UTD 2016 Spring

CS6301.5U1 Advanced Computational Methods for Data Science

Assignment 2 – Linear Regression and PCA

Name: Yu Zhang, Mingyue Sun

NetID: yxz141631, mxs151730

Overview:

In this homework, we are going to deal with two datasets. First, the Auto-MPG dataset, in which we are going to apply the multiple linear regression analysis method to look for a suitable model for it. Then in the second Wine-Quality dataset, we are going to apply the principle component analysis technique and try to explore how PCA works in R with this particular dataset.

Part 1. Multiple Linear Regression analysis on the Auto-MPG dataset.

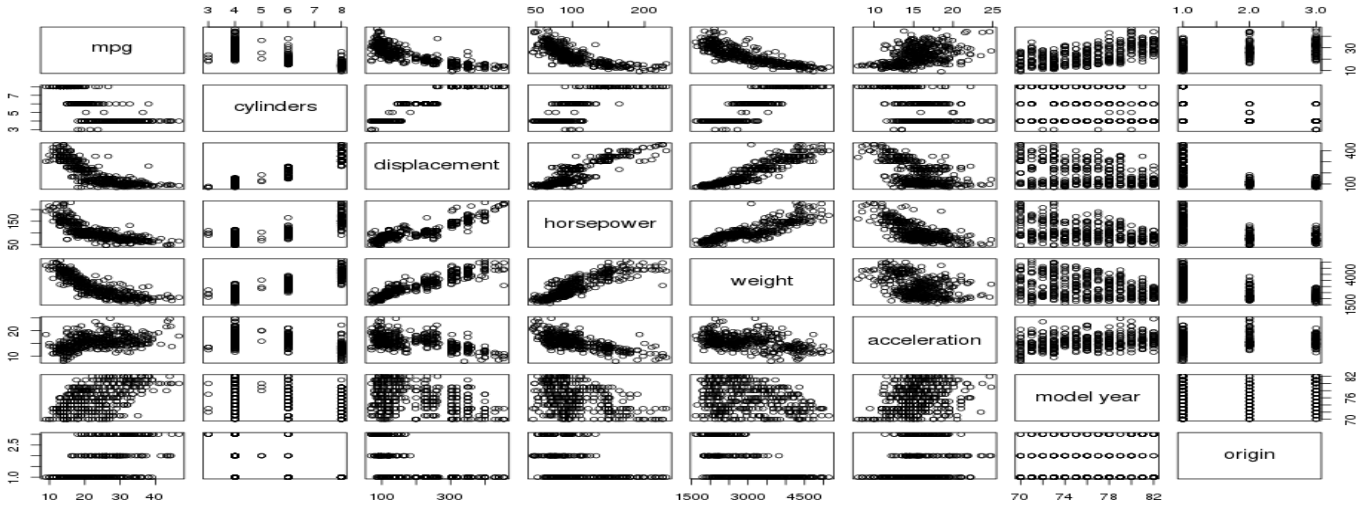
This dataset concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multi-valued discrete and 5 continuous attributes.

1.1Data cleaning and discovery in linearity of predicator attributes and the class label

At the beginning of our research on this dataset, we did the cleaning of data such removing the last attribute in the dataset which is a unique label “car name” for each instance as we think it’s totally irrelevant with a car’s MPG and removing several instances whose “horsepower” attribute contains simply a ‘?’ sign and we could not perform any analysis on such value.

So with this “cleaned” dataset, we simply plot it out as shown in Scheme 1 and we would like to discover some linear relationship between those seven predicator attributes with class attribute. From the scheme we can see for these four continuous attributes: “displacement”, “horsepower”, “weight” and “acceleration”, three of them have strong linear relationship between each of them and the class attribute “MPG”. They are “displacement”, “horsepower” and “weight”, and we could tell from it that the larger value these attributes have, the lower the “MPG” value will be. And that’s also how we think that the fact would be: a car with larger “displacement”, “horsepower” and higher “weight” will surely have a poor performance in the “MPG” part. Only the “acceleration” we decided it’s a non-linear term for the class label.

