Student ID: 1920253
Student Name: Chan, Hsin-Ju

7CCMMS61T Statistics for Data Analysis Coursework

Question 1

Exploratory Data Analysis

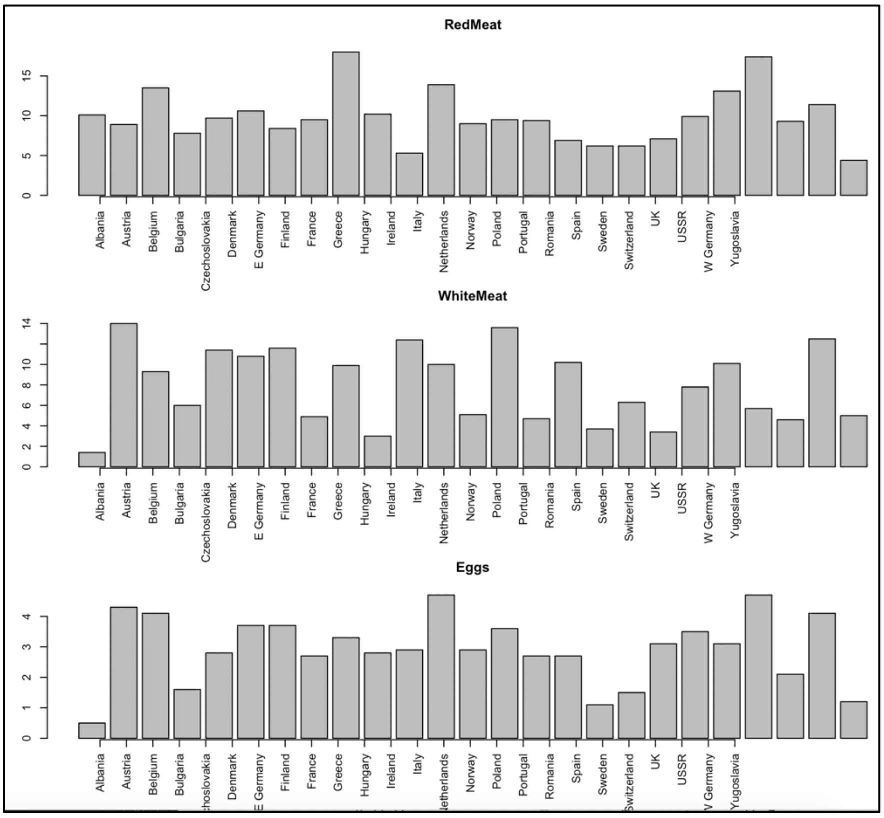
(a) For each variable, calculate appropriate summary statistics to show the level and spread of the data (one statistic for each is enough).

```
> sd(RedMeat)  #3.347078
[1] 3.347078
> var(WhiteMeat)   #13.6423
[1] 13.64623
> range(Eggs)  #0.5 4.7
[1] 0.5 4.7
> IQR(Milk)    #12.2
[1] 12.2
> min(Fish)  #0.2
[1] 0.2
> max(Cereals) #56.7
[1] 56.7
> mean(Starch)  #4.276
[1] 4.276
> median(Nuts) #2.4
[1] 2.4
> quantile(Fr.Veg,probs = 0.25)  #2.9
25%
2.9
```

```
> summary(protein)
           Country       RedMeat          WhiteMeat           Eggs             Milk
Albania       : 1   Min.   : 4.400   Min.   : 1.400   Min.   :0.500   Min.   : 4.90
Austria       : 1   1st Qu.: 7.800   1st Qu.: 4.900   1st Qu.:2.700   1st Qu.:11.10
Belgium       : 1   Median : 9.500   Median : 7.800   Median :2.900   Median :17.60
Bulgaria      : 1   Mean   : 9.828   Mean   : 7.896   Mean   :2.936   Mean   :17.11
Czechoslovakia: 1   3rd Qu.:10.600   3rd Qu.:10.800   3rd Qu.:3.700   3rd Qu.:23.30
Denmark       : 1   Max.   :18.000   Max.   :14.000   Max.   :4.700   Max.   :33.70
(Other)       :19
      Fish            Cereals          Starch           Nuts            Fr.Veg
Min.   : 0.200   Min.   :18.60   Min.   :0.600   Min.   :0.700   Min.   :1.400
1st Qu.: 2.100   1st Qu.:24.30   1st Qu.:3.100   1st Qu.:1.500   1st Qu.:2.900
Median : 3.400   Median :28.00   Median :4.700   Median :2.400   Median :3.800
Mean   : 4.284   Mean   :32.25   Mean   :4.276   Mean   :3.072   Mean   :4.136
3rd Qu.: 5.800   3rd Qu.:40.10   3rd Qu.:5.700   3rd Qu.:4.700   3rd Qu.:4.900
Max.   :14.200   Max.   :56.70   Max.   :6.500   Max.   :7.800   Max.   :7.900
```

(b) For each variable, plot the data in a suitable way to illustrate the level and the spread.

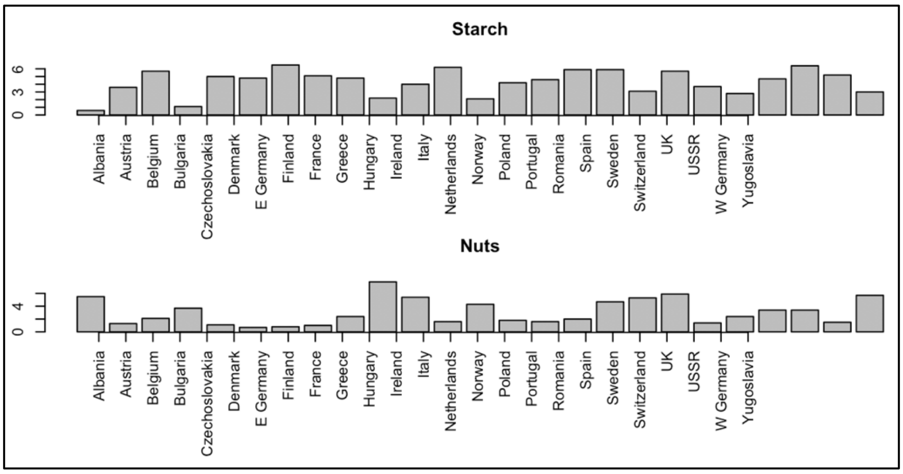
The barplot below shows the consumption of RedMeat, WhiteMeat, Eggs in different countries.

