Assignment 2 Submission

Ian Chen (24227644) and Marco Gunawan(23780778)

16/05/2025

Statement of Contribution

Both people did both questions and discussed before submission.

Question 1: Air Pollution and Mortality. Does pollution kill people? (30 marks)

(a) (5 marks) Carry out exploratory data analysis (EDA) of this dataset.

Answer:

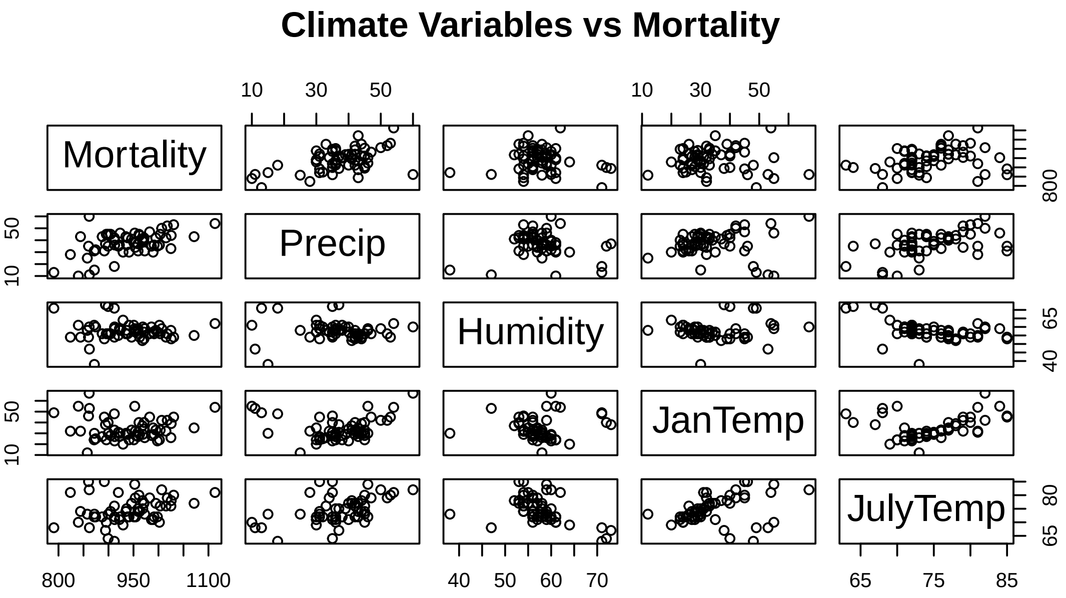
Climate EDA: The Mortality-Precip plot shows a strong postive linear relationship, indicating that as the precipitation increases, so does the Mortality. The Mortality-Humidity plot shows no clear relationship as datapoints appear randomly scattered. The Mortality-JanTemp plot depeicts either no clear relationship or a very weak negative relationship. Lastly the Mortality-JulyTemp plot depicts a moderate positive relationship with a few outliers.

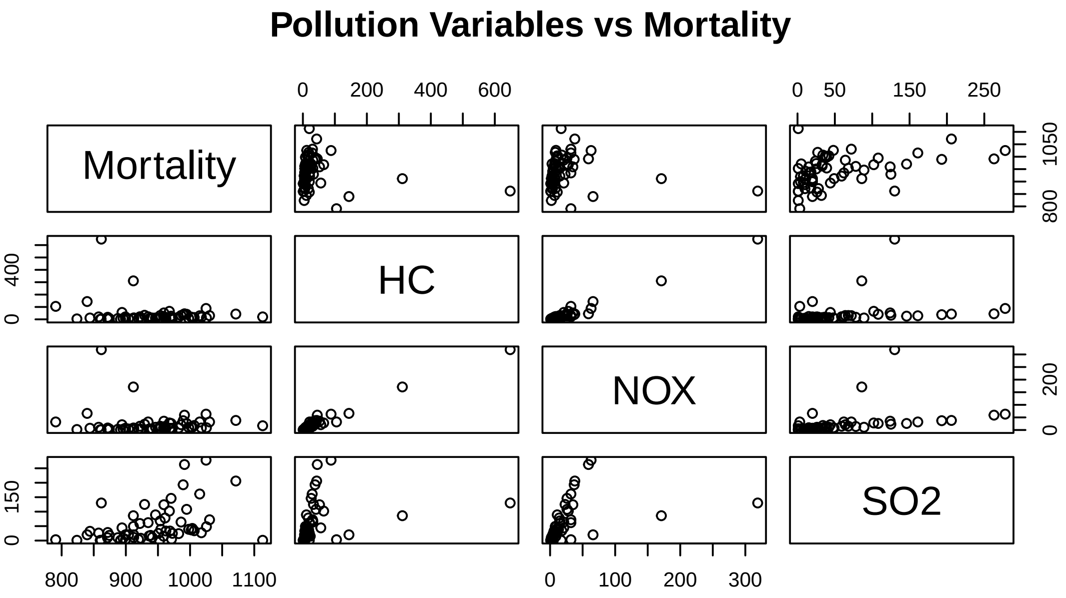
Socioeconomic Variables EDA: From the pairs plot, we can see that the Mortality-Over65 graph appears to have no apparent relationship.The Mortality-House plot shows a weak to moderate positive relationship. The Mortality-Educ plot shows a strong negative relationship. The Mortality-Sound plot shows a moderate to strong negative relationship. The Mortality-Density plot has a Moderate positive relationship. The Mortality-NonWhite plot shows a strong positive relationship. The Mortality-WhiteCol plot shows no clear relationship. Lastly the Mortality-Poor plot shows a moderate positive relationship.

