Assignment 7

Robert Ivill 46012819

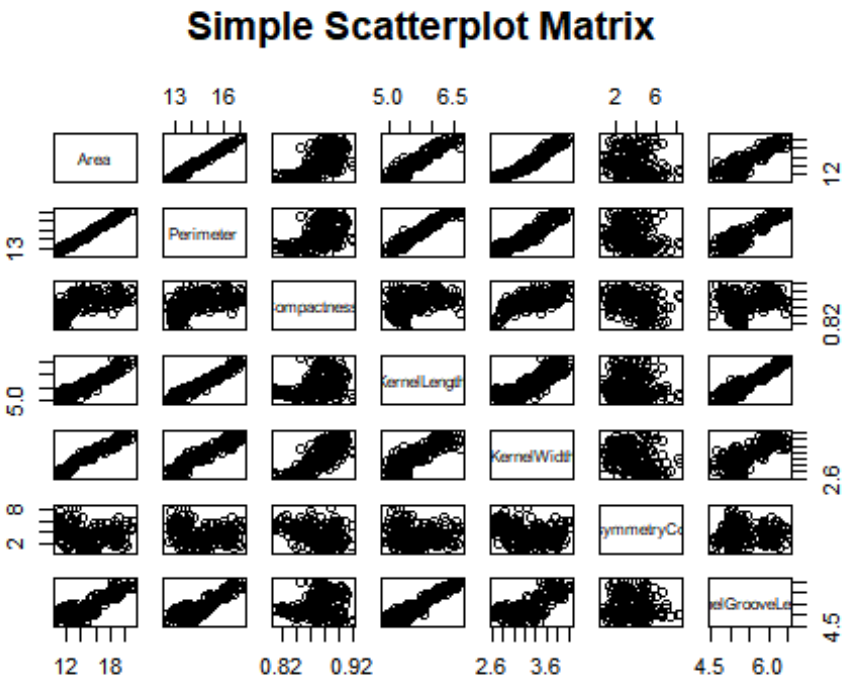
Question 1 [80% of the marks] The data file seeds.csv contains measurements of geometrical properties of kernels belonging to three different varieties of wheat. The measurements are: Area, Perimeter, KernelLength, KernelWidth, KernelGrooveLength (in cm on an X-ray image) and two measures of shape: Compactness, and AsymmetryCoef.

a. Load the data, print the summary, remove Variety variable and plot pairwise scatterplots. Comment on a few features you notice in the scatterplot matrix.

```
seeds <- read.csv("/Users/robbi/Downloads/seeds.csv")
summary(seeds)
```

```
##       Area         Perimeter      Compactness      KernelLength
##  Min.   :10.59   Min.   :12.41   Min.   :0.8081   Min.   :4.899
##  1st Qu.:12.27   1st Qu.:13.45   1st Qu.:0.8569   1st Qu.:5.262
##  Median :14.36   Median :14.32   Median :0.8734   Median :5.524
##  Mean   :14.85   Mean   :14.56   Mean   :0.8710   Mean   :5.629
##  3rd Qu.:17.30   3rd Qu.:15.71   3rd Qu.:0.8878   3rd Qu.:5.980
##  Max.   :21.18   Max.   :17.25   Max.   :0.9183   Max.   :6.675
##   KernelWidth    AsymmetryCoef    KernelGrooveLength    Variety
##  Min.   :2.630   Min.   :0.7651   Min.   :4.519      Min.   :1
##  1st Qu.:2.944   1st Qu.:2.5615   1st Qu.:5.045      1st Qu.:1
##  Median :3.237   Median :3.5990   Median :5.223      Median :2
##  Mean   :3.259   Mean   :3.7002   Mean   :5.408      Mean   :2
##  3rd Qu.:3.562   3rd Qu.:4.7687   3rd Qu.:5.877      3rd Qu.:3
##  Max.   :4.033   Max.   :8.4560   Max.   :6.550      Max.   :3
```

```
seeds_df <- seeds[!names(seeds) %in% "Variety"]
pairs(~ .,data=seeds_df,
main="Simple Scatterplot Matrix")
```



From the scatterplot matrix, we can see that quite a few of the variables share a positive linear relationship with each other. This includes area and perimeter, kernel length and kernel width, and kernel groove length and perimeter. The asymmetry coefficient seems to have little to no pattern when compared to other variables. Compactness also has a less clear relationship with the other variables.

b. Calculate the variance of each variable. Explain why you would scale the variables before performing principal component analysis (PCA)?

