*This is the main submission document****. Save and rename this document filename with your <u>registered full name</u> as Prefix before submission.***

| | |
|---|---|
| Full Name | Lee Fang Hui Jesslyn |
| Email Address | leef0015@e.ntu.edu.sg |

** : Delete and replace as appropriate.*

**Declaration of Academic Integrity**
By submitting this assignment for assessment, I declare that this submission is my own work, unless otherwise quoted, cited, referenced or credited. I have read and understood the Instructions to CBA.PDF provided and the Academic Integrity Policy.
I am aware that failure to act in accordance with the University's Academic Integrity Policy may lead to the imposition of penalties which may include the requirement to revise and resubmit an assignment, receiving a lower grade, or receiving an F grade for the assignment; suspension from the University or termination of my candidature.

I consent to the University copying and distributing any or all of my work in any form and using third parties to verify whether my work contains plagiarised material, and for quality assurance purposes.

*Please insert an "X" within the square brackets below to indicate your selection.*
**[X] I have read and accept the above.**

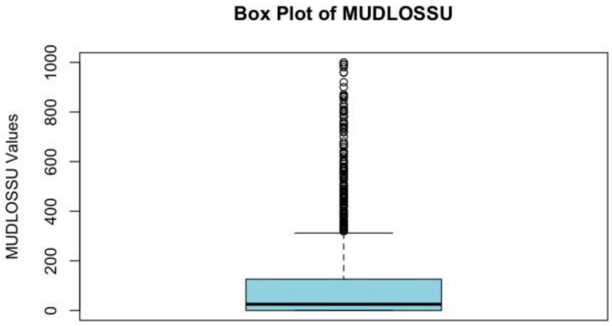
Table of Contents

*For each question, please start your answer in a new page.*

<u>Answer to Q1:</u>

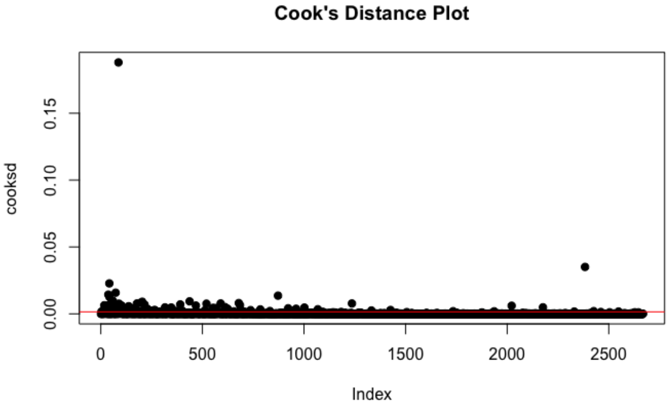
1. Explore the data and report 3 notable findings.

#1 Outliers
The "outliers" were very extreme for the predictor variable: MUDLOSSU
This is seen from a single box plot of MUDLOSSU, which shows that there are many data points that is outside the IQR. If there are that many data points they generally won't be considered as true outliers but rather extreme values within the distribution. This suggests that the outliers should not be removed as they are valuable data points that can provide important insights.

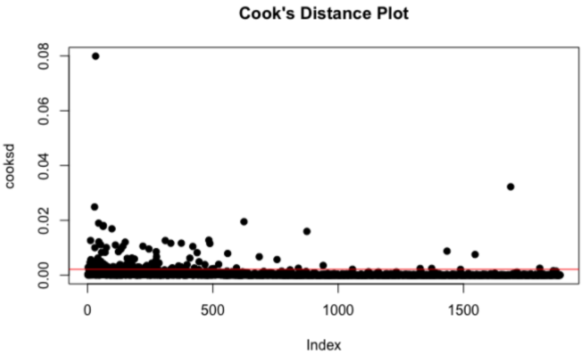


Additionally, after executing both my initial and optimal linear regression model, I used Cook's distance to determine if there were outliers involved that could potentially affect my model. I used a threshold
The results are as shown.

