# XN Project: Individual Draft Presentation

Esha Mulki
Mohammad Mohavedhi
Ajoy Kumar Nandakumar
Taiye Murtala
College of Professional Studies,
Northeastern University
ALY6080: Integrated Experiential Learning
Instructor: Dr. Matthew Goodwin

# Introduction

- Danish-based global corporation called Danfoss.
- Products made by Danfoss include those for the automotive, building, energy and natural resource, food and beverage, marine, and offshore industries.
- Solutions from Danfoss include those for power, temperature control, and drives..

# Executive Summary

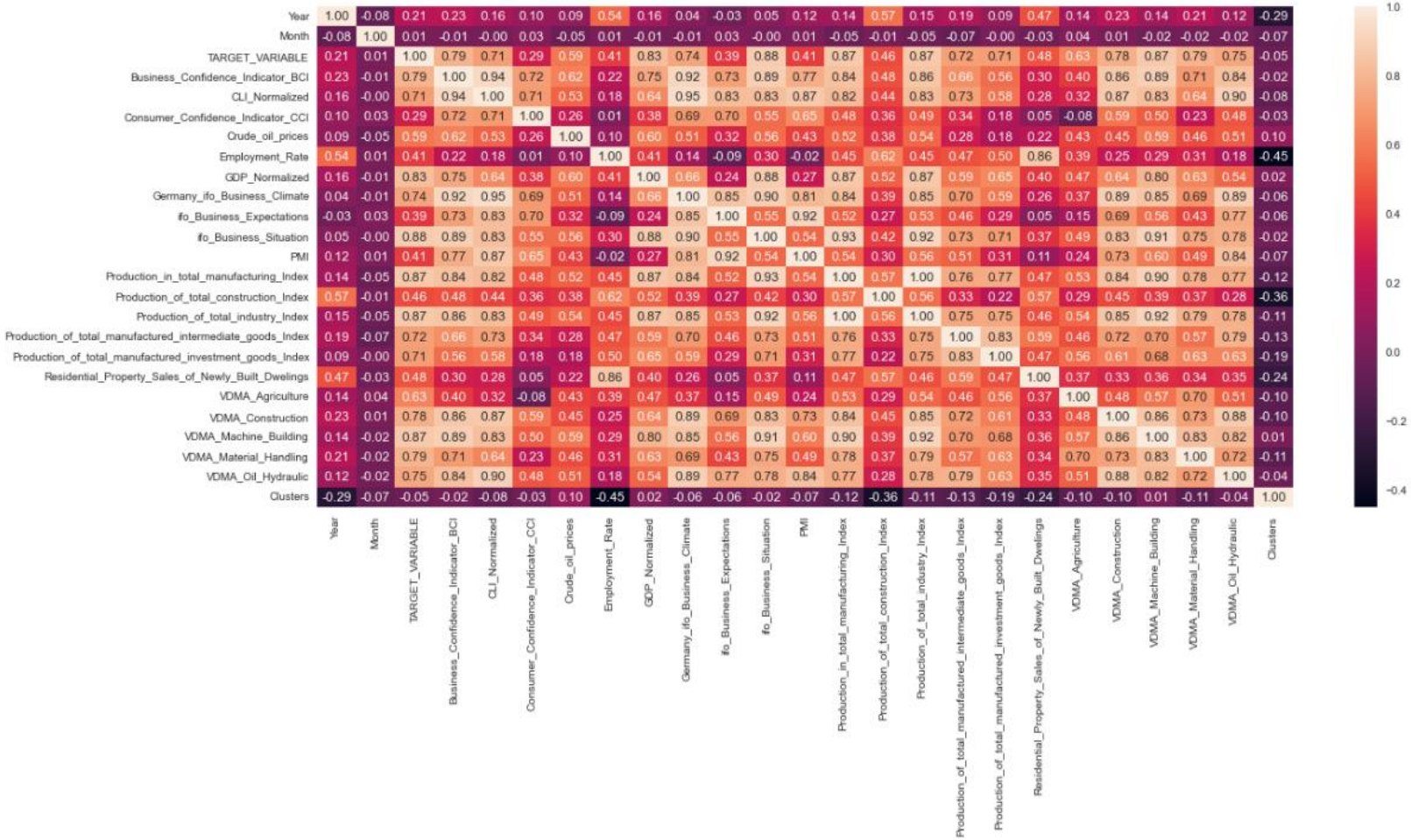
- Improve Danfoss client's sales forecasting.
- Study of the sponsor business sector to discuss the dependent and independent variables with the sponsor.
- Data cleaning, reformatting, and exploratory analysis comprise the initial analysis.
- Used language: R
- Model Evaluation- computed RMSE and MAPE values

# Business Problem

- Due to the pandemic's uncertainty, accurately forecast the sales.
- Not able to control its inventories and avoid both stock-out and overstock problems.
- Managers can estimate revenue and profit using data from accurate sales forecasting.
- The project's objective is to create a precise sales forecasting model for Danfoss.