While for those three multi-valued discrete attributes, although it’s not so easy to tell the linearity of each of one on the class attribute, we can however after removing several instances which have a “cylinders” value of 3 or 5 since they are rare cases in the dataset and we think these cases could not represent the relationship between the “cylinders” and “MPG” class and then looking for the medium in each categories of every multi-valued discrete attributes, we could see a linear tendency between all these three attributes and the class label. And the less the number of cylinder, the newer the model year and the car with origin as “3” (probably Japan☺), the higher the “MPG” value would be. It also coincides with what we would expect in the real world: a newer car with only 4 cylinders and Japan-made (these cars are famous for its “reliability” and “fuel economic performance”) would definitely outperform another one with older year, 8 cylinders and US-made, just in term of “MPG” (not considering other factors like “ride comfort”, “safety” and so on).



Scheme 1. Potential linearity of predicator attributes with class attribute

So after we made some judgments on the linearity of predicator attributes and the class label from the plot of original data. We tried to build up a multiple linear regression model still with all predicator attributes regardless of its linearity. The result is summarized in Scheme 2.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
`model year`   0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

Scheme 2. Summary of multiple linear regression model on the original cleaned dataset

From this scheme we could tell that the R^2 value is 0.8215 and the adjusted R^2 value is 0.8182 which is quite acceptable for us. In the coefficients part, we did notice a quite high p-value for that non-linear term “acceleration” which is not a tolerable value (0.41548) to us and therefore in the follow-up step, we simply kicked this non-linear term out and constructed the linear model again. The result is summarized in Scheme 3 down there.

1.2 Re-modeling

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.556e+01  4.175e+00  -3.728 0.000222 ***
cylinders    -5.067e-01  3.227e-01  -1.570 0.117236
displacement  1.927e-02  7.472e-03   2.579 0.010287 *
horsepower   -2.389e-02  1.084e-02  -2.205 0.028031 *
weight       -6.218e-03  5.714e-04 -10.883  < 2e-16 ***
`model year`  7.475e-01  5.079e-02  14.717  < 2e-16 ***
origin        1.428e+00  2.780e-01   5.138 4.43e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.326 on 385 degrees of freedom
Multiple R-squared:  0.8212,    Adjusted R-squared:  0.8184
F-statistic: 294.6 on 6 and 385 DF,  p-value: < 2.2e-16
```

Scheme 3. Summary of multiple linear regression model on the dataset after removing the “acceleration” attribute.

From this scheme, first of all the R^2 value is 0.8212 which decreased a little bit while and the adjusted R^2 value is 0.8184 which increased a little bit and both of them are still quite acceptable to us since a too high R^2 value might sometimes mean overfitting of the model. And for coefficients part, we did notice a slight decrease in “cylinders” attribute’s p-value, a slight decrease in “origin”’s p-value and a slight increase in “displacement”’s p-value. What was the most meaningful discovery for us here was the huge decrease in p-value of “horsepower” (almost a magnitude of 10 deduction from 0.21963 which was not so acceptable as well to 0.028031 which is now quite good in terms of p-value consideration). And this discovery demonstrated that the removing of the non-linear term “acceleration” from the original dataset did actually make the new model a better fit for the whole dataset. Furthermore, given the fact that p-value for “cylinders” was still quite high (0.117236), we would like to further remove this attribute and to see how the new model will be. The result this time is summarized in Scheme 4 below.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.669e+01  4.120e+00  -4.051 6.16e-05 ***
displacement  1.137e-02  5.536e-03   2.054   0.0406 *
horsepower   -2.192e-02  1.078e-02  -2.033   0.0428 *
weight       -6.324e-03  5.685e-04 -11.124  < 2e-16 ***
`model year`  7.484e-01  5.089e-02  14.707  < 2e-16 ***
origin        1.385e+00  2.772e-01   4.998 8.80e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.333 on 386 degrees of freedom
Multiple R-squared:   0.82,     Adjusted R-squared:  0.8177
F-statistic: 351.7 on 5 and 386 DF,  p-value: < 2.2e-16
```

Scheme 4. Summary of multiple linear regression model on the dataset after removing the “cylinders” attribute.

From this scheme, we can see that the R^2 value is 0.82 while the adjusted R^2 value is 0.8177. Both of them got decreased a little bit while they are still in the OK range. However for coefficients part, all the “displacement”, “horsepower” and “origin” attributes’ p-value got increased (from double to quadruple). Therefore we decided that it’s probably not a wise idea to try to remove the “cylinders” attribute in order to get a better model. And in all, we found that the most fit multiple linear regression model for this dataset would contain “cylinders”, “displacement”, “horsepower”, “weight”, “model year” and “origin” as predicator attributes and the constructed model tells us that a new “origin 1 (Japanese?)”car with three or four cylinders and low in displacement/ horsepower/weight will have a better MPG and vice versa.

