**Project Report**

**Project Title**

**Estimated ultimate recovery prediction in Shale Reservoirs**

**Team Leader:**

Shivanoor Vignesh 21R11A0597

**Team Members:**

G . Sandhya Rani 21R11A0569

R. Dhathri 21R11A0545

G.Sri Hasini 21R11A0518

A.Pooja Goud 21R11A0501

PSBN Sriya 21R11A0595

MD.Akhbar Ali 21R11A0586

N sai Charan 21R11A0588

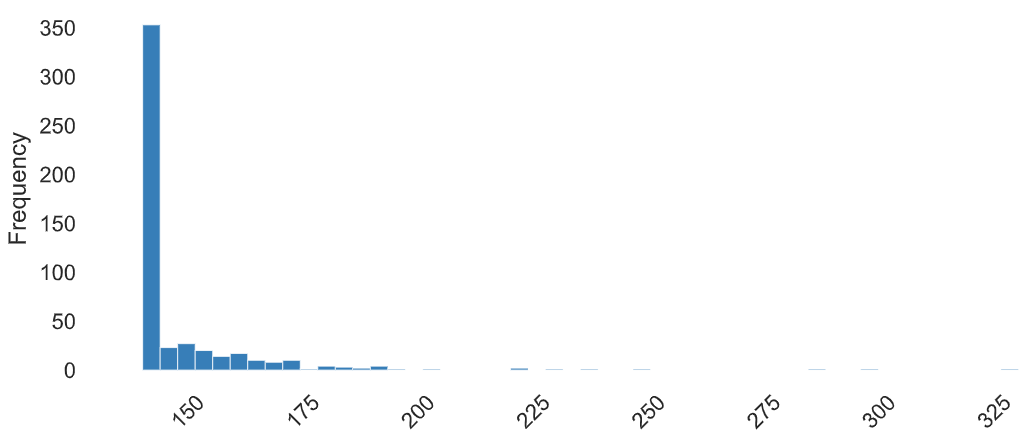
**Exploratory Data Analysis**

**Import data set**

For importing the dataset and to perform Exploratory Data Analysis we have to import some packages or library which are essential.

* import pandas
* import NumPy
* import seaborn
* import matplotlib

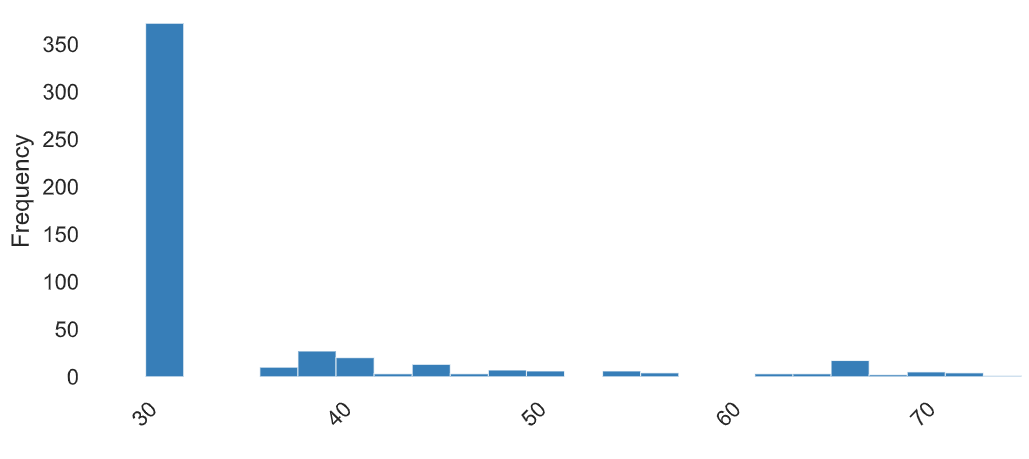
1. **State Spacing**



**OBSERVATION**

* The data shows a wide range of values from a minimum of 140 to a maximum of 330, with a median of 141 and a mean of 147.64.
* The data exhibits significant positive skewness (skewness value of 5.21) and high kurtosis (kurtosis value of 36.99), indicating a heavy tail and potential outliers. The standard deviation is relatively high at 18.39, representing notable variability in the data.
* The coefficient of variation is 0.12, suggesting moderate relative variability compared to the mean. The interquartile range (IQR) is 8, indicating that the middle 50% of the data is relatively compact. However, the data is not monotonic, indicating non-linear or complex patterns in its distribution.