Pollution Variables EDA: The Mortality-HC plot shows no correlation between the two variables, same with the Mortality-NOX plot.This is because the datapoints appear randomly scattered. However the Mortality SO2 plot depoicts a Moderate positive relationship. This means as the SO2 increases so does mortality.

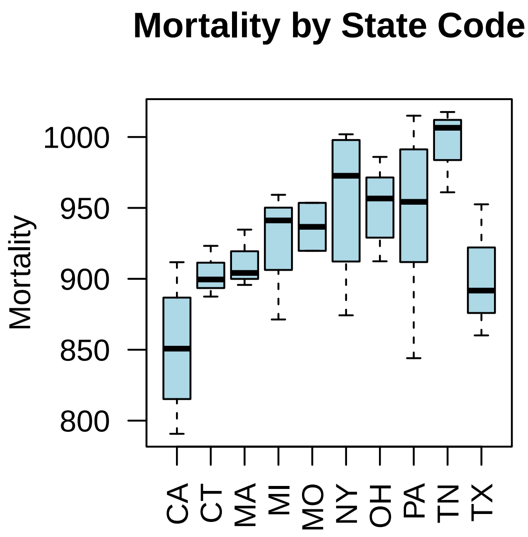
Mortality by Region Analysis: The south region has the highest rates of mortality, the highest median (975) as well as the widest range. The median and range are notably higher than then rest of the regions. The West region has the lowest average mortality rate (875) as well as the most narrow spread. The Midwest and Northeast are similar in average values (~940) and an outlier in the Midwest data. The regional differences stated suggests that the region a person lives in may influence their mortality rate.

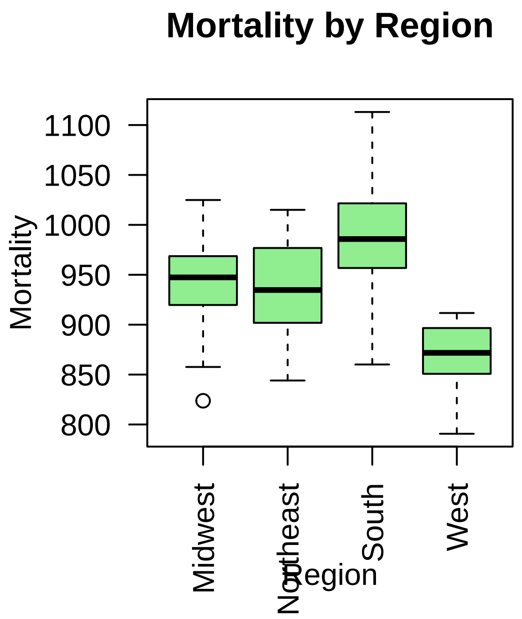
Mortality by State Code Analysis: Before we plotted out graph, I first examined the data, using table(pollution$State.code). Many of our states only have a singular datapoint and therefore are too small a sample size. Excluding the states with one datapoint we are left with 10 that we plot. The mean mortality rate differs greatly from each state. The states NY, PA, and OH have higher median mortality rates while CA and TX have significantly lower ones. States such as CA and PA having much wider interquartile range indicating a high variance. This suggests that the state an individual lives is relavent to their mortality rate.

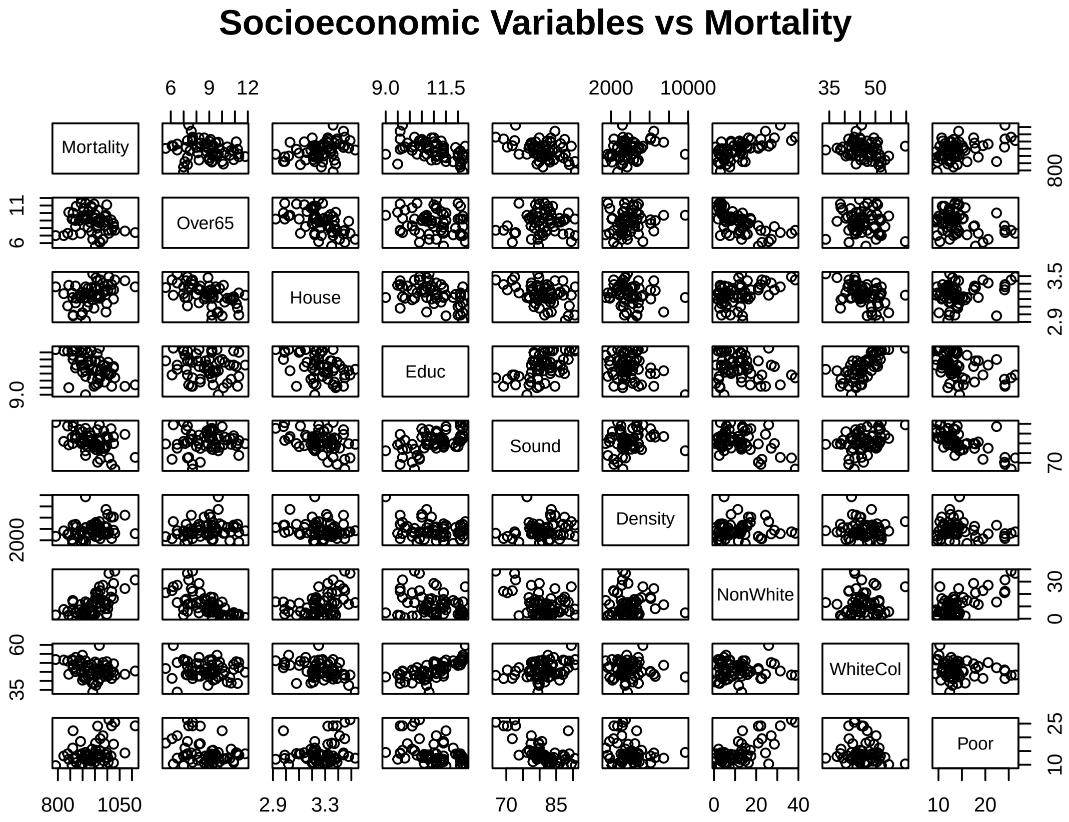




```
AL CA CO CT DC DE FL GA IL IN KS KY LA MA MD MI MN MO NC NY OH OR PA RI TN TX
 1  4  1  3  1  1  1  1  1  1  1  1  1  3  1  3  1  2  1  6  8  1  6  1  3  3
VA WA WI
 1  1  1
```







