TANIYA PAL(373094)

**Task 1:**

**Methodology**

At the beginning of the task, the dataset “ASP.csv” was loaded into the R workspace. While reading the csv file provided, the file was read into a data frame. The corresponding data frame was converted into a matrix, as for plotting histograms, a matrix needs to be provided as input. Hereafter, the histograms were plotted by creating a function. The histograms of each variable are shown in Figure 1.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Figure 1: Histograms of each variable in “ASP.csv” file

For further analysis, the data frame was used instead of the matrix. From the data frame, the NAs were omitted. Descriptive Statistics, which included mean for each variable and displayed in a data frame. The data frame can be seen in Figure 2. Following that, the sum of all the variables that represent sugar were calculated. The sum was calculated by multiplying the mean of each sugar variable (calculated in the previous steps) by the number of observations (23). The sum of abscisic acid and its metabolite was calculated in a similar way.

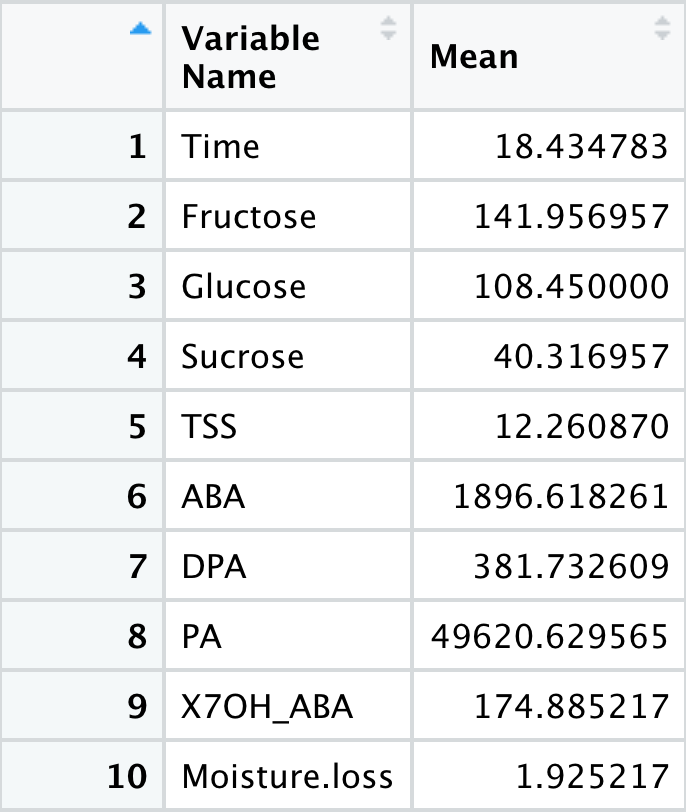


Figure 2: Data frame representing mean of all variables (columns in the ASP data frame)

The two data frames representing sum of sugars and ABA metabolites were merged into one data frame. Figure 3 represents the data frame.

Table

Description automatically generated

Figure 3: Data frame containing values of sum of each sugar and each ABA metabolite. The column “sum\_sugar” represents sum of sugars and column “ABA\_metab” represents sum of ABA metabolites.

Chart, box and whisker chart

Description automatically generatedA boxplot was created which helps to visualize mean of the “sum\_sugar” column, “ABA\_metab” column and the distribution of TSS and Moisture loss variables. The boxplot can be visualized in Figure 4.

Figure 4: Boxplot of sugars, ABA metabolites, TSS and Moisture loss.