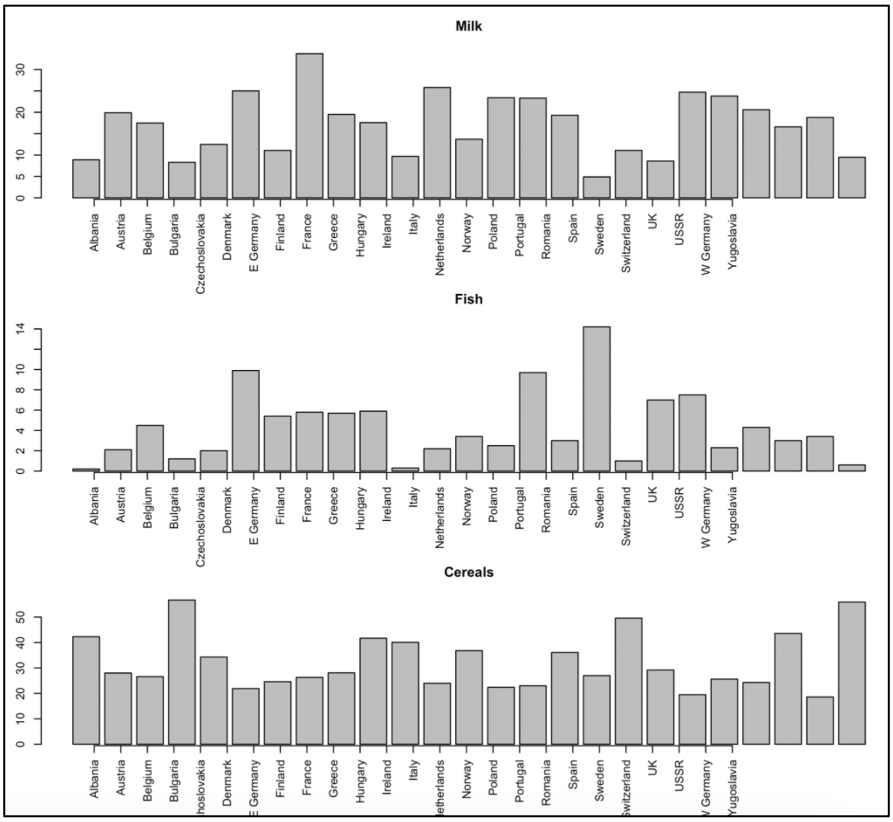



```
##(b)##
plot(protein,col=Country)
par(family=("Heiti TC Light"))
par(mfrow=c(3,1))
## RedMeat, WhiteMeat, Eggs ##
for (i in 2:4){
  barplot(protein[,i], main = names(protein)[i])
  countryNames <- as.vector(Country)
  axis(1,at=1:25,labels=countryNames, las=2)}
## Milk, Fish, Cereals ##
for (i in 5:7){
  barplot(protein[,i], main = names(protein)[i])
  countryNames <- as.vector(Country)
  axis(1,at=1:25,labels=countryNames, las=2)}
## Starch, Nuts ##
for (i in 8:9){
  barplot(protein[,i], main = names(protein)[i])
  countryNames <- as.vector(Country)
  axis(1,at=1:25,labels=countryNames, las=2)}
```

Student ID: 1920253
Student Name: Chan, Hsin-Ju

The barplot on the right shows the consumption of Starch, Nuts in different countries.





The barplot on the left shows the consumption of Milk, Fish, Cereals in different countries.

(c) Calculate a summary statistic to show the association of the consumption of fruit and vegetables with each of the other food categories.

