STATISTICAL ANALYSIS ON BODY FAT IN RELATION TO PERSONS BODY MEASUREMENTS

Detailed Project Report: -

BY- SHUBH MEHTA

BS21DMU039

Table of Contents:

Contents

Introduction

The body fat percentage (BFP) of a human or other living being is the total mass of fat divided by total body mass, multiplied by 100; body fat includes essential body fat and storage body fat. Essential is necessary to maintain life and reproductive functions. Excessive stored fat can be harmful to your health. Having too much stored fat can lead to chronic diseases and conditions.

The objective of this analysis is to predict the body fat percentage in a person based on his age and body measurements. This will allow us to predict their ***level of health***, which could help us identify different health issues the person might be having. We have a data set consisting of several people with their body structure measurements.

The data consist of following attributes:

1. Fat
2. Age
3. Weight
4. Height
5. BMI
6. Neck
7. Chest
8. Abdomen
9. Hip
10. Thigh
11. Knee
12. Ankle
13. Biceps
14. Forearm
15. Wrist

Fat is the dependant variable to be studied based on the input variables.

(There is no unit for Fat since its shown in percentage form)

Data Understanding

Sample Data:

The sample data is shown below:

```
> head(mydata)
   Fat Age Weight Height      BMI Neck Chest Abdomen   Hip Thigh Knee Ankle Biceps Forearm Wrist
1 12.3  23 154.25  67.75 3.360521 36.2  93.1    85.2  94.5  59.0 37.3  21.9   32.0    27.4  17.1
2  6.1  22 173.25  72.25 3.318926 38.5  93.6    83.0  98.7  58.7 37.3  23.4   30.5    28.9  18.2
3 25.3  22 154.00  66.25 3.508722 34.0  95.8    87.9  99.2  59.6 38.9  24.0   28.8    25.2  16.6
4 10.4  26 184.75  72.25 3.539230 37.4 101.8    86.4 101.2  60.1 37.3  22.8   32.4    29.4  18.2
5 28.7  24 184.25  71.25 3.629424 34.4  97.3   100.0 101.9  63.2 42.2  24.0   32.2    27.7  17.7
6 20.9  24 210.25  74.75 3.762821 39.0 104.5    94.4 107.8  66.0 42.0  25.6   35.7    30.6  18.8
```

- Top 6 observations from a dataset of 252 observations is shown above.
- Data isn't biased as people from age group of 22 to 81 were all included in this analysis.
- All body measurements are in standard units. (Weight in pounds etc)
- As per the NOIR classification (Nominal, Ordinal, Interval and Ratio classification) the data in dataset can be classified into **Interval data of continuous type.**
- NULL value test was performed on the dataset. **No NULL values** were present in the dataset

Key Statistics for Data:

Before we proceed let us find the key parameters of the data attribute:

```
> psych::describe(mydata)
        vars   n   mean    sd median trimmed   mad    min    max  range  skew kurtosis   se
Fat        1 252  19.15  8.37  19.20   19.05  9.27   0.00  47.50  47.50  0.14    -0.37 0.53
Age        2 252  44.88 12.60  43.00   44.44 11.86  22.00  81.00  59.00  0.28    -0.45 0.79
Weight     3 252 178.92 29.39 176.50  177.41 28.73 118.50 363.15 244.65  1.19     5.08 1.85
Height     4 252  70.15  3.66  70.00   70.27  2.97  29.50  77.75  48.25 -5.32    57.86 0.23
BMI        5 252   3.69  1.36   3.57    3.57  0.45   2.56  23.56  20.99 12.44   178.13 0.09
Neck       6 252  37.99  2.43  38.00   37.96  2.37  31.10  51.20  20.10  0.55     2.60 0.15
Chest      7 252 100.82  8.43  99.65  100.28  8.38  79.30 136.20  56.90  0.67     0.91 0.53
Abdomen    8 252  92.56 10.78  90.95   92.00 10.90  69.40 148.10  78.70  0.83     2.14 0.68
Hip        9 252  99.90  7.16  99.30   99.49  5.78  85.00 147.70  62.70  1.48     7.22 0.45
Thigh     10 252  59.41  5.25  59.00   59.17  4.60  47.20  87.30  40.10  0.81     2.55 0.33
Knee      11 252  38.59  2.41  38.50   38.50  2.22  33.00  49.10  16.10  0.51     0.99 0.15
Ankle     12 252  23.10  1.69  22.80   22.98  1.33  19.10  33.90  14.80  2.23    11.57 0.11
Biceps    13 252  32.27  3.02  32.05   32.24  2.89  24.80  45.00  20.20  0.28     0.44 0.19
Forearm   14 252  28.66  2.02  28.70   28.68  2.08  21.00  34.90  13.90 -0.22     0.80 0.13
Wrist     15 252  18.23  0.93  18.30   18.21  0.89  15.80  21.40   5.60  0.28     0.34 0.06
```