Hereafter, two way anova was performed on the dataset to infer the effect of Treatment variable and Time variable (independent variables) on the following variables: sum of sugars, ABA metabolites, TSS and moisture loss(dependent variables). On performing the Tukey Test for each anova result of the four variables, the result is listed in Table 1.

|  |  |  |
| --- | --- | --- |
| Variable | Number of Significant Interaction | Significant interaction variables |
| Sugar | 6 | DCA 8  AIR 8  AIR 28  AIR 8  DCA 28  AIR. 8  AIR 28  DCA 8  DCA 28  DCA 8  DCA 28  AIR 28 |
| ABA | 6 | DCA 8  AIR 8  AIR 28  AIR 8  DCA 28  AIR. 8  AIR 28  DCA 8  DCA 28  AIR 8  DCA 28  AIR 28 |
| TSS | 1 | DCA 8  AIR 8 |
| Moisture | 1 | DCA 8  AIR 8 |

Table 1: List of significant interactions after Tukey Test

The distribution of residuals after performing two way anova was also visualized for each variable. Figure 5 shows the histogram plots of distribution of residuals for each variable.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Figure 5: Distribution of residuals after performing two way Anova test on each variable.

Top left: for sugar, Top right: for ABA metabolites, Bottom left: for TSS, Bottom right: for Moisture loss

PCA plots were generated using the “FactoMineR” library. Following that, individual and biplots were generated using “factoextra” library. Scree plot was also generated for deciding how many principal components to keep.

Figure 6 shows the different plots derived after PCA. Figure 7 shows the scree plot for individual principal components.

Diagram

Description automatically generated with medium confidenceChart, scatter chart

Description automatically generated

Chart

Description automatically generated (a) (b)

Chart, scatter chart

Description automatically generated

(c) (d)

Figure 6: (a) PCA individual plot grouped by variable Time (b) PCA biplot grouped by variable Treatment (c) PCA individual plot grouped by variable Treatment (d) PCA biplot grouped by variable Time

Chart, histogram

Description automatically generated

Figure 7: Scree plot displaying importance of each principal components

Bar plots were created to visualize the contribution of each variable in principal component dimension 1 and 2. Figure 8 is a representation of the bar plots generated.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure 8: bar plots demonstrating contribution of each variable in the first two principal components

Regression plots were generated from linear models between TSS and the variables: sum of sugars, Fructose, Glucose and Sucrose. The R-squared value and the p value was also calculated for each model. The respective regression plots were generated with the regression line represented in each plot with blue color. Figure 9 shows the four regression plots for the four models.

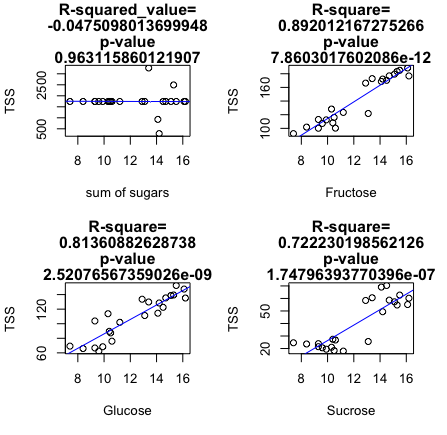
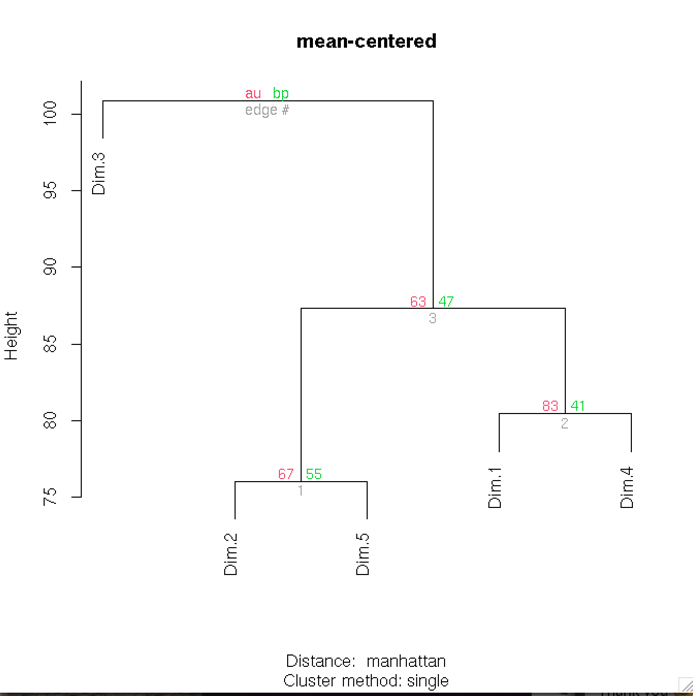
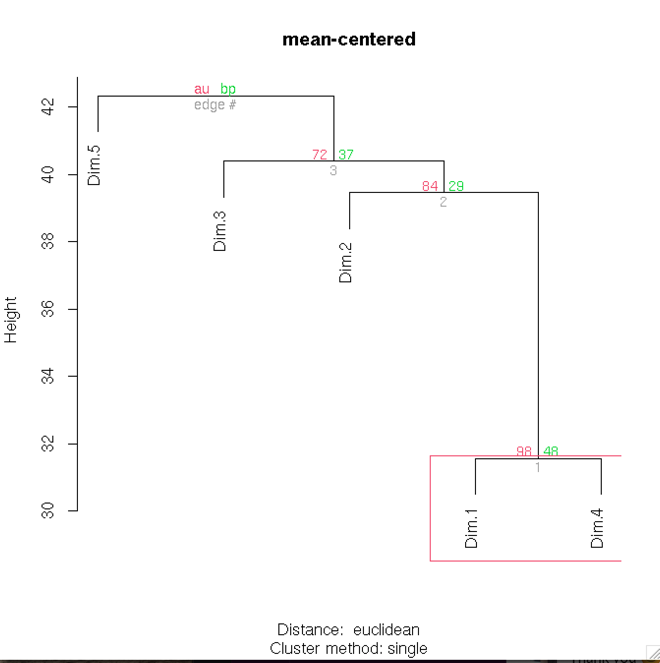