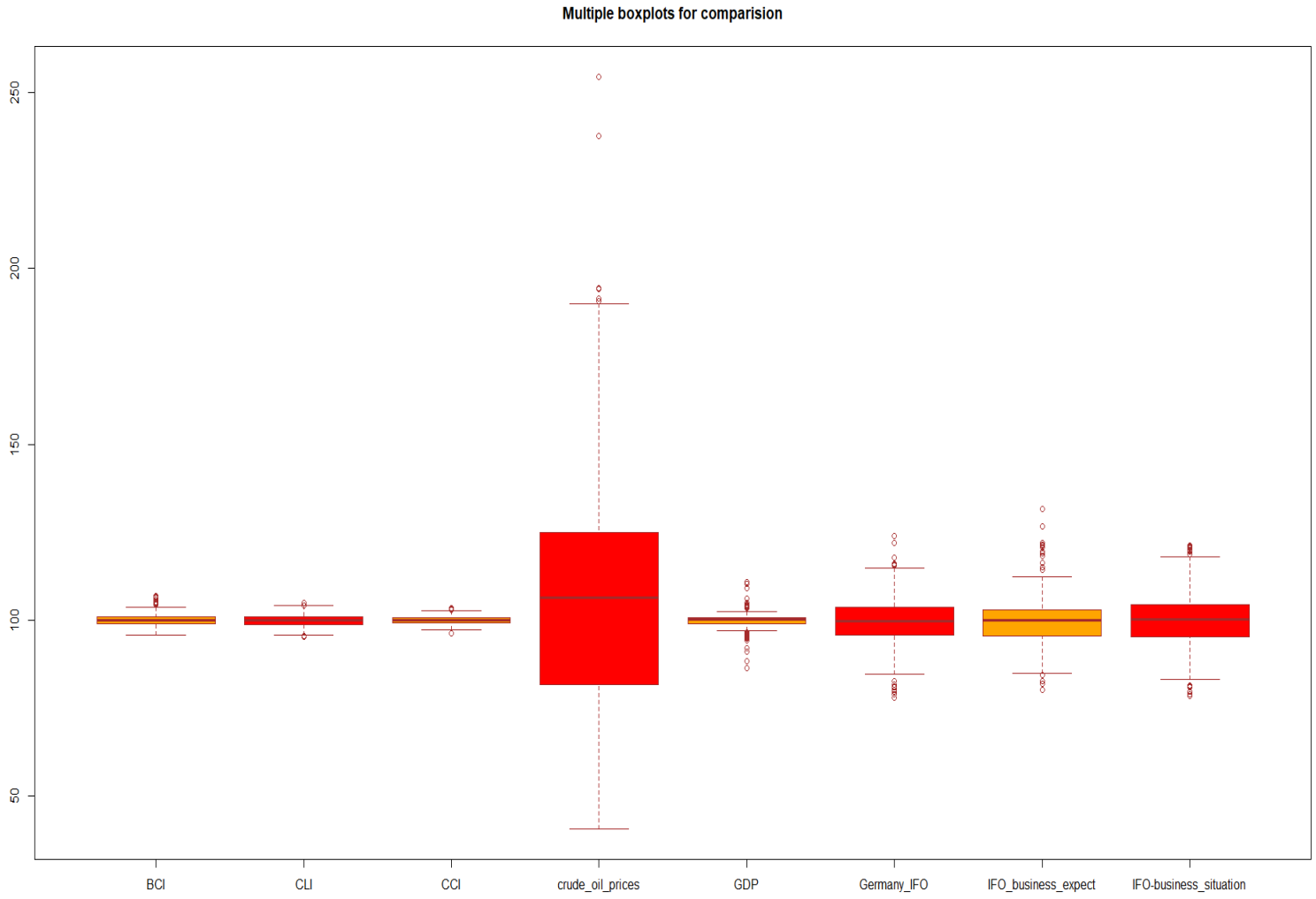
# Visualizations

# Correlation Plot



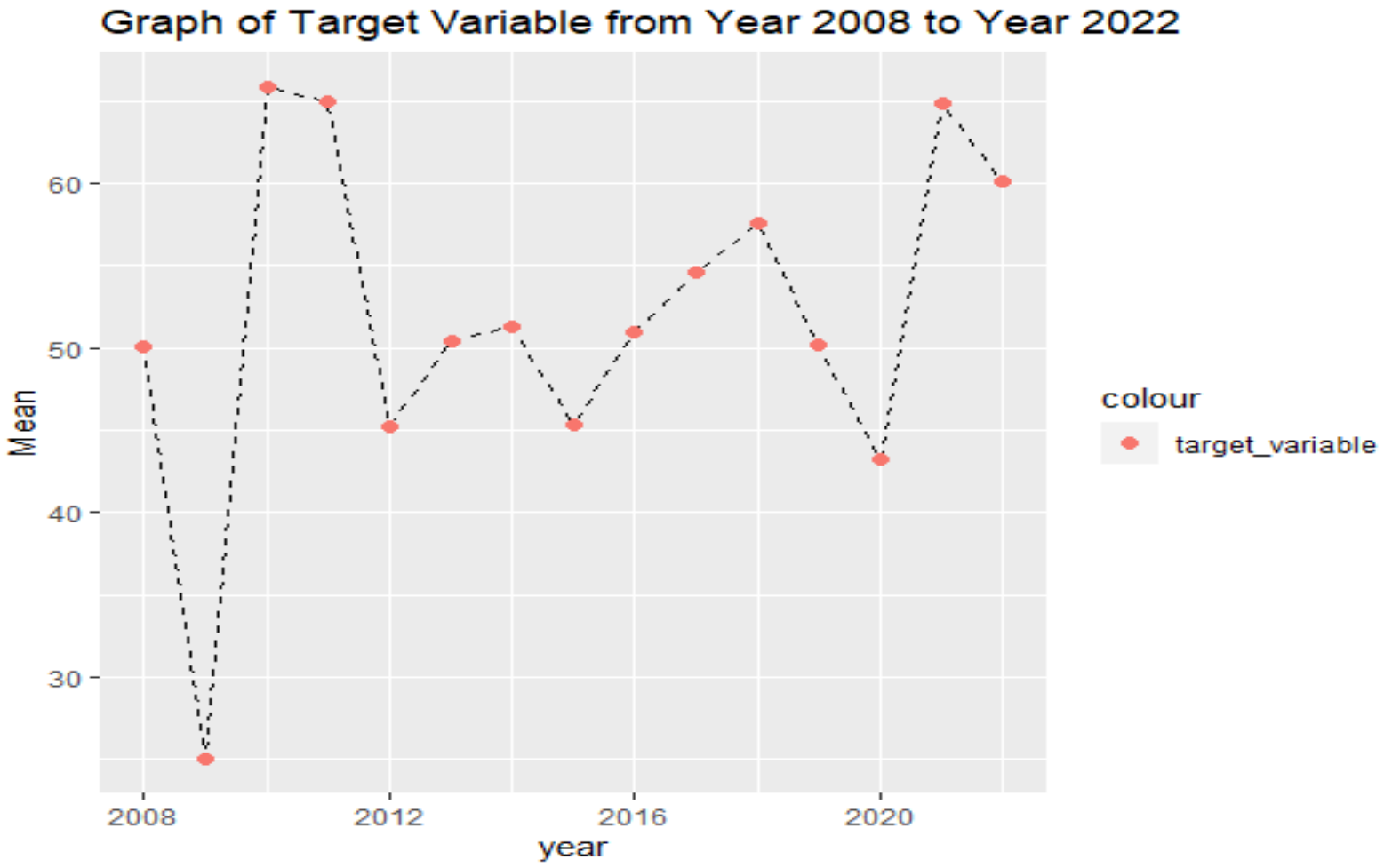

- Target Variable strong correlation 0.88 and 0.87 with VDMA Machine Building and IFO Business Solutions
- 0.83 with GDP Normalized, 0.79 with Business Confidence Index, and 0.79 with VDMA Material Building

