**Project Report**

**Project Title**

**Estimated ultimate recovery prediction in Shale Reservoirs**

**Team Leader:**

A Harshith 21R11A0504

**Team Members:**

P Rohith 21R11A0537

G Harshit 21R11A0519

G Pranav 21R11A0515

D Nissi 21R11A0511

**Exploratory Data Analysis**

**Import data set**

To start Exploratory Data Analysis (EDA) and import the dataset, it's essential to include key libraries:

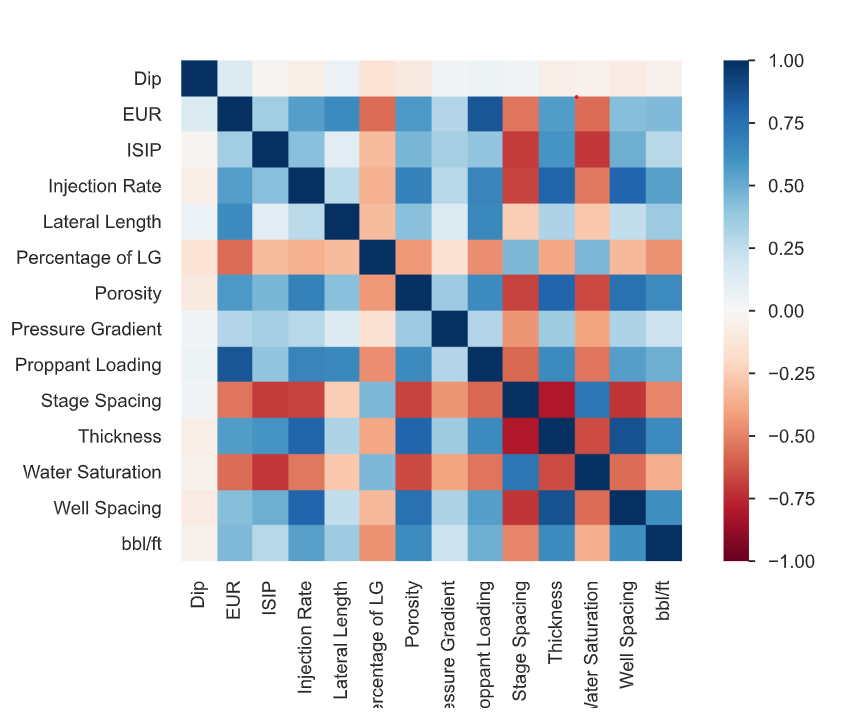
- Pandas: A library for data manipulation and analysis.

- NumPy: A fundamental package for numerical computing.

- Seaborn: A visualization library based on matplotlib for statistical plotting.

- Matplotlib: A comprehensive library for creating static, animated, and interactive visualizations in Python.

**1. Correlation**



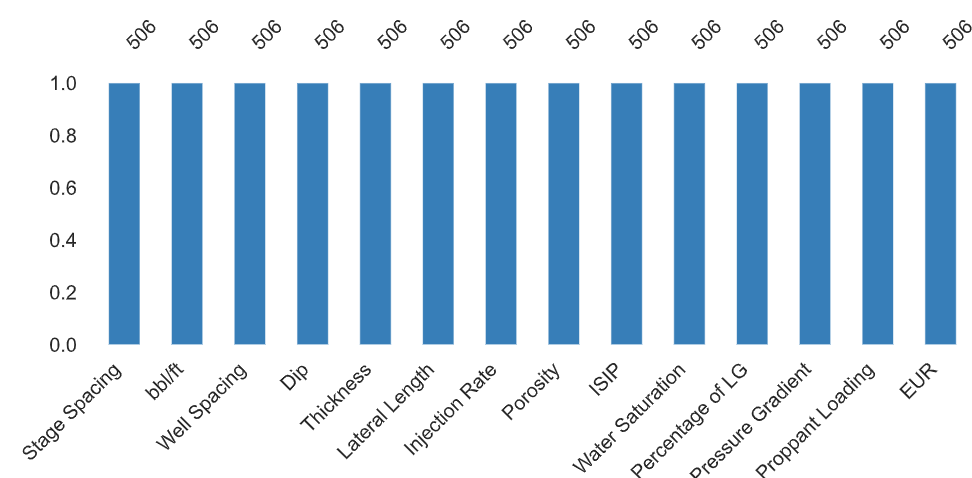
**Observations**

EUR and Proppant Loading (0.852): A robust positive correlation exists between estimated ultimate recovery (EUR) and the quantity of proppant introduced into the well during hydraulic fracturing. This indicates that greater proppant loading correlates with higher anticipated recovery rates.