Figure 9: Regression plot for measuring correlation of TSS with sum of sugars, Glucose, Fructose, Sucrose respectively

Chart, diagram, box and whisker chart

Description automatically generated

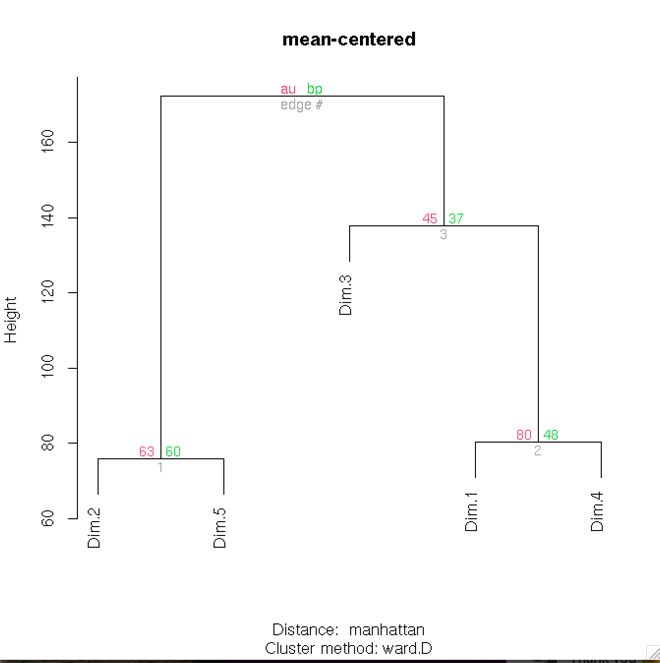


Figure 10: HCA on PCA scores on all the dimensions with different linkage and distance method

HCA was done on the PCA scores of all the dimensions, and dendograms were visualized by using two linkage and distance methods, ward.D and single among the linkage methods and Euclidean and Manhattan among the distance methods.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

(b)

(a)

Figure 11: k-means clustering on scores of first three dimensions of PCA, with k=3. (a) k-means clustering on raw data (b) k-means clustering on scaled data

