A Study on the Relationship between Height and Foot Size

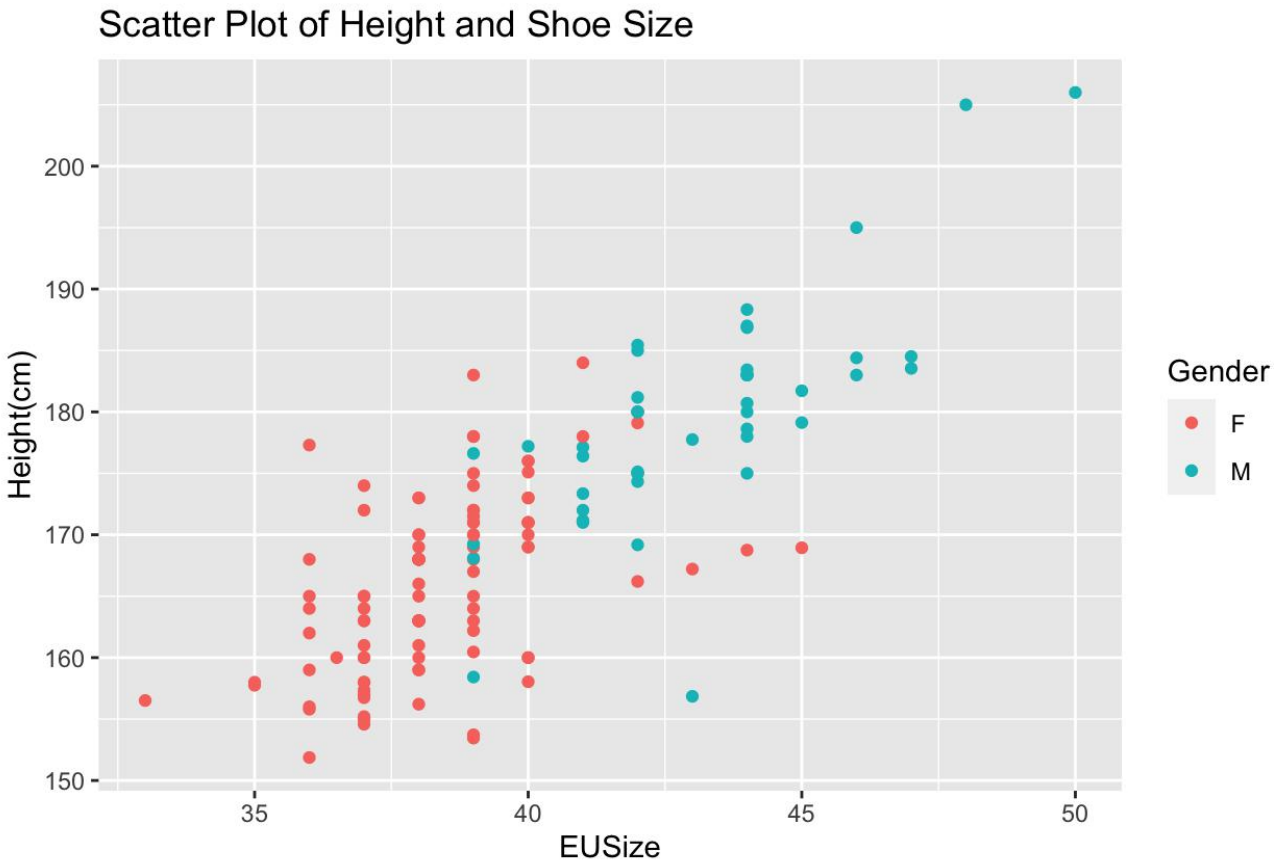
1) Data pre-processing: consolidate the two body height and shoe/foot size data files in one data file containing relevant attributes. The data may contain some imperfection that requires some data cleansing activities. Please describe these activities and provide necessary justifications and assumptions in the report.

The first step in the data analysis process is data pre-processing. In this step, the raw data is converted into a clean and structured format for further study.

