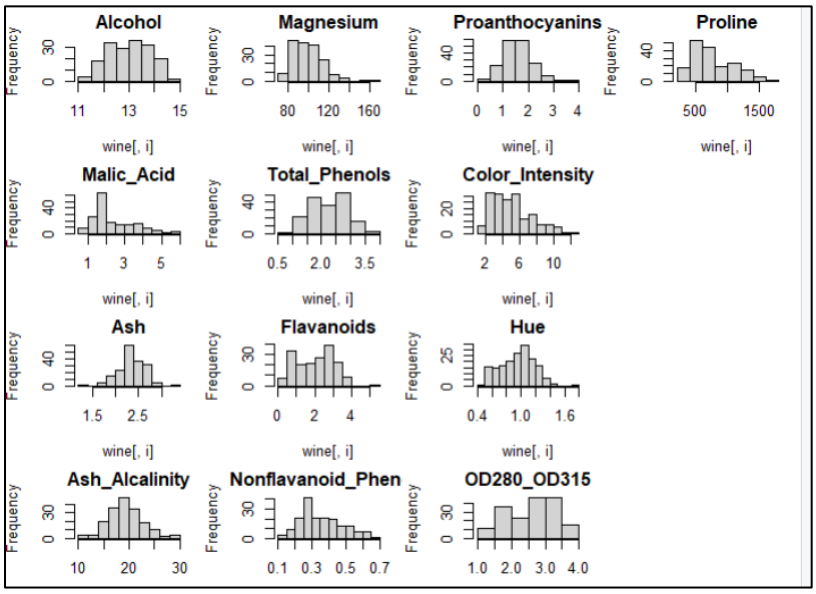
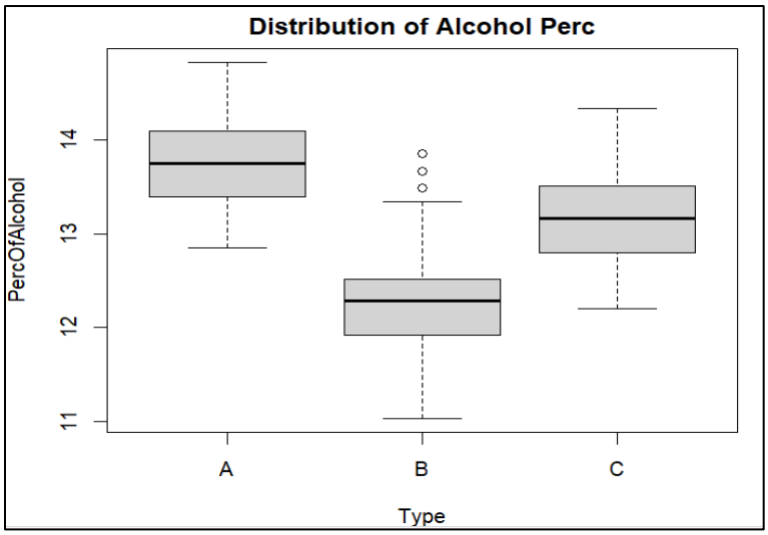
# Exploratory Data Analysis for Wine Dataset