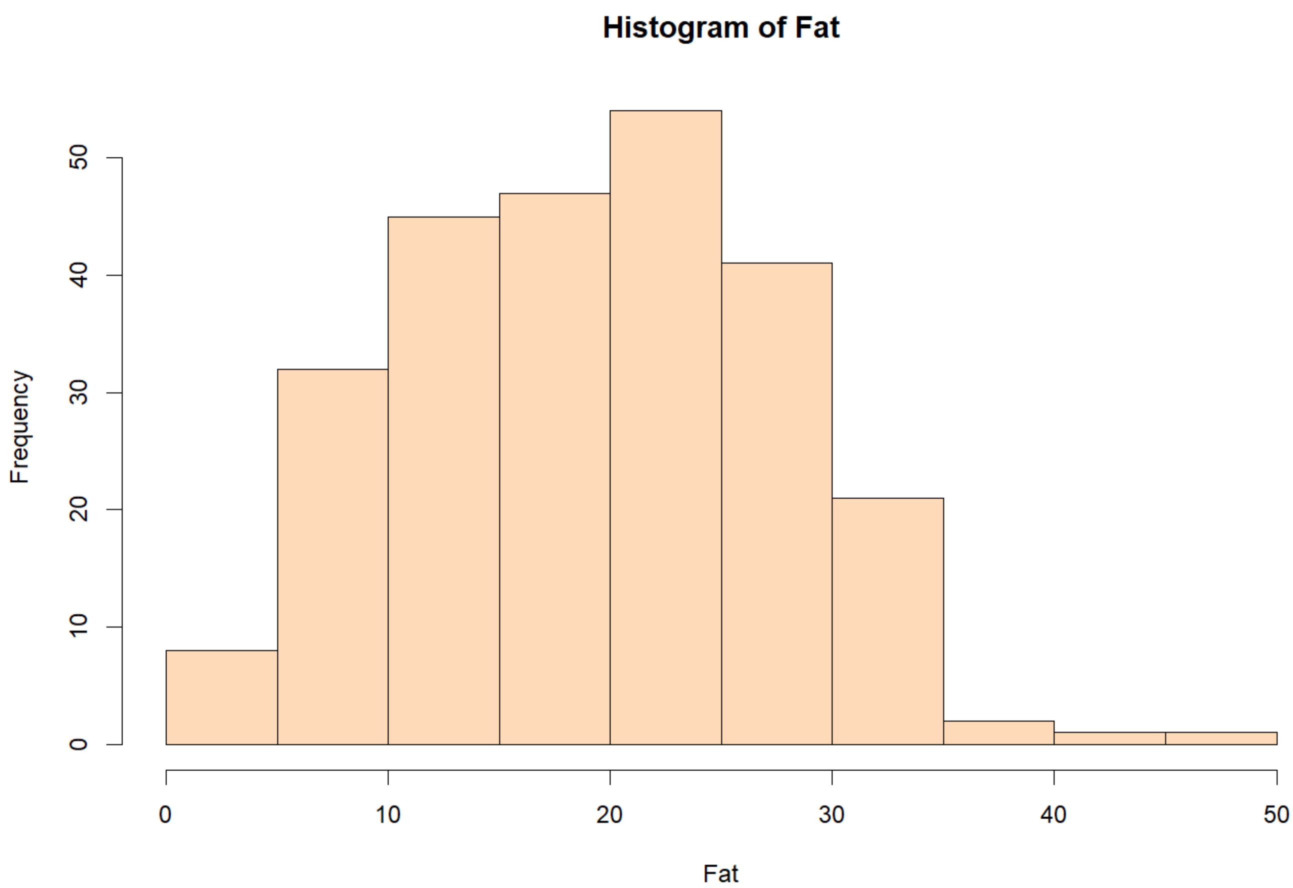
From the above table output:

Following key observations can be made:

1. Mean of Fat (19.15) is slightly less than median of Fat (19.20), indicating a negative skewness (0.14).
    a. We will check for outliers and make the mean closer to median
2. This skewness values conveys that data is fairly symmetrical. This also tells that tail of the left side of Fat distribution is longer or fatter than the tail on the right side.
3. BMI, Height, and Ankle have high kurtosis compared to other variables.

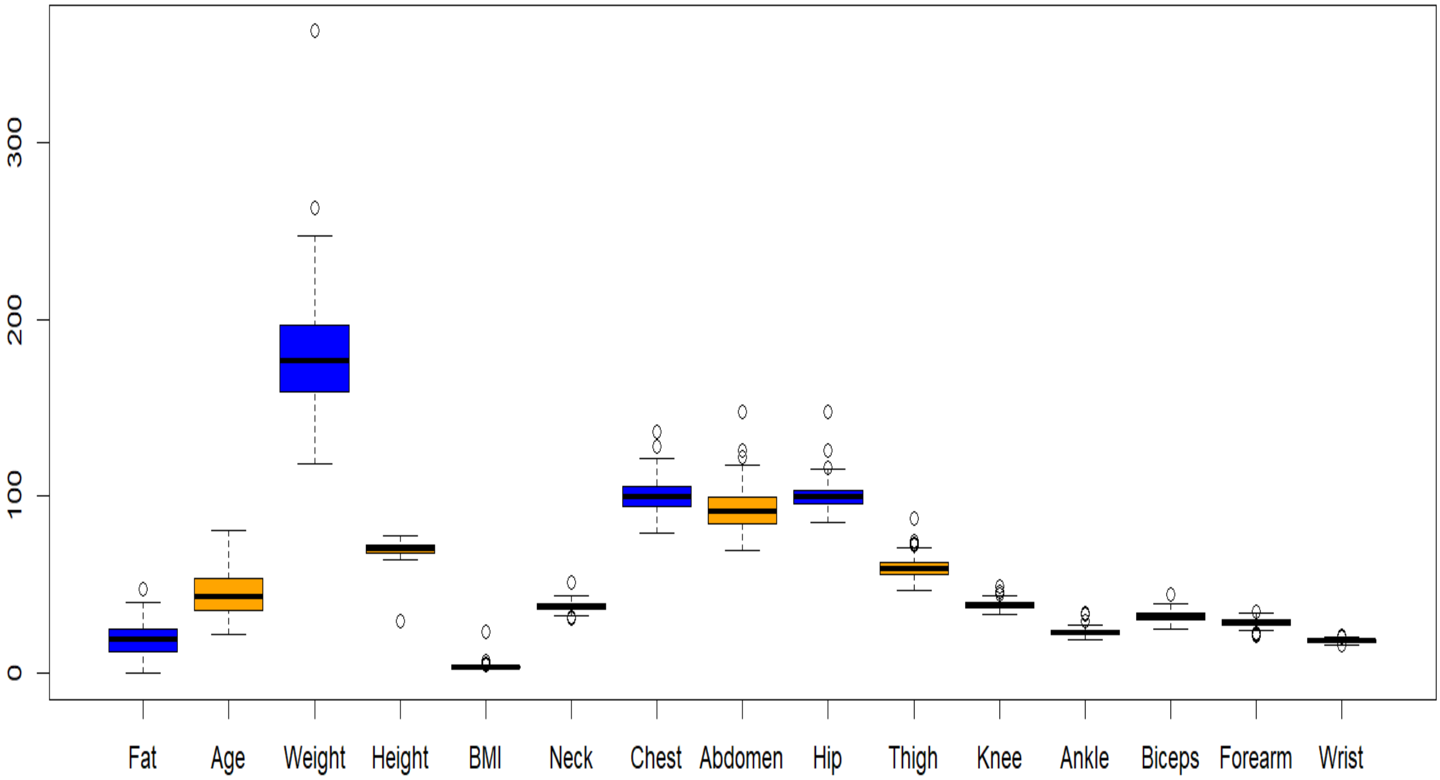
Histograms and Boxplots:

Let us draw some important plots to understand the distribution of our dependent variable:



The above Histogram shows that Fat is normally distributed with a mild negative skewness.

The box plot below shows how each attribute is classified and how much outliers are present.



There are very few outliers in the data set. After examining the dataset, it was found that for no parameter the outliers can be removed because to some extent there was correlation observed between the output (Fat) and for the outlier values. Hence the dataset remains same.

Correlation Scatterplot and Correlation Matrix:

- Correlation coefficient between two random variables X and Y, usually denoted by $r\ (X, Y)$ $or\ r_{XY}$ is a numerical measure of linear relationship between them and is defined as:

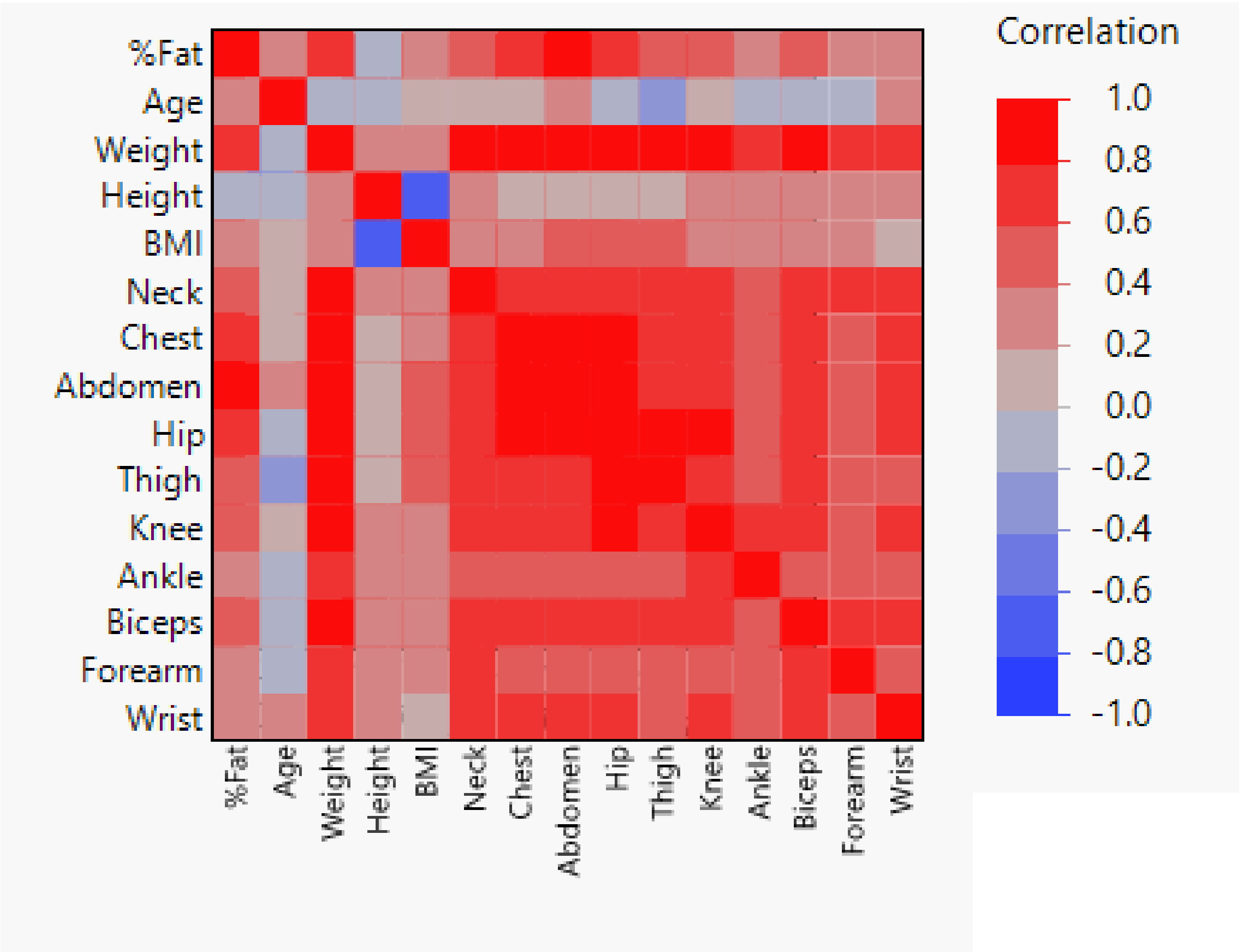
$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- r_{XY} provided a measure of linear relationship between X and Y.
- It is a measure of degree of relationship.
- Correlation Matrix is an n×n matrix (square, upper triangular or lower triangular consisting of a total of ‘n’ attributes) of correlation coefficient between all variables of the dataset taken 2 at a time.
- Correlation Scatterplot is exactly same as correlation matrix except that instead of having values of correlation coefficient, it has scatterplot for variables taken 2 at a time.

