Pair assessment INFO634

Michelle Visscher 84632465 and Kenneth Rosal 54668165

08 September 2022

Question 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.

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
## Rows: 101 Columns: 4
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): time, sex
## dbl (2): height, shoe_size
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Next we need to clean the data frame: 1. remove N/A values in the PDF file 2. rename column names 3. remove student column and time column (unnecessary)

Cleaning the Data:

Removing any variables with a height less than 20cm. Removing row 68 - outlier with height 364.00, shoe size 88. We are removing these outliers because these are assumed to be a data entry errors.

```
##     gender height shoe_size
## 1    woman 160.00      40.0
## 2    woman 171.00      39.0
## 3    woman 174.00      39.0
## 4    woman 176.00      40.0
## 5      man 195.00      46.0
## 6    woman 157.00      37.0
## 7    woman 160.00      38.0
## 8    woman 178.00      39.0
## 9    woman 168.00      38.0
## 10     man 171.00      41.0
## 11   woman 165.00      39.0
## 12     man 175.00      44.0
## 13   woman 163.00      38.0
## 14   woman 158.00      37.0
## 15   woman 159.00      38.0
## 16     man 183.00      44.0
## 17   woman 155.00      37.0
## 18   woman 172.00      39.0
## 19   woman 164.00      39.0
## 20   woman 158.00      35.0
## 21   woman 174.00      37.0
## 22   woman 164.00      37.0
## 23   woman 168.00      38.0
## 24   woman 168.00      38.0
## 25   woman 163.00      37.0
## 26   woman 160.00      37.0
## 27     man 183.00      46.0
## 28   woman 161.00      38.0
## 29   woman 162.00      36.0
## 30   woman 165.00      37.0
## 31   woman 164.00      36.0
## 32   woman 161.00      37.0
## 33   woman 163.00      39.0
## 34   woman 169.00      40.0
## 35   woman 171.00      39.0
## 36   woman 163.00      38.0
## 37   woman 159.00      36.0
## 38   woman 180.00      42.0
## 39   woman 168.00      38.0
## 40   woman 170.00      38.0
## 41   woman 168.00      38.0
## 42     man 180.00      42.0
## 43     man 183.00      44.0
## 44   woman 170.00      40.0
## 45   woman 172.00      39.0
## 46   woman 163.00      38.0
## 47   woman 168.00      38.0
## 48   woman   1.84      41.0
## 49   woman 169.00      38.0
## 50     man 206.00      50.0
## 51   woman 165.00      38.0
## 52   woman 171.00      40.0
## 53   woman 165.00      37.0
## 54   woman 168.00      38.0
## 55     man 180.00      44.0
## 56   woman 160.00      40.0
## 57     man 183.00      44.0
## 58   woman 160.00      36.5
## 59   woman 171.00      40.0
## 60   woman 167.00      39.0
## 61   woman 172.00      37.0
## 62   woman   1.63      38.0
## 63   woman 173.00      38.0
## 64     man 187.00      44.0
## 65   woman 176.00      40.0
## 66     man 180.00      42.0
## 67   woman 171.50      39.0
## 68   woman 364.00      88.0
## 69   woman 168.00      36.0
## 70   woman 175.00      39.0
## 71     man 185.00      42.0
## 72     man 205.00      48.0
## 73   woman 165.00      36.0
## 74     man 175.00      42.0
## 75     man 175.00      42.0
## 76     man 172.00      41.0
## 77   woman 156.00      36.0
## 78   woman   1.68      38.0
## 79   woman 163.00      37.0
## 80   woman 163.00      38.0
## 81   woman   1.73      38.0
## 82   woman 169.00      39.0
## 83   woman 178.00      39.0
## 84   woman 170.00      38.0
## 85   woman 168.00      38.0
## 86    <NA>     NA        NA
## 87   woman 170.00      39.0
## 88   woman 173.00      40.0
## 89   woman 171.00      40.0
## 90   woman 163.00      38.0
## 91   woman 166.00      38.0
## 92   woman 159.00      38.0
## 93   woman 178.00      41.0
## 94     man 178.00      44.0
## 95   woman 169.00      40.0
## 96   woman 158.00      37.0
## 97   woman 170.00      39.0
## 98   woman 183.00      39.0
## 99   woman 173.00      40.0
## 100  woman 160.00      37.0
## 101  woman 168.00      39.0
##     gender              height        shoe_size    
##  Length:96          Min.   :155.0   Min.   :35.00  
##  Class :character   1st Qu.:163.0   1st Qu.:38.00  
##  Mode  :character   Median :169.0   Median :39.00  
##                     Mean   :170.0   Mean   :39.31  
##                     3rd Qu.:174.5   3rd Qu.:40.00  
##                     Max.   :206.0   Max.   :50.00  
##                     NA's   :1       NA's   :1
```

Converting shoe size to foot length for dataset footlength1 We used the merge function to merge the datasets for FL to EU size and Foot length 1. During this process, some variables were excluded from the data set as the foot length 1 data set included shoe sizes not included in the official FL to EU size conversion. Due to this, the cases with a shoe size greater than 47 and any half shoe sizes were excluded from the data set.

```
## Rows: 32 Columns: 2
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): foot length
## dbl (1): EU Size
## 
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
##     EU Size      foot length       
##  Min.   :16.00   Length:32         
##  1st Qu.:23.75   Class :character  
##  Median :31.50   Mode  :character  
##  Mean   :31.50                     
##  3rd Qu.:39.25                     
##  Max.   :47.00
```

```
##     gender              height        shoe_size    
##  Length:96          Min.   :155.0   Min.   :35.00  
##  Class :character   1st Qu.:163.0   1st Qu.:38.00  
##  Mode  :character   Median :169.0   Median :39.00  
##                     Mean   :170.0   Mean   :39.31  
##                     3rd Qu.:174.5   3rd Qu.:40.00  
##                     Max.   :206.0   Max.   :50.00  
##                     NA's   :1       NA's   :1
```

cleaning the merged data set:

Merging the two data sets footlength2.1 and mergeddataFL1

Question 2:

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

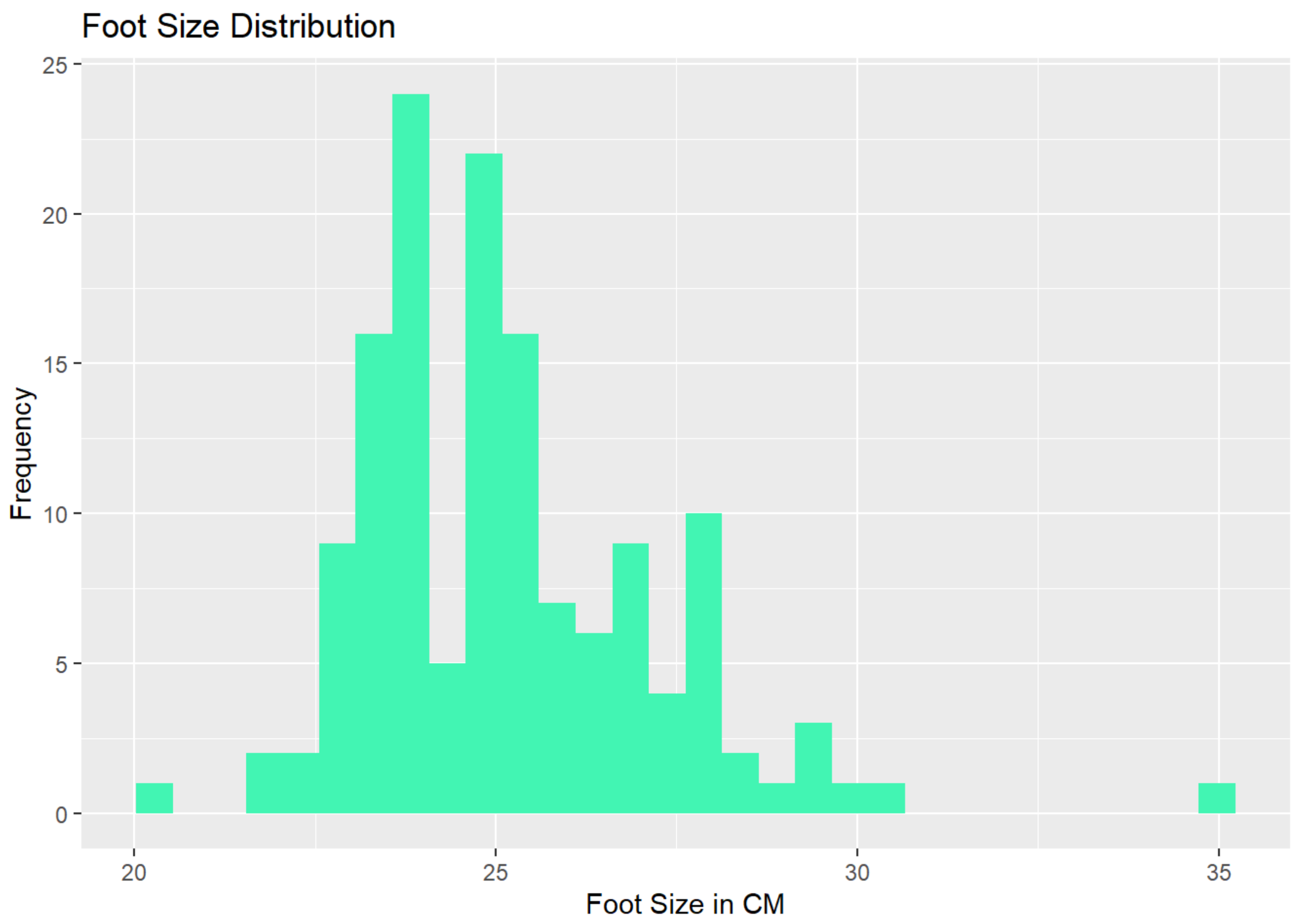
```
## [1] 0.6969386
```

```
##
##  Pearson's product-moment correlation
##
## data:  footlength_height_comb$height and footlength_height_comb$foot_length
## t = 11.499, df = 140, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6012361 0.7729298
## sample estimates:
##       cor
## 0.6969386
```

From the correlation coefficient test result of 0.693 (3 dp), there is a evidence to suggest there is a positive relationship between the variables height and foot length for the sample set provided. This suggests that as body height increases so too does foot size.