(b) (9 marks) Perform model selection process using the all subset method to arrive at good regression models that account for variation in mortality between the cities that can be attributed to differences in climate and socioeconomic factors. Identify optimal model(s) based on each of the adjusted R2 , BIC and C_p. Use a maximum of 10 variables for the selection. For each criterion, state the number and names of the selected variables.

For the adjusted Rˆ2 model, we have 7 variables. Those variables are Precip, JanTemp, JulyTemp, House, Educ, Density and NonWhite. For the BIC model, we have 4 variables. Those variables are JanTemp, House, Educ and NonWhite. For the Cp model, we have 6 variables. Those variables are Precip, JanTemp, JulyTemp, Educ, Density and NonWhite.

```
[1] "Precip"   "JanTemp"  "JulyTemp" "House"    "Educ"     "Density"  "NonWhite"
```

```
[1] "JanTemp"  "House"    "Educ"     "NonWhite"
```

```
[1] "Precip"   "JanTemp"  "JulyTemp" "Educ"     "Density"  "NonWhite"
```

(c) (4 marks) Fit the model with the lowest C_p as obtained in (b). Write the equation of this fitted model. Interpret this model.

Answer:

Equation: Mortality = 1242.0+ 1.401 * Precip -1.684 * JanTemp - 2.8 * JulyTemp -16.16 * Educ + 0.00757 * Density + 5.275 * Nonwhite

Interpretation: From viewing the model summary, the Rˆ2 value tells us that the model explains approximately 70.86% of the variation in the data's mortality rate. The model summary also tells us the F-Statistic and the p-value. Those being 21.48 and 1.305e-12 respectively. The large F-statistic (1<) suggests that the model has significant variables chosen. The small p-value (0.05>) also is evidence against the null hypothosies and that our model is statically significant. We then interpretthe coefficients for our model. A unit increase in Precip increases the rate by 1.401 deaths per 100,000. For each degree increase in JanTemp, mortality rate decreases by 1.684 deaths per 100,000. For each degree increase in JulyTemp, mortality rate decreases by 2.840 deaths per 100,000. Every year spend in education decreases the mortality rate by 16.16 per 100,000. Per unit increase in Density, the mortality rate increases by 0.00757 per 100,000. For every percentage increase in NonWhite, the mortality increases by 5.275 deaths per 100,000.

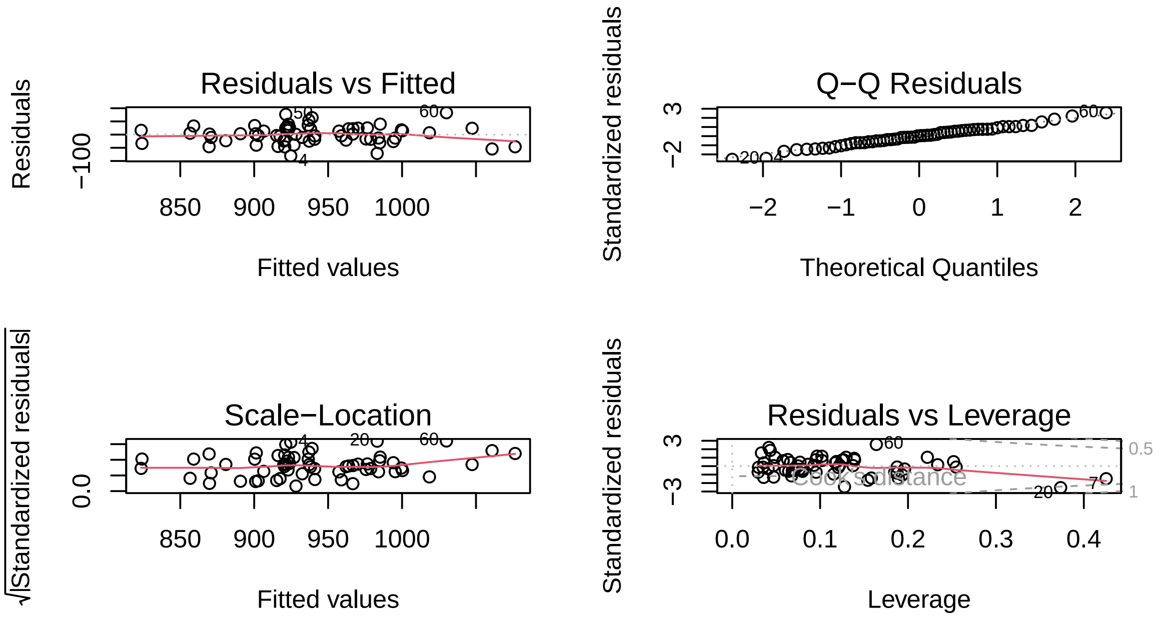
```
Call:
lm(formula = Mortality ~ Precip + JanTemp + JulyTemp + Educ + 
    Density + NonWhite, data = poll_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-80.685 -21.529   1.422  22.777  83.055 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.242e+03  1.233e+02  10.078 6.41e-14 ***
Precip       1.401e+00  6.074e-01   2.307   0.0250 *  
JanTemp     -1.684e+00  5.330e-01  -3.161   0.0026 ** 
JulyTemp    -2.840e+00  1.289e+00  -2.203   0.0319 *  
Educ        -1.616e+01  6.652e+00  -2.429   0.0186 *  
Density      7.570e-03  3.316e-03   2.283   0.0265 *  
NonWhite     5.275e+00  6.906e-01   7.639 4.24e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.43 on 53 degrees of freedom
Multiple R-squared:  0.7086,    Adjusted R-squared:  0.6756 
F-statistic: 21.48 on 6 and 53 DF,  p-value: 1.305e-12
```