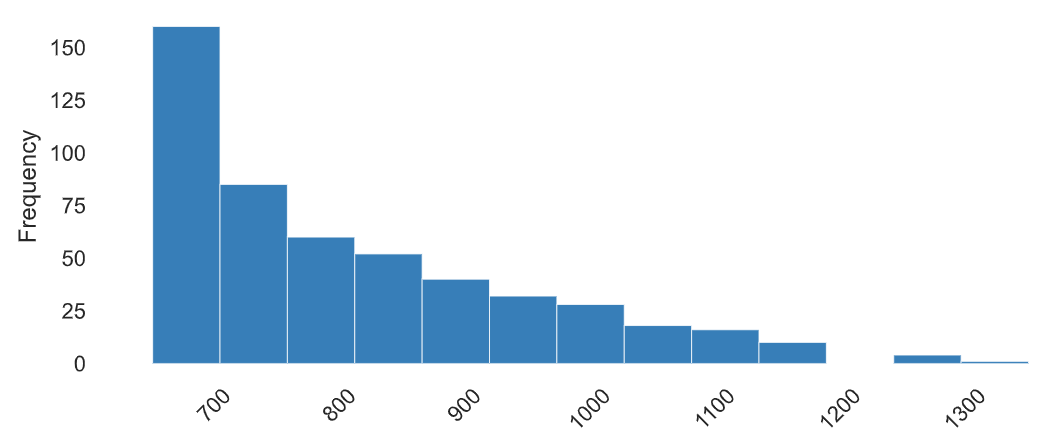
**2.** **Analysis of bbl/ft**



**Observation:**

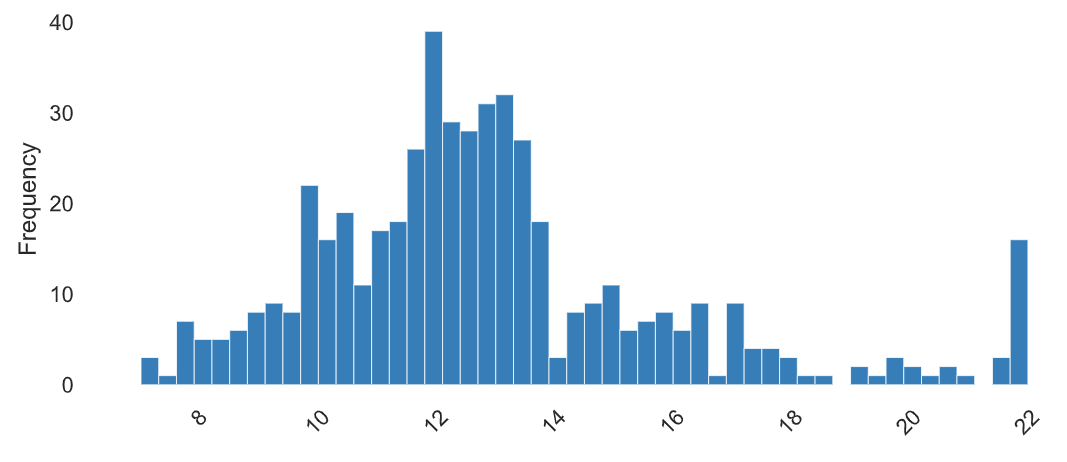
* The data has a relatively narrow range from 30 to 75, with a median, first quartile (Q1), and fifth percentile all at 30, and a third quartile (Q3) at 36.
* The mean is 35.13, slightly higher than the median, indicating a positive skewness of 2.23.
* The standard deviation is 10.53, representing moderate variability around the mean. The coefficient of variation is 0.30, indicating moderate relative variability.
* The kurtosis is 4.04, suggesting a moderately peaked distribution. The interquartile range (IQR) is 6, showing that the middle 50% of the data is relatively compact.
* The data is not monotonic, indicating non-linear or complex patterns in its distribution.