# Boxplot



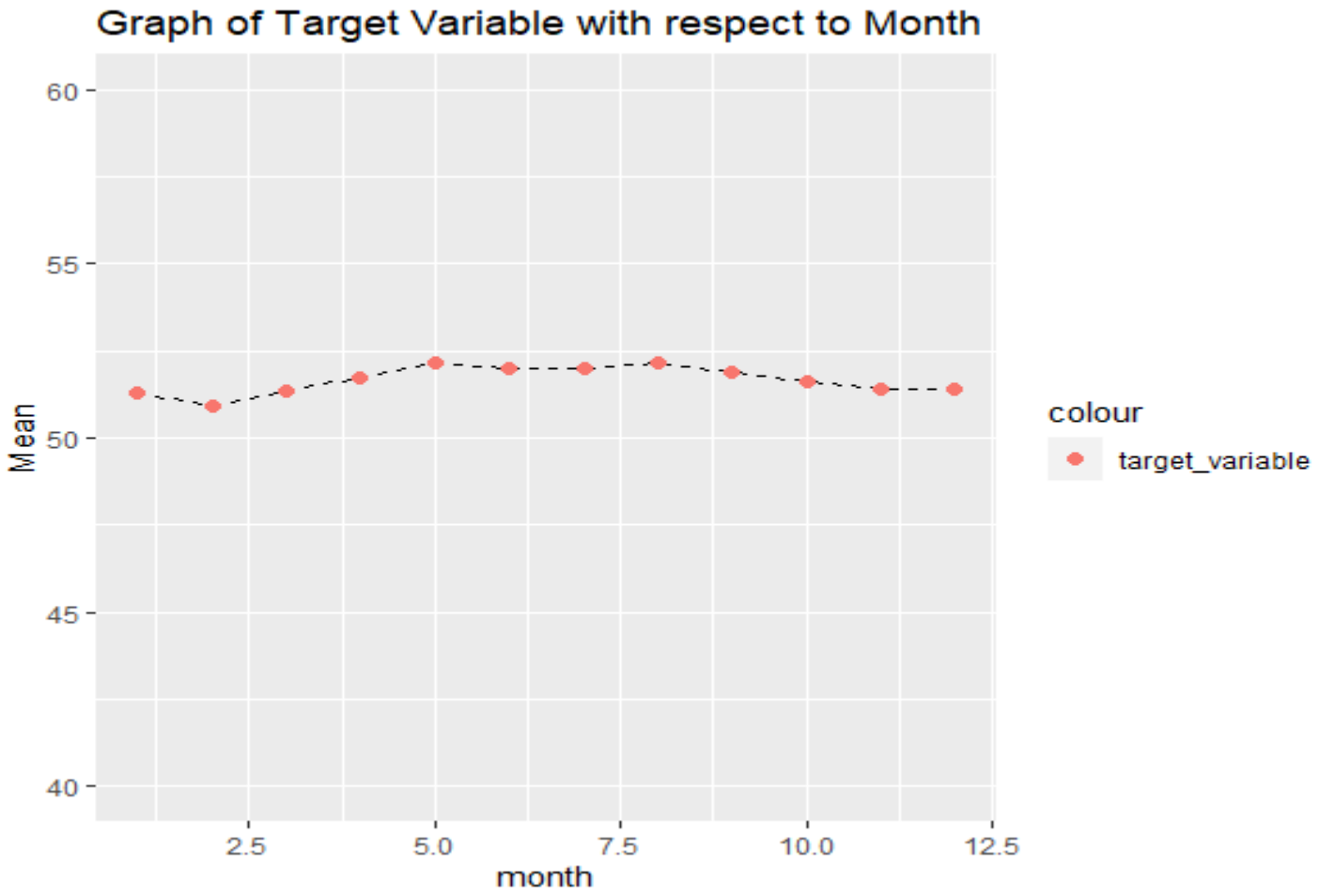

Crude oil prices has huge outliers

# Mean Target Variable vs Year



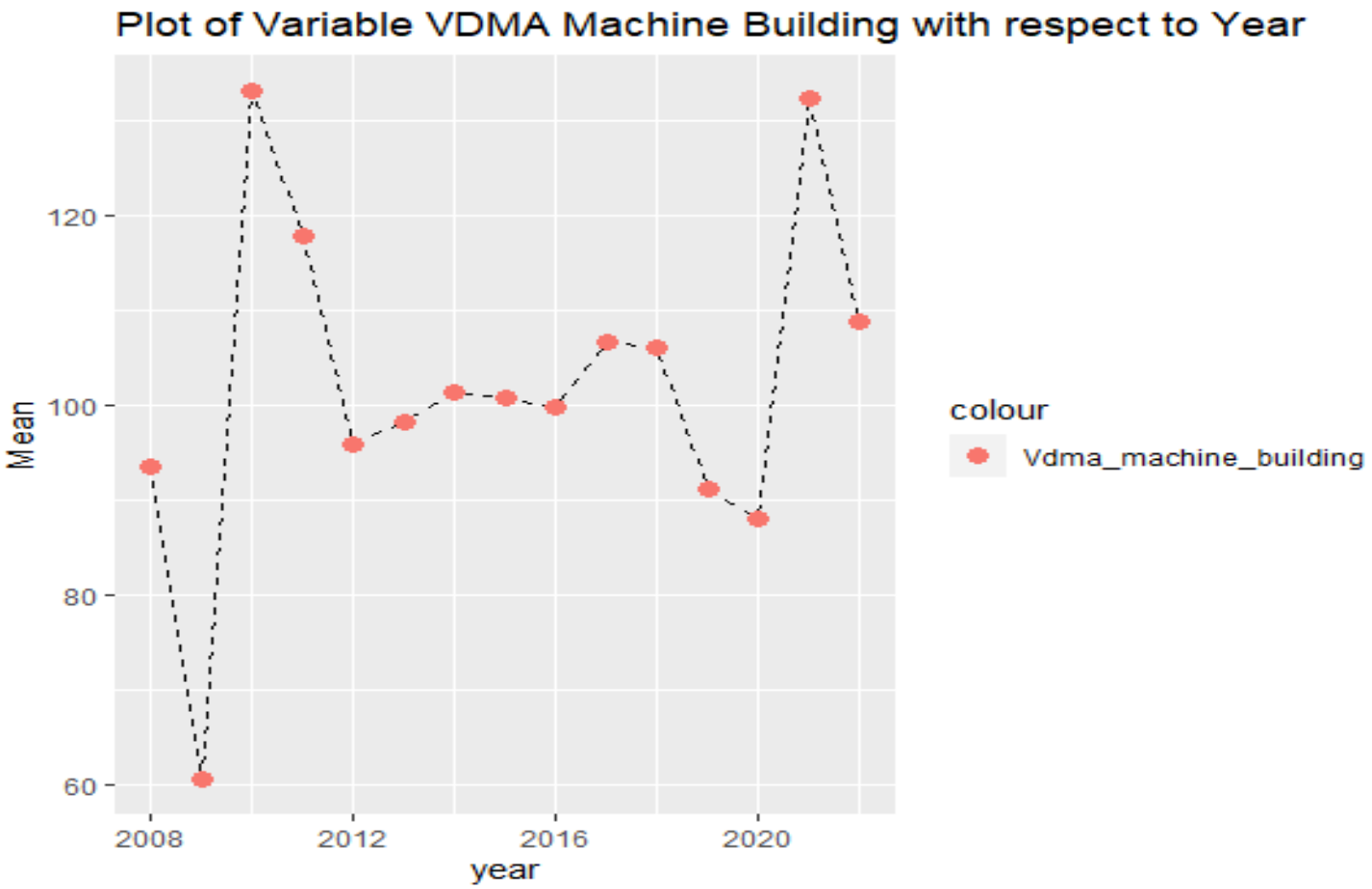

- The mean target variable values fluctuate from 2008 to 2012, but after that point, they become more or less stable until 2018, after which they start to fluctuate till 2022.