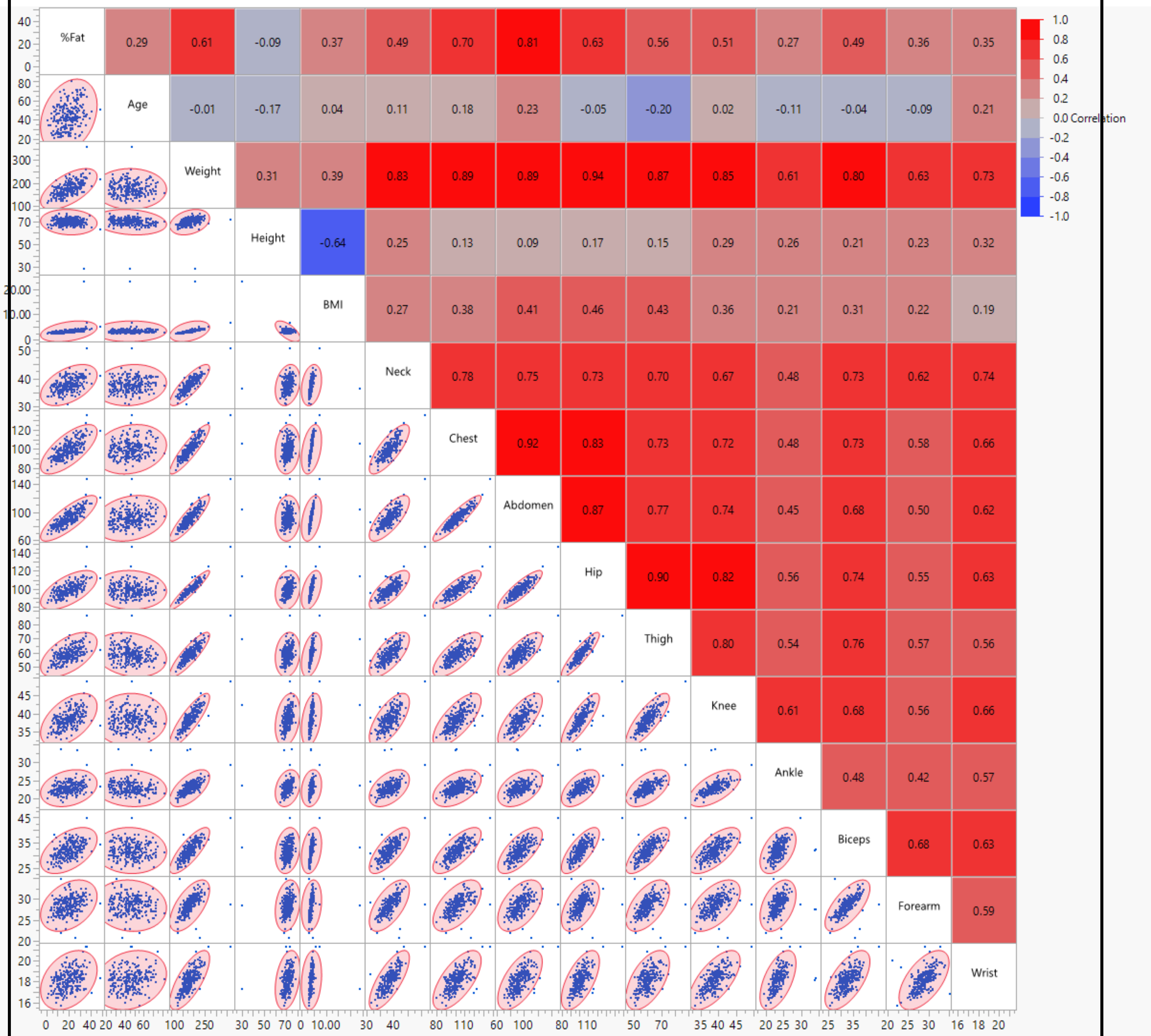
Correlation Matrix:

Row	%Fat	Age	Weight	Height	BMI	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
%Fat	1.000	0.291	0.612	-0.089	0.371	0.491	0.703	0.813	0.625	0.560	0.509	0.266	0.493	0.361	0.347
Age	0.291	1.000	-0.013	-0.172	0.040	0.114	0.176	0.230	-0.050	-0.200	0.018	-0.105	-0.041	-0.085	0.214
Weight	0.612	-0.013	1.000	0.308	0.391	0.831	0.894	0.888	0.941	0.869	0.853	0.614	0.800	0.630	0.730
Height	-0.089	-0.172	0.308	1.000	-0.638	0.254	0.135	0.088	0.170	0.148	0.286	0.265	0.208	0.229	0.322
BMI	0.371	0.040	0.391	-0.638	1.000	0.266	0.383	0.415	0.462	0.433	0.364	0.210	0.311	0.215	0.190
Neck	0.491	0.114	0.831	0.254	0.266	1.000	0.785	0.754	0.735	0.696	0.672	0.478	0.731	0.624	0.745
Chest	0.703	0.176	0.894	0.135	0.383	0.785	1.000	0.916	0.829	0.730	0.719	0.483	0.728	0.580	0.660
Abdome	0.813	0.230	0.888	0.088	0.415	0.754	0.916	1.000	0.874	0.767	0.737	0.453	0.685	0.503	0.620
Hip	0.625	-0.050	0.941	0.170	0.462	0.735	0.829	0.874	1.000	0.896	0.823	0.558	0.739	0.545	0.630
Thigh	0.560	-0.200	0.869	0.148	0.433	0.696	0.730	0.767	0.896	1.000	0.799	0.540	0.761	0.567	0.559
Knee	0.509	0.018	0.853	0.286	0.364	0.672	0.719	0.737	0.823	0.799	1.000	0.612	0.679	0.556	0.665
Ankle	0.266	-0.105	0.614	0.265	0.210	0.478	0.483	0.453	0.558	0.540	0.612	1.000	0.485	0.419	0.566
Biceps	0.493	-0.041	0.800	0.208	0.311	0.731	0.728	0.685	0.739	0.761	0.679	0.485	1.000	0.678	0.632
Forearm	0.361	-0.085	0.630	0.229	0.215	0.624	0.580	0.503	0.545	0.567	0.556	0.419	0.678	1.000	0.586
Wrist	0.347	0.214	0.730	0.322	0.190	0.745	0.660	0.620	0.630	0.559	0.665	0.566	0.632	0.586	1.000

