Stat 7350 Final Project Report

Jingyu Wang 7701969

April 19, 2019 9:00am

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0       v purrr   0.2.5
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.3.1
## v readr   1.3.1       v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at C:/Users/summe/Downloads/Stat-7350-class
```

```
here()
```

```
## [1] "C:/Users/summe/Downloads/Stat-7350-class"
```

Section I: Introduct the dataset and the questions that you will answer with your report

Introduction

Wine is one of the most common types of alcohols purchased to all countries of world. It is also a major export of Portugal which is a top ten wine exporting country in the world. In 2015, they exported 738 billion Euros worth of wine[1]. Different types of wine products have varying properties which can affect the quality and flavour of wine. Also, producers need certificate of their products to sell their wine legally. Physiochemical tests is commonly used method for wine's certification, which includes measuring pH, density, alcohol levels etc., as well as human tests which relies on human experts to taste and evaluate the quality of the wine. Measurement data could help to figure out variables attected quality of wine in order to improve wine production. Usually, higher quality of wine has higher price in the marketing.

The motivation is that finding the variables affects the quality, then we can improve the qulaity of wine and set higher price.

The dataset in our study was obtained from the University of California Irvine, Machine Learing Repository. Dr. Paulo Cortez et al. who are from the University of Minho in Portugal studied and had paper such that Modeling wine preferences by data mining from physicochemical properties based on this data in 2009. There

are two types of vinho verde wine, red wine and white wine, included in this dataset and wine vinho verde is unique product from the northwest region of Portugal. Data was collected by the official certification entity (CVRVV), which is an organization looking to improve the quality of vinho verde.

In the dataset, there are 1599 red wines samples and 4898 white wine samples. For each wine, it includes 11 relative variables and one respone variable(quality) and shows as following:

- Fixed acidity(g(tartaric acid)/dm3): a numeric vector
- Volatile acidity(g(acetic acid)/dm3): a numeric vector
- Citric acid(g/dm3): a numeric vector
- Residual sugar(g/dm3): a numeric vector
- Chlorides(g(sodium chloride)/dm3): a numeric vector
- Free sulfur dioxide(mg/dm3): a numeric vector
- Total sulfur dioxide(mg/dm3): a numeric vector
- Density(g/cm3): a numeric vector
- pH: a numeric vector
- Sulphates(g(potassium sulphate)/dm3): a numeric vector
- Alcohol(vol%): a numeric vector
- Quality: a factor with levels: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

The question we are trying to solve is which kinds of variables really affects the quality wine as the important variables and how those variables affects the quality of wine together.

Section 2: Analysis Methods

In our study, the first thing we considered is how samples distributed in different levels of quality. The histogram shows the results with counting of both wine in different levels of quality. Secondly, the line chart was given to explaine how mean of each variable changes for different quality for both wines. Scatter polt and correlation show that there exist relationship between few variables for both wine.

Finally, principle component regression model was used to find which variables can be linearly combined and affect the reponse variable(quality) together. Few steps need to be completed in PCA. Firstly, VIF checking can ensure whether multicolinearity exists in our dataset. PCA also can be used to decorrelate the dataset, then principle component regression model can be fitted by all standardized PCs. At the end of this method, important PCs are kept and show the linear combination in each PC. Here, avova also used for checking whether there is significant difference between two models.

Section 3: Results

First thing after get the dataset is to see how many samples in each level of quality for both wine. The histogram displayed that in the Figure 1.

