848	48.306469
1	8.666	353	3.542	55.404	3.56	1.504	41.300912
2	9.814	259	4.411	87.360	3.56	2.176	49.688356
3	12.369	675	2.893	47.772	4.32	1.504	59.132694
4	12.264	457	3.498	13.128	6.04	1.520	39.503121

Figure 1 : Sample of the geological dataset

Porosity is a measure of the void spaces in a material, and is a fraction of the volume of voids over the total volume, between 0 and 1, or as a percentage between 0% and 100%

Matrix Perm "Matrix Permeability" is one of the most important parameters for characterizing a source rock reservoir and for predicting hydrocarbon production. The low permeability value and the presence of induced fractures during core retrieval and transportation make the accurate measurement of the true permeability values for source rocks a significant challenge for the industry

Acoustic impedance are measures of the opposition that a system presents to the acoustic flow resulting from an acoustic pressure applied to the system.

Brittleness Ratio is the ratio of uniaxial compressive strength to tensile strength.

TOC "Total Organic Carbon" is the amount of carbon found in an organic compound and is often used as a non-specific indicator of water quality or cleanliness of pharmaceutical manufacturing equipment. TOC may also refer to the amount of organic carbon in soil, or in a geological formation, particularly the source rock for a petroleum play;

Vitrinite Reflectance is the proportion of incident light reflected from a polished vitrinite surface.

$\boldsymbol{A\sqrt{K}}$ is a productivity metric obtained from rate transient analysis (RTA) in unconventional reservoirs and it is equivalent to kh in conventional reservoirs. AOK is simply the cross-sectional area multiplied by the square root of permeability

From the dataset we can observe that all features including the target / output feature is numeric and that there are no categorical features.

PLAN

Python, seaborn, pandas and numpy will be the tools used to conduct the Exploratory Data Analysis, First we will use the .describe() function to understand the dataset more, know the mean, standard deviation, minimum values, maximum values, medians for each features.

Techniques to apply

- Histogram and density plots to check if there are skewed data
- Boxplots to check features that might have outliers
- Scatter plots will be used to understand the relationship between a feature and another
- Heatmap plot to check if there are any features that are linearly correlated

INSIGHTS

From the Histogram and Density plots shown in Figure 2, we can see that all features including the target feature are normally distributed, however Matrix Perm (nd) has a positive skew, Brittleness ratio and TOC has a negative skew