After checking the summary of the foot01 data set, we found that there is one NA which is removed by na.omit function.
In the height, there are some data recorded by meter and centimeter.
Then the data in meters is multiplied by 100 to unified the standard of height.
Also there is a maximum data point for height is 364cm, according to the record the tallest person in the world is 272cm, then we remove the outlier which height is bigger than 272. We also set the column name that prepares for a combined data set later.

Since the second dataset is in pdf, we use pdftools to extract text from PDF, and use stringr packages which is a member of the tidyverse collection of R packages to manipulate strings of text. and create a neat dataset.
First, use pdt_text command to read the text of the files, create new objects pdfData, and read_lines() function to read the lines of the file.
We want to get the useful info, this is the lines 4 to 56 and 60 to 112 of our file(2 tables).
Row 4 &60 contains the column names of the data we generated, naming the data frame as data_lines1and data_lines2.
Use the complete.cases function to filter out all data rows in the dataframe that do not contain missing values and use str_split function to split the elements of each string into substrings.
After names using colnames and rownames, we use the rbind function to join these two tables together. We can see they are characters. Convert them to integer and numeric objects using as.integer and as.numeric.
For the file FLtoEUsize.csv, by checking variable type, we can see the foot.length is character.
Then we extract the number part from the foot.length, we change the variable foot length to numeric.
Using the table of foot length and EU size, we convert the Foot variable from the second data set into EU size.
Then we found that there are foot lengths that don't have a corresponding EU size, therefore we treat them as outliers and remove them.
By setting the same column name, then we can combine the two datasets together for further analysis.

2) What is the correlation between body height and foot size, and explain your results.



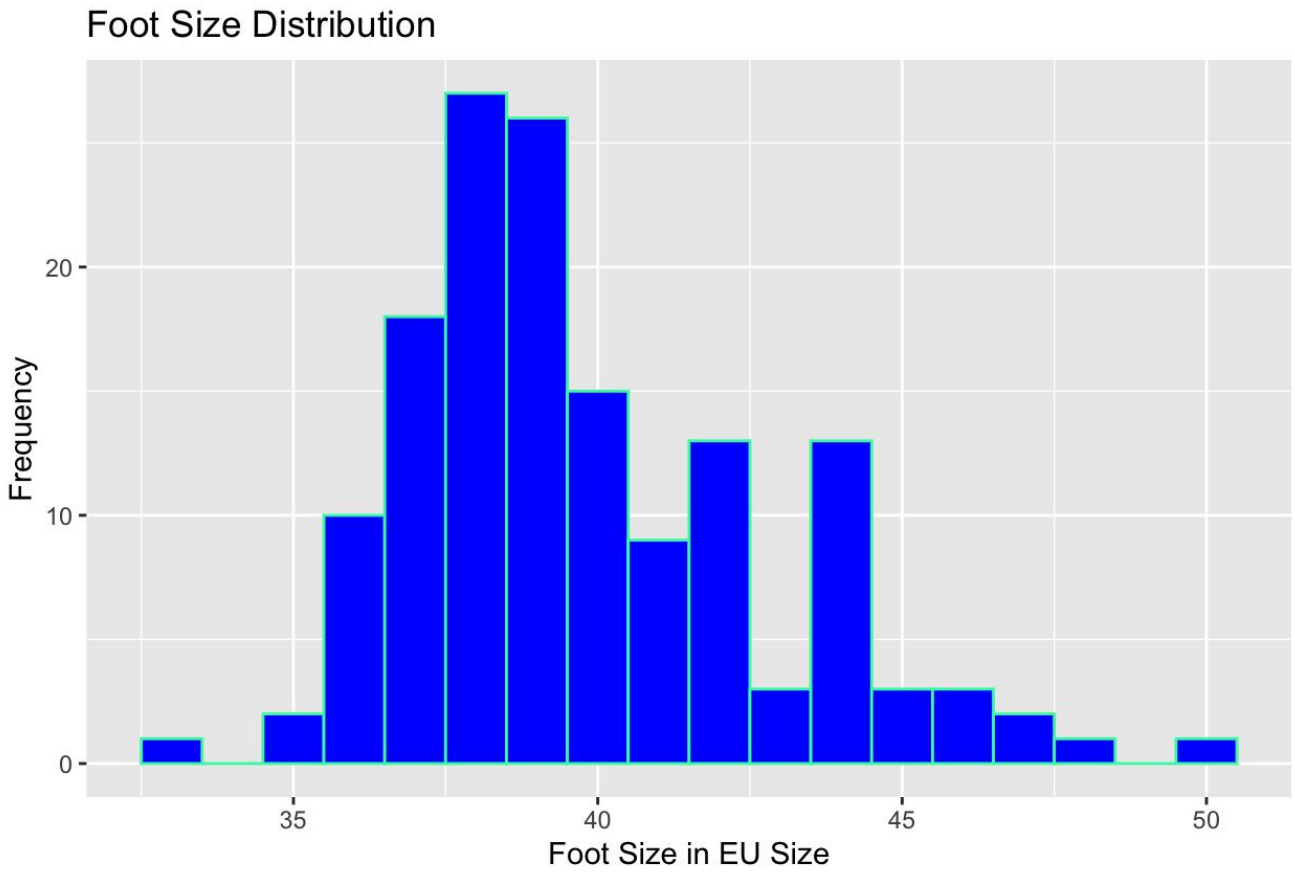
correlation between height and foot size

