Table of Contents

[Answer to Q1: 2](#_Toc148421606)

[Answer to Q2: 8](#_Toc148421607)

[Answer to Q3: 9](#_Toc148421608)

[Answer to Q4: 10](#_Toc148421609)

[Answer to Q5: 14](#_Toc148421610)

[Answer to Q6: 15](#_Toc148421611)

# Answer to Q1:

**Finding 1: summary(marun) and Missing Values**

Before diving into the values of the dataset, I wanted to first find out the datatype for each variable and confirm if it was accurate. To get this, I used sapply() function to identify the datatype and we can observe that the variables are either integers or numeric. However, there was a variable that stood out which was Formation as the research article indicated it as “Formation type”, which meant the values were categorical, which meant it had to be changed to a factor datatype. The datatypes of the rest of the variables were accurate and did not need any amendments.

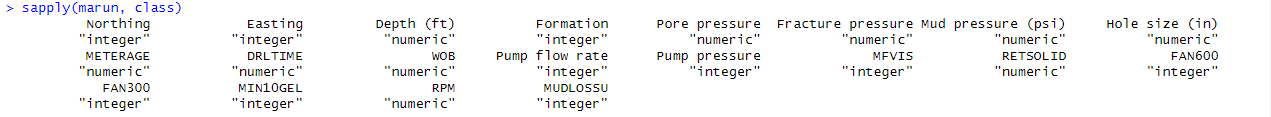


Figure 1: Datatype for each variable

To convert “Formation” from a numeric value to a factor, I used the factor() function (Figure 2).



Figure 2: Factoring "Formation" variable

Subsequently, this is the result depicting the different variables we are working with for the assignment (Table 1):

Table 1: Variables and datatype

|  |  |  |
| --- | --- | --- |
| **Variables** | **Class** | **Levels** |
| **Northing** | Integer | - |
| **Easting** | Integer | - |
| **Depth (ft)** | Numeric | - |
| **Formation** | Factor | 15 (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15) |
| **Pore pressure** | Numeric | - |
| **Fracture pressure** | Numeric | - |
| **Mud pressure (psi)** | Numeric | - |
| **Hole size (in)** | Numeric | - |
| **METERAGE** | Numeric | - |
| **DRLTIME** | Numeric | - |
| **WOB** | Numeric | - |
| **Pump flow rate** | Integer | - |
| **Pump pressure** | Integer | - |
| **MFVIS** | Integer | - |
| **RETSOLID** | Numeric | - |
| **FAN600** | Integer | - |
| **FAN300** | Integer | - |
| **MIN10GEL** | Integer | - |
| **RPM** | Numeric | - |
| **MUDLOSSU** | Integer | - |

Furthermore, the two variables Northing and Easting are geographical coordinates which sole purpose is to identify the location of the specific instance a drilling was carried out in the Marun field. Hence the values in both Northing and Easting do not contribute to the severity of lost circulation. This then leaves us with the following:

1. 1 continuous Y variable (MUDLOSSU)
2. 16 continuous X variables
3. 1 categorical X variable

To get a quick overview of the dataset we are working with, I used the summary() function and the results are shown in Figure 3.

A close-up of a text

Description automatically generated

Figure 3: Overview of summary(marun)

Since we have a large number of continuous variables to work with, just by looking at the numbers in Figure 3 will not suffice. Therefore, this will be further discussed in part 2 of my findings with plots. However, we also observe that there is a total of 4 NAs in the dataset from the following columns: FAN600, FAN300, MIN10GEL and MUDLOSSU. To find out which row the missing values belonged to, I used the which(is.na()) function, and stored the 4 results into a dataframe for easier readability.

A screen shot of a computer code

Description automatically generated

Figure 4: Results of missing data

From Figure 4, we can observe that the missing values from FAN600, FAN300, MIN10GEL belonged to row 3 while the missing value from MUDLOSSU belonged to row 6. This will be useful information for Question 2.

**Finding 2: Outliers** **in Variables**

To better analyse the data from summary(marun), it is apt that we create a boxplot for each variable to better observe the behaviours of the values (Figure 5, Figure 6).

A chart of different sizes and shapes

Description automatically generated with medium confidenceA chart of different sizes and shapes

Description automatically generated with medium confidence

Figure 5: Boxplot of variables

A graph of a diagram

Description automatically generated with medium confidence

Figure 6: Boxplot of variables (cont.)