Stage Spacing and ISIP (-0.698): A significant negative correlation is observed between stage spacing (the distance between clusters during fracturing) and initial shut-in pressure (ISIP). This implies that reducing stage spacing tends to elevate ISIP.

Water Saturation exhibits negative correlations with various factors such as EUR, Injection Rate, and Proppant Loading. This implies that lower levels of water saturation may enhance recovery and fracturing effectiveness.

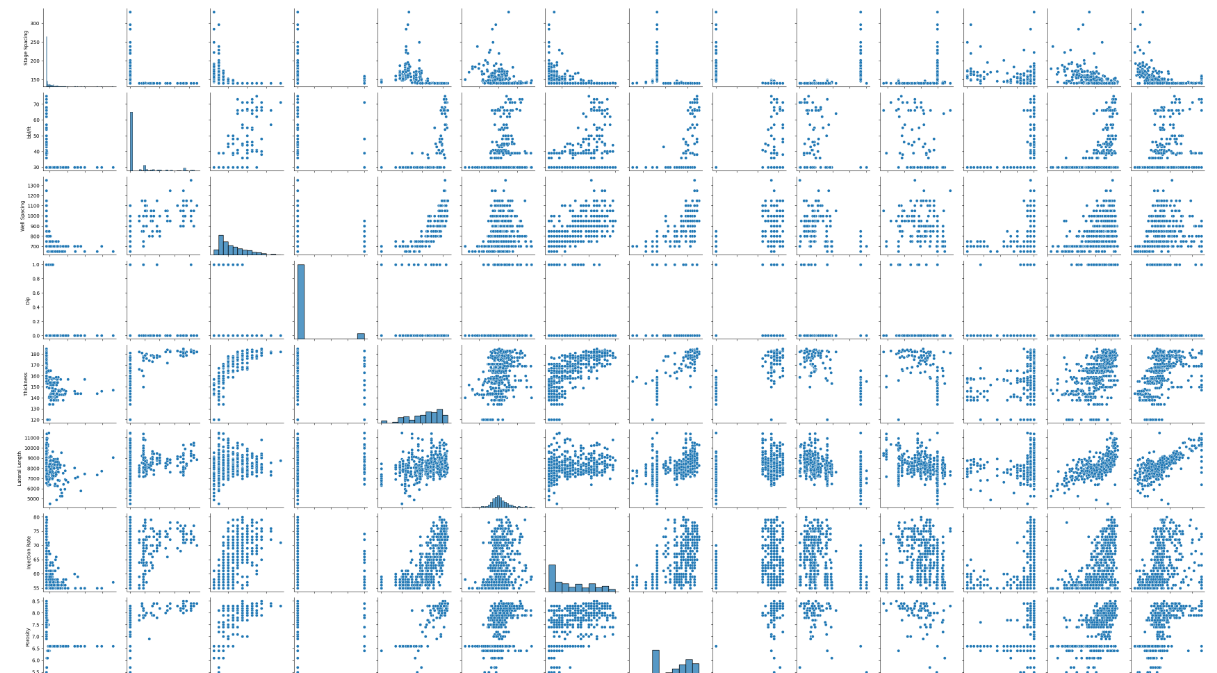
**2.Missing values**



**Observation:**

Addressing missing values is crucial as they can significantly affect the accuracy and reliability of analyses and models. It's essential to assess the extent of missing data in each variable before proceeding with any analysis or modeling. Strategies such as imputation or deletion should be considered based on the context and their impact on the analysis results. Ignoring or mishandling missing values can result in biased or misleading conclusions.

3**. Multivariate Data Analysis and Insights Derived from Scatter Plot Matrices**





**Observations**

The scatter plot matrix provides a comprehensive overview of the multivariate data, presenting scatter plots of each variable pair.