(d) (7 marks) Perform diagnostics checking on the model in (c). Do you think there are influential points in the data? Identify the cities which are influential points using leverage and Cook's distance respectively.

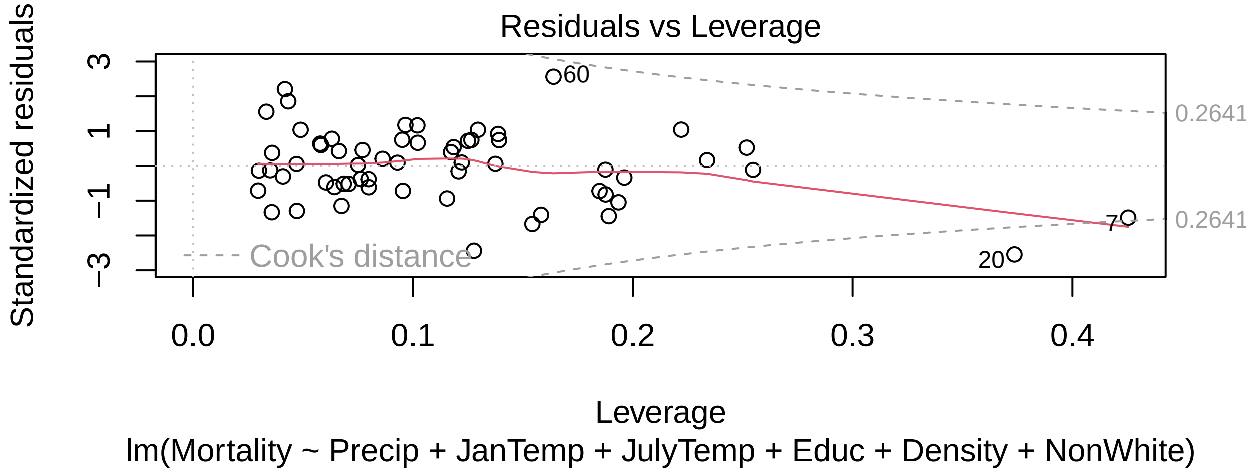
Answer: We first generate our diagnostic plots, checking the Residuals vs Fitted, Q-Q Residuals , Scale-Location and Residuals vs Leverage. The Residuals vs Fitted has a mostly random scatter of data points. There's minor curvature at the right end but overall this plot suggests the relationship is approximately linear. The Q-Q Residuals follows the diagonal line closely with minor deviations at the ends. Overall, the plot suggests that the residuals are approximately normally distributed.The Scale Location is fairly horizontal however towards the right end it appears to curve upwards. This suggests the variance may be slightly uneven. The Residuals vs Leverage plot depicts data points that require further investigation as they have a high leverage and are far from x=0.

Taking a further look at influential points using leverage and Cook's distance, I use the equations Leverage cutoff = (2(p+1))/ n and cook's cutoff 2 * (p + 1) / (n - p - 1) to find influential points. Using these equations as well as the linear model we can find the points above the cutoffs. The data points that are above the cook's cutoff is data point 20, the city being York. The data points above the leverage cutoff is data points 3,7,8,19 and 20. Their cities are San Diego, Miami, Los Angeles,San Francisco and York.

```r
poll_data <- read.csv("Pollution.csv")
cp_model <- lm(Mortality ~ Precip + JanTemp + JulyTemp + Educ +
                 Density + NonWhite, data = poll_data)
# Model dimensions
p <- length(coef(cp_model)) - 1
n <- nrow(poll_data)
#Diagnostics Plot
par(mfrow = c(2, 2))
plot(cp_model)
```



```r
#The influential points are those with Cook's distance
    par(mfrow = c(1, 1))
    cook_cutoff <- 2 * (p + 1) / (n - p - 1)
    plot(cp_model, which = 5, cook.levels = cook_cutoff)
```



```
#Find Influential Points using Cook's Distance
cooks_d <- cooks.distance(cp_model)
cook_cutoff_points <- which(cooks_d > cook_cutoff)
cat(cook_cutoff_points)
```

```
20
```

```
#Find high leverage
lev_cutoff <- 2 * (p + 1) / n
high_leverage <- which(hatvalues(cp_model) > lev_cutoff)
cat(high_leverage)
```

```
3 7 8 19 20
```

```
#Find the Cities
print(poll_data[cook_cutoff_points, "City"])
```

```
[1] "York, PA"
```

```
print(poll_data[high_leverage, "City"])
```

```
[1] "San Diego, CA"     "Miami, FL"         "Los Angeles, CA"
[4] "San Francisco, CA" "York, PA"
```

(e) (5 marks) Using the model obtained in (c), add the three pollution variables (transformed to their natural logarithm) and obtain the p-value from the extrasum-of-squares F-test due to their addition. Summarise your findings in a few concise sentences.