Colour Map for Coefficient Correlations:



Correlation Scatterplot:



- Above plot has density ellipses shaded at α=0.01, along with Correlation Coefficients as heat map.
- The above plot conveys strong and positive relationship between multiple variables (dependent and independent).
- Attribute *'Weight'* seems to have highest correlation with other variables.
- Correlations make it possible to use the value of one variable to predict the value of another, hence reducing number of attributes used in model.
- Though a correlation can indicate the possibility of a cause-effect relationship, it certainly does not prove the direction of the influence.

Splitting Data for Training and Testing:

Now data is split into training and testing data with following commands:

```
install.packages('caTools')
library(caTools)

set.seed(123)
split=sample.split(mydata,SplitRatio=0.70)

train_data<-subset(mydata,split==T) #Created Training Data for Analysis
test_data<-subset(mydata,split==F) #Created Testing Data for Final Verification
```

Data is split in 70:30 ratio, 70 % data is considered for training and remaining 30% data is considered for testing.

test_data	85 obs. of 15 variables
train_data	167 obs. of 15 variables

- As seen above, 30% of 252 observation = 85 is for testing data, rest is for training data.
- For future Regression Analysis, we will use the training data.
- The models ahead and the final model will be based on Training data only.
- In the end we will check with testing data the accuracy of our model.

Regression Analysis

Now that we have visualized the data and seen their respective correlations, we take this data for our regression model

- A multiple linear regression model takes the form

$$\boxed{Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \varepsilon}$$

- For Hypothesis testing and the setting of confidence limits, we also assume that ε is normally distributed.
- Hence "**Σ $\boldsymbol{\varepsilon}$ =0** ".
- The linearity of the model above is defined with respect to the regression coefficients.

X variables β_1, β_2 $etc.$ … in the test are as follows:

1. Age
2. Weight
3. Height
4. BMI
5. Neck
6. Chest
7. Abdomen
8. Hip
9. Thigh
10. Knee
11. Ankle
12. Biceps
13. Forearm
14. Wrist

Y variable for the model is:

1. Fat

Regression Model – Trial 1:

Fitting Multiple Linear Regression with all X variables included:

```
> # Fitting Multiple Linear Regression
> model = lm(Fat ~ Age+Height+Weight+BMI+Neck+Chest+Abdomen+Hip+Thigh+Knee+Ankle+Biceps+Forearm+Wrist)
> summary(model)
Call:
lm(formula = Fat ~ Age + Height + Weight + BMI + Neck + Chest + 
    Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + 
    Wrist)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.1415  -2.8555  -0.1037   3.1865  10.0913 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -24.246009  24.500167  -0.990  0.32337    
Age           0.062253   0.032412   1.921  0.05598 .  
Height        0.001679   0.224781   0.007  0.99405    
Weight       -0.102668   0.067226  -1.527  0.12805    
BMI           0.191510   0.545920   0.351  0.72605    
Neck         -0.456368   0.236404  -1.930  0.05474 .  
Chest        -0.014460   0.102884  -0.141  0.88835    
Abdomen       0.963379   0.090015  10.702  < 2e-16 ***
Hip          -0.203456   0.146643  -1.387  0.16661    
Thigh         0.244859   0.146764   1.668  0.09656 .  
Knee         -0.002546   0.247694  -0.010  0.99181    
Ankle         0.180612   0.222676   0.811  0.41812    
Biceps        0.188934   0.172711   1.094  0.27509    
Forearm       0.449118   0.199669   2.249  0.02541 *  
Wrist        -1.617543   0.536007  -3.018  0.00282 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.313 on 237 degrees of freedom
Multiple R-squared:  0.7492,    Adjusted R-squared:  0.7344 
F-statistic: 50.56 on 14 and 237 DF,  p-value: < 2.2e-16
```

This model has R^2=0.7492, indicating that it’s a good model but can further improved

From the above summary for our model, we can derive the Multiple Regression Equation to be:

$$Fat = -24.246009 + 0.062253 * Age + 0.001679 * Height - 0.102668 * Weight + 0.191510 * BMI - 0.456368 * Neck - 0.014460 * Chest + 0.963379 * Abdomen - 0.203456 * Hip + 0.244859 * Thigh - 0.002546 * Knee + 0.180612 * Ankle + 0.188934 * Biceps + 0.449118 * Forearm - 1.617534 * Wrist$$

Multiple Regression Equation with all variables.

