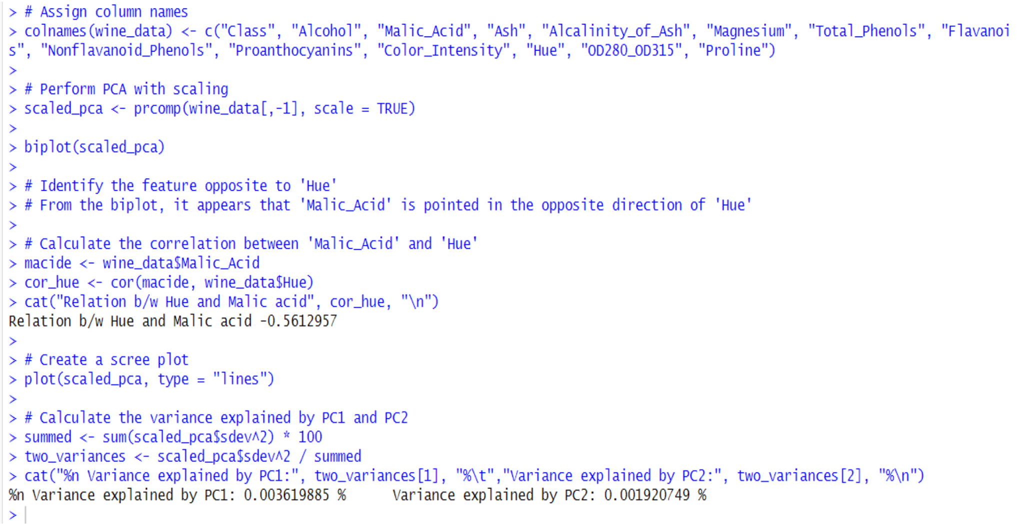
Homework 5

O Sai ram

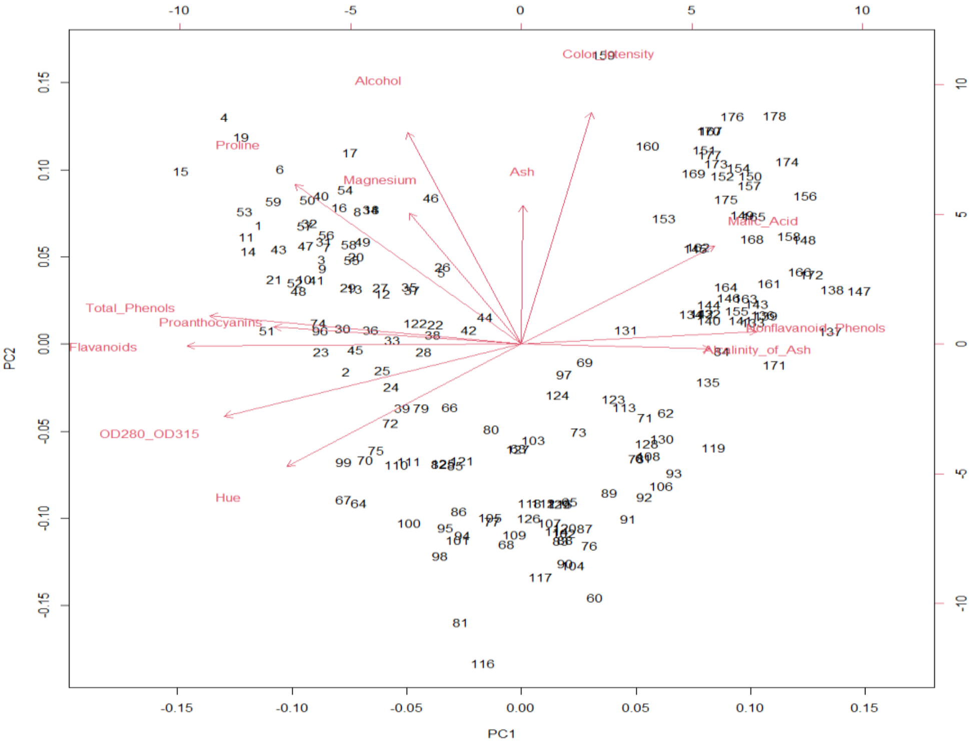
A20522183

Practicum Problems

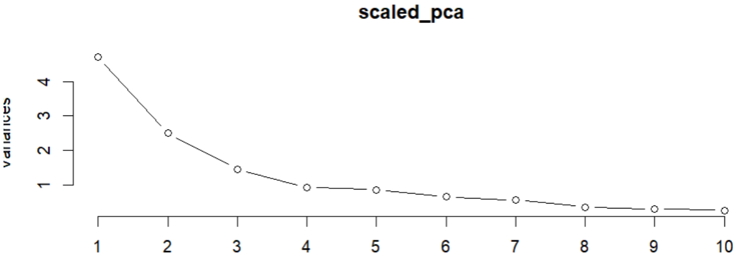
Q1.