Initial



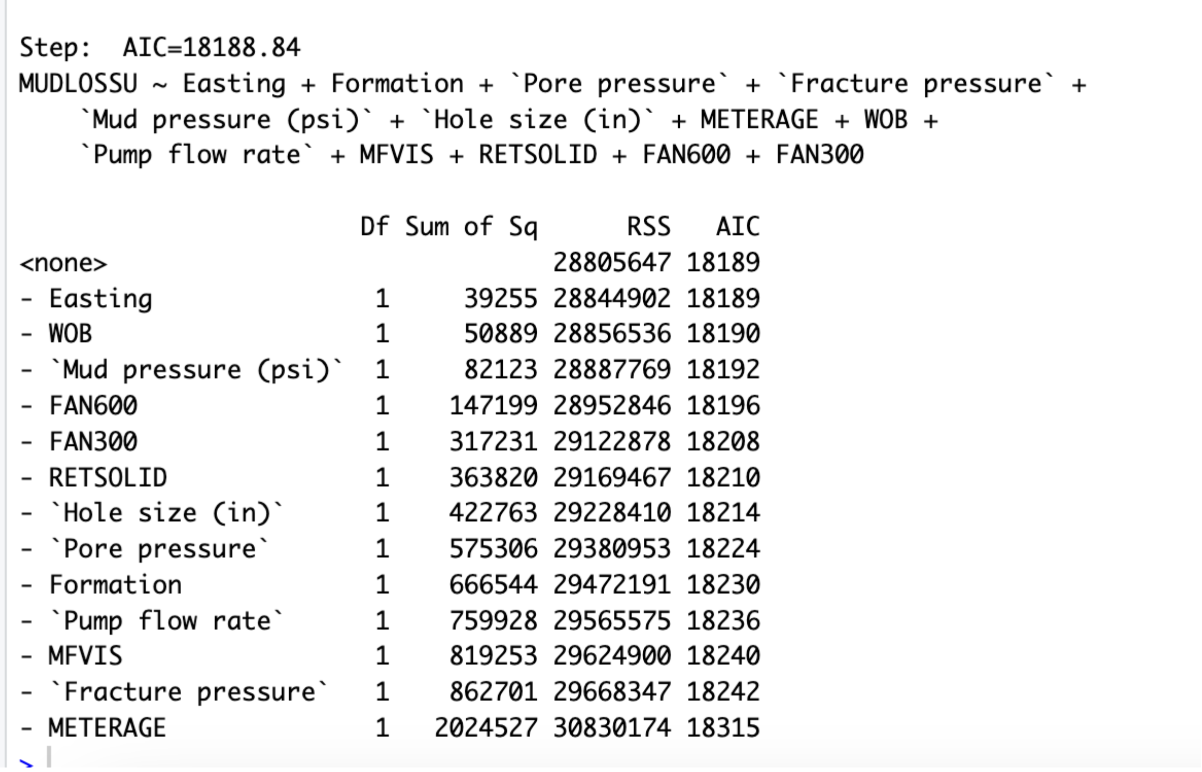
Optimal



There are only a handful of "outliers" and hence I have determined not to remove the outliers as it will not significantly affect my model's predictions.

#2  Overall, using both backward elimination method and multicollinearity detection has helped me obtain a lower train and test set error in linear regression model refinement.

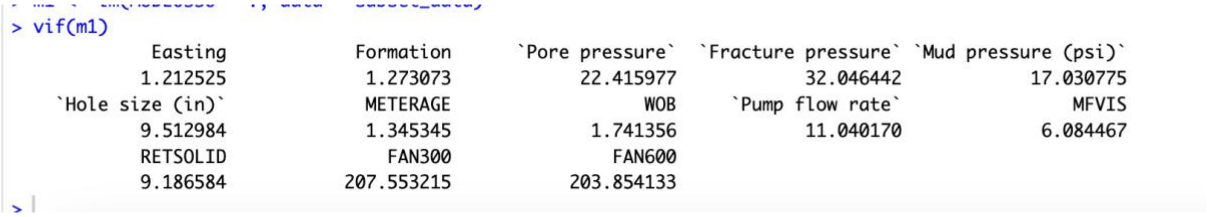
When backward elimination is executed, this was my final model, with the AIC = 18188.84



```
Step:  AIC=18188.84
MUDLOSSU ~ Easting + Formation + `Pore pressure` + `Fracture pressure` +
    `Mud pressure (psi)` + `Hole size (in)` + METERAGE + WOB +
    `Pump flow rate` + MFVIS + RETSOLID + FAN600 + FAN300

                        Df Sum of Sq      RSS   AIC
<none>                               28805647 18189
- Easting                1     39255 28844902 18189
- WOB                    1     50889 28856536 18190
- `Mud pressure (psi)`   1     82123 28887769 18192
- FAN600                 1    147199 28952846 18196
- FAN300                 1    317231 29122878 18208
- RETSOLID               1    363820 29169467 18210
- `Hole size (in)`       1    422763 29228410 18214
- `Pore pressure`        1    575306 29380953 18224
- Formation              1    666544 29472191 18230
- `Pump flow rate`       1    759928 29565575 18236
- MFVIS                  1    819253 29624900 18240
- `Fracture pressure`    1    862701 29668347 18242
- METERAGE               1   2024527 30830174 18315
>
```

It is observed that the variables : **Northing, Depth(ft), DRLTIME, Pump Pressure, MIN10GEL and RPM** were removed. These variables were removed indicating that they are not contributing significantly to the model's performance.

After performing vif() with the remaining variables to identify variables with high gif values, this output is shown.

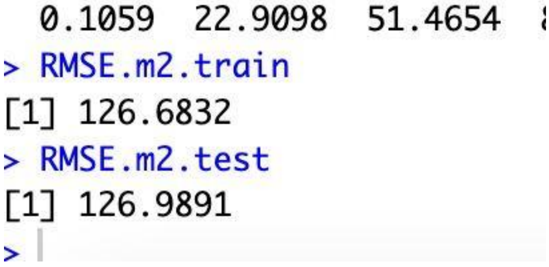


```
> vif(m1)
            Easting           Formation     `Pore pressure` `Fracture pressure` `Mud pressure (psi)`
           1.212525            1.273073           22.415977           32.046442            17.030775
   `Hole size (in)`            METERAGE                 WOB    `Pump flow rate`                MFVIS
           9.512984            1.345345            1.741356           11.040170             6.084467
           RETSOLID              FAN300              FAN600
           9.186584          207.553215          203.854133
>
```

We can note that FAN600 and FAN300 have extremely high values, as well as Pore pressure and Fracture Pressure and Mud pressure(psi).

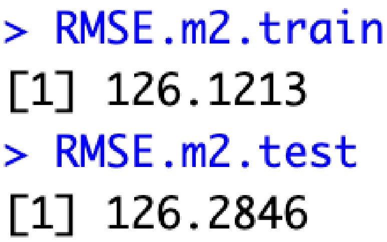
Hence, between FAN600 and FAN300, I removed FAN600 as I managed to get a lower and smaller difference between the train and test set error.