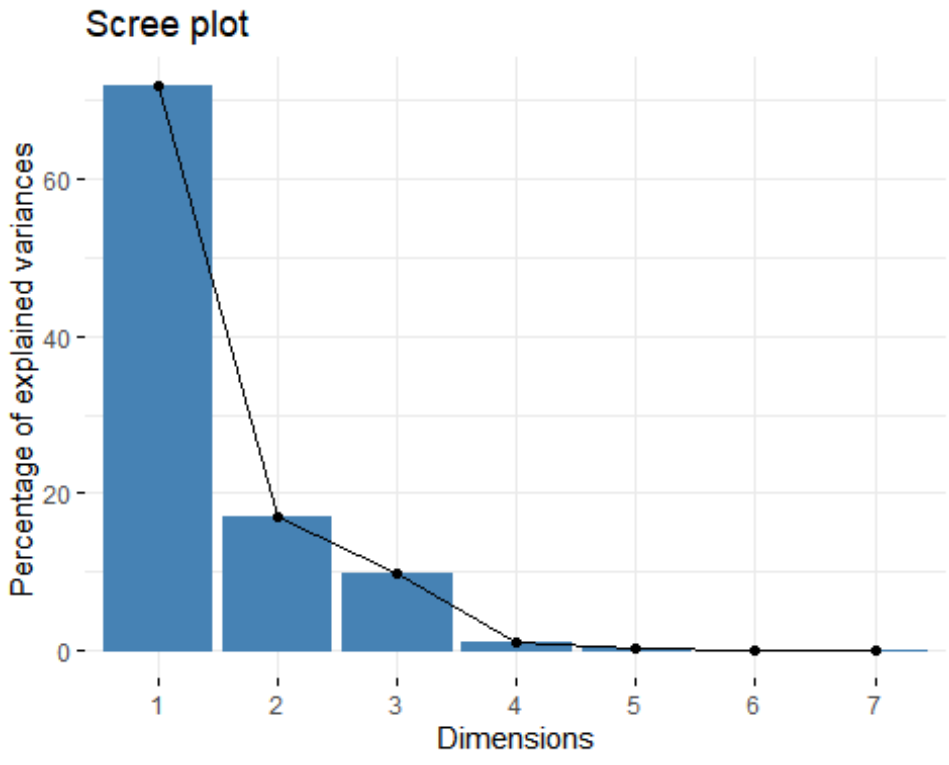
```
apply(seeds_df, 2, var)

##                Area           Perimeter         Compactness
KernelLength
##       8.4663507769       1.7055281955       0.0005583493
0.1963052453
##        KernelWidth      AsymmetryCoef KernelGrooveLength
##       0.1426682019       2.2606840456       0.2415530810
```

By calculating the variance of each variable, we can see the variability within the dataset. It is important to scale the variables before performing PCA because PCA is affected heavily by the relative scales of each variable. The difference in variance of Area and Compactness is not necessarily because area is more variable, but because it is on a different scale to compactness. By scaling, we make all of the variables to have a standard deviation of one, making it easier for PCA to use all of the variables equally and remove bias in it.

c. Perform PCA. Print the summary of the PCA and create a scree plot to answer the questions: • How much of the variance is explained by the first four principal

components? • How many principal components would you choose according to the elbow rule? Is this enough to explain at least 80% of the variance in the data? • Assuming you want to explain at least 80% of your data, how many principal components would you choose according to the elbow rule?

```
seeds_pca <- prcomp(seeds_df, scale=TRUE)
library(factoextra)

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

summary(seeds_pca)

## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
PC7
## Standard deviation     2.2430 1.0943 0.82341 0.26147 0.13680 0.07302
0.02850
## Proportion of Variance 0.7187 0.1711 0.09686 0.00977 0.00267 0.00076
0.00012
## Cumulative Proportion  0.7187 0.8898 0.98668 0.99645 0.99912 0.99988
1.00000

fviz_eig(seeds_pca)
```

