MATPMD1: Statistics for Data Science Assignment

1. Introduction

In this project, we will analyze the Athlete dataset to transform data into useful information for athletes. The analysis will help athletes make better decisions. The data provides information on athlete sex, sport type, red cell count, white cell count, Hematocrit, Hemoglobin, plasma ferritin concentration, body mass index, sum of skin folds, body fat percentage, lean body mass, weight, and height which describe the athletes physical markers. The Athlete dataset comprises data on 100 female and 102 male athletes from a total of 202.

The project will involve exploratory data analysis, hypothesis testing, and developing a linear regression model using R and install necessary packages like dplyr and ggplot2 for a user-friendly coding environment.

```
> #install required packages
  install.packages("xlsx")
  install.packages("dplyr")
  install.packages("ggplot")
  install.packages("nortest")

  # loading the packages
  library("readxl")
  libray("dplyr")
  library("nortest")
```

2. Data and Methods

We will perform exploratory data analysis on the Athlete data to understand how the variables in the given Athlete dataset are distributed and identify any problematic values before performing a hypothesis test. One of the first steps of any data analysis project is exploratory data analysis. This involves exploring a dataset in three steps:

- utilizing descriptive statistics to create a dataset summary.
- using charts to visualize a dataset.
- Finding the values that are missing.

  First, let us load and view the data. It is simple to import data into R. For Excel files, use the readxl package's read_excel() function. After importing the Athlete dataset, using the

head () function, we can view the dataset's first six rows:

```
> # Importing a Excel file
> Athlete_data <- read_excel("C:/Users/Hibat/Downloads/Athlete Data.xlsx")
> # Viewing the first six rows
> head(Athlete_data)
# A tibble: 6 × 13
  Sex    Sport  RCC   WCC   Hc    Hg    Ferr  BMI   SSF  `%Bfat`  LBM   Ht    Wt
  <chr>  <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
1 female BBall  3.96  7.5   37.5  12.3    60  20.6 109.     19.8  63.3  196.  78.9
2 female BBall  4.41  8.3   38.2  12.7    68  20.7 103.     21.3  58.6  190.  74.4
3 female BBall  4.14  5     36.4  11.6    21  21.9 105.     19.9  55.4  178.  69.1
4 female BBall  4.11  5.3   37.3  12.6    69  21.9 126.     23.7  57.2  185   74.9
5 female BBall  4.45  6.8   41.5  14      29  19.0  80.3    17.6  53.2  185.  64.6
6 female BBall  4.1   4.4   37.4  12.5    42  21.0  75.2    15.6  53.8  174   63.7
> |
```

To get the dimensions of the dataset we use the dim () function and see that the dataset has 202 rows and 13 columns.

```
> #view dimensions of the Athlete data
> dim(Athlete_data)
[1] 202  14
```

To Summarize the data, we use the summary () function to quickly summarize each variable in the dataset and get the basic statistical analysis and provide an overview of the data.

```
> # summary statistics
> summary(Athlete_data)
     Sex          Sport          RCC              WCC              Hc       
 female:100   Row    :37   Min.   :3.800   Min.   : 3.300   Min.   :35.90  
 male  :102   T400m  :29   1st Qu.:4.372   1st Qu.: 5.900   1st Qu.:40.60  
              BBall  :25   Median :4.755   Median : 6.850   Median :43.50  
              Netball:23   Mean   :4.719   Mean   : 7.109   Mean   :43.09  
              Swim   :22   3rd Qu.:5.030   3rd Qu.: 8.275   3rd Qu.:45.58  
              Field  :19   Max.   :6.720   Max.   :14.300   Max.   :59.70  
              (Other):47                                                  
       Hg              Ferr             BMI             SSF             %Bfat       
 Min.   :11.60   Min.   :  8.00   Min.   :16.75   Min.   : 28.00   Min.   : 5.630  
 1st Qu.:13.50   1st Qu.: 41.25   1st Qu.:21.08   1st Qu.: 43.85   1st Qu.: 8.545  
 Median :14.70   Median : 65.50   Median :22.72   Median : 58.60   Median :11.650  
 Mean   :14.57   Mean   : 76.88   Mean   :22.96   Mean   : 69.02   Mean   :13.507  
 3rd Qu.:15.57   3rd Qu.: 97.00   3rd Qu.:24.46   3rd Qu.: 90.35   3rd Qu.:18.080  
 Max.   :19.20   Max.   :234.00   Max.   :34.42   Max.   :200.80   Max.   :35.520  

      LBM              Ht              Wt        
 Min.   : 34.36   Min.   :148.9   Min.   : 37.80  
 1st Qu.: 54.67   1st Qu.:174.0   1st Qu.: 66.53  
 Median : 63.03   Median :179.7   Median : 74.40  
 Mean   : 64.87   Mean   :180.1   Mean   : 75.01  
 3rd Qu.: 74.75   3rd Qu.:186.2   3rd Qu.: 84.12  
 Max.   :106.00   Max.   :209.4   Max.   :123.20  
```

This table shows the summary statistics for the variables: Sex, Sport, RCC, WCC, Hc, Hg, Ferr, BMI, SSF, %Bfat, LBM, Ht, and Wt. The table has 202 rows, each representing a different observation or case. The table has 13 columns, each representing a different variable. The table contains numerical data: RCC, WCC, Hc, Hg, Ferr, BMI, SSF, %Bfat, LBM, Ht, and Wt. We can view the following information for each of the numerical variables:

Min: The lowest possible number.

First Qu: First quartile value (25th percentile).

Median: The value in the middle.

Mean: The average amount.

Third Qu: The third quartile's value, or the 75th percentile.

Max: The highest possible number.

We can see a frequency count for each value for the categorical variables (sex, sport) in the dataset. For the sex variable, as an illustration:

Female: 100 observations of this value have been made

Male: 102 instances of this value are found.

We need to convert Sex and Sport data type from a character to a categorical type using a factor and rename %Bfat to Bfat as below.

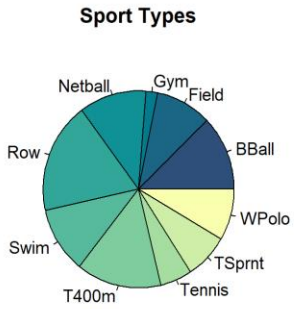
```
> # convert Sex and Sport data type from character to a factor vector
  Athlete_data$Sex<- as.factor(Athlete_data$Sex)
  Athlete_data$Spor<- as.factor(Athlete_data$Sport)

  # change %Bfat name
  names(Athlete_data)[names(Athlete_data)=="%Bfat"] <- "Bfat"
```

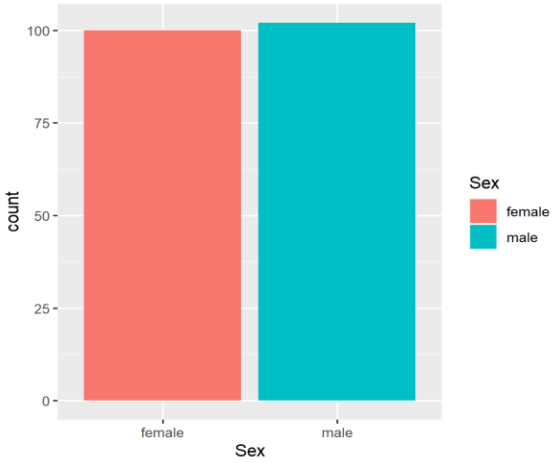
```
> contrasts(as.factor(Athlete_data$Sex))
       male
female    0
male      1
> contrasts(as.factor(Athlete_data$Sport))
        Field Gym Netball Row Swim T400m Tennis TSprnt WPolo
BBall       0   0       0   0    0     0      0      0     0
Field       1   0       0   0    0     0      0      0     0
Gym         0   1       0   0    0     0      0      0     0
Netball     0   0       1   0    0     0      0      0     0
Row         0   0       0   1    0     0      0      0     0
Swim        0   0       0   0    1     0      0      0     0
T400m       0   0       0   0    0     1      0      0     0
Tennis      0   0       0   0    0     0      1      0     0
TSprnt      0   0       0   0    0     0      0      1     0
WPolo       0   0       0   0    0     0      0      0     1
>
```

To depict the gender and sport of the athletes, we created a bar chart to show the number of athletes per gender and a pie chart to show the sports that the athletes participate in.

```
> # creat bar chart of sex
> ggplot(Athlete_data, aes(x= Sex, fill = Sex)) + geom_bar()
>
```

```
> # Create a table from the sport
>
> sport_n <- table(Athlete_data$Sport)
> sport_n

  BBall   Field     Gym Netball     Row    Swim   T400m  Tennis  TSprnt   WPolo
    25      19       4      23      37      22      29      11      15      17
> # Pie
>
> pie(sport_n,main = "Sport Types",
+     col = hcl.colors(length(sport_n), "BluYl"))
>
```