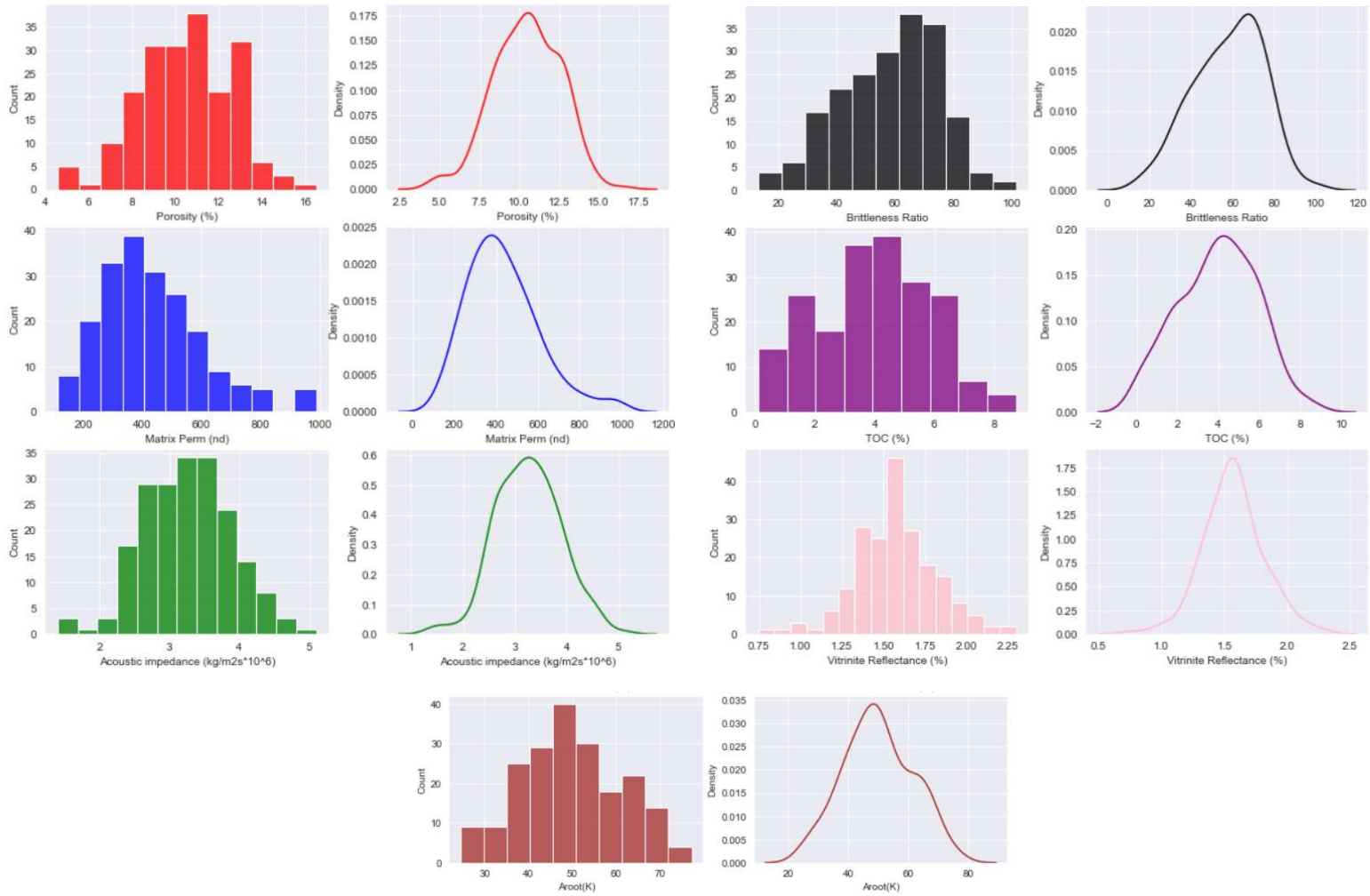

Figure 2 : Histogram and Density plots of all features

From the Boxplot as shown below in Figure 3, Outliers exist in the Matrix Permeability feature as well as the Vitrinite Reflectance Feature, this can be treated by transforming the features.