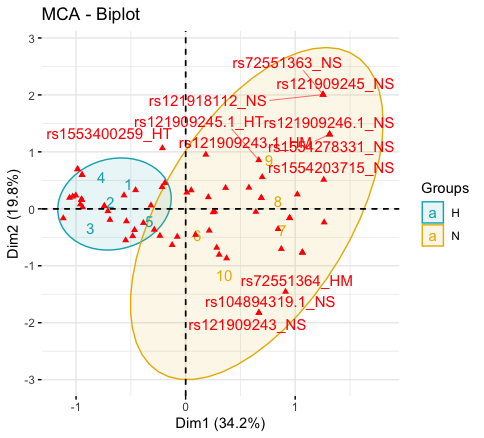
K-means clustering was performed on the raw data first with a k value of 3. Following that, the data was scaled, and k-means clustering was performed using the same parameters.

**Results and Discussion:** The histogram shown in Figure 1 of all the variables represents the distribution of the values of each variable. It can be seen some variables have missing values, which was fixed in the later part of the analysis. As we can see, the distribution of observations of each variable is close to normal distribution. ABA metabolites shows some visible outliers in boxplot in Figure 4. As we can see from the boxplot, the overall distribution of observations of ABA metabolites has a higher value than other variables such as sugar, TSS and moisture loss, whose observations lie on the same level in the boxplot. By seeing the histograms of residuals in Figure 5, we can see that the residuals for ABA and sugar is not normally distributed, for TSS, close to normal distribution, for moisture loss, the distribution is more normal like. The PCA plots in Figure 6, demonstrated clustering of samples based on similarity. We can see in Figure 6(c) for Treatment variable; proper clustering cannot be seen between the two groups AIR and DCA. This may be due to the data being not linear, or the there is some correlation between the two groups AIR and DCA. Though, for Time, clustering can be seen for the observations at 8th day and 28th day. Some of the observations in 28th day are highly correlated to the first dimension of PCA. DCA observations has the highest correlation to first and second dimensions of PCA as seen in Figure 6(c). We can see in the biplot, there is clustering of the observations in two different days. In the two biplots, glucose vector is at right angle to dimension 1, hence it contributes most to first principal component. ABA vector is close to a right angle with dimension 2, hence we can say that it contributes most to dimension 2. These observations are supported by Figure 8. From the scree plot, in Figure 7, we can infer the two principal components will be enough to keep. Also, in Figure 10, we can see that on performing HCA on the PCA score of first three dimensions, two dimensions are close in distance than the third dimension. By applying different linkages and distance methods, we see that the performance of the HCA analysis was better in case of applying Euclidean linkage and single distance method. This is because the dendrogram shows Dimension 1, 2, 3,4 in one cluster and Dimension 5 in another cluster, whereas in the other dendograms this is not the case. From Figure 9, we can see all the four regression models, TSS with sum of sugars, Glucose, Fructose, Sucrose each, are linear. Here the independent variables is TSS and dependent variables are sum of sugars, Glucose, Fructose, Sucrose. By looking at the R square value, we can interpret TSS and sum of sugars are negatively correlated, i.e. increase in value of TSS will cause a decrease in value of Sum of sugars, whereas TSS and all the other three variables, are positively correlated, increase in value of TSS will cause an increase in the value of the variables. Since, all the variables follow a linear regression model, we can conclude TSS can be used to predict the values of sum of sugars or any individual sugar. Figure 11 displays results after performing k-means clustering on raw data of asparagus and scaled data. We see the same results for both raw and scaled data. The first cluster shows all the sugars together, along with variables such as Moisture loss, TSS and DPA. DPA should have been in the same cluster as ABA, also PA and ABA should have been together too. Thus we can conclude k-means has not successfully clustered together samples belonging to the same group.