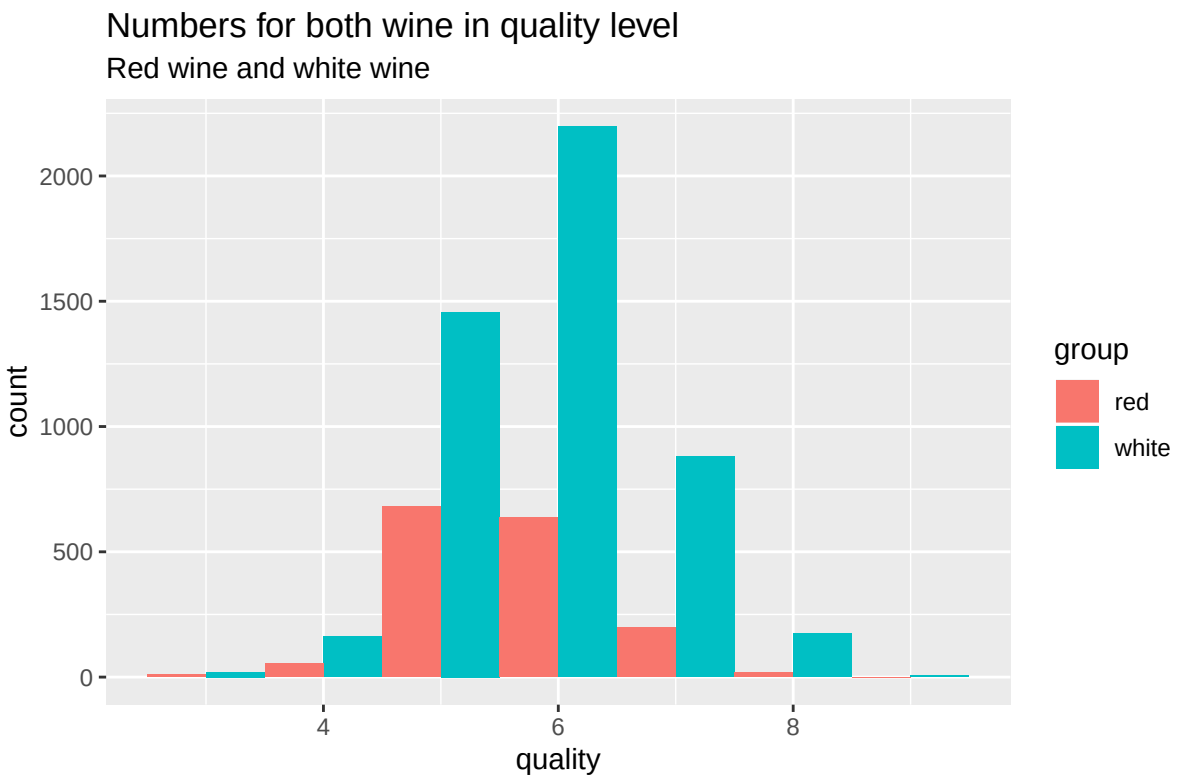

Figure 1: Numbers for both wine in different quality level

From the histogram, red wine and white wine samples distributed in different quality levels. Red wine samples are distributed in the the level from 3 to 8. Most of samples shows in the quality 5,6 and 7. However, there are few samples in the level 9 for white wine. It looks like a normal distribution for both wine, the midiem quality have more samples and few samples fall into very low or very high levels of quality.

Next, the line chart displays the changing of mean for variables in different level of quality for red and white wine reperately as Figure 2 and Figure 3.

From above both line chart, it is obviously to see that some varlables value is increasing with increasing of quality, such as citric acid and sulphates for red wine and pH for white wine. Some of them are decreasing when the quality is getting larger, such as volatile acidity in red wine and chlorides in white wine. However, this graph cannot show whether there are relationship between variables.

Thus, we also make scatter plot to see how the data looks like between variables for both wine and check whether there is relationship between them(The graphs shows in the supporting file, because it gives same information with correlation matrix). The correlation matrix for both wine gives the relationship between different variables as Figure 4 and Figure 5.