**3. Comparative Analysis of well spacing**



**Observations**

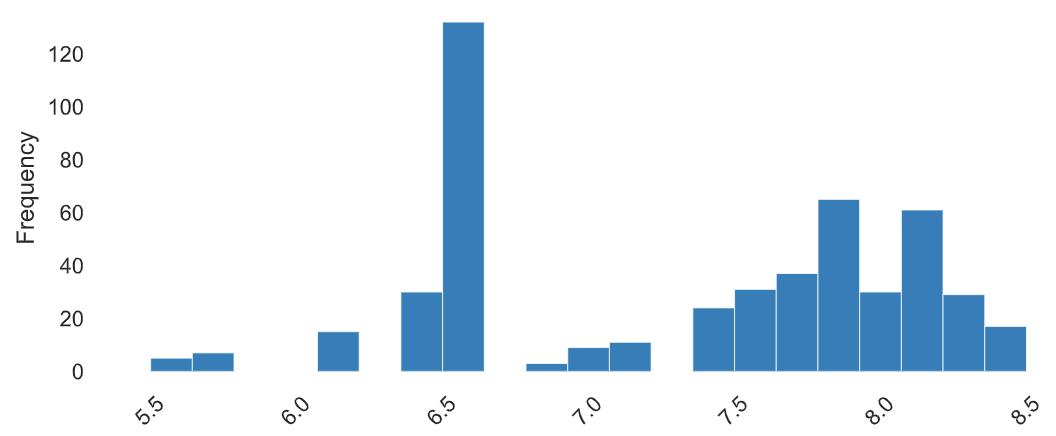
* The data has a range from 650 to 1350, with a median at 800 and a mean of 820.16. The data is slightly positively skewed with a skewness of 0.98.
* The standard deviation is 135.74, indicating moderate variability around the mean.
* The coefficient of variation is 0.17, suggesting moderate relative variability. The kurtosis is 0.40, indicating a relatively normal distribution with a slightly flattened peak.
* The interquartile range (IQR) is 200, showing that the middle 50% of the data is relatively compact. The data is not monotonic, indicating non-linear or complex patterns in its distribution.

**4. Analysis of EUR** 

**Observation:**

* The data shows a relatively narrow range from 7 to 22, with a median at 12.4 and a mean of 12.85.
* The data is positively skewed with a skewness of 1.10 and has a moderately peaked distribution with a kurtosis of 1.49.
* The standard deviation is 3.07, indicating moderate variability around the mean. The coefficient of variation is 0.24, suggesting moderate relative variability. The interquartile range (IQR) is 2.7, showing that the middle 50% of the data is relatively compact.
* The data is not monotonic, indicating non-linear or complex patterns in its distribution.

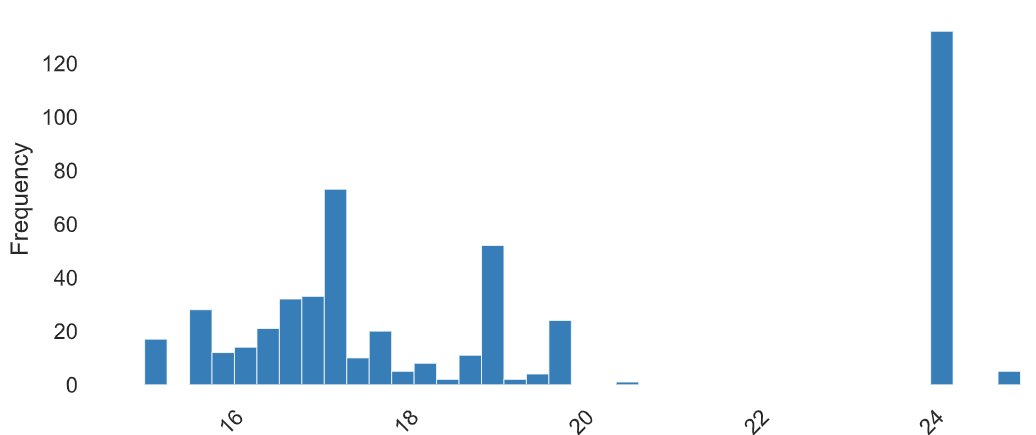
**5. Analyzing Porosity**



**Observations**

* The data ranges from 5.5 to 8.5, with a median at 7.5 and a mean of 7.34.
* The data is negatively skewed with a skewness of -0.32 and has a moderately flat distribution with a kurtosis of -1.16.
* The standard deviation is 0.75, indicating relatively low variability around the mean. The coefficient of variation is 0.10, suggesting low relative variability. The interquartile range (IQR) is 1.4, showing that the middle 50% of the data is relatively compact.
* The data is not monotonic, indicating non-linear or complex patterns in its distribution.