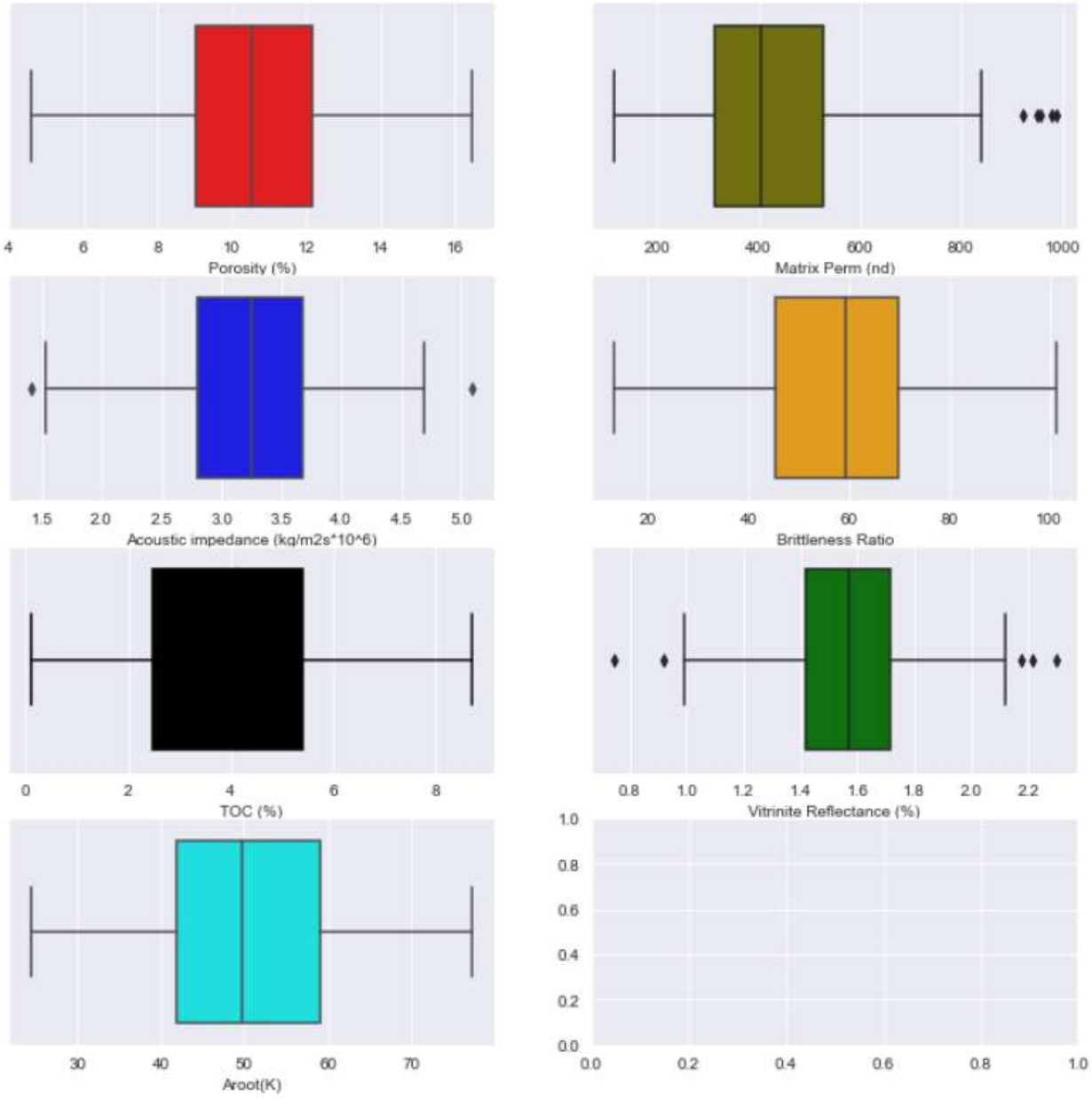

Figure 3 : Box plots of all features

From the pair plot shown in Figure 4, we can observe that Porosity, Matrix Perm are strongly correlated with the target feature Aroot(K), we can also observe that Porosity and Matrix Perm are linearly correlated, TOC and Porosity are also linearly correlated, this means that we can drop the Matrix Perm and TOC columns as they will not provide significant value