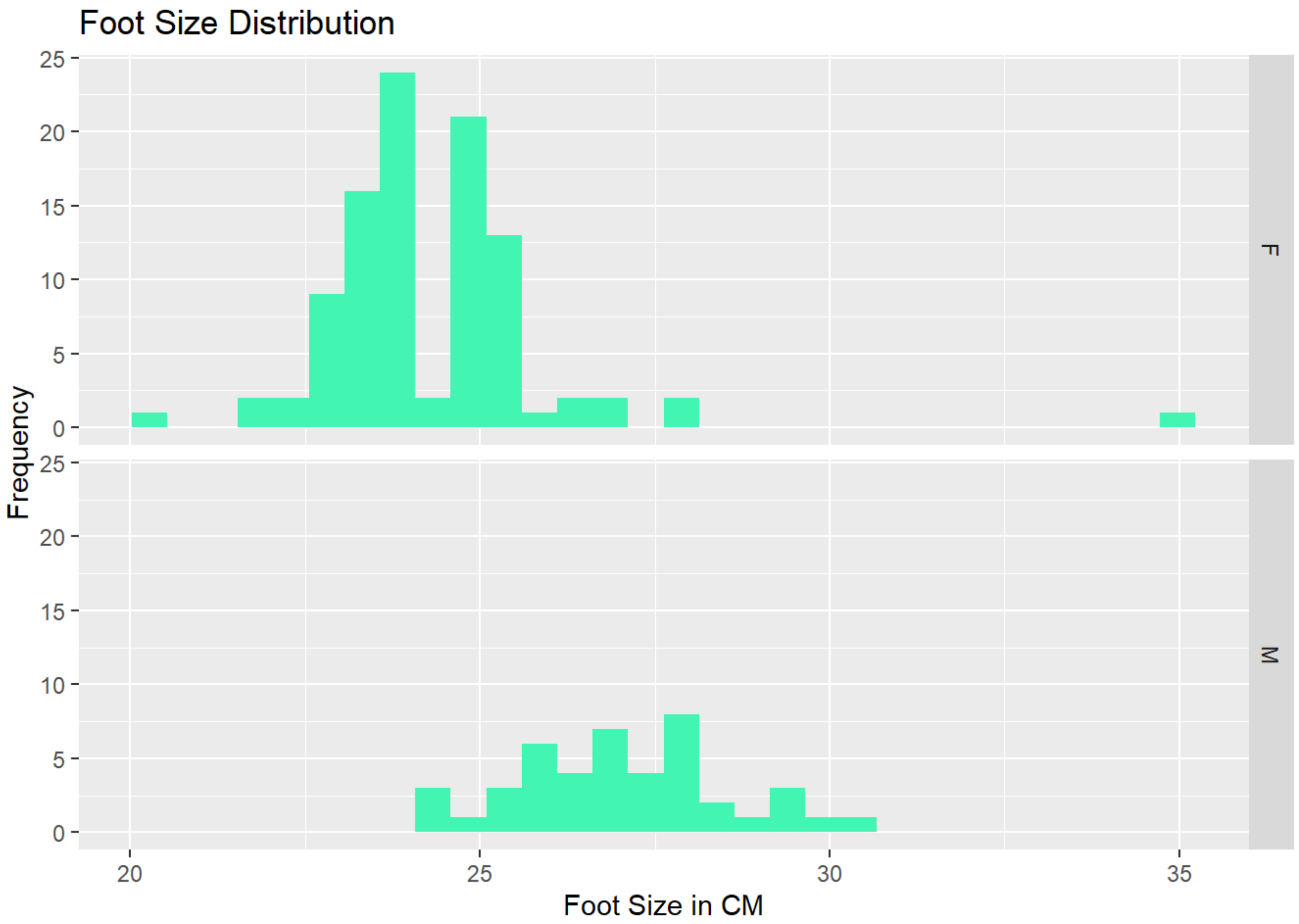
Question 3:

Create a histogram based on foot size values. Requirements: i) Figure title: “Foot Size Distribution” ii) X-axis is labelled with “Foot Size in CM” or “Foot Size in EU Size” iii) Y-axis is labelled with “Frequency” iv) The bins should be coloured with hex colour code “#42f5b3”



Question 4:

Enhance the figure generated in 3) i) Create a facet chart based on genders



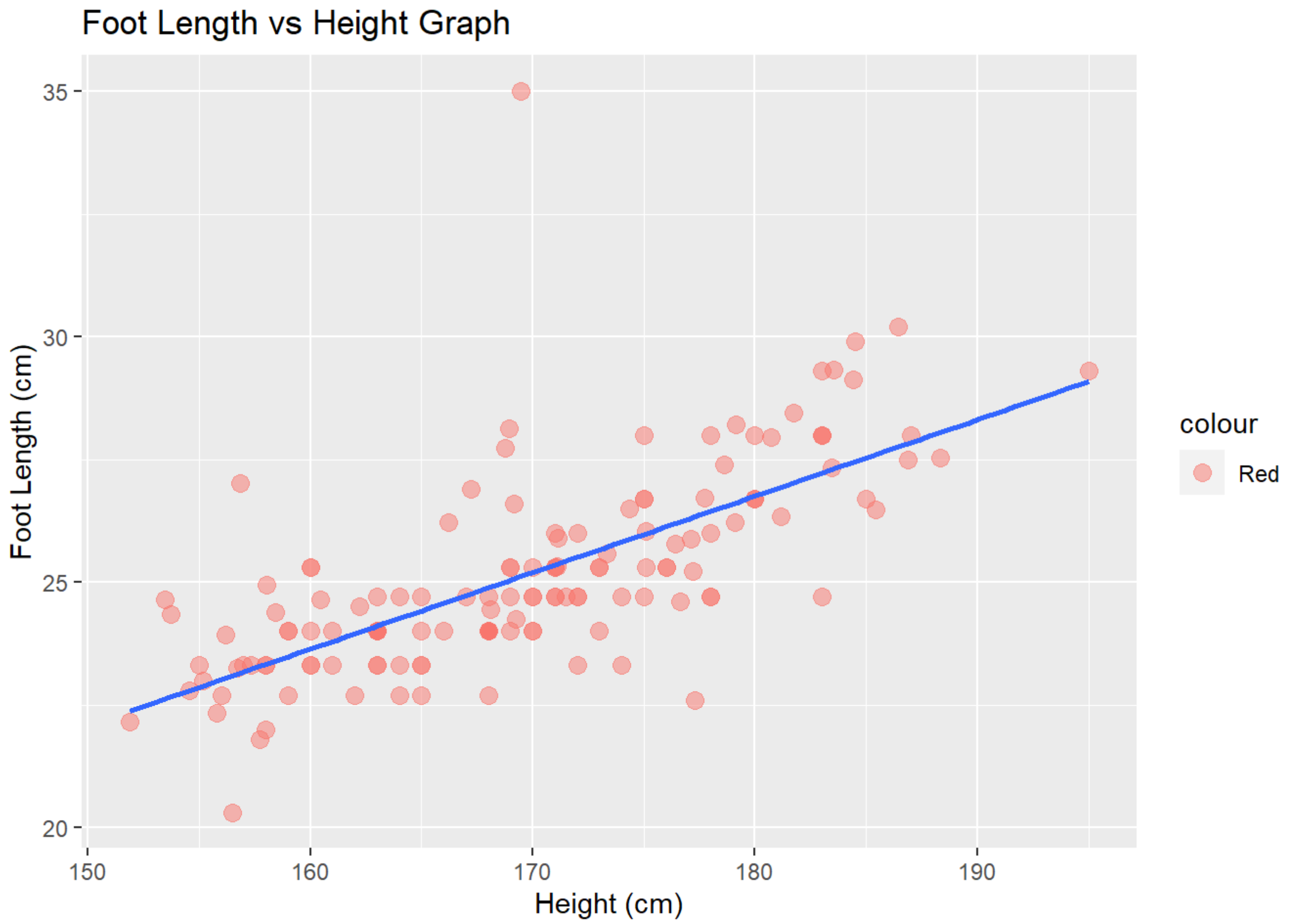
ii. Provide descriptive descriptions and insights of the visualisations, not less than 200 words.