	Correlation Coefficient	95% confidence interval	Significance p-value
All	0.780	0.707-0.836	<2.2e-16
Female	0.503	0.342-0.636	<6.975e-8
Male	0.757	0.596-0.859	<1.807e-9

The height is considered as independent variable and shoe size as dependent variable. Through the scatter plot, we can see that there is a positive linear correlation between them.

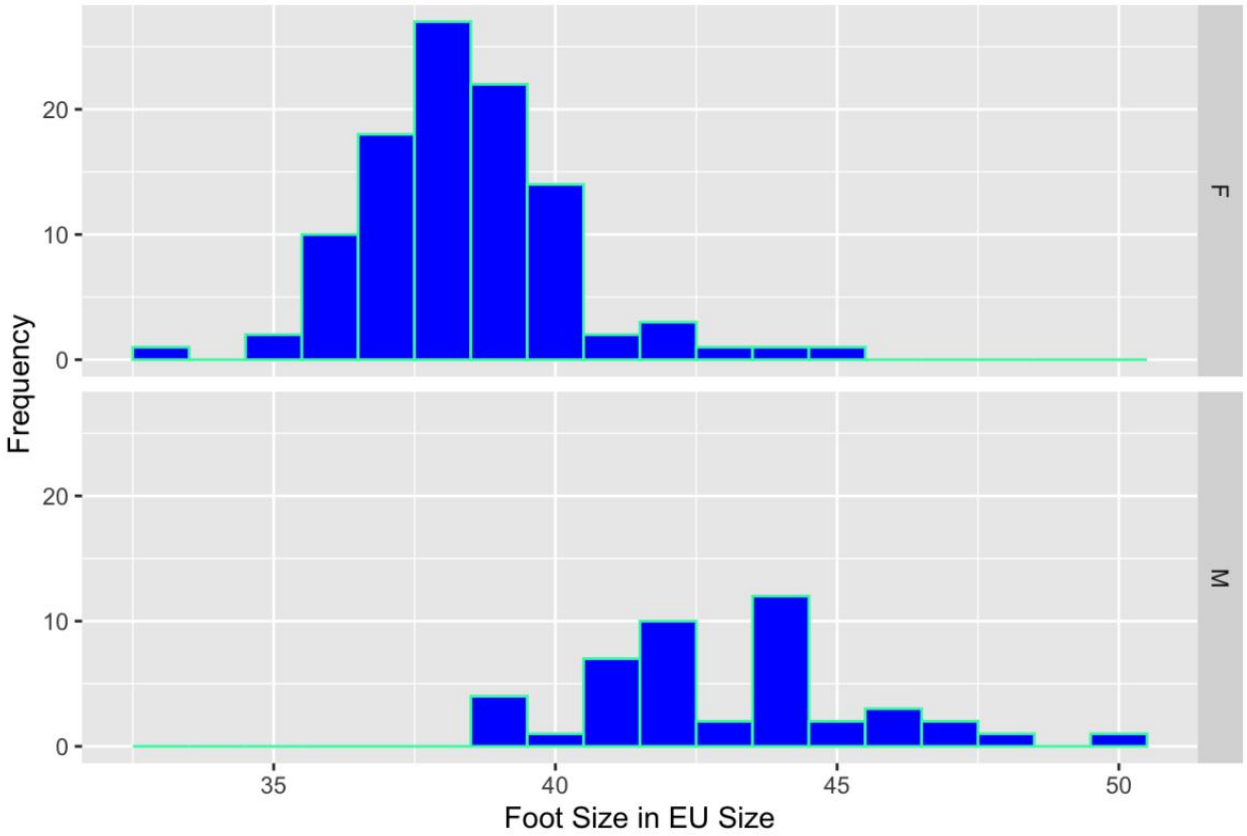
The correlation coefficient measures the strength and direction of the relationship between two variables. The above table indicates a positive relationship between height and foot size. As the $p<0.05$, the correlation is statistically significant.