# Mean Target Variable vs Month




- There is not much variation in the target variable's graph with regard to the months.

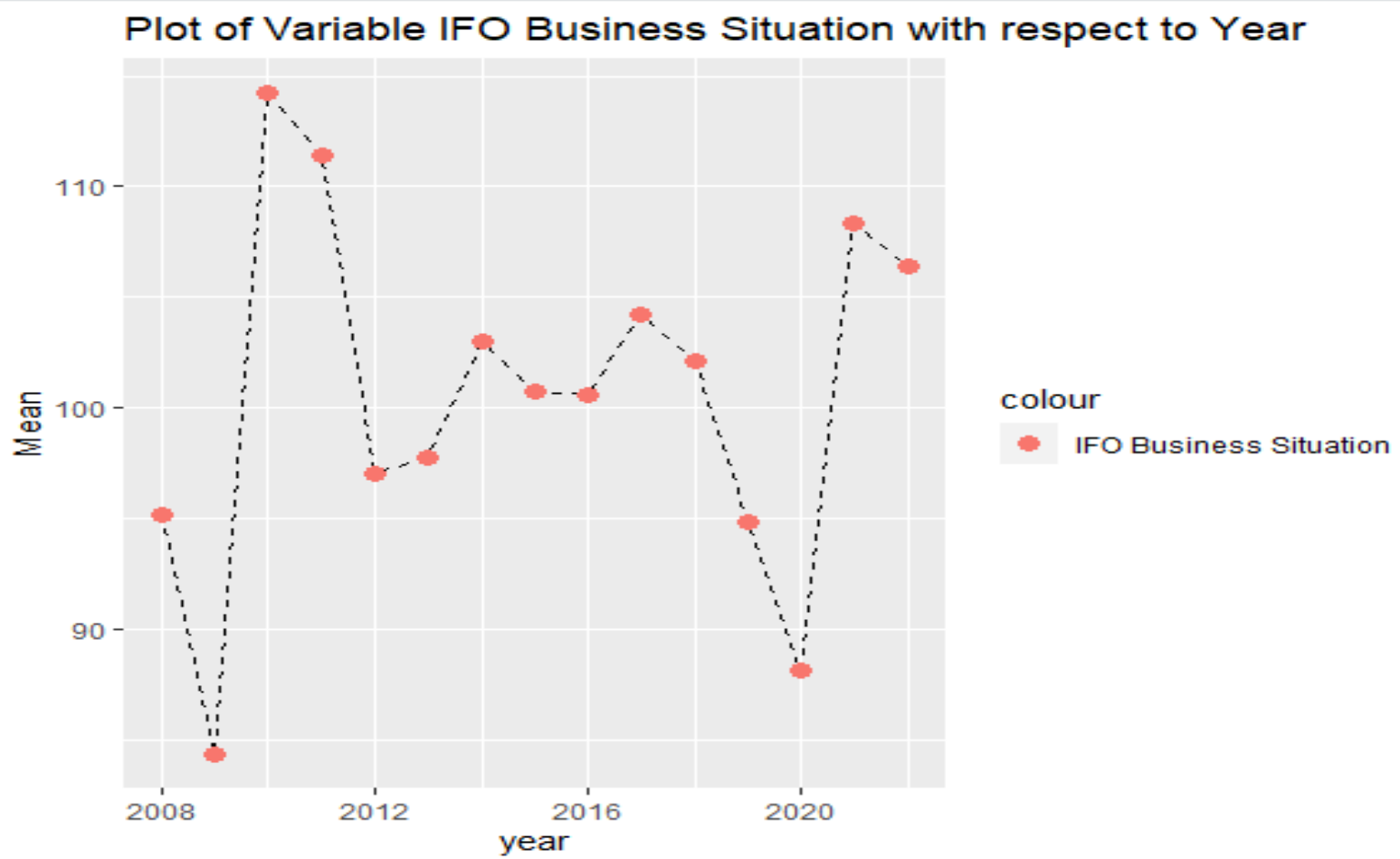
# Mean value VDMA Machine Building vs Years




The mean values of the VDMA machine building variables vary from 2008 to 2012 and from 2019 to 2022.

The values are more or less stable from 2012 to 2018.