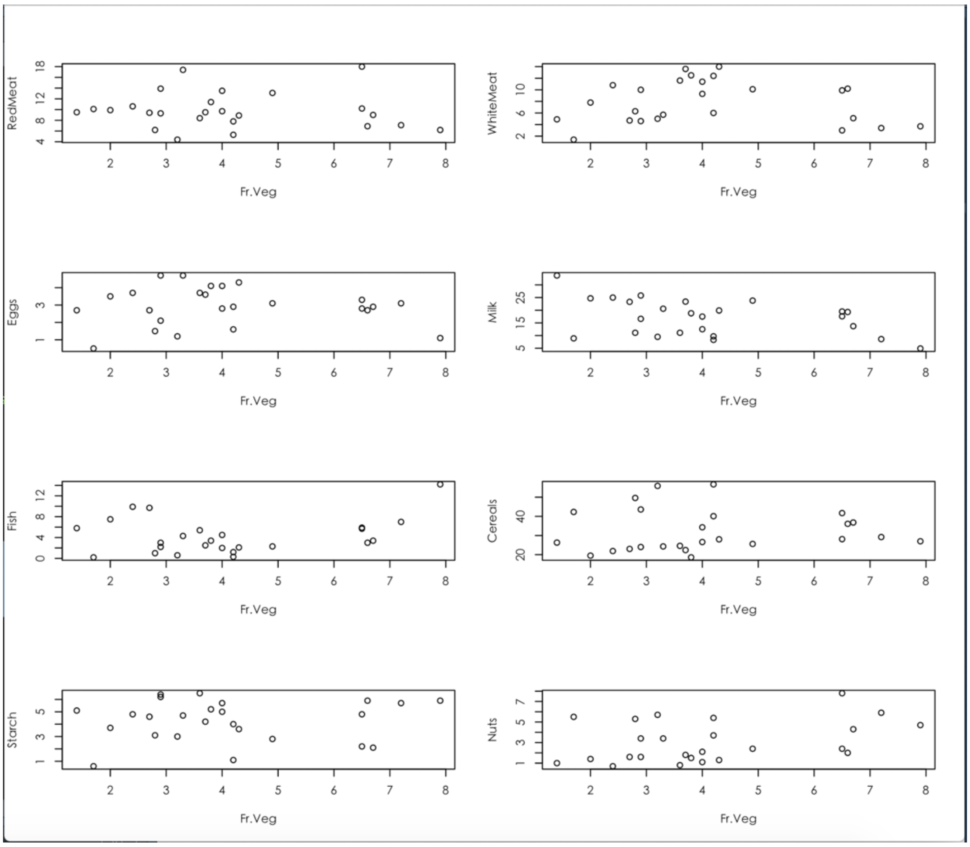
```
> ##(c)##
> cor(Fr.Veg,protein[,2:9])
        RedMeat  WhiteMeat        Eggs       Milk      Fish    Cereals     Starch      Nuts
[1,] -0.07422123 -0.0613167 -0.04551755 -0.4083641 0.2661387 0.04654808 0.08440956 0.3749697
```

The picture shows the correlation between fruit and vegetables and each of the other food categories.

(d) Show a plot illustrating the association of the consumption of fruit and vegetables with each of the other food categories.

```
##(d)##
par(mfrow=c(4,2))
fd_cat <- as.vector(2:9)
for (k in fd_cat){
  plot(Fr.Veg, protein[,k],ylab = names(protein)[k])}
```

Student ID: 1920253
Student Name: Chan, Hsin-Ju



The plot shows the consumption of fruit and vegetables with each of the other food categories.

Inference

(e) Provide confidence intervals at level 95% for the mean consumption of each category of food.

```
> ##(e)##
> for (i in fd_cat){
+   print(paste("95% Confidence interval for", names(protein)[i]))
+   print(t.test(protein[,i],conf.level = 0.95)$conf.int)
+ }
[1] "95% Confidence interval for RedMeat"
[1]  8.446394 11.209606
attr(,"conf.level")
[1] 0.95
[1] "95% Confidence interval for WhiteMeat"
[1] 6.371158 9.420842
attr(,"conf.level")
[1] 0.95
[1] "95% Confidence interval for Eggs"
[1] 2.474671 3.397329
attr(,"conf.level")
[1] 0.95
[1] "95% Confidence interval for Milk"
[1] 14.17903 20.04497
attr(,"conf.level")
[1] 0.95
[1] "95% Confidence interval for Fish"
[1] 2.879503 5.688497
attr(,"conf.level")
[1] 0.95
[1] "95% Confidence interval for Cereals"
[1] 27.71783 36.77817
attr(,"conf.level")
[1] 0.95
[1] "95% Confidence interval for Starch"
[1] 3.601483 4.950517
attr(,"conf.level")
[1] 0.95
[1] "95% Confidence interval for Nuts"
[1] 2.252351 3.891649
attr(,"conf.level")
[1] 0.95
```

(f) Carry out the appropriate test of hypothesis to check if the average consumption of starch is larger than the average consumption of nuts. Also check if the assumptions behind this test are reasonable in this case.

Student ID: 1920253
Student Name: Chan, Hsin-Ju

```
> t.test(Starch,Nuts,var.equal = T)

	Two Sample t-test

data:  Starch and Nuts
t = 2.3409, df = 48, p-value = 0.02344
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1698872 2.2381128
sample estimates:
mean of x mean of y 
    4.276     3.072 
```

The average consumption of starch (4.276) is larger than the average consumption of nuts (3.072).

T = 2.3409 ; p-value:0.02344 < 0.05

This shows the consumption of Starch and Nuts has significant difference.

Question 2

Exploratory Data Analysis

(a) State the scaling of each of the above variables.

```
> z
          Length       Width   Thickness     B.Width     J.Width     H.Length      Weight
 [1,] -0.51277250 -1.21732249 -0.96088667 -0.83146927 -1.75816080 -0.451242516 -0.96096335
 [2,] -0.69336022 -0.90702460 -0.96088667          NA -0.62373234 -0.127126003 -0.74703865
 [3,] -0.92890943 -1.12035440 -0.76497774 -0.56004938 -1.50199953 -1.298931856 -0.96096335
 [4,] -0.70906350 -1.15914164 -0.63437179 -0.08506457 -1.35562167 -1.273999817 -0.86588571
 [5,] -1.47067260 -0.96520546 -2.13634028 -0.39041195 -1.53859400 -1.124407580 -1.26996570
 [6,] -0.59128890 -1.02338631 -2.07103730 -0.35648446 -1.42881060 -0.600834752 -1.10357982
 [7,] -0.70906350 -0.26703520 -0.89558370 -0.69575932 -1.46540507 -1.448524092 -0.88965512
 [8,] -0.06522901 -0.65490757 -0.24255392  0.32206527 -0.73351574 -1.049611462 -0.34295865
 [9,] -0.12804213 -0.88763098 -0.04664498  0.18635532 -1.31902720 -1.124407580 -0.60442218
[10,] -1.23512339 -1.21732249 -1.41800752          NA -1.06286594 -0.476174555 -1.15111864
[11,] -1.32934307 -1.46943953 -1.35270454 -0.86539676 -1.46540507 -0.426310476 -1.22242688
[12,] -0.55988234 -1.45004591 -0.43846285          NA -0.80670468 -0.226854161 -0.67573041
[13,] -1.24297503 -1.06217355 -1.54861348 -0.45826692 -0.91648807 -0.725494949 -1.05604100
[14,] -0.59128890 -1.04277993 -1.41800752  0.15242784 -0.40416555 -0.725494949 -0.91342453
```