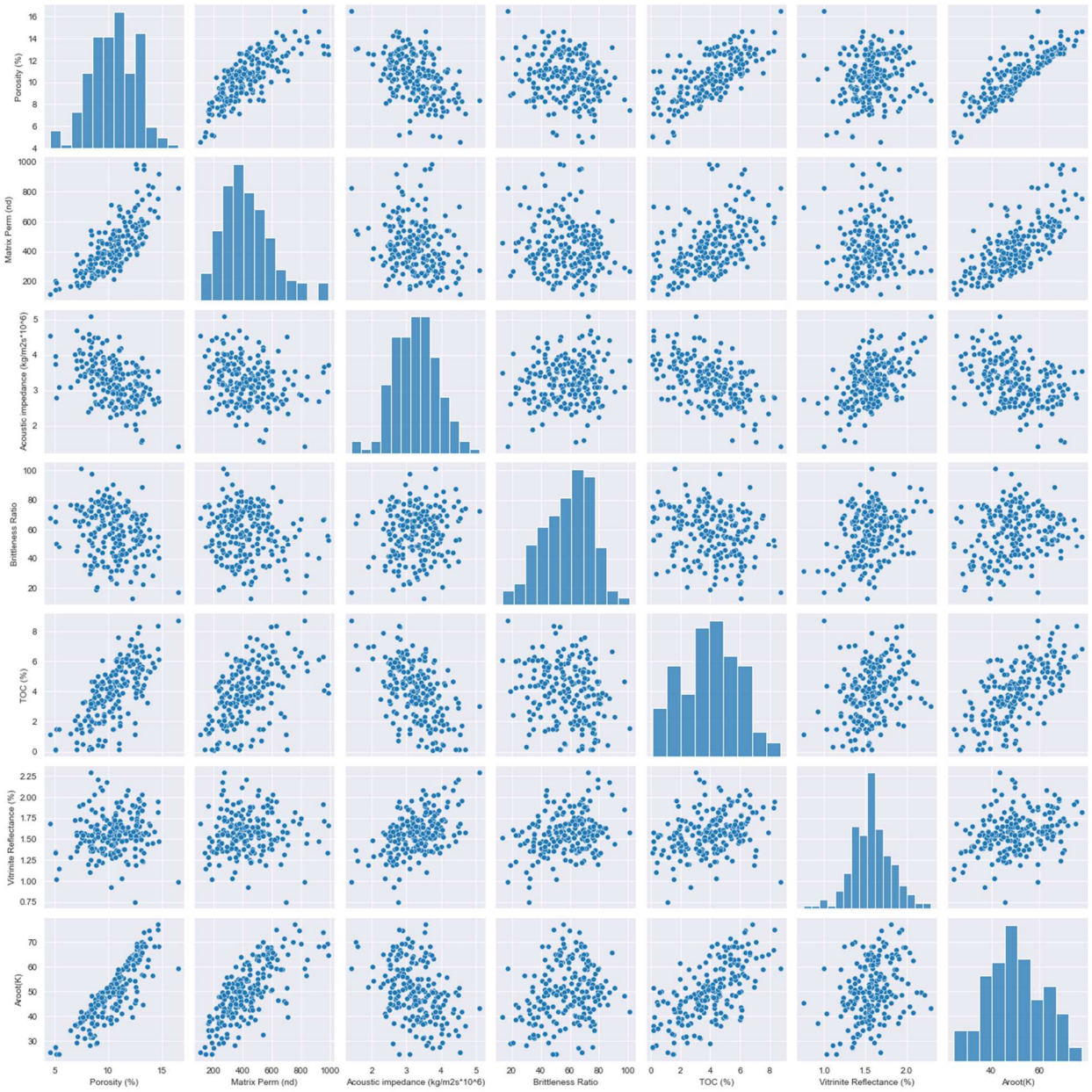

Figure 4: Pair plot of all features

DATA CLEANING & FEATURE ENGINEERING

As observed in the pair plot Porosity and Matrix Perm are linearly correlated, TOC and Porosity are also linearly correlated, However we want to quantify the Correlation thus we will use the pearson correlation, a threshold we will set is that features that have a person number above 0.7 will be considered linearly correlated and can be dropped from the dataframe.

Using the Heatmap function from the seaborn library as shown in Figure 5, we can observe that TOC and Porosity have a pearson correlation number of 0.71, and Matrix Perm and Porosity have a pearson correlation number of 0.76, this means these features are linearly correlated and we can drop any of these features.