Removing FAN300



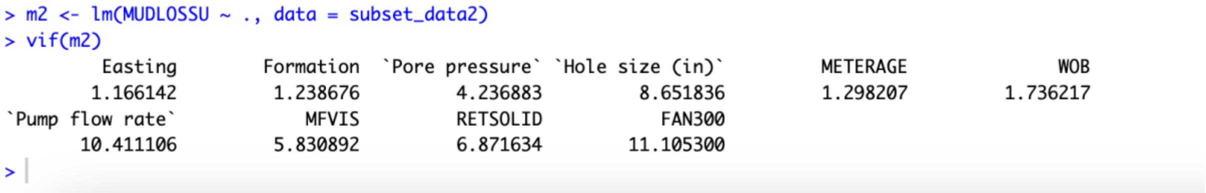
```
  0.1059   22.9098   51.4654
> RMSE.m2.train
[1] 126.6832
> RMSE.m2.test
[1] 126.9891
>
```

Removing FAN600



```
> RMSE.m2.train
[1] 126.1213
> RMSE.m2.test
[1] 126.2846
```

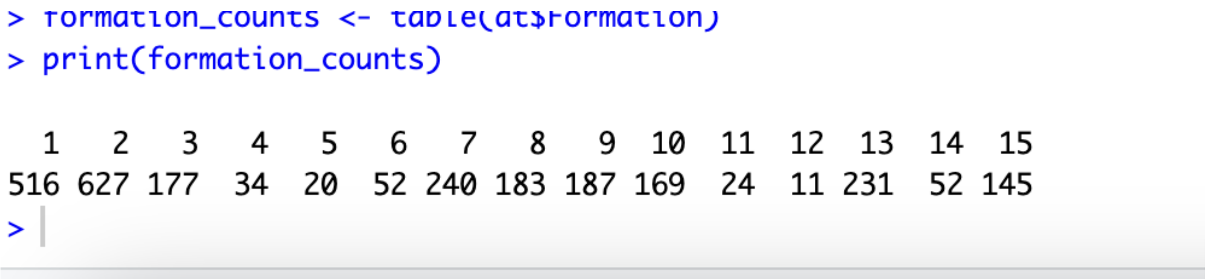
Between Pore pressure and Fracture pressure, I removed Fracture pressure as it had a higher vif value than Pore Pressure.
I removed Mud Pressure (psi) completely.
These were my new vif values for the remaining variables. It is observed that the numbers for Pore Pressure and FAN300 have dropped significantly.



```
> m2 <- lm(MUDLOSSU ~ ., data = subset_data2)
> vif(m2)
        Easting         Formation  `Pore pressure` `Hole size (in)`          METERAGE               WOB 
       1.166142          1.238676          4.236883          8.651836          1.298207          1.736217 
`Pump flow rate`             MFVIS          RETSOLID            FAN300 
      10.411106          5.830892          6.871634         11.105300 
> 
```

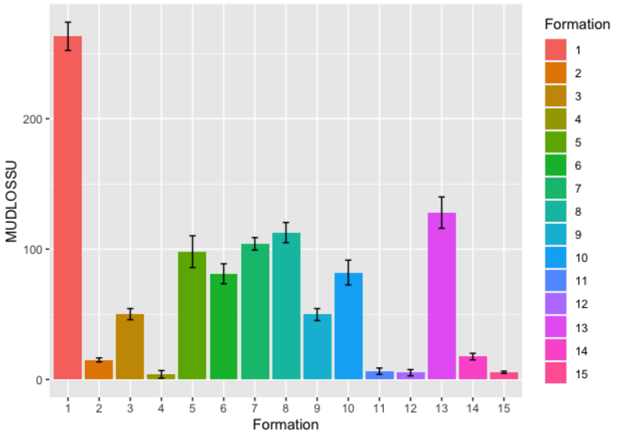
Therefore, backwards elimination would help simplify the model while retaining the relevant variables. After backward elimination, I assessed the VIF values for the remaining variables. High VIF values might indicate multicollinearity which could cause potential issues in my model. Hence, I removed the variables that I deemed fit, which resulted in a more interpretable and effective linear regression model.

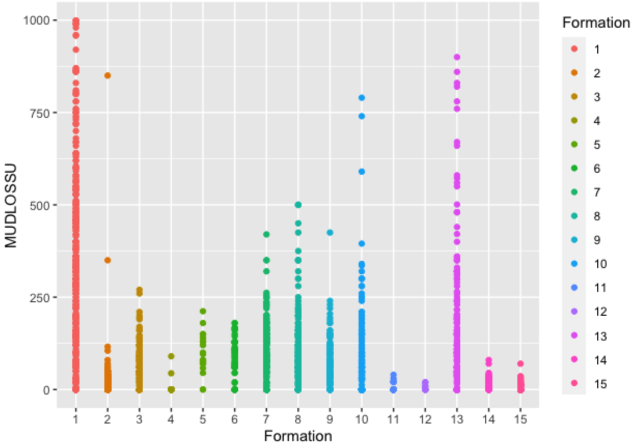
#3
Formation is a categorical variable. This will be further explained in Question 2.
The output below shows the number of data points for each formation. If there were no biasness in collection of sample data, we can infer that Formation 1 and 2 are the most common types of formation.



```
> formation_counts <- table(df$Formation)
> print(formation_counts)

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15 
516 627 177  34  20  52 240 183 187 169  24  11 231  52 145 
> 
```

Below are the Bar Charts & Scatter Plots, with x as my Formation and y as my MUDLOSSU (loss circulation)



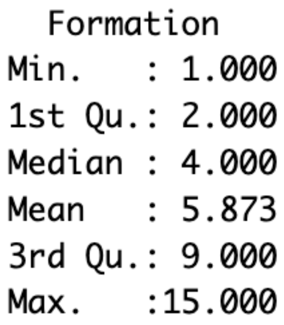


It can be interpreted that Formation 1 has the highest loss circulation, followed by Formation 13. Additionally, Formation 2 exhibits a relatively lower occurrence of loss circulation. This observation is considered to be more reliable due to the larger sample size associated with this formation.
As a result, it is advisable to avoid Formation 1 during drilling operations. The limited sample size for formations like 11, 12, and 14 makes it uncertain whether they genuinely have a lower occurrence of loss circulation. In contrast, Formation 2 provides a more dependable basis for analysis and should be a preferred choice for drilling.