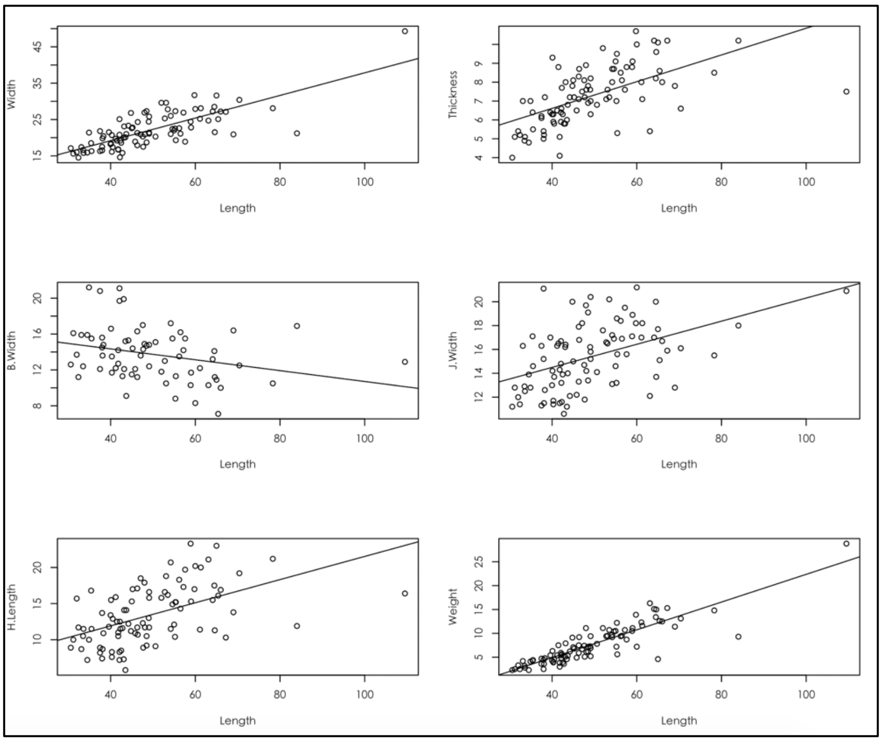
```
> z <- scale(DP[,3:9])
> summary(z)
     Length             Width            Thickness            B.Width             J.Width        
 Min.   :-1.4707   Min.   :-1.4694   Min.   :-2.13634   Min.   :-2.25642   Min.   :-1.75816  
 1st Qu.:-0.6659   1st Qu.:-0.6840   1st Qu.:-0.66702   1st Qu.:-0.69576   1st Qu.:-0.83415  
 Median :-0.1752   Median :-0.1895   Median :-0.04664   Median :-0.05114   Median : 0.05327  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000  
 3rd Qu.: 0.5079   3rd Qu.: 0.5960   3rd Qu.: 0.63904   3rd Qu.: 0.59349   3rd Qu.: 0.61133  
 Max.   : 4.7243   Max.   : 5.2795   Max.   : 2.23896   Max.   : 2.52735   Max.   : 2.12085  
                                                        NA's   :22         NA's   :1         
    H.Length           Weight       
 Min.   :-1.8973   Min.   :-1.2700  
 1st Qu.:-0.7255   1st Qu.:-0.7352  
 Median :-0.2269   Median :-0.2003  
 Mean   : 0.0000   Mean   : 0.0000  
 3rd Qu.: 0.7206   3rd Qu.: 0.5722  
 Max.   : 2.4658   Max.   : 5.0289  
```

The picture on the left side shows the value of each variables after scaling.

The picture on the right side shows the summary of the variables after scaling.

There is no Blade.Sh, Should.Sh, Should.Or, Haft.Sh, Haft.Or because these variables are shown in categorization.

(b) First consider the variable Length. Represent graphically the relationship between Length and the other variables and describe any interesting patterns.



These plots show that **Length** has **positive correlation** with **Width, Thickness, J.Width, H.Length and Weight**.

It is obvious that **Width, Thickness and Weight** has **highly** positive correlation with Length than the other two variables.

While **B.Width** is the only one variable that has **negative correlation** with Length.

(c) For the variables which seems to be associated with Length calculate a summary statistic which will describe the strength of the association, if possible.

Student ID: 1920253
Student Name: Chan, Hsin-Ju

```
> ##(c)##
> for (p in 4:9){
+   print(names(DP)[p])
+   print(cor(Length,DP[,p],use = "complete.obs"))}
[1] "Width"
[1] 0.7689932
[1] "Thickness"
[1] 0.5890989
[1] "B.Width"
[1] -0.283949
[1] "J.Width"
[1] 0.4540618
[1] "H.Length"
[1] 0.5091807
[1] "Weight"
[1] 0.879953
```

To describe the strength of the association, calculating the correlation coefficient of the variables.

It is true that only **B.Width** has **negative correlation** since the correlation coefficient is negative, while others (**Width, Thickness, J.Width, H.Length, Weight**) has **positive correlation coefficient**.