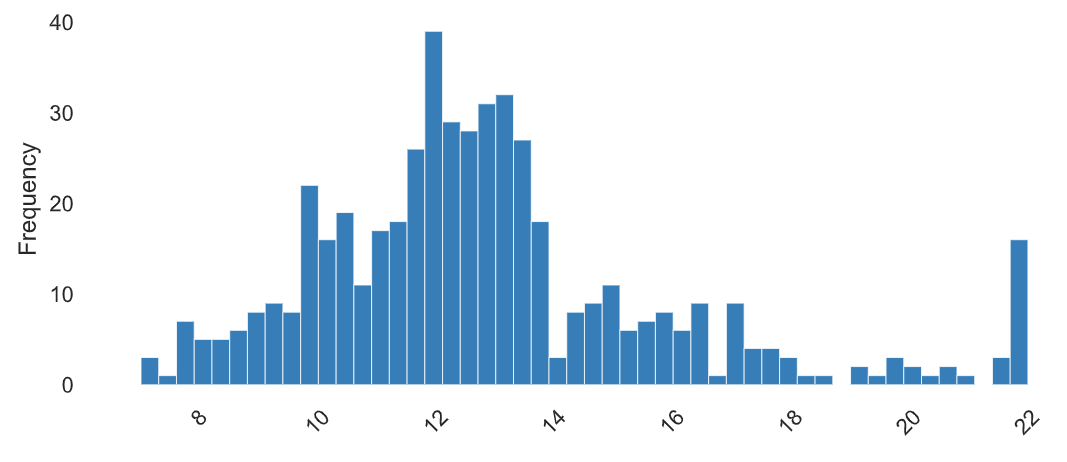
Positive Correlation between EUR and Proppant Loading: The scatter plot matrix reveals a clear positive correlation between Estimated Ultimate Recovery (EUR) and Proppant Loading. As proppant loading increases, EUR also tends to rise, suggesting that higher levels of proppant during hydraulic fracturing could lead to enhanced recovery from shale reservoirs.

EUR and Injection Rate Relationship: While there is a positive correlation between EUR and Injection Rate, it appears less pronounced compared to the relationship with proppant loading. This implies that higher injection rates may moderately boost estimated ultimate recovery.

Impact of Lateral Length on EUR: The pair plot illustrates a moderate positive correlation between Lateral Length and EUR. This indicates that longer lateral lengths in shale wells are linked to higher estimated ultimate recovery, aligning with industry knowledge that longer horizontal sections can enhance production.

Water Saturation and Recovery: The analysis reveals a negative correlation between Water Saturation and EUR. Lower water saturation levels correspond to higher estimated ultimate recovery, emphasizing the significance of managing water content in the reservoir for optimal recovery rates.

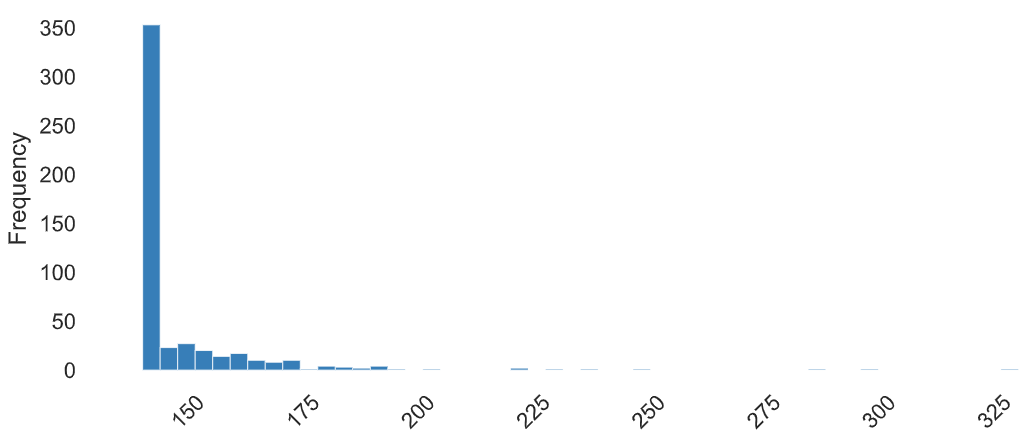
In summary, the pair plot analysis provides valuable insights into the relationships between various parameters and their influence on shale gas reservoir performance and recovery rates.

**4. Analysis of EUR** 

**Observation:**

The dataset exhibits a relatively narrow range, spanning from 7 to 22, with a median of 12.4 and a mean of 12.85. It demonstrates positive skewness (1.10) and a moderately peaked distribution with a kurtosis of 1.49. With a standard deviation of 3.07, there's moderate variability around the mean. The coefficient of variation is 0.24, indicating moderate relative variability. The interquartile range (IQR) is 2.7, suggesting a relatively compact middle 50% of the data. Non-monotonicity in the data suggests the presence of non-linear or complex patterns within its distribution.