Answer: First I created the expanded model by adding the log-transformed variables of HC, NOX and SO2 into the base model.I used the extrasum-of-squares F-test on both the base and expanded linear models, giving me the p-value of 0.008313. This p-value is small, being less than 0.05. This means that the expanded model is statistically significant. This suggests the addition of the logarithm of the three pollution variables improved the model's ability to predict mortality rates. Therefore, we reject the null hypothosies.

```
# Answer code here
poll_data <- read.csv("Pollution.csv")
cp_model <- lm(Mortality ~ Precip + JanTemp + JulyTemp + Educ
               + Density + NonWhite,
               data = poll_data)
cp_model_expanded <- lm(Mortality ~ Precip + JanTemp + JulyTemp + Educ
                        + Density + NonWhite + log(HC) + log(NOX) +log(SO2),
                        data = poll_data)
anova(cp_model, cp_model_expanded)
```

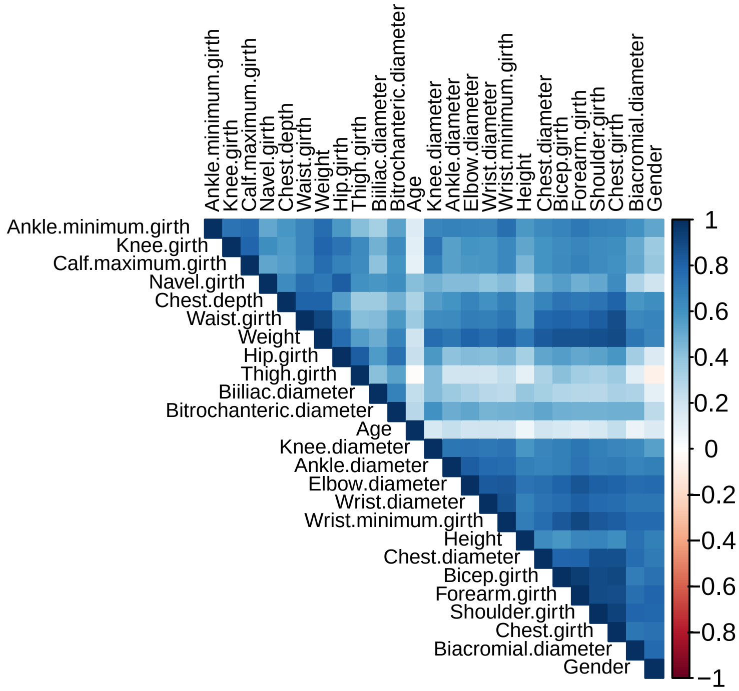
```
Analysis of Variance Table

Model 1: Mortality ~ Precip + JanTemp + JulyTemp + Educ + Density + NonWhite
Model 2: Mortality ~ Precip + JanTemp + JulyTemp + Educ + Density + NonWhite +
    log(HC) + log(NOX) + log(SO2)
  Res.Df   RSS Df Sum of Sq     F   Pr(>F)
1     53 66518
2     50 52712  3     13806 4.365 0.008313 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Question 2: Body Measurements (25 marks)

(a) (4 marks) Carry out exploratory data analysis (EDA) of this dataset before you do any modelling.

Using str() and with the information provided with the task, we know that the dataset is mostly numeric, with 26 variables. With two variables standing out, primarily, X (the row index) and gender (a factor with two levels). I cleaned the data by removing X and performing EDA from then on.

As there are 26 variables, and producing a pairs plot would result in 625 scatterplots, resulting in a plot which is too dense to view meaningfully, we have opted for a heatmap analysis instead.



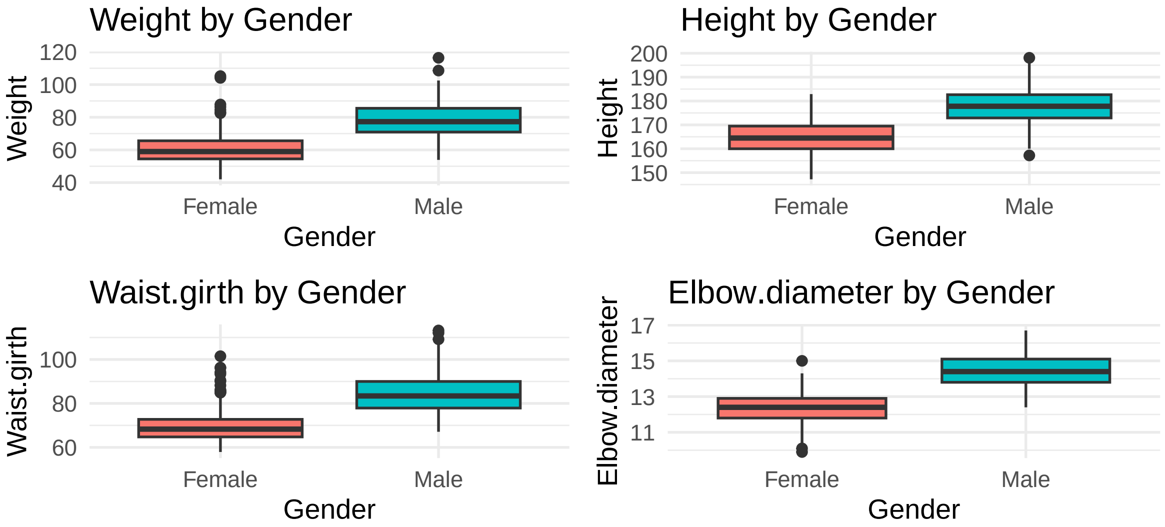
Having a look at the heatmap itself, we can see that particularly weight, waist girth, chest girth and various other girth measurements appears to be strongly correlated.

As gender is a categorical variable, I decided to create boxplots for further analysis with the numeric variables that are most strongly correlated to gender.

For the main response variable, weight, we can see that male has a notably higher median weight (~78kg) to female's median at (~60kg). Both groups show some outliers at the upper end, particularly among females. These measurements show clear sexual dimorphism in human body proportions, with minimal overlap in distributions for measurements like height and waist girth.

(b) (10 marks) After the exploratory analysis has been carried out, construct two multiple linear regression models for this dataset using the training set.

We begin by splitting the dataset into a training set and a testing set. Using both forward and backwards selection, we yield the following model equations.