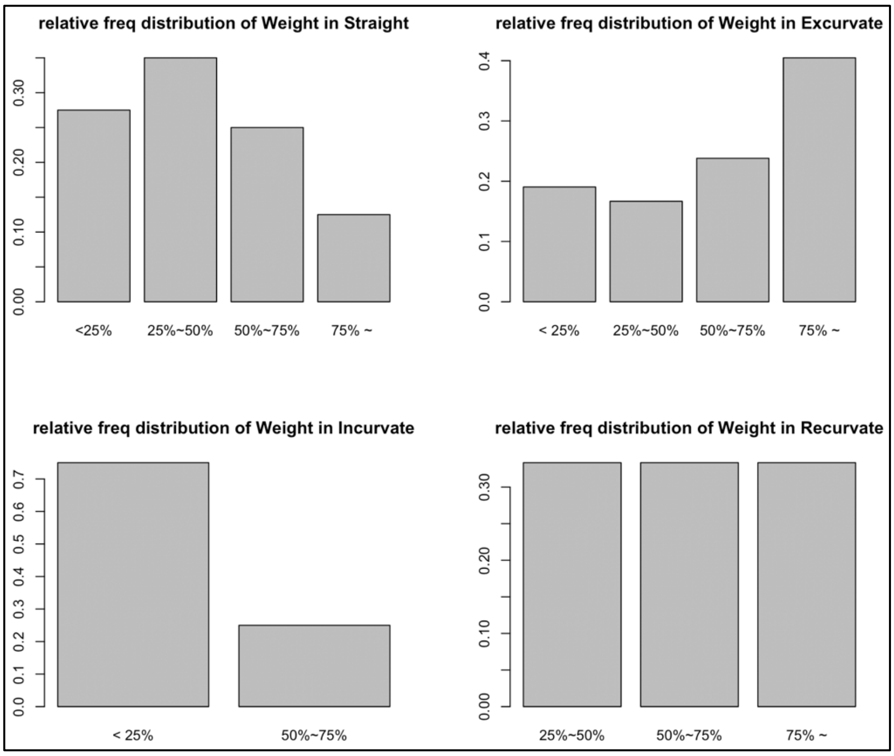
(d) Compute and represent graphically the relative frequency distribution of Weight conditionally to the various types of blade shape.

```
> data.frame('X'=Str,'Weight'=Weight[Str],'Straight'=S)
    X Weight Straight
1   1    3.6   0.0900
2   2    4.5   0.1125
3   3    3.6   0.0900
4   4    4.0   0.1000
5   5    2.3   0.0575
6   6    3.0   0.0750
7  17    2.5   0.0625
8  20    3.3   0.0825
9  21    3.6   0.0900
10 27    6.2   0.1550
11 29    4.6   0.1150
12 30    5.4   0.1350
13 32    5.1   0.1275
14 33    4.7   0.1175
15 34    7.2   0.1800
16 36    3.9   0.0975
17 44   28.8   0.7200
18 45   13.9   0.3475
19 46    9.4   0.2350
20 47    5.3   0.1325
21 48    7.9   0.1975
22 49    7.3   0.1825
23 59    9.2   0.2300
24 60    9.4   0.2350
25 61    6.7   0.1675
```

```
> data.frame('X'=Exc, 'Weight'=Weight[Exc],'Excurvate'=E)
    X Weight  Excurvate
1   8    6.2 0.14761905
2   9    5.1 0.12142857
3  10    2.8 0.06666667
4  11    2.5 0.05952381
5  12    4.8 0.11428571
6  13    3.2 0.07619048
7  14    3.8 0.09047619
8  15    4.5 0.10714286
9  16    4.4 0.10476190
10 19    4.2 0.10000000
11 22    7.4 0.17619048
12 23    5.6 0.13333333
13 24    4.8 0.11428571
14 25    7.8 0.18571429
15 26    9.2 0.21904762
16 31    5.9 0.14047619
17 37    4.1 0.09761905
18 39   10.7 0.25476190
19 40   12.5 0.29761905
20 41   13.4 0.31904762
21 42   11.1 0.26428571
22 50   12.2 0.29047619
23 51    9.3 0.22142857
24 52   11.1 0.26428571
25 53   14.8 0.35238095
26 54   10.7 0.25476190
27 55   11.1 0.26428571
28 56   12.3 0.29285714
```

```
> data.frame('X'=Inc, 'Weight'=Weight[Inc], 'Incurvate'=I)
   X Weight Incurvate
1  7    3.9     0.975
2 18    2.3     0.575
3 35    2.5     0.625
4 68    6.8     1.700
> data.frame('X'=Rec, 'Weight'=Weight[Rec], 'Recurvate'=R)
   X Weight Recurvate
1 43    7.2  2.400000
2 58    6.1  2.033333
3 62   15.3  5.100000
```

These three pictures show the value of relative frequency distribution of Weight in different blade shape.



The picture on the right side is a boxplot shows the relative frequency distribution of Weight in different type of Blade Shape.

Calculate the quantile in different shape.

Group the data by lower than Q1, Q1~Q2, Q2~Q3 and over Q3.

Plot it as four barplots which show the relative frequency distribution of them.