**Histogram:**



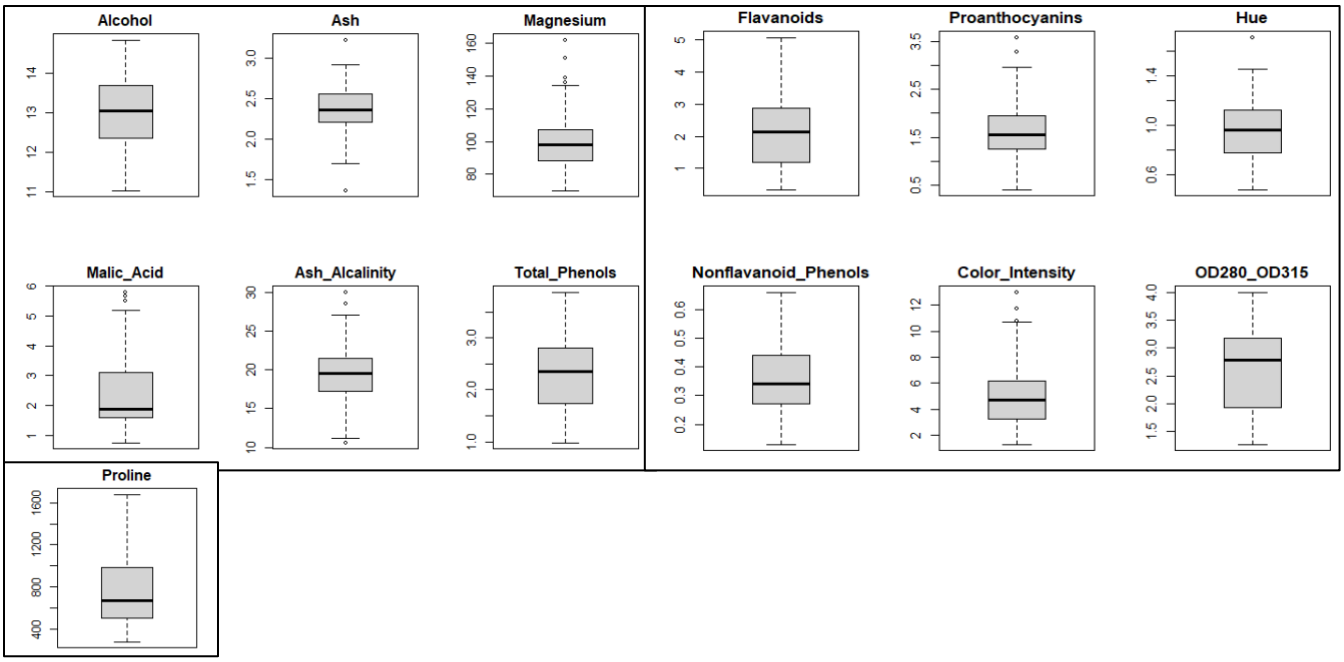

Most of histograms are approximately symmetric, but some show a possible right-skew and outliers. For example, Proline, Color Intensity, Ash Alcalinity, Malic Acid, and Magnesium all have main cluster of values at one end and a tail toward the other, meaning that there is an unusually high observation. For others, they look bell-shaped, indicating their value seems normally distributed.

**Distribution of Alcohol Percentage:**



The box plot shows the distribution of alcohol percentage between wine types. Type A has the highest alcohol percentage at around 14, having values between 13.5 and 14 while Type B has the lowest alcohol percentage at around 12 with most values falling between 12 and 12.5 and some are outliers. Type C stays between Type A and Type B as the median is higher than Type B but lower than Type A with values at around 13.

**Box Plot:**




The box plots show that some variables contain outliers. For example, Ash has one outlier above 3.0, Magnesium has four above 130, Proanthocyanins have two around 3.5, Hue has one above 1.6, Malic Acid have a few above 5, Ash Alcalinity have two above 27, and Color Intensity have 2 around 12. Other than that, everything looks fairly distributed with no extreme outliers but some variables like Proline have long whiskers, meaning a possible skewness.

# PCA

**1.**

**a) Provide the PCA outputs.**

**Without Normalization:**

```
Importance of components:
                            PC1      PC2     PC3     PC4     PC5     PC6   PC7   PC8
Standard deviation     314.963 13.13527 3.07215 2.23409 1.10853 0.91710 0.528 0.389
Proportion of Variance   0.998  0.00174 0.00009 0.00005 0.00001 0.00001 0.000 0.000
Cumulative Proportion    0.998  0.99983 0.99992 0.99997 0.99998 0.99999 1.000 1.000
                            PC9  PC10  PC11  PC12   PC13
Standard deviation     0.335 0.268 0.194 0.145 0.0906
Proportion of Variance 0.000 0.000 0.000 0.000 0.0000
Cumulative Proportion  1.000 1.000 1.000 1.000 1.0000
```

From the proportion of variance of above picture, we can see that almost all variation is explained by PC1 with 99.8%.

**With Normalization:**

```
Importance of components:
                          PC1   PC2   PC3    PC4    PC5    PC6    PC7    PC8    PC9
Standard deviation     2.169 1.580 1.203 0.9586 0.9237 0.8010 0.7423 0.5903 0.5375
Proportion of Variance 0.362 0.192 0.111 0.0707 0.0656 0.0494 0.0424 0.0268 0.0222
Cumulative Proportion  0.362 0.554 0.665 0.7360 0.8016 0.8510 0.8934 0.9202 0.9424
                          PC10   PC11  PC12    PC13
Standard deviation     0.5009 0.4752 0.411 0.32152
Proportion of Variance 0.0193 0.0174 0.013 0.00795
Cumulative Proportion  0.9617 0.9791 0.992 1.00000
```

After normalized, this is more balanced than the previous picture. The cumulative proportion from PC1 to PC5 explains around 80% of variance collectively.