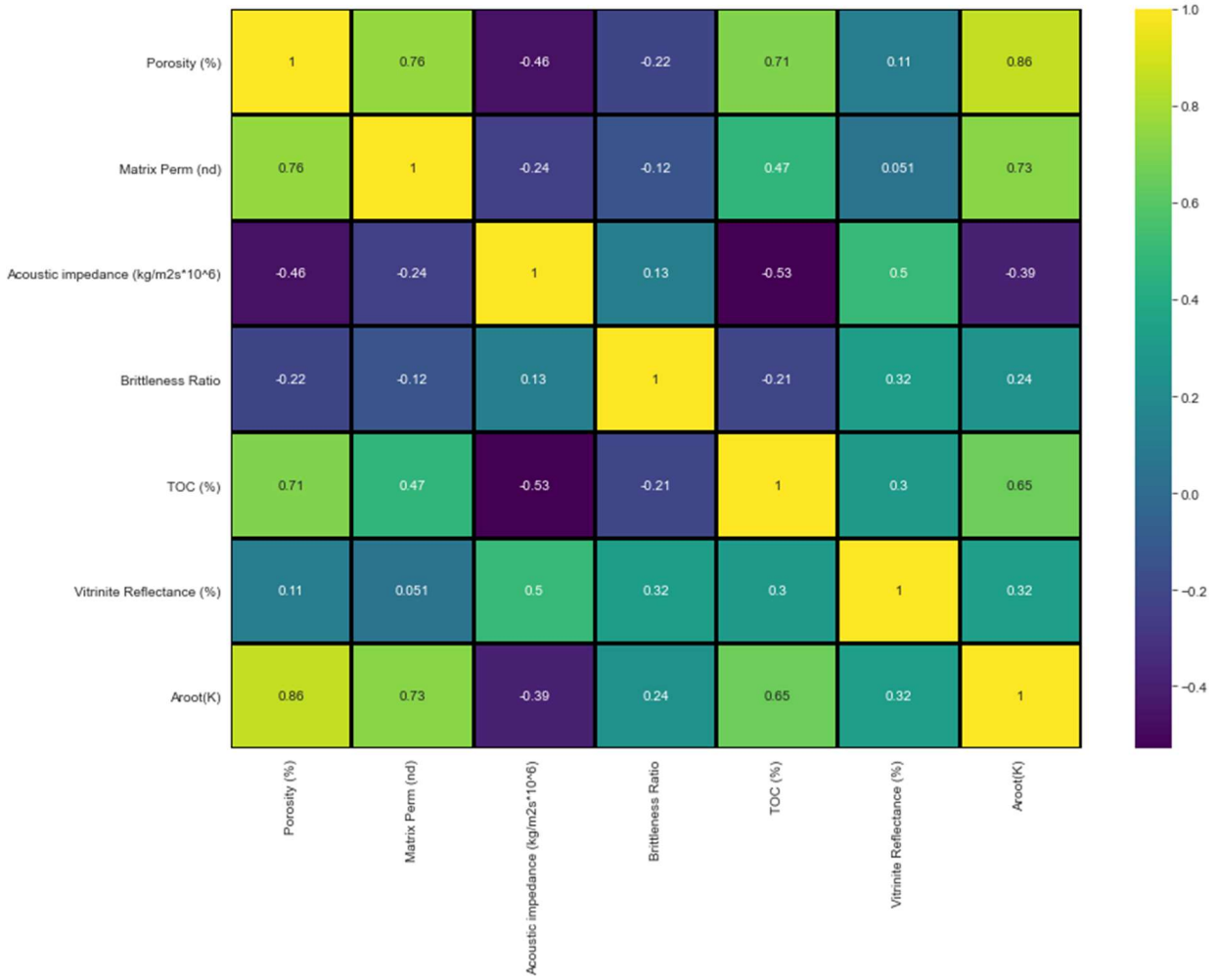

Figure 5: Heat map of all features

Dropping Features

We will drop the Matrix Permeability and the TOC feature as they are already linearly correlated with the Porosity feature

```
Matrix_Perm, TOC = df.columns[1], df.columns[4]

df.drop([Matrix_Perm, TOC], axis = 1, inplace = True)
```

```
df.head()
```