1.3 Model Performances Comparosion

```
> sum(resid(MPG_model)^2)
[1] 4252.213
> sum(resid(MPG_model1)^2)
[1] 4259.571
> sum(resid(MPG_model2)^2)
[1] 4286.842
> anova(MPG_model,MPG_model1)
```

Scheme 5. Summary of RSSs of all previous constructed and mentioned LMs

Summarized in Scheme 5 are the computed RSSs for models summarized in Scheme 2, 3 and 4 which correspond to original model, model with “acceleration” removed and model with “acceleration”, “cylinders” removed respectively. We could see that for the original model, the RSS is 4252.213, for the model with “acceleration” removed, the RSS is 4259.571 and for the model with “acceleration”, “cylinders” removed, the RSS is 4286.842. These RSS values were also verified by applying the ANOVA model comparison between the original model and the modified models as shown in Scheme 6 and 7

```
Analysis of Variance Table

Model 1: mpg ~ cylinders + displacement + horsepower + weight + acceleration +
    `model year` + origin
Model 2: mpg ~ cylinders + displacement + horsepower + weight + `model year` +
    origin
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    384 4252.2
2    385 4259.6 -1   -7.3584 0.6645 0.4155
```

Scheme 6. ANOVA comparison between the original model and the model with “acceleration” removed

```
Analysis of Variance Table

Model 1: mpg ~ cylinders + displacement + horsepower + weight + acceleration +
    `model year` + origin
Model 2: mpg ~ displacement + horsepower + weight + `model year` + origin
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    384 4252.2
2    386 4286.8 -2    -34.63 1.5636 0.2107
```

Scheme 7. ANOVA comparison between the original model and the model with “acceleration” and “cylinders” removed

And from these schemes we can tell the RSS for the original model is indeed 4252.2, the RSS for the model with “acceleration” removed is indeed 4259.6 and the RSS for the model with “acceleration”, “cylinders” removed is indeed 4286.8.

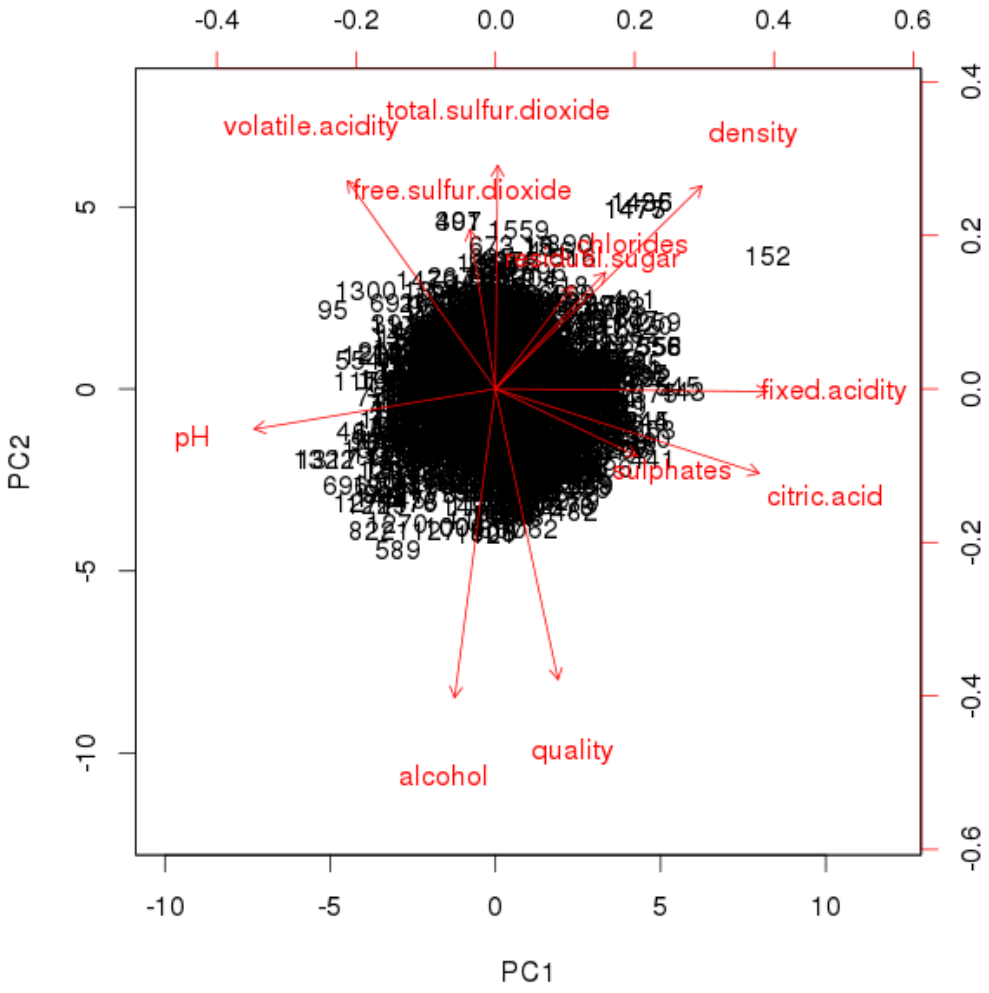
Part2 Visualization of Multi-dimension Dataset through PAC