3) Create a histogram based on foot size values.



4) Enhance the figure generated in 3)

i) Create a facet chart based on genders

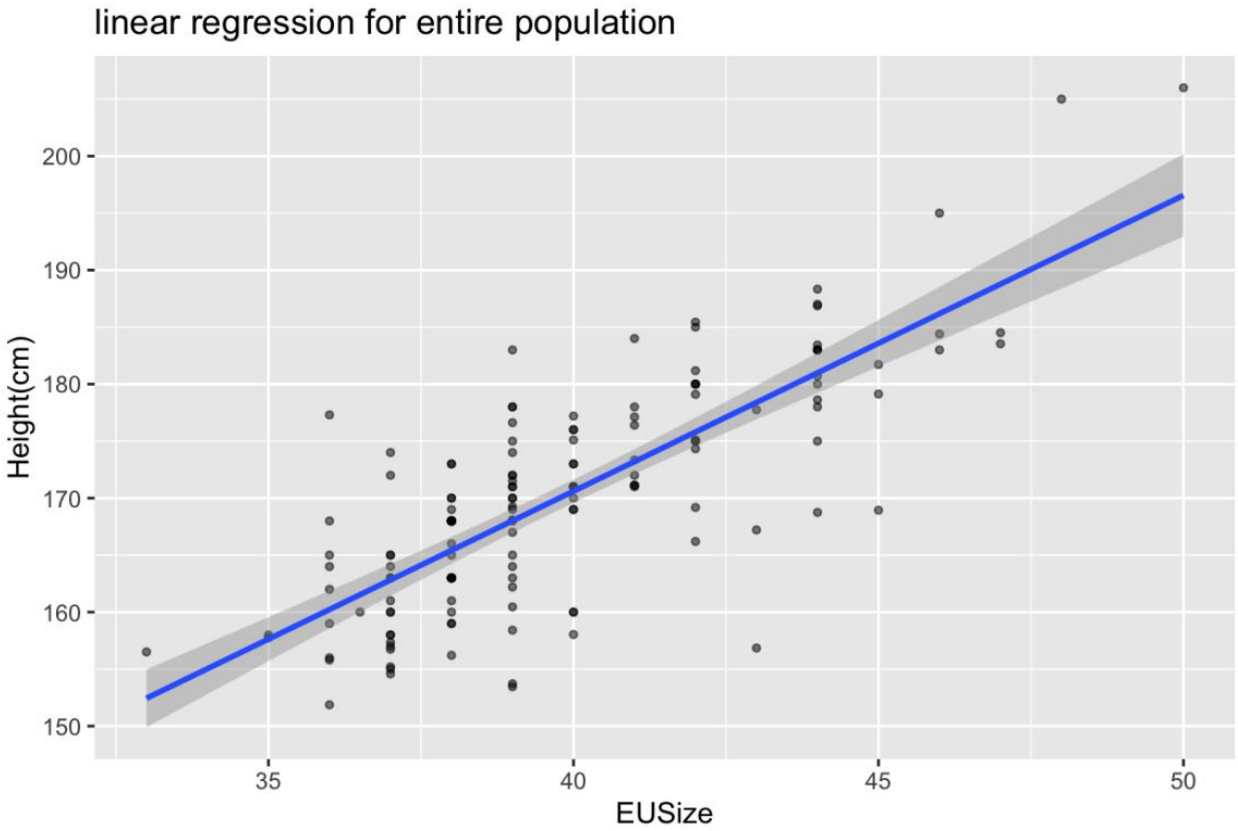


ii) **Provide descriptive descriptions and insights of the visualisations, not less than 200 words.**

From the plot, we can see that the distribution of female shoe size is approximately normal distribution. Male tends to have bigger shoe size than female on average.