```r
# Data Selection
set.seed(2401)
test_index <- sample(nrow(body_clean), floor(0.2 * nrow(body_clean)))
BodyMeasurementsTrain <- body_clean[-test_index, ]
BodyMeasurementsTest <- body_clean[test_index, ]
```

The first model given by the forward AIC step function (Model 1) is shown below:

```
Call:
lm(formula = Weight ~ Waist.girth + Height + Thigh.girth + Forearm.girth +
    Shoulder.girth + Calf.maximum.girth + Hip.girth + Chest.girth +
    Knee.diameter + Age + Chest.depth + Gender + Knee.girth +
    Chest.diameter + Bicep.girth + Elbow.diameter + Wrist.minimum.girth +
    Wrist.diameter, data = BodyMeasurementsTrain)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3979 -1.2832 -0.0807  1.2514  9.0969

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)         -118.94354    3.00524 -39.579  < 2e-16 ***
Waist.girth            0.35664    0.02917  12.228  < 2e-16 ***
Height                 0.30191    0.01889  15.986  < 2e-16 ***
Thigh.girth            0.23965    0.06154   3.894 0.000116 ***
Forearm.girth          0.48622    0.15402   3.157 0.001720 **
Shoulder.girth         0.07093    0.03335   2.127 0.034038 *
Calf.maximum.girth     0.34923    0.07048   4.955 1.08e-06 ***
Hip.girth              0.22429    0.04525   4.957 1.07e-06 ***
Chest.girth            0.13685    0.04120   3.321 0.000981 ***
Knee.diameter          0.46740    0.14904   3.136 0.001843 **
Age                   -0.06130    0.01386  -4.423 1.27e-05 ***
Chest.depth            0.30357    0.07640   3.974 8.45e-05 ***
Gender                -1.43739    0.57360  -2.506 0.012623 *
Knee.girth             0.18447    0.08431   2.188 0.029273 *
Chest.diameter         0.14876    0.08756   1.699 0.090117 .
Bicep.girth            0.15346    0.09535   1.609 0.108331
Elbow.diameter         0.22482    0.19554   1.150 0.250959
Wrist.minimum.girth   -0.42530    0.21911  -1.941 0.052983 .
Wrist.diameter         0.39274    0.25308   1.552 0.121510
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 2.137 on 387 degrees of freedom
Multiple R-squared:  0.976, Adjusted R-squared:  0.9749
F-statistic: 874.8 on 18 and 387 DF,  p-value: < 2.2e-16
```

And similarly, for the backward AIC step function (Model 2)

```
Call:
lm(formula = Weight ~ Chest.depth + Chest.diameter + Wrist.diameter +
    Knee.diameter + Shoulder.girth + Chest.girth + Waist.girth +
    Hip.girth + Thigh.girth + Bicep.girth + Forearm.girth + Knee.girth +
    Calf.maximum.girth + Wrist.minimum.girth + Age + Height +
    Gender, data = BodyMeasurementsTrain)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3908 -1.3082 -0.1104  1.2070  9.1297

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -119.42408    2.97727 -40.112  < 2e-16 ***
Chest.depth            0.30916    0.07627   4.053 6.11e-05 ***
Chest.diameter         0.14989    0.08759   1.711 0.087823 .
Wrist.diameter         0.47684    0.24238   1.967 0.049855 *
Knee.diameter          0.49901    0.14654   3.405 0.000730 ***
Shoulder.girth         0.07223    0.03334   2.167 0.030874 *
Chest.girth            0.14158    0.04101   3.452 0.000618 ***
Waist.girth            0.35145    0.02883  12.192  < 2e-16 ***
Hip.girth              0.22937    0.04505   5.091 5.56e-07 ***
Thigh.girth            0.23220    0.06123   3.792 0.000173 ***
Bicep.girth            0.15422    0.09538   1.617 0.106735
Forearm.girth          0.51681    0.15177   3.405 0.000730 ***
```