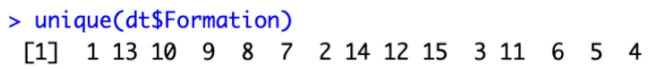
<u>Answer to Q2:</u>

2. State the categorical variable(s) (if any) and explain how you will handle the missing values in the data (if any).

The categorical variable I have identified is the variable: **Formation.**
When finding the unique values for formation, the output is as shown. It is noted than the unique values is rather small, for a dataset of 2668 points.



```
  Formation
Min.   : 1.000
1st Qu.: 2.000
Median : 4.000
Mean   : 5.873
3rd Qu.: 9.000
Max.   :15.000
```

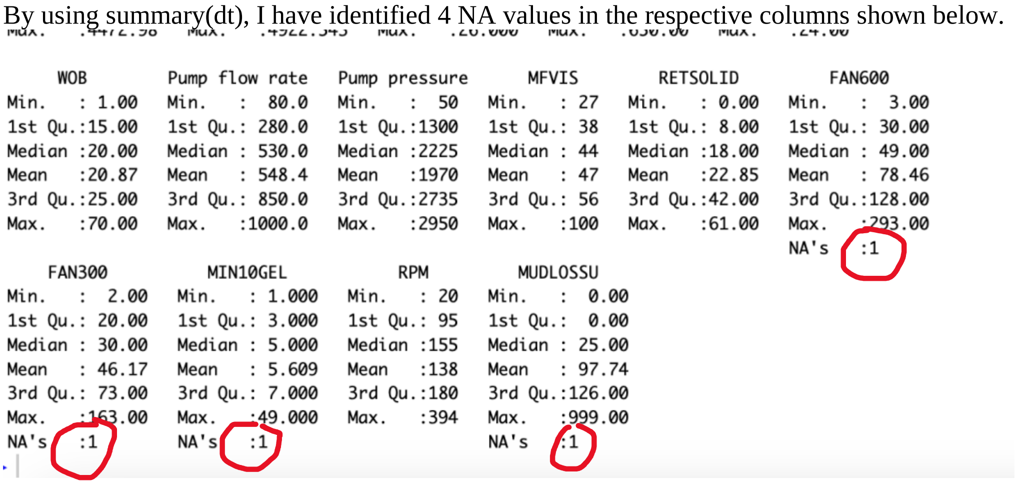


```
> unique(dt$Formation)
 [1]  1 13 10  9  8  7  2 14 12 15  3 11  6  5  4
```

In table 1 of Sabah 2019, under item 16 is Formation Type. It is the only variable that has no measurement units indicated in the table, hence I believe that "Formation Type" in the paper refers to "Formation" in the dataset, Moreover, it was stated in the article "Key Formations drilled to penetrate the reservoir of Marun oil field are Aghajery, Mishan, Gachsaran (GS7, GS6, GS5, GS4, GS3, GS2, Cap rock) and Asmari" , which enhances my belief that formation is a categorical variable.

Hence, the variable is categorical because they represent distinct categories or groups rather than a continuous numeric measurement. In the context of my data, I would treat "Formation" as a categorical variable, and these numeric values are just labels for the categories. So, "Formation" would be analysed as a categorical variable with 15 levels or categories. When constructing the models, I have used as.factor to change them into categorical variables.

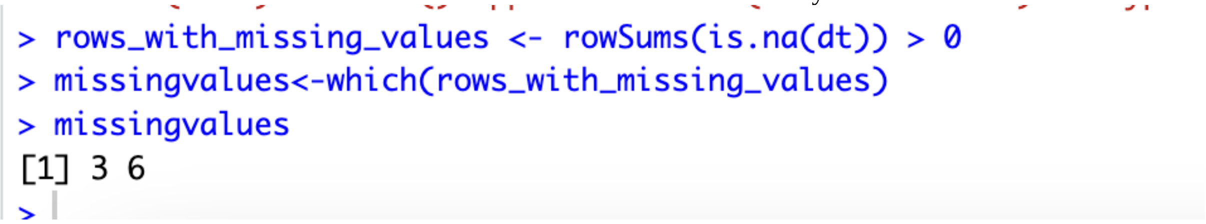
<u>Handling Of NA Values</u>
I have tested 2 methods of handling my NA values.
By using summary(dt), I have identified 4 NA values in the respective columns shown below.



```
      WOB          Pump flow rate   Pump pressure      MFVIS         RETSOLID          FAN600      
 Min.   : 1.00   Min.   :  80.0   Min.   :  50   Min.   : 27   Min.   : 0.00   Min.   :  3.00  
 1st Qu.:15.00   1st Qu.: 280.0   1st Qu.:1300   1st Qu.: 38   1st Qu.: 8.00   1st Qu.: 30.00  
 Median :20.00   Median : 530.0   Median :2225   Median : 44   Median :18.00   Median : 49.00  
 Mean   :20.87   Mean   : 548.4   Mean   :1970   Mean   : 47   Mean   :22.85   Mean   : 78.46  
 3rd Qu.:25.00   3rd Qu.: 850.0   3rd Qu.:2735   3rd Qu.: 56   3rd Qu.:42.00   3rd Qu.:128.00  
 Max.   :70.00   Max.   :1000.0   Max.   :2950   Max.   :100   Max.   :61.00   Max.   :293.00  
                                                                               NA's   :1       
     FAN300          MIN10GEL            RPM          MUDLOSSU     
 Min.   :  2.00   Min.   : 1.000   Min.   : 20   Min.   :  0.00  
 1st Qu.: 20.00   1st Qu.: 3.000   1st Qu.: 95   1st Qu.:  0.00  
 Median : 30.00   Median : 5.000   Median :155   Median : 25.00  
 Mean   : 46.17   Mean   : 5.609   Mean   :138   Mean   : 97.74  
 3rd Qu.: 73.00   3rd Qu.: 7.000   3rd Qu.:180   3rd Qu.:126.00  
 Max.   :163.00   Max.   :49.000   Max.   :394   Max.   :999.00  
 NA's   :1        NA's   :1                      NA's   :1       
```