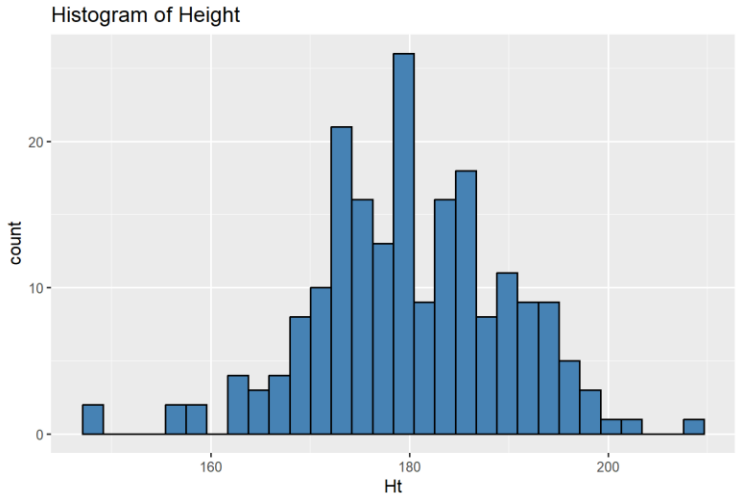




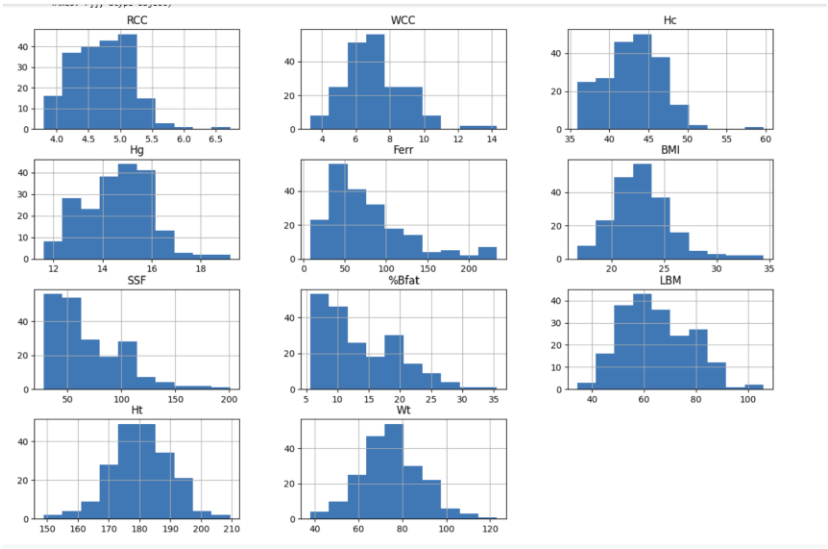
The bar graph indicates that the number of athletes in each gender is balanced. In the pie chart, the percentage of athletes who participate in ten different sports is shown.

To visualize the distribution of our variables, geom_histogram() function is utilized to generate a histogram based on the height values of the athletes.

```
> #create histogram of values for height
> ggplot(data= Athlete_data, aes(x=Ht)) +
+     geom_histogram(fill="steelblue", color="black") +
+     ggtitle("Histogram of Height")
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

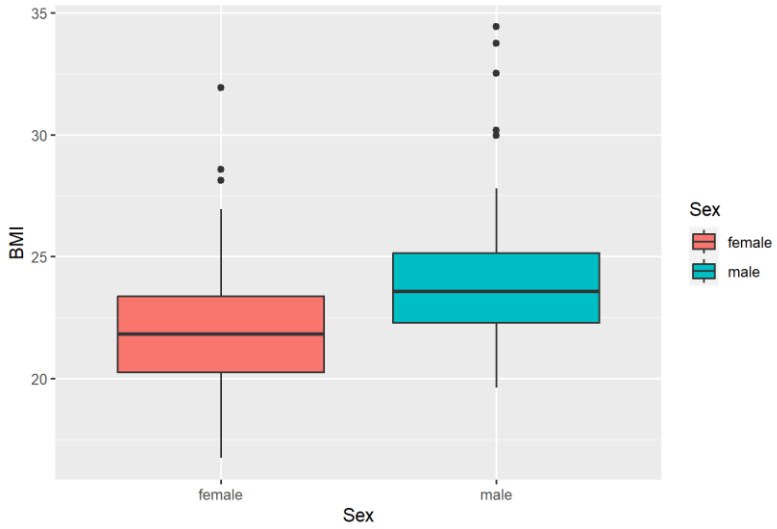


The distribution of height and weight are normal distribution, while for the other numerical variables are right skewed as shown in the figure below



To present the body mass index per sex, the geom_boxplot() method is used to create a boxplot of BMI.

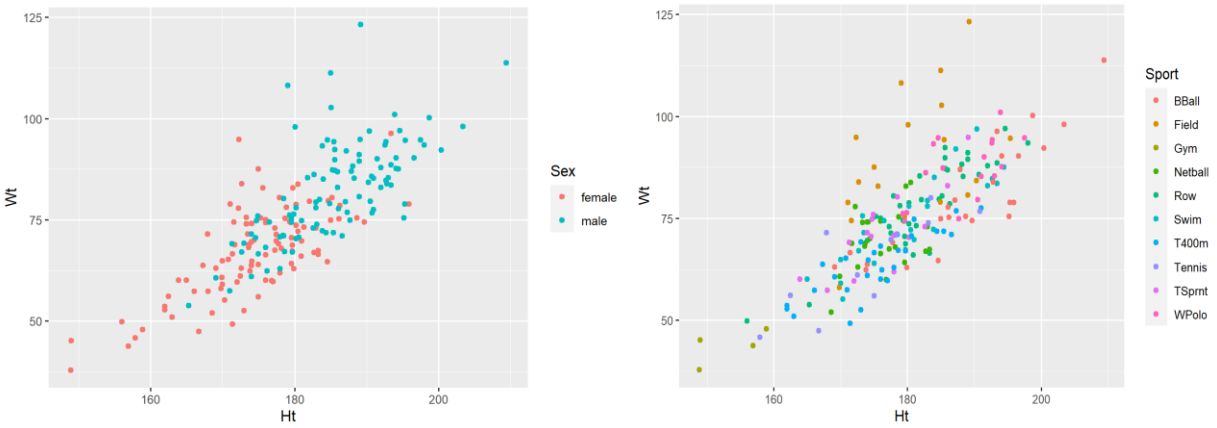
```
> #create  a boxplot of BMI, grouped by Sex
> ggplot(data=Athlete_data, aes(x=Sex, y=BMI, fill=Sex)) +  geom_boxplot()
>
```



It is clearly seen that the body mass index for male is greater than female

Additionally, we plot a scatterplot of the height and weight variables for each sex and sport using the geom_point() function.

```
> #create scatterplot of height and weight and Sex as color variable
> ggplot(data=Athlete_data, aes(x=Ht, y=Wt, color=Sex)) +
+     geom_point()
> #create scatterplot of height and weight and Sport types as color variable
> ggplot(data=Athlete_data, aes(x=Ht, y=Wt, color=Sport)) +   geom_point()
>
```



Both scatterplots show a linear relationship between height and weight.

The correlation coefficient matrix is used to determine the strength and direction of a linear relationship between two continuous variables such as height and weight.

```
> #create correlation matrix
> round(cor(Athlete_data[c('RCC', 'WCC', 'Hc', 'Hg',  'Ferr', 'BMI',   'SSF', '%Bfat',
'LBM', 'Ht', 'Wt')]), 2)
         RCC  WCC    Hc    Hg  Ferr  BMI   SSF %Bfat   LBM    Ht   Wt
RCC     1.00 0.15  0.92  0.89  0.25 0.30 -0.40 -0.49  0.55  0.36 0.40
WCC     0.15 1.00  0.15  0.13  0.13 0.18  0.14  0.11  0.10  0.08 0.16
Hc      0.92 0.15  1.00  0.95  0.26 0.32 -0.45 -0.53  0.58  0.37 0.42
Hg      0.89 0.13  0.95  1.00  0.31 0.38 -0.44 -0.53  0.61  0.35 0.46
Ferr    0.25 0.13  0.26  0.31  1.00 0.30 -0.11 -0.18  0.32  0.12 0.27
BMI     0.30 0.18  0.32  0.38  0.30 1.00  0.32  0.19  0.71  0.34 0.85
SSF    -0.40 0.14 -0.45 -0.44 -0.11 0.32  1.00  0.96 -0.21 -0.07 0.15
%Bfat -0.49 0.11 -0.53 -0.53 -0.18 0.19  0.96  1.00 -0.36 -0.19 0.00
LBM     0.55 0.10  0.58  0.61  0.32 0.71 -0.21 -0.36  1.00  0.80 0.93
Ht      0.36 0.08  0.37  0.35  0.12 0.34 -0.07 -0.19  0.80  1.00 0.78
Wt      0.40 0.16  0.42  0.46  0.27 0.85  0.15  0.00  0.93  0.78 1.00
>
```

A correlation coefficient close to 1 or -1 indicates a strong positive or negative relationship, while a coefficient close to 0 indicates a weak or no relationship.

To find missing values in the data, we use the following code:

```
> #count total missing values in each column
> sapply(Athlete_data, function(x) sum(is.na(x)))
   Sex Sport   RCC   WCC    Hc    Hg  Ferr   BMI   SSF %Bfat   LBM    Ht    Wt
     0     0     0     0     0     0     0     0     0     0     0     0     0
>
```

It is shown in the output that each column has exactly zero missing values.

3. Results

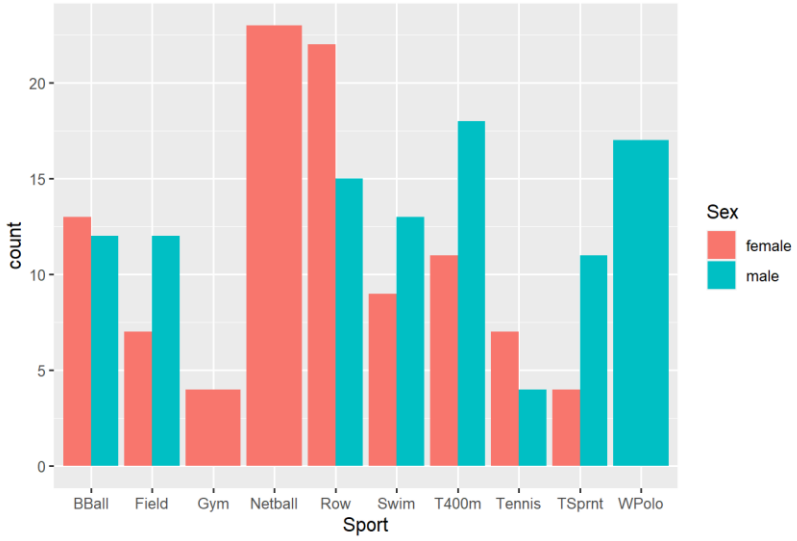
Hypothesis testing is a crucial procedure for understanding the relationship between athletes' and their physical markers. It involves defining the research question and hypothesis, selecting

appropriate statistical tests like t-test, ANOVA, and regression, performing the test using software like R, and interpreting the results to draw conclusions based on the formulated hypothesis.

chi-square test

The chi-square test is a statistical method used to determine the association between sex and sport type, comparing different sport types and sex groupings using a bar chart.

```
> ggplot(Athlete_data, aes(x = Sport, fill = Sex)) + geom_bar(position = "dodge")
>
```



The bar chart shows that males and females play 10 sports, with netball being the most popular sport for female. Boys prefer T400m, while girls prefer Netball. Gym is least preferred by female. Females participate in more sports than boys. Tennis has minimal gender differences, but netball and Gym, Wpolo sport show significant gender differences.

A chi-square test involves creating a contingency table displaying observed frequencies of categories, comparing them with expected frequencies under the null hypothesis of no association.

```
> # create a contingency table
> ct<- table(Athlete_data$Sex, Athlete_data$Sport)
> # view the contingency table
> ct

        BBall Field Gym Netball Row Swim T400m Tennis TSprnt WPolo
  female    13     7   4      23  22    9    11      7      4     0
  male      12    12   0       0  15   13    18      4     11    17
> #add margins to the contingency table
> ct_m <- addmargins(ct)
> # view the contingency table
> ct_m

        BBall Field Gym Netball Row Swim T400m Tennis TSprnt WPolo Sum
  female    13     7   4      23  22    9    11      7      4     0 100
  male      12    12   0       0  15   13    18      4     11    17 102
  Sum       25    19   4      23  37   22    29     11     15    17 202
>
```

The table shows 202 observations of female and male sports preferences, with 13 females playing BBall and 22 practicing Row sport, among other sports.

The research question is whether the sport an athlete participates in is influenced by gender classification.

The Hypotheses: we are going to test are:

The null hypothesis H_0: sport and sex are not associated, they are independent
The alternative hypothesis H_1: sport and sex are related

Test statistic: the chi-square test statistic is defined as

$$X_2 = \sum_{i=1}^{c} \frac{(O_i - E_i)}{E_i}$$

Where:

C the number of classes

X_2 is the chi-square test statistic

O is the observed frequency

E is the expected frequency

To test if there is a relationship between sex and sport at a significance level of 5%, we apply the chisq.test () function

```
> # perform a chi-square test,
> chi_test <- chisq.test(ct)
> chi_test

        Pearson's Chi-squared test

data:  ct
X-squared = 53.167, df = 9, p-value = 0.00000002717

> |
```

The output shows the test statistic 53.167, the degrees of freedom 9, and the p-value of the test .00000002717 which < 0.05, indicates that sex and sport are dependent, rejecting the null hypothesis.

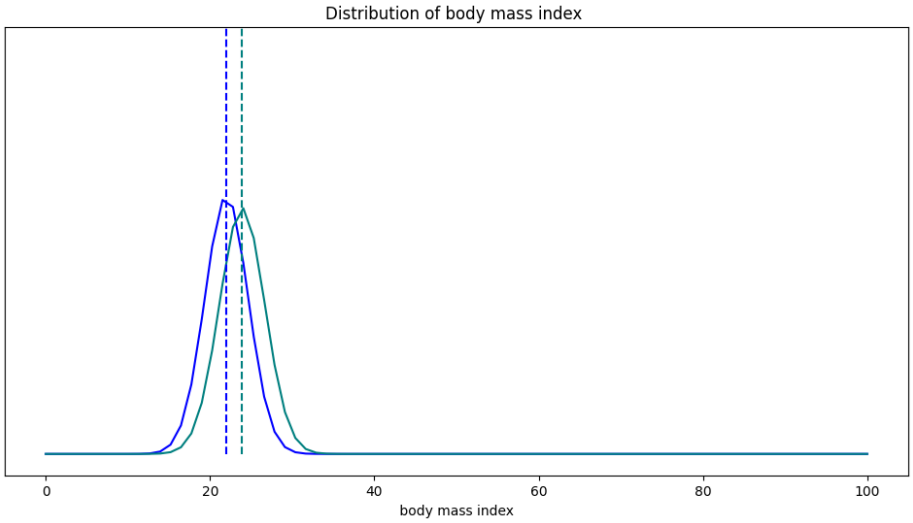
T-test:

The t-test compares the means of two groups, male and female athletes, on a continuous variable like body mass index, indicating if there is a significant difference.
Research question: is there difference in body mass index between male and female athletes.
The mean and standard deviation for each group is calculated.

```
> df_stat <- Athlete_data %>% group_by(Sex) %>% summarize( count = n(), mean = mean(BMI,
na.rm = TRUE), sd = sd(BMI, na.rm = TRUE))
> df_stat
# A tibble: 2 × 4
  Sex    count  mean    sd
  <chr>  <int> <dbl> <dbl>
1 female   100  22.0  2.64
2 male     102  23.9  2.77
>
```

```
[1] import numpy as np
    import matplotlib.pyplot as plt
    import scipy.stats as stats
```

```
sd1 = 2.64
sd2=2.77
mean1=22
mean2=23.9

# plot normal distribution for both means body mass index
maxy=0.25
plot1 = plt.figure(figsize=(12,6))
x = np.linspace(0, 100, num=80)
plt.plot(x, stats.norm.pdf(x, mean1, sd1),color='Blue')
plt.plot(x, stats.norm.pdf(x, mean2, sd2),color='Teal')
plt.vlines(mean1, 0, maxy, colors='Blue', linestyle='dashed')
plt.vlines(mean2, 0, maxy, colors='Teal', linestyle='dashed')
plt.ylim(top=maxy)
plt.xlabel(' body mass index ')
plt.title('Distribution of body mass index ')
# remove y axis
frame = plt.gca()
frame.axes.get_yaxis().set_visible(False)

plt.show(plot1)
```



The Hypotheses:

The null hypothesis H_0: there is no difference in means of body mass index between male and female athletes

The alternative hypothesis H_1: there is a difference in body mass index between male and female

Test statistic: The T test formula is as shown

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\left(S_p^2\left(\frac{1}{n_1} - \frac{1}{n_2}\right)\right)}}$$

x_1, x_2 represent the two groups' mean

S_p represent the standard error of the two groups

n_1 and n_2 are the number of observations in each group

It assumes normal distributions and equal variances.

