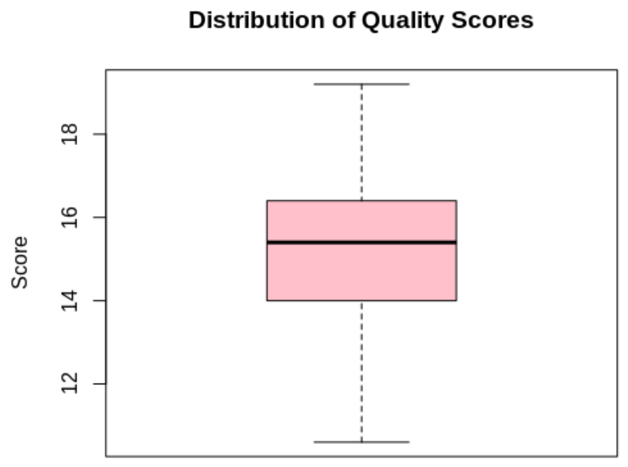
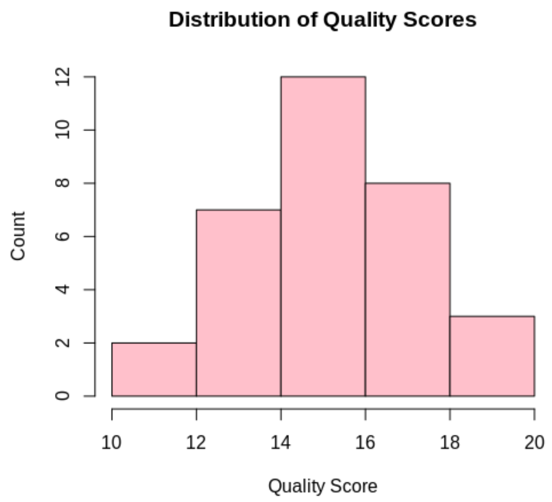
Regression Analysis Assignment

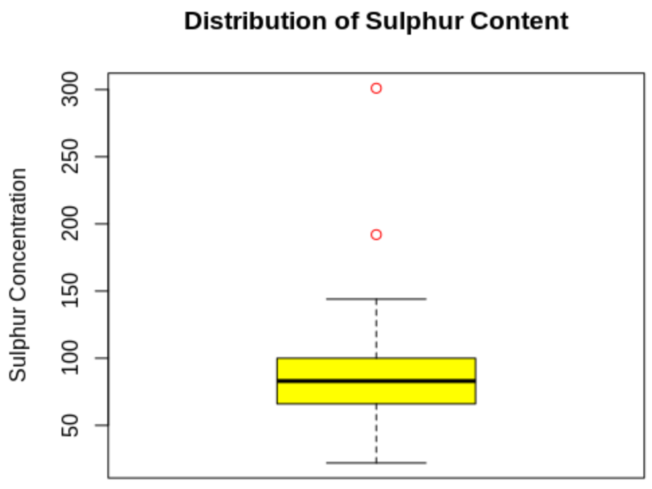
Liam De Barra

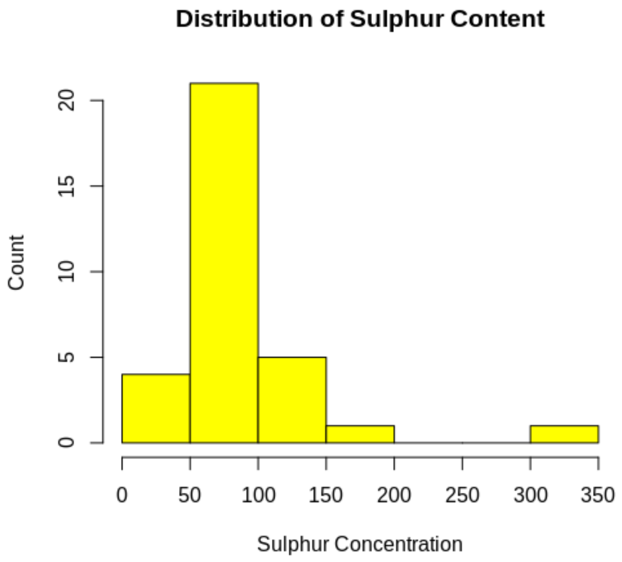
Question 1

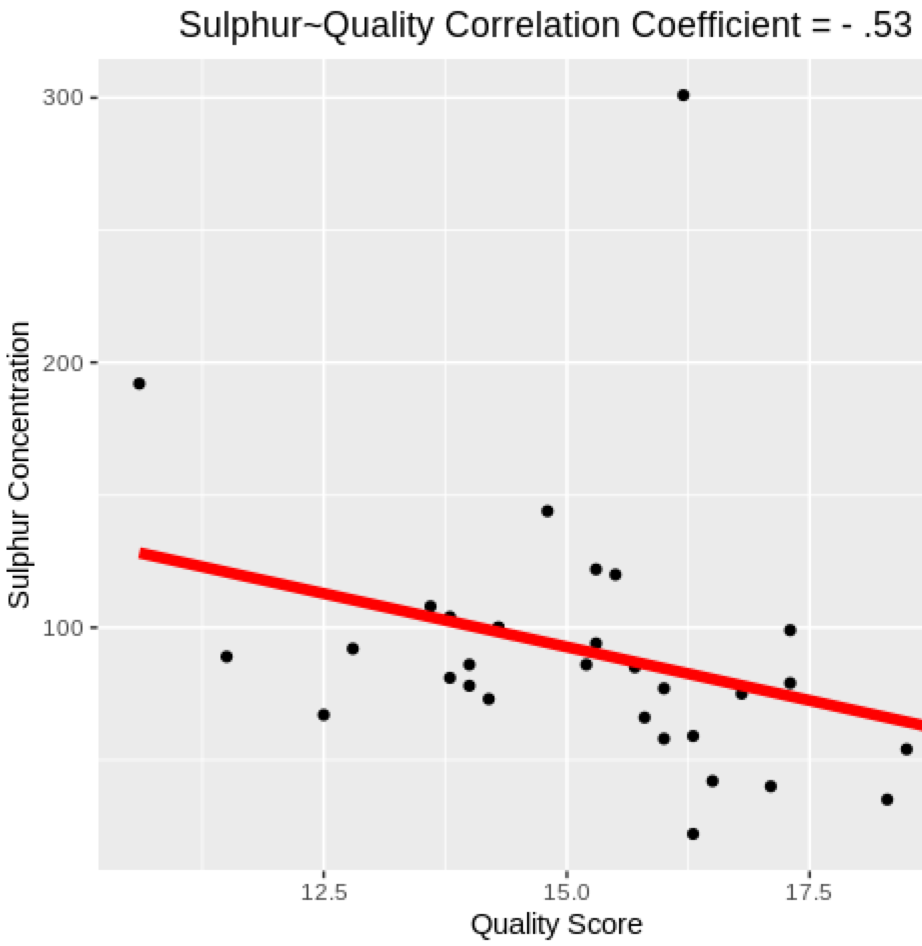
Data was loaded into R using base read.csv function and checked for missing values using the VIM ‘aggr’ function, none were found. The dataset consists of 32 observations of the quality & sulphur content of wine. Although the variable quality is not truly continuous; both variables were treated as such for the purpose of analysis. The distribution of both variables was first visualised using boxplots and histograms to assess normality. While both variables had approximately normal distributions there was slight left skewness in Sulphur content and the boxplot identified two outliers which were noted. A scatterplot of quality versus sulphur content indicated a possible weak correlation between the two variables and the Spearman correlation coefficient was calculated at ρ = - .5289 (p-value = 0.001853). See figure in section e



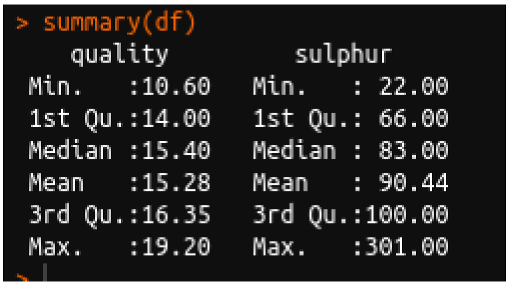






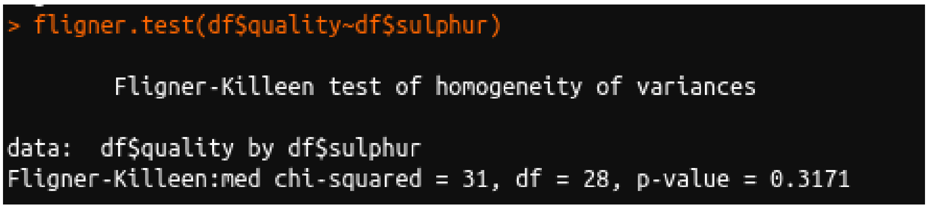


Sumary statistics were also output using R’s summary function and showed the large range in Sulphur content as well as the discrepancy in mean & median both reflective of outliers in the extreme high range.



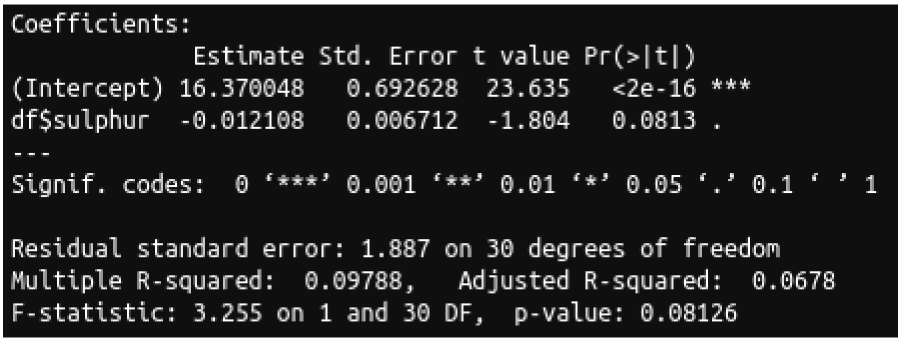
```
> summary(df)
    quality          sulphur     
 Min.   :10.60   Min.   : 22.00  
 1st Qu.:14.00   1st Qu.: 66.00  
 Median :15.40   Median : 83.00  
 Mean   :15.28   Mean   : 90.44  
 3rd Qu.:16.35   3rd Qu.:100.00  
 Max.   :19.20   Max.   :301.00  
>
```

B.
Prior to fitting a regression model the homogenity of variances was investigated using the Fligner-Kileen Test. A p-value of .3171 means we fail to reject the null hypothesis and conclude the variances are not significantly different and we may apply ANOVA.



```
> fligner.test(df$quality~df$sulphur)

        Fligner-Killeen test of homogeneity of variances

data:  df$quality by df$sulphur
Fligner-Killeen:med chi-squared = 31, df = 28, p-value = 0.3171
```

The model was fit using base R’s ‘aov’ function and the co-efficients taken from the output of summary.lm() of the resulting model.

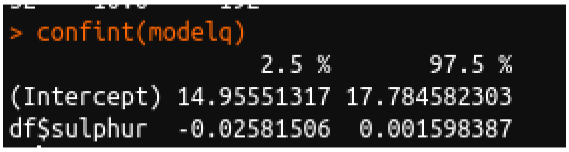


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.370048   0.692628  23.635   <2e-16 ***
df$sulphur  -0.012108   0.006712  -1.804   0.0813 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.887 on 30 degrees of freedom
Multiple R-squared:  0.09788,	Adjusted R-squared:  0.0678 
F-statistic: 3.255 on 1 and 30 DF,  p-value: 0.08126
```

Quality(y) = 16.37 -0.0121*(Sulphur) + e
β0 = 16.37 or y intercept of regression line
β1 = -0.0121 or the regression coefficient of Sulphur
e = residual error

The β1 co-efficient implies that the sulphur content has a mild negative effect on quality. Specifically; -0.0121 reflects the effect of a unit increase in Sulphur concentration on wine quality.

C.
The 95% confidence interval for 1 was calculated using base R’s ‘confint() function and had a lower bound of -0.02581 and upper bound of 0.001598. *Note the change of sign as the confidence interval spans 0.



