STAT 4051 HW3

Ruwiada Al Harrasi

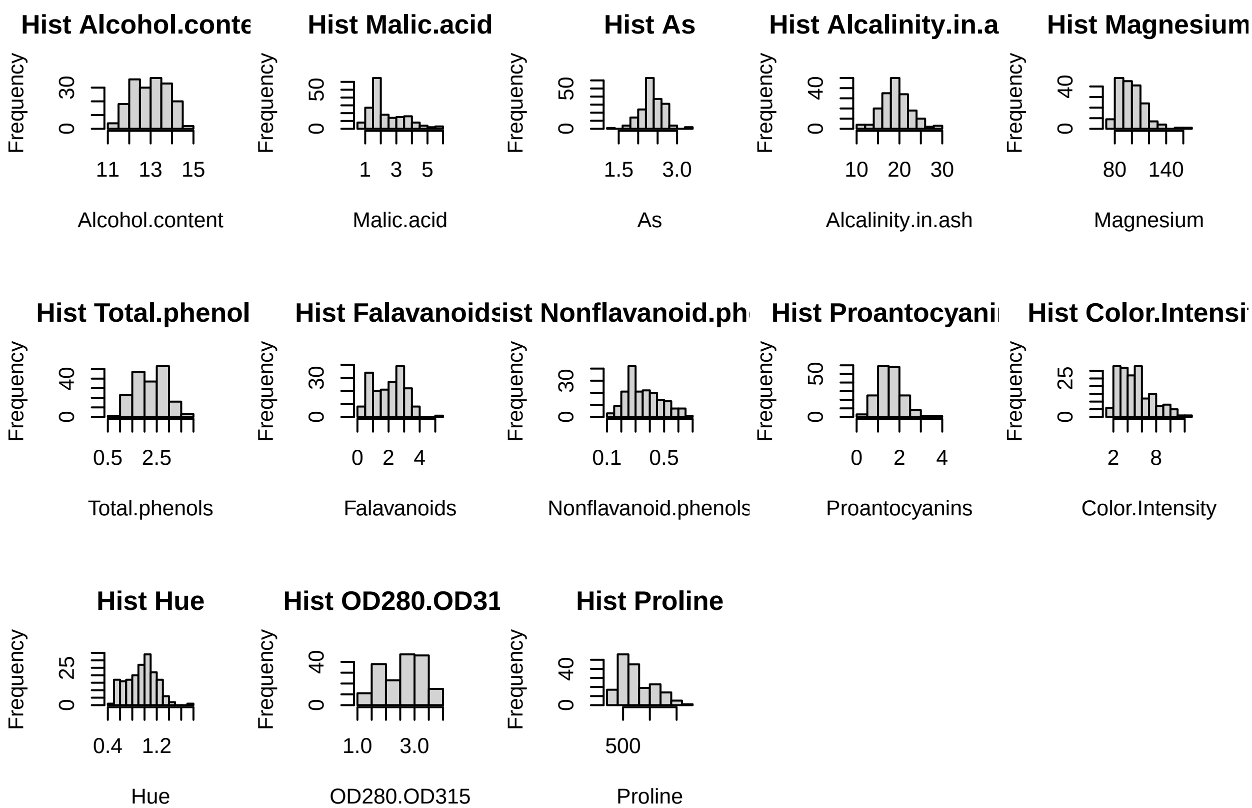
2023-10-13

```r
wine <- read.csv("~/Downloads/wine.csv")
summary(wine)
```

```
##  Alcohol.content   Malic.acid          As        Alcalinity.in.ash
##  Min.   :11.03   Min.   :0.740   Min.   :1.360   Min.   :10.60
##  1st Qu.:12.36   1st Qu.:1.607   1st Qu.:2.210   1st Qu.:17.20
##  Median :13.05   Median :1.870   Median :2.360   Median :19.50
##  Mean   :12.99   Mean   :2.341   Mean   :2.366   Mean   :19.49
##  3rd Qu.:13.67   3rd Qu.:3.047   3rd Qu.:2.553   3rd Qu.:21.50
##  Max.   :14.83   Max.   :5.800   Max.   :3.230   Max.   :30.00
##    Magnesium      Total.phenols    Falavanoids    Nonflavanoid.phenols
##  Min.   : 70.00   Min.   :0.980   Min.   :0.340   Min.   :0.1300
##  1st Qu.: 88.00   1st Qu.:1.715   1st Qu.:1.097   1st Qu.:0.2700
##  Median : 98.00   Median :2.335   Median :2.120   Median :0.3400
##  Mean   : 99.64   Mean   :2.284   Mean   :2.012   Mean   :0.3627
##  3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.865   3rd Qu.:0.4425
##  Max.   :162.00   Max.   :3.880   Max.   :5.080   Max.   :0.6600
##  Proantocyanins  Color.Intensity       Hue            OD280.OD315
##  Min.   :0.410   Min.   : 1.280   Min.   :0.4800   Min.   :1.270
##  1st Qu.:1.235   1st Qu.: 3.240   1st Qu.:0.7775   1st Qu.:1.905
##  Median :1.545   Median : 4.750   Median :0.9600   Median :2.780
##  Mean   :1.581   Mean   : 5.087   Mean   :0.9536   Mean   :2.602
##  3rd Qu.:1.950   3rd Qu.: 6.213   3rd Qu.:1.1200   3rd Qu.:3.170
##  Max.   :3.580   Max.   :13.000   Max.   :1.7100   Max.   :4.000
##     Proline
##  Min.   : 278
##  1st Qu.: 500
##  Median : 666
##  Mean   : 744
##  3rd Qu.: 985
##  Max.   :1680
```