In red wine correlation figures, there is strong positive relationship between fixed acidity and density, as well as between fixed acidity and citric acid. However, pH have strong negative relationship with fixed acidity. In addition, the value which is between 0.6 and 0.79 express the strong relationship between compared variables in statistics[2]. Also, in white wine, residual sugar and density have very strong positive relationship, and density and alcohol have very strong negative relationship. However, there is no relationship between free

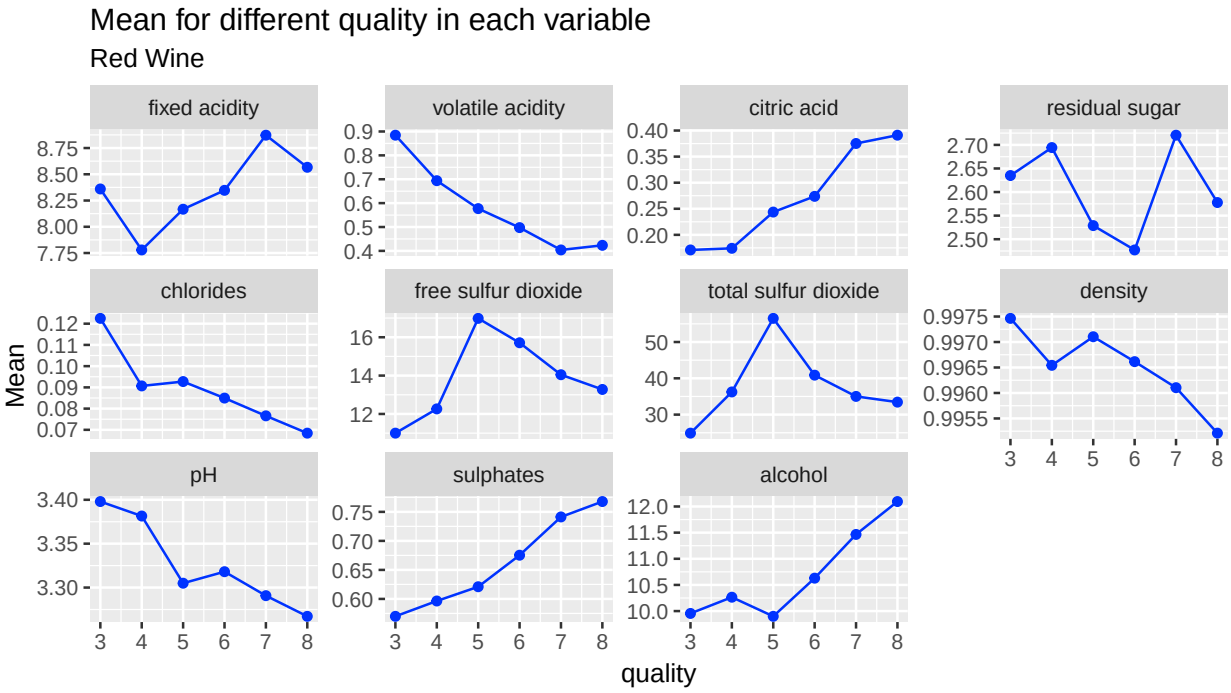

Figure 2: Mean for different quality in each variable(red wine)

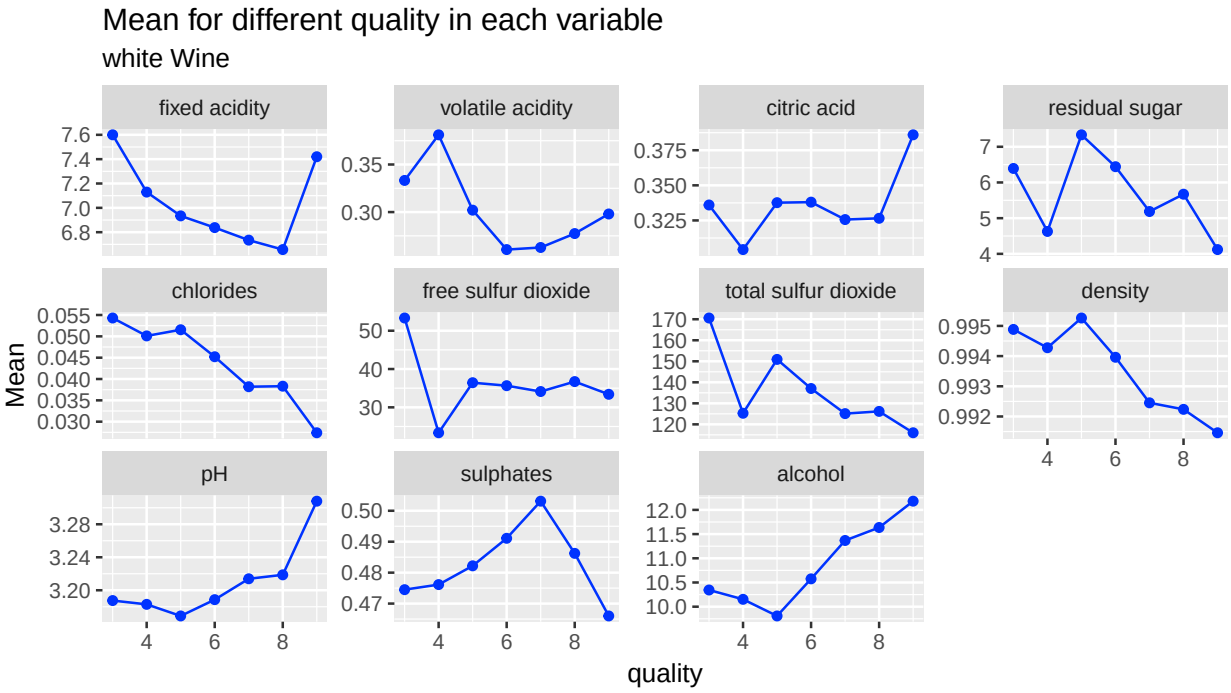

Figure 3: Mean for different quality in each variable(white wine)

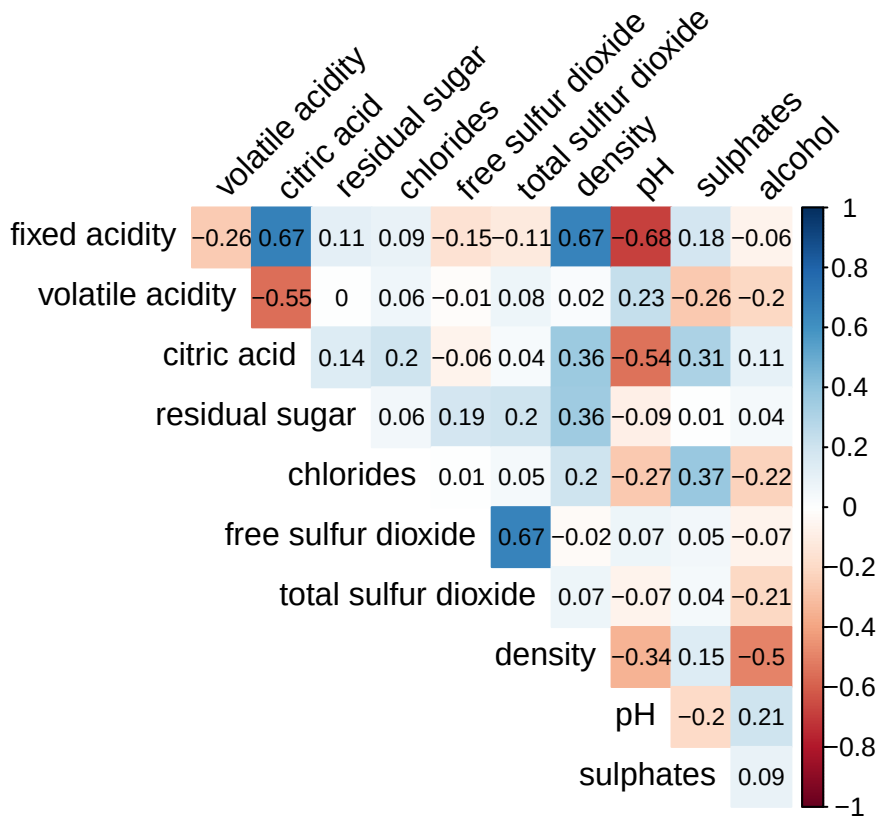

Figure 4: Correlation between different variables(red wine)

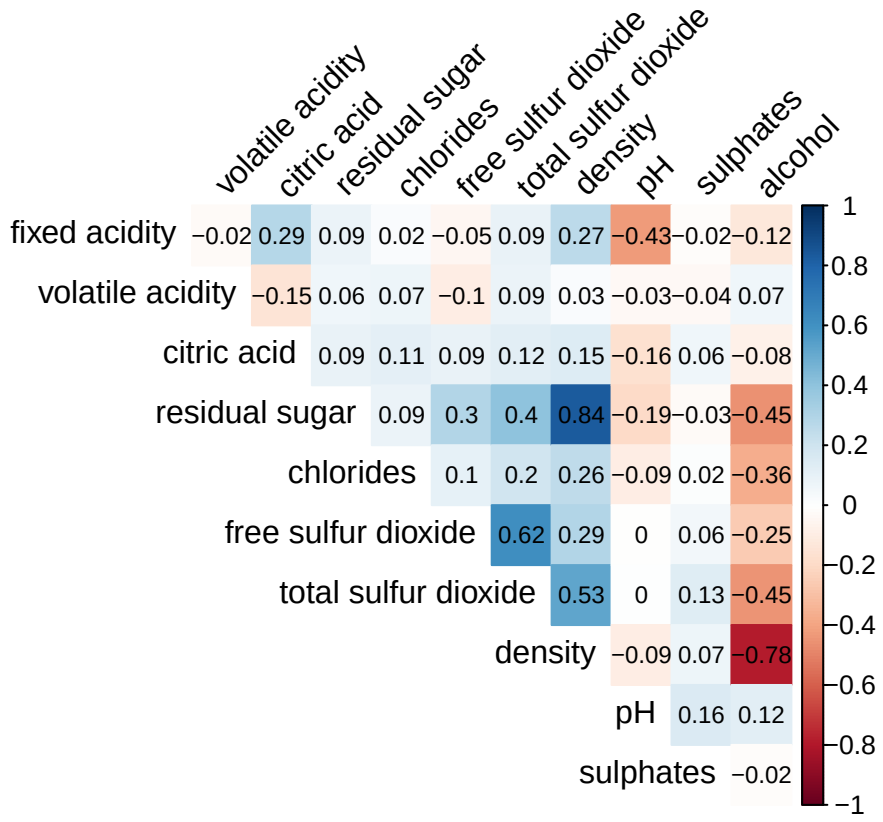

Figure 5: Correlation between different variables(white wine)

sulfur dioxide and pH, as well as between total sulfur dioxide and pH.