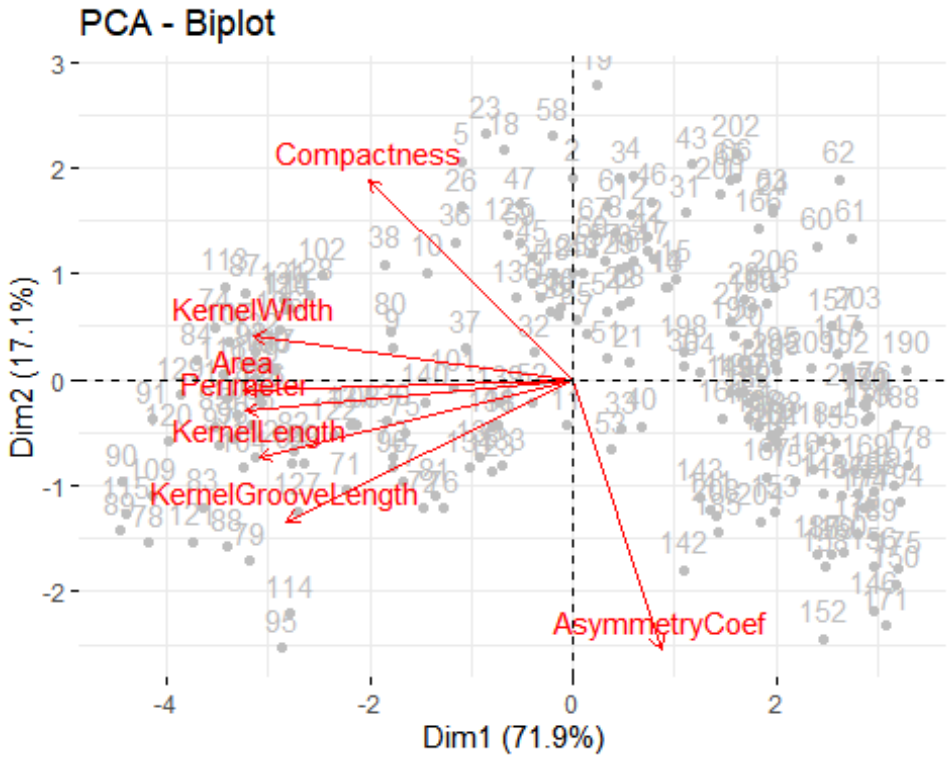


To find the proportion of variance explained by the first four principal components, we add the values of PC1-4 for the proportion of variance in the summary output. These values are 0.7187,

0.1711, 0.09686, and 0.00977. These sum to 0.99643 or approximately 99.6% of the variance. To find how many principal components to choose according to the elbow rule, we find the elbow of the graph where the plot starts to level off after steep declines. In this plot, the elbow point is at 4, implying that the first three principal components explain the most variation in the data. We know that the first 3 components make up around 98.6% of the variance according to our first answer in this question. As this proportion is greater than 80%, the three components can explain at least 80% of the variance in the data. As we have confirmed that the first three principal components can explain at least 80% of the variance, we would choose these three to explain at least 80% of our data.

d. Visualise the results using fviz_pca_biplot function. Referring to the biplot answer the following questions: • Name three variables that are highly correlated with the first principal component. Are they negatively or positively correlated? • What do you think the first principal component explains? • Which two variables have the highest absolute loadings on the second principal component? What does this mean? • Does seed 152 (located in the upper right corner) has high or low Compactness?

```
fviz_pca_biplot(seeds_pca, col.var = "red", col.ind="grey")
```