(a) Describe the data and present some initial pictorial and numerical summaries, such as scatterplots, histograms, etc.

```r
# Set up a 3x3 grid for histograms
par(mfrow=c(3,5))
# Loop through each variable in the wine dataset and create histograms with overlaid normal curves
for (col in colnames(wine)) {
  hist(wine[[col]], xlab=col, main=paste("Hist", col ))
}
```

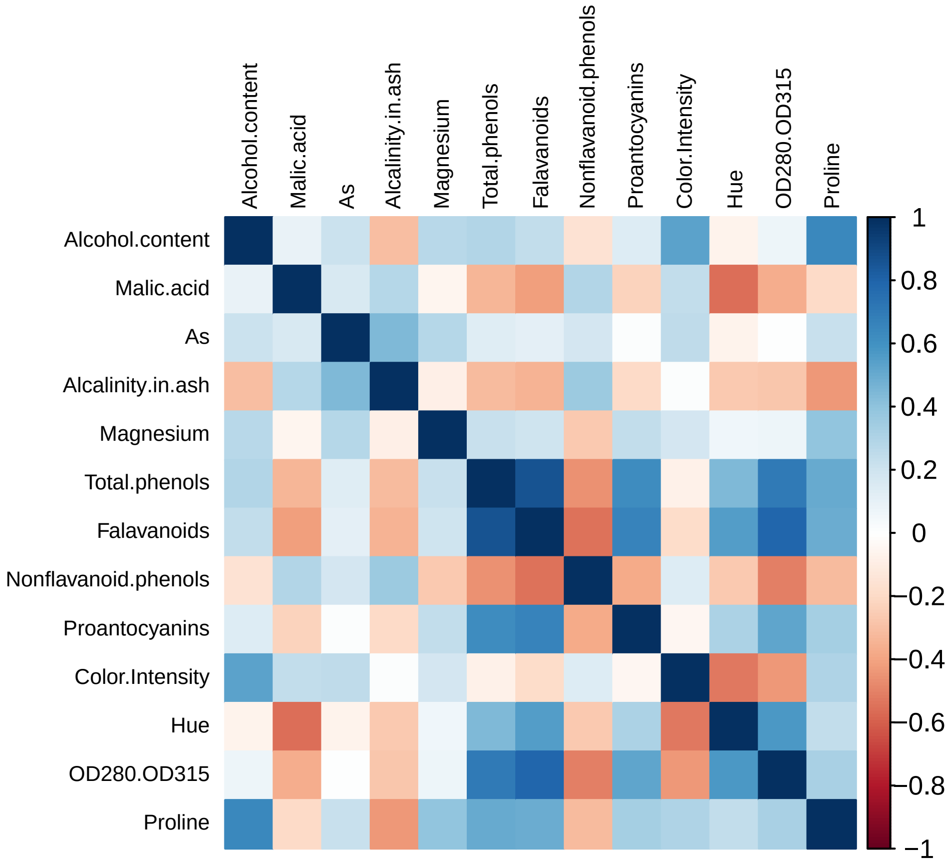


I can tell that the histogram for Alcalinity.in.ash and Proantocyanins appears to follow a roughly normal distribution, while the histograms for the remaining variables do not exhibit a distinct pattern. Some of them resemble a uniform distribution, while others appear to be skewed.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(wine), method = "color",tl.cex = 0.7, tl.col = "black")
```



I can discern a moderately negative correlation (-0.558) between Hue and Malic.acid. In contrast, there exists a relatively strong positive correlation (0.868) between Falavanoids and Total.phenols.content, and a substantial positive correlation (0.791) between Falavanoids and OD280.OD315. Additionally, there is a notable positive correlation (0.6450) between Proline and Alcohol. However, the remaining variables do not display notably robust correlations.

(b) Comment on whether PCA on covariances or correlations makes more sense for this dataset. Make your decision and proceed. PCA is applied to the covariance matrix of centered data (mean 0) Alternative: standardize all variables first (mean 0, sd 1); This is equivalent to applying PCA to the correlation matrix. First I will check the mean and var of the varibles

```
apply(wine, 2, mean)
```

```