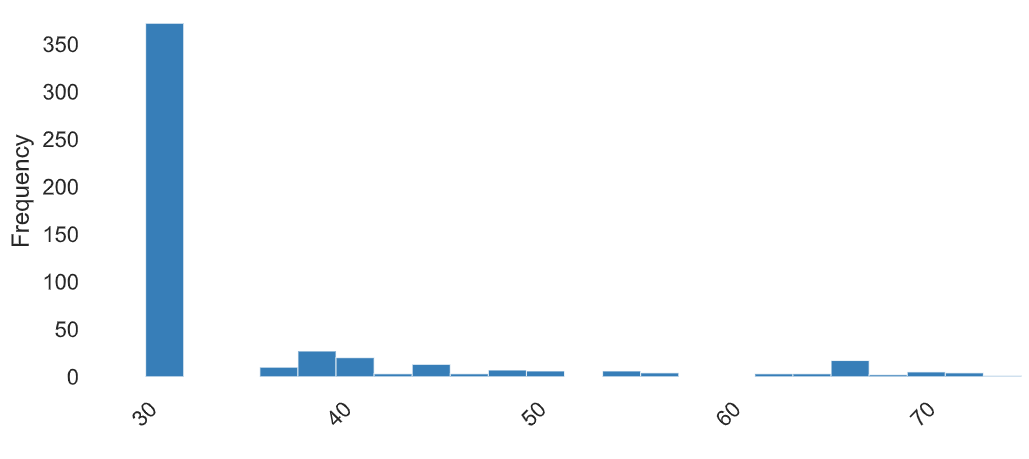
**5.State Spacing**



**OBSERVATION**

The dataset displays a broad range of values, spanning from a minimum of 140 to a maximum of 330. Its middle point, as indicated by the median, stands at 141. Meanwhile, the average value, represented by the mean, is 147.64, though this is slightly skewed by outliers on the higher end. This skewness is evident with a positive skewness value of 5.21, indicating a distribution leaning towards higher values. Furthermore, the kurtosis value of 36.99 suggests that the dataset has heavier tails and potentially more extreme values than a normal distribution. The standard deviation, at 18.39, illustrates notable variability within the dataset. When considering this variability relative to the mean, represented by the coefficient of variation at 0.12, it appears moderately variable. The interquartile range (IQR), measuring the spread of the middle 50% of the data, is relatively compact at 8. However, the non-monotonic nature of the data indicates the presence of complex or non-linear patterns within its distribution.

**6.** **Analysis of bbl/ft**

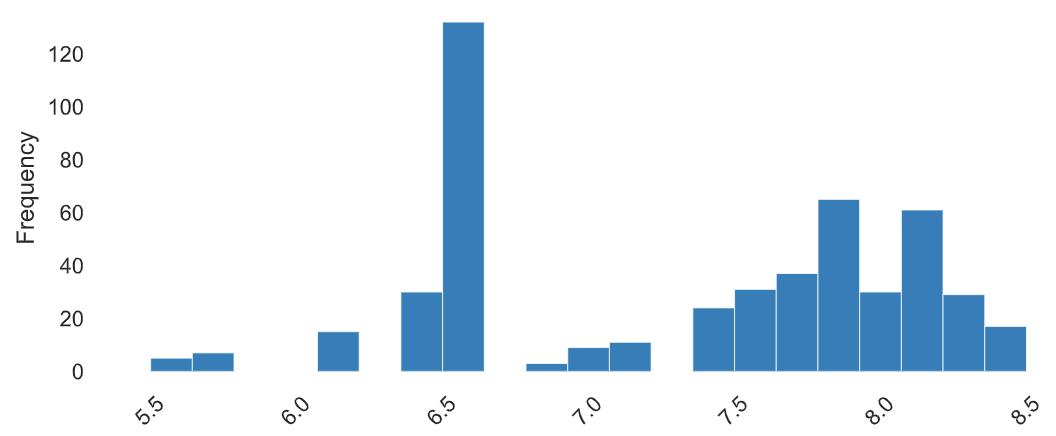


**Observation:**

The dataset spans a narrow range from 30 to 75, with the median, first quartile (Q1), and fifth percentile all at 30, and the third quartile (Q3) at 36. The mean slightly surpasses the median at 35.13, indicating a positive skewness of 2.23. With a standard deviation of 10.53, there's moderate variability around the mean. The coefficient of variation stands at 0.30, signaling moderate relative variability. The kurtosis of 4.04 suggests a moderately peaked distribution. The interquartile range (IQR) is 6, highlighting a relatively compact middle 50% of the data. Non-monotonicity in the data suggests the presence of non-linear or complex patterns within its distribution.