For a large dataset with 2668 rows of values, it is not uncommon to see outliers due to the natural variation. Therefore, having a boxplot allows us to see at a glance if there are any extreme outliers in the variables. There are a few observations that stand out from the rest:

1. METERAGE has an extreme outlier of 650m
2. RPM has two outliers of 394rpm and 375.5rpm
3. MUDLOSSU (Y variable) has a huge number of outliers

My planned approach is to remove the rows containing the outliers from METERAGE and RPM since these outliers are questionable and deviate too far from the outliers that appear to be of natural variation. On the other hand, I plan to look deeper into the outliers of MUDLOSSU through the use of a histogram (Figure 7).

A graph of a number of objects

Description automatically generated

Figure 7: Histogram of Severity of Lost Circulation (MUDLOSSU)

Figure 7 indicates that the severity of lost circulation creates a right-skewed histogram, which indicates the majority of records (1916) do have little to no lost circulation in the process of drilling. However, there are definitely instances when lost circulation occur which will be useful in our predictive modelling subsequently. If there are no examples of lost circulation occurring, there will be not enough data to help predict the severity of lost circulation in the future. Therefore, these outliers are critical to our analysis and shall not be removed.

**Finding 3:** **Correlation between Variables**

To find the correlation between variables, I figured it was the most straightforward to use the corrplot() function to visualise it for better understanding. A corrplot can be interpreted this way: the area of the circle shows the absolute value of corresponding correlation coefficients - the bigger the circle, the greater the absolute value of correlation coefficient – while blue indicates a positive correlation while red indicates a negative correlation. The greater the intensity of the colour, the more positively or negatively correlated it is, depending on the colour.

A diagram of a diagram

Description automatically generated with medium confidence

Figure 8: Corrplot of variables

This corrplot is very useful in the sense that it gives us a fair bit of insights into the correlation between variables. I have singled out the more insightful ones as follows:

1. Pore pressure, fracture pressure and mud pressure all have positive correlations with one another and this suggests an increase in pore pressure will reduce the effective stress on the rock formation, making it more likely to fracture and also require an increase in mud pressure to maintain control to prevent lost circulation. Similarly, the increase fracture pressure will also require an increase in mud pressure.
2. Viscosity (measure of the resistance of a fluid to flow) and solid percent obtained from retort test (measure of the percentage of solids in the drilling fluid) are also positively correlated at 0.77 which suggests an increase in solid percent obtained from retort test will increase the viscosity of the drilling fluid.
3. In addition to point 2, these variables are also positively correlated to the shear stresses (the stress that acts on a surface parallel to the surface), indicating an increase in either variable will result in an increase the shear stresses on the drilling fluid.
4. Shear stresses at shear rates of 600rpm (FAN600) and shear stresses at shear rates of 300rpm (FAN300) are perfectly correlated at 1.00 and this makes sense because they are the measurement of the same thing at different rates.

However, if we focus on the relationships between the target Y variable (MUDLOSSU) against the other X variables, we will realise that there is not much of a strong positive/negative correlation, with the highest positive correlation at 0.52 against pump flow rate, and the highest negative correlation at -0.47 against depth. Therefore, the relationship between the severity of lost circulation and other variables is much more complex and cannot be captured by a simple correlation model and requires a deeper analysis into it.

# Answer to Q2:

With reference to Figure 2 and Table 1, the categorical variable in this dataset is Formation.

Previously in Figure 4, we have identified the missing values and the different rows these values belonged to, with FAN600, FAN300 and MIN10GEL belonging to row 3 and MUDLOSSU belonging to row 6. Considering the fact that row 3 was missing 3 different variables, it will not be wise to use any form of models to predict and fill in the missing values as it would be difficult to determine the relationship between these missing variables and other variables which may reduce the accuracy or introduce new biases which is not what we want. Couple that with the fact that we are working with a large dataset, dropping this row entirely will not affect our results too much. Therefore, we will drop row 3 (Figure 9).