Since there are many variables in one equation, we perform ‘P-TEST’ and reduce the dimension for this equation

Criteria:

We perform *anova* on our model to find p-value for each variable. The variables with P value < 0.05 hold good for predicting the output, hence we keep them. We remove rest of the variables from equation.

Anova – for Model 1:

```
> # Regression Model Performance
> anova(model)
Analysis of Variance Table

Response: Fat
           Df Sum Sq Mean Sq  F value    Pr(>F)    
Age         1 1493.3  1493.3  80.2675 < 2.2e-16 ***
Height      1   28.2    28.2   1.5166 0.2193594    
Weight      1 7689.1  7689.1 413.3042 < 2.2e-16 ***
BMI         1  509.4   509.4  27.3824 3.674e-07 ***
Neck        1  293.0   293.0  15.7513 9.578e-05 ***
Chest       1  287.3   287.3  15.4412 0.0001117 ***
Abdomen     1 2461.2  2461.2 132.2923 < 2.2e-16 ***
Hip         1   22.8    22.8   1.2229 0.2699172    
Thigh       1  111.7   111.7   6.0036 0.0150006 *  
Knee        1    0.2     0.2   0.0083 0.9276089    
Ankle       1    0.2     0.2   0.0081 0.9285046    
Biceps      1   47.7    47.7   2.5660 0.1105160    
Forearm     1   56.4    56.4   3.0302 0.0830249 .  
Wrist       1  169.4   169.4   9.1069 0.0028245 ** 
Residuals 237 4409.2    18.6                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The ANOVA table shows the P and F value for each variable along with their degree of freedom.
- Based on above table, we can say that since variables [*Age, Weight, BMI, Neck, Chest, Abdomen, Thigh, Wrist*] have P value < 0.05, apart from these variables, we can remove other ones.
- We remove variables with height having p>0.05 as it indicates that it has no effect on y variable so based on that we are rejecting the null hypothesis and, hence discarding that variable.

Let us proceed further and improve the model in the following steps:

Model 2: (After removing the insignificant attributes)

X variables $\beta_1, \beta_2\ etc. \ldots$ in the test are as follows:

1. Age
2. Weight
3. ~~Height~~
4. BMI
5. Neck
6. Chest
7. Abdomen
8. ~~Hip~~
9. Thigh
10. ~~Knee~~
11. ~~Ankle~~
12. ~~Biceps~~
13. ~~Forearm~~
14. Wrist

Y variable for the model is:

1. Fat

Regression Model – Trial 2:

Fitting Multiple Linear Regression with updated X variables:

Summary of new model:

```
> model_m = lm(Fat~ Age+Weight+BMI+Neck+Chest+Abdomen+Thigh+Wrist)
> summary(model_m)

Call:
lm(formula = Fat ~ Age + Weight + BMI + Neck + Chest + Abdomen + 
    Thigh + Wrist)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.4782  -3.0092  -0.2148   3.2706  10.3940 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -31.66028    9.86533  -3.209  0.00151 ** 
Age           0.05448    0.03134   1.739  0.08335 .  
Weight       -0.11998    0.03694  -3.248  0.00133 ** 
BMI           0.12818    0.23193   0.553  0.58100    
Neck         -0.28860    0.22371  -1.290  0.19826    
Chest         0.06464    0.09474   0.682  0.49568    
Abdomen       0.87635    0.08138  10.769  < 2e-16 ***
Thigh         0.24312    0.12351   1.968  0.05016 .  
Wrist        -1.19297    0.51178  -2.331  0.02057 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.37 on 243 degrees of freedom
Multiple R-squared:  0.736,	Adjusted R-squared:  0.7273 
F-statistic: 84.67 on 8 and 243 DF,  p-value: < 2.2e-16
```

From the above summary for our model, we can derive the New Multiple Regression Equation to be:

$$
\begin{aligned}
Fat = {} & -31.66028 + 0.05448 * Age - 0.11998 * Weight \\
& + 0.12818 * BMI - 0.28860 * Neck - 0.06464 \\
& * Chest + 0.87635 * Abdomen + 0.24312 \\
& * Thigh - 1.19297 * Wrist
\end{aligned}
$$

New Multiple Regression Equation

Anova for Model 2:

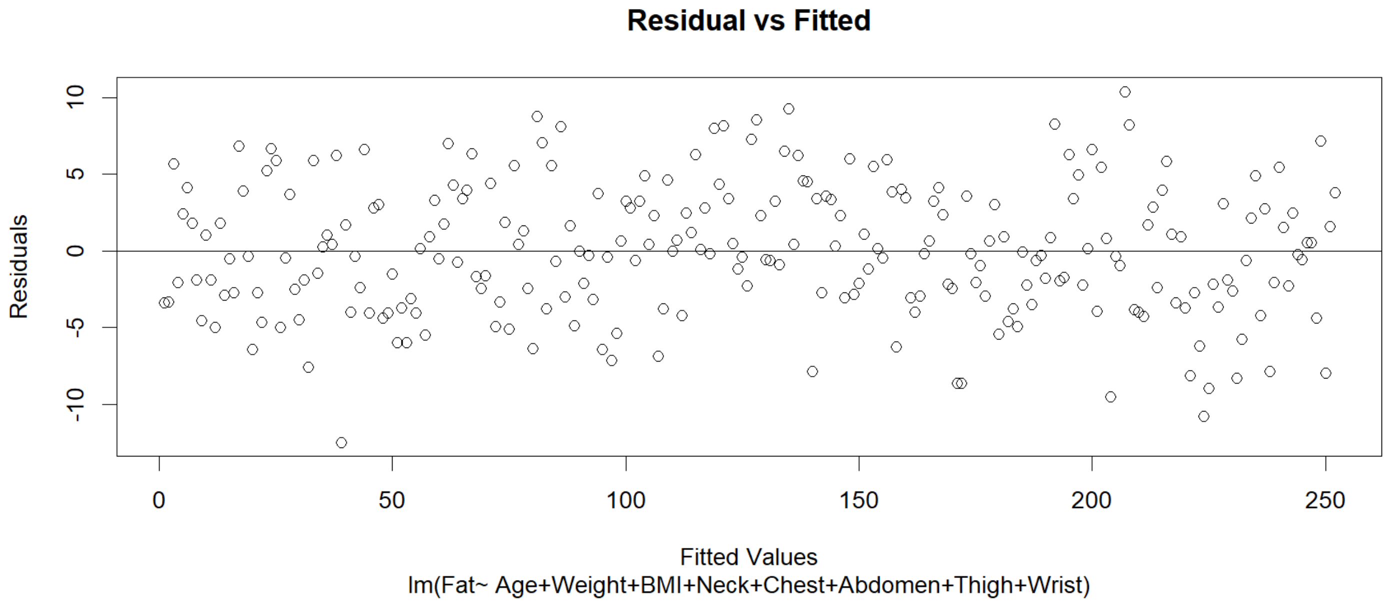
```
> # Regression Model Performance
> anova(model_m)
Analysis of Variance Table