```
> confint(modelq)
                  2.5 %      97.5 %
(Intercept) 14.95551317 17.784582303
df$sulphur  -0.02581506  0.001598387
```

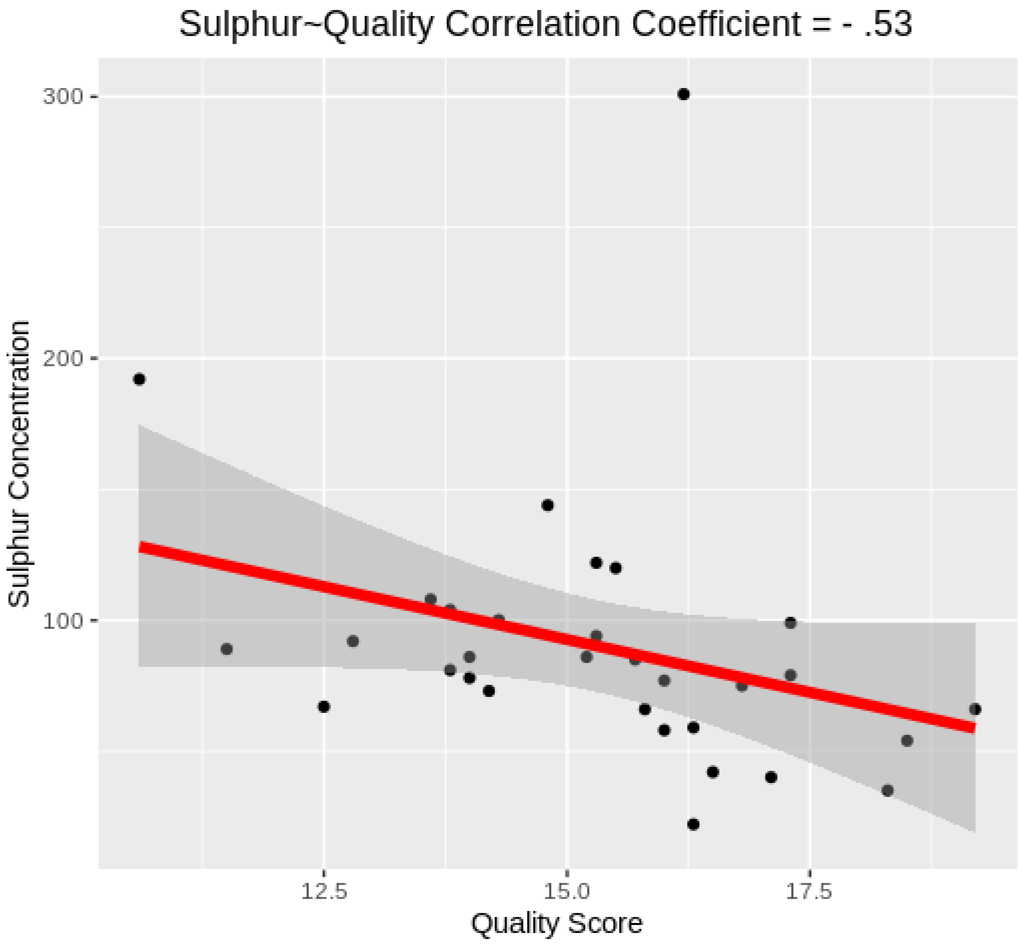
D.

The large p-value of .0813 output from the model summary indicates that the effect of sulphur on quality **detected in this analysis** is not significant at the commonly accepted 95% confidence level. We fail to reject the null hypothesis that sulphur content has no effect on wine quality. We reject the alternative hypothesis that sulphur content affects wine quality.

These results severely limit the applicability and robustness of our model and it's use to predict wine quality is not advised. A major limitation of this analysis however is the small sample size and it should be reported to end-users that failure to reject the null hypothesis in this sample does not necessarily indicate the absence of a correlation between sulphur content and wine quality in the real-world.

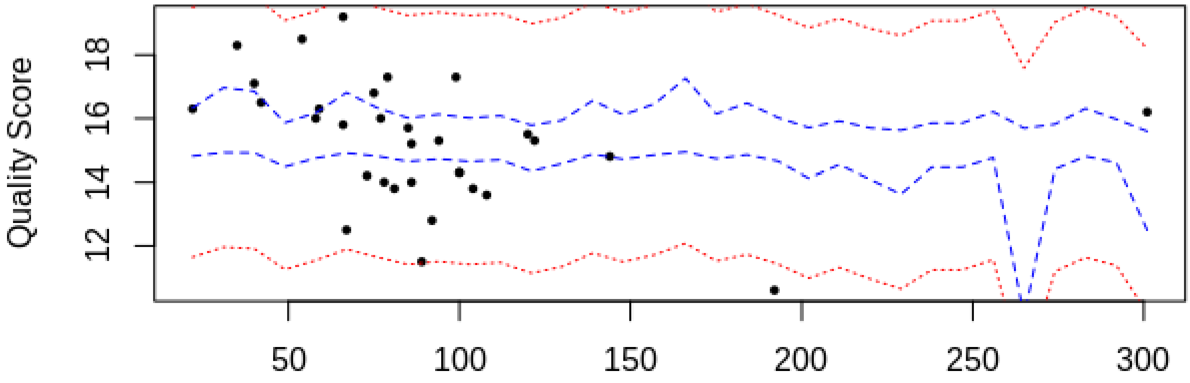
E.

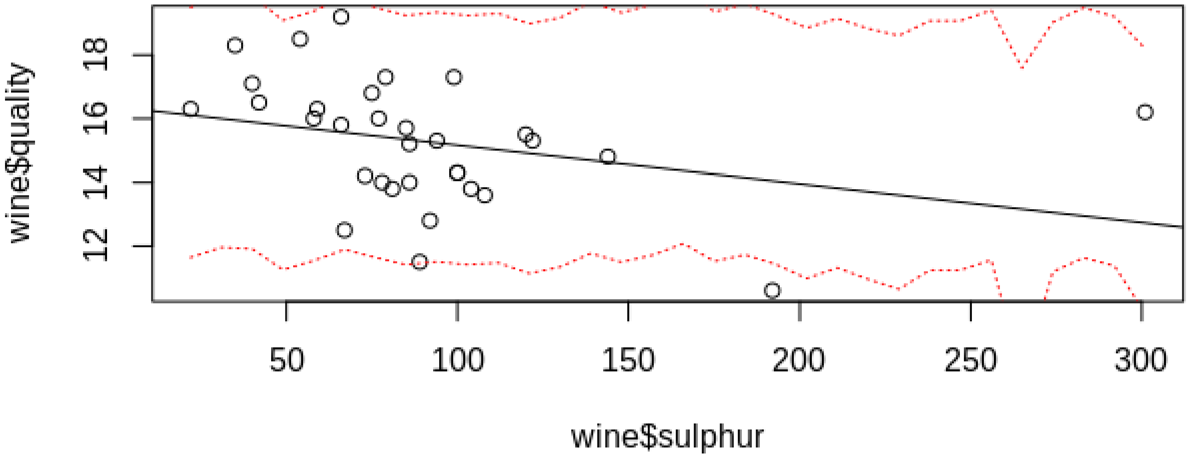
The regression line and 95% confidence intervals were calculated & plotted using ggplot2

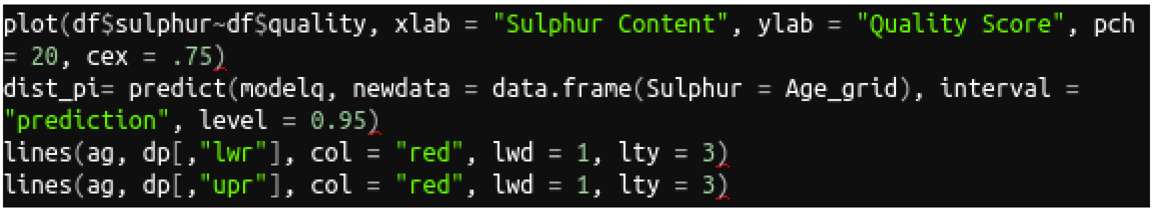


95% confidence intervals are indicated by grey-fill.

Plotting of prediction Intervals was also attempted using base R but performed erratically. The large intervals are reflective of the low confidence in the predictions arising from the small sample size.



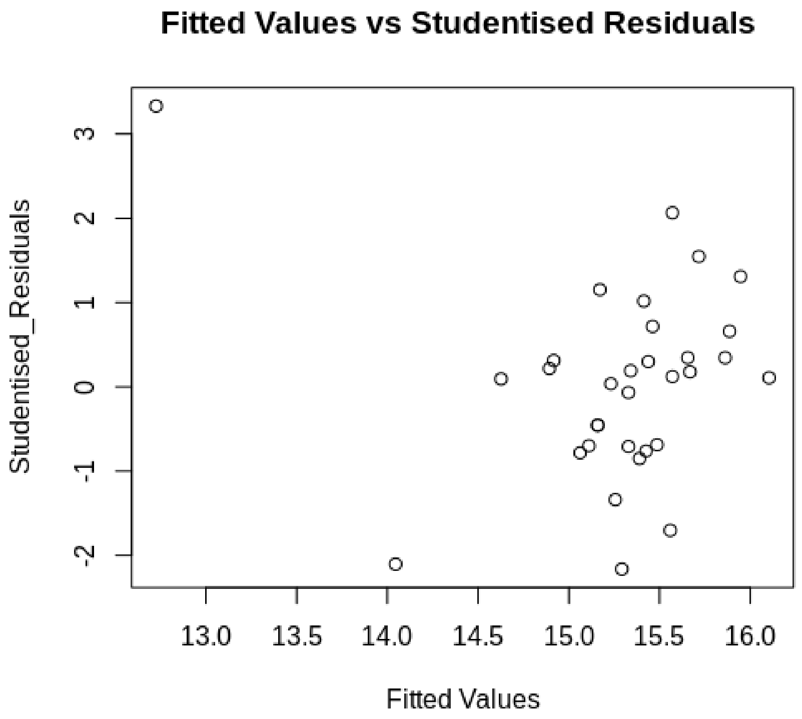


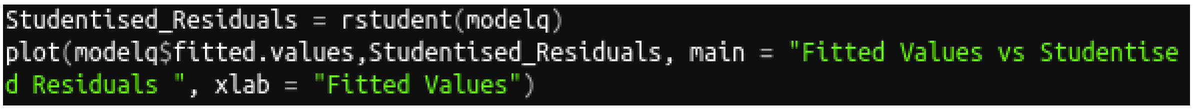