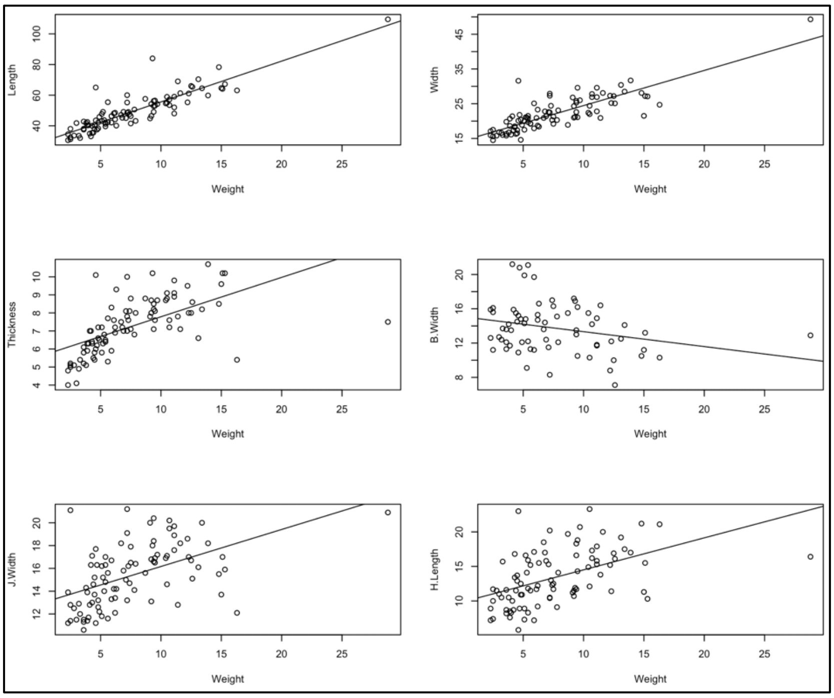
Student ID: 1920253
Student Name: Chan, Hsin-Ju

Multiple Linear Regression

(e) Select an appropriate multiple regression model, which can be used to predict the weight of the dart, using some or all (after appropriate selection) of the variables listed above as explanatory variables (with the exclusion of the weight itself, of course).

```
> for (p in 3:8){
+   print(names(DP)[p])
+   print(cor(Weight,DP[,p],use = "complete.obs"))}
[1] "Length"
[1] 0.879953
[1] "Width"
[1] 0.8263948
[1] "Thickness"
[1] 0.6001054
[1] "B.Width"
[1] -0.2641345
[1] "J.Width"
[1] 0.5025996
[1] "H.Length"
[1] 0.486397
```

First, calculate the correlation coefficient of all the numeric variables, we find out that Length (0.879953) and Width (0.8263948) are much higher than other variables.



This scatter plot can observe easily that **Length** and **Width** has strong relation with Weight, since it has obvious positive linear relation.

```
par(mfrow=c(3,2))
for (j in 3:8){
  plot(Weight, DP[,j], ylab =names(DP)[j])
  abline(lm(DP[,j]~Weight))}
```

Decide which variable to fit the linear regression model

```
lm1 <- lm(Weight~Length)   #highly relevant
summary(lm1) #R-Square:0.7743, p:<2.2e-16 rse:2.01,
anova(lm1)

lm2 <- lm(Weight~Width)  #worse than lm1
summary(lm2)  #R-squared: 0.6829, p: <2.2e-16, rse:2.382
anova(lm2)

lm3 <- lm(Weight~Thickness)   #worse
summary(lm3)   #R-squared: 0.3601, p-value:3.242e-10, se:3.384
anova(lm3)

lm4 <- lm(Weight~Length+Width)  #better than lm1
summary(lm4) #R-squared: 0.8292  p-value: < 2.2e-16, se:1.759
anova(lm4)
```

In lm1,lm2 and lm3, it is to calculate the model between Length and Weight, Width and Weight, Thickness and Weight. In lm1, **Length is highly relevant to Weight**, while in lm2, R-squared is a little **bit lower than lm1**. Lm3 is the worst one in these three models, which means **Thickness is not that relevant to Weight**.
lm4 is the model with Length+Width, and R-squared is 0.8292. It is the highest one means that **Weight is highly relevant to Length+Width**.

Student ID: 1920253
Student Name: Chan, Hsin-Ju

Anova-Two-Way Test

There are 21 possible models in Anova-Two-Way Test, and the picture on the right side is the two models I chose.

```
atw4<-lm(Weight~Name*Should.Or)   ##highest!
summary(atw4)  #R-squared:  0.6088, p-value: 2.835e-11, se: 2.853

atw20<-lm(Weight~Should.Or*Haft.Or)  #good good
summary(atw20)  #R-squared: 0.483, p-value: 3.793e-07, se: 3.28
```

```
> summary(atw4)

Call:
lm(formula = Weight ~ Name * Should.Or)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.0000 -1.4889 -0.1846  1.2154  7.7111 

Coefficients: (7 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                20.9179     4.1829   5.001 3.57e-06 ***
NameEnsor                 -16.3179     5.0632  -3.223 0.001871 ** 
NamePedernales              0.8821     3.6643   0.241 0.810423    
NameTravis                  4.3821     1.1034   3.971 0.000161 ***
NameWells                   4.2043     1.1034   3.810 0.000280 ***
Should.OrH                -15.8179     5.0632  -3.124 0.002525 ** 
Should.OrT                -16.5333     4.1453  -3.988 0.000152 ***
Should.OrX                -17.5000     3.4942  -5.008 3.47e-06 ***
NameEnsor:Should.OrH       15.6429     5.9842   2.614 0.010784 *  
NamePedernales:Should.OrH   4.3179     4.8161   0.897 0.372778    
NameTravis:Should.OrH           NA         NA      NA       NA    
NameWells:Should.OrH        0.1957     4.1829   0.047 0.962802    
NameEnsor:Should.OrT       17.2083     5.2305   3.290 0.001520 ** 
NamePedernales:Should.OrT   4.7667     3.6678   1.300 0.197671    
NameTravis:Should.OrT           NA         NA      NA       NA    
NameWells:Should.OrT            NA         NA      NA       NA    
NameEnsor:Should.OrX            NA         NA      NA       NA    
NamePedernales:Should.OrX       NA         NA      NA       NA    
NameTravis:Should.OrX           NA         NA      NA       NA    
NameWells:Should.OrX            NA         NA      NA       NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.853 on 76 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.6088,	Adjusted R-squared:  0.5471 
F-statistic: 9.857 on 12 and 76 DF,  p-value: 2.835e-11
```

Atw4 shows the highest relation to Weight, the non-numeric variables are **Name and Should.Or**.