Now. we know that some of variables do have relationship between each other. Thus, the principle component analysis will be used to dicorrelate the data, which the one of the features of PCA. Firstly, variance inflation factors(VIF) value are calculated to ensure that multicolinearity exists in the both wine dataset, because scales of VIF are quite different, expecially white wine. For example, density have 28.23 but sulphates only have 1.14 for VIF. Covariance matrix is also checked after that, and shows that the scales of covariance are quite different. Thus, we ensure the multicolinearity for both wine and decide to use correlation matrix for the PCA at the sme time.

To do PCA, eigenvalues and eigenvectors are calculated by using correlation mateix. From the information of proportion of variation by each PC, 7 components and 8 components should be kept for red wine and white wine to express information of data, respectively. Also, standardized PCs was obtained by using correlation matrix times z, where $z = \frac{x-\bar{x}}{\sigma}$. Then, we have a new dataset with all standardized PCs and correlation checking shows there is no relationship between each other. Thus, this dataset was used to fit the principle components regession model.

For the red wine, there are 7 PCs(PC1-3,PC5 and PC7-9) are extremely significant(PC4 and PC10), two PCs are significant and two PCs are not significant. After reducing two not significant PCs, the model gave 0.3559 for adjusted R-squared which did not change so much with that of full model(0.3561). That means the nine components we kept was correct and did not lose much information of dataset. Also, we can check the importance in each variables from eigenvector matrix. For example, in PC1, it has fixed acidity(0.49) citric acid(0.46) and pH(-0.44) which contrbutes for the first PC and fixed acidity and citric acid have positive effects and pH have oppsite effects. Free sulfur dioxide and total sulfur dioxide have improtant affect for second PC and alcohol ahve negative effects. The following table shows how each variable contributed in each PC we kept.

```
# All variables contribute for PCs in red wine
read_csv(here("Assignemnt 4_Final project","Data","redwPC.csv"),
         col_types = cols())
```

```
## # A tibble: 11 x 10
##    X1                 PC1    PC2   PC3   PC4   PC5    PC7   PC8   PC9  PC10
##    <chr>            <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
##  1 fixed acidity     0.49 -0.11  -0.12  0.23  0.08  0.35   0.18 -0.19 -0.25
##  2 volatile acidity -0.24  0.27  -0.45 -0.08 -0.22  0.53   0.08  0.13  0.36
##  3 citric acid       0.46 -0.15   0.24  0.08  0.06 -0.1    0.38  0.38  0.62
##  4 residual sugar    0.15  0.27   0.1   0.37 -0.73 -0.290 -0.3  -0.01  0.09
##  5 chlorides         0.21  0.15  -0.09 -0.67 -0.25 -0.37   0.36 -0.11 -0.22
##  6 free sulfur dio~ -0.04  0.51   0.43  0.04  0.16  0.12   0.2  -0.64  0.25
##  7 total sulfur di~  0.02  0.570  0.32  0.04  0.22  0.09  -0.02  0.59 -0.37
##  8 density           0.4   0.23  -0.34  0.17 -0.16  0.17   0.24 -0.02 -0.24
##  9 pH               -0.44  0.01   0.06  0    -0.27  0.02   0.56  0.17 -0.01
## 10 sulphates         0.24 -0.04   0.28 -0.55 -0.23  0.45  -0.37  0.06  0.11
## 11 alcohol          -0.11 -0.39   0.47  0.12 -0.35  0.33   0.22 -0.04 -0.3
```

For white wine, there are 9 PCs(PC1-5 and PC 8-11) are extremely significant and two PCs are not significant.

