Exploratory Data Analysis including Dimension Reduction using PCA:

Data Set: White Wine Factors (33 observations, 12 variables)

Environment Used for Analysis: RStudio

Analysis done by: SAHANA MUKHERJEE, UBIT: mukherj3

Mail : mukherj3@buffalo.edu

- **Step 1:**

My first step in the exploratory data analysis is to **load the data** into our concerned environment and read the head of it (which usually consists of first 5 observations).

We read our data using the following command.

setwd("C:\\Users\\SAHANA\\Documents\\Statistical Data Mining 1\\RdataSets")

WhiteWineData <- read.csv("wine.csv", sep=";", header = TRUE)

head(WhiteWineData)

- Observation: Our data set named as WhiteWineData consists of a total of 12 variables, where the first 11 are the independent variables and the last column being the 12th one, ‘quality’ is the dependent variable which depends on the 11 independent variables.

- **Step 2:**

We begin exploring our data. The first step in any exploration would be to see what are the measures of central tendency and how the data is spread around the center.

To do the above, we do the following:

summary(WhiteWineData)

- Observation: The above command gives a very good overview of our data.It tells us about the measures of central tendency of the individual variables, and how they are spread around the mean. This summary also helps us to identify if the data contains outliers. When mean >> median, we can say, that the data is right skewed, and there are outliers to the right, and if mean << median the data is left skewed. A point is said to be a outlier if it is below [Q1- (1.5)IQR] or above [Q3+(1.5)IQR].

The below table gives us the summary of the data set. Two variables have been identified to have outliers, and marked in red, after scanning through the summary.

Furthermore, we do a box-plot of the independent variables, which helps us to visualize the outliers in our variable (Figure 1)

Statistics	fixed.acidity	volatile.acidity	citric.acid	residual.sugar
Min	6.2	0.14	0.04	1
1st Quartile	6.6	0.24	0.34	1.3
Median	7	**0.27**	0.38	**1.7**
Mean	7.191	**0.305**	0.36	**5.08**
3rd Quartile	7.9	0.31	0.41	7.5
Max	8.6	0.67	0.62	20.7

Statistics	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
Min	0.029	7	47	0.98	2.98	0.35	8.8	5
1st Quartile	0.04	17	100	0.9917	3.17	0.44	9.7	6
Median	0.045	30	**132**	0.9938	3.22	0.48	10.1	6
Mean	0.044	29.27	**129.2**	0.994	3.229	0.48	10.48	6
3rd Quartile	0.05	37	146	0.9955	3.3	0.53	11	6
Max	0.07	56	245	1	3.54	0.71	12.8	8

Sample code for boxplot:

```
boxplot(WhiteWineData$residual.sugar, col="slategray2", pch=19)
mtext("Residual Sugar", cex=0.8, side=1, line=2)
```

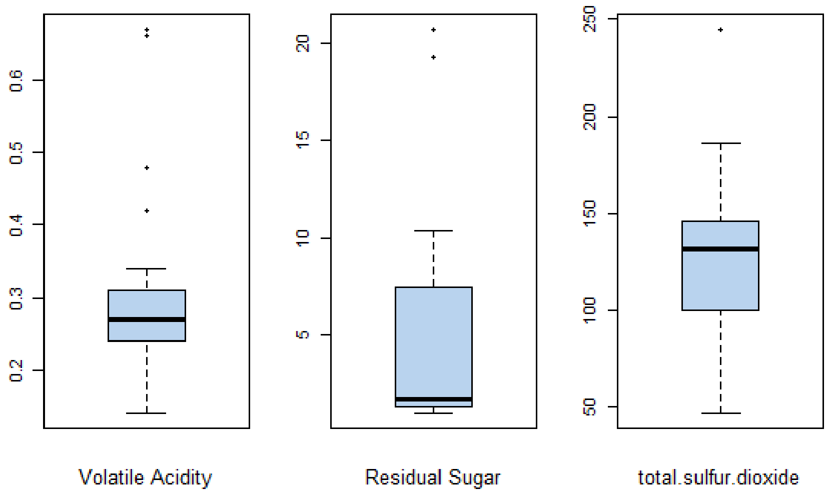


Figure 1

So we see, that the above three components are having some outliers, which we could get rid of, however, our data set is pretty small with only 33 observations, and the outliers are not that much significant, so we do not proceed to get rid of them, but prefer to keep them.

- **Step 3:**

Now that we have explored our independent variables, we would proceed to analyse which are the most significant variables that contribute to the dependent variable ‘quality’. For this we would do **Principal Component Analysis**. But before doing that we would first see, how the independent variables are corelated among themselves and visualize the same.

```
cor(WhiteWineData[,-12])
pairs(WhiteWineData[,-12], gap=0.9, pch=19, cex=0.4, col="darkblue")
```

The above two commands let us know that the following independent variables are having corelation greater than 0.5, which we consider as significant.