The graph above of foot size distribution as faceted by gender shows a general trend of women (F) having a smaller foot size than men (M). The range of foot sizes is larger in women than in men, notably the largest smallest foot size in women is 20.31cm, while the largest foot size in women is 35.01cm. This data point of a foot size of 35.01cm for a woman looks to be an outlier, and it may be the case that this data point is incorrectly measured. The range in foot sizes for men is less than the range for women, with the minimum foot size for men being 24.25cm and the largest foot size at 30.21. The overall distribution for foot size distribution in men is approximately normally distributed, whereas the overall distribution for foot size in women is approximately right skewed (inclusive of the outlier). We suspect that by excluding the outlier in the women data set, we would see a more normally distributed sample. We also note that there is a larger sample size for women than for men, with 44 samples taken for men and 97 samples taken for women. If we have a larger data set of samples, we expect our data set to be more equally (normally) distributed.

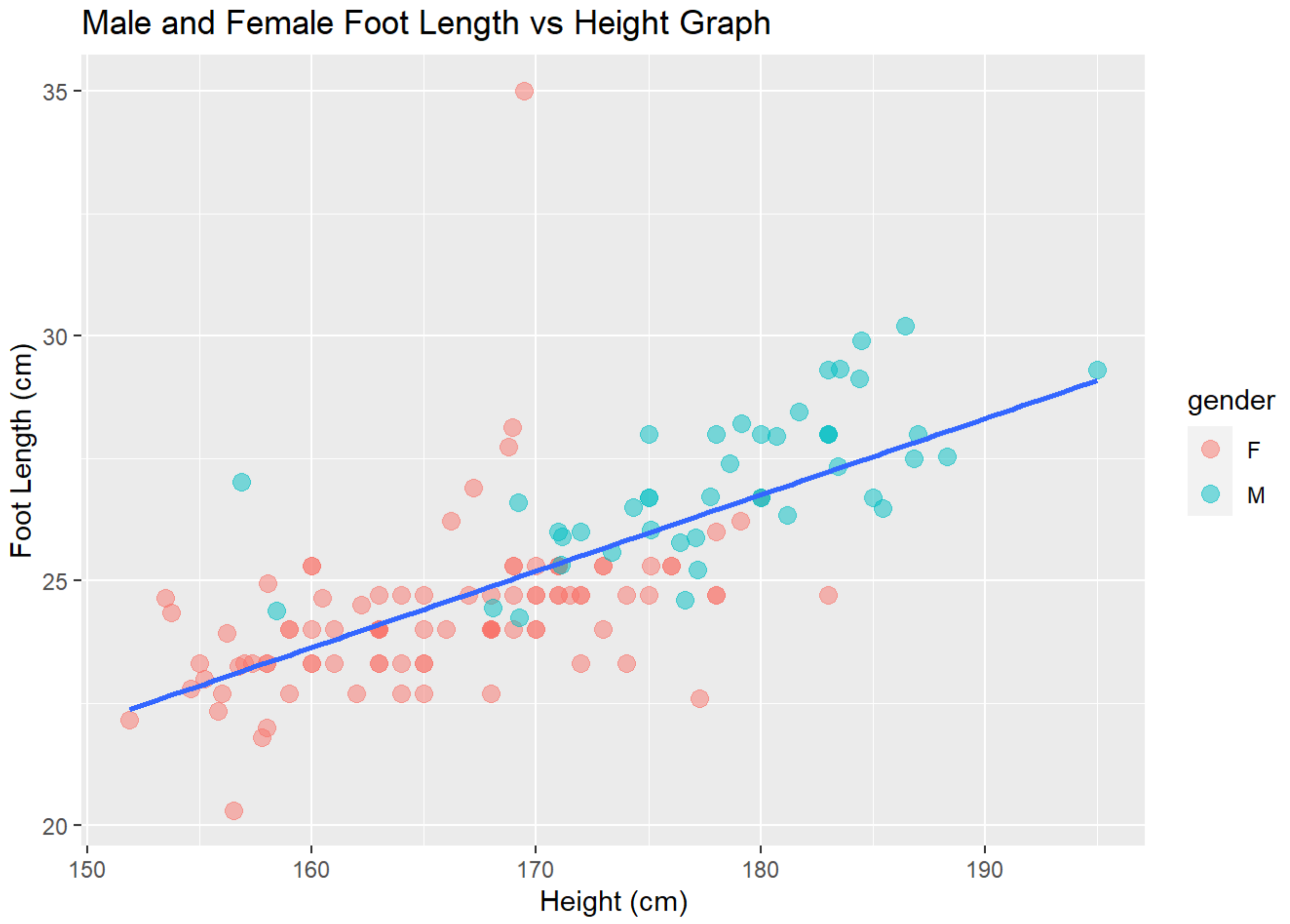
Question 5:

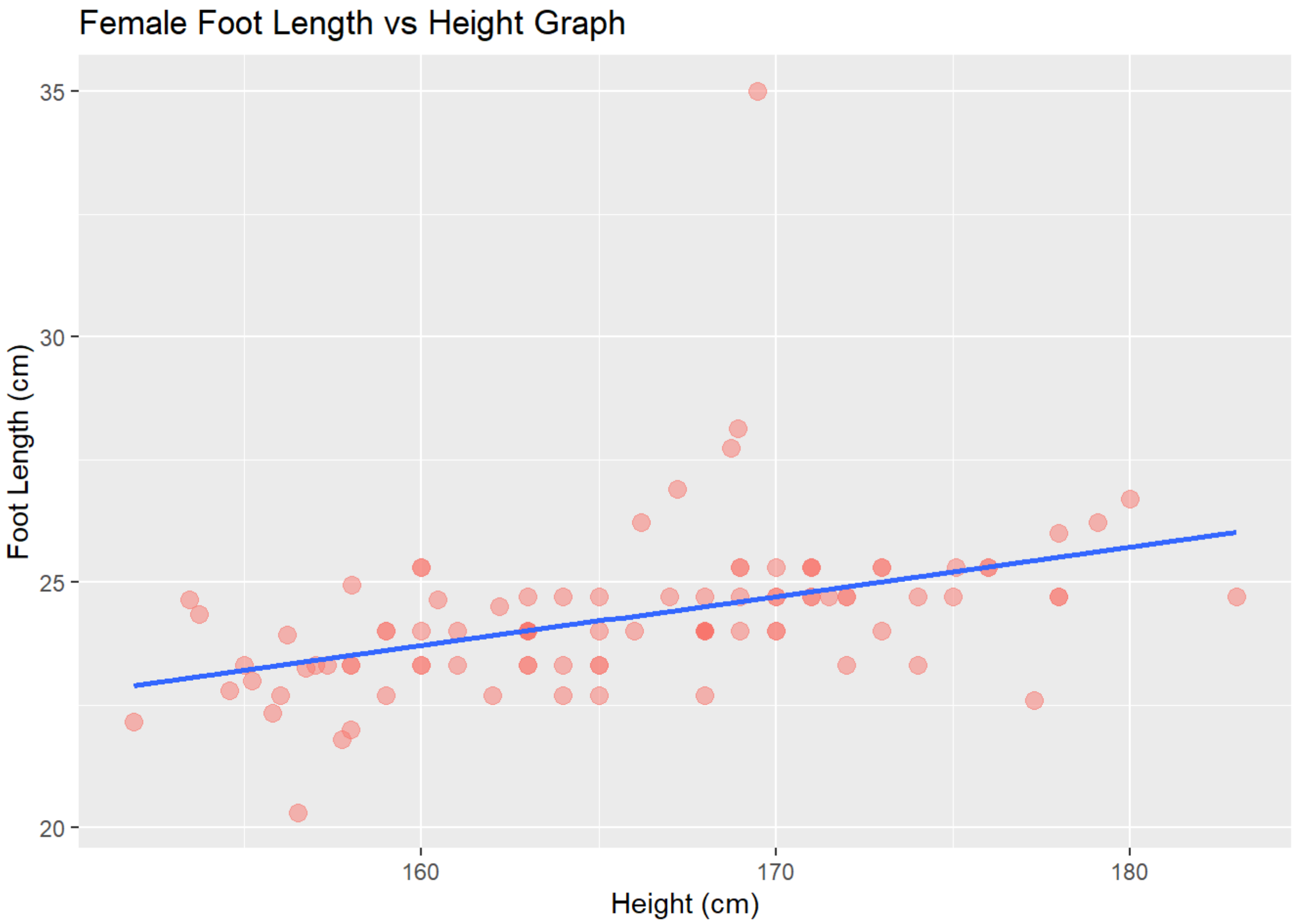
Create linear regression models of human body heights and shoe sizes for the entire population, female population and male population respectively. Generate plots of the models over the samples. Justify comprehensively your answer using the model summaries.

