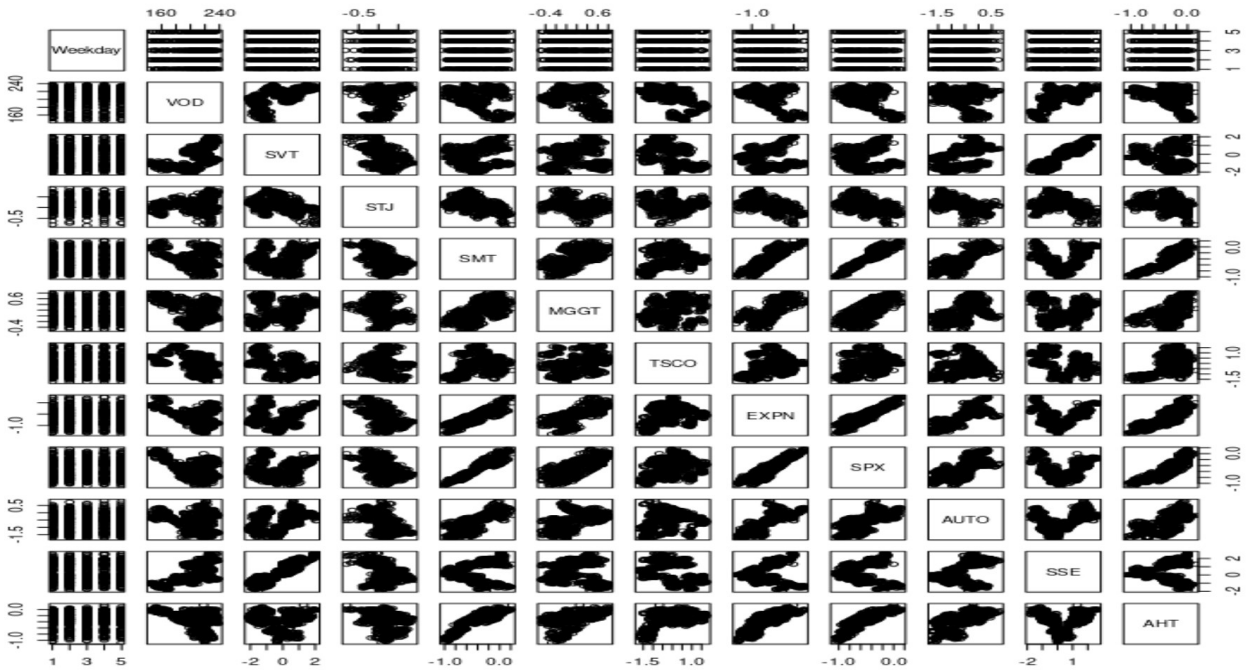
Regression Modelling Project

Explore the data

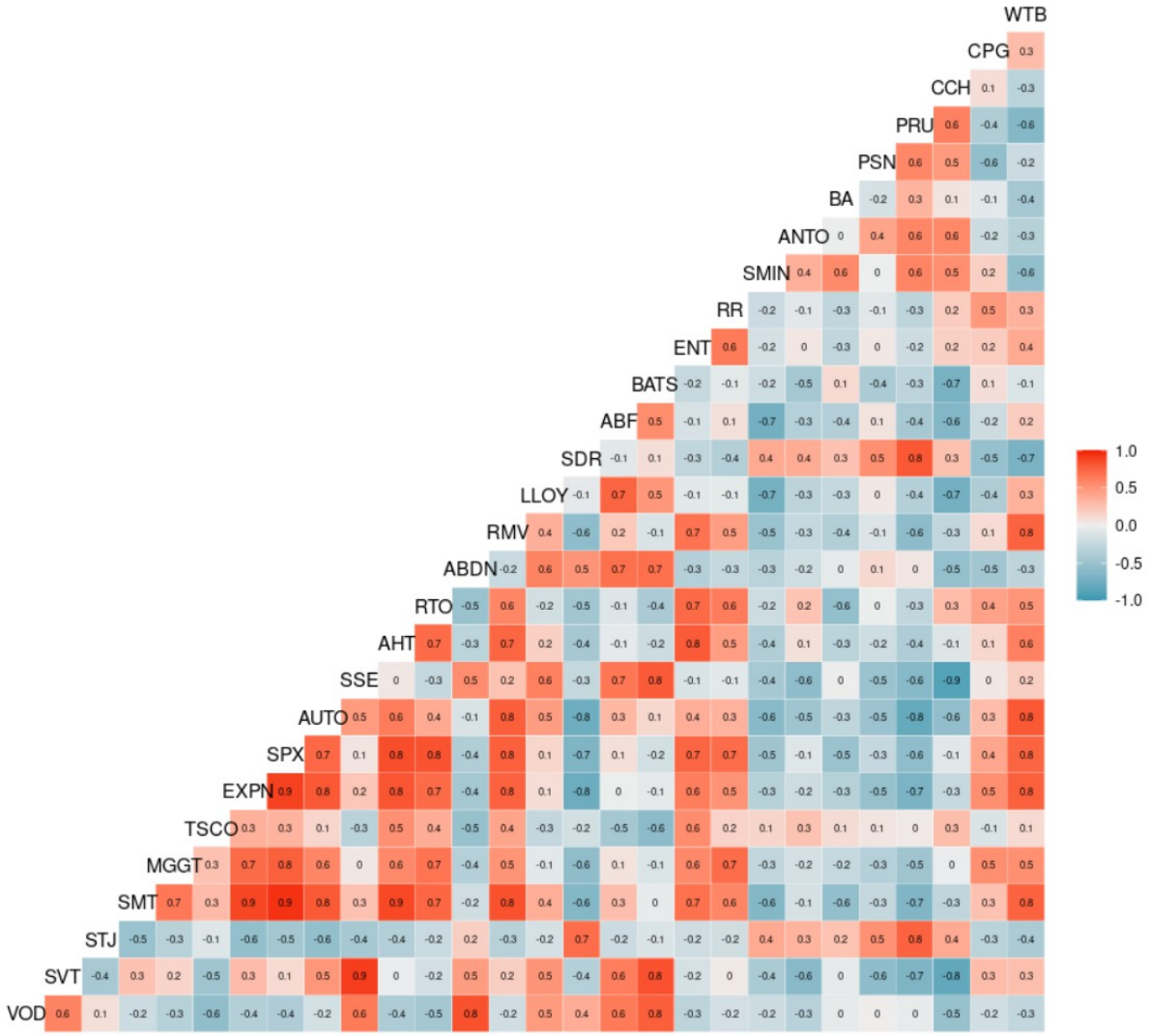
The Financial Times Stock Exchange (FTSE 100) is a stock market index that measures the performance of the top 100 businesses listed on the London Stock Exchange.The lse dataset in this study examines Vodafone's and 27 other FTSE 100 Index businesses' closing share prices. The information was gathered between January 2016 and January 2019 and is organised into a 757-row, 32-column dataset. The first four columns are "Date", "Year", "Month", and "Weekday". The closing prices of the other firms are listed in the next 28 columns. The purpose of this research is to develop a linear regression model that uses stock price data from other firms to explain the variability in Vodafone's daily stock price. Apart from developing a model to explain stock price volatility, the goal of this study is to figure out what factors and elements influence the price of VOD.

I constructed a correlation matrix between all of the firms using the corr() function. In statistics, the correlation coefficient is used to determine how strong a link between two variables is. The correlation coefficient is always in the range of 1 to -1.



I then plotted a more visually aesthetic graph to show the correlation between each company and VOD in the Dataset. Companies having the strongest correlation with VOD are ABDN, BATS, ABF, SVT and SSE.

Looking at the correlogram, it's evident that several firms are highly associated with one another, implying multicollinearity. When independent variables in a multivariate regression model are correlated, this is known as multicollinearity. This is a problem in a regression model since the independent variables must be independent. When a set of independent variables is highly correlated, the outcomes of hypothesis tests, coefficient estimates, and, as a result, the model's interpretation might be influenced.