```
plot(df$sulphur~df$quality, xlab = "Sulphur Content", ylab = "Quality Score", pch
= 20, cex = .75)
dist_pi= predict(modelq, newdata = data.frame(Sulphur = Age_grid), interval =
"prediction", level = 0.95)
lines(ag, dp[,"lwr"], col = "red", lwd = 1, lty = 3)
lines(ag, dp[,"upr"], col = "red", lwd = 1, lty = 3)
```

F

When trying to identify outliers, one problem that can arise is when there is a potential outlier that influences the regression model to such an extent that the estimated regression function is "pulled" towards the potential outlier, so that it isn't flagged as an outlier using the standardized residual criterion. To address this issue, **studentized residuals** offer an alternative criterion for identifying outliers. The basic idea is to delete the observations one at a time, each time refitting the regression model on the remaining points. A data point having a large deleted residual suggests that the data point is influential. If an observation has a studentized residual that is larger than 3 (in absolute value) we can call it an **outlier (https://newonlinecourses.science.psu.edu/stat462/node/247/)** Plotting the fitted values against studentised residuals identified one such outlier; indicating accuracy of the model could be greatly improved by removing this value.

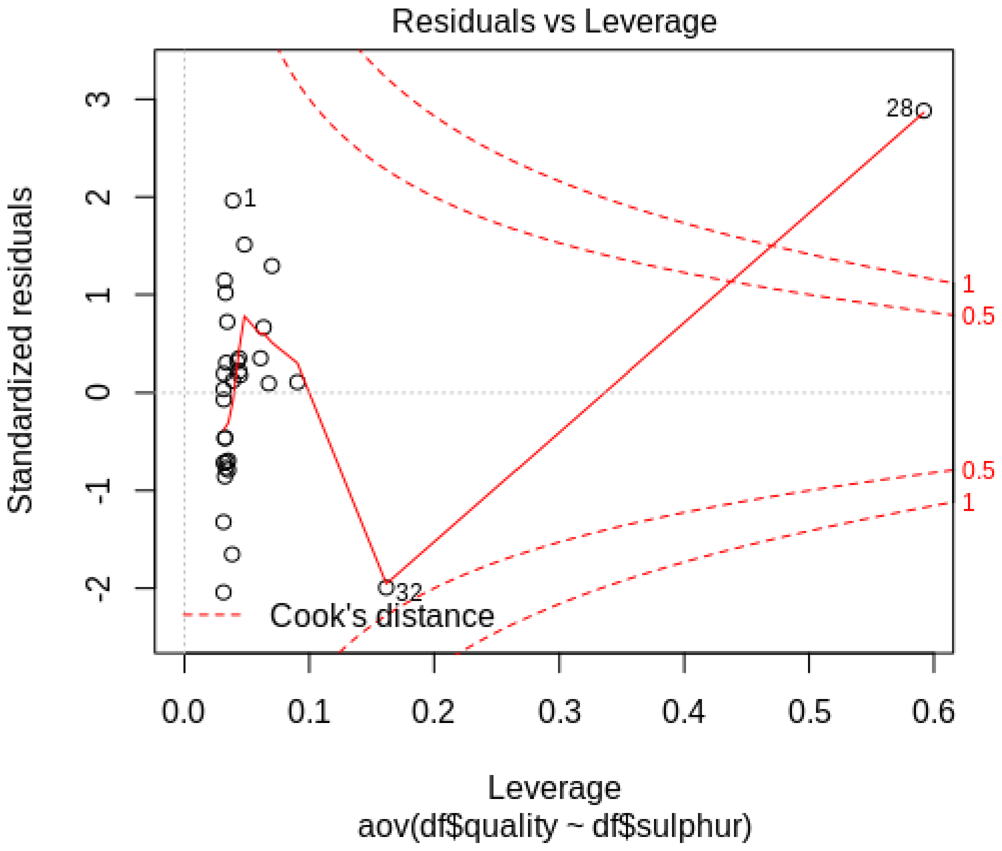




```
Studentised_Residuals = rstudent(modelq)
plot(modelq$fitted.values,Studentised_Residuals, main = "Fitted Values vs Studentise
d Residuals ", xlab = "Fitted Values")
```

G

Leverage points are those observations made at extreme or outlying values of the independent variables such that lack of neighbouring observations means that the fitted regression model will pass close to that particular observation. Investgation of the leverage of each datapoint revealed two potential leverage points (datapoint 28 & 32).



H

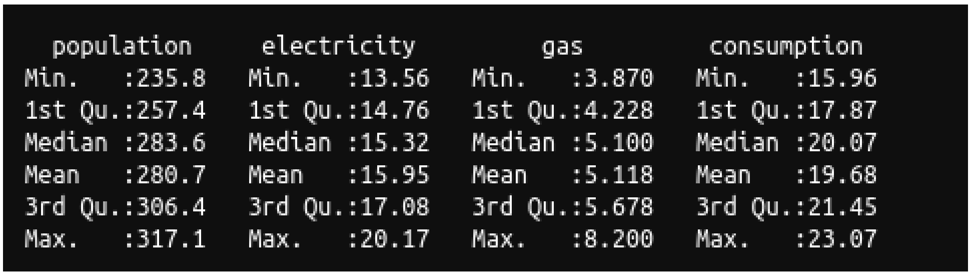
Only observation 28 had a Cooks distance of greater than 1 and inspection of the data showed this observation had a sulphur concentration of more than double the next nearest value. By examining the range and interquartile values of the ‘quality’ variable (Range: 10.6-19.2, Median = 15.4, IQ3 = 16.35) and comparing this with the quality score for observation 28 (16.2,) it was clear that this observation could have markedly increased the B1 co-efficient i.e. the negative correlation between high sulphur content and quality was offset by this observation. Indeed; re-doing the model with this observation excluded caused B1 to change from -0.0121 to -.0348; albeit with B0 also increasing from 16.37 to 18.15

This plot helps us to find influential cases (i.e., subjects) if any. Not all outliers are influential in linear regression analysis; athough data may have extreme values they might not be influential in determining a regression line. That means, the results wouldn’t be much different if we either include or exclude them from analysis. They follow the trend in the majority of cases and they don’t drastically alter the fitting of a model; they are not influential. On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. Another way to put it is that they don’t get along with the trend in the majority of the cases.

In interpreting the leverage plot we watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. Look for cases outside of a dashed line, Cook’s distance as these cases are influential to the regression results. The resulting model will be altered if we exclude those cases as was the case for this dataset.

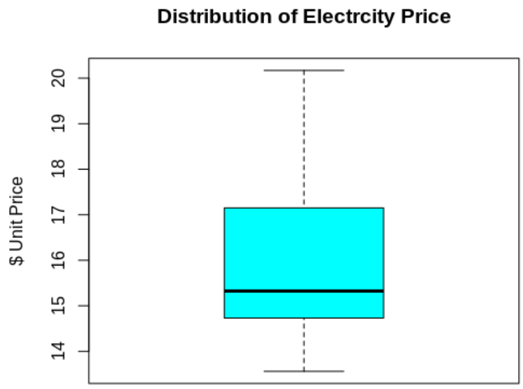
Section 2

This dataset consists of 34 yearly population, electricity & gas price observations recorded between 1984 and 2017; and the response variable consumption. Summary information for each variable is output below where quite similar means & medians for each variable indicated a normal distribution which was supported by graphical analysis. All variables were assessed for missing values using the VIM package and none were found.

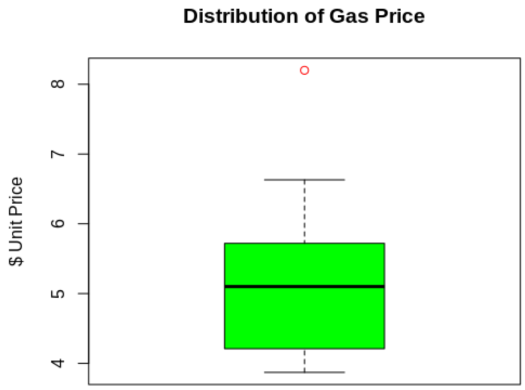


```
  population      electricity         gas          consumption
 Min.   :235.8   Min.   :13.56   Min.   :3.870   Min.   :15.96
 1st Qu.:257.4   1st Qu.:14.76   1st Qu.:4.228   1st Qu.:17.87
 Median :283.6   Median :15.32   Median :5.100   Median :20.07
 Mean   :280.7   Mean   :15.95   Mean   :5.118   Mean   :19.68
 3rd Qu.:306.4   3rd Qu.:17.08   3rd Qu.:5.678   3rd Qu.:21.45
 Max.   :317.1   Max.   :20.17   Max.   :8.200   Max.   :23.07
```

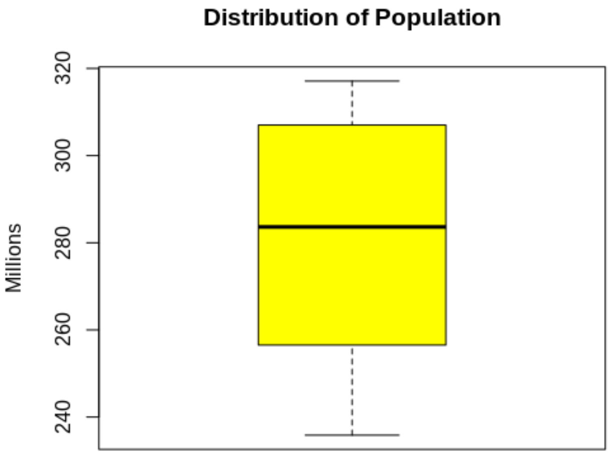
Distribution of each variable was visualised by boxplot to identify outliers but given the temporal nature of this dataset scatterplots were deemed more informative than histograms etc.



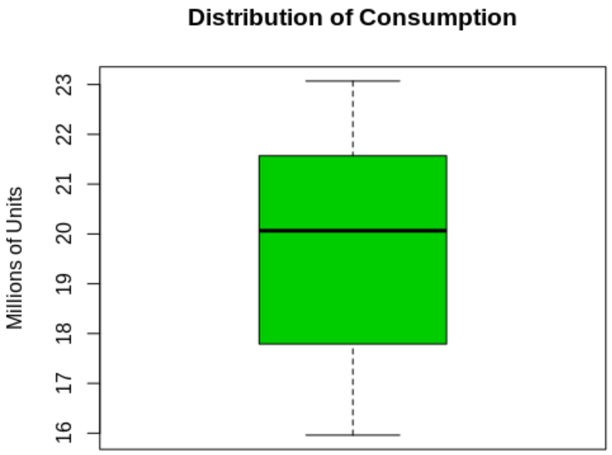
Distribution of Electrcity Price



Distribution of Gas Price

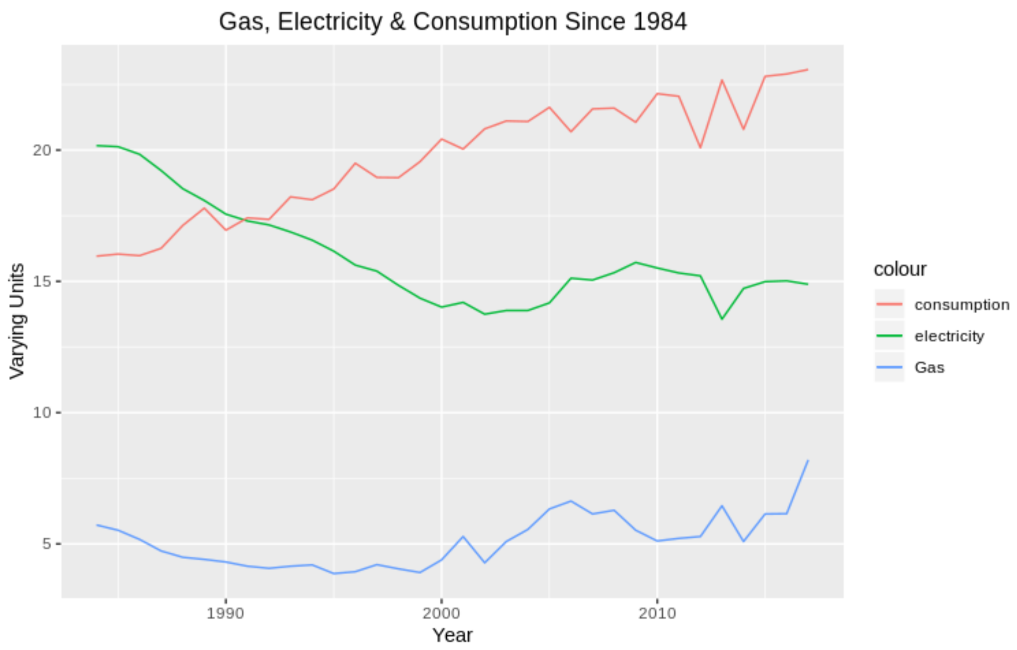


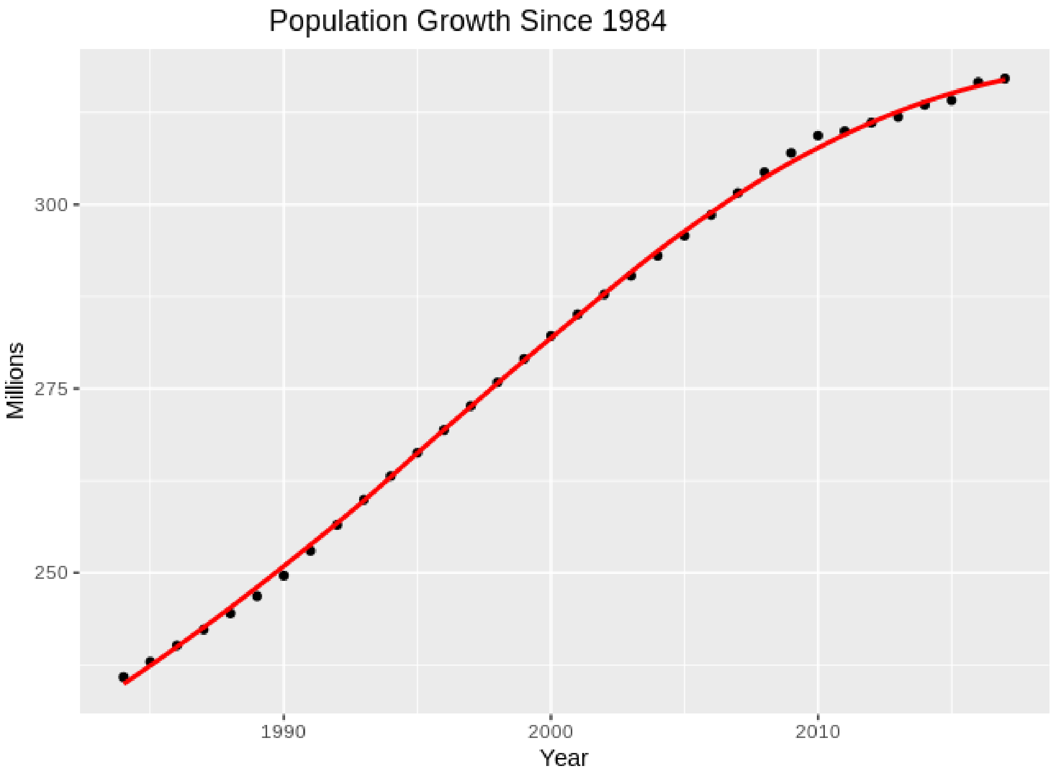
Distribution of Population



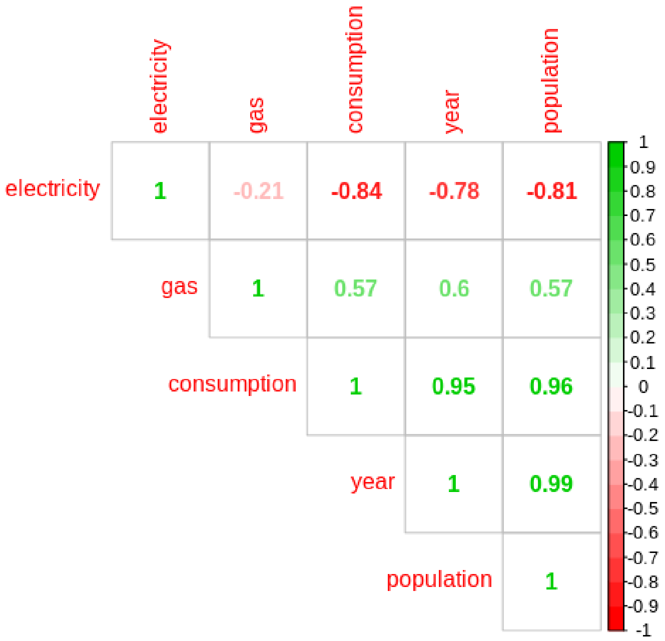
Distribution of Consumption

Given their similar scales, the variables gas, electricity and consumption were plotted together using ggplot and geom_line() to better visualise the underlying pattern. Population was plotted seperately.



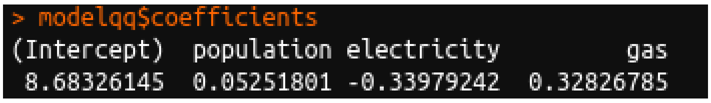


The correlation co-efficients between variables were calculated using base R’s ‘cor()’ function and visualised in the form of a correlation matrix. The positive correlation between advancing year and the variables population & consumption was extremely strong (correlation co-efficients of .99 and .95 respectively). Price of electricity was strongly negatively correlated with year (-0.78) while gas price was positively correlated (0.6)



	electricity	gas	consumption	year	population
electricity	1	-0.21	-0.84	-0.78	-0.81
gas		1	0.57	0.6	0.57
consumption			1	0.95	0.96
year				1	0.99
population					1

B.(i)



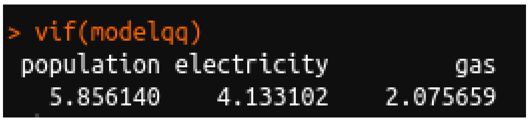
```
> modelqq$coefficients
(Intercept)  population electricity         gas
 8.68326145  0.05251801 -0.33979242  0.32826785
```

Consumption(y) = 8.68 +.0525*(Population) -0.3398*(Electricity) +0.3283*(Gas) + e
β0 = 8.68 or y intercept of regression line
β1 = 0.0525 or the regression coefficient of Population
β2 = -0.3398 or the regression coefficient of Electricity
β3 = 0.3283 or the regression coefficient of Gas
e = residual error

The co-efficient for population of .0525 implies population size has a positive correlation with consumption. Specifically; 0.0525 reflects the effect of a unit increase in population on consumption. Although the co-efficient may seem small; population is measured in the range of 100’s of millions and actually has a substantial impact on consumption as you would expect.

(ii)

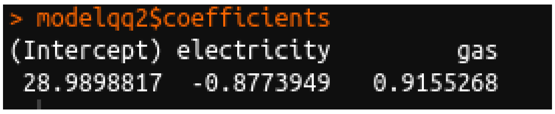
The striking correlation between pairs of variables in this dataset is an indicator of co-linearity and discerning the contributions of such variables to the response can be easily obscured. Collinearity can be detected by calculating the variance inflation factors for each co-efficient.