- Observation: We find that ‘density’ and ‘residual sugar’ are strongly corelated positively, ‘density’ and ‘alcohol’ are strongly negatively corelated, ‘free.sulfur.dioxide’ and ‘total.sulfur.dioxide’ are also sharing strong positive corelation. For example, density and alcohol are having as large as -0.75 correlation, alcohol and citric acid having as much as 0.60 correlation and so on, which can be found out from the output of the above command.

The ideal scenario should be, the independent variables should share no corelation among themselves, that is there should not be any co-linearity among the independent variables, they should only be related to output variable. But in real data sets, like this one, we can see that there is corelation between the independent variables. Below is the scatter plot of the independent variables, showing how they are correlated with each other (Figure 2).

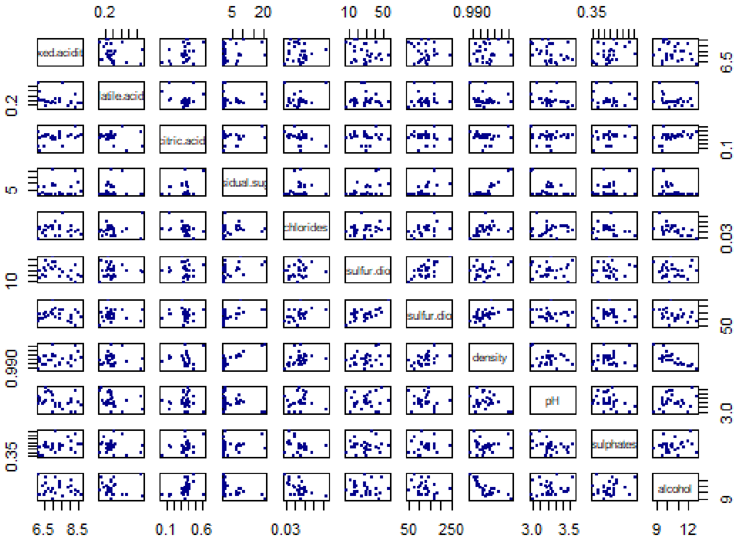


Figure 2

- **Step 4:**

Now, we proceed to do PCA on our data set, to identify which variables most significantly contribute to the dependent variable ‘quality.’ We normalise the data before performing PCA.

```
PerformPCA <-prcomp(WhiteWineData[,-12],center = TRUE,scale. = TRUE)

summary(PerformPCA)
```

The above command performs PCA, and gives us 11 principal components each one of which are composed of linear combination of all the 11 independent variables, weighed with different weights, as per respective Eigen values.

Below is the summary output:

```
Importance of components:

                         PC1    PC2    PC3    PC4     PC5     PC6     PC7

Standard deviation     1.9151 1.4600 1.1913 1.0986 0.92248 0.81927 0.67681

Proportion of Variance 0.3334 0.1938 0.1290 0.1097 0.07736 0.06102 0.04164

Cumulative Proportion         0.3334 0.5272 0.6562 0.7659 0.84330 0.90432 0.94596

                          PC8     PC9    PC10    PC11

Standard deviation     0.58609 0.41764 0.25796 0.09994

Proportion of Variance 0.03123 0.01586 0.00605 0.00091

Cumulative Proportion 0.97719 0.99304 0.99909 1.00000
```

- ✓ Observation: Maximum variation is explained by PC1(33%), followed by PC2 and so on. This is visualised using the following plots: The right plot is a scree plot, showing that maximum variance is explained by PC1, followed by the rest, and plot on the right is a biplot, showing how the individual components are correlated to PC1 and PC2.