Top of Form

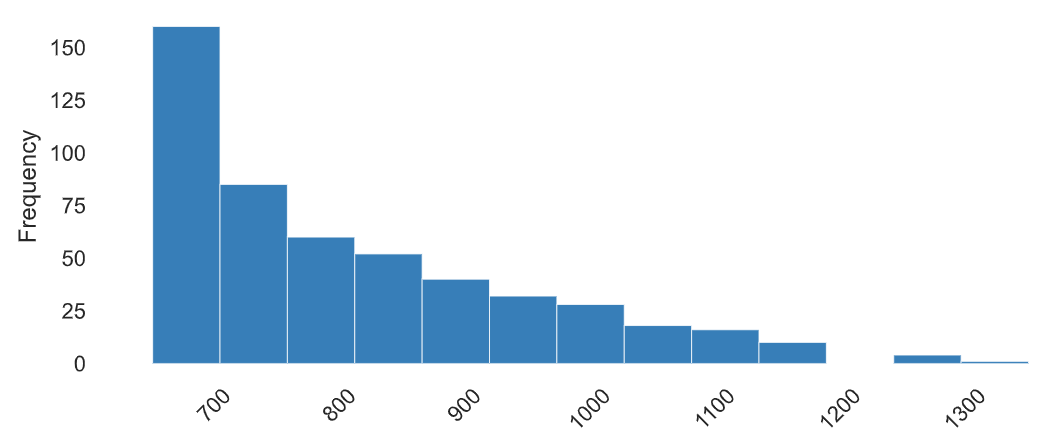
**7. Analyzing Porosity**



**Observations**

The dataset spans from 5.5 to 8.5, with a median of 7.5 and a mean of 7.34. It exhibits negative skewness (-0.32) and a moderately flat distribution with a kurtosis of -1.16. With a standard deviation of 0.75, there's relatively low variability around the mean. The coefficient of variation, at 0.10, indicates low relative variability. The interquartile range (IQR) is 1.4, suggesting a relatively compact middle 50% of the data. Non-monotonicity in the data suggests the presence of non-linear or complex patterns within its distribution.

**8. Comparative Analysis of well spacing**

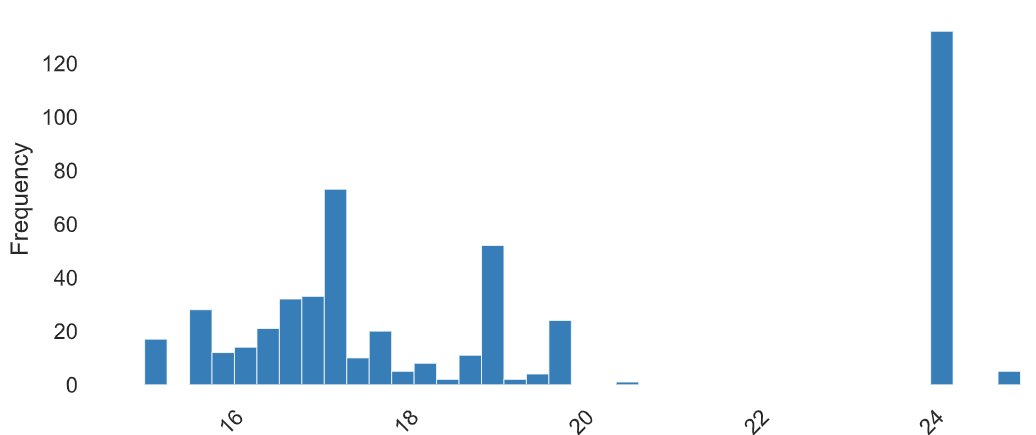


**Observations**

The dataset ranges from 650 to 1350, with a median of 800 and a mean of 820.16. It exhibits slight positive skewness (0.98). With a standard deviation of 135.74, there's moderate variability around the mean. The coefficient of variation stands at 0.17, suggesting moderate relative variability. The kurtosis of 0.40 indicates a distribution with a slightly flattened peak, resembling a relatively normal distribution. The interquartile range (IQR) is 200, indicating a relatively compact middle 50% of the data. Non-monotonicity in the data suggests the presence of non-linear or complex patterns within its distribution.