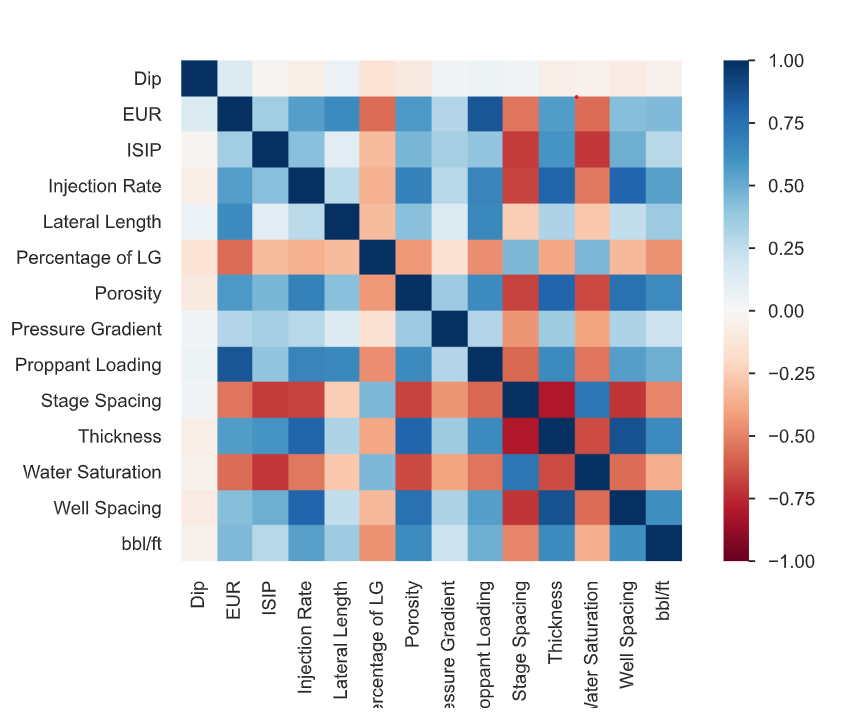
**6. Water Saturation**



**Observation**

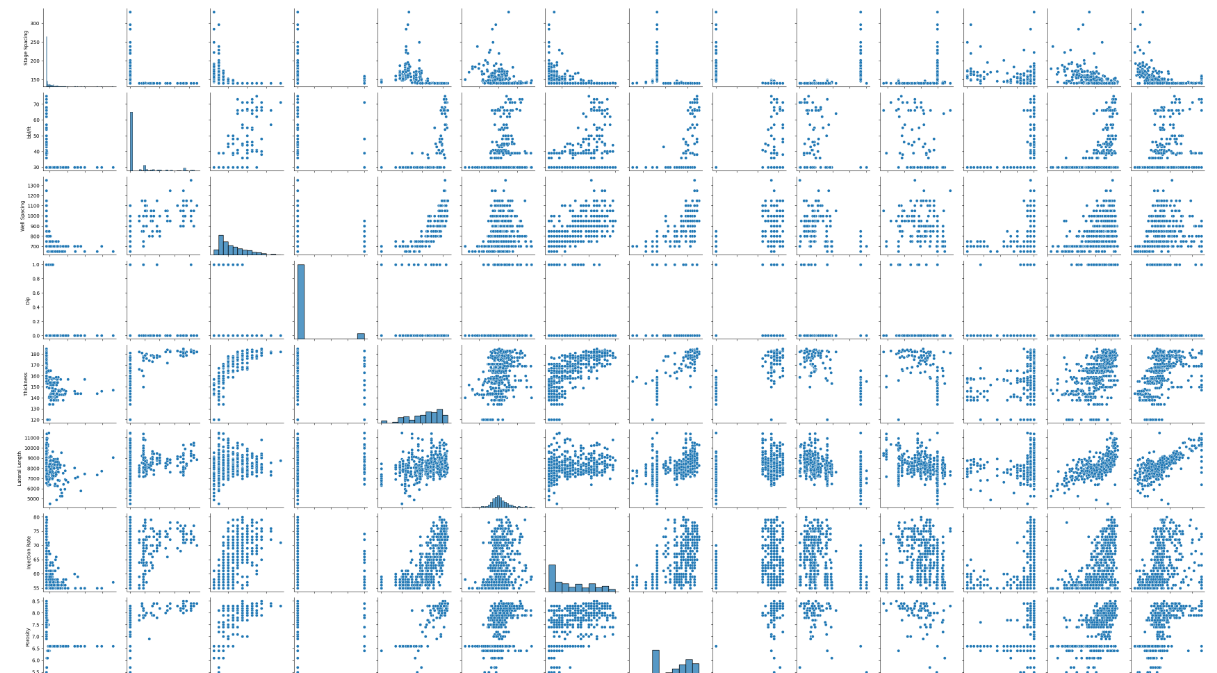
* The data ranges from 15 to 25, with a median at 17.7 and a mean of 19.21.
* The data is positively skewed with a skewness of 0.67 and has a moderately flat distribution with a kurtosis of -1.14.
* The standard deviation is 3.20, indicating moderate variability around the mean. The coefficient of variation is 0.17, suggesting moderate relative variability. The interquartile range (IQR) is 7.3, showing that the middle 50% of the data is relatively spread out.
* The data is not monotonic, indicating non-linear or complex patterns in its distribution.