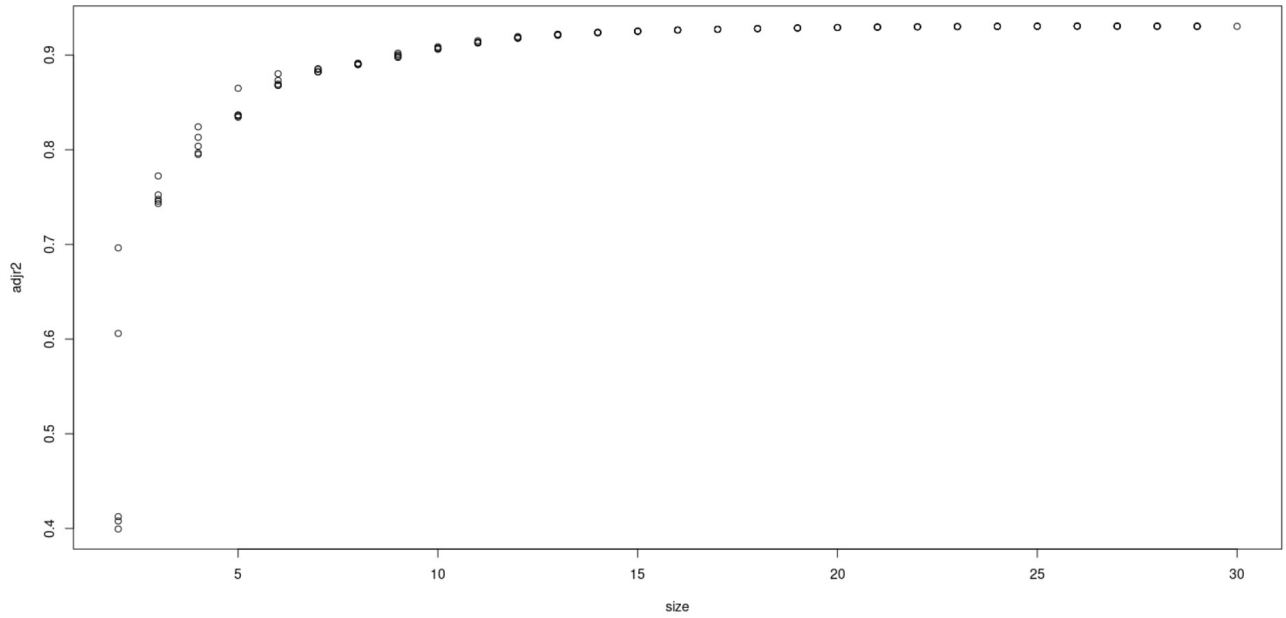


Model Building

We start by determining which variables have an association with VOD, so we aim to exclude variables that are not related with VOD or have a misleading correlation with the response, i.e., there appears to be a relationship but there is no logical justification for it. The objective is to strike a compromise between include enough variables in the model to describe the data trend while avoiding fitting to noise. In this scenario, we're attempting to include everything that's relevant while removing variables that aren't. The factors with a significant association on model fit are included, whereas variables with a little impact are excluded. In this case, it's was a good idea to exclude out variables having spurious associations, as the relationship might appear by accident in the data set.

I utilised leaps and bounds to cope with this balancing act, which is effectively akin to fitting every potential model and selecting the best one. The primary drawback of this strategy is that it needs some human input and might be subjective.

I started by creating a series of models of varied sizes and calculating one of the selection criteria for each of them. After plotting the values of the selection criteria, a model is chosen depending on where the plot begins to level out. The following is the result of a plot using adjusted R-squared as the criterion:



I felt the best model of size 15 would be appropriate since things have started to level off. Alternatively, I can choose the model of size 14 that appears to be the most similar. I believe neither is correct nor incorrect, so I took these models forward and investigated them further using additional criteria.

Forward selection:

Forwards selection regression begins with an intercept-only model and improves the regression by gradually adding variables. When there is no more progress to be made by adding more variables, the process comes to an end.