# Mean value of IFO Business Situation vs Years




High fluctuation in the value in the period from 2008 to 2012 and between 2020 and 2022

# Milestones of Project

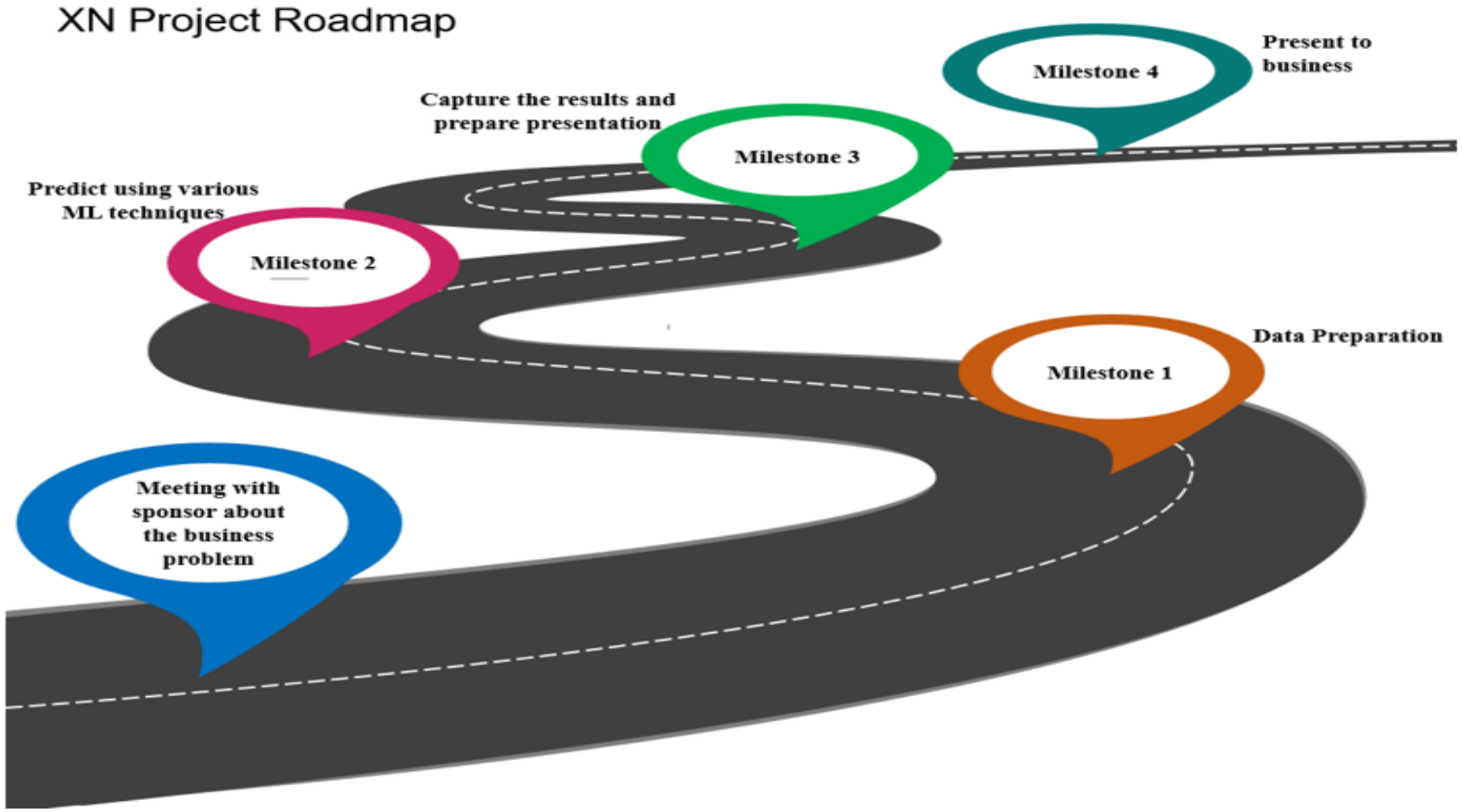



**Fig 1: Milestones for the project**

# Analysis and Findings

- The goal variable and certain other factors, such as the IFO business scenario, VDMA machine building, Production in Total Manufacturing Index, and VDMA Material Building, have a significant link.
- The high collinearities of some of the variables also contribute to multicollinearity.
- The price of crude oil has changed significantly throughout the years, as can be seen from the value of crude oil prices over the years.
- Between the years of 2008 and 2012, it was discovered that the mean IFO business scenario value varied; however, after that, the value remained essentially constant until 2018, at which time it started to vary once more.

# Analysis and Findings(contd)

- The mean target variable values varied between the years of 2008 and 2012, but after that, they stayed largely stable until 2018, when they started to vary once more.
- When compared to the months, it was observed that the target variable does not vary significantly.
- In general, it can be seen that Danfoss's revenue was significantly impacted by both the pandemic that occurred between 2020 and 2022 and the world recession that occurred between 2008 and 2009.