```
> T_test<-t.test(Athlete_data$BMI ~ Athlete_data$Sex, data = Athlete_data)
> T_test

        Welch Two Sample t-test

data:  Athlete_data$BMI by Athlete_data$Sex
t = -5.0313, df = 199.85, p-value = 1.083e-06
alternative hypothesis: true difference in means between group female and group male is
not equal to 0
95 percent confidence interval:
 -2.664750 -1.164105
sample estimates:
mean in group female   mean in group male
            21.98920             23.90363
```

The t-test results show a mean difference of -5.0313, with a confidence interval of -2.664750 to -1.64050, p-value 0.000001083 < 0.05, confirming the alternative hypothesis that there is significant difference in means of body mass index between male and female athletes.

Furthermore, we use t- test to see if there is no difference in means of bodyfat percentage between male and female athletes as shown

```
> Athlete_data %>% group_by(Athlete_data$Sex) %>%summarize(count = n(),mean = mean(Athlete_data$Bfat, na.rm
= TRUE),sd = sd(Athlete_data$Bfat, na.rm = TRUE))
# A tibble: 2 × 4
  `Athlete_data$Sex` count  mean    sd
  <fct>              <int> <dbl> <dbl>
1 female               100  13.5  6.19
2 male                 102  13.5  6.19
>
```

```
> # run t- test to see if there is difference in means of bodyfat percentage between male and female athlete
s
> BFat_test<-t.test(Bfat~Sex, data= Athlete_data)
> BFat_test

        Welch Two Sample t-test

data:  Bfat by Sex
t = 13.65, df = 158.87, p-value < 0.00000000000000022
alternative hypothesis: true difference in means between group female and group male is not equal to 0
95 percent confidence interval:
 7.354159 9.842276
sample estimates:
mean in group female   mean in group male
           17.849100            9.250882

>
```

The test indicates there is a significant different in body fate percentage between male and female.

Analysis of Variance ANOVA

A one-way ANOVA test compares the means of two groups of observations, using a single independent variable. It can determine significant differences or chance-based differences, like sport types, on a continuous variable like lean body mass. A two-way ANOVA test compares average values across multiple groups, like sport type, sex, and plasma ferritin concentration.

The research question is whether there is a relationship between sport type and lean body mass.

The lean body mass statistical summary by sport is displayed in the table below.

```
> LBM_stat <- Athlete_data %>% group_by(Sport) %>% summarize( count = n(), mean = mean(L
BM, na.rm = TRUE), sd = sd(LBM, na.rm = TRUE))
> LBM_stat
# A tibble: 10 × 4
   Sport   count  mean    sd
   <chr>   <int> <dbl> <dbl>
 1 BBall      25  68.3 14.2
 2 Field      19  76.5 14.0
 3 Gym         4  38.7  3.52
 4 Netball    23  54.3  3.92
 5 Row        37  66.6 12.4
 6 Swim       22  67.4 11.3
 7 T400m      29  58.6  8.09
 8 TSprnt     15  65.7  9.62
 9 Tennis     11  56.2 11.2
10 WPolo      17  75.9  5.97
> |
```

The hypotheses are:

The null hypothesis H_0: is the mean of lean body mass between sports kinds is equal.

The alternative hypothesis H_1: One of the groups' means differs noticeably from the others.

The test statistic: the F statistic is defined as

$$\frac{Betwwen\ Group\ Variance}{Within\ Group\ Variance}$$

The ANOVA test makes the following assumptions:

Observational independence: the data were acquired using statistically acceptable sampling methods, and there are no hidden correlations between observations.

To perform an ANOVA, the command aov() is used. We will model the differences in the mean of the response variable, lean body mass, as a function of sport type in this example.

```
> # One-way ANOVA
> one.way_LBM <- aov(LBM~ Sport, data = Athlete_data)
> summary(one.way_LBM)
             Df Sum Sq Mean Sq F value               Pr(>F)
Sport         9  12487  1387.4   12.19 0.00000000000000356 ***
Residuals   192  21850   113.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> |
```

The one-way ANOVA result is revealed that there is a statistically significant difference in the mean of lean body mass between at least two sport groups (F-value = 12.19, df =9 and p = 0.00000000000000356 <.05).

Tukey's HSD Test can be used to identify means that are significantly different from one another.

Two-way ANOVA

We perform two-way ANOVA to ascertain whether the athlete's sex and sport are important variables in determining the plasma ferritin concentration.

The research question: if sport type, gender, or the interaction between sport and gender affects plasma ferritin concentration.

The Hypotheses are:

The null hypothesis (H_0) states that there is no difference in the mean hematocrit level among sport types. The alternative hypothesis (H_1) is that one of the groups' means differs significantly from the others.

```
> # Two-way ANOVA whereWe want to know if sport type, gender, or the interaction between sport and gender de
termining the plasma ferritin concentration
> aov2_test<-aov(Ferr~Sex+Sport+Sex*Sport, data= Athlete_data)
> summary(aov2_test)
             Df Sum Sq Mean Sq F value         Pr(>F)
Sex           1  78554   78554  46.380 0.000000000132 ***
Sport         9  50683    5631   3.325       0.000865 ***
Sex:Sport     6  10956    1826   1.078       0.377153
Residuals   185 313337    1694
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

The implications we may derive from this result are that the gender factor and sport type are significant at the 5% significance level, while the interaction is not significant at the 5% significance level.

Multiple Linear Regression

To make a model that predicts plasma ferritin levels in an athlete given the physical markers, we will need to use a multiple linear regression. This can model the relationship between a dependent variable, such as plasma ferritin concentration, and one or more independent variables, such as sport, sex, height, weight, lean body mass, etc. The multiple linear regression model can tell us how much the dependent variable changes when the independent variables change, and how well the model fits the data.

In this section, we will build the multiple regression model to predict plasma ferritin level in an athlete given the following physical markers: Sex, Sport, RCC, WCC, Hc, Hg, BMI, SSF, Bfat, LBM, Ht, Wt, and Wt.

The multiple linear regression model formula is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where

Y is the dependent variable

α is the intercept of the regression line

$\beta_{1, \ldots}, \beta_k$ are the coefficient of the independent variables

$X_1, \ldots, X_k$ are the independent variables

We are looking to find the values of α and β which give the highest possible correlation coefficient R between the observed values and the predicted values of Y.

The Y variable's total sum of squares is

$$\sum (Y_i - \bar{Y})^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 + \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$

The values of α, β_1, β_2,... β_k that minimize the residual sum of squares are what we are looking for.

The best estimate of α, $\hat{\alpha}$ is defined as follows:

$$\hat{\alpha} = Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - \cdots - \hat{\beta}_k X_k + \epsilon$$

The following assumptions are made by the model:

Linearity: the relationship between the dependent variable and independent variables is linear.

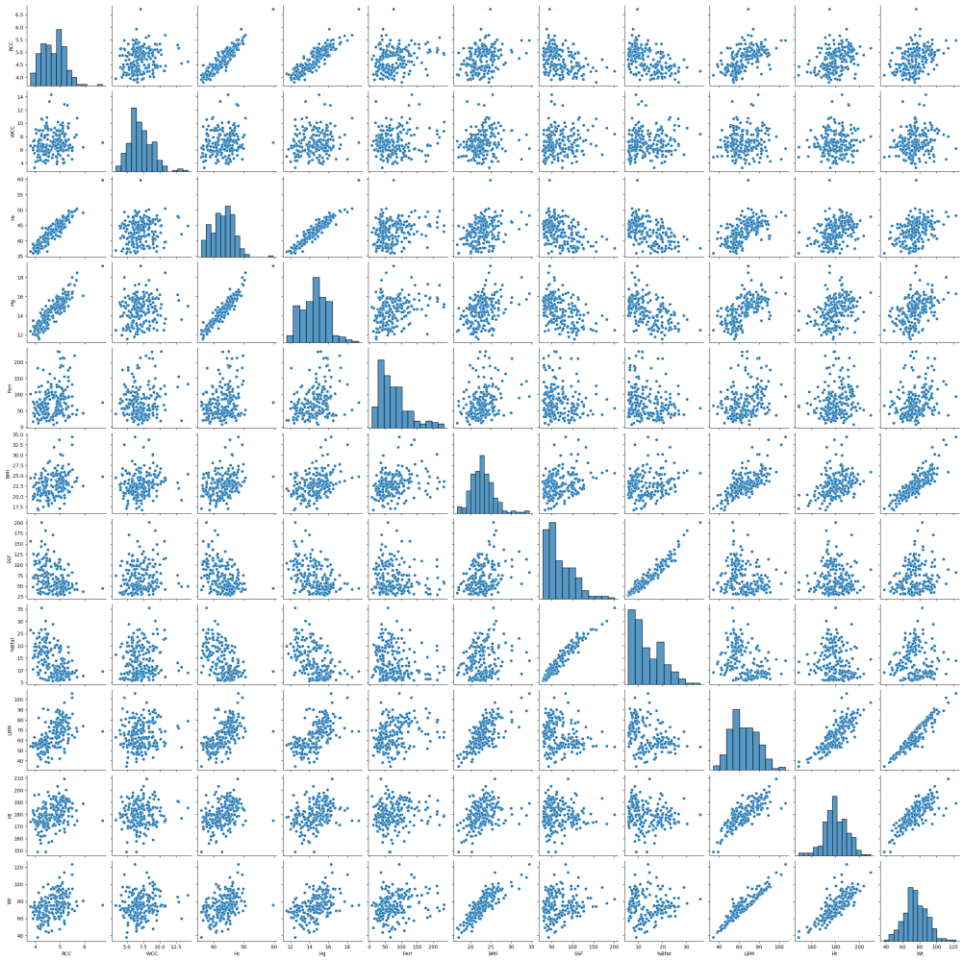
Normality: The residuals have a normal distribution with a mean of 0.