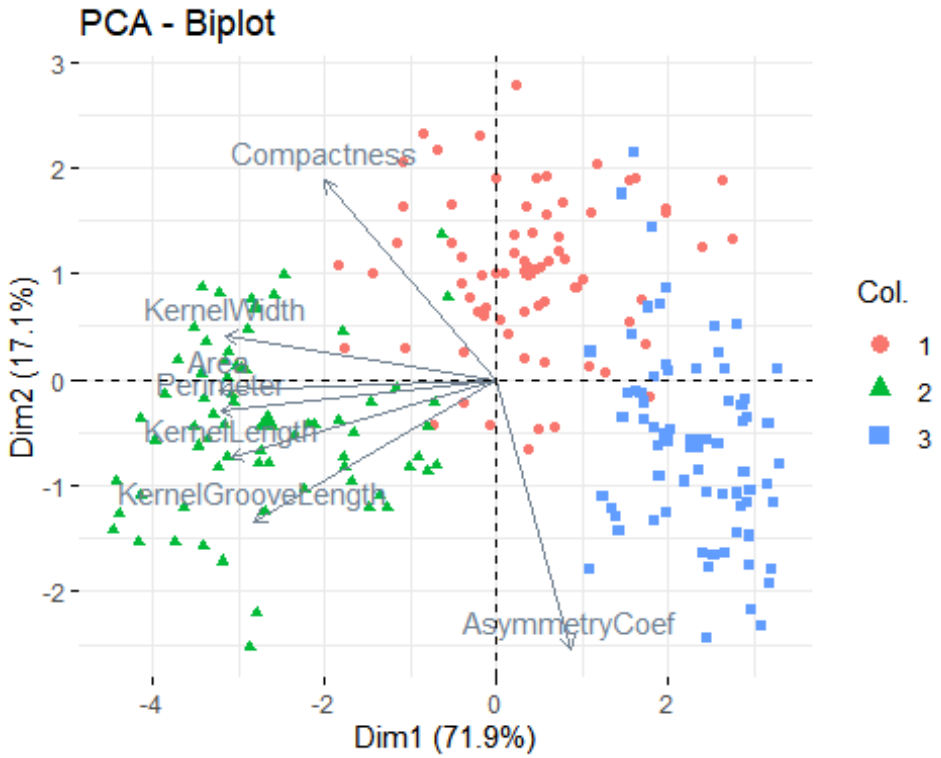


From the PCA biplot, we can see that the variables that are highly correlated with the first principal component are area, perimeter and kernel width. All of these are negatively correlated to PC1. The first principal component explains the majority of proportion of variance in the dataset, as we can see with how many variables are correlated to it. The variables with the highest absolute loadings on the second principal component are the Asymmetry Coefficient and compactness. This means that these variables account for most of the

variance in PC2. Seed 152 has a low compactness, as it is on the opposite side of the biplot to the compactness variable vector.

e. Now create a biplot where the seeds are grouped by variety. Explain the differences in the varieties in terms of the kernel sizes and asymmetry coefficients.

```
fviz_pca_biplot(seeds_pca, col.var = "slategrey",
col.ind=as.factor(seeds$Variety), label="var")
```