Figure 9: Omitting missing values

On the other hand, in row 6, the missing variable is MUDLOSSU, which is the target Y variable. With all the other variables intact in the row, we are able to predict the missing value using regression imputation with the codes (Figure 10), which gives us a predicted value of **131.5214.**

A close up of text

Description automatically generated

Figure 10: Predicting MUDLOSSU value

One thing to note: the prediction of MUDLOSSU value is done before omitting the missing values in row 3 as Figure 9 is a line of code that deletes all missing values from the dataset and not just row 3. It is important that there is a new value in the originally missing MUDLOSSU variable first before omitting all missing values from dataset.

# Answer to Q3:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Complexity | Trainset RMSE | Testset RMSE |
| Linear Regression | With reference to Figure 14,  1 Categorical X Variable:   1. Formation   11 Continuous X Variable:   1. Depth (ft) 2. Pore pressure 3. Fracture pressure 4. Hole size (in) 5. METERAGE 6. Pump flow rate 7. MFVIS 8. RETSOLID 9. FAN600 10. FAN300 11. MIN10GEL | 123.3481 | 119.3640 |
| CART | 4 Terminal Nodes    Figure 11: Summary of cart2 | 122.1758 | 114.0043 |

# Answer to Q4:

**Train-Test Split**

Before attempting to develop the different models, I did a 70-30 train-test split (Figure 12) to prevent overfitting or underfitting of the models.

A black and white text

Description automatically generated with medium confidence

Figure 12: Train-Test split

**Linear Regression**

As mentioned earlier, I will be excluding Northing and Easting variables when developing the linear regression model. I created the first linear regression model (m1) as shown in Figure 12, and it shows that the statistically significant variables are Depth (ft), Formation, Pore pressure, Fracture pressure, Hole size, Meterage, Pump flow rate, MFVIS, RETSOLID, FAN600, FAN300, MIN10GEL.

A screenshot of a computer

Description automatically generated

Figure 13: Summary of first linear regression model

After attaining the results of the above, I removed the statistically insignificant variables and re-ran the linear regression model on the trainset data, with the results shown in Figure 13.

A screenshot of a computer

Description automatically generated

Figure 14: Summary of second linear regression model

According to the p-values, the top 5 most statistically significant variable is order is drilling meterage, followed by formation, then viscosity, then shear stresses at rates of 300rpm, lastly shear rates at 600rpm.

With this, I proceeded to calculate the RMSE of the trainset which resulted in **123.3481**. Next, I used the trainset model to predict the testset and calculated the RMSE based on the testset which gave me a result of **119.364**.

A number of numbers and symbols

Description automatically generated with medium confidence

Figure 15: RMSE results of linear regression model

**CART**

The CART model was first developed on the trainset and the output is the maximal tree which takes all the variables into consideration.

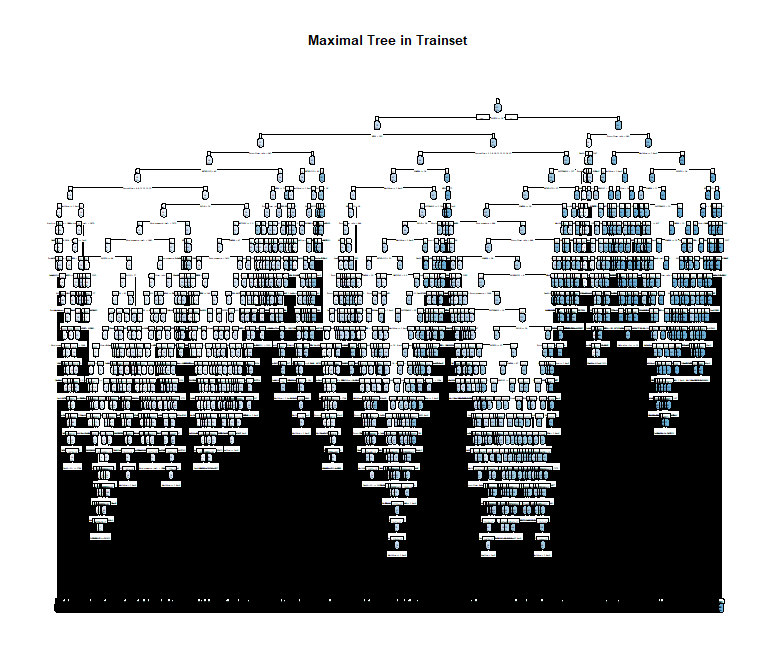


Figure 16: Maximal tree in trainset

The CV error cap is **0.601076**. The optimal CP region whose CV error is just below the CV error cap in the maximal tree is where we will prune the maximal tree to get our optimal tree. After calculations, the optimal CP value is **0.036327**.