##      Alcohol.content           Malic.acid                   As
##           12.9948889            2.3405000            2.3657778
##    Alcalinity.in.ash            Magnesium        Total.phenols
##           19.4922222           99.6444444            2.2842778
##          Falavanoids Nonflavanoid.phenols       Proantocyanins
##            2.0122778            0.3627222            1.5808333
##      Color.Intensity                  Hue          OD280.OD315
##            5.0874333            0.9536444            2.6018889
##              Proline
##          744.0388889
```

Magnesium has a significantly higher mean compared to Hue and Nonflavanoid.phenols, with the difference being several orders of magnitude.Proline also has a much higher mean compared to other variables, being approximately 7 times higher than Magnesium and 14 times higher than Color.Intensity.

```
apply(wine, 2, var)
```

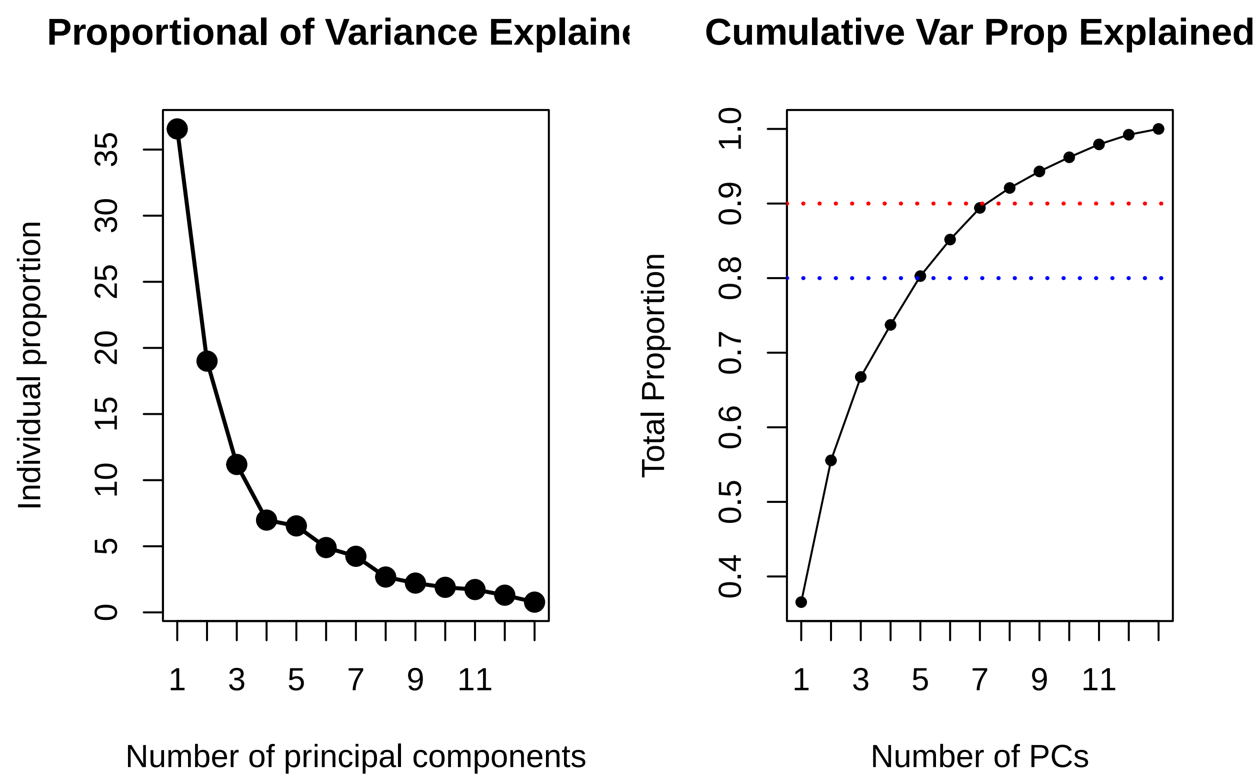
```
##      Alcohol.content           Malic.acid                   As
##         6.555436e-01         1.236758e+00         7.447705e-02
##    Alcalinity.in.ash            Magnesium        Total.phenols
##         1.102944e+01         2.028338e+02         3.978738e-01
##          Falavanoids Nonflavanoid.phenols      Proantocyanins
##         1.012412e+00         1.540093e-02         3.330244e-01
##      Color.Intensity                  Hue          OD280.OD315
##         5.445489e+00         5.297958e-02         5.071551e-01
##              Proline
##         9.879235e+04
```

The variables also have vastly different variances. If I failed to scale the variables before performing PCA, then most of the principal components that I observed would be driven by the Proline variable, since it the largest mean and variance. Thus, it is important to standardize the variables to have mean zero and standard deviation one before performing PCA to makes the analysis independent of units. (c) Comment on the percentage of variance explained and number of principal com- ponents to retain. Include a scree plot.