Top of Form

**9.Water Saturation**

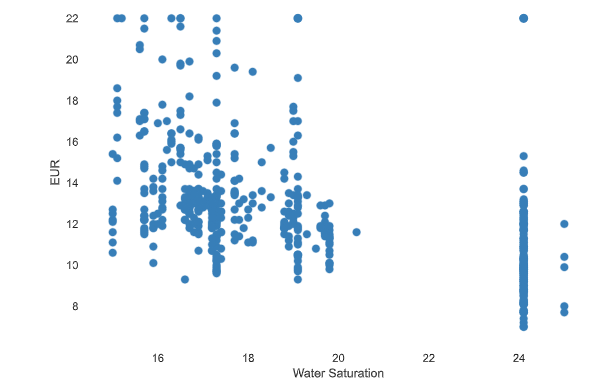


**Observation**

The dataset spans from 15 to 25, with a median of 17.7 and a mean of 19.21. It displays positive skewness (0.67) and a moderately flat distribution with a kurtosis of -1.14. With a standard deviation of 3.20, there's moderate variability around the mean. The coefficient of variation is 0.17, indicating moderate relative variability. The interquartile range (IQR) is 7.3, suggesting a relatively spread-out middle 50% of the data. Non-monotonicity in the data suggests the presence of non-linear or complex patterns within its distribution.

Top of Form

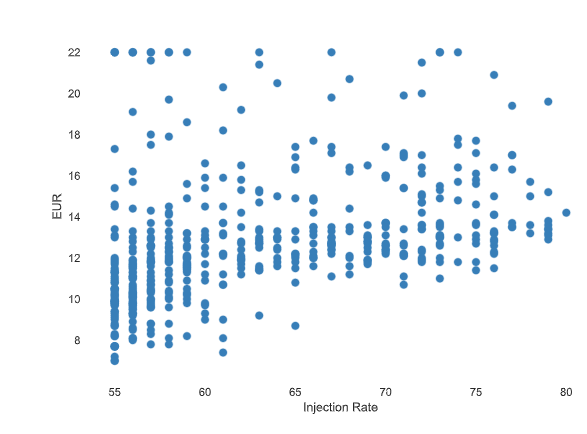
**10.Water Saturation and EUR**



**Observation:**

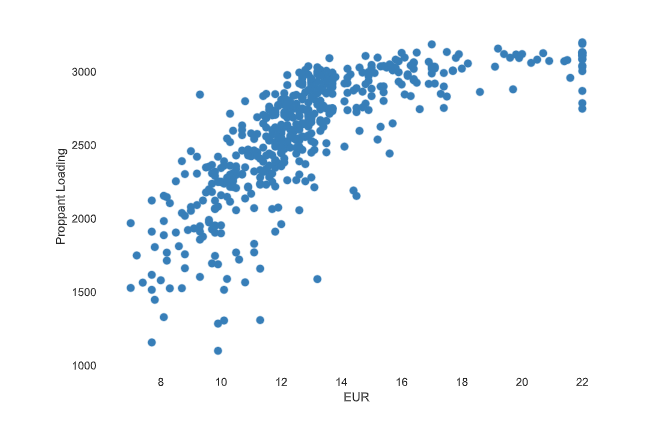
The negative correlation suggests that reducing water saturation levels could enhance recovery and improve fracturing efficiency. Exploring strategies to control water saturation becomes pertinent in optimizing production processes.

**11.Injection Rate and EUR**



**Observation:**

The moderate positive correlation observed between injection rate and EUR suggests that elevated injection rates might enhance recovery. This connection holds relevance in refining injection strategies to improve reservoir performance.

**12.Interactions between EUR and Proppant loading**

**Observation:**

This connection holds significant importance as it signifies that higher proppant loading correlates with amplified Estimated Ultimate Recovery (EUR). Emphasizing this correlation can provide valuable insights into the impact of hydraulic fracturing methods on reservoir performance.