Independence of observations

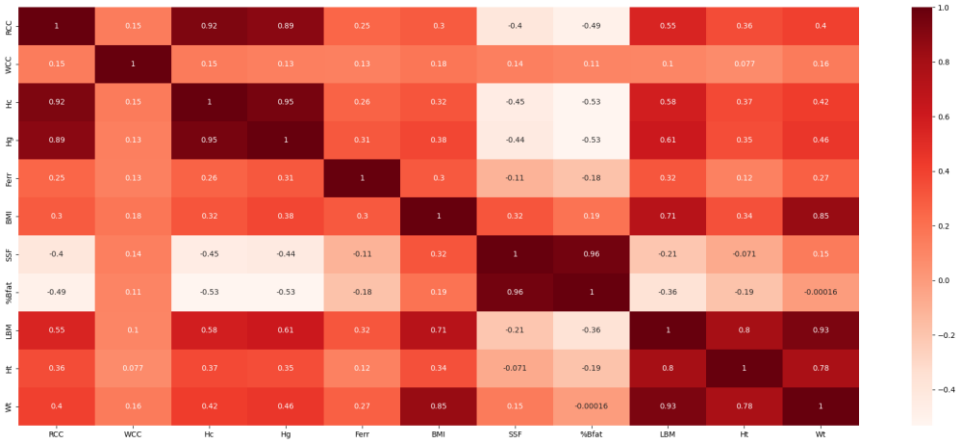
Constant spread: For all x values, the residuals have the same standard deviation.

We will check whether there is a linear relationship between the independent and dependent variables: height, weight, lean body mass, red cell count, white cell count, Hematocrit, Hemoglobin, body mass index, sum of skin folds, and body fat percentage. If the distribution of data points could be characterized by a straight line, we could visually evaluate this using a scatter plot. The relationship is roughly linear as in the figure below

```
#create pairs plot for all numeric variables
sns.pairplot(Athlete_df)
```



The heatmap visualizes the correlation between variables as below.



In the heatmap, there are high correlation coefficients as in the dark color between such as Rcc and Hc (r=0.92), Hc and Hg (r=0.95), Hc and RCC (r=0.92), Ht and Wt (r=0.78), body fat and SSF (r=0.96), Wt and BMI (r=0.501), Wt and LBM (r=0.93).

The statistical tests: we will use the t-test and the F-test to see if the model coefficients are different from zero and if the model can explain a significant amount of variability in the response variable, accordingly.

Here, we are assuming a linear relationship between the concentration of plasma ferritin and each of the independent factors. To test this, we use the t-test to evaluate the model coefficients for each of the ten covariates (j = 1–10).

The hypotheses are:

The null hypothesis: H_0: $\beta_j=0$

The alternative hypothesis: H_1: $\beta_j\neq0$

The statistic test: t-test

$$T = \frac{\hat{\beta_i}}{s\beta_i} \sim t(n-k-1)$$

For the F-test, the hypotheses are:

The null hypothesis: $H_{0:}\ \beta_1 =\beta_2=\ldots=\beta_{10}=0$

The alternative hypothesis: $H_{1:}\ \beta_j\neq0$, for at least one j value

The statistical test:

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

where

The square of the multiple correlation coefficient is represented as R^2

k and n − k − 1 degree of freedom, where k is the number of independent variables and n is the number of response variable observations.

The significance level: 0.05, to determine if the results are statistically significant or not.

The model is built by splitting the dataset into a training and test set, with 80% used for training and 20% for testing.

```
> #set 80% of dataset as training set and 20% as test set
> sample <- sample(c(TRUE, FALSE), nrow(Athlete_data), replace=TRUE, prob=c(0.8,0.2))
> train  <- Athlete_data[sample, ]
> test   <- Athlete_data[!sample, ]
> #view dimensions of the training set
> dim(train)
[1] 166  14
> #view dimensions of the test set
> dim(test)
[1] 36 14
```

The training set data is fitted into a multiple regression model using the lm() function in R.

```
> #set up a mutiple  linear regression model where Plasma ferritin  is the dependent variable which varies w
ith independent variables: Sex, Sport, RCC, WCC, Hc, Hg, BMI, SSF, Bfat, LBM, Ht, Wt and Wt
> model_Ferr<-lm(data= train, Ferr ~ Sex+Sport+RCC+WCC+Hc+Hg+BMI+SSF+ Bfat+LBM+Ht+ Wt)
> summary(model_Ferr)

Call:
lm(formula = Ferr ~ Sex + Sport + RCC + WCC + Hc + Hg + BMI +
    SSF + Bfat + LBM + Ht + Wt, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-80.30 -23.77  -6.70  19.05 118.05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  663.7900   515.6019   1.287 0.200003
Sexmale       66.6046    16.7121   3.985 0.000106 ***
SportField    34.0718    18.8511   1.807 0.072770 .
SportGym       5.6340    32.4908   0.173 0.862577
SportNetball  19.5415    15.3485   1.273 0.204990
SportRow      25.4937    13.2487   1.924 0.056283 .
SportSwim     29.1423    15.4081   1.891 0.060570 .
SportT400m    12.3067    16.6130   0.741 0.460019
SportTennis   49.3752    19.0829   2.587 0.010653 *
SportTSprnt   44.4124    20.3185   2.186 0.030434 *
SportWPolo    37.3598    16.4666   2.269 0.024756 *
RCC          -17.3316    20.9676  -0.827 0.409828
WCC            1.3601     2.1810   0.624 0.533869
Hc            -4.1435     4.0296  -1.028 0.305529
Hg            12.0684     8.8127   1.369 0.172982
BMI           -9.3697    11.5289  -0.813 0.417714
SSF           -0.2060     0.5414  -0.381 0.704118
Bfat           2.7880     4.8144   0.579 0.563424
LBM           -0.2165     5.7097  -0.038 0.969811
Ht            -3.5123     2.8663  -1.225 0.222416
Wt             3.5758     5.9057   0.605 0.545804
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.32 on 145 degrees of freedom
Multiple R-squared:  0.3524,    Adjusted R-squared:  0.263
F-statistic: 3.945 on 20 and 145 DF,  p-value: 0.0000006202
```

The summary indicates that male sex and specific sports, particularly field, row, swim, tennis, TSPnt, and WPolo, are significantly associated with plasma ferritin levels. The model explains only 35.24 % of the variability in the response (R2).

And the result of F-test is:

F=3.945, df= 20,145 p-value= 0.000000006202

This shows that the regression model is significant in explaining the response variability.

To assess the prediction model's performance, we must validate it on the test dataset, which was not utilized to estimate the model parameters.

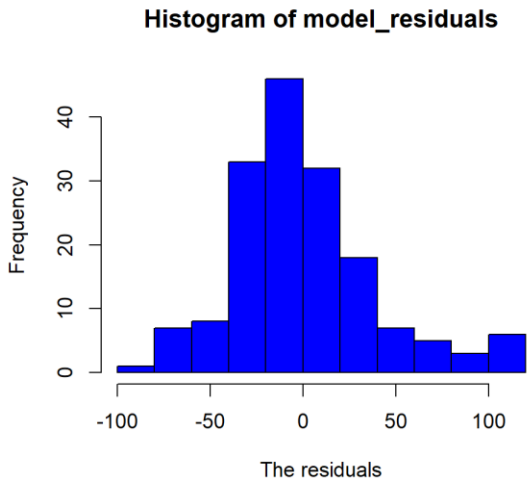
```
> # predicting the Ferr variable
> predicted_Ferr <- predict(model_Ferr, test)
> ## computing the ferrmodel performance metrics
> data.frame(R2 = R2(predicted_Ferr, test$Ferr),
+            RMSE = RMSE(predicted_Ferr, test$Ferr),
+            MAE = MAE(predicted_Ferr, test$Ferr))
         R2     RMSE      MAE
1 0.2106952 40.45568 31.90933
> 
```

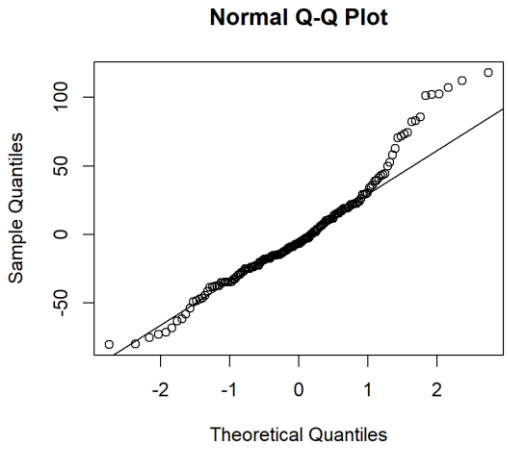
After we have developed the model, we will go over the assumptions and evaluate the results.