**b) Provide the components and weights for the first 5 principal components for both the PCA with and without normalization.**

**Without Normalization:**

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| Alcohol | -0.0017 | -0.00120 | -0.017 | -0.141 | -0.020 |
| Malic_Acid | 0.0007 | -0.00215 | -0.122 | -0.160 | 0.613 |
| Ash | -0.0002 | -0.00459 | -0.052 | 0.010 | -0.020 |
| Ash_Alcalinity | 0.0047 | -0.02645 | -0.939 | 0.331 | -0.064 |
| Magnesium | -0.0179 | -0.99934 | 0.030 | 0.005 | 0.006 |
| Total_Phenols | -0.0010 | -0.00088 | 0.040 | 0.075 | -0.315 |
| Flavanoids | -0.0016 | 0.00005 | 0.085 | 0.169 | -0.525 |
| Nonflavanoid_Phenols | 0.0001 | 0.00135 | -0.014 | -0.011 | 0.030 |
| Proanthocyanins | -0.0006 | -0.00500 | 0.025 | 0.050 | -0.251 |
| Color_Intensity | -0.0023 | -0.01510 | -0.291 | -0.879 | -0.332 |
| Hue | -0.0002 | 0.00076 | 0.026 | 0.060 | -0.052 |
| OD280_OD315 | -0.0007 | 0.00350 | 0.070 | 0.178 | -0.261 |
| Proline | -0.9998 | 0.01777 | -0.005 | 0.003 | 0.002 |

Proline explained entirely in PC1, PC2 with Magnesium, PC3 with Ash Alcalinity, PC4 with Color Intensity, PC5 with Malic Acid and Flavanoids.

**With Normalization:**

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| Alcohol | -0.144 | -0.484 | -0.21 | -0.02 | 0.27 |
| Malic_Acid | 0.245 | -0.225 | 0.09 | 0.54 | -0.04 |
| Ash | 0.002 | -0.316 | 0.63 | -0.21 | 0.14 |
| Ash_Alcalinity | 0.239 | 0.011 | 0.61 | 0.06 | -0.07 |
| Magnesium | -0.142 | -0.300 | 0.13 | -0.35 | -0.73 |
| Total_Phenols | -0.395 | -0.065 | 0.15 | 0.20 | 0.15 |
| Flavanoids | -0.423 | 0.003 | 0.15 | 0.15 | 0.11 |
| Nonflavanoid_Phenols | 0.299 | -0.029 | 0.17 | -0.20 | 0.50 |
| Proanthocyanins | -0.313 | -0.039 | 0.15 | 0.40 | -0.14 |
| Color_Intensity | 0.089 | -0.530 | -0.14 | 0.07 | 0.08 |
| Hue | -0.297 | 0.279 | 0.09 | -0.43 | 0.17 |
| OD280_OD315 | -0.376 | 0.164 | 0.17 | 0.18 | 0.10 |
| Proline | -0.287 | -0.365 | -0.13 | -0.23 | 0.16 |

With Normalized PCA, the output gives a fair contribution between variables. PC1 are Flavanoids, Total Phenols, and Proanthocyanins. PC2 with Color Instensity and Alcohol, PC3 with Ash and Ash Alcalinity. PC4 with Malic Acid and Proanthocyanins. PC5 with Magnesium.

**2. Consider the rows labeled "Proportion of Variance."**

**a) Explain why the value for PC1 is so much greater than that of any other column. (This should be in paragraph form.)**

PCA works on the covariance matrix, meaning if one variable has huge variance, its diagonal value will dominate the matrix. In this case, the unnormalized PCA has PC1 takes up entirely due to a variable named Proline. Due to the difference in measurement, Proline

has a much larger value that is in hundreds while other small variables are either in tens or ones, leading Proline's variance also numerically higher than another variable's variance.

**b) Comment on the use of normalization (standardization) in part (a). Comparing the Proportion of Variance between the PCA outputs with and without normalization.**

After we used standardization, the effect of scale is removed. PCA now is based on the correlation matrix instead of the covariance matrix. Each variable now contributes to the variance equally, regardless of its measurement unit. With unnormalized PCA, PC1 takes entirely the proportion of variance by 98% while other PC have almost 0%. Normalized PCA gives more balanced distribution as PC1 has 36%, PC2 with 19%, PC3 with 11%, PC4 with 7% and PC5 with 6%. These five components together explain around 80% of the total variance.