In this part we will use PCA to analysis Wine dataset (red and white). The object is to find potential cluster from projection on PC1 and PC2, also the relation between those 12 attributes. Because the dataset is too large that makes the plot on PC1 and PC2 too dense to see clear cluster, we will replace index number with points and try to find out whether a clear cluster exist.

2.1 Red Wine Dataset

The red wine dataset consists of 1599 instances with 12 variables.

2.1.1 PCA Plot on Red Wine Dataset



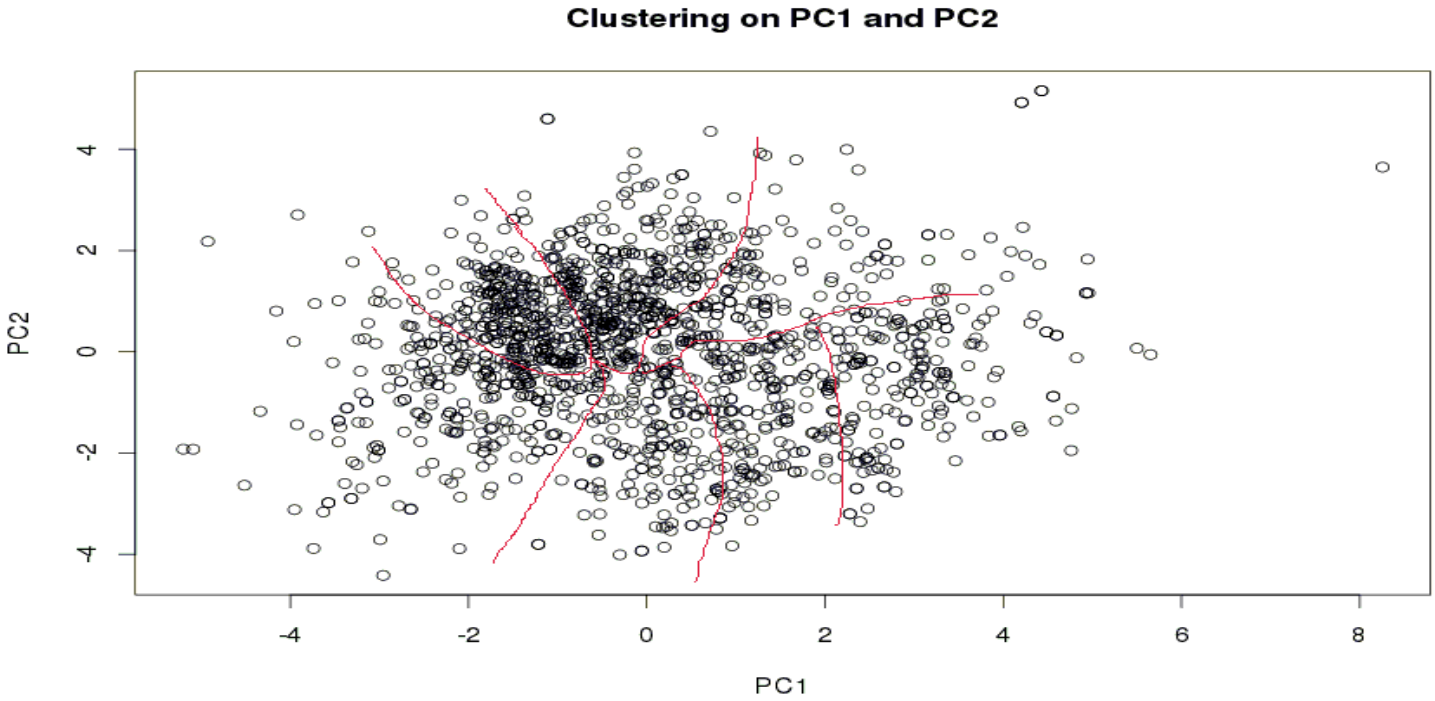
This is the PCA plot of red wine dataset. The X-axis and Y-axis is the score value on PC1 and PC2. Black points are the indexes of data that located on the projection on plane consists of PC1 and PC2. The red arrows represent the vectors that indicate the contribution of reach attributes to PC1 and PC2. From this plot we can easily find following results:

1) Quality: from the arrows we can see that quality is negatively correlated with volatile acidity, free sulfur dioxide and total sulfur dioxide. This means volatile acidity, sulfur dioxide volume will make quality worse. Likely red wine quality is positively correlated with alcohol volume. Also, because the vector of quality is perpendicular with PH and fixed acidity, it indicates that quality is independent with PH and fixed acidity. This means the quality is rarely influenced by them.

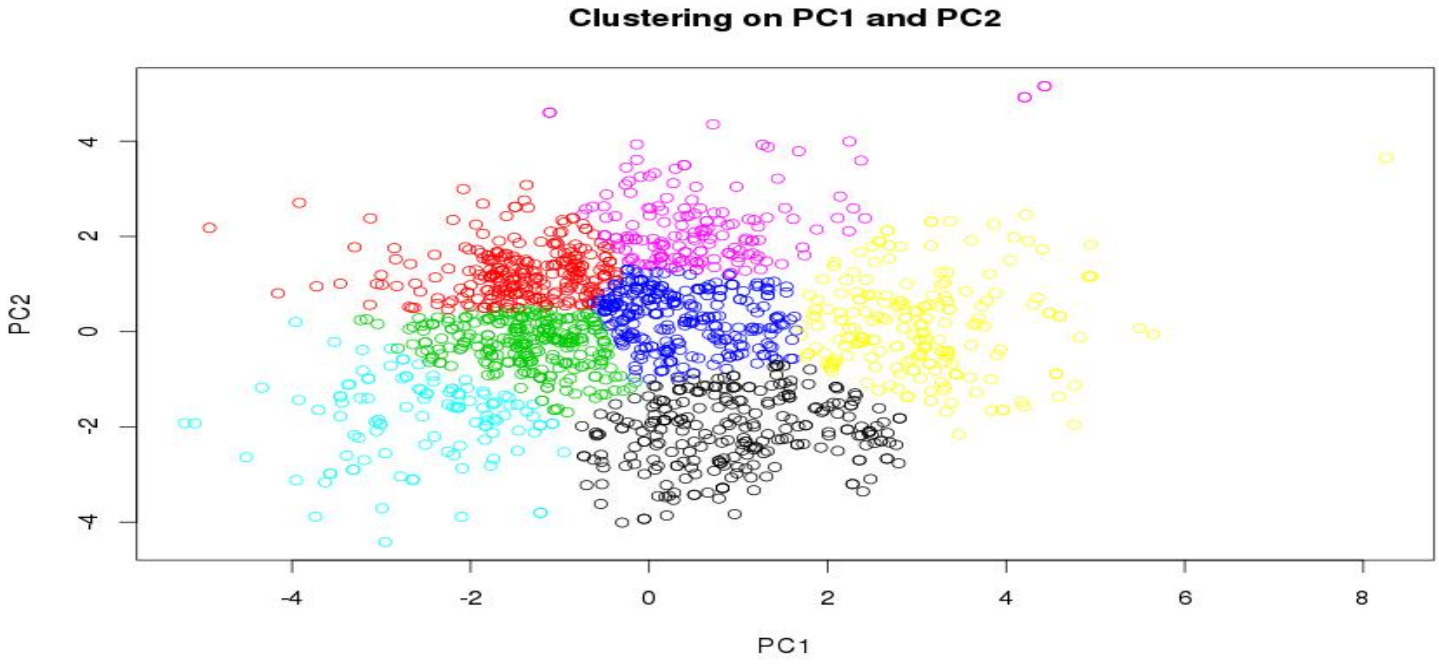
2) PH: PH value is independent with alcohol, volatile acidity and volume of sulfur dioxide (total and free). Further, it's negatively correlated with fixed acidity, citric acid and sulphates. This matches the rule of PH value – higher density of H^+ the lower the PH value.
3) Density: Density is positively related with chlorides, residual sugar. This may indicate that red wine with higher sugar volume may make the wine denser and denser wine may have higher chlorides. Also the density is barely related with volatile acidity and quality.