	Porosity (%)	Acoustic impedance (kg/m2s*10^6)	Brittleness Ratio	Vitrinite Reflectance (%)	Aroot(K)
0	8.456	3.080	97.680	1.848	48.306469
1	8.666	3.542	55.404	1.504	41.300912
2	9.814	4.411	87.360	2.176	49.688356
3	12.369	2.893	47.772	1.504	59.132694
4	12.264	3.498	13.128	1.520	39.503121

We can observe that the new set of features are:

1. Porosity
2. Acoustic Impedance
3. Brittleness Ratio
4. Vitrinite Reflectance

Scaling

We can also observe that we need to scale the features as there are big differences between the values of Brittleness Ratio column and Vitrinite Reflectance column.

We will use the MinMax scaler from the sklearn library to scale the features so that they are between the values (0 and 1)

```
from sklearn import preprocessing
```

```
scaler = preprocessing.MinMaxScaler(feature_range=(0,1))
scaler.fit(df)
df_scaled = scaler.transform(df)
df_scaled = pd.DataFrame(df_scaled, columns=[df.columns[0], df.columns[1], df.columns[2], df.columns[3], df.columns[4]])
```

```
df_scaled
```

	Porosity (%)	Acoustic impedance (kg/m2s*10^6)	Brittleness Ratio	Vitrinite Reflectance (%)	Aroot(K)
0	0.325294	0.453731	0.960076	0.711340	0.451776
1	0.342941	0.579104	0.480038	0.489691	0.319177
2	0.439412	0.814925	0.842894	0.922680	0.477932
3	0.654118	0.402985	0.393378	0.489691	0.656690
4	0.645294	0.567164	0.000000	0.500000	0.285149
...	...	...	...	...	...
195	0.317647	0.504478	0.766317	0.582474	0.382052
196	0.672941	0.626866	0.454830	0.592784	0.767007
197	0.327647	0.668657	0.628560	0.412371	0.285905
198	0.529412	0.358209	0.644638	0.731959	0.654256
199	0.843529	0.582090	0.480038	0.690722	1.000000

HYPOTHESIS TESTING

- H0 : Brittleness Ratio is highly correlated to the Aroot(K)

- H1 : Brittleness Ratio are not correlated to the Aroot(K)

We will use the scipy library to determine the p-value and the pearson correlation values of our hypothesis.

A 0.05 Significance level is set as a threshold.

```
# To find the correlation value and p-value

r, p = stats.pearsonr(df['Aroot(K)'], df['Brittleness Ratio'])
```

```
print(f"The Correlation value is {round(r, 4)}")
print(f"The P-Value is : {round(p, 4)}")
```

```
The Correlation value is 0.2372
The P-Value is : 0.0007
```

```
# Defining the Signifigance level and evaluation the hypothesis

significance_level = 0.05
if p < significance_level:
    print("Reject null hypothesis: Brittleness Ratio are not highly correlated to the Aroot(K)")
else:
    print("Accept null hypothesis: Brittleness Ratio are highly correlated to the Aroot(K)")
```

```
Reject null hypothesis: Brittleness Ratio are not highly correlated to the Aroot(K)
```

We can observe from the above code that the null hypothesis is rejected as the p-value (0.007) was less than the significance value (0.05).

NEXT STEPS

Our data is now ready to be used in training Machine Learning models, we can use the following dataset to predict the Aroot(K) by training a Multivariate Linear Regression model on the dataset.

SUMMARY

- Extracted insights through Exploratory Data Analysis on the Dataset
- Feature Engineered the Dataset by removing linearly correlated features & normalizing the features
- Formulated 2 hypothesis about the data and conducted significance test to evaluate the hypothesis by calculating p-values