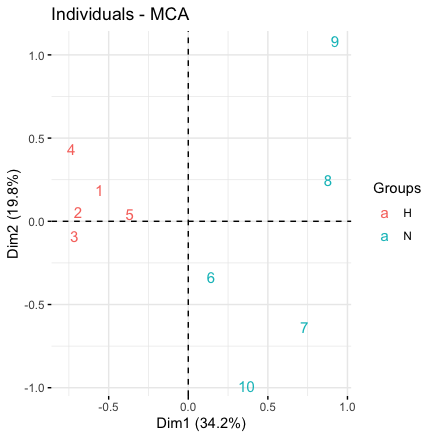
**Task 2:**

**Methodology:**

The csv file was loaded, and the columns were converted to factors. MCA plot of the data was generated, followed by plotting the individual scores plot and biplot of the MCA, grouping the variables according to BMI. Figure 11 displays the MCA individual scores plot and biplot generated.



(a)



(b)

Figure 12: (a) MCA individual scores plot (b) MCA biplot

**Results and discussion:** Multi-dimensional scaling plots are required when correlation between categorial variables needs to be visualized. From the individual plot, we can visualise 2nd, 3rd and 5th individual has the most correlation with dimension 1 and 4th individual has least correlation with dimension 2. From the biplot, we can see normal BMI group is more parallel to Dimension 1 than high BMI, hence normal BMI contributes more to first principal component than high BMI. Since all the categories of SNPs i.e. no SNP (NS), homozygous SNP (HS) and heterozygous SNP (HT) fall into the same cluster i.e. normal BMI, there is no difference between the effect of three categories of SNP on BMI, and since they correlate to normal BMI cluster, we can infer that presence of SNPs has no effect on BMI.

**Task 3**

**Methodology:**

The data was loaded in the workspace and moving average smoothing was performed on the raw data. The raw data was plotted to see general distribution of the data. Figure 13 (a) shows the plot. The raw data was filtered with a threshold 600, thereafter the highest incidences were visualized by plotting the filtered data. The plot can be visualized on Figure 13(b).

Chart, histogram

Description automatically generated

Fig 13: (a) Plot of raw data (b) Plot of filtered data of number of people at a threshold of 600

Moving average smoothing technique with window size 5 and 11 were applied on the raw data. The three graphs of raw data, moving average smoothing at a window size 5 and at a window size 11 were overlayed and plotted in a single graph. The graph is displayed at Figure 14.

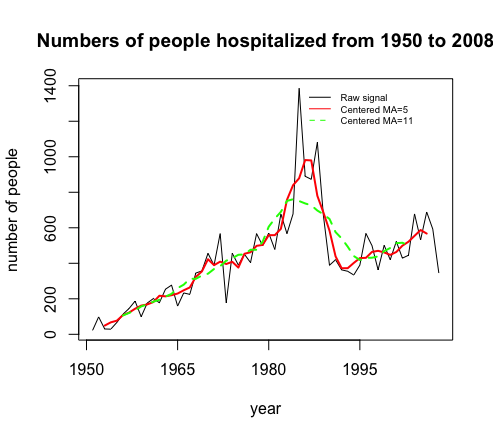


Figure 14: Plot of raw data (black), moving average at a window size 5 (red), moving average at a window size 11 (green)

Kernel smoothing was applied to the data henceforth. The bandwidth sizes for kernel smoothing were 5 and 2. The plot of raw data, kernel smooth at bandwidth sizes 5 and 2 were visualized in a single plot. The plot can be viewed at Figure 15.

Chart, histogram

Description automatically generated

Figure 15: Plot of raw data (black), kernel smoothed data at bandwidth 5 (red), at bandwidth 2 (green)

**Results and Discussion:** Smoothing techniques are used to exclude noise from a raw data. Usually, the wider the bandwidth we choose, smoother is the graph. As we can see in Figure 13, at a window size of 11 of Moving Average smoothing technique, the graph was smoother compared to a window size of 5. Similarly, for kernel smoothing technique, a bandwidth of 5 gave smoother results than a bandwidth of 2. Figure 12 (a) plotted the raw data, after using a threshold of 600 for the number of people the highest incidences were visualized around 1985 and 1990 by seeing the Figure 12 (b). Even after applying smoothing techniques, we can observe highest peak after the year 1980 in the graph, so we can conclude the ministry’s claim was false.