To check the distribution of the model residuals. This may be demonstrated in R by using the hist () function.

```
> # Get the Ferr model residuals
> model_residuals = model_Ferr$residuals
> # Plot the  Ferr model residuals
> hist(model_residuals, xlab="The residuals",ylab="Frequency", col = "blue")
> 
```

```
> # Plot the residuals
> qqnorm(model_residuals)
> # Plot the Q-Q line
> qqline(model_residuals)
```





The histogram looks nearly bell-shaped; hence we can conclude the normality with enough confidence. Instead of the histogram, consider the residuals along the standard Q-Q plot. If the readings are normal, they should follow a straight line.

To make sure of the normality of our residuals we use the Anderson-Darling test.

The hypotheses are as follows:

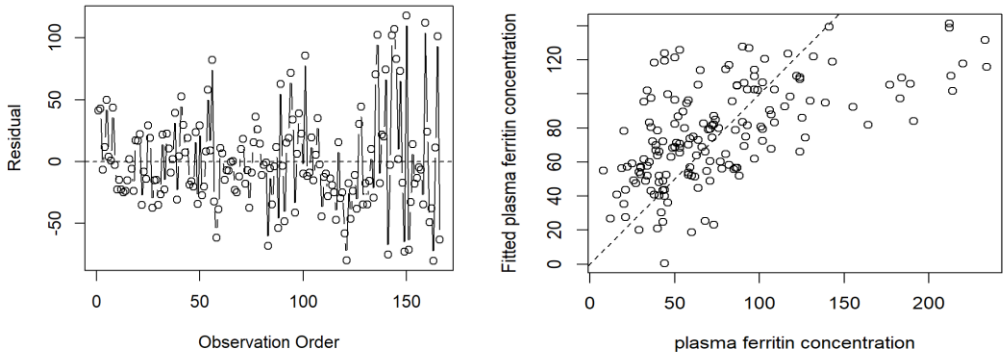
The null hypothesis H_0: the data have a normal distribution.

The alternative hypothesis H_1: the data are not normally distributed.

```
>
> library("nortest")
> # perform the Anderson-Darling test
> ad.test(model_residuals)

        Anderson-Darling normality test

data:  model_residuals
A = 2.2626, p-value = 0.00000929

> |
```

Test statistic A=2.2626, p-value = 0.00000929 for the Anderson-Darling test. There is enough evidence to show that the residuals are normally distributed, as p-value=0.00000063210.05. As a result, the assumption of residual normalcy is correct.

```
> #plot predicted values vs residuals
> plot(predict.lm(model_Ferr),model_residuals,xlab="Fitted values",ylab="Residual")
> #  view value of residual = 0 with  horizontal line t
> abline(h=0)
> #and to test independence of observations plot residual in order of observation
> plot (seq(1,length(model_residuals),1),model_residuals,xlab="Observation Order",ylab="Residual",type="b")
> # dashed horizontal line to show where residual = 0
> abline(h=0,lty=2)
> # plot fitted (predicted)  plasma ferritin concentration  vs actual  plasma ferritin concentration
> plot(train$Ferr,predict.lm(model_Ferr),xlab="plasma ferritin concentration",ylab="Fitted plasma ferritin c
oncentration")
> abline(0,1,lty=2)
> |
```

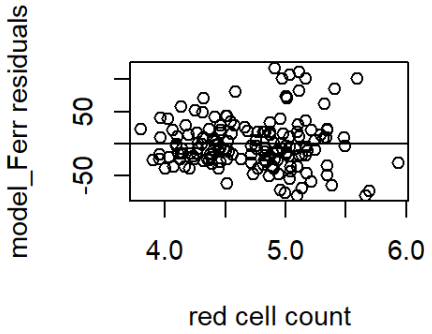


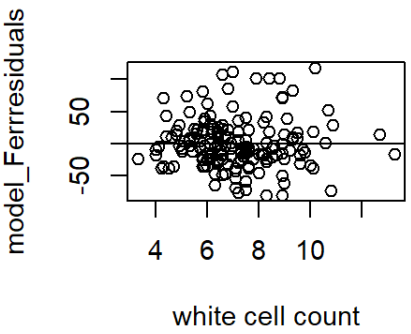
The Residual vs Order diagnostic plot reveals no clear trend, allowing us to conclude that the residuals are independent of one another.

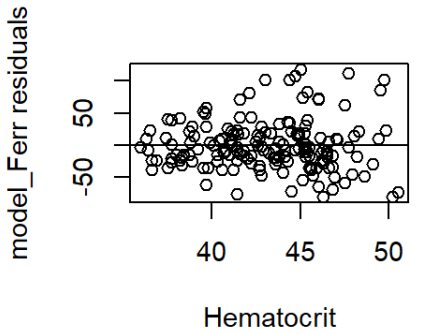
We assess the model's goodness of fit as in the right figure. The fitted values are significantly underestimating the enormous values of the reported Plasma ferritin levels, which could be related to the existence of multicollinearity. That is explained why the R squared is 35.24.

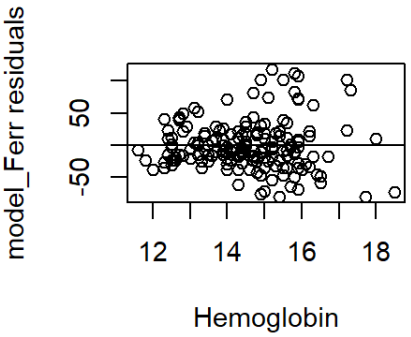
Lastly, to ensure a linear relationship between each independent variable and the dependent variable. We plot the residuals against each of the covariates to allow us to check this. The residuals will be dispersed at random above and below the zero line if the relationship is linear.

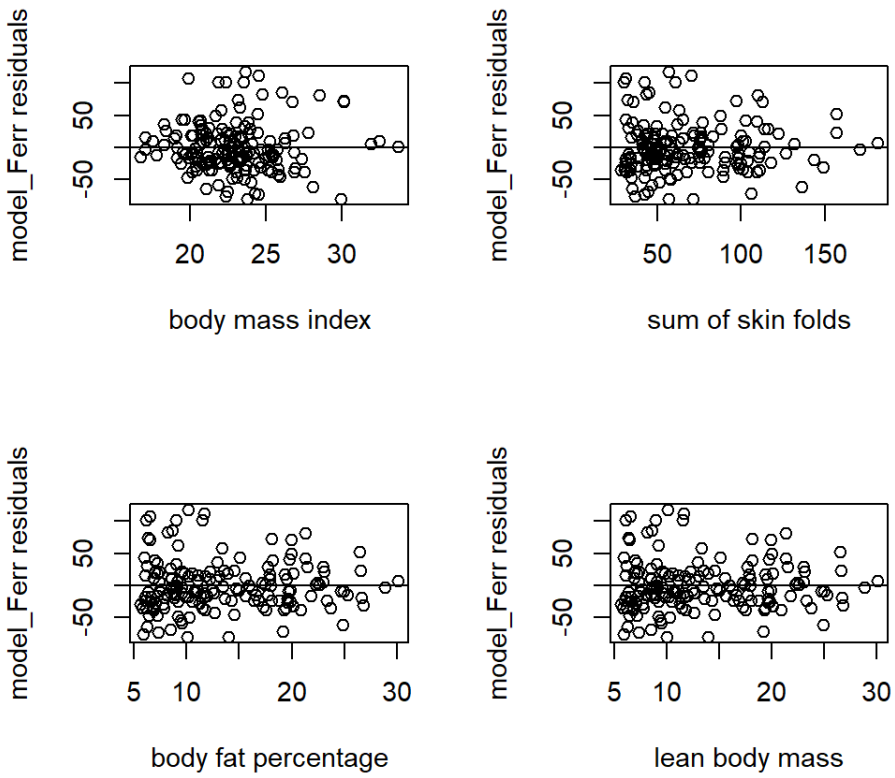
```
> with(train, {
+     # RCC
+     plot(RCC,model_residuals,xlab="red cell count",ylab="model_Ferr residuals")
+     # horizontal line to show value of residual = 0
+     abline(h=0)
+     # WCC
+     plot(WCC,model_residuals,xlab="white cell count",ylab="model_Ferrresiduals")
+     # horizontal line to show value of residual = 0
+     abline(h=0)
+     # Hc
+     plot(Hc,model_residuals,xlab="Hematocrit",ylab="model_Ferr residuals")
+     # horizontal line to show value of residual = 0
+     abline(h=0)
+     # Hg
+     plot(Hg,model_residuals,xlab="Hemoglobin",ylab="model_Ferr residuals")
+     # horizontal line to show value of residual = 0
+     abline(h=0)
+     # BMI
+     plot(BMI,model_residuals,xlab="body mass index",ylab="model_Ferr residuals")
+     # horizontal line to show value of residual = 0
+     abline(h=0)
+     # SSF
+     plot(SSF,model_residuals,xlab="sum of skin folds",ylab="model_Ferr residuals")
+     # horizontal line to show value of residual = 0
+     abline(h=0)
+     # Bfat
+     plot(Bfat,model_residuals,xlab="body fat percentage",ylab="model_Ferr residuals")
+     # horizontal line to show value of residual = 0
+     abline(h=0)
+     # LBM
+     plot(Bfat,model_residuals,xlab="lean body mass",ylab="model_Ferr residuals")
+     # horizontal line to show value of residual = 0
+     abline(h=0)
+     # Ht
+     plot(Ht,model_residuals,xlab="Height",ylab="model_Ferrresiduals")
+     # horizontal line to show value of residual = 0
+     abline(h=0)
+     # Wt
+     plot(Wt,model_residuals,xlab="Weight",ylab="model_Ferr residuals")
+     # horizontal line to show value of residual = 0
+     abline(h=0)
+ })
```











