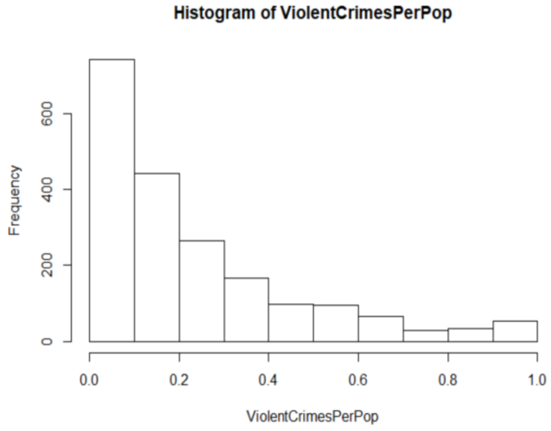
1.Exploratory data analysis for Communities and Crime dataset:

Since the dataset from the UCI database has no headers/names, I manually add headers according to the attributes names from the documentation. There are many missing data labeled as '?' in this dataset, so I change them to N/A in order to avoid some errors .

Since the communities and crime dataset have 128 attributes/columns, I need to do some exploratory data analysis to see what attributes will contribute to my model.

First, I make a historical diagram for the crime per capita to see the distribution of the crime(figure 1).

I thought household income is related to the crime rate, so I make a plot that shows the relationship between violent crimes per population and median house income. As what shows in figure2, communities with higher median income has lower violent crimes per population.



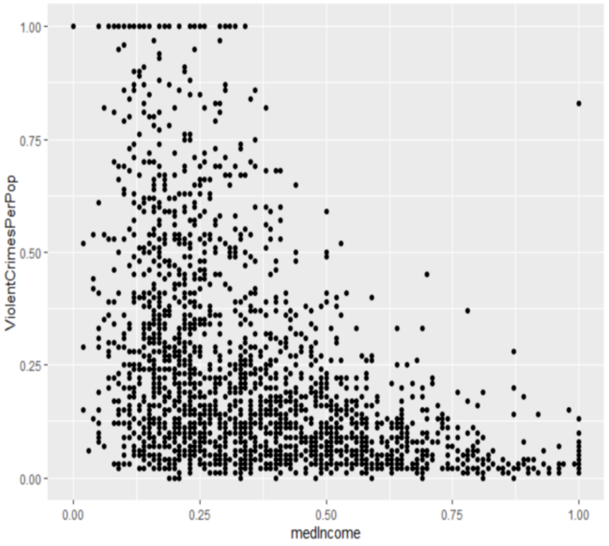


Figure 1

Figure2

I also make a point graph of different age groups against the crime per population as the figure 3 shows(red point: age 12-29,blue point: age 12-21, green point: age 16-24, yellow point: age 65+). We can see from the graph that the average ages of the community do influence the crime rate: more people with 12-29 in the community reflects a higher crime rate.