First Method: Removing the NA rows

I checked the rows of these columns that contained these NA values and they were rows 3 and 6.

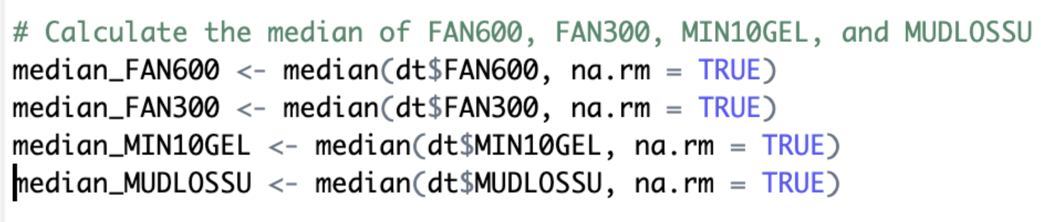


```
> rows_with_missing_values <- rowSums(is.na(dt)) > 0
> missingvalues<-which(rows_with_missing_values)
> missingvalues
[1] 3 6
> 
```

There are 2668 rows in the dataset, however the NULL values are just in 2 of those rows. If these rows were removed, it will not significantly impact my analysis and sample size.

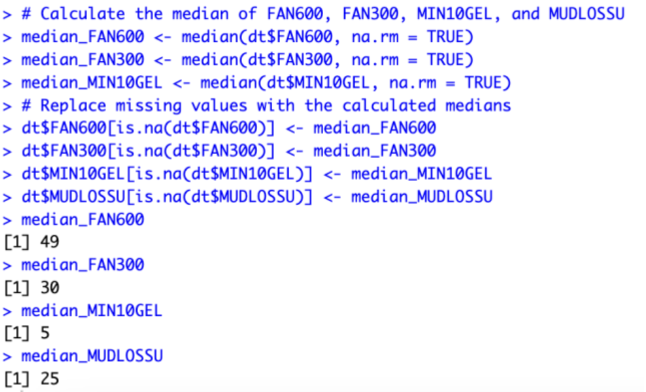
Second method: Replacing NULL values with median

I computed the median values of the columns which has NULL values.



```
# Calculate the median of FAN600, FAN300, MIN10GEL, and MUDLOSSU
median_FAN600 <- median(dt$FAN600, na.rm = TRUE)
median_FAN300 <- median(dt$FAN300, na.rm = TRUE)
median_MIN10GEL <- median(dt$MIN10GEL, na.rm = TRUE)
median_MUDLOSSU <- median(dt$MUDLOSSU, na.rm = TRUE)
```

I then replaced the NULL values with the calculated medians.



```
> # Calculate the median of FAN600, FAN300, MIN10GEL, and MUDLOSSU
> median_FAN600 <- median(dt$FAN600, na.rm = TRUE)
> median_FAN300 <- median(dt$FAN300, na.rm = TRUE)
> median_MIN10GEL <- median(dt$MIN10GEL, na.rm = TRUE)
> # Replace missing values with the calculated medians
> dt$FAN600[is.na(dt$FAN600)] <- median_FAN600
> dt$FAN300[is.na(dt$FAN300)] <- median_FAN300
> dt$MIN10GEL[is.na(dt$MIN10GEL)] <- median_MIN10GEL
> dt$MUDLOSSU[is.na(dt$MUDLOSSU)] <- median_MUDLOSSU
> median_FAN600
[1] 49
> median_FAN300
[1] 30
> median_MIN10GEL
[1] 5
> median_MUDLOSSU
[1] 25
```

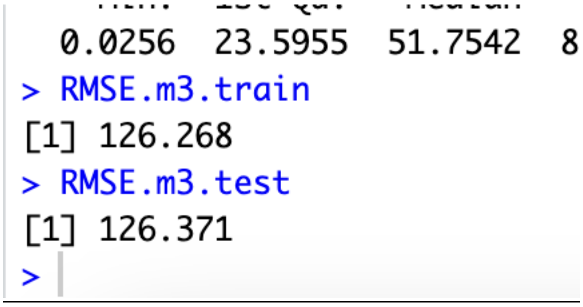
I have decided to use the second method because I managed to obtain a better performance in my RMSE errors in the linear regression and CART models.