2.1.2 Project on PC1 and PC2 of Red Wine Dataset

The plot above project data with indexes which makes the plot too dense to find the potential cluster. So, we will try to plot it in dots instead of indexes.



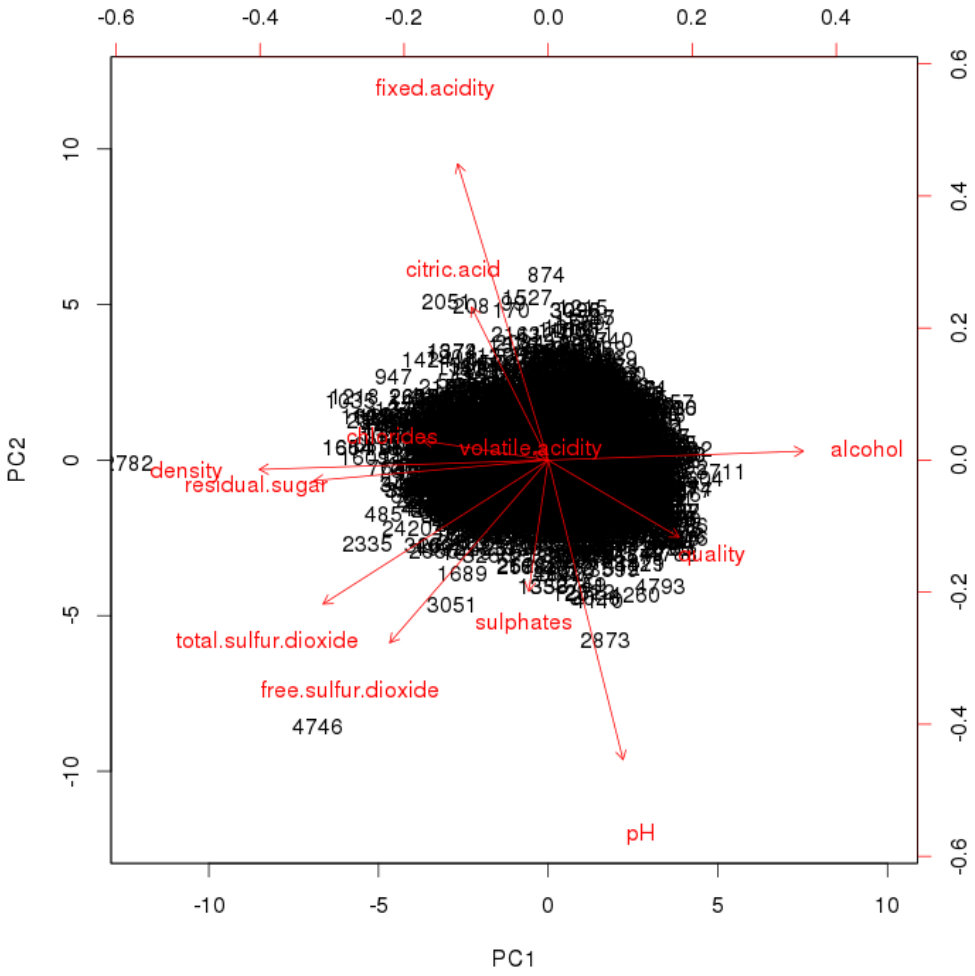
Unfortunately we cannot find very clear clustering in this plot. But we can still observe some patterns that can cluster the result. The red line is the line we drew for the potential clusters. According to the description of dataset, there are 7 quality classes. So we use k-means with k=7 to visualize the potential clusters.



2.2 White Wine Dataset

The white wine dataset consists of 4898 instances with 12 variables.