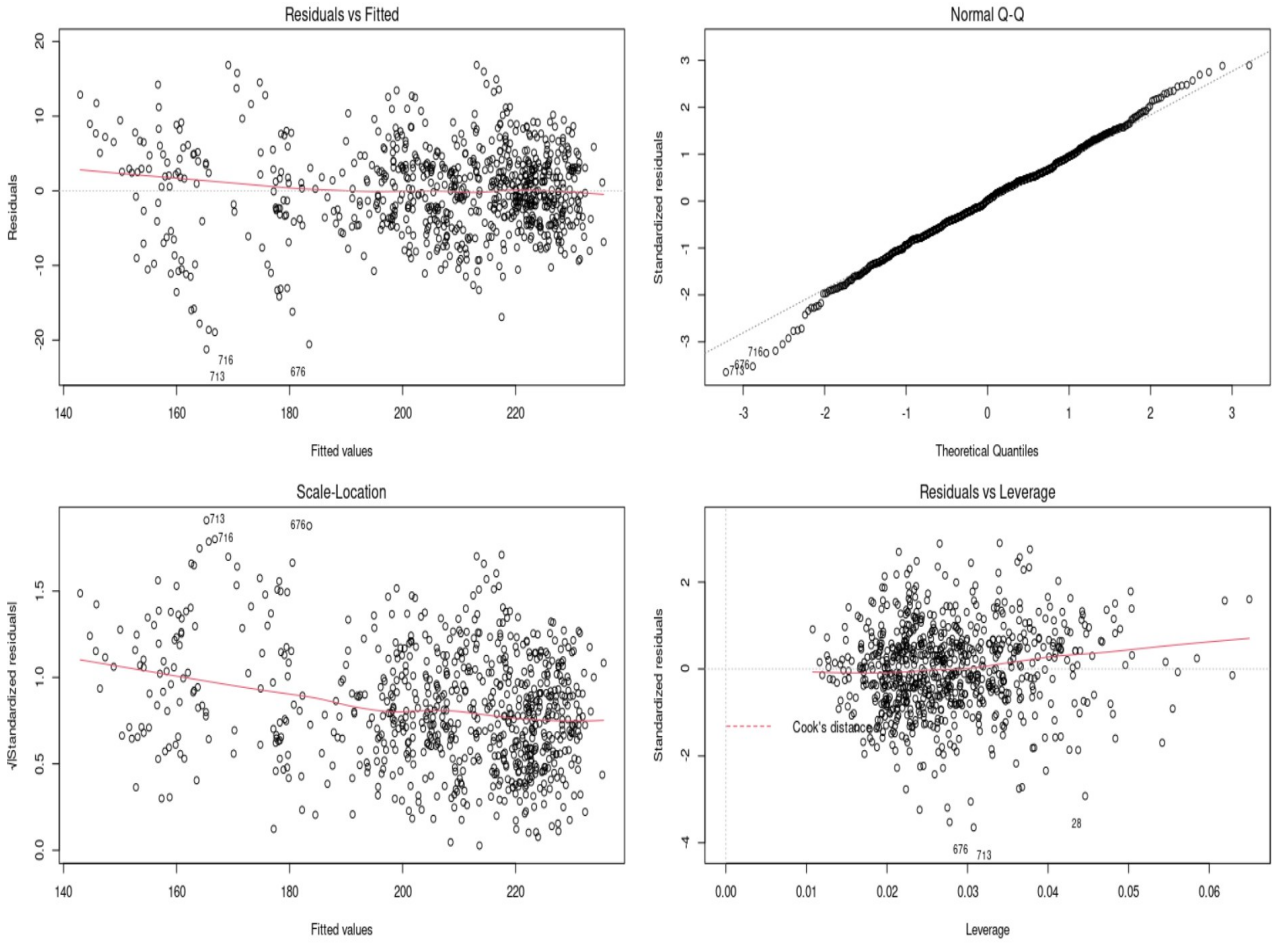
Backward Selection:

The goal of this selection regression is to discover the model with the most predictors by using backwards regression. We begin by including all variables in the model and subsequently exclude the variables that are not significant or do not enhance model fit.

Step-wise selection:

Iterate between forward and backward steps until no more progress is possible. Only important predictors remain in the simplified model which were :

VOD ~ ABDN + BATS + AHT + SSE + AUTO + BA + RMV + CPG + ANTO + PRU + SMIN + SPX + SMT + STJ + SVT + RR + SDR + LLOY + EXPN + TSCO