The scatterplots of residuals versus each covariate reveal no pattern indicating that there is a linear relationship.

Our model has many variables; therefore, we might have to decide which are crucial.

To do this, another statistical test (F-test) will be used to decide if a variable can be included in the model. When the parameter "test=F" is passed to the R function drop1(), it performs this test. Think of two models: Model0, a subset of Model1, which has k 0 variables, and Model1, which has n observations and k 1 variables.

The hypotheses are:

The null hypothesis H_0: All the extra coefficients in outer Model 1 equal zero

The alternative Hypothesis H_1: The outer Model 1 extra coefficients do not equal 0

To do this, we will utilize the R function step () shown below.

```
> with(train, {
+     # try the drop1() function
+     drop1(model_Ferr,test="F")
+ })
Single term deletions

Model:
Ferr ~ Sex + Sport + RCC + WCC + Hc + Hg + BMI + SSF + Bfat +
    LBM + Ht + Wt
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              247605 1255.1
Sex     1   27122.9 274728 1270.3 15.8834 0.0001063 ***
Sport   9   21724.2 269329 1251.0  1.4135 0.1873302
RCC     1    1166.7 248772 1253.8  0.6832 0.4098282
WCC     1     664.1 248269 1253.5  0.3889 0.5338686
Hc      1    1805.6 249410 1254.3  1.0574 0.3055289
Hg      1    3202.4 250807 1255.2  1.8753 0.1729824
BMI     1    1127.9 248733 1253.8  0.6605 0.4177145
SSF     1     247.3 247852 1253.2  0.1448 0.7041180
Bfat    1     572.6 248178 1253.5  0.3353 0.5634244
LBM     1       2.5 247607 1253.1  0.0014 0.9698114
Ht      1    2564.1 250169 1254.8  1.5016 0.2224158
Wt      1     626.0 248231 1253.5  0.3666 0.5458035
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Thus, at last, we have identified a model that has one variable that is significant according to our F-test.

```
> #set up a multiple  linear regression model where Plasma ferritin  is the dependent variable which varies
with independent variables: Sex
> second_Ferrmodel<-lm(data= train, Ferr ~ Sex)
> summary(second_Ferrmodel)

Call:
lm(formula = Ferr ~ Sex, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-88.690 -27.207  -6.449  17.930 137.310

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)   56.207      4.835  11.626 < 0.0000000000000002 ***
Sexmale       40.483      6.796   5.957         0.0000000152 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.78 on 164 degrees of freedom
Multiple R-squared:  0.1779,    Adjusted R-squared:  0.1729
F-statistic: 35.48 on 1 and 164 DF,  p-value: 0.00000001523
```

Let ‘s find out if backward selection will choose the same model.

```
> with(train, {
+
+     # try the step() function backwards to test our model selection
+     step(model_Ferr,direction="backward")
+
+ })
Start:  AIC=1255.06
Ferr ~ Sex + Sport + RCC + WCC + Hc + Hg + BMI + SSF + Bfat +
    LBM + Ht + Wt

       Df Sum of Sq    RSS    AIC
- Sport  9   21724.2 269329 1251.0
- LBM    1       2.5 247607 1253.1
- SSF    1     247.3 247852 1253.2
- Bfat   1     572.6 248178 1253.5
- Wt     1     626.0 248231 1253.5
- WCC    1     664.1 248269 1253.5
- BMI    1    1127.9 248733 1253.8
- RCC    1    1166.7 248772 1253.8
- Hc     1    1805.6 249410 1254.3
- Ht     1    2564.1 250169 1254.8
<none>              247605 1255.1
- Hg     1    3202.4 250807 1255.2
- Sex    1   27122.9 274728 1270.3

Step:  AIC=1251.02
Ferr ~ Sex + RCC + WCC + Hc + Hg + BMI + SSF + Bfat + LBM + Ht +
    Wt

       Df Sum of Sq    RSS    AIC
- LBM    1       0.0 269329 1249.0
- BMI    1      37.4 269366 1249.0
- RCC    1      58.7 269388 1249.1
- Wt     1     143.3 269472 1249.1
- WCC    1     453.7 269783 1249.3
- SSF    1     698.7 270028 1249.5
- Bfat   1     888.8 270218 1249.6
- Ht     1    1935.9 271265 1250.2
- Hc     1    2352.3 271681 1250.5
- Hg     1    2571.4 271901 1250.6
<none>              269329 1251.0
- Sex    1   26441.8 295771 1264.6
```

```
Step:  AIC=1249.02
Ferr ~ Sex + RCC + WCC + Hc + Hg + BMI + SSF + Bfat + Ht + Wt

       Df Sum of Sq    RSS    AIC
- BMI   1      37.3 269366 1247.0
- RCC   1      58.7 269388 1247.1
- WCC   1     454.3 269783 1247.3
- Wt    1     494.2 269823 1247.3
- SSF   1     746.8 270076 1247.5
- Bfat  1    1915.6 271245 1248.2
- Ht    1    1948.1 271277 1248.2
- Hc    1    2354.4 271684 1248.5
- Hg    1    2571.8 271901 1248.6
<none>               269329 1249.0
- Sex   1   26499.3 295828 1262.6

Step:  AIC=1247.05
Ferr ~ Sex + RCC + WCC + Hc + Hg + SSF + Bfat + Ht + Wt

       Df Sum of Sq    RSS    AIC
- RCC   1      59.8 269426 1245.1
- WCC   1     454.9 269821 1245.3
- SSF   1     747.4 270114 1245.5
- Bfat  1    1884.5 271251 1246.2
- Hc    1    2384.8 271751 1246.5
- Hg    1    2573.1 271940 1246.6
<none>               269366 1247.0
- Wt    1    7135.7 276502 1249.4
- Ht    1   19715.7 289082 1256.8
- Sex   1   26470.5 295837 1260.6

Step:  AIC=1245.08
Ferr ~ Sex + WCC + Hc + Hg + SSF + Bfat + Ht + Wt

       Df Sum of Sq    RSS    AIC
- WCC   1     463.1 269889 1243.4
- SSF   1     716.2 270142 1243.5
- Bfat  1    1860.2 271286 1244.2
- Hg    1    2598.7 272025 1244.7
- Hc    1    2778.8 272205 1244.8
<none>               269426 1245.1
- Wt    1    7090.9 276517 1247.4
- Ht    1   19677.6 289104 1254.8
- Sex   1   27115.6 296542 1259.0

Step:  AIC=1243.37
Ferr ~ Sex + Hc + Hg + SSF + Bfat + Ht + Wt

       Df Sum of Sq    RSS    AIC
- SSF   1     655.5 270545 1241.8
- Bfat  1    1903.4 271793 1242.5
- Hc    1    2516.4 272406 1242.9
- Hg    1    2545.1 272434 1242.9
<none>               269889 1243.4
- Wt    1    6996.7 276886 1245.6
- Ht    1   19677.1 289566 1253.0
- Sex   1   27222.2 297112 1257.3