Figure 17 displays the optimal tree with 4 terminal nodes alongside the cp chart.

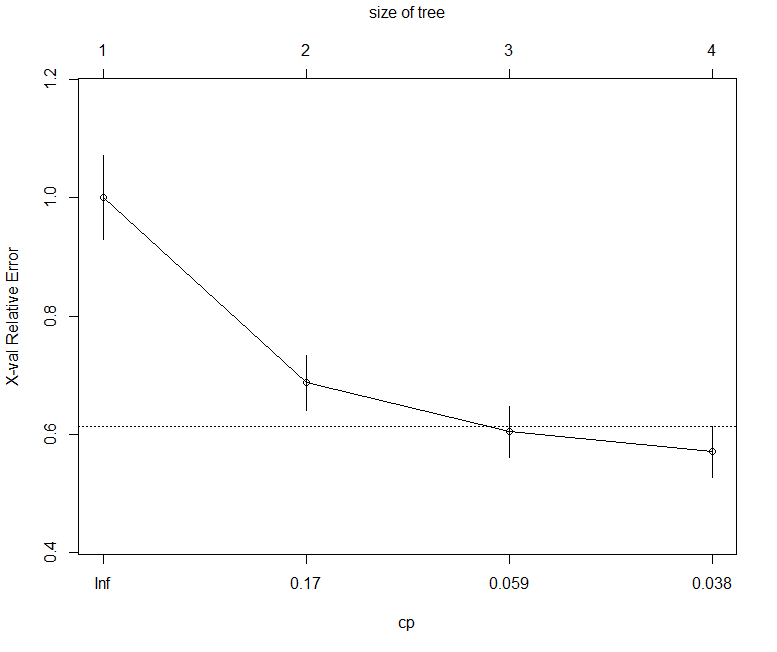
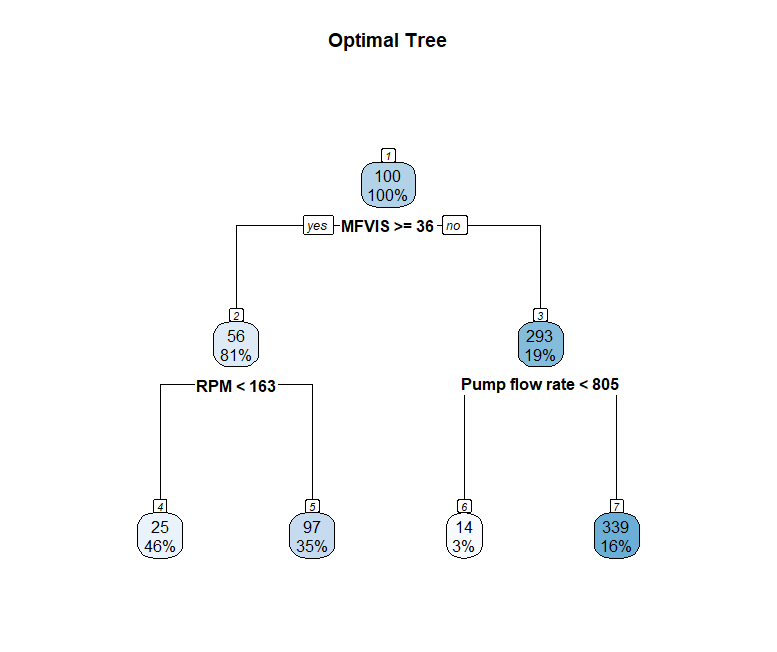


Figure 17: Optimal tree and CP chart

In addition, I also calculated the variable importance of each X variable and the results are shown in Figure 18. The results indicate that the viscosity holds the most weightage at 19%, followed by pore and fracture pressure at 13% each and so on.