Response: Fat
           Df Sum Sq Mean Sq  F value    Pr(>F)    
Age         1 1493.3  1493.3  78.1834 < 2.2e-16 ***
Weight      1 6674.3  6674.3 349.4419 < 2.2e-16 ***
BMI         1  293.2   293.2  15.3496 0.0001162 ***
Neck        1  130.3   130.3   6.8213 0.0095684 ** 
Chest       1 1046.7  1046.7  54.7989 2.179e-12 ***
Abdomen     1 3117.4  3117.4 163.2171 < 2.2e-16 ***
Thigh       1   78.7    78.7   4.1214 0.0434327 *  
Wrist       1  103.8   103.8   5.4336 0.0205712 *  
Residuals 243 4641.3    19.1                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- As seen, the P value < 0.05 for the above variables indicating that Model-2 holds good for predicting the output.
- Hence, the above results convey that all the p values are significant.

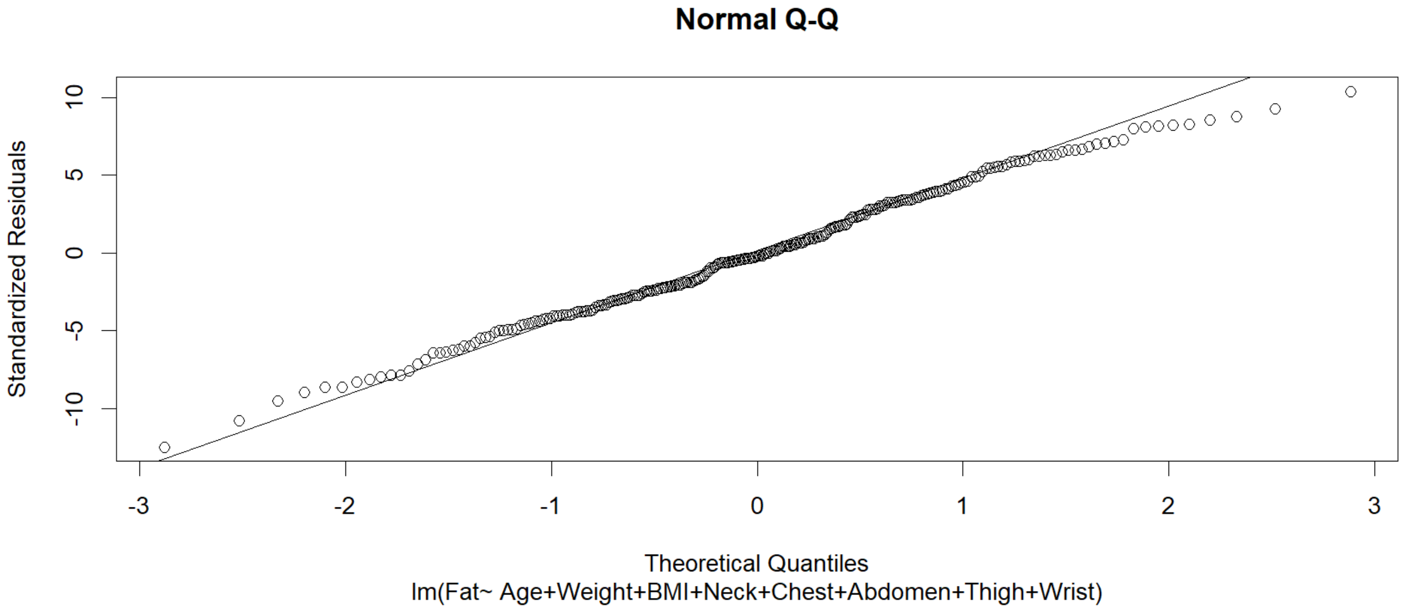
Hence, now that our final model is ready with the final Regression Equation, we can do *‘Residual Analysis’* (draw QQ plots and Residuals Plots).

Residual Analysis

Residual Plot:



Normal Q-Q Plot:



Inference:

- Initially we had assumed that Residuals followed Normal Distribution, and from both residual vs fitted plot and Normal QQ plot, we can see the assumption that data is linearly distributed is correct.
- QQ Line indicates that residuals follow normal distribution as the points lie very near to the line.
- Though, there are few outliers, but their effect is a bit significant, hence we don't remove them.