Step:  AIC=1241.77
Ferr ~ Sex + Hc + Hg + Bfat + Ht + Wt

       Df Sum of Sq    RSS    AIC
- Hc    1    2352.3 272897 1241.2
- Hg    1    2504.8 273050 1241.3
- Bfat  1    3237.0 273782 1241.7
<none>               270545 1241.8
- Wt    1    6395.2 276940 1243.7
- Ht    1   19131.4 289676 1251.1
- Sex   1   27958.6 298503 1256.1

Step:  AIC=1241.21
Ferr ~ Sex + Hg + Bfat + Ht + Wt

       Df Sum of Sq    RSS    AIC
- Hg    1     216.5 273114 1239.3
<none>               272897 1241.2
- Bfat  1    3562.9 276460 1241.4
- Wt    1    8022.2 280919 1244.0
- Ht    1   24064.8 296962 1253.2
- Sex   1   28050.9 300948 1255.5


Step:  AIC=1239.34
Ferr ~ Sex + Bfat + Ht + Wt

       Df Sum of Sq    RSS    AIC
<none>               273114 1239.3
- Bfat  1      3354 276467 1239.4
- Wt    1     10076 283189 1243.3
- Ht    1     28058 301172 1253.6
- Sex   1     33865 306979 1256.7

Call:
lm(formula = Ferr ~ Sex + Bfat + Ht + Wt, data = train)

Coefficients:
(Intercept)      Sexmale         Bfat           Ht           Wt
    348.098       60.305        1.360       -2.262        1.160

> |
```

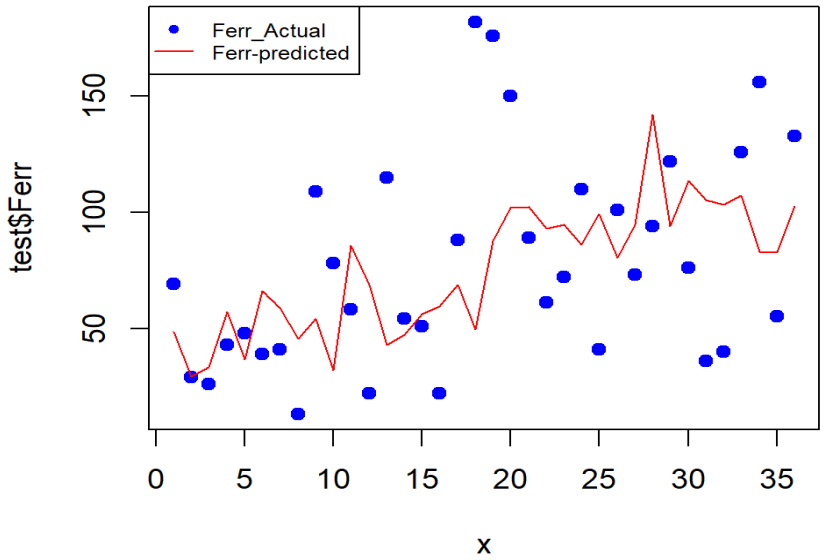
```
> # Finally set up a multiple  linear regression model where Plasma ferritin  is the dependent variable whic
h varies with independent variables: Sex+Bfat+Ht+Wt
> Final_Ferrmodel<- lm (data= train, Ferr ~ Sex+Bfat+Ht+Wt)
> summary(Final_Ferrmodel)

Call:
lm(formula = Ferr ~ Sex + Bfat + Ht + Wt, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-83.761 -25.104  -8.406  20.657 130.091

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 348.0980    81.4198   4.275 3.26e-05 ***
Sexmale      60.3052    13.4970   4.468 1.48e-05 ***
Bfat          1.3604     0.9675   1.406   0.1616
Ht           -2.2623     0.5563  -4.067 7.44e-05 ***
Wt            1.1598     0.4759   2.437   0.0159 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.19 on 161 degrees of freedom
Multiple R-squared:  0.2857,    Adjusted R-squared:  0.2679
F-statistic:  16.1 on 4 and 161 DF,  p-value: 4.168e-11

> |
```

The step () backward selects the Sex, Bfat, Ht and Wt as model selection which is:

Ferr = 348.0980 – 60.3052 sexmale –2.3604 Bfat-2.2623 Ht+ 1.598 Wt

```
> # predicting the Ferr variable from the final model
>
> predicted_Ferr_ <- predict(Final_Ferrmodel, test)
```

The plot displays the actual ferritin compared to the predicted ferritin

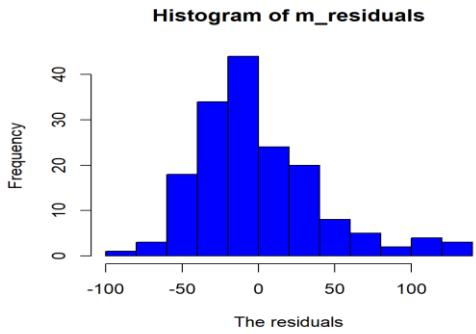


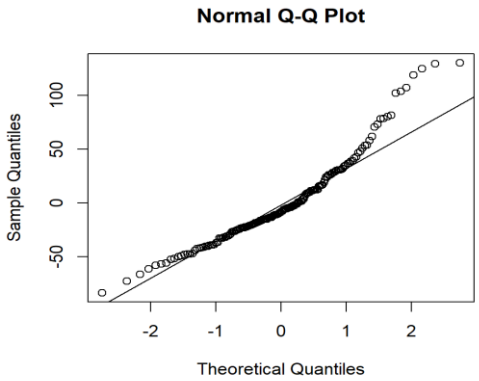
The model appears to be well-fit, but it is unable to accurately capture the outlier.

```
> # computing the final  ferr model performance metrics
> data.frame(R2 = R2(predicted_Ferr_, test$Ferr),RMSE = RMSE(predicted_Ferr_, test$Ferr), MAE = MAE(predicte
d_Ferr_, test$Ferr))
         R2     RMSE      MAE
1 0.0823939 45.01684 36.24485
>
```

```
> # Plot the  Ferr model residuals
> hist(m_residuals, xlab="The residuals",ylab="Frequency", col = "blue")
> # Get the final Ferr model residuals
> m_residuals = Final_Ferrmodel$residuals
> # Plot the  Ferr model residuals
> hist(m_residuals, xlab="The residuals",ylab="Frequency", col = "blue")
>
```

```
> # Plot the residuals
> qqnorm(m_residuals)
> # Plot the Q-Q line
> qqline(m_residuals)
>
```

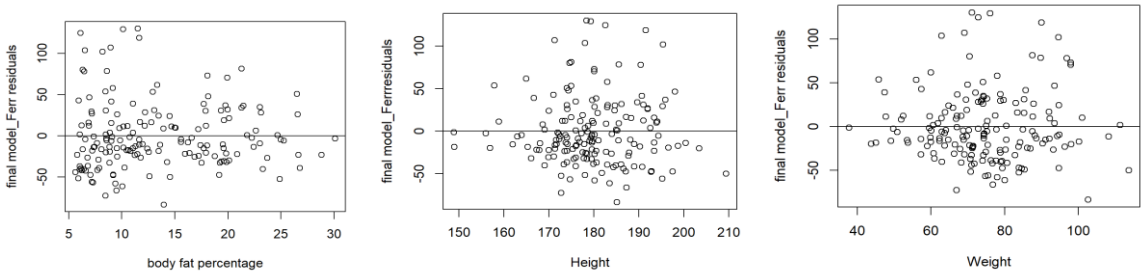




The residual looks skewed to the right, let test its normality

```
> # perform the Anderson-Darling test
> ad.test(m_residuals)

        Anderson-Darling normality test

data:  m_residuals
A = 3.0546, p-value = 1.07e-07

>
```

From the test, the residuals follow the normal distribution



There is no trend visible in the scatterplots of the residuals versus each covariate, suggesting a linear relationship.

So, the model that can predict plasma ferritin levels:

Ferr = 348.0980 – 60.3052 sexmale –2.3604 Bfat-2.2623 Ht+ 1.598 Wt

This model illustrates the calculation of Ferritin is influenced by factors: sex, body fat, height, and weight. It reflects real life

5. Conclusion

The study found that male athletes play similar sports to females, with netball being the most significant sport. There is evidence of gender-related differences in body mass index and body fate percentage between male and female athletes. There is also a significant difference in lean body mass between at least two sport groups. The gender factor and sport type are significant at the 5% significance level, but not when interacting.

The gender, body fat percentage, height, and weight factors are significant at the 5% significance level, but not when interacting.

The model accurately represents the reality that factors: sex, body fat, height, and weight all influence ferritin levels. The model explains 8% of response variable variation, but small and high R-squared values are not always problematic, as unexplainable variation is common in practice.

The model's conclusions can be significantly influenced by the sample size used in statistical analysis, suggesting the need for a larger sample size.

6. Reference

[1] Nora. Tanner. (2023). MATPMD1: Learning and Teaching: MATPMD1 - Statistics for Data Science (2023/4) (stir.ac.uk)