```
Knee.girth              0.18553    0.08434   2.200 0.028412 *
Calf.maximum.girth      0.34924    0.07051   4.953 1.09e-06 ***
Wrist.minimum.girth    -0.43639    0.21899  -1.993 0.046993 *
Age                    -0.05981    0.01380  -4.333 1.88e-05 ***
Height                  0.30696    0.01837  16.705  < 2e-16 ***
Gender                 -1.36753    0.57061  -2.397 0.017020 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 2.138 on 388 degrees of freedom
Multiple R-squared:  0.9759,    Adjusted R-squared:  0.9749
F-statistic: 925.5 on 17 and 388 DF,  p-value: < 2.2e-16
```

We notice that there are some variables tha tare a apart of this model that are not statistically significant, that is, their p-value is greater than 0.1. To yield two seperate models to be able to be compared however, we instead decided to remove all non-significant variables at once. And doing so sequentially, until all non-significant variables have been removed or until the overall model fit (adjusted R-squared, AIC, BIC) significantly worsen.

Thus, in the end we yield Model 1:

Weight = -120.9683 + 0.3591 × Waist.girth + 0.3101 × Height + 0.2724 × Thigh.girth + 0.5641 × Forearm.girth + 0.0936 × Shoulder.girth + 0.3385 × Calf.maximum.girth + 0.2197 × Hip.girth + 0.1818 × Chest.girth + 0.5451 × Knee.diameter - 0.0562 × Age + 0.2679 × Chest.depth - 1.2184 × Gender + 0.1559 × Knee.girth

We can see that all predictors in our model are statistically significant at the $p < 0.1$ level, with the intercept being -121, which has no real meaning as someone could never have negative weight. Model 1’s final adjusted R-squared value was 0.9744, with a F-statistic 1189 on 13 and 392 DF. Model 1’s final variables includes Waist.girth, Height, Thigh.girth, Forearm.girth, Shoulder.girth, Calf.maximum.girth, Hip.girth, Chest.girth, Knee.diameter, Age, Chest.depth, Gender and Knee.girth.

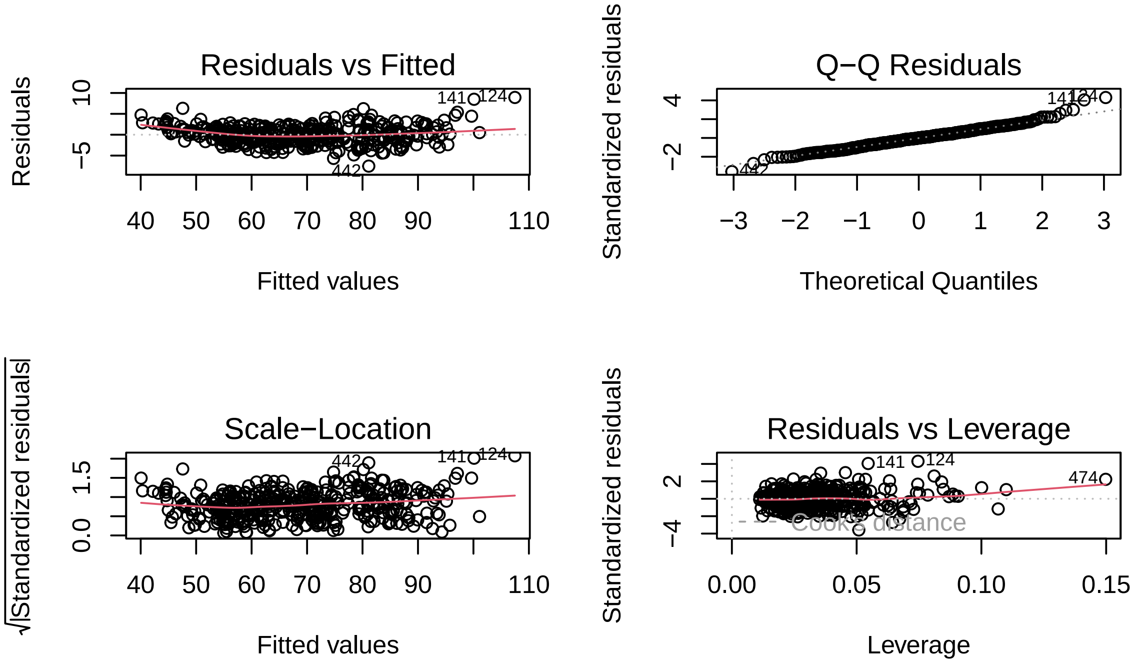
And similarly, for Model 2:

Weight = -120.2155 + 0.2838 × Chest.depth + 0.4900 × Wrist.diameter + 0.5195 × Knee.diameter + 0.0963 × Shoulder.girth + 0.1767 × Chest.girth + 0.3566 × Waist.girth + 0.2193 × Hip.girth + 0.2631 × Thigh.girth + 0.6449 × Forearm.girth + 0.1701 × Knee.girth + 0.3488 × Calf.maximum.girth - 0.4509 × Wrist.minimum.girth - 0.0573 × Age + 0.3073 × Height - 1.2544 × Gender