Answer to Q3:

3. Using 70-30 train-test, conduct (a) Linear Regression and (b) CART to compare the trainset and testset errors. Display the results in a table. Example:

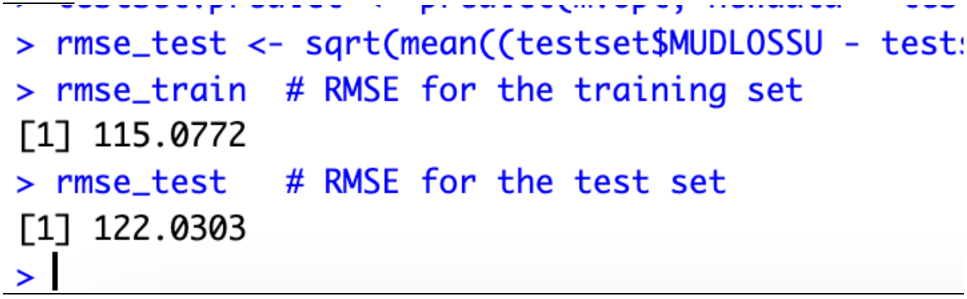
| Model | Complexity | Train Set RMSE | Test Set RMSE |
|---|---|---|---|
| Linear Reg | Categorical: 1<br>Continuous: 7 | 126.268 | 126.371 |
| CART | 5 | 115.0772 | 122.0303 |

Linear Regression



```
  0.0256  23.5955  51.7542  8
> RMSE.m3.train
[1] 126.268
> RMSE.m3.test
[1] 126.371
> 
```

CART



```
> rmse_test <- sqrt(mean((testset$MUDLOSSU - test
> rmse_train  # RMSE for the training set
[1] 115.0772
> rmse_test   # RMSE for the test set
[1] 122.0303
> 
```

Answer to Q4:

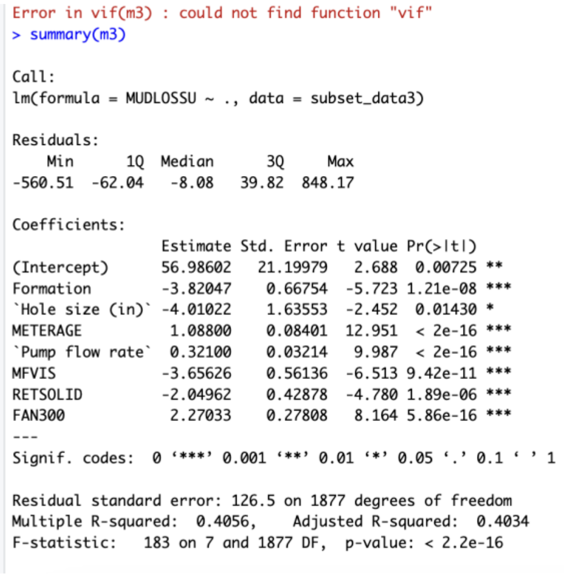
4. Comment on your models' results and provide insights for the business application.

It can be observed that the linear regression model uses 7 predictor variables whereas the CART model has only 5 terminal nodes. This indicates that the CART model is a simpler model. However, the simplest model might not always be the best. We need to juggle between inaccuracy and complexity.

When comparing the test errors between both models, it is observed that the CART model has a lowest test set and train set error than the linear regression model, indicating that the CART model is fitting the training data better. In a nutshell, the CART model is performing better in terms of both train and test set errors.
However, it is noticed that for linear regression, the difference between test and train set errors are smaller, suggesting that the model is generalising well to new unseen data. It is ideal for training and test set errors to be as close as possible.

In terms of business application, the better model to use would be based on the primary objective. If you want a model that **performs consistently well on new unseen data,** the goal would be to minimise the difference between the train and test set errors, which would indicate better generalisation. Hence, in this case, **a linear regression model would be the better choice.**
However, if the goal is **to have model make the most accurate predictions**, the model which has a lower test and train set error should be chosen, which would be **CART** in this case.
Ultimately, the choice of the best model depends on the business objective. It is important to evaluate multiple metrics and access the trade-offs involved before coming to a consensus.

a) *If the linear regression model is used: We can use this regression equation to predict the amount of loss circulation.*