5) Create linear regression models of human body heights and shoe sizes for the entire population, female population and male population respectively. Generate plots of the models over the samples. Justify comprehensively your answer using the model summaries.

Regression based on the entire population



Call:
lm(formula = footdata$`Height(cm)` ~ footdata$EUSize)

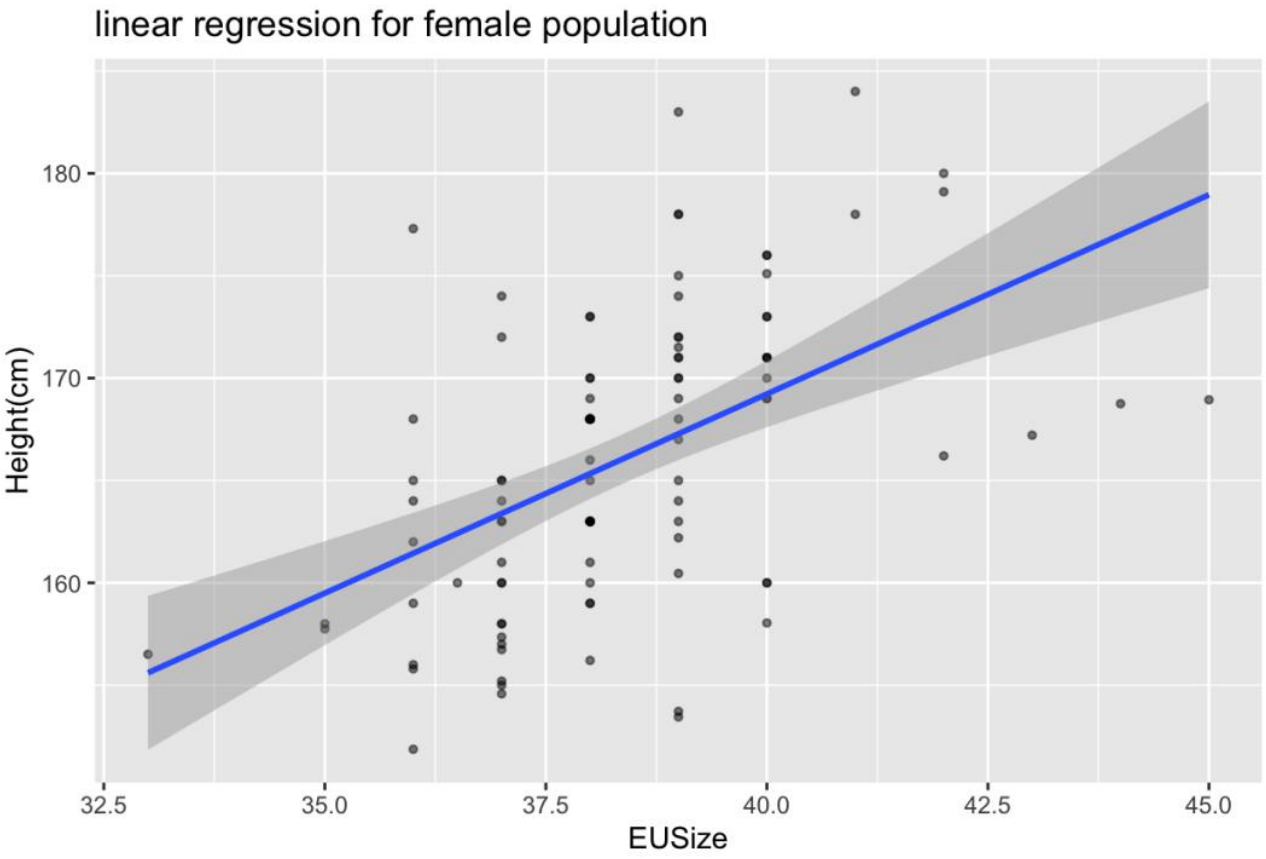
Residuals:
    Min   1Q Median   3Q   Max
-21.5423 -3.0949   0.1767   3.8444 17.0715