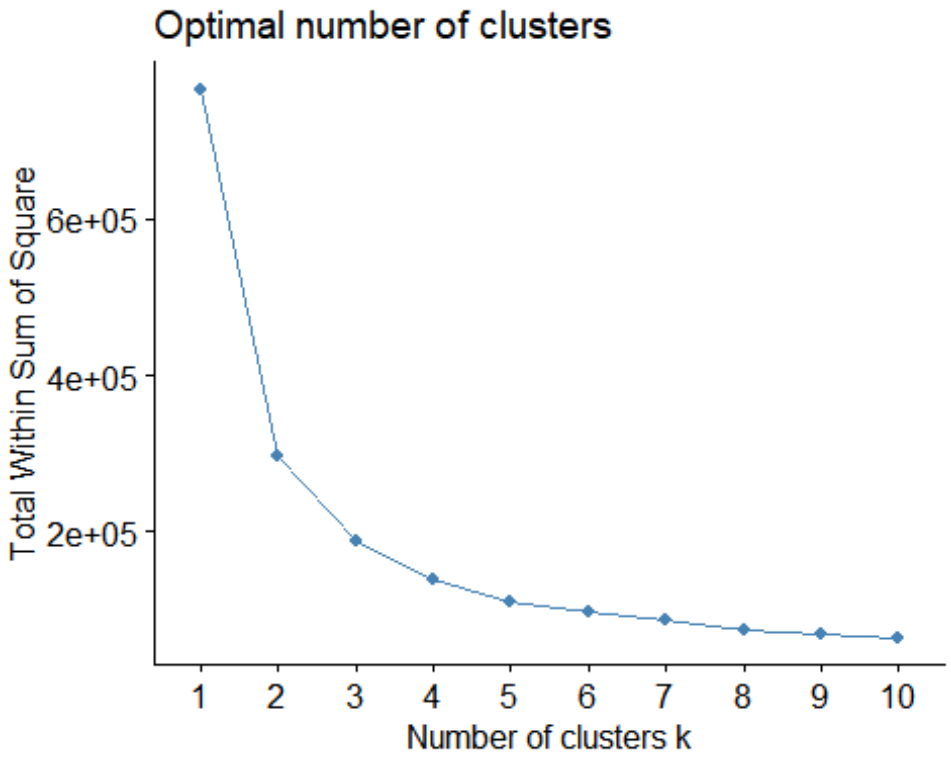
The seeds with green variability have larger kernel sizes as they are all in the direction of the kernel size vectors. They have a bit of variability in these sizes as the points are spread out a bit. They have quite an average assymetry coefficient as most of the points are roughly perpendicular to the asymmetry coefficient with some variability. The red seeds are quite variable for both asymmetry and kernel size as the points are spread out quite a bit. They have smaller kernel sizes and asymmetry coefficient as they are on the opposite side to those vectors. The blue seeds have the smallest kernel size as they are on the opposite side to the kernel size vectors. They have slighly larger asymmetry coefficients to the green seeds. They have lower variability with each other as they are quite compactly spread.

Question 2 [20% of the marks]

Weather dataset contains different weather related characteristics on 244 different days in Algeria forest.

a. Load the data, print the structure. Scale the data and plot the withiness against the number of clusters.How many clusters would you choose for a k-mean clustering analysis?

```
weather <- read.csv("/Users/robbi/Downloads/weather.csv")
str(weather)

## 'data.frame':    244 obs. of  10 variables:
##  $ Temperature: int  29 29 26 25 27 31 33 30 25 28 ...
##  $ RH         : int  57 61 82 89 77 67 54 73 88 79 ...
##  $ Ws         : int  18 13 22 13 16 14 13 15 13 12 ...
##  $ Rain       : num  0 1.3 13.1 2.5 0 0 0 0 0.2 0 ...
##  $ FFMC       : num  65.7 64.4 47.1 28.6 64.8 82.6 88.2 86.6 52.9 73.2 ...
##  $ DMC        : num  3.4 4.1 2.5 1.3 3 5.8 9.9 12.1 7.9 9.5 ...
##  $ DC         : num  7.6 7.6 7.1 6.9 14.2 22.2 30.5 38.3 38.8 46.3 ...
##  $ ISI        : num  1.3 1 0.3 0 1.2 3.1 6.4 5.6 0.4 1.3 ...
##  $ BUI        : num  3.4 3.9 2.7 1.7 3.9 7 10.9 13.5 10.5 12.6 ...
##  $ FWI        : num  0.5 0.4 0.1 0 0.5 2.5 7.2 7.1 0.3 0.9 ...

weather_scaled <- as.data.frame(scale(weather))
fviz_nbclust(weather, kmeans, method = "wss")
```