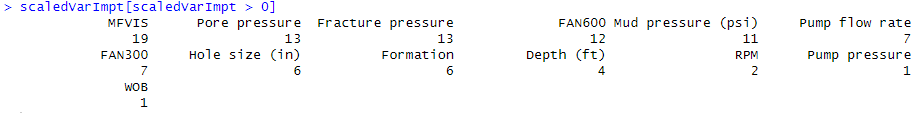


Figure 18: Variable importance of X variables

Subsequently, I proceeded to calculate the RMSE of the trainset which gave me a result of **122.1758** by square rooting the mean squared error (MSE) from the root node error of 26847. Next, I used the trainset model to predict the testset and calculated the RMSE which had the result of **114.0043**.

A number with numbers and numbers

Description automatically generated with medium confidence

Figure 19: RMSE results of CART model

**Analysis of Results and Insights**

Firstly, a train-test split allows us to train the model using the trainset and the testset is used to evaluate the performance of the model on unseen data. In both my linear regression model and CART model, the RMSE of both testsets were lower than the trainset, which proves that it is able to generalise the data and there was no overfitting. A lower RMSE was observed in the CART testset model as compared to the linear regression testset model, which signifies that the CART model may be a better fit for the marun dataset.

Hence, if we focus on the CART model, the variable importance scores (Figure 18) suggest that the following variables are most important for predicting the severity of lost circulation (MUDLOSSU):

1. MFVIS (viscosity)
2. Pore pressure
3. Fracture pressure
4. FAN600 (shear stresses at shear rates of 600rpm)
5. Mud pressure
6. Pump flow rate
7. FAN300 (shear stresses at shear rates of 300rpm)

These variables are all related to the drilling process and the properties of the drilling fluid. Therefore, businesses can use these insights to improve their drilling processes and operations by carefully managing and controlling these variables. For instance, if they are experiencing high lost circulation, they can try to reduce the viscosity of the drilling fluid, reduce pore pressure or increase the mud pressure. Furthermore, with the simplicity of a pruned tree, it aids the businesses by providing a clear decision route to minimise lost circulation. By better understanding the factors that contribute to the severity of lost circulation, businesses can then develop better strategies to prevent this problem, leading to significant cost savings.

# Answer to Q5:

Sabah M. et. al. (2019) chose to employ a regression tree with 5-fold cross-validation over a classification tree which was the correct model since the Y variable – severity of lost circulation – was a continuous variable. The first split at the root node was based on viscosity with a splitting value of 35.5 cp, followed by RPM and pump flow rate and this process continues until reaching the terminal nodes with no further splitting. The first few splits are identical to the optimal tree that I have generated in Figure 17. Furthermore, Sabah M. et. al’s (2019) decision tree was then evaluated by 4 different performance measures, and it demonstrated the highest applicability in lost circulation prediction as compared to other models developed, with a RMSE value of 0.091, a R2 value of 0.93 and a PI value of 1.8 (Figure 19). This indicates a high level of accuracy and a good fit of the model to the data, meaning the CART model is effective in predicting lost circulation in drilling operations. The decision tree allows for businesses to weigh the possible actions after studying the possible outcomes of a series of choices.

A table of numbers with text

Description automatically generated with medium confidence

Figure 20: Performance indices of models

# Answer to Q6:

Sabah M. et al. used 4 different performance measures (RMSE, R2, VAF (%), PI) to evaluate the predictive model outputs for lost circulation (Figure 19):

* RMSE (Root Mean Square Error) was used to measure the average magnitude of the differences between the predicted and actual values of lost circulation.
* R2 (Coefficient of Determination) was used to assess how well the model predicts or explains an outcome. A higher R2 indicates a better fit of the model to the data.
* VAF (Variance Accounted For) was used to quantify the proportion of variance in the predicted values that can be explained by the model. It provides an indication of how well the model fits the data.
* PI (Performance Index) was used as a comprehensive measure of the overall performance of the prediction models, taking into account all the above 3 performance measures.

These performance measures were utilised to assess the effectiveness of the models in predicting lost circulation, with the decision tree model outperforming the other models. Sabah M. et. al. did a remarkable job by using a combination of performance measures instead of limiting himself to one as different performance measure individually has its own advantages and disadvantages, for instance, the R2 alone does not provide the magnitude of errors which RMSE is able to help determine. Therefore, by assessing multiple performance measures, it allows for a less biased and well-evaluated analysis of the performance of the different prediction models.