In this model, **R-squared gets the highest result (0.6088)**.

Among the coefficients, **Travis** and **Wells** in Name, **Tapered** and **None** in Shoulder orientation (Should.Or) has three significant code. **Ensor** in Name, **Horizontal** in Shoulder.Or has two significant code.

That is these variables are highly relevant to Weight, therefore I chose them to process another Anova-Two-Way test.

Student ID: 1920253
Student Name: Chan, Hsin-Ju

```
> summary(atw20)  #R-squared: 0.483, p-value: 3.793e-07, se: 3.28

Call:
lm(formula = Weight ~ Should.Or * Haft.Or)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.0000 -2.0462 -0.2462  1.7750  7.4692 

Coefficients: (7 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          22.0359     2.8285   7.791 2.77e-11 ***
Should.OrH          -16.1942     4.4335  -3.653 0.000475 ***
Should.OrT          -12.9692     2.4914  -5.206 1.60e-06 ***
Should.OrX          -14.2359     3.6578  -3.892 0.000212 ***
Haft.OrE            -17.4359     4.3312  -4.026 0.000133 ***
Haft.OrP             -3.5000     4.0173  -0.871 0.386366    
Haft.OrT             -0.2359     1.6189  -0.146 0.884531    
Haft.OrV              1.3583     2.1173   0.642 0.523098    
Should.OrH:Haft.OrE  16.1542     5.7067   2.831 0.005939 ** 
Should.OrT:Haft.OrE  13.7154     4.1689   3.290 0.001521 ** 
Should.OrX:Haft.OrE       NA         NA      NA       NA    
Should.OrH:Haft.OrP  10.8917     5.9132   1.842 0.069389 .  
Should.OrT:Haft.OrP   2.3551     4.2894   0.549 0.584588    
Should.OrX:Haft.OrP       NA         NA      NA       NA    
Should.OrH:Haft.OrT   1.4442     4.4335   0.326 0.745504    
Should.OrT:Haft.OrT       NA         NA      NA       NA    
Should.OrX:Haft.OrT       NA         NA      NA       NA    
Should.OrH:Haft.OrV       NA         NA      NA       NA    
Should.OrT:Haft.OrV       NA         NA      NA       NA    
Should.OrX:Haft.OrV       NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.28 on 76 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.483,	Adjusted R-squared:  0.4013 
F-statistic: 5.916 on 12 and 76 DF,  p-value: 3.793e-07
```

Atw20 gets the second-high R-square (0.483) in all 21 models, this model is to calculate the relation **Should.Or*Haft.Or** to Weight.

Among the coefficients, **Horizontal, Tapered** and **None** in Should.Or, **Expanding** in Haft.Or got three significant code.

These variables are highly relevant to Weight, therefore I chose them to process another Anova-Two-Way test.

Choose the specific type in the above selected values to process Anova-Two-Way Test

```
NameTravis <- as.double(Name == 'Travis')
Should.OrH <- as.double(Should.Or == 'H')
Should.OrT <- as.double(Should.Or == 'T')
d1 <- lm(Weight~Length+Width+Should.OrH*NameTravis+Should.OrT*NameTravis)
summary(d1)  ## 0.8577, 1.657

Haft.OrE <- as.double(Haft.Or == 'E')
Should.OrH <- as.double(Should.Or=='H')
Should.OrT <- as.double(Should.Or=='T')
d2 <- lm(Weight~Length+Width+Should.OrH*NameTravis+Should.OrT*NameTravis
        +Should.OrT*Haft.OrE+Should.OrH*Haft.OrE)
summary(d2)  #0.8604, 1.672
```

d1 model is the original model (Length + Width) plus the interaction between Travis (Name) and Horizontal, Tapered (Should.Or). Its R-squared is 0.8577.

d2 model add the interaction between Horizontal, Tapered (Should.Or) and Expanding (Haft.Or). Its R-squared become higher than d1 model, which is 0.8604.

d3 model add the interaction between Ensor, Travis (Name) and Tapered (Should.Or). Its R-squared is also higher than the previous one, which is 0.8632.

d4 model add the interaction between Wells (Name) and Tapered, None (Should.Or). Its R-squared is a little bit higher than the previous one, which is 0.8634.