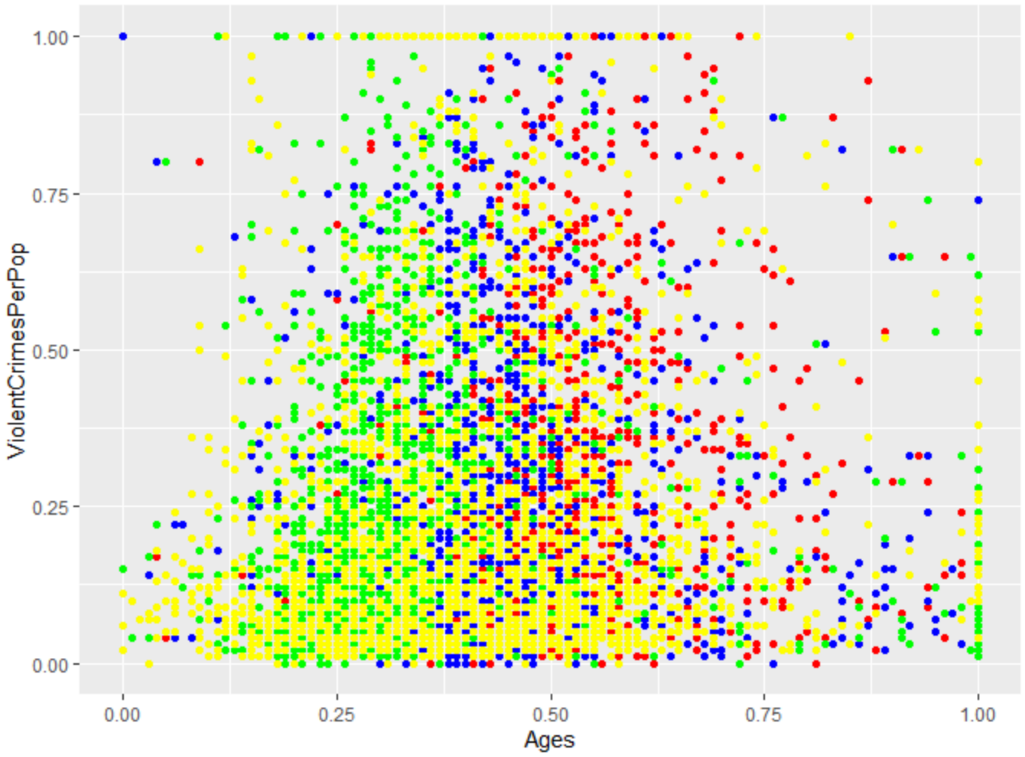


Figure 3

In addition, I want to perform a PCA analysis to turn my original variables into a smaller number of “Principle Components” as what shows in figure 4. Since there are many missing elements marked as question mark in the data, I remove those column that have a majority of missing data to do the PCA analysis.

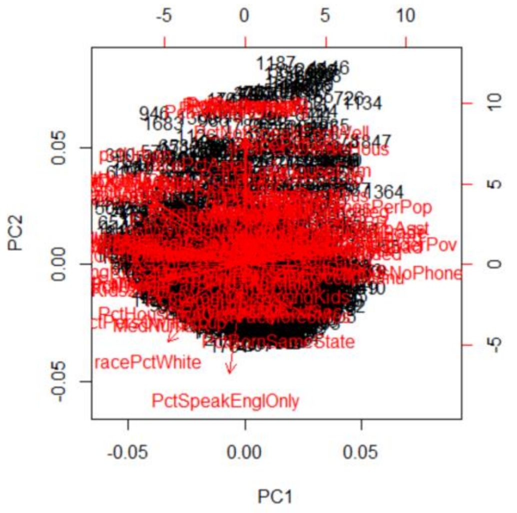


Figure 4

From the exploratory analysis, the violent crime per population is highly related to other attributes in the model. So I decided to make models to predict the violent crime per population based on some chosen variables in the dataset. We find this dataset will be good to use a regression model.

Exploratory data analysis Wine quality dataset:

The dataset I am using the red wine quality dataset. First, I make a historical diagram for quality attribute to see the distribution of the quality(figure 5). From the diagram, we can see that most wines have a quality of 5 or 6.

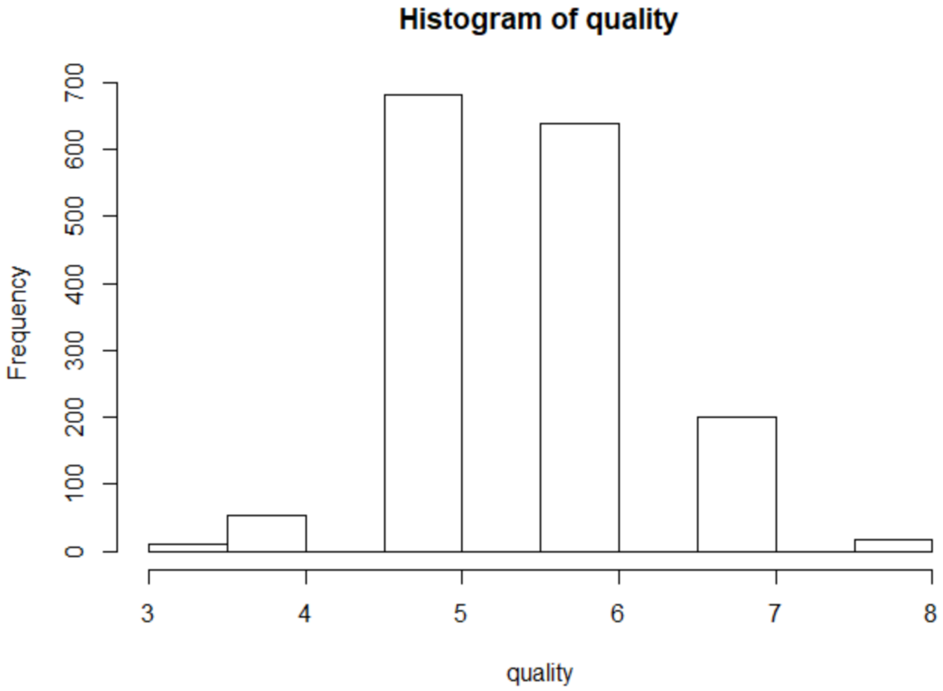