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   66.8149   6.9005   9.683   <2e-16 ***
footdata$EUSize   2.5948   0.1729 15.007   <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.201 on 145 degrees of freedom
Multiple R-squared: 0.6083, Adjusted R-squared: 0.6056
F-statistic: 225.2 on 1 and 145 DF, p-value: < 2.2e-16

Regression based on female population



Call:
lm(formula = datafemale$`Height(cm)` ~ datafemale$EUSize)

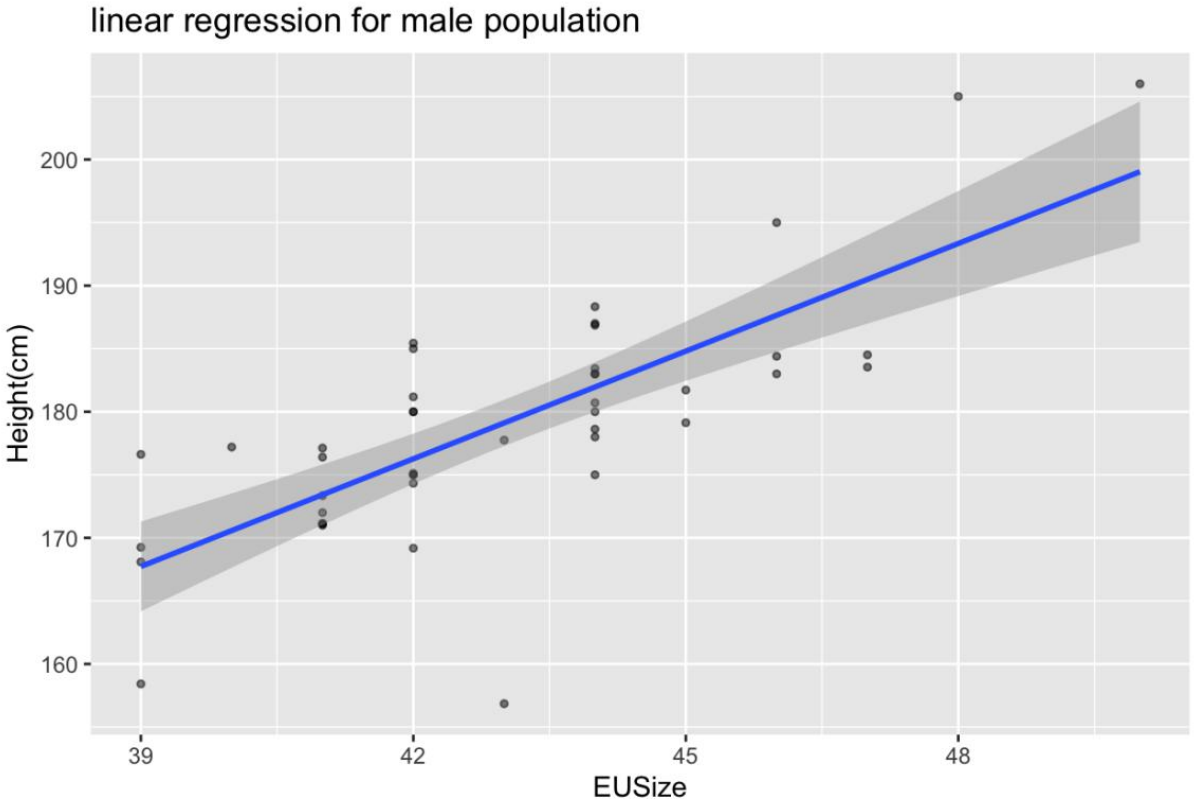
Residuals:
    Min   1Q Median   3Q   Max
-13.8292 -4.8928   0.6394   3.7208 15.8579