```
> vif(modelqq)
 population electricity         gas 
   5.856140    4.133102    2.075659 
```

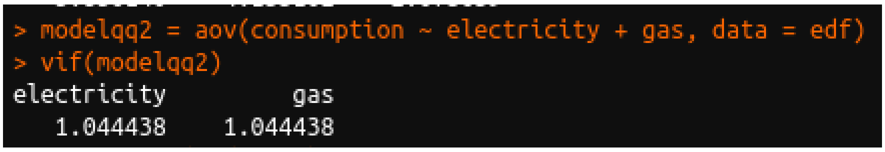
Although none of the variable inflation factors exceed the threshold value of 8, the large value for population in particular are cause for further investigation.

(iii)
By refitting the model with population removed we observe substantial changes in the co-efficients for electricity & gas but the signs of both co-efficients remain the same which is re-assuring.



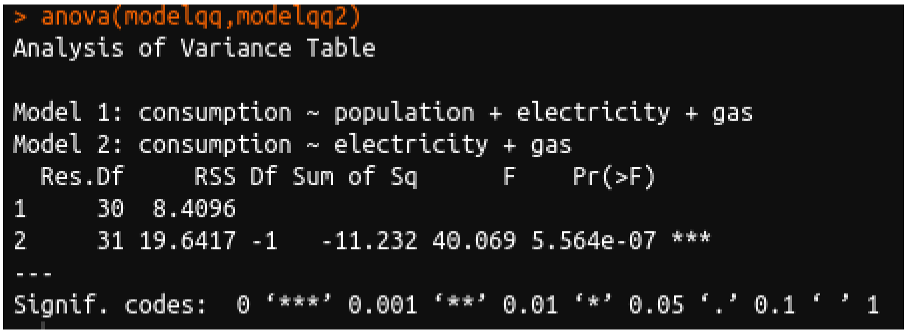
```
> modelqq2$coefficients
(Intercept) electricity         gas 
 28.9898817  -0.8773949   0.9155268 
```

Re-applying the vif() function to the new model results in greatly reduced variable importance factors



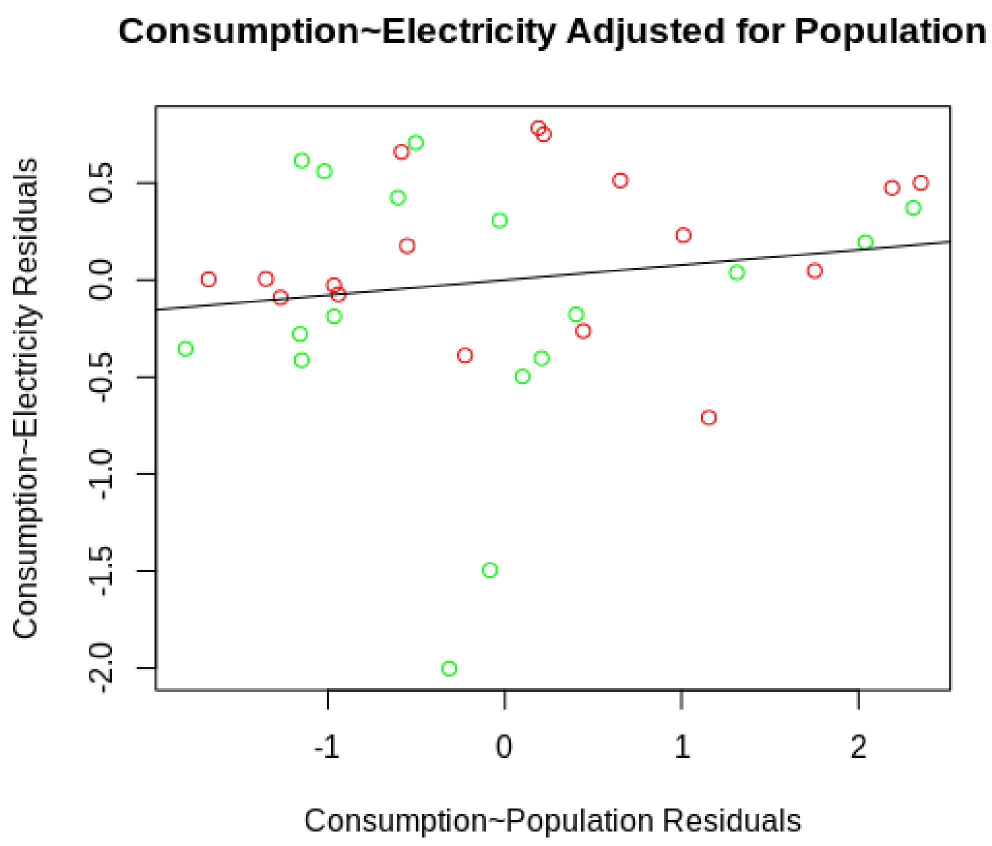
```
> modelqq2 = aov(consumption ~ electricity + gas, data = edf)
> vif(modelqq2)
electricity         gas 
   1.044438    1.044438 
```

The fit of the two models was compared using the ANOVA likelihood ratio test:



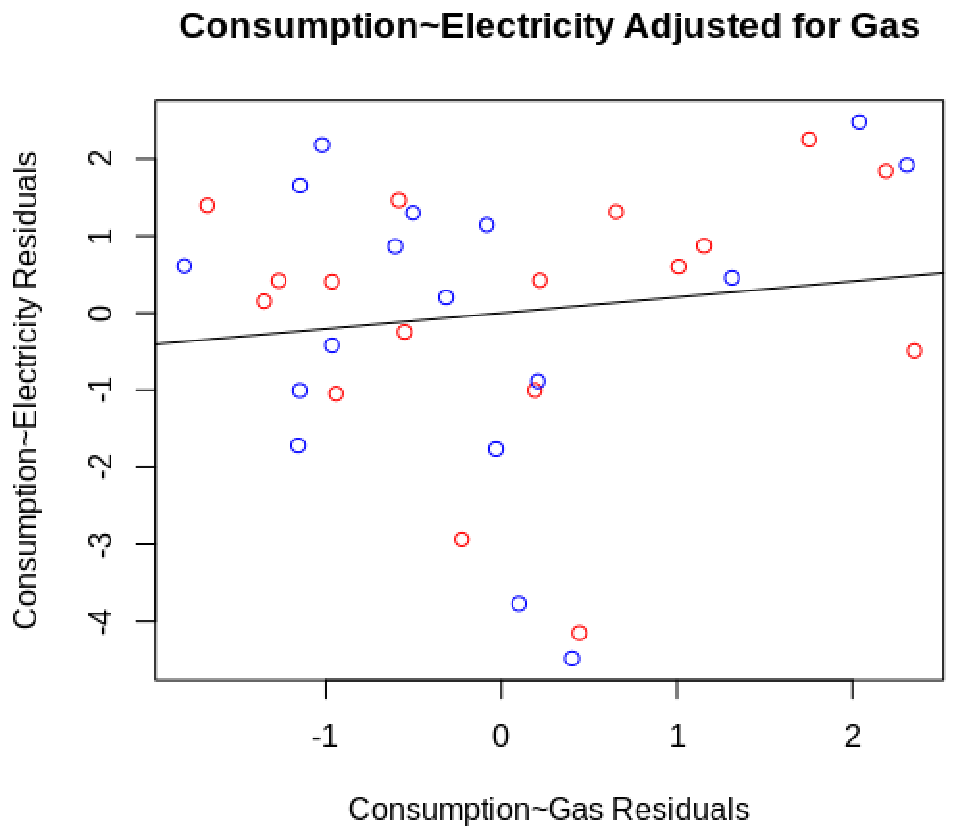
```
> anova(modelqq,modelqq2)
Analysis of Variance Table

Model 1: consumption ~ population + electricity + gas
Model 2: consumption ~ electricity + gas
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)    
1     30  8.4096                                  
2     31 19.6417 -1   -11.232 40.069 5.564e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This analysis indicated the fit of the models was indeed significantly different and the more elaborate Model 1 taking population into account (modelqq above) is preferable.