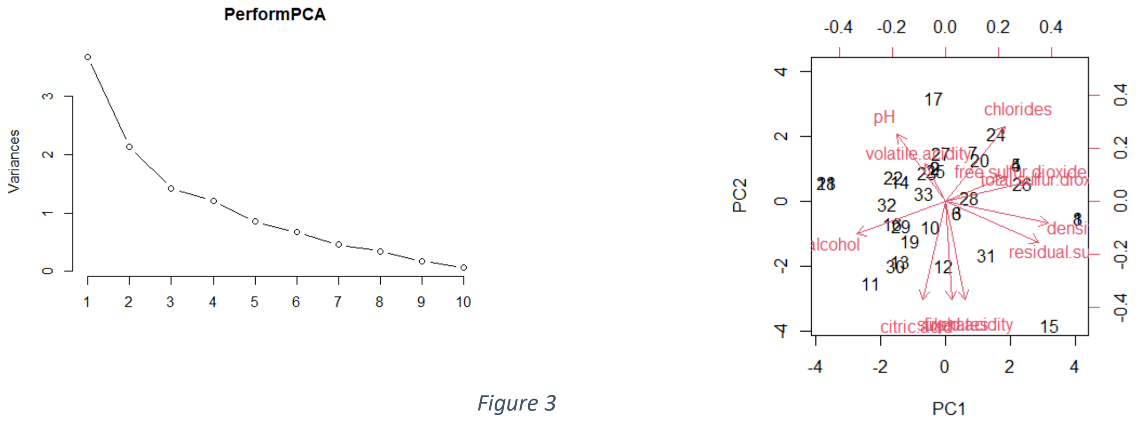


Figure 3

Commands used for figure 3

```
screeplot(PerformPCA, type = "lines")
biplot(PerformPCA,scale = 0)
```

The plots are explained further in consequent steps.

- **Step 5:**

Interpreting the PCA:

From the above table, we observe that in order to obtain around 90% variability, we need 6 principal components. This implies, that with the linear combination of the 11 variables, and weighing them with respective weights, we can reduce the dimension of the data set from 11 independent variables to 5 or 6 independent variables, to get around 84-90% variability. These few principal components are sufficient to explain 84-90% variability in the data set, and we can choose the optimum number of PCAs depending on business requirement.

Understanding the contribution of individual components:

Command Used:

```
Print(PerformPCA)
```

This command helps us understand the loading and can conclude which individual components are contributing more to the PCAs. For example, following is the loading for PC1.

fixed.acidity	0.09187657
volatile.acidity	-0.09565107
citric.acid	-0.10721373
residual.sugar	0.44097603
chlorides	0.27973962
free.sulfur.dioxide	0.28696547
total.sulfur.dioxide	0.39040733
density	0.48561952
pH	-0.22678291
sulphates	0.03065707
alcohol	-0.41861311

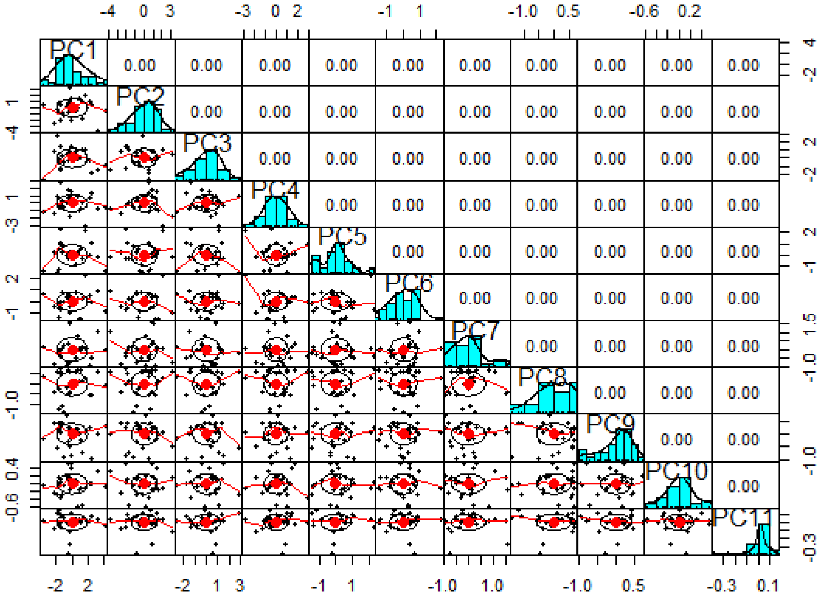
here we can conclude that density, residual.sugar contribute mostly (positive loading), and alcohol (negative loading)for PC1 (as observed in biplot in step 4)as their weights are more meaning PC1 is primarily having those components as major contributors. Similar analysis for PC2 lets us know that fixed.acidity, citric.acid and sulphates(negative loading) are mostly contributing for PC2. In this way we can observe how the individual components are weighed and then combined to form the PCs, so as to capture maximum information.

Final Conclusion and Path Forward:

As discussed in Step 3; there should not be any collinearity between the independent variables, meaning 0 correlation between them. PCA allows us to get rid of that, the PCAs formed do not have any collinearity between them. From about 0.70 correlation among two independent variables in step 3, now we have 0 correlation among the independent variables. Following graph shows this observation:

Command for scatter plot of scores of PCA:

pairs.panels(PerformPCA$x,gap = 0)



So now that we have the principal components which are independent of each other, we can choose only 6 PCs, which will be sufficient to explain 90% of the variability in our data, retaining as much information as possible. Thus our data set of 12 variables can be reduced to 7 or 6 variables (depending on the total variability demanded by requirement, say we want 80% variability, then we can go for only 6 PCs) and with this reduced data set, our next steps for predictive analysis would be simpler and convenient to visualize.