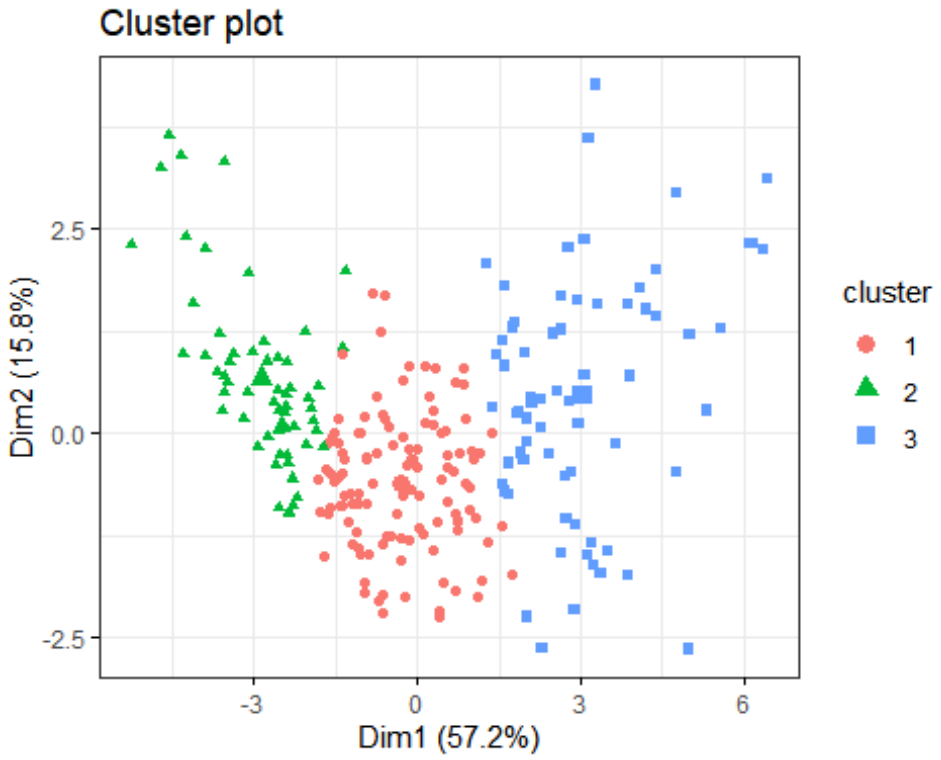


To find the number of clusters to choose for a k-mean cluster analysis, we must inspect the plot and choose the elbow point where the within sum of squares begins to level off. For this dataset, the elbow point we would choose is 3, as that is where the initial drop begins to level off. Therefore we choose clusters for the analysis.

b. Perform a k-mean clustering analysis with the number of clusters chosen in a. (replace K in the code below with the corresponding number). Visualise the clusters in the first two principal component plane using fviz_cluster function and the code below. Is the first principal component enough to separate the clusters?

In the code below, we replaced K with 3 as that is what we chose in a.

```
weather_ca <- kmeans(weather_scaled, 3)
fviz_cluster(weather_ca, data = weather_scaled, geom = "point", ellipse.type
= "none", ggtheme = theme_bw())
```



To see if the first principal component is enough to seperate the clusters, we look for distinct seperations or patterns among the clusters. The clusters look quite well separated with little overlap, therefore we can conclude that PC1 is enough to separate the clusters.

c. Print the centers of the clusters. Does the cluster with highest average temperatures also have highest average rain fall?

```
weather_ca$centers

##   Temperature          RH          Ws       Rain       FFMC        DMC
## 1  0.07599247 -0.07869306 -0.26834537 -0.1980591  0.2929524 -0.2741423
## 2 -1.07474764  0.91337180  0.45521487  0.7932412 -1.4664993 -0.8903347
## 3  0.81867254 -0.67167520  0.05610546 -0.3620531  0.7942276  1.2532439
##           DC        ISI        BUI        FWI
## 1 -0.3026896 -0.1926414 -0.2845164 -0.3085852
## 2 -0.7537807 -0.9535157 -0.8764260 -0.9009403
## 3  1.1814535  1.1699609  1.2586686  1.3213573
```

From the centers of the clusters, we can see that the 3rd cluster has the highest average temperature, with 0.808. If we then look at the rain variable, we can see that the 2nd cluster has the highest average rainfall with 0.7854. Therefore, the cluster with the highest average temperature does not have the highest average rainfall.