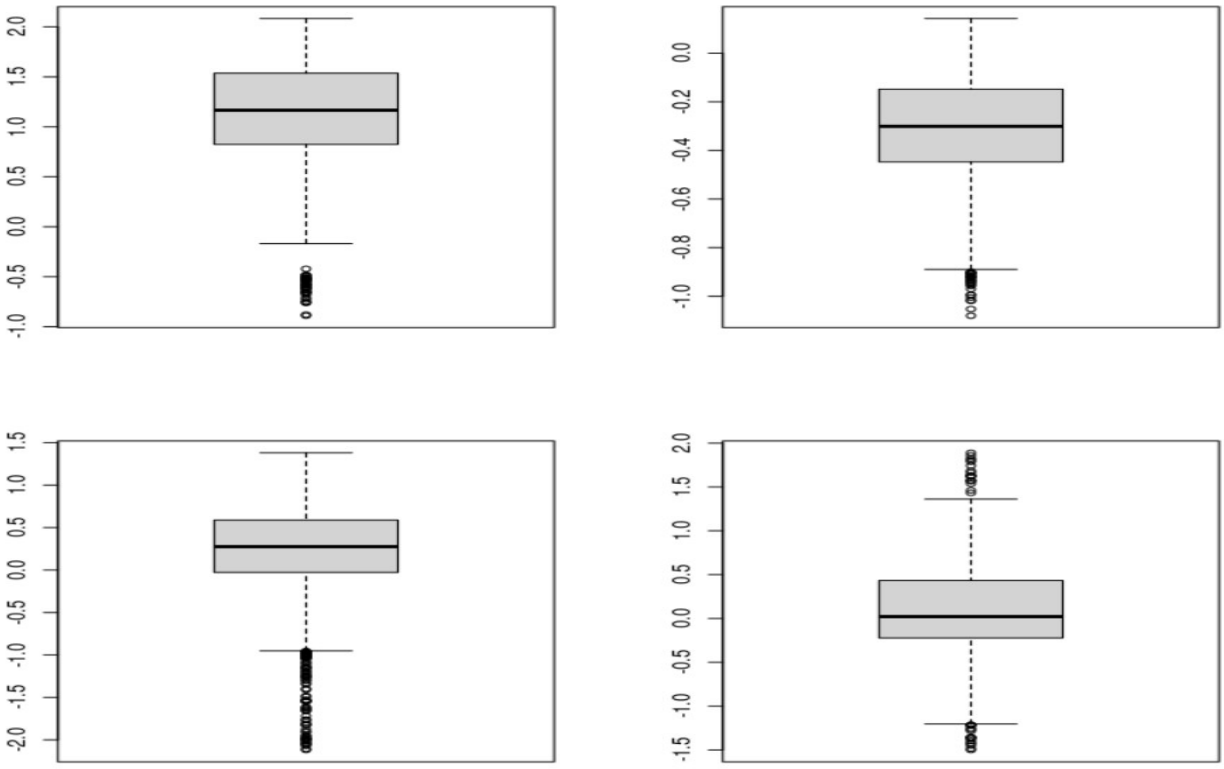


The R squared value of the step wise model is 0.9304813, which is higher than both the backward and forward selection models. Backwards and forwards stepwise selection regression create almost identical models. Above are the graphs that resulted from the stepwise selection.

The thin tails phenomenon is visible in the Normal Q-Q plot, where less data is distributed in the teails and more data is concentrated in the centre of the distribution . The thin tails correspond to the initial quantiles, which occur at bigger than anticipated values, and the last quantiles, which occur at lower than expected values. There is no evident over or under dispersion in the random scatter plot, hence the variance is constant. The figure below shows that there are no data points that are considerably skewing the model findings with Cooks distance, showing that there are no data points that are significantly skewing the model result

If the independence condition is violated, the model is likely to underestimate the variability of our data. As a result, the confidence intervals would be excessively small, and the hypothesis tests would be invalid. Because dependence can emerge in data obtained in a sequential manner, this assumption must be explored. There was no weaving or curving seen in the scatter plots created with plot(stepwise_model). As a result, there is no compelling evidence against independence.

I next looked at all of the predictor variables' boxplots to see which ones had outliers. A substantial number of outliers were discovered in ABDN, SDR,BA, SMIN and many more. As a result, I tried eliminating each variable one at a time, and simply removing BA reduced the RMSPE value.



Discussion of Model

After removing the variables with considerate amount of outliers the final model I think is the best fit is:

```
>   VOD.lm <- lm(VOD ~ ABDN + BATS + AHT + SSE + AUTO  + RMV + CPG + ANTO + PRU + SMIN + SPX + SMT + STJ
 + SVT + RR + SDR + LLOY + EXPN + TSCO, data = lse)
>
```

The RMSPE value was lowered after the natural log of VOD was removed. This was a considerable reduction, indicating that y should not be treated as a natural log. The RMSPE value was lowered after an interaction term was included.