**7. Correlation**

 **Observations**

* EUR and Proppant Loading (0.852): There is a strong positive correlation between the estimated ultimate recovery (EUR) and the amount of proppant loaded into the well during hydraulic fracturing. This suggests that higher proppant loading is associated with higher expected recovery.
* Stage Spacing and ISIP (-0.698): There is a strong negative correlation between stage spacing (cluster spacing during fracturing) and initial shut-in pressure (ISIP). This suggests that as stage spacing decreases, ISIP tends to increase.
* Water Saturation shows negative correlations with several factors such as EUR, Injection Rate, and Proppant Loading, suggesting that lower water saturation levels may be beneficial for recovery and fracturing efficiency.

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

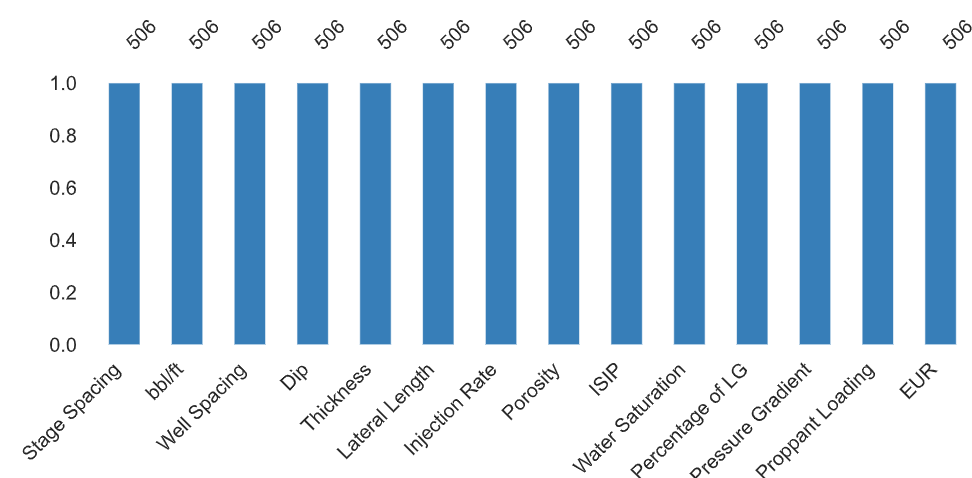




**Observations**

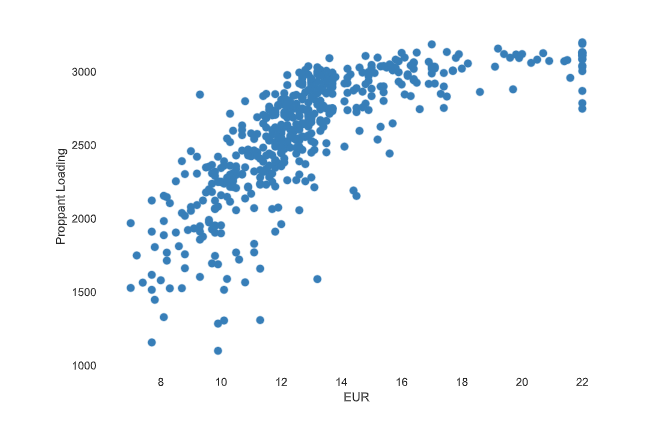
* The scatter plot matrix offers a comprehensive view of the multivariate data, with each cell displaying a scatter plot of two variables.
* Positive Correlation between EUR and Proppant Loading: The pair plot shows a clear positive correlation between EUR (Estimated Ultimate Recovery) and Proppant Loading. As the proppant loading increases, the estimated ultimate recovery also tends to increase, indicating that higher proppant levels during hydraulic fracturing may lead to greater recovery from the shale reservoirs.
* EUR and Injection Rate Relationship: There appears to be a positive correlation between EUR and Injection Rate, although the relationship is not as strong as with proppant loading. This suggests that higher injection rates may have a moderate positive impact on estimated ultimate recovery. Impact of Lateral Length on EUR: The pair plot shows a moderate positive correlation between Lateral Length and EUR. This indicates that longer lateral lengths in shale wells are associated with higher estimated ultimate recovery, which aligns with industry knowledge that longer horizontal sections can lead to increased production.
* Water Saturation and Recovery: There seems to be a negative correlation between Water Saturation and EUR. Lower water saturation levels are associated with higher estimated ultimate recovery, highlighting the importance of managing water content in the reservoir for optimal recovery rates. Overall, the pair plot analysis provides valuable insights into the interrelationships between various parameters and their impact on shale gas reservoir performance and recovery rates.