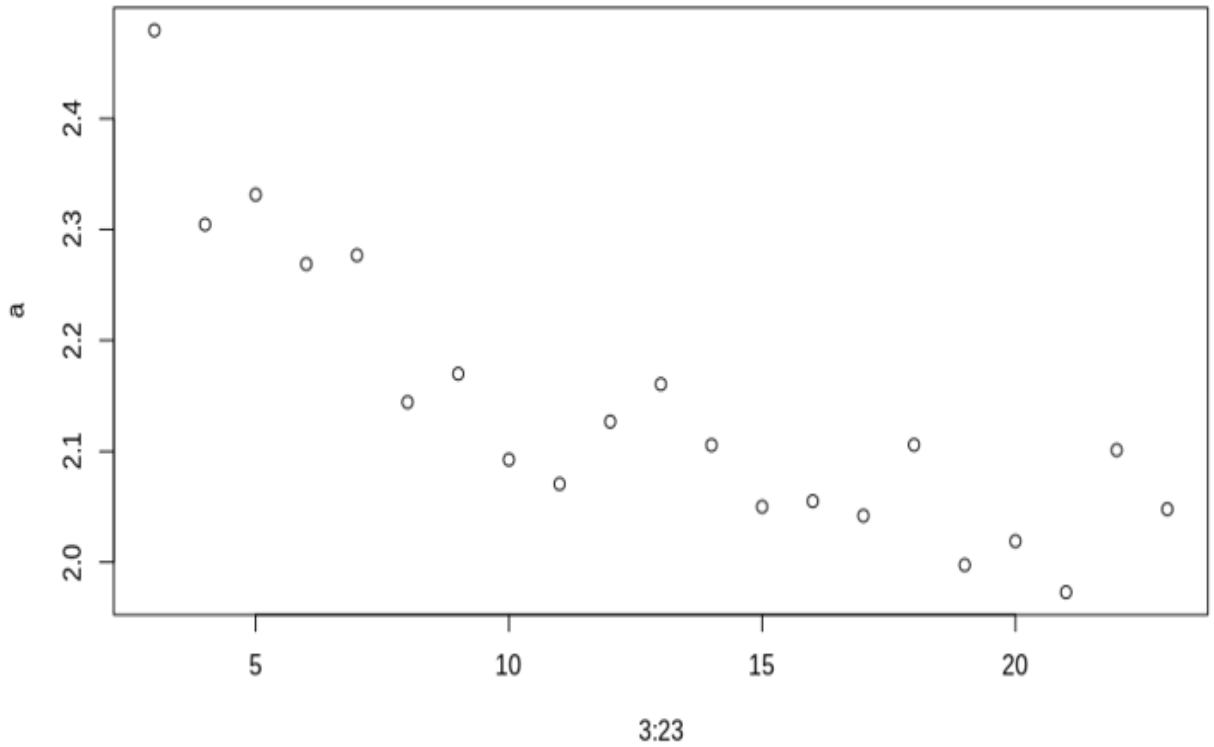
# Model Built

# Prediction using Random Forest

Random Forests - Used in a wide range of industries. Its foundation is the creation of several decision trees, each of which is built using a different subset of your training data.

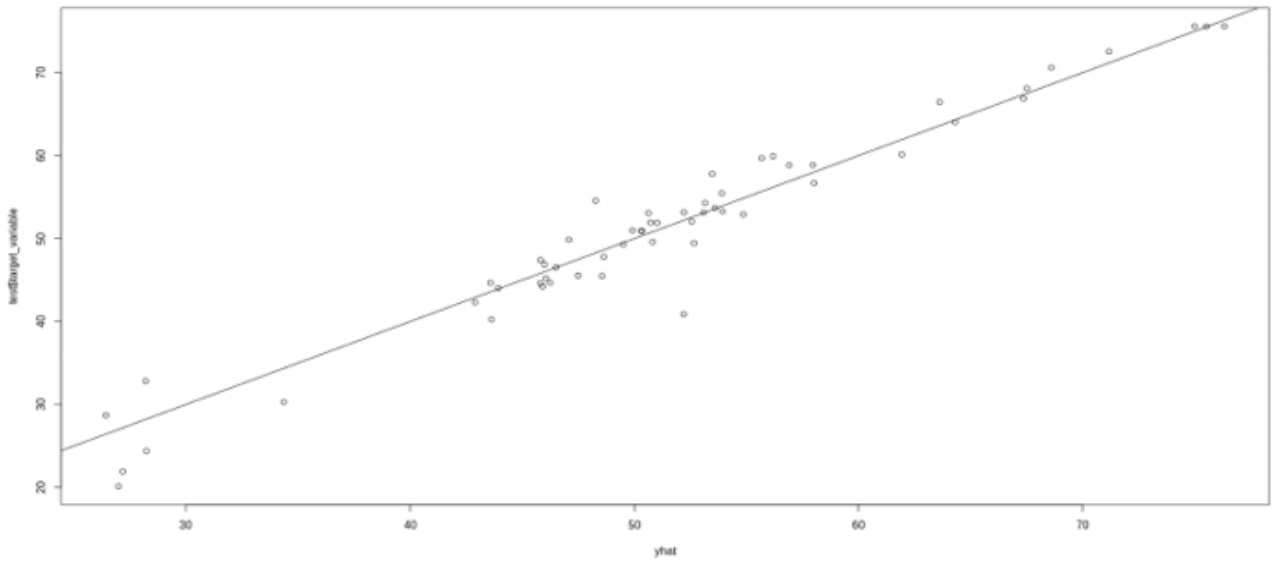
By choosing the most frequent output, a decision trees is created to determine a categorization model (Pires, 2017).