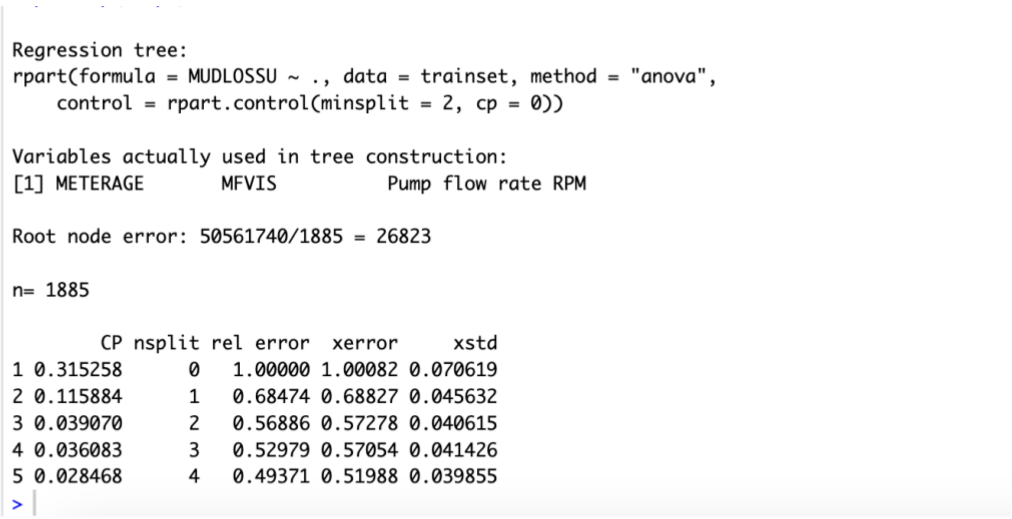
```
Error in vif(m3) : could not find function "vif"
> summary(m3)

Call:
lm(formula = MUDLOSSU ~ ., data = subset_data3)

Residuals:
    Min      1Q  Median      3Q     Max 
-560.51  -62.04   -8.08   39.82  848.17 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      56.98602   21.19979   2.688  0.00725 ** 
Formation        -3.82047    0.66754  -5.723 1.21e-08 ***
`Hole size (in)` -4.01022    1.63553  -2.452  0.01430 *  
METERAGE          1.08800    0.08401  12.951  < 2e-16 ***
`Pump flow rate`  0.32100    0.03214   9.987  < 2e-16 ***
MFVIS            -3.65626    0.56136  -6.513 9.42e-11 ***
RETSOLID         -2.04962    0.42878  -4.780 1.89e-06 ***
FAN300            2.27033    0.27808   8.164 5.86e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 126.5 on 1877 degrees of freedom
Multiple R-squared:  0.4056,	Adjusted R-squared:  0.4034 
F-statistic:   183 on 7 and 1877 DF,  p-value: < 2.2e-16
```

**MUDLOSSU = 56.98602−3.82047*Formation−4.01022*Hole size (in)+1.08800*METERAGE+0.32100*Pump flow rate−3.65626*MFVIS−2.04962*RETSOLID+2.27033*FAN300**

We can observe that the variables: Formation, Hole Size, MFVIS and RETSOLID has a negative relationship with MUDLOSSU. This equation helps the business to better plan ways to reduce MUDLOSSU, by controlling these variables in the equation.
(Other predictor variables were removed due to the backwards elimination method and by using vif() to identify potential multicollinearity issues.)

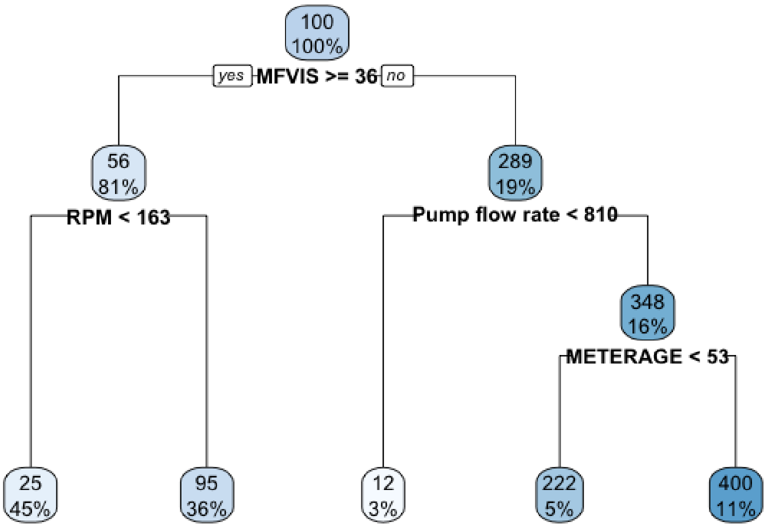
b) *If the CART model is used: We can use this decision tree below to predict the amount of loss circulation.*

From the results shown in the table below, we can observe that the 5th tree is optimal with a CP value of 0.028468.



```
Regression tree:
rpart(formula = MUDLOSSU ~ ., data = trainset, method = "anova",
    control = rpart.control(minsplit = 2, cp = 0))

Variables actually used in tree construction:
[1] METERAGE       MFVIS          Pump flow rate RPM

Root node error: 50561740/1885 = 26823

n= 1885

        CP nsplit rel error  xerror     xstd
1 0.315258      0   1.00000 1.00082 0.070619
2 0.115884      1   0.68474 0.68827 0.045632
3 0.039070      2   0.56886 0.57278 0.040615
4 0.036083      3   0.52979 0.57054 0.041426
5 0.028468      4   0.49371 0.51988 0.039855
>
```

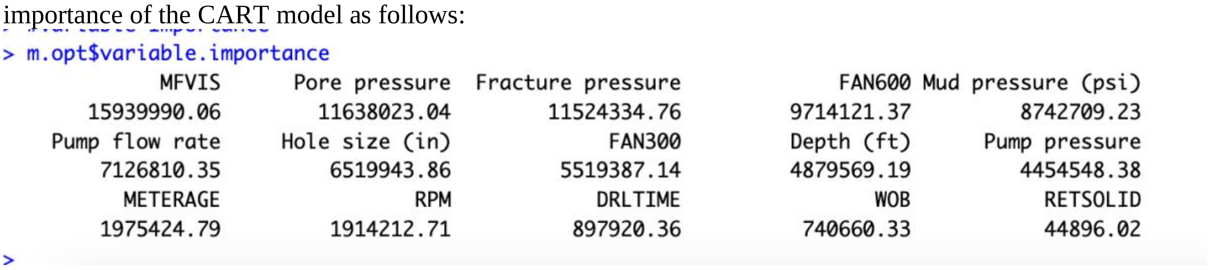
My decision tree diagram is as follows:



<u>Interpreting the results</u>
It shows that "MFVIS" is used as the root node, suggesting that "MFVIS" is a highly effective variable in discrimination amongst the data. Similarly, "RPM" or "Pump flow rate" is used for splitting at deeper nodes, indicating that these variables are also effective at further separating the data within specific branches of the tree.
Based on the decision tree, if the MFVIS < 36, it will continue into the right branch. On the right branch, if the Pump flow rate < 819, we can assume that the amount of loss circulation will be 12bb/hr. If it is > = 819, it continues onto the right branch, to another node. Here, if METERGAE <53, loss circulation will be 222bb/hr whereas if METERAGE >= 53, loss circulation will be 400bb/hr
If the MFVIS >= 36, it will continue onto the left branch onto another node. If RPM <163, loss circulation is predicted to be 25 bb/hr whereas if RPM >= 163, loss circulation is predicted to be 95 bb/hr.
We can also notice that majority of the data falls under the left branch of the RPM condition. (Given 45% + 36% = 81% ). Hence, it is more likely for the loss circulation to be ranging around these values, rather than extremes such as 12, 222 or 400.

Variable Importance

Furthermore, we can observe that the variable importance has been ranked by the model, indicating that MFVIS is the most important with RETSOLID having the least importance. It is interesting to note that the variable importance of the CART model as follows:



```
> m.opt$variable.importance
            MFVIS    Pore pressure Fracture pressure             FAN600 Mud pressure (psi)
      15939990.06      11638023.04       11524334.76         9714121.37         8742709.23
   Pump flow rate    Hole size (in)            FAN300        Depth (ft)      Pump pressure
       7126810.35       6519943.86        5519387.14         4879569.19         4454548.38
         METERAGE               RPM           DRLTIME                WOB           RETSOLID
       1975424.79       1914212.71         897920.36          740660.33           44896.02
>
```

- MFVIS," "Pore pressure," "Fracture pressure," "FAN600," and "Mud pressure (psi)" have the highest importance values. When splitting the data into nodes, these variables have the most substantial impact on reducing impurity or error.
- "Pump flow rate," "Hole size (in)," "FAN300," "Depth (ft)," and "Pump pressure" also have relatively high importance values, indicating their importance in the decision tree.
- "METERAGE," "RPM," "DRLTIME," "WOB," and "RETSOLID" have lower importance values, suggesting that they have less impact on the tree's decision-making process.

From this, we can infer the business should aim to reduce high importance values such as **MFVIS, Pore pressure and Fracture pressure,** which has a significant impact in reducing MUDLOSSU.

Answer to Q5:

5. Comment on Sabah M. et. al. (2019) use of the CART model.

The CART model was used with a 5-fold cross validation. The regression tree consists of a root node, 35 intermediate nodes and 37 terminal nodes. Firstly, the initial split at the root node is based on the viscosity, and has a splitting value of 35.5 cp. This means that if the viscosity is < than 35.5cp, the split will then depend on the flow rate of the drilling fluid whereas if >35.5cp, the nodes will be further divided based on the bit rotational speed. This process continues until it reaches the terminal nodes.

The use of CART model has limitations, one being that it is prone to over fitting.  The more variables used, the higher the chances of multicollinearity and cause an issue of overfitting. This may lead to a tree that is too complex and not suitable for making accurate predictions.

Additionally, the CART model has a bias towards variables with many categories. This is due to the tree building algorithm which may find it beneficial to split the data based on the various level of categories. In this context, formation may be considered as a categorical variable with 15 levels, hence caution must be exercised to reduce the impact of categorical variables with numerous categories on the regression tree.

By addressing these aspects, the study can provide a more robust and insightful analysis of the CART model's performance and its relevance to the problem of predicting lost circulation.

Answer to Q6:

6. Comment on Sabah M. et. al. (2019) use of the 4 performance measures in section 6 and table 5 of the research paper.

Comparing the use of the 4 performance measures:

RMSE quantifies how far, on average, the predicted values are from the actual values. It is used commonly as a performance indicators to evaluate the accuracy of a prediction model. From Table 5, we can observe that the Decision Tree produced the lowest RMSE, hence it is the most accurate model out of the 5 models tested.

R^2 is also used as a performance measure and is similar to RMSE. A higher R^2 value, would mean that the model is explaining the variability well. Table 5 shows that again, the Decision Tree has the highest R^2 value, making it the best fit model.

VAF is used to quantify the proportion of variance in a outcome variable that is explained by one or more predictor variables. VAF is typically expressed as a percentage. A higher VAF value, would mean that the Variance that was present in the regression model has been accounted for. Similarly, the Decision Tree has the highest VAF(%) value, making it the most accurate regression model

Lastly, PI is a metric used to assess how the model's overall performance and is also used to evaluate accuracy. It considers the 3 performance measures that were mentioned previously when evaluating the accuracy. We can conclude that Decision Tree has the highest PI value, and is the most accurate model.

These 4 performance indicators has proven that Decision Tree is the most accurate model. However, one limitation is that the 4 performance indicators only takes into account the accuracy of the prediction against actual values. It fails to take account multicollinearity, which is measured by VAF. A high VAF value could potentially mean multicollinearity amongst predictor variables.

Additionally, The study lacks a proper comparison between the initial and optimized models. This comparison could have shed light on how the variable selection and optimization process impacted the model's ability to explain the variance in the target variable, rather than just doing a comparison against the optimal models of the 5 different models executed.

By failing to compare the performance indices between initial and optimal, it fails to answer questions like eg. Certain variables becoming more or less important after optimisation, or whether the optimisation process helped improve the model's ability to explain variance. The author would have been able to make more informed decisions on the variables to include or exclude in the final model, which could enhance accuracy in predicting loss circulation.