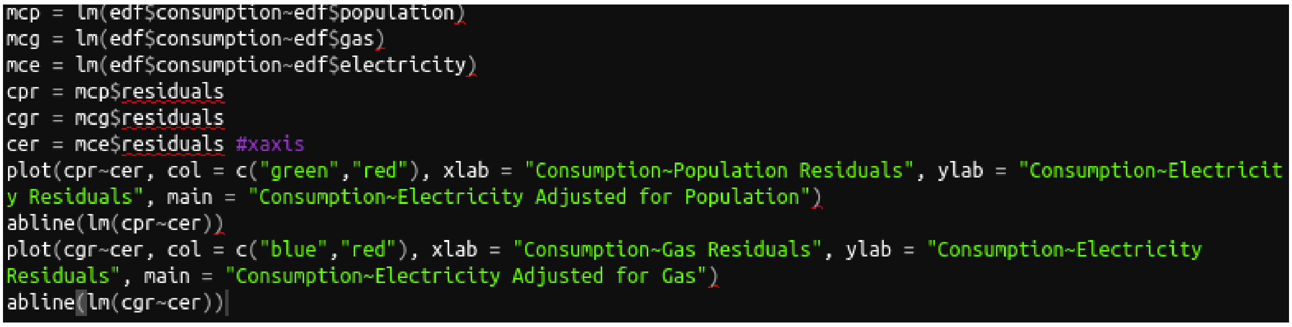


(iv)

Above: Population residuals shown in green, electricity in red
Below: Gas residuals shown in blue, electricity in red

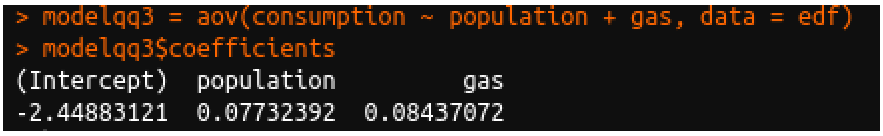


Code used for part iv:



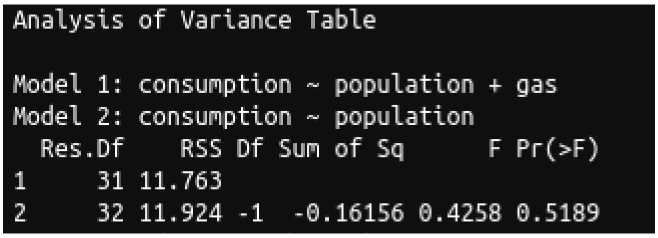
```
mcp = lm(edf$consumption~edf$population)
mcg = lm(edf$consumption~edf$gas)
mce = lm(edf$consumption~edf$electricity)
cpr = mcp$residuals
cgr = mcg$residuals
cer = mce$residuals #xaxis
plot(cpr~cer, col = c("green","red"), xlab = "Consumption~Population Residuals", ylab = "Consumption~Electricit
y Residuals", main = "Consumption~Electricity Adjusted for Population")
abline(lm(cpr~cer))
plot(cgr~cer, col = c("blue","red"), xlab = "Consumption~Gas Residuals", ylab = "Consumption~Electricity
Residuals", main = "Consumption~Electricity Adjusted for Gas")
abline(lm(cgr~cer))
```

C
(i)

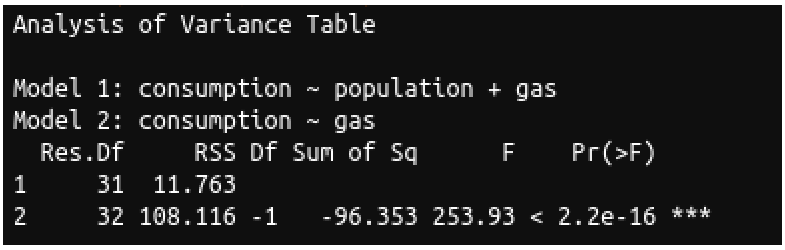


```
> modelqq3 = aov(consumption ~ population + gas, data = edf)
> modelqq3$coefficients
(Intercept)  population         gas
-2.44883121  0.07732392  0.08437072
```

As above, assessing if individual co-efficients make a statistically significant contribution to the model fit is done by comparing model fits with & without each corresponding variable.



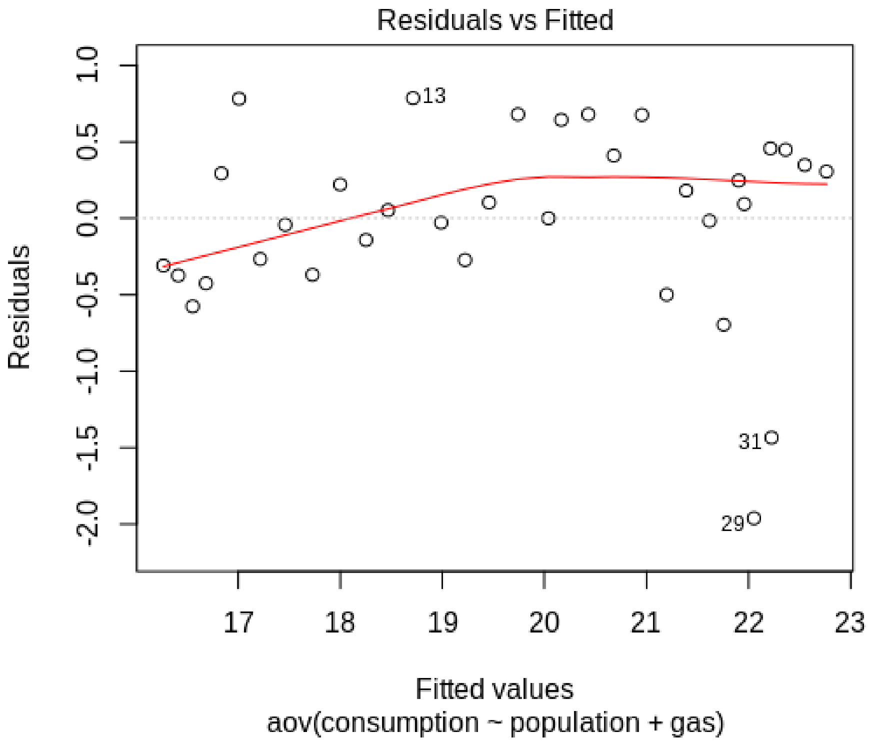
```
Analysis of Variance Table

Model 1: consumption ~ population + gas
Model 2: consumption ~ population
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     31 11.763
2     32 11.924 -1  -0.16156 0.4258 0.5189
```



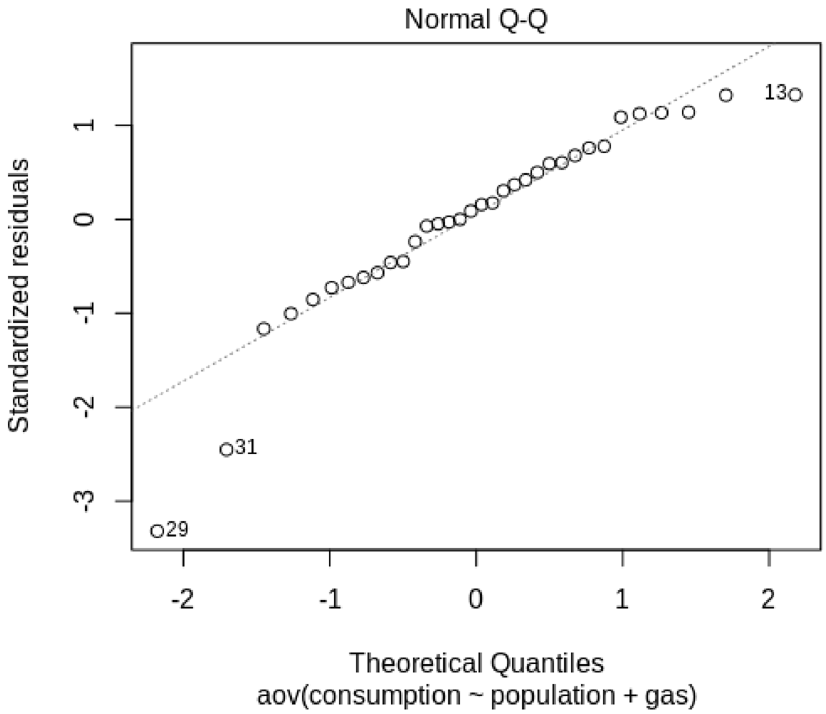
```
Analysis of Variance Table

Model 1: consumption ~ population + gas
Model 2: consumption ~ gas
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     31  11.763
2     32 108.116 -1   -96.353 253.93 < 2.2e-16 ***
```

The analysis indicated the null hypothesis should be rejected and that at least 1 of the co-efficients makes a statistically significant contribution to the model fit (population). Removing ‘gas’ from the model did not have a statistically significant effect on the fit and the simpler model without gas is preferable which supported earlier analysis showing the importance of the variable population.

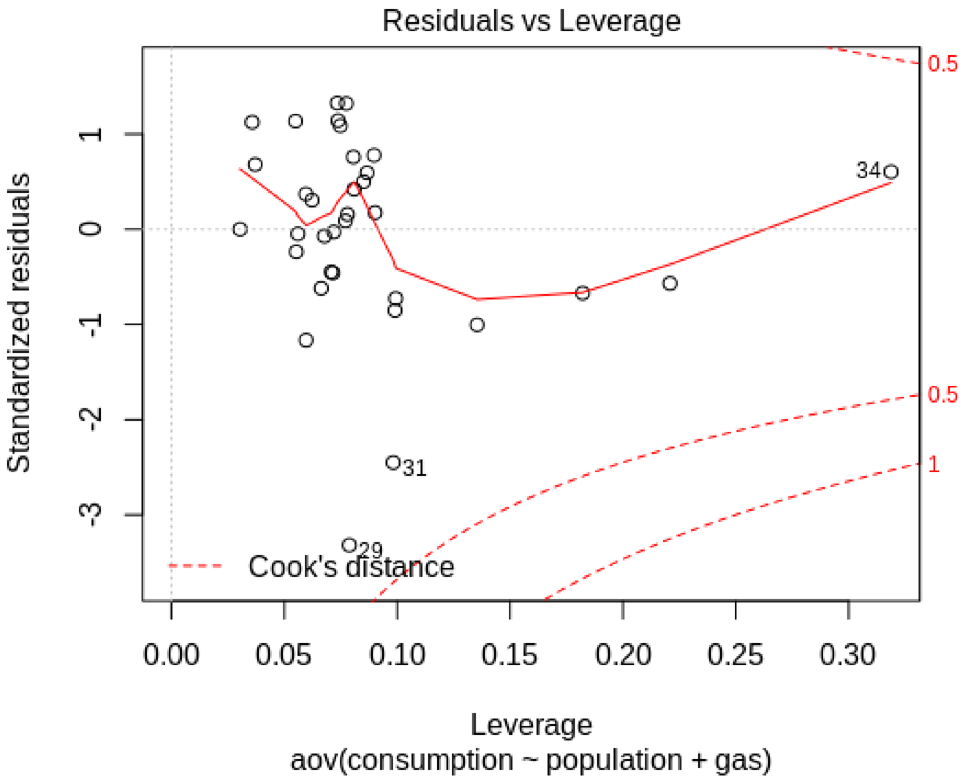
(ii)
Plot of residuals vs fitted values appears normal (approximately 0 mean and no sharp trend), although it does highlight two points of interest (29 & 31)



A QQ plot of standardised residuals again shows an approximate straight line distribution with three points of interest that deviate from the rest.