**9.Missing values**



**Observation:**

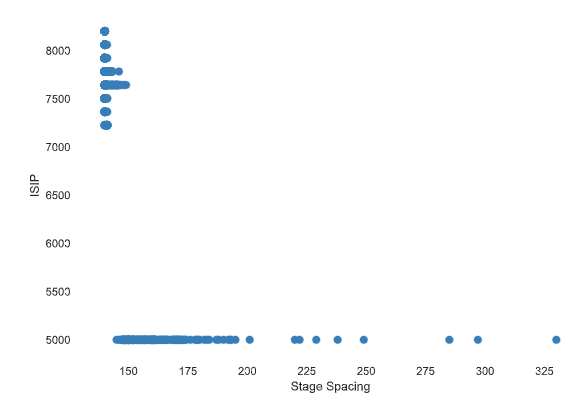
* There are missing values that need to be addressed. Missing values can impact the accuracy and reliability of analyses and models. Therefore, it's essential to handle them appropriately.
* Before proceeding with any analysis or modeling, it's important to identify the extent of missing in each variable and consider strategies such as imputation or deletion based on the context and impact on the analysis results. Ignoring missing values or handling them improperly can lead to biased or misleading conclusions.

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

**Observation:**

* This interaction is crucial as it indicates that higher proppant loading is associated with increased estimated ultimate recovery (EUR).
* Highlighting this relationship can be valuable for understanding how hydraulic fracturing practices affect reservoir performance.
* A horizontal cluster near the bottom of the graph, around the 5000 mark on the SIP axis.
* A vertical cluster on the left side of the graph, between approximately 150 and 175 on the Stage Spacing axis.

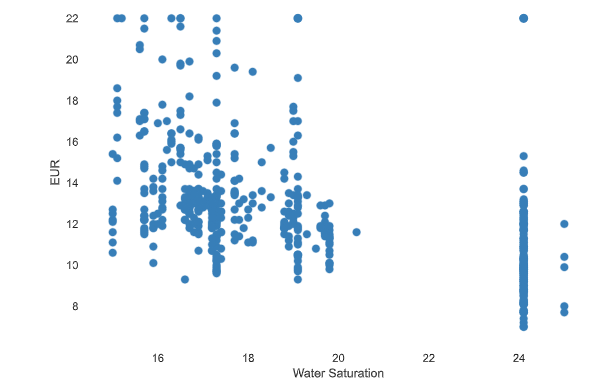
**11.State spacing and ISIP**



**Observation:**

* **T**he strong negative correlation between stage spacing and initial shut-in pressure (ISIP) suggests an important interaction in fracturing performance.
* Discussing how stage spacing impacts ISIP and, consequently, reservoir stimulation efficiency can be informative.
* A horizontal cluster near the bottom of the graph, around the 5000 mark on the SIP axis.
* A vertical cluster on the left side of the graph, between approximately 150 and 175 on the Stage Spacing axis.

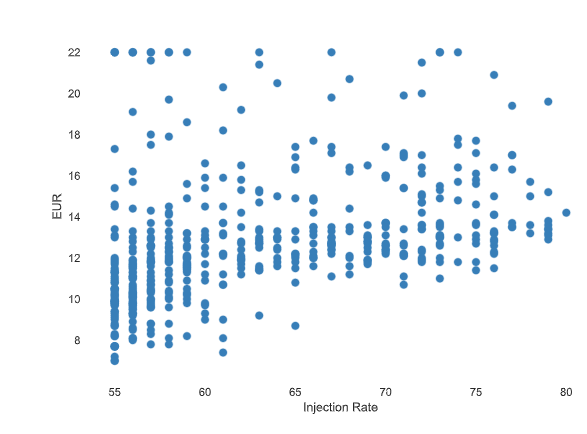
**12.Water Saturation and EUR**



**Observation:**

* The negative correlation implies that lower water saturation levels may be beneficial for recovery and fracturing efficiency.
* Discussing strategies to manage water saturation can be relevant for optimizing production.
* Most data points cluster between Water Saturation values of roughly 16 to 20 and EUR values of approximately 10 to 18.