Linear regression model of human body heights and foot sizes for the entire population:

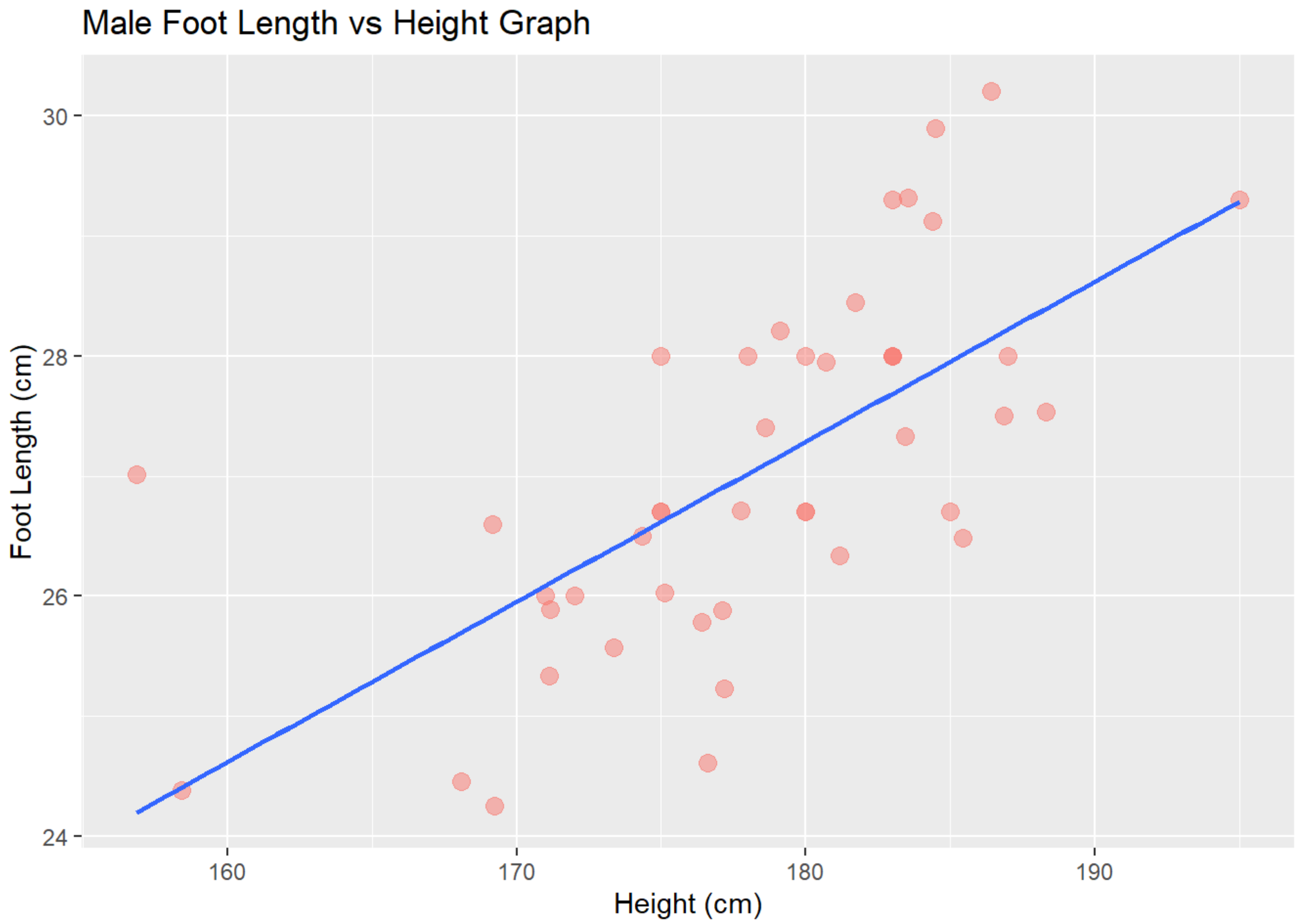


Plotting linear regression models of human body heights and shoe sizes for the entire population, color separated into female sample population and male sample population respectively.





```
##
## Call:
## lm(formula = female$height ~ female$foot_length)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9741  -4.0788   0.5046   4.1851  16.3657
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        121.5471     9.5687  12.703  < 2e-16 ***
## female$foot_length   1.8254     0.3929   4.646 1.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.297 on 96 degrees of freedom
## Multiple R-squared:  0.1836, Adjusted R-squared:  0.1751
## F-statistic: 21.58 on 1 and 96 DF,  p-value: 1.077e-05
```