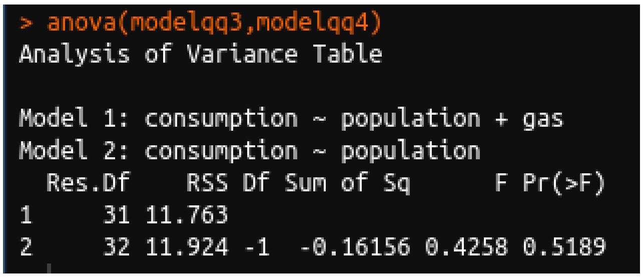


The residuals vs leverage plot allows the effect of the outlying values highlighted in the previous plots to be thoroughly investigated. The potential leverage points 28 & 31 both had Cook's distances of less than .5 and do not justify removal & refitting of the model.



(iii)
The F test allows comparison of two models where the second model contains a subset of the explanatory variables in the first.

H0: There is no difference in fit of the two models
HA: The larger model is a better fit



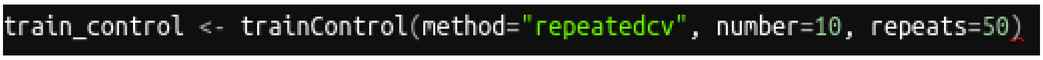
```
> anova(modelqq3,modelqq4)
Analysis of Variance Table

Model 1: consumption ~ population + gas
Model 2: consumption ~ population
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     31 11.763
2     32 11.924 -1  -0.16156 0.4258 0.5189
```

The calculated F value of 0.4258 and p value of .5189 allows to conclude that gas price is not significantly correlated with consumption i.e. we fail to reject the null hypothesis. Generally, an F statistic of 1 or less is expected if the null hypothesis is true.

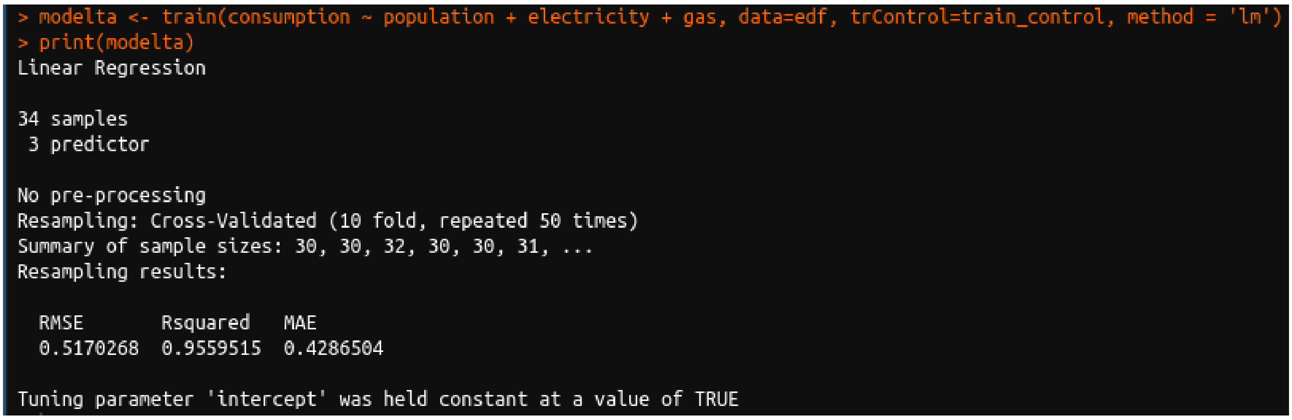
d

The performance of each model was assessed using the caret & klar packages to cary out 10-fold cross-validation with 50 repeats per model. The most comlex model taking into account population, electricity and gas had the highest R squared (.956) but it's performance was only marginally better than the simplest model (.9494) which relied on population alone. Surprisingly, the intermediate model had an inferior R squared to the simpler model. Generally speaking the R squared should only increase with the addition of more explanatory variables but the differences were extremely small.

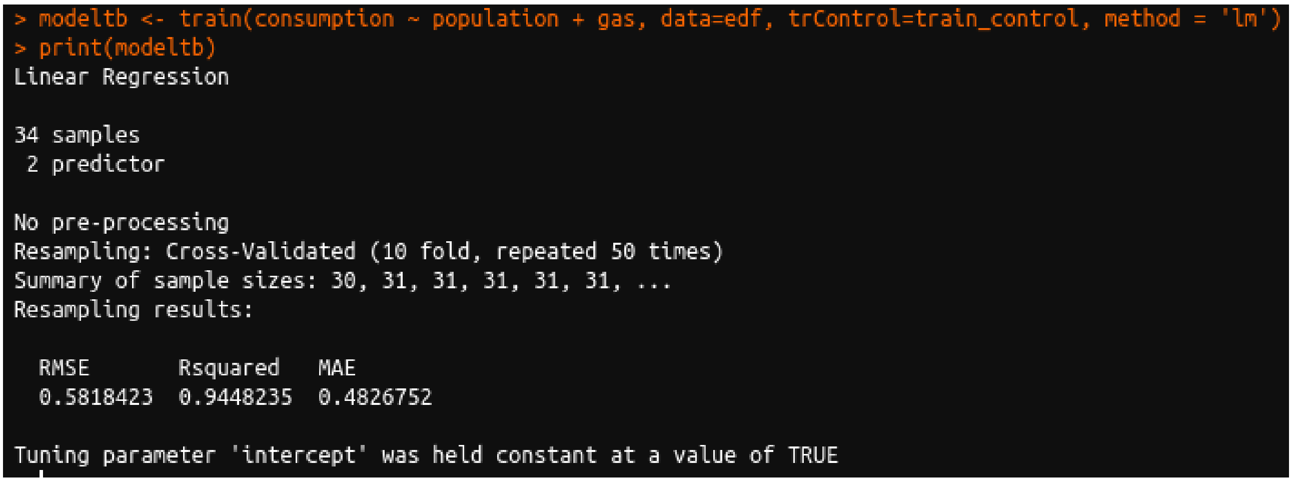
The variable train_control used in model training was defined as follows:



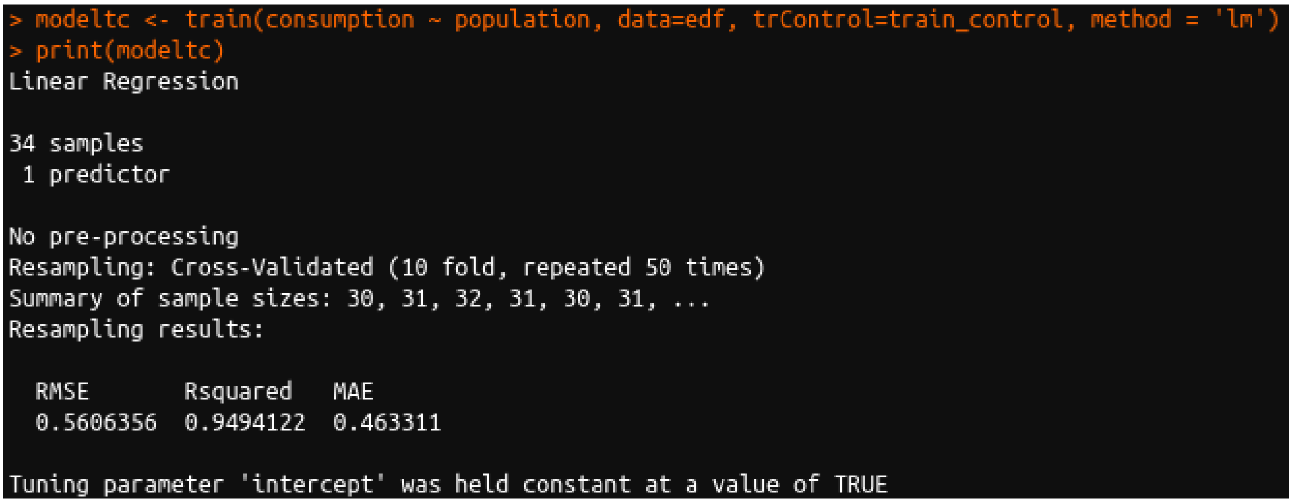
```
train_control <- trainControl(method="repeatedcv", number=10, repeats=50)
```



```
> modelta <- train(consumption ~ population + electricity + gas, data=edf, trControl=train_control, method = 'lm')
> print(modelta)
Linear Regression

34 samples
 3 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 50 times)
Summary of sample sizes: 30, 30, 32, 30, 30, 31, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.5170268  0.9559515  0.4286504

Tuning parameter 'intercept' was held constant at a value of TRUE
```



```
> modeltb <- train(consumption ~ population + gas, data=edf, trControl=train_control, method = 'lm')
> print(modeltb)
Linear Regression

34 samples
 2 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 50 times)
Summary of sample sizes: 30, 31, 31, 31, 31, 31, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.5818423  0.9448235  0.4826752

Tuning parameter 'intercept' was held constant at a value of TRUE
```



```
> modeltc <- train(consumption ~ population, data=edf, trControl=train_control, method = 'lm')
> print(modeltc)
Linear Regression

34 samples
 1 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 50 times)
Summary of sample sizes: 30, 31, 32, 31, 30, 31, ...
Resampling results:

  RMSE       Rsquared   MAE     
  0.5606356  0.9494122  0.463311

Tuning parameter 'intercept' was held constant at a value of TRUE
```

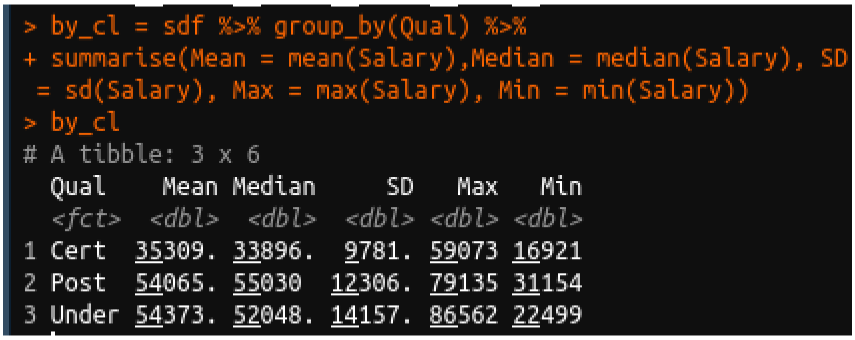
E.

The additional complexity associated with including extra variables in a model should be justified by increased predictive performance; which was not clearly the case in this analysis. The simplest model predicting electricity consumption based on population alone is intuitively reasonable and consideration of extra variables did not confer much benefit; the third model is therefore my recommendation.