2.1.1 PCA Plot on White Wine Dataset



The component of plot on white wine dataset is the same as that on wine dataset we mentioned in part 2.1.1. What we found are as follow:

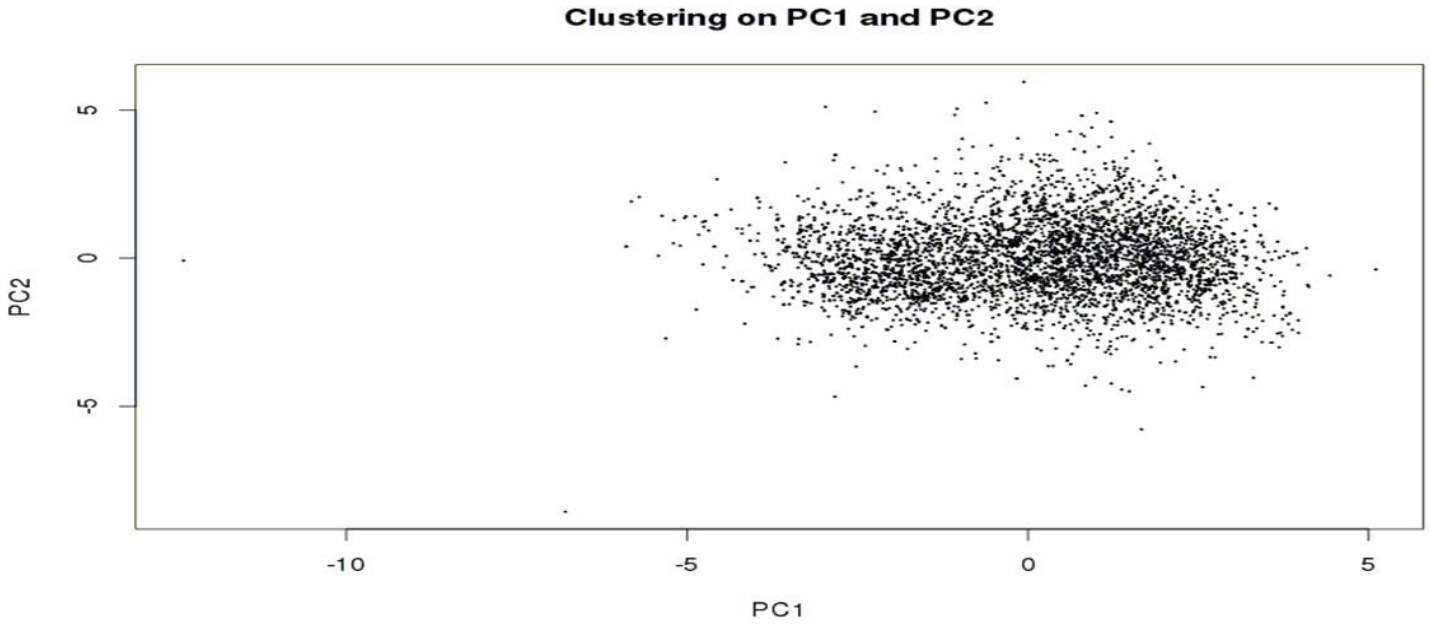
1) Quality: The vector of quality is perpendicular with free sulfur dioxide and total sulfur dioxide. This means the quality of red wine is not influenced by sulfur dioxide volume which is different from that of red wine. We can also find out that quality of red wine is negatively correlated with volatile acidity. In this perspective, the red and white wine are the same, their qualities are all negatively influenced by volatile acidity.

2) Density: Density of red wine is positively correlated with residual sugar and chlorides. But alcohol volume is negatively influenced with density. This may indicate that red wine with higher sugar could be denser and has higher chlorides level. But red wine with high degree of alcohol usually has lower density and sugar.

3) PH: PH is negatively correlated with fixed acidity and citric acid. This is similar with that of red wine. Also, we can find PH value is positively correlated with quality and sulphates. This shows that 1. Good red wine usually has higher PH value 2. High sulphates rate can increase PH value.

2.2.2 Project on PC1 and PC2 of White Wine Dataset

The approach is the same as the way we analysis red wine dataset.



Through the plot we can hardly find cluster because the plot dots are too dense to identify the clusters manually. Thus, we use k-means to visualize the potential clusters in the white wine dataset using k=7.