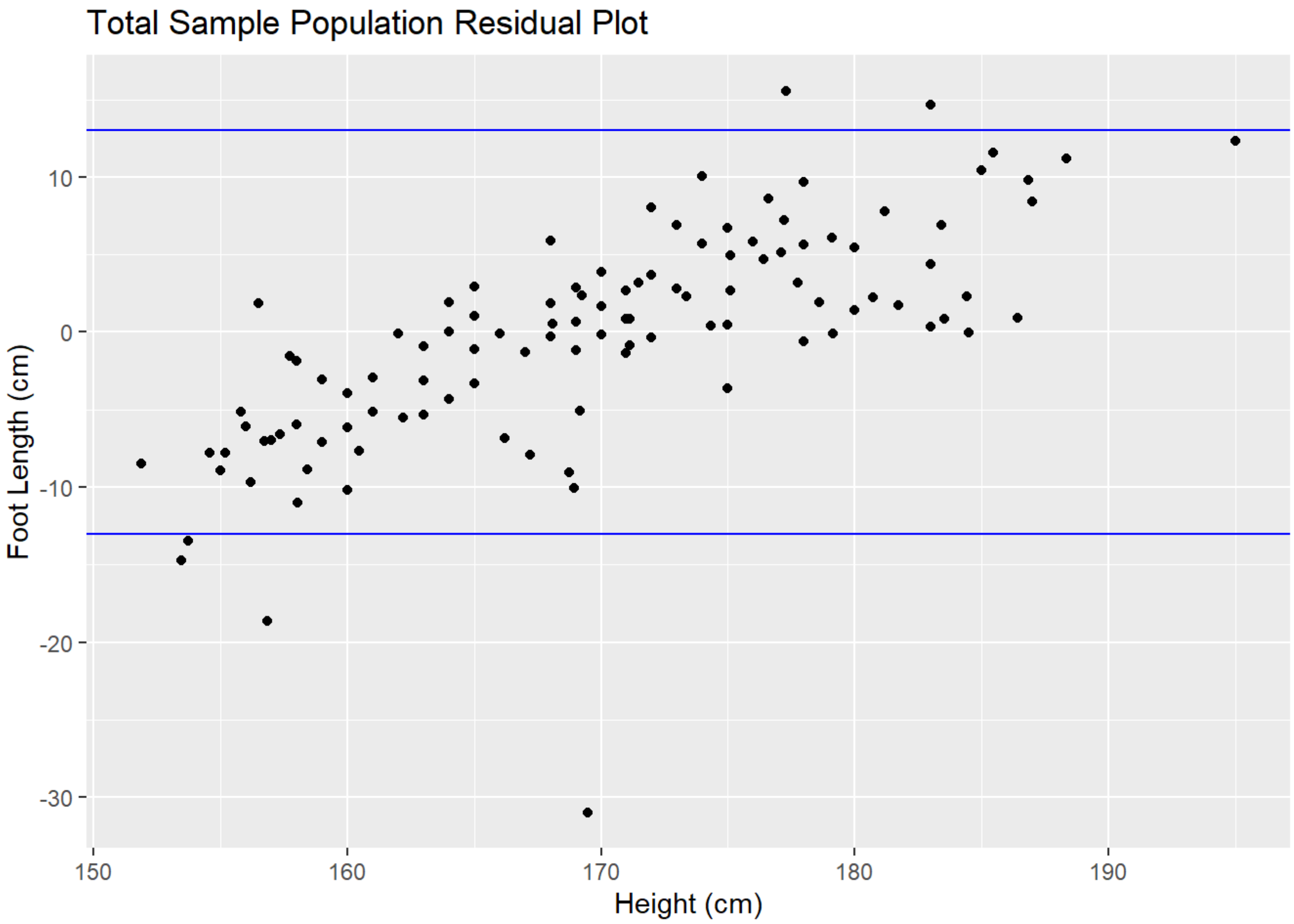
```
##
## Call:
## lm(formula = male$height ~ male$foot_length)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2378  -2.4024  -0.6052   2.9434   9.2952
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        88.2470    15.5617   5.671 1.18e-06 ***
## male$foot_length    3.3262     0.5745   5.790 7.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.585 on 42 degrees of freedom
## Multiple R-squared:  0.4438, Adjusted R-squared:  0.4306
## F-statistic: 33.52 on 1 and 42 DF,  p-value: 7.972e-07
```

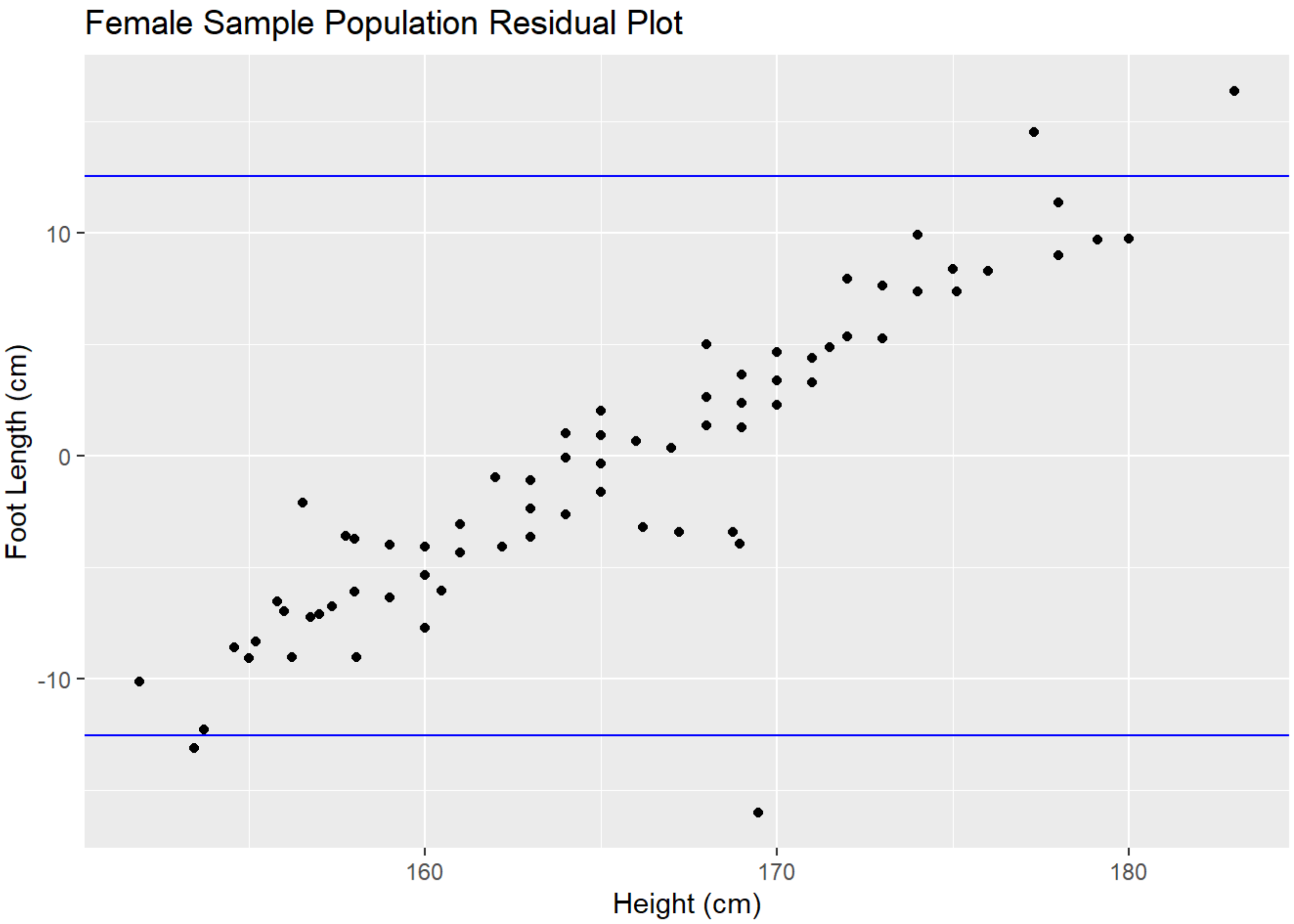
Based on the above linear regression models for body height and foot size in the sample population, we reject the null hypothesis that there is no relationship between the variables body height and foot size within our sample population. The p value of < 2.2e-16 being below alpha = 0.05, suggests our finding is statistically significant. This is accompanied with a large F- statistic of 128.2 on 1 and 139 DF. The same holds true when we look at body heights and foot size based on gender. For female body height and foot size, we can also see a statistically significant relationship between the two variables, with a p value of 1.957e-05. For male body height and foot size, our p value of 7.972e-07 is also low, and statistically significant. Our sample population model suggests that for a one unit increase in height there is a 3.107 increase in foot size. Height explains 48% of the variability in foot size.