Hypothesis Testing on Model 2 (final model) outcome:

Null Hypothesis:

$$H_0: \beta_0 = \beta_1 = \cdots = \beta_{k-1} = 0$$

$$H_1: \beta_j \neq 0, for\ atleast\ one\ j$$

Anova Output:

```
> # Regression Model Performance
> anova(model_m)
Analysis of Variance Table

Response: Fat
           Df Sum Sq Mean Sq  F value    Pr(>F)    
Age         1 1493.3  1493.3  78.1834 < 2.2e-16 ***
Weight      1 6674.3  6674.3 349.4419 < 2.2e-16 ***
BMI         1  293.2   293.2  15.3496 0.0001162 ***
Neck        1  130.3   130.3   6.8213 0.0095684 ** 
Chest       1 1046.7  1046.7  54.7989 2.179e-12 ***
Abdomen     1 3117.4  3117.4 163.2171 < 2.2e-16 ***
Thigh       1   78.7    78.7   4.1214 0.0434327 *  
Wrist       1  103.8   103.8   5.4336 0.0205712 *  
Residuals 243 4641.3    19.1                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the above results show all the p values are significant. We can reject the NULL hypothesis. Therefore, **"Model can be used for prediction".**

Prediction for Test Data

1) Using Test Data to make predictions

```
predictions=predict(model_m,test_data)
```

- This line of R code predicts the value for testing data, which was divided in 70/30 ratio before starting regression analysis.
- It fits the test data into our final model (Model 2) and stores the predictions as a variable *'predictions'*.

2) Creating a file containing Actual and Predicted values combined

```
actuals_preds <- data.frame(cbind(actuals=test_data$Fat, predicteds=predictions))
head(actuals_preds)
```

- This code creates a variable which stores *'Actual'* and 'Predicted' values for the testing data, by the name *'actual_preds'*, together for seeing the accuracy of model.
- Below are some top observations of this variables.

```
> actuals_preds <- data.frame(cbind(actuals=test_data$Fat, predicteds=predictions))
> head(actuals_preds)
   actuals predicteds
2      6.1   9.413117
4     10.4  12.447039
5     28.7  26.253035
8     12.4  14.260468
11     7.1   8.997783
17    29.0  22.160649
```

- Note: - The numbers [2,4,5,8,11,17] show that splitting of data was done randomly.

3) Finding Accuracy of our Model:

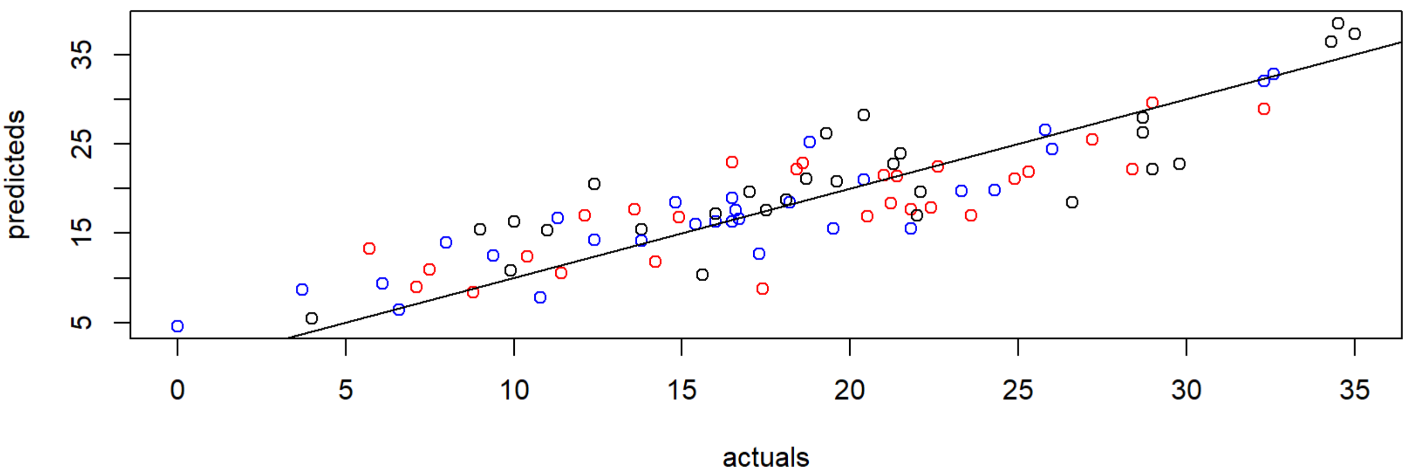
```
correlation_accuracy <- cor(actuals_preds)
head(correlation_accuracy)
```

- This code makes a matrix similar to confusion matrix, which tells us the accuracy of the model.

```
> correlation_accuracy <- cor(actuals_preds)
> head(correlation_accuracy)
             actuals predicteds
actuals    1.0000000  0.8625479
predicteds 0.8625479  1.0000000
```

- The above image shows that our model has an accuracy of **<u>86.255%</u>.**

Actual vs Predicted Graph:



- The line drawn is y=x, which is taken as we want to see the distribution of point when actual=predicted.
- We notice that maximum points near to the line, hence again indicating that our model is pretty good.

References

[1] Kaya Uyanık, Gülden & Güler, Neşe. (2013). A Study on Multiple Linear Regression Analysis. Procedia - Social and Behavioral Sciences. 106. 234–240. 10.1016/j.sbspro.2013.12.027.

[2] Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. The Review of Economic and Statistics, 92-107.

[3] Turóczy Zsuzsannaa, Liviu Mariana. (2012). Multiple regression analysis of performance indicators in the ceramic industry. Procedia Economics and Finance 3 (2012) 509 – 514

[4] Yunus Koloğlu, Hasan Birinci, Sevde Ilgaz Kanalmaz, Burhan Özyılmaz. (2018). A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position.

→ Have used R, SAS JMP and Excel to create and plot tables and graphs.