The summary of model gave 0.2801 for adjusted R-squared which did not change so much with that of full model(0.2803) after reducing that two PCs. Avova method was also use to compare to models and the p-value is 0.2116, which means there is no significant difference between two models. Then, the importance in each variables from eigenvector matrix gave that residual sugar, total sulfur dioxide and density attibute the first PCs and alcohol attibute with opposite effects. Fixed acidity and pH attribute the eighth PC and residual sugar ahve opposite effects at the same time. Each variable contributed in each PC we kept display as below.

```
# All variables contribute for PCs in white wine
read_csv(here("Assignemnt 4_Final project","Data","whitewPC.csv"),
         col_types = cols())
```

```
## # A tibble: 12 x 10
##    X1          PC1     PC2    PC3    PC4 PC5      PC8    PC9    PC10   PC11
##    <chr>     <dbl>   <dbl>  <dbl>  <dbl> <chr>  <dbl>  <dbl>   <dbl>  <dbl>
##  1 fixed a~   0.16   0.59   -0.12  -0.02 0.25    0.59  -0.33   0.13   -0.17
##  2 volatil~   0     -0.05    0.59  -0.28 0.64    0.03   0.15   0.22   -0.02
##  3 citric ~   0.14   0.35   -0.5   -0.15 0.05   -0.15   0.2    0.04   -0.01
##  4 residua~   0.43  -0.01    0.21   0.27 0.01   -0.39  -0.41  -0.09   -0.49
##  5 chlorid~   0.21   0.01    0.1   -0.71 -0.32  -0.1   -0.39  -0.05   -0.02
##  6 free su~   0.3   -0.290  -0.28   0.31 0.19   -0.08  -0.14   0.570   0.03
##  7 total s~   0.41  -0.24   -0.13   0.06 0.3     0.25   0.15  -0.71   -0.04
##  8 density    0.51  -0.01    0.13   0.02 -0.09   0.07  -0.09   0.07    0.76
##  9 pH        -0.13  -0.580  -0.13  -0.1  -0.13   0.53  -0.26   0.11   -0.14
## 10 sulphat~   0.04  -0.22   -0.43  -0.44 0.39   -0.27   0.01   0.06   -0.04
## 11 alcohol   -0.44   0.04   -0.11   0.14 0.34   -0.2   -0.62  -0.27    0.36
## 12 <NA>      NA     NA      NA     NA    .      NA     NA     NA      NA
```

Limitation/Discussion

When we obtain an multivariate dataset, the first thing we want to try is fitting the linear model. I also tried that, the linear regression model gave the variables which played important role in the data. However, in our case, there is multicolinearity exist and correlation between few variables. Thus, the regression model may not be accurate. This is the reason why we need to use principle component regression model.

Also, note that the quality as respone variable is ordinal, therefore the ordinal regression model(clm in R) may need to use and compare the different outcomes from general lieane model(lm in R). There are very similar result as the only difference is that pH is found to be significant in the linear regression model for red wine. White wine gives the exact same variables to be significant.

Section 4: Conclusion& Reflection

In our study, some of 11 variables have relationshp and affect quality of wine together. There are 7 and 9 extremely significant components for red wine and white wine ,respectively. Also, for each componet, there are different variables contributed in different ways(positive or nigative). There, all of them will give the

results for which kinds ot variables affect quality of wine together and how to affect it.

For this assignment, I know there is no missing data in our dataset. Also, it is a multivariable datset with one response variable. Before doing this assignmnt, I will fit the linear model directly when Got this kind of data. However, for now, I know multicolinearity may exist in this kind of dataset. Thus, the linear model fitting may not accurate. Also mention this point for the future student who want to use this dataset, the relationship between variables should be check first. Even the linear model fitting give nice result, the assumptions(independent between variables) are not satified, which means the result can be not correct.

References

[1] State of the Vitiviniculture World Market. 2016. *Organisation internation.*

[2] Peason's Correlation. www.statstutor.ac.uk/resources/uploaded/pearsons.pdf