```
pca1<- prcomp(wine, scale = TRUE)
pca1.var <- pca1$sdev^2
pve <- (pca1.var / sum(pca1.var))*100
pve
```

```
##  [1] 36.5604190 19.0024805 11.1858532  6.9794715  6.5411569  4.9007825
##  [7]  4.2420645  2.6739649  2.2163832  1.8954638  1.7265574  1.2996523
## [13]  0.7757503
```

```
par(mfrow=c(1,2))
## pve plot
pci = summary(pca1)$importance
plot(y=pve, x=c(1:13), pch=20, cex=2, xaxt='n',
     main='Proportional of Variance Explained',
     ylab='Individual proportion', xlab='Number of principal components')
lines(y=pve, x=c(1:13), lwd=2)
axis(1, at=1:13)
pve <- pci[3, ]
plot(y=pve, x=c(1:13), pch=20, cex=1, xaxt='n',
     main='Cumulative Var Prop Explained',
     ylab='Total Proportion', xlab='Number of PCs')
lines(y=pve, x=c(1:13), lwd=1)
axis(1, at=1:13)
abline(h=0.9, lwd=2, lty=3, col='red')
abline(h=0.8, lwd=2, lty=3, col='blue')
```



We see that the first principal component explains 36.56% of the variance in the data, the next principal component explains 19 % of the variance, and the third explains 11% and so forth.5 PCs are needed to explain 80% of the variance, and 7 PCs are needed to explain 90% of the variance (d) Comment on variable loadings and their potential interpretations.