We can see that all predictors are statistically significant with the intercept being -120, which also has no real meaning. Model 2’s final adjusted R-squared value was 0.9747, with a F-statistic 1041 on 15 and 390 DF. Model 2’s final variables includes Chest.depth , Wrist.diameter , Knee.diameter , Shoulder.girth , Chest.girth , Waist.girth , Hip.girth , Thigh.girth , Forearm.girth , Knee.girth , Calf.maximum.girth , Wrist.minimum.girth , Age , Height , Gender.

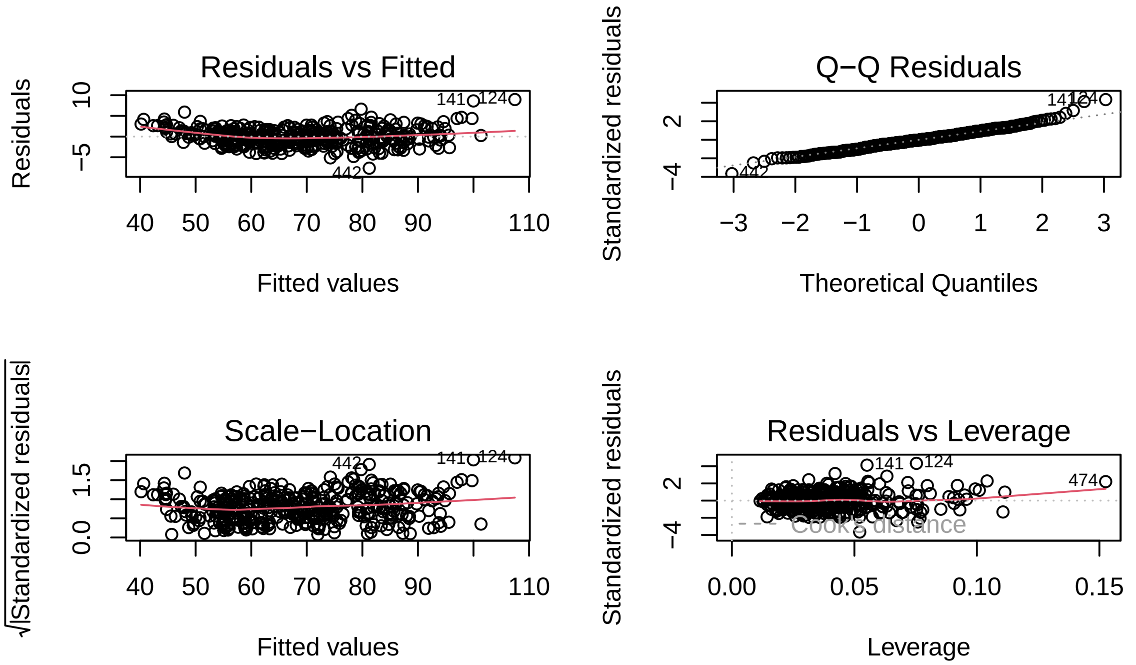
(c) (5 marks) Perform diagnostics checking for each of the final fitted models, Model 1 and Model 2 respectively.

For Model 1:



The diagnostic plots indicate the the residuals are not mean zero. The Q-Q plot The points follow the diagonal line fairly well indicating the residuals are approximately normally distributed. The scale-location plot suggests the variance may also not be constant. Finally, no influential points are detected.

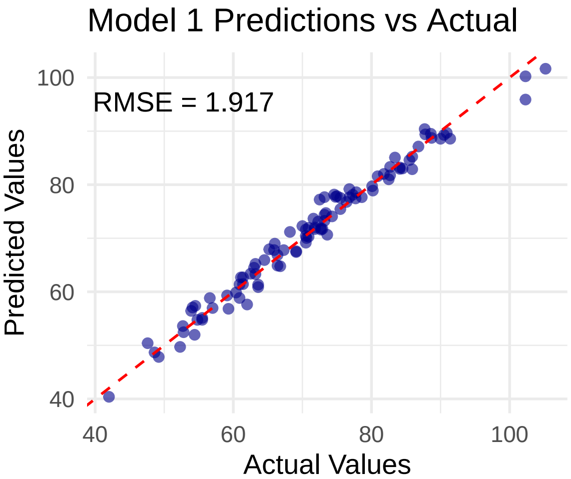
For Model 2:

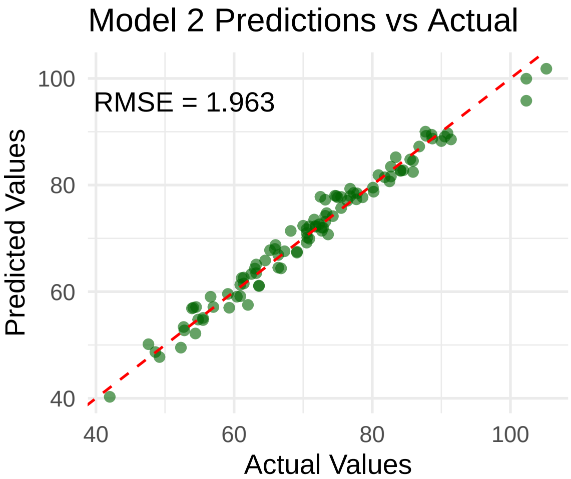


The diagnostic plots indicate the the residuals are not mean zero though slighlty better than Model 1. The Q-Q plot The points follow the diagonal line fairly well indicating the residuals are approximately normally distributed. The scale-location plot suggests the variance may also not be constant. And no influential points are detected.

(d) (6 marks). Despite any inadequacies that you may or may not have identified above, you use the two models obtained in (b) to make predictions of Weight in the test set.

(i) Produce a correctly drawn and labelled plot of predicted values against the actual values in the test set, and obtain the root mean squared error of prediction (RMSEP) based on each fitted model.


Model 1 Predictions vs Actual


Model 2 Predictions vs Actual

(ii) Using the RMSEPs and the plots you produced, comment on how well the models performed.

Both models appear to perform very similarly based on the visual plots and their RMSE (Root Mean Square Error of Prediction) values:

- Model 1 has an RMSE of 1.917
- Model 2 hsa a RMSE of 1.963

Given the small difference in RMSE values and the similar visual patterns, both models appear to be performing well, with Model 1 having a marginally better overall performance based on the slightly lower RMSE.