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   91.396   12.831   7.123 1.65e-10 ***
datafemale$EUSize 1.946   0.334   5.825 6.97e-08 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.154 on 100 degrees of freedom
Multiple R-squared: 0.2533, Adjusted R-squared: 0.2459
F-statistic: 33.93 on 1 and 100 DF, p-value: 6.975e-08

Regression based on male population



Call:

lm(for

mula =

datamale$`Height(cm)` ~ datamale$EUSize)

Residuals:

Min	1Q	Median	3Q	Max
−22.264	−3.085	−1.149	3.731	11.659

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	56.7640	16.1579	3.513	0.00106	**
datamale$EUSize	2.8453	0.3748	7.592	1.81e−09	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.089 on 43 degrees of freedom

Multiple R−squared: 0.5727, Adjusted R−squared: 0.5628

F-statistic: 57.64 on 1 and 43 DF, p-value: 1.807e-09

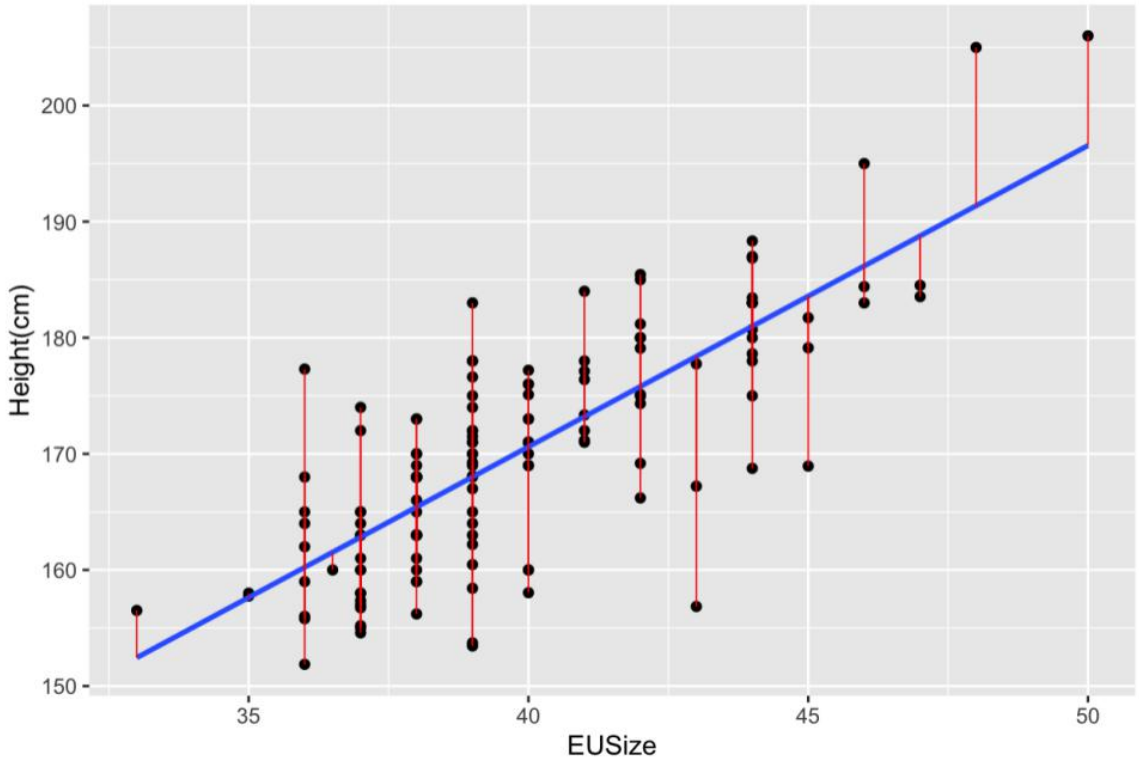
6) Based on the results from 5) above, analyse the residuals to determine if the assumptions underlying your regression analysis are valid. You need to provide a visualisation for this purpose and justify your answer.