```
pci <- summary(pca1)$importance
pci
```

```
##                              PC1      PC2      PC3       PC4       PC5       PC6
## Standard deviation     2.180104 1.571726 1.205886 0.9525394 0.9221445 0.7981865
## Proportion of Variance 0.365600 0.190020 0.111860 0.0697900 0.0654100 0.0490100
## Cumulative Proportion  0.365600 0.555630 0.667490 0.7372800 0.8026900 0.8517000
##                              PC7       PC8       PC9      PC10      PC11
## Standard deviation     0.7426092 0.5895892 0.5367773 0.4963973 0.4737641
## Proportion of Variance 0.0424200 0.0267400 0.0221600 0.0189500 0.0172700
## Cumulative Proportion  0.8941200 0.9208600 0.9430300 0.9619800 0.9792500
##                             PC12     PC13
## Standard deviation     0.4110411 0.317565
## Proportion of Variance 0.0130000 0.007760
## Cumulative Proportion  0.9922400 1.000000
```

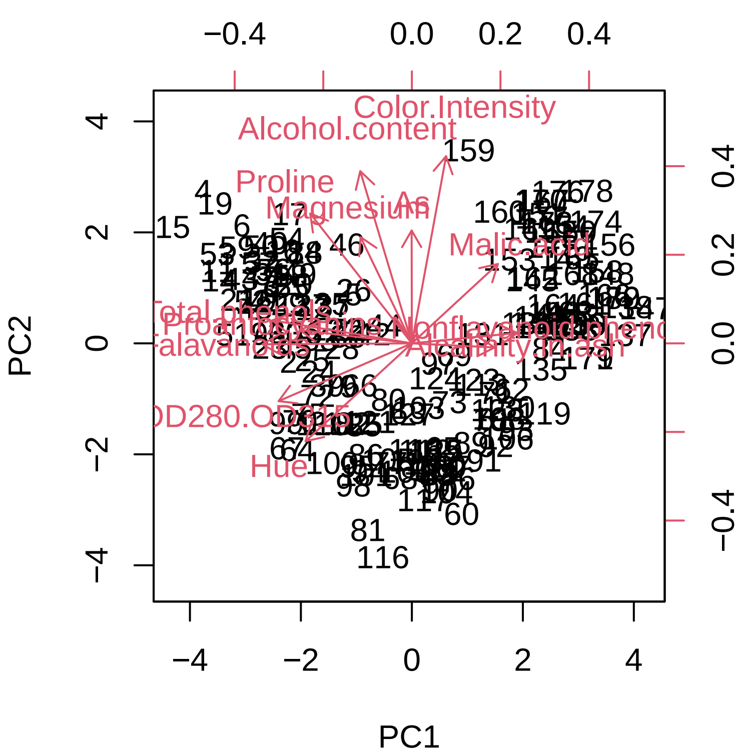
```
# Access the loadings from the PCA results
pca1$rotation[, 1:4]
```

```
##                                    PC1          PC2         PC3          PC4
```

```
## Alcohol.content      -0.1457288808  0.485827735 -0.20382051  0.004072664
## Malic.acid            0.2428354514  0.224032880  0.09609523 -0.550126299
## As                   -0.0007060265  0.318472138  0.62360105  0.208865746
## Alcalinity.in.ash     0.2334324671 -0.011585548  0.61597578 -0.059936867
## Magnesium            -0.1447089251  0.299734167  0.13255852  0.366250653
## Total.phenols        -0.3940942090  0.064829048  0.14148803 -0.200468292
## Falavanoids          -0.4216916487 -0.002769256  0.14529847 -0.154741411
## Nonflavanoid.phenols  0.2972878956  0.029408946  0.17228021  0.196104431
## Proantocyanins       -0.3158148542  0.037506016  0.14841814 -0.388020419
## Color.Intensity       0.0964965691  0.528046077 -0.14426881 -0.061824262
## Hue                  -0.3001334193 -0.276220055  0.08779118  0.415655715
## OD280.OD315          -0.3761487697 -0.163121690  0.16104563 -0.189495491
## Proline              -0.2867566693  0.366279206 -0.12669382  0.224361123
```

PC1 has the highest standard deviation, which means it explains the most variance in the data. Subsequent PCs have decreasing standard deviations, explaining less variance.PC1 explains approximately 36.56% of the total variance, PC2 explains about 19.00%, and so on. It provides insights into how much information is retained by each PC.Cumulative Proportion allows us to assess how much information is retained when considering a subset of PCs.In this case, the first two PCs (PC1 and PC2) explain approximately 55.56% of the total variance, while the first four PCs capture about 73.73%. (e) Make a plot of the data projected on the first two PCs. Comment on any interesting features, including potential outliers, if any.