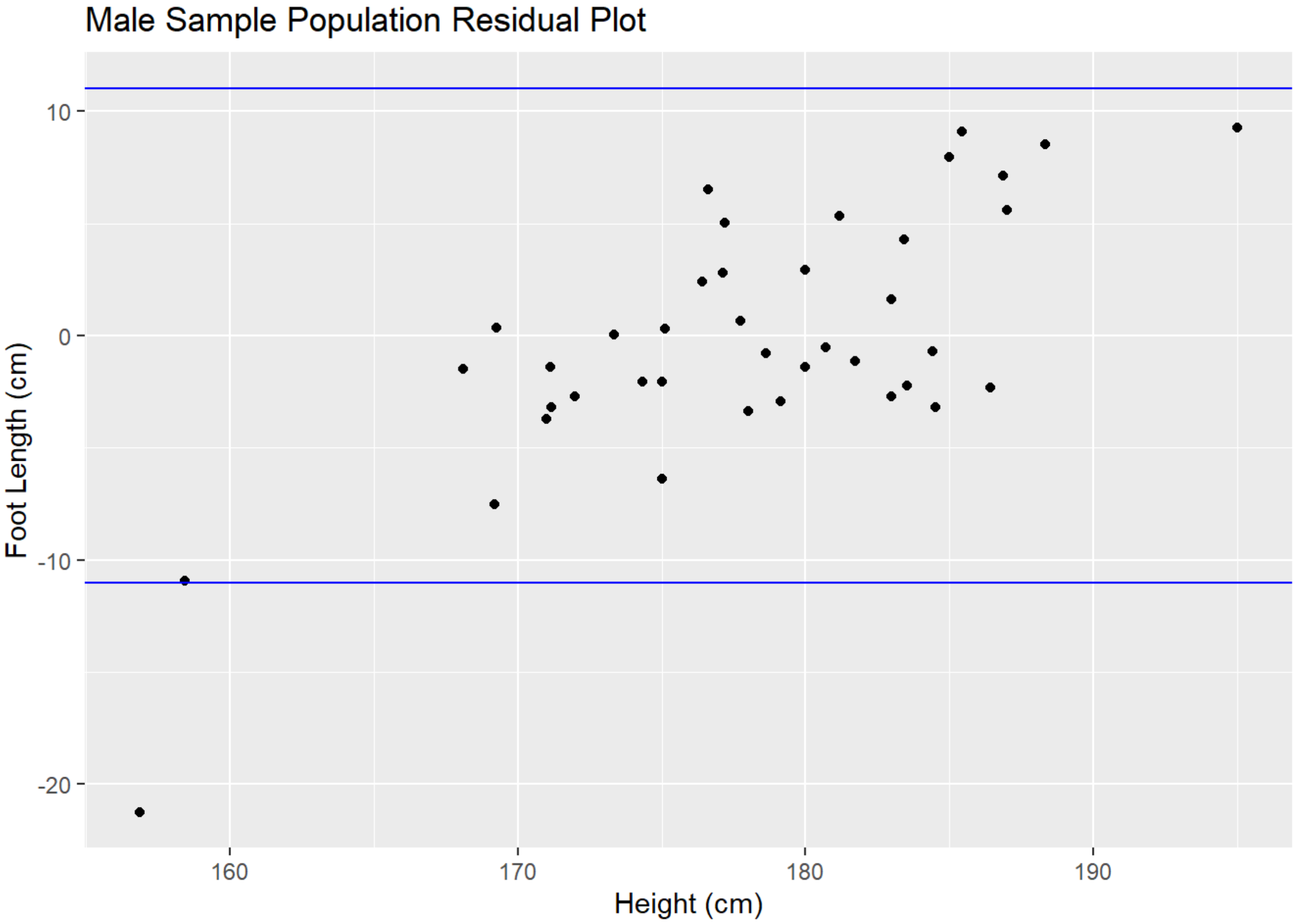
Question 6:

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.

Plotting the residuals:







Based on our analysis of the residuals, the majority of the data of the sample population does fall between +-2 Standard Deviations of the mean, however the data does not meet the assumption of linearity in that not all the residuals are distributed within +- 2 Standard Deviations.

The female sample population has more outliers when compared with the male sample population as seen on the graphs above. However, the outlier residual for the female sample is smaller than the male outlier.