**Modeling for the Estimated ultimate recovery prediction in Shale Reservoirs**

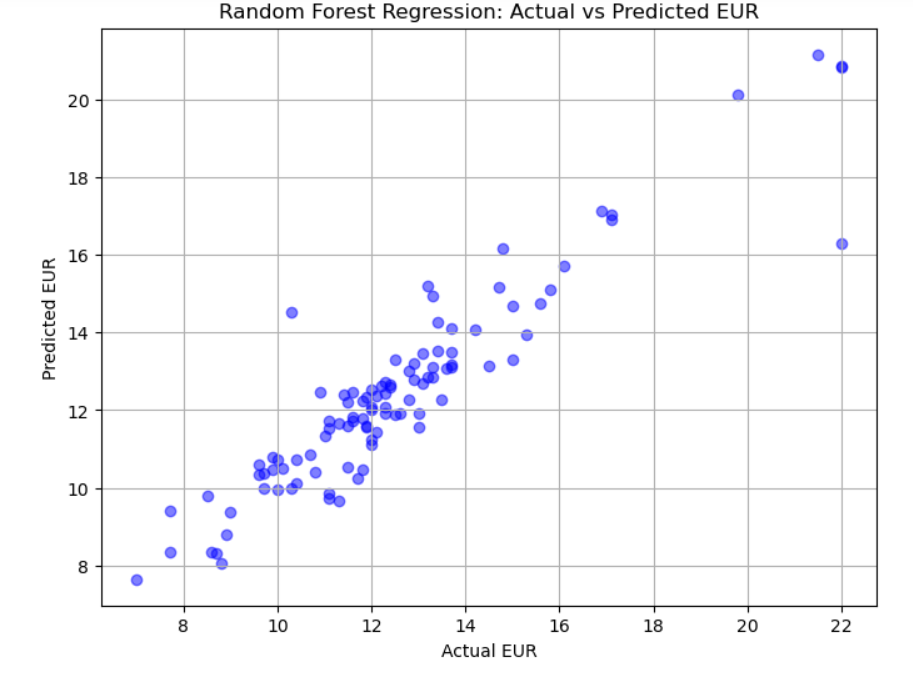
**1.Random Forest Regressor**

1. **feature importance**
2. **Neural Networks**

**1.RandomForestRegressor**

Target variable=’EUR’

Test size has been of the 0.2 meaning that trained size has 80% and test size has 20% of the data.



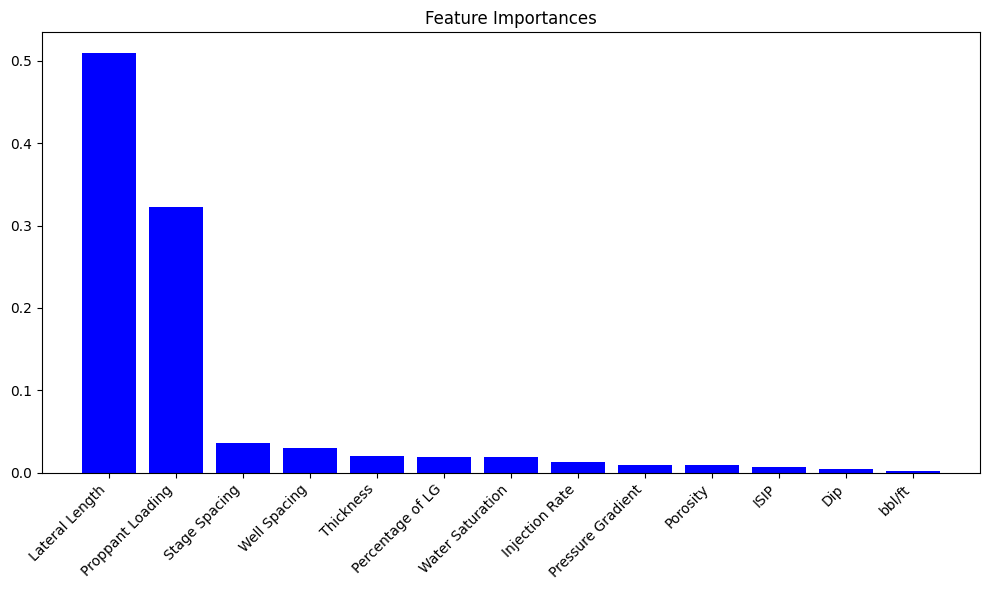
**Observation:**

The scatter plot provides a visualization of the performance of the Random Forest Regressor model in predicting actual values. Ideally, the points should align along a diagonal line, indicating perfect prediction where actual and predicted values match precisely. Any deviations from this line signify predictive errors of the model. A scattered or dispersed pattern observed away from the diagonal may indicate regions where the model encounters challenges in accurately predicting the target variable. This highlights potential areas for model enhancement or additional analysis.**Accuracy of the Random Forest Regressor:**