```r
library(ggplot2)
biplot(pca1, scale = 0)
```

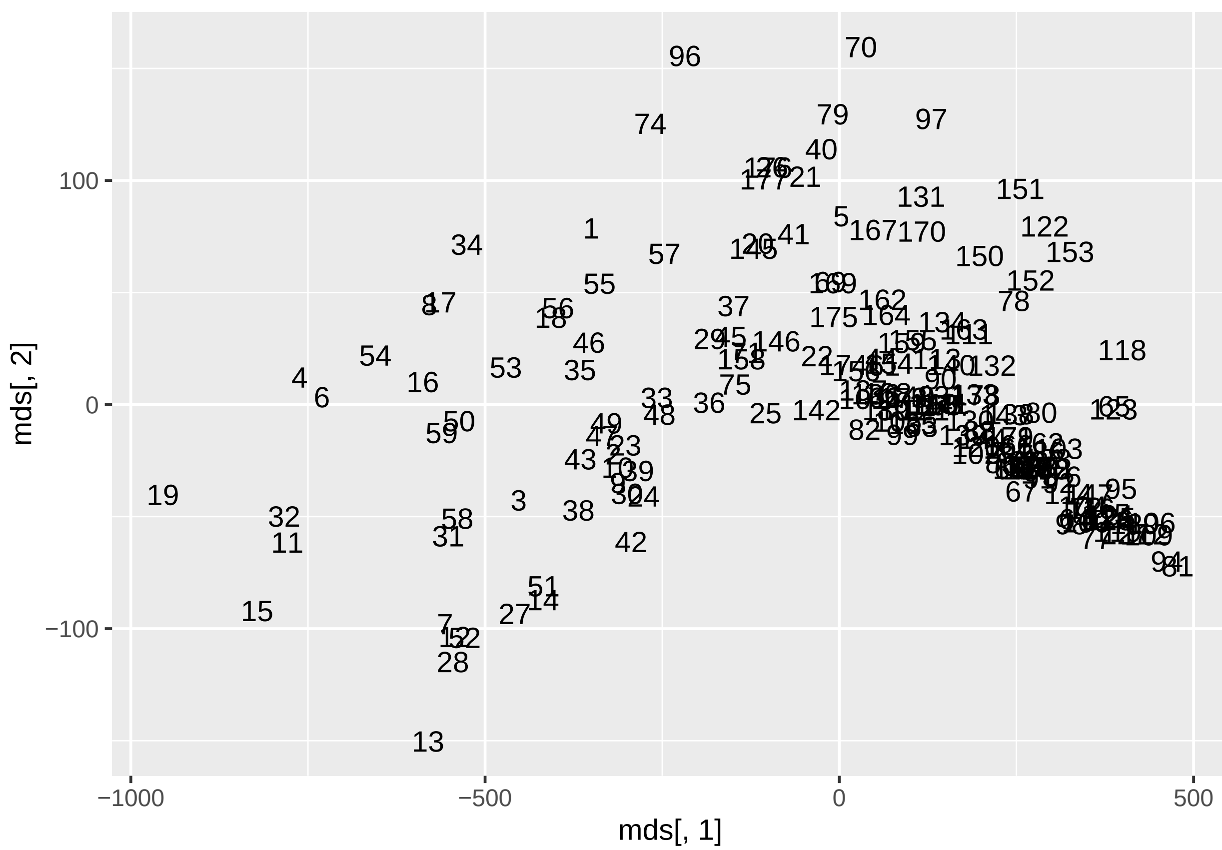


Total phenols, Proanthocyanins, and Flavanoids: These three variables appear close to each other on the biplot, suggesting a strong positive relationship. States with high values of these variables tend to cluster together, sharing a similar profile. This indicates that wines with high levels of total phenols are also likely to have high levels of proanthocyanins and flavanoids.Falavanoids and Alcalinity appear in opposite directions from the origin on PC1, which indicates a negative correlation. States with high falavanoid values tend to have low alcalinity. This suggests an inverse relationship between these variables. Alcohol Content and Color Intensity are positively correlated in PC2. States with high alcohol content also tend to have high color intensity in their wines.Hue and OD280.OD315: Hue and OD280.OD315 are positioned far from the center, indicating that there may be outliers.

2.On the same data set, perform multidimensional scaling to 2-D by selecting a dissimilarity other than the Euclidean distance. You can specify other aspects of the algorithm. Plot the resulting 2-D results. Comment on its comparison with your PCA 2-D plot.

```
# Calculate the distance matrix using Manhattan distance
d <- dist(wine, method = "manhattan")
mds <- cmdscale(d, k = 2)
par(mfrow=c(2,1))
qplot(x = mds[, 1], y = mds[, 2], label=rownames(wine), geom="text")
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



I can see that both PCA and MDS clustered data points 19 and 4 close to each other, and they also clustered data points 140 and 141 close to each other. There are more data points clustered together in mds than pca. And the outliers are different. This suggests that these data points share similarities, despite the different shapes of the plots. PCA and MDS are used for dimensionality reduction and data visualization, their underlying methodologies and objectives differ. PCA is primarily focused on variance maximization, while MDS is focused on distance preservation. As a result, they can produce different plots and emphasize different aspects of the data.