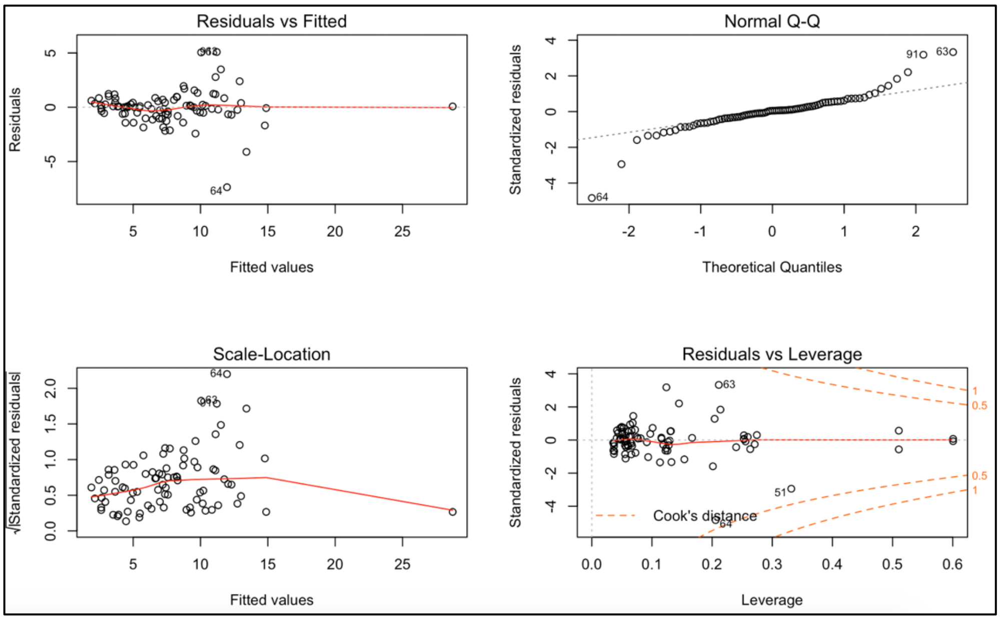
```
NameEnsor <- as.double(Name =='Ensor')
NameTravis <- as.double(Name == 'Travis')
Should.OrT <- as.double(Should.Or == 'T')
d3 <- lm(Weight~Length+Width+Should.OrH*NameTravis+Should.OrT*NameTravis
        +Should.OrT*Haft.OrE+Should.OrH*Haft.OrE
        +Name.E*Should.OrT+Name.T*Should.OrT)   #0.8632
summary(d3)

Should.OrX <- as.double(Should.Or=='X')
NameWells <- as.double(Name == 'Wells')
d4 <- lm(Weight~Length+Width+Should.OrH*NameTravis+Should.OrT*NameTravis
        +Should.OrT*Haft.OrE+Should.OrH*Haft.OrE
        +NameEnsor*Should.OrT+NameTravis*Should.OrT
        +NameWells*Should.OrT+NameWells*Should.OrX)   #0.8634
summary(d4)
```

```
d_final <- lm(Weight~Length+Width+Should.OrH*NameTravis+Should.OrT*NameTravis
             +Should.OrT*Haft.OrE+Should.OrH*Haft.OrE
             +NameEnsor*Should.OrT+NameTravis*Should.OrT
             +NameWells*Should.OrT+NameWells*Should.OrX)
summary(d_final)  #0.8634, 1.709
```

As a result, d4 model is the final regression model to predict the weight of the dart.
(I changed its name as d_final)

Student ID: 1920253
Student Name: Chan, Hsin-Ju

(f) Check and describe the fit of your model using whatever graphical or numerical methods seem appropriate.



(f) Interpret the fitted model in practical terms. What does it tell you about predicting the dart weight?

```
> summary(d_final)  #0.8634, 1.709

Call:
lm(formula = Weight ~ Length + Width + Should.OrH * NameTravis +
    Should.OrT * NameTravis + Should.OrT * Haft.OrE + Should.OrH *
    Haft.OrE + NameEnsor * Should.OrT + NameTravis * Should.OrT +
    NameWells * Should.OrT + NameWells * Should.OrX)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3661 -0.6184  0.0304  0.6026  5.0904

Coefficients: (2 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -9.79384    2.41635  -4.053 0.000124 ***
Length                 0.17101    0.02538   6.737 3.03e-09 ***
Width                  0.40146    0.06659   6.029 5.96e-08 ***
Should.OrH            -2.04148    1.56685  -1.303 0.196641
NameTravis            -0.23196    2.12923  -0.109 0.913544
Should.OrT             0.32954    1.54261   0.214 0.831425
Haft.OrE              -0.29914    3.03896  -0.098 0.921854
NameEnsor             -0.81506    1.92890  -0.423 0.673846
NameWells              0.37178    1.88582   0.197 0.844254
Should.OrX             0.61328    2.41043   0.254 0.799871
Should.OrH:NameTravis       NA         NA      NA       NA
NameTravis:Should.OrT  1.52957    2.20435   0.694 0.489925
Should.OrT:Haft.OrE    0.32192    3.00415   0.107 0.914953
Should.OrH:Haft.OrE    2.05198    2.46677   0.832 0.408170
Should.OrT:NameEnsor  -0.31581    2.12637  -0.149 0.882335
Should.OrT:NameWells  -0.40437    2.00701  -0.201 0.840879
NameWells:Should.OrX        NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.709 on 74 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8634,    Adjusted R-squared:  0.8376
F-statistic: 33.41 on 14 and 74 DF,  p-value: < 2.2e-16
```

The multiple linear regression is:

$$Y_i = -9.79 + 0.17x_{i1} + 0.4x_{i2} - 2.04x_{i3} - 0.23x_{i4} + 0.33x_{i5} - 0.3x_{i6} - 0.82x_{i7} + 0.37x_{i8} + 0.61x_{i9} + 1.53x_{i10} + 0.32x_{i11} + 2.05x_{i12} - 0.32x_{i13} - 0.4x_{i14} + 1.709$$

We wish to test the hypothesis:

H_0 : μ *in each variables are equal at level* $\alpha = 0.05$

In this model, the multiple R-squared is 0.8634 and the Adjusted R-squared is 0.8376. Both of them are quite close, which means this model may have higher possibility to be accurate. The p_value is really low and smaller than 0.05, that is a non-rejection to H_0.

(h) Predict the expected dart weight for a dart point of type Travis, with maximum length 70 mm, H.Length 60mm, Thickness 50 mm, B.Width 50 mm, J.Width 50 mm, Width 60 mm and with both blade shape and base shape recurvate, straight shoulder shape, barbed shoulder orientation, excurvate shape for the lateral haft element and parallel orientation of the lateral haft element. Give a 95% confidence interval for this expected weight. Is there any reason to be cautious about your estimate?

```
> newdata<-data.frame(Length=70,Width=60,NameTravis=1,Should.OrH=0,Should.OrT=0,Haft.OrE
=0,NameEnsor=0,NameWells=0,Should.OrX=0)
> predict(d_final,newdata,df=74,interval = "confidence",level=0.95)
       fit      lwr      upr
1 26.03204 19.58275 32.48132
```

The expected Weight is 26.0324.

The 95% confidence interval for this expected weight is [19.58275, 32.48132].