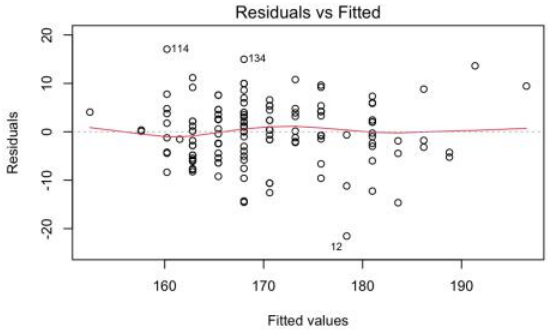
The key assumptions of the regression including:

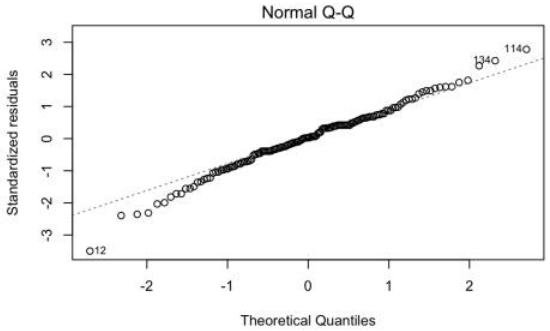
1. Linearity of the data
2. Normality of residuals
3. Homogeneity of residual variance

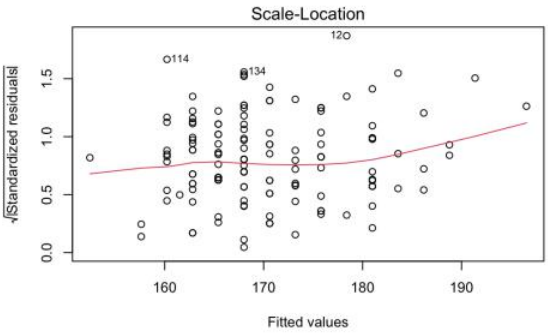
4.

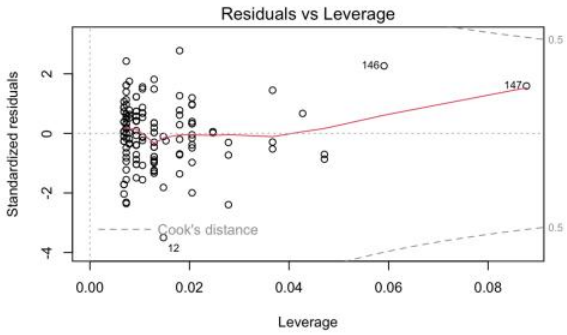
Independence











The above diagnostic plots shows:

1. Residuals vs Fitted

It is used to check linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship, in our case, it is good.

2. Normal Q–Q

It is used to examine whether the residuals are normally distributed. It's good if residual points follow the straight dashed line. In our case, it is good.

3. Scale–Location

It is used to check the homogeneity of variance of the residuals. Horizontal line with equally spread points is a good indication. However, it is not the case in our question, where we have a heteroscedasticity problem where points are at widely varying distances from the regression line. (##not too sure about this part, do you think the residual is randomly spread?)

4. Residuals vs Leverage

It is used to identify influential cases, which are extreme values that may influence the regression results when included or excluded from the analysis. The top 3 extreme data points(#12, #146, #147) are labeled in the plot. #12 has a standardized residual below –3 which indicates it may be an outlier. #146 and #147 have a high leverage compared to other data points.