**13.Injection Rate and EUR**



**Observation:**

* The moderate positive correlation between injection rate and EUR implies that higher injection rates may contribute to increased recovery.
* This interaction is relevant for optimizing injection strategies for better reservoir performance.

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

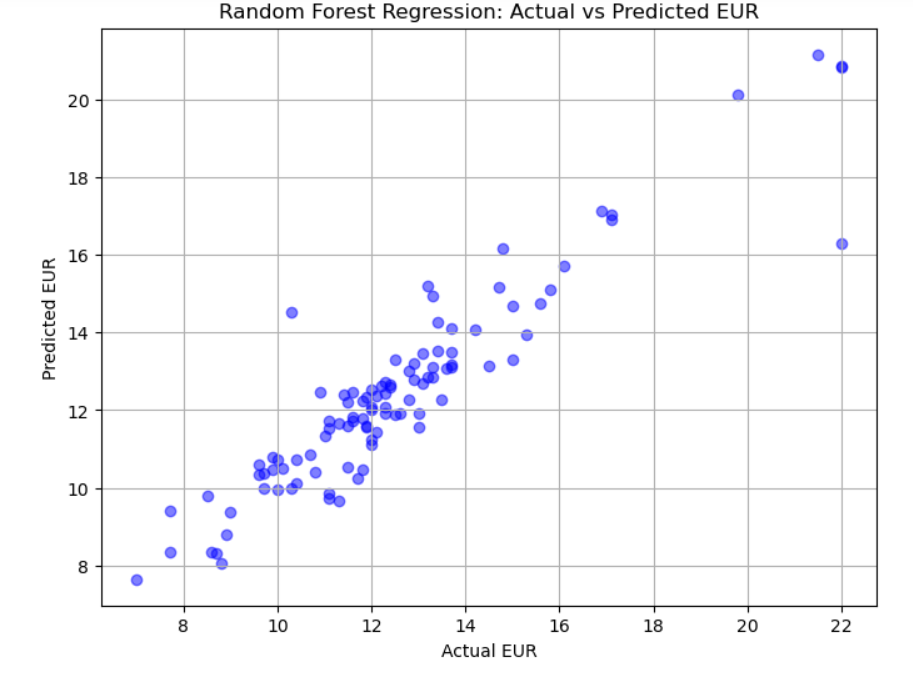
**1. Random Forest Regressor**

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:**

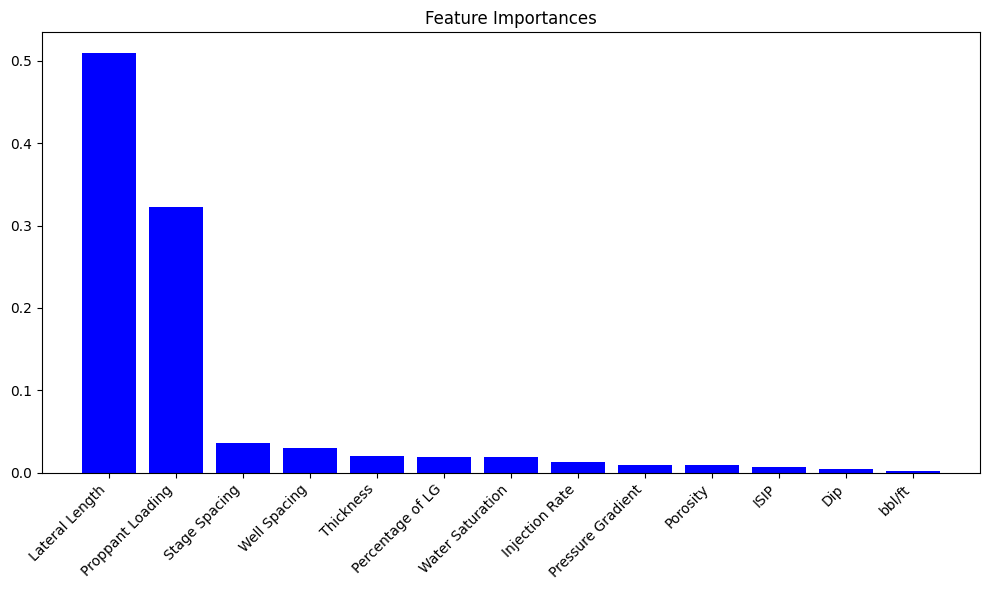
* The scatter plot visualizes how well the Random Forest Regressor model predicts the actual values. Ideally, the points should form a diagonal line, indicating a perfect prediction where actual and predicted values are identical. Deviations from this line indicate the model's predictive errors.
* A scattered or dispersed pattern away from the diagonal may suggest areas where the model struggles to accurately predict the target variable, highlighting potential areas for model improvement or further analysis.
* There’s a trend indicating that as the Actual EUR increases, so does the Predicted EUR, but not perfectly linearly.