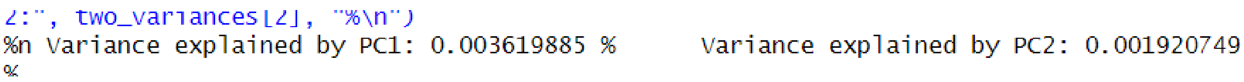
```
> # Assign column names
> colnames(wine_data) <- c("Class", "Alcohol", "Malic_Acid", "Ash", "Alcalinity_of_Ash", "Magnesium", "Total_Phenols", "Flavanoi
s", "Nonflavanoid_Phenols", "Proanthocyanins", "Color_Intensity", "Hue", "OD280_OD315", "Proline")
> 
> # Perform PCA with scaling
> scaled_pca <- prcomp(wine_data[,-1], scale = TRUE)
> 
> biplot(scaled_pca)
> 
> # Identify the feature opposite to 'Hue'
> # From the biplot, it appears that 'Malic_Acid' is pointed in the opposite direction of 'Hue'
> 
> # Calculate the correlation between 'Malic_Acid' and 'Hue'
> macide <- wine_data$Malic_Acid
> cor_hue <- cor(macide, wine_data$Hue)
> cat("Relation b/w Hue and Malic acid", cor_hue, "\n")
Relation b/w Hue and Malic acid -0.5612957
> 
> # Create a scree plot
> plot(scaled_pca, type = "lines")
> 
> # Calculate the variance explained by PC1 and PC2
> summed <- sum(scaled_pca$sdev^2) * 100
> two_variances <- scaled_pca$sdev^2 / summed
> cat("%n Variance explained by PC1:", two_variances[1], "%\t","Variance explained by PC2:", two_variances[2], "%\n")
%n Variance explained by PC1: 0.003619885 %      Variance explained by PC2: 0.001920749 %
> 
```



The scree plot, is the correlation between malic acid and hue.