**Mean Squared Error: 1.0744639019607833**

**R^2 Score: 0.8677581481942844**

**Feature Importance**



Values:  
Feature Importance

0 Lateral Length 0.509277

1 Proppant Loading 0.322331

2 Stage Spacing 0.036085

3 Well Spacing 0.030172

4 Thickness 0.020174

5 Percentage of LG 0.019101

6 Water Saturation 0.018486

7 Injection Rate 0.012802

8 Pressure Gradient 0.009364

9 Porosity 0.008673

10 ISIP 0.007010

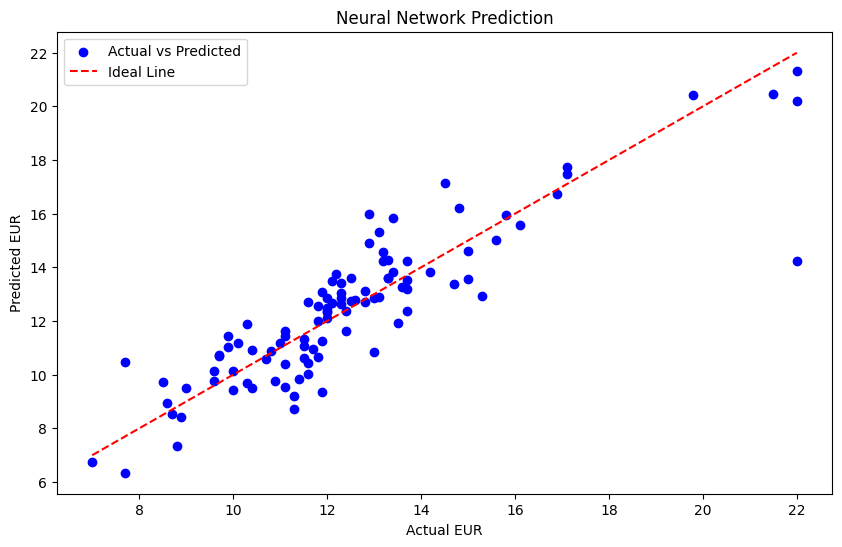
11 Dip 0.004078

12 bbl/ft 0.002448

**Observations:**

1. The bar graph displays the significance of various features, likely associated with geological or environmental data, with scores ranging from 0 to 1 on the y-axis.
2. "Lateral Length" emerges as the most crucial feature, with a score surpassing 0.5, indicating its substantial influence on outcomes.
3. Following "Lateral Length," "Proppant Loading" is notable as the second most influential feature, albeit with a lesser impact compared to the former.
4. Other features such as "Stage Spacing" and "Well Spacing" demonstrate minimal importance, suggesting their limited effect on outcomes.
5. The findings derived from this analysis can inform decision-making processes, highlighting the importance of optimizing "Lateral Length" and exploring strategies to enhance "Proppant Loading" for improved results.

**Neural Network**



**Observations**

The scatter plot above visualizes the performance of neural networks in predicting EUR values.

Interpreting the Scatter Plot: Each point on the plot represents an observation, comparing the actual and predicted EUR values. It provides a visual representation of the model's predictive accuracy.

Ideal Line Representation: The diagonal ideal line serves as a benchmark for perfect prediction. Any deviation from this line indicates prediction errors.

Assessing Accuracy: The dispersion of points around the ideal line reflects the model's accuracy. A tight cluster suggests precise predictions, while a wider spread implies variability in the model's performance.

Evaluating Model Performance: A well-fitting model is characterized by points closely aligned with the ideal line, indicating accurate predictions. Conversely, widely scattered points suggest inconsistent predictions and areas for potential improvement.

Further Examination: Analyzing residuals (the differences between actual and predicted values) offers deeper insights into the model's performance. Additionally, R-squared (R²) can quantify how well the model explains the variance in the data, with higher values indicating a better fit.

**Mean Squared Error: 1.8676594495773315**

**Total Observation:**

The Random Forest Regressor model demonstrated strong performance with a mean squared error of 1.074 and an R-squared score of 0.868. These metrics indicate a solid fit to the data and successful prediction of Estimated Ultimate Recovery (EUR) in shale reservoirs.

The analysis identified "Lateral Length" as the most influential feature for EUR prediction, followed by "Proppant Loading." Conversely, features such as "Stage Spacing" and "Well Spacing" had minimal impact, suggesting they could be given less priority in the modeling process.

In the scatter plot generated by the neural network model, the comparison between actual and predicted EUR values reveals areas of deviation, highlighting the model's accuracy and potential areas for enhancement.

The scatter plot analysis aids in evaluating the models' accuracy, with a tight cluster around the ideal line indicating precise predictions and a scattered pattern suggesting variability in predictions.