**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

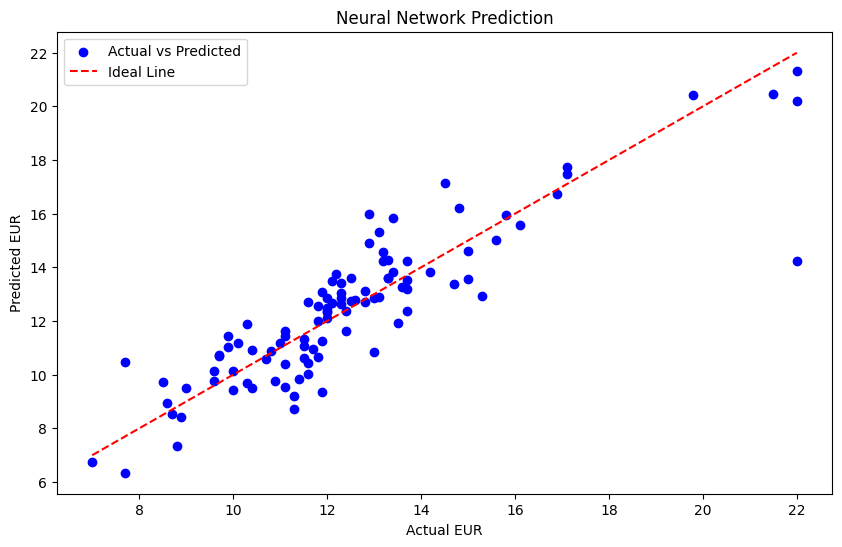
11 Dip 0.004078

12 bbl/ft 0.002448

**Observations:**

* The bar graph illustrates the importance of different features, likely related to geological or environmental data, with the y-axis showing scores from 0 to 1.
* "Lateral Length" emerges as the most significant feature, with a score exceeding 0.5, indicating its strong influence on outcomes.
* Following "Lateral Length," "Proppant Loading" stands out as the second most important feature, though with a lower impact compared to the former.
* Other features, such as "Stage Spacing" and "Well Spacing," exhibit minimal importance, suggesting they have little effect on the outcomes.
* The insights gleaned from this analysis can guide decision-making, emphasizing the need to optimize "Lateral Length" and explore ways to enhance "Proppant Loading" for improved result.

**Neural Network**



**Observations**

The above scatterplot predicts that of neural networks:

Scatter Plot Visualization: The scatter plot compares actual and predicted EUR values, with each point representing an observation. It helps visualize how well the model's predictions align with the actual values.

Ideal Line Interpretation: The diagonal ideal line represents perfect prediction, where all points would lie if the model's predictions were exact. Deviations from this line indicate prediction errors.

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

Model Performance Evaluation: A good fit is indicated by most points aligning closely to the ideal line, signifying accurate predictions. Widely scattered points indicate less consistent predictions and potential areas for improvement.

Further Analysis: Calculating residuals (differences between actual and predicted values) helps assess the model's performance in more detail. Additionally, R-squared (R²) can provide a metric for 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 achieved an accuracy with a mean squared error of 1.074 and an R-squared score of 0.868, indicating a good fit to the data and effective prediction of estimated ultimate recovery (EUR) in shale reservoirs.

The analysis revealed that "Lateral Length" is the most influential feature for predicting EUR, followed by "Proppant Loading." Other features like "Stage Spacing" and "Well Spacing" have minimal impact, suggesting they can be deprioritized in the modeling process.

The neural network model's scatter plot compares actual and predicted EUR values, highlighting areas where the model's predictions deviate from actual values. It indicates the model's accuracy and areas for potential improvement.

The scatter plot analysis helps assess the accuracy of the models, with a tight cluster around the ideal line indicating accurate predictions and a scattered pattern suggesting variability in predictions.