Section 3

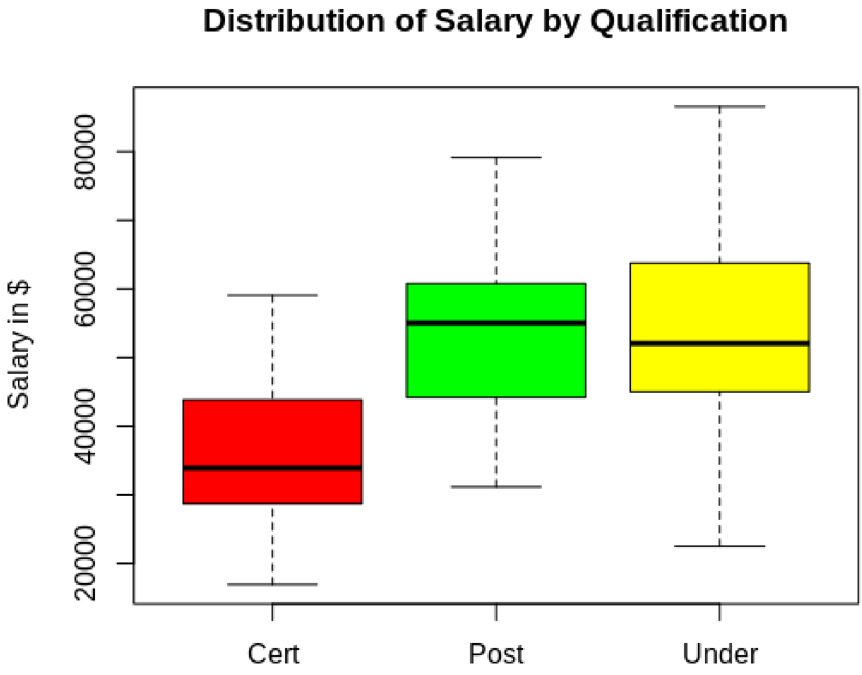
a

This dataset consisted of two observations of two explanatory variables (years & qualification) and the response variable salary for 150 engineers. Salary and years were treated as numeric variables while qualification was categorical. On read-in, the variable ‘qual’ had to be manually converted to a factor. The number of observations for each class was investigated first and the data was found to be perfectly balanced (50 observations for each class). Basic EDA was carried out using a combination of base R, tidyr and dplyr. Summary statistics grouped by qualifation were collected first.

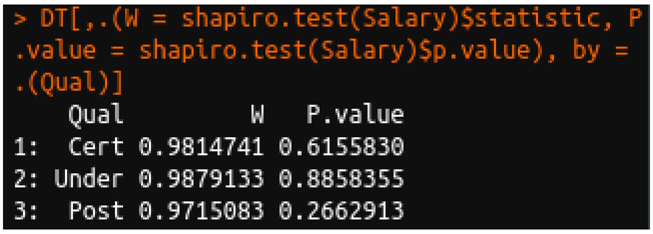


```
> by_cl = sdf %>% group_by(Qual) %>%
+ summarise(Mean = mean(Salary),Median = median(Salary), SD
 = sd(Salary), Max = max(Salary), Min = min(Salary))
> by_cl
# A tibble: 3 x 6
  Qual     Mean Median     SD   Max   Min
  <fct>   <dbl>  <dbl>  <dbl> <dbl> <dbl>
1 Cert   35309. 33896.  9781. 59073 16921
2 Post   54065. 55030  12306. 79135 31154
3 Under  54373. 52048. 14157. 86562 22499
```

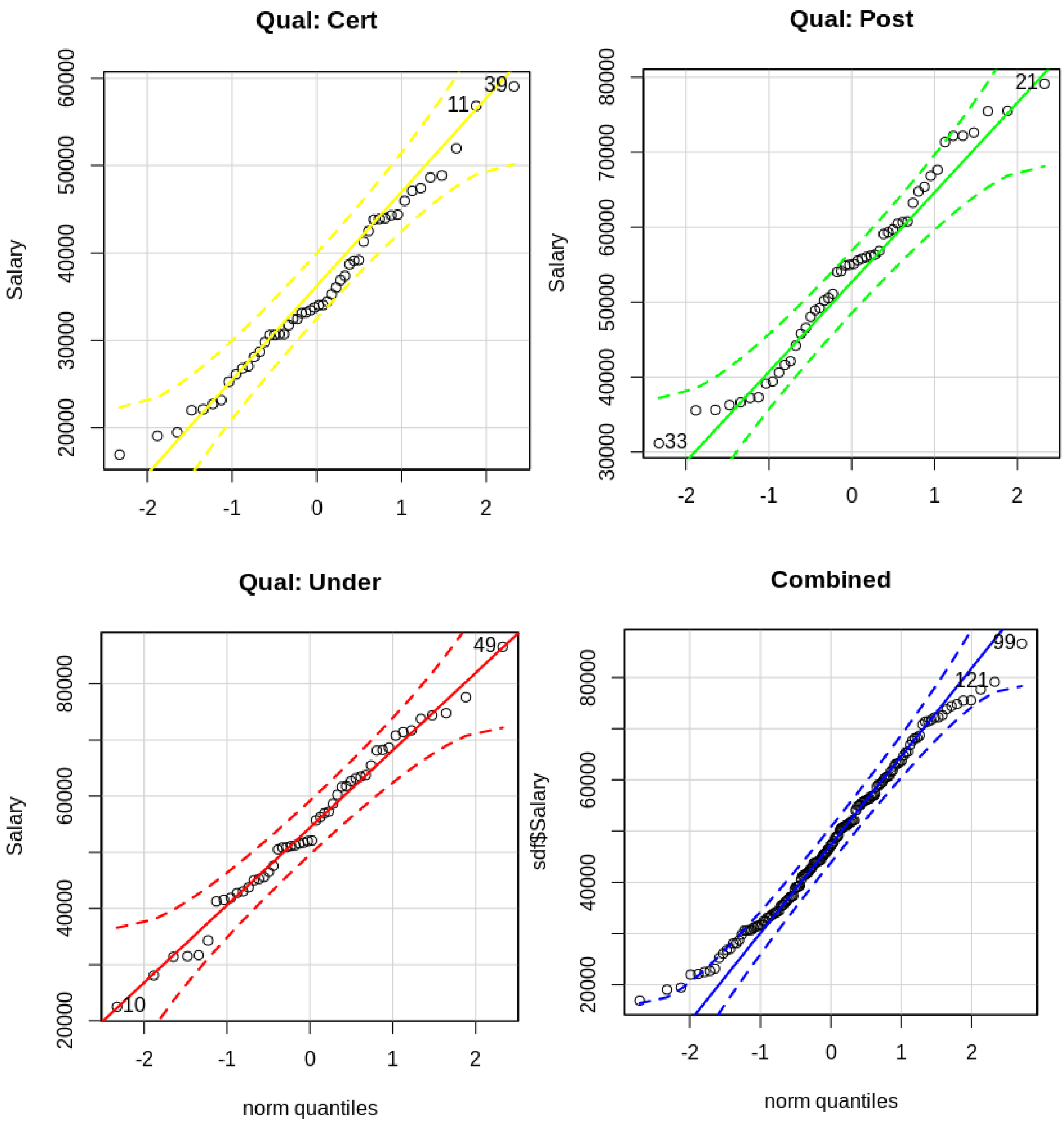
Given the similar ranges for all three classes; a boxplot provided the clearest method of visualising their distribution. No outlying values were detected for any class.



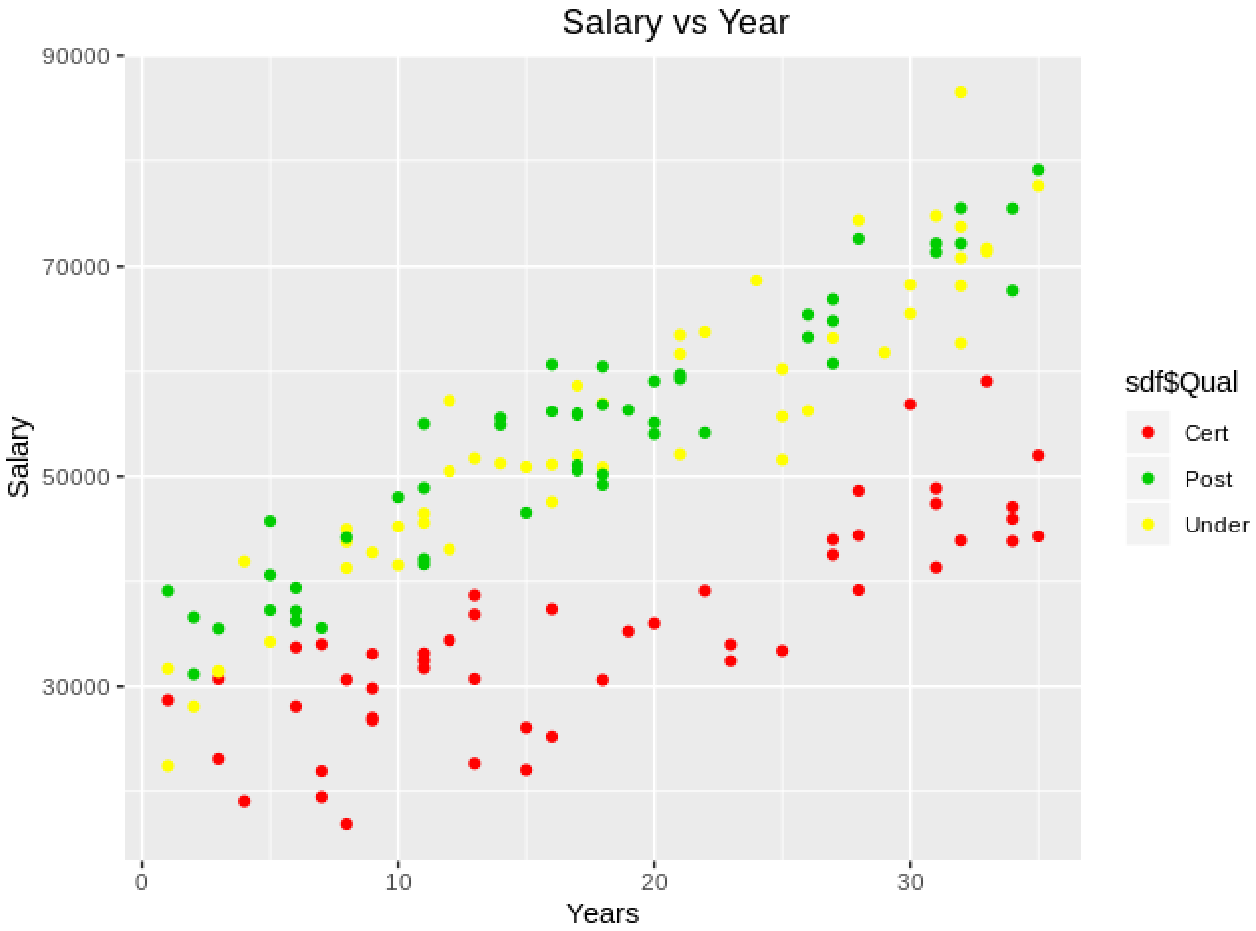
The boxplot and summary statistics showing little disparity between mean & median for salary in all three classes suggested normal distributions but this was confirmed using qqplots and a class wise Shapiro-Wilk normality test.