# Prediction using Random Forest



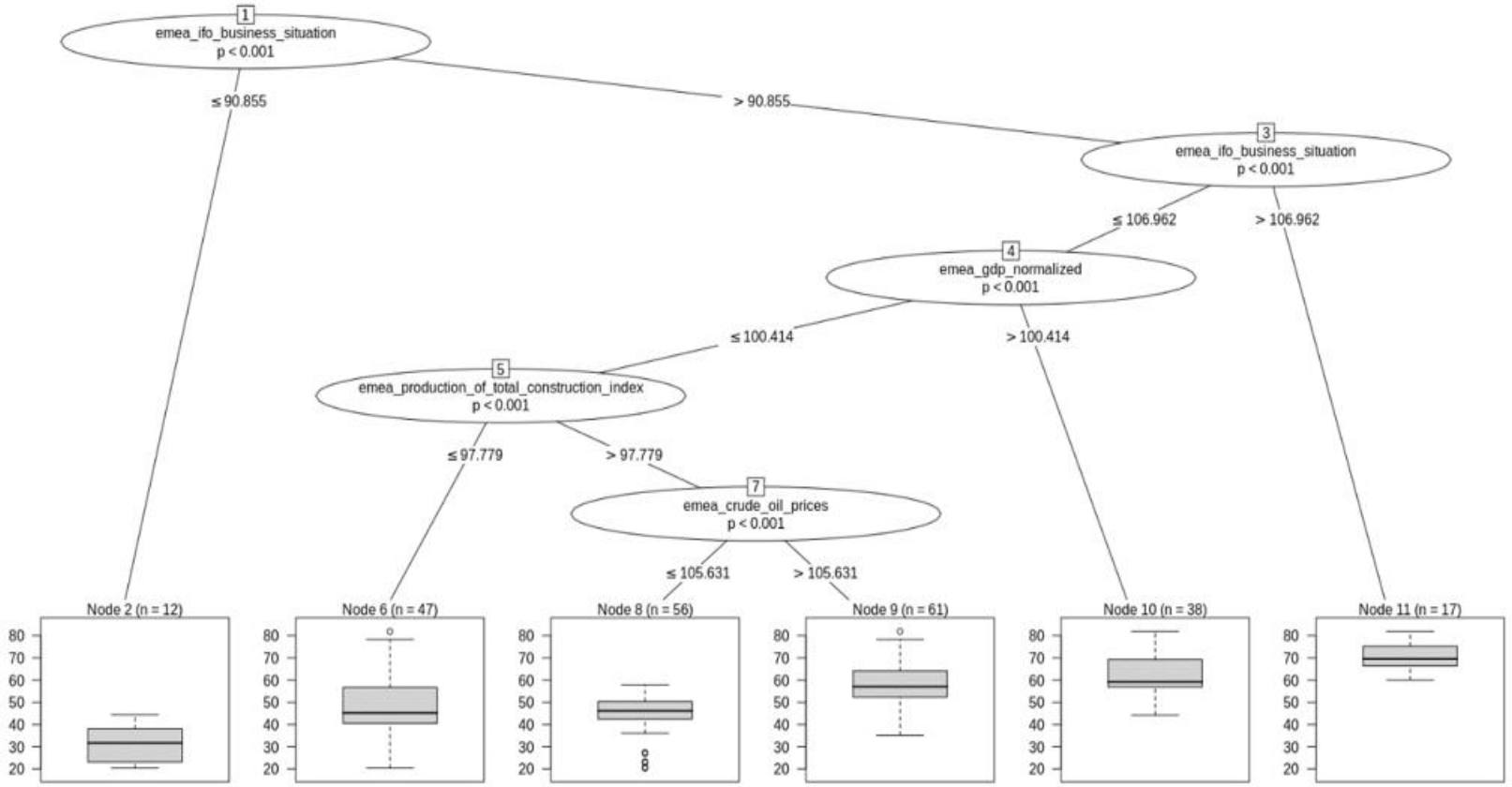

- The code on the loop to tune the random forest.
- The plot below shows the mean of error for each forest based on different parameter or number of variables picked up in each tree.

# Prediction using Random Forest




- Most of the points are close to y=x line indicating that the residuals are almost 0.
- The RMSE value for final random forest is 8.056074

# Prediction using Random Forest




- Decision tree is generated by the randon forest algorithms.
- It's not the most successful tree but it's impressive how it has done a great job in some nodes.

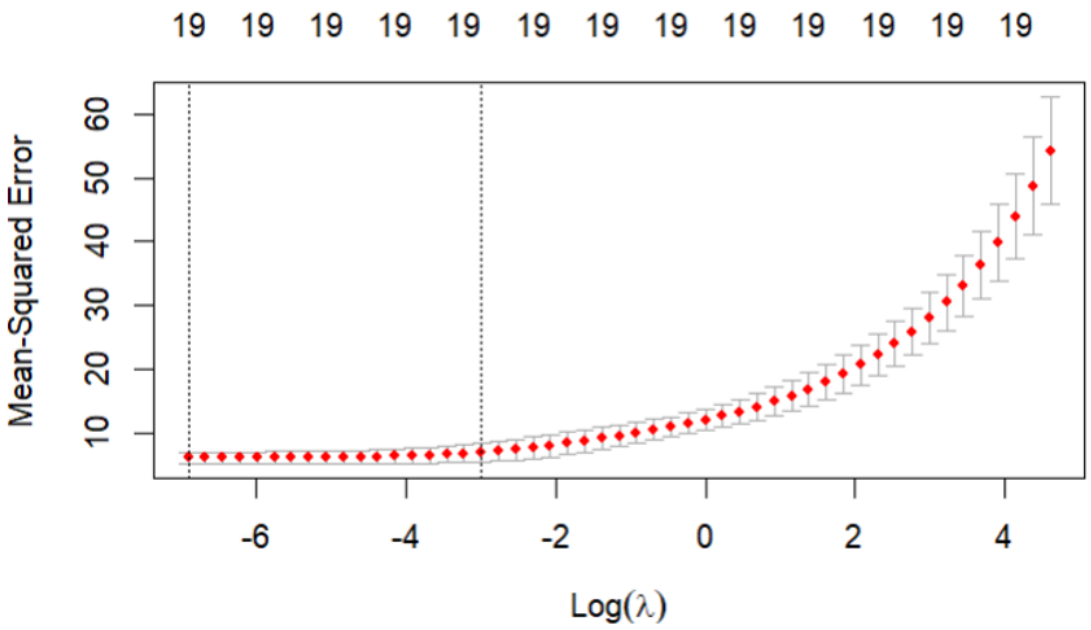
# Prediction using Ridge Regression(Kassambara, 2018)

If there are more variables in a multivariate data set than there are samples, the basic linear model does not perform well.

Regression techniques are penalised when there are too many variables in the model, and this is done by adding a constraint to the equation.

This regularization or shrinkage procedure causes the less significant variables coefficient value to decrease to zero.

# Ridge Regression(L2 Norm)




- The mean-squared error -y-axis, and the log - x-axis.
- The numbers at the top of the graph display the number of variables (non-zero coefficients) that would be kept in the model for the specified value of.
- The dashed line on the right represents the greatest value of within one standard error (1se) of the minimum, which is -6.907755, while the dashed line on the far left represents the least value of lambda that minimises out-of-sample loss, which is -2.993361.
- There are 19 variables in both the minimal and initial models.

# Ridge Regression(L2 Norm)

```
> predictions_train <- predict(ridge_reg, s = optimal_lambda, newx = train_x)
> eval_results(train_y, predictions_train, train_x)
      RMSE   Rsquare
1 1.976778 0.9728505
>
> # Prediction and evaluation on test data
> predictions_test <- predict(ridge_reg, s = optimal_lambda, newx = test_x)
> eval_results(test_y, predictions_test, test_x)
      RMSE   Rsquare
1 2.657362 0.9449558
```

- The RMSE and R square values for the train and test dataset is shown above.
- It can be seen that the RMSE for the test data is higher than the train and the R squared value is lower for the test than the train data.