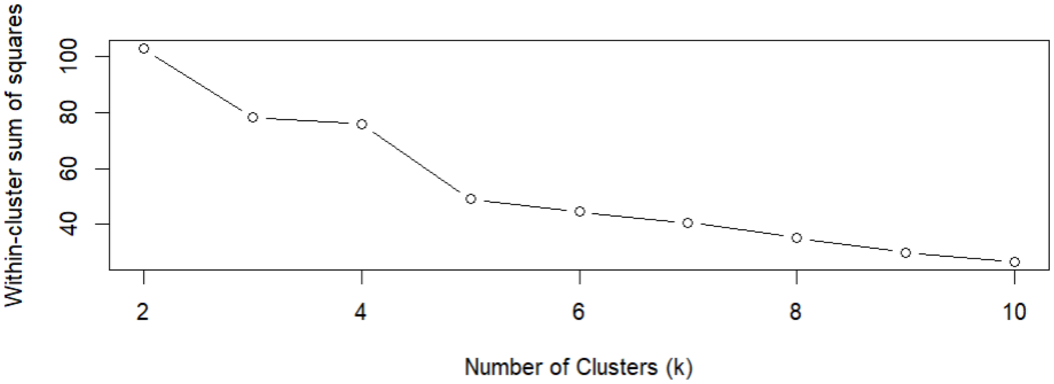


The variance explained by PCA 1 and PCA2 is given by

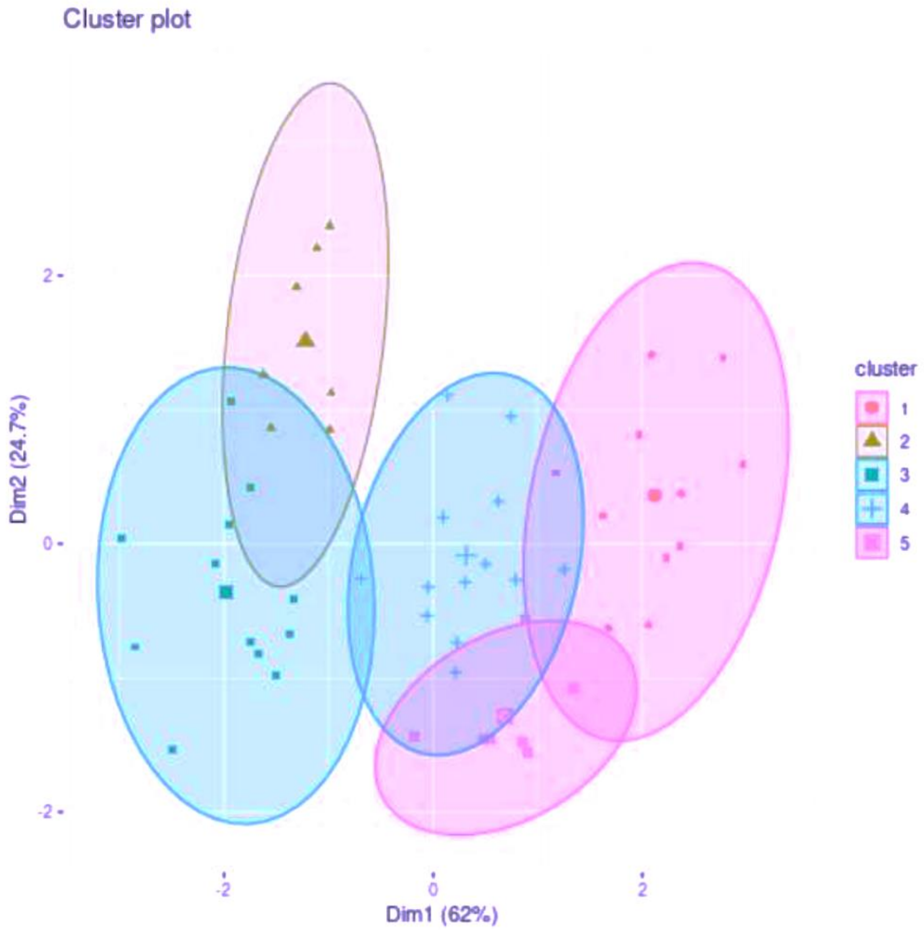


```
2:", two_variances[2], "%\n")
%n Variance explained by PC1: 0.003619885 %      Variance explained by PC2: 0.001920749
%
```

Q2.



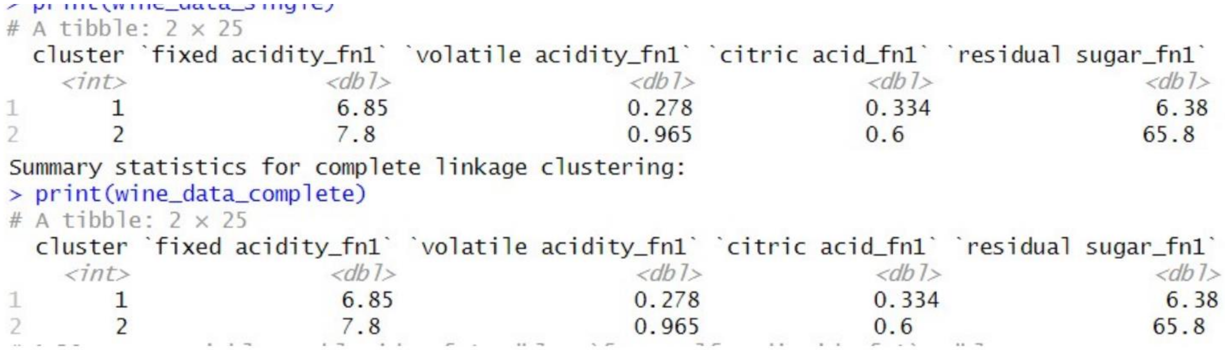
The optimal number of clusters is 5



Q3.

Single linkage penultimate distance: 9.698919

Complete linkage penultimate distance : 21.07833


```
> print(wine_data_single)
# A tibble: 2 × 25
  cluster `fixed acidity_fn1` `volatile acidity_fn1` `citric acid_fn1` `residual sugar_fn1`
    <int>               <dbl>                  <dbl>             <dbl>                <dbl>
1       1                6.85                  0.278             0.334                 6.38
2       2                7.8                   0.965             0.6                  65.8
Summary statistics for complete linkage clustering:
> print(wine_data_complete)
# A tibble: 2 × 25
  cluster `fixed acidity_fn1` `volatile acidity_fn1` `citric acid_fn1` `residual sugar_fn1`
    <int>               <dbl>                  <dbl>             <dbl>                <dbl>
1       1                6.85                  0.278             0.334                 6.38
2       2                7.8                   0.965             0.6                  65.8
```

Largest differences for both hierarchical clustering is residual sugar_fn1 with absolute difference equal to 59.41201. Both complete and single linkage are produces as balanced clustering