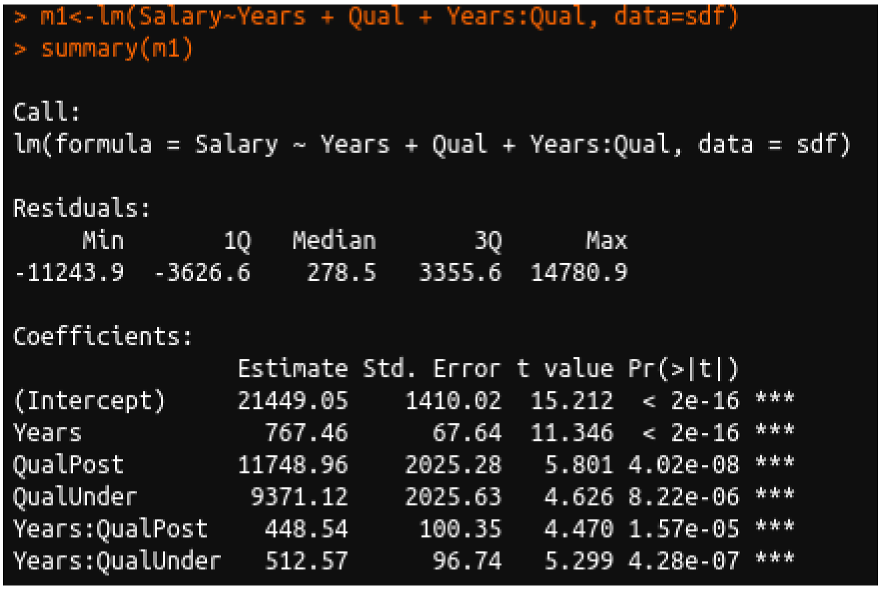
```
> DT[,.(W = shapiro.test(Salary)$statistic, P
.value = shapiro.test(Salary)$p.value), by =
.(Qual)]
    Qual         W   P.value
1:  Cert 0.9814741 0.6155830
2: Under 0.9879133 0.8858355
3:  Post 0.9715083 0.2662913
```



Finally, salary was scatter-plotted against years with color coding for each class. This suggested a positive correlation between years and salary for each class. It suggested lower salaries in the cert group but little to no difference between post & under

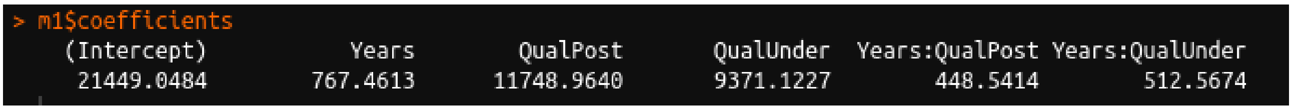


b



```
> m1<-lm(Salary~Years + Qual + Years:Qual, data=sdf)
> summary(m1)

Call:
lm(formula = Salary ~ Years + Qual + Years:Qual, data = sdf)

Residuals:
     Min       1Q   Median       3Q      Max 
-11243.9  -3626.6    278.5   3355.6  14780.9 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     21449.05    1410.02  15.212  < 2e-16 ***
Years             767.46      67.64  11.346  < 2e-16 ***
QualPost        11748.96    2025.28   5.801 4.02e-08 ***
QualUnder        9371.12    2025.63   4.626 8.22e-06 ***
Years:QualPost    448.54     100.35   4.470 1.57e-05 ***
Years:QualUnder   512.57      96.74   5.299 4.28e-07 ***
```

c

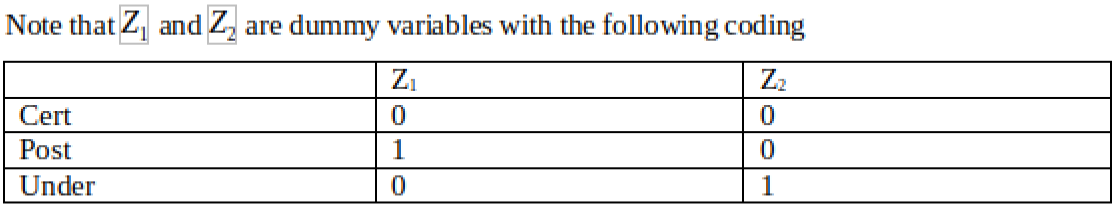


```
> m1$coefficients
   (Intercept)          Years       QualPost      QualUnder  Years:QualPost Years:QualUnder
    21449.0484       767.4613     11748.9640      9371.1227        448.5414        512.5674
```

$$y=\beta_0+\beta_1 Years+\beta_2 Z_1+\beta_3 Z_2+\beta_4 Years\, Z_1+\beta_5 Years\, Z_2+e$$

y = 21449 + 767.4613*Years + 11748.96**Z1* + 9371**Z2* + 448.54*Years**Z1* + 512*Years**Z2*

β0 = Intercept for salaries with Cert qualification
β1 = The effect of a unit increase in Year on Salary for Cert qualification
β2 = The effect on the intercept from switching to Post qualification
β3 = The effect on the intercept from switching to Under qualification
β4 = The effect on the slope of the salary vs years regression line when switching from cert to post
β5 = The effect on the slope of the salary vs years regression line when switching from cert to under



Note that Z_1 and Z_2 are dummy variables with the following coding

	Z_1	Z_2
Cert	0	0
Post	1	0
Under	0	1

d

By inserting Years = 15, Z1 = 1 and Z2 = 0 the above model returns a predicted salary of: 51437.98

Comparing this with the scatterplot confirms it is similar to other datapoints with similar characteristics.

E

y = 21449 + 767.4613*(Years) + 11748.96*Z1 + 9371*Z2 + 448.54*Years*Z1 + 512*Years*Z2

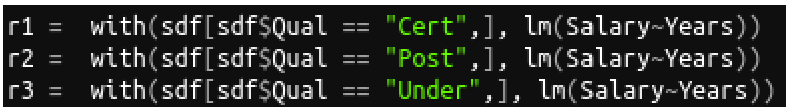
F.

To test whether the three regression lines are parallel the three equations were computed seperately in R.

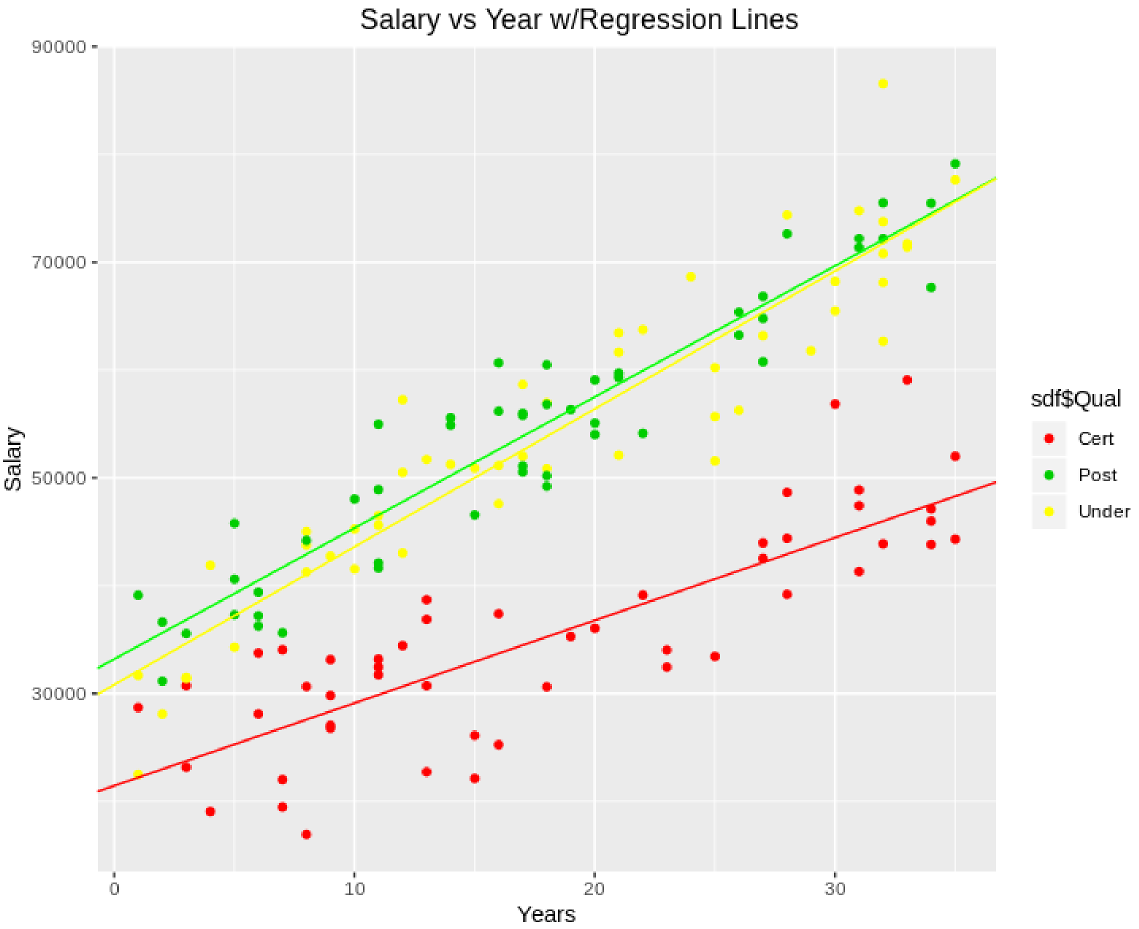
H0

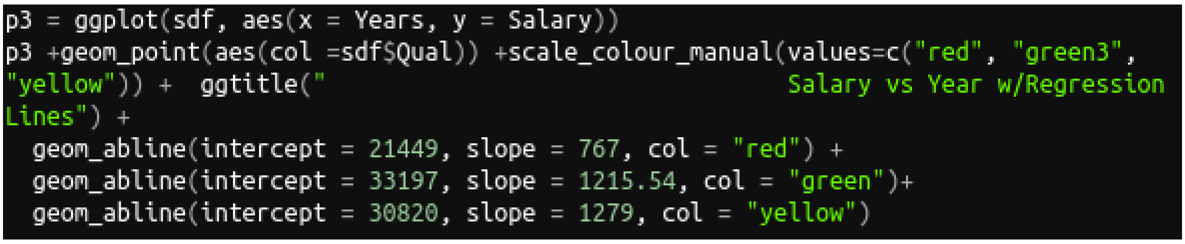
The null hypothesis was that all three regression lines are parallel i.e. the regression co-efficients for each model are equal.

H1. The alternative hypothesis is that at least one of the regression co-efficients is different to the others.



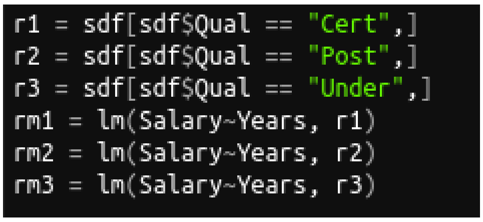
```
r1 =  with(sdf[sdf$Qual == "Cert",], lm(Salary~Years))
r2 =  with(sdf[sdf$Qual == "Post",], lm(Salary~Years))
r3 =  with(sdf[sdf$Qual == "Under",], lm(Salary~Years))
```



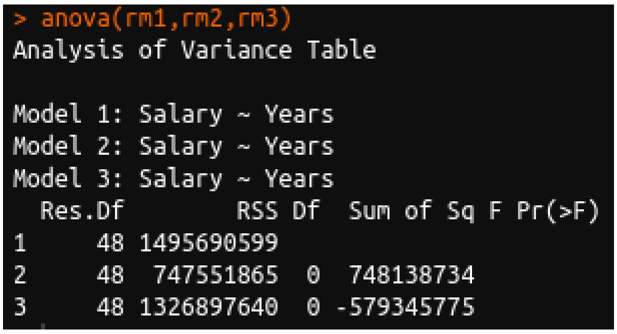


```
p3 = ggplot(sdf, aes(x = Years, y = Salary))
p3 +geom_point(aes(col =sdf$Qual)) +scale_colour_manual(values=c("red", "green3",
"yellow")) +  ggtitle("                                    Salary vs Year w/Regression
Lines") +
  geom_abline(intercept = 21449, slope = 767, col = "red") +
  geom_abline(intercept = 33197, slope = 1215.54, col = "green")+
  geom_abline(intercept = 30820, slope = 1279, col = "yellow")
```

An ANOVA test on the three sets of models as well as each pair in isolation was attempted in R but the p-value was not reported by the ANOVA function and the cause of this could not be found. Interpretation of the regression co-efficients and visual inspection strongly suggests the regression lines for ‘Post’ and ‘Under’ are parallel while ‘Under’ appears to have a much more gradual slope than the other two.



```
r1 = sdf[sdf$Qual == "Cert",]
r2 = sdf[sdf$Qual == "Post",]
r3 = sdf[sdf$Qual == "Under",]
rm1 = lm(Salary~Years, r1)
rm2 = lm(Salary~Years, r2)
rm3 = lm(Salary~Years, r3)
```



```
> anova(rm1,rm2,rm3)
Analysis of Variance Table

Model 1: Salary ~ Years
Model 2: Salary ~ Years
Model 3: Salary ~ Years
  Res.Df        RSS Df  Sum of Sq F Pr(>F)
1     48 1495690599
2     48  747551865  0  748138734
3     48 1326897640  0 -579345775
```