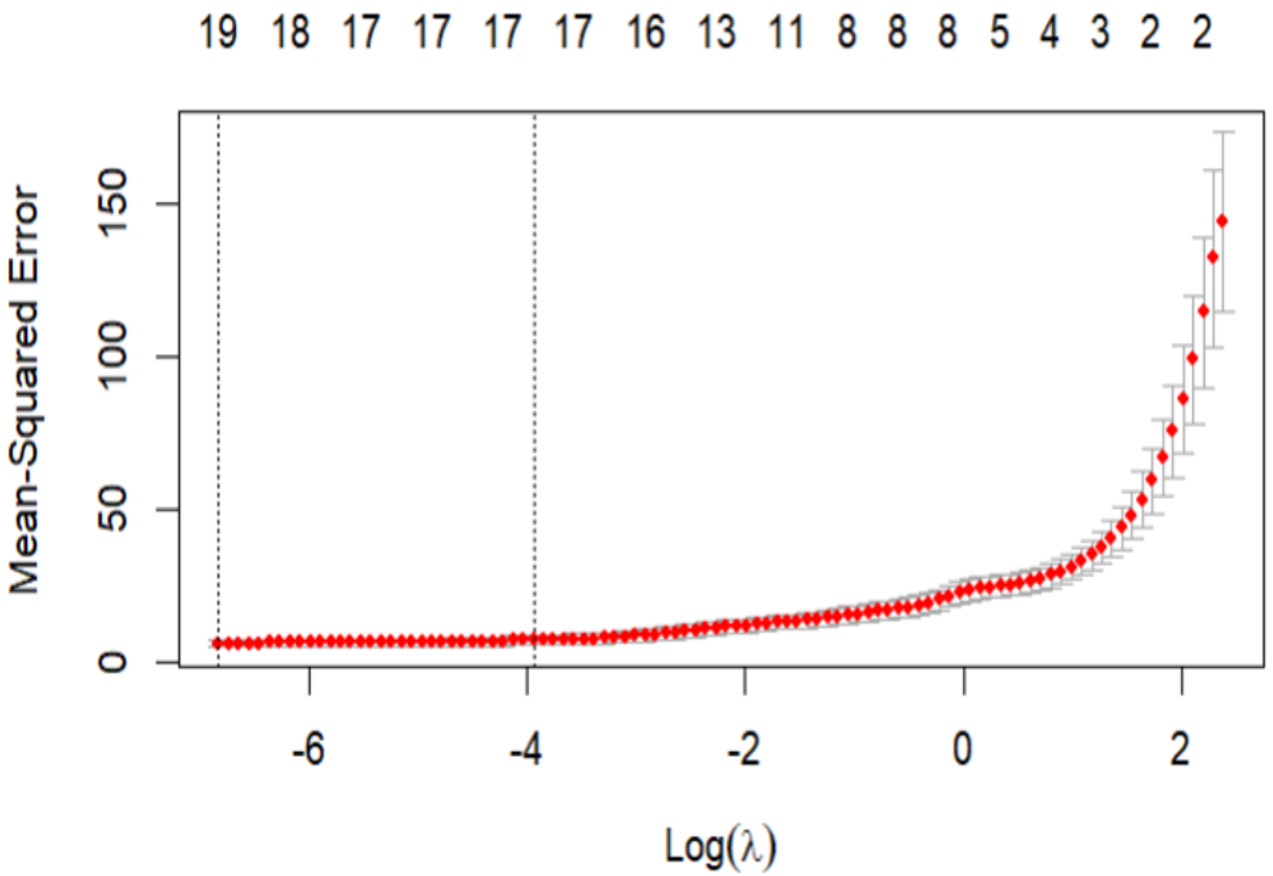
# Lasso Regression(L1 Norm) (Kassambara, 2018)

The regression coefficients are brought closer to zero by punishing the regression model with a penalty term called L1 norm, which is the sum of the absolute coefficients.

With regard to lasso regression, the penalty has the result of setting some coefficient estimates that barely affect the model exactly equal to zero.

To reduce the complexity of the model, variable selection using lasso might be viewed as an alternative to subset selection methods.

# Lasso Regression(L1 Norm)




- Cross validation's output shows that there are 19 nonzero coefficients for the Lamda.min model and 17 for the Lamda.1se model.
- As a result, the model will not include 2 out of 19 variables and 8 out of 17 variables whose coefficients are zero.
- The value of log lamda min is -6.826579, and log lamda.1se is equal to -3.942533.
- Furthermore, when comparing LASSO regression to Ridge regression, we can observe a decline in the actual lambda values.

# Lasso Regression(L1 Norm)

```
> predictions_train <- predict(lasso_model, s = lambda_best, newx = train_x)
> eval_results(train_y, predictions_train, train_x)
      RMSE   Rsquare
1 1.975359 0.9728894
>
> predictions_test <- predict(lasso_model, s = lambda_best, newx = test_x)
> eval_results(test_y, predictions_test, test_x)
      RMSE   Rsquare
1 2.681538 0.9439497
```

- The RMSE and R square values for the train and test dataset is shown above.
- It can be seen that the RMSE for the test data is higher than the train and the R squared value is lower for the test than the train data.

# Summary

- The random forest and the lasso & ridge regression models has been used predicting the sales.
- The random forest gives the RMSE value of 8.05
- The lasso and ridge gives RMSE value of  2.657 and 2.68
- The ridge and lasso helps in the better prediction by using the penalization method for large number of variables

# Future Research

- Future research will be done to reduce the RMSE values below 1.8 by using hyperparameter tunning techniques and other tunning techniques.
- Other machine learning algorithms will also be tried to see which ML algorithm gives the best result and suits best for our sales prediction.

# References


- Kelwig,D.(2022, June 24).The definitive guide to sales forecasting methodologies. Zendesk. Retrieved from https://www.zendesk.com/blog/5-essential-sales-forecasting-techniques/
- GeeksforGeeks. (2021). Design a Learning System in Machine Learning. GeeksforGeeks. Retrieved from https://www.geeksforgeeks.org/design-a-learning-system-in-machine-learning/
- Logallo, N. (2019, December). Data Science Methodology 101. Towards Data Science. https://towardsdatascience.com/data-science-methodology-101-ce9f0d660336

# References


- Pires, S. (2017, April 10). *A very basic introduction to Random Forests using R | Oxford Protein Informatics Group*. Blopig.com. https://www.blopig.com/blog/2017/04/a-very-basic-introduction-to-random-forests-using-r/

- Singh. (2019, November 12). *Linear, Lasso, and Ridge Regression with R*. PluralSight. https://www.pluralsight.com/guides/linear-lasso-and-ridge-regression-with-r

- Kassambara. (2018, November 11). *Penalized Regression Essentials: Ridge, Lasso & Elastic Net*. STHDA. http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/153-penalized-regression-essentials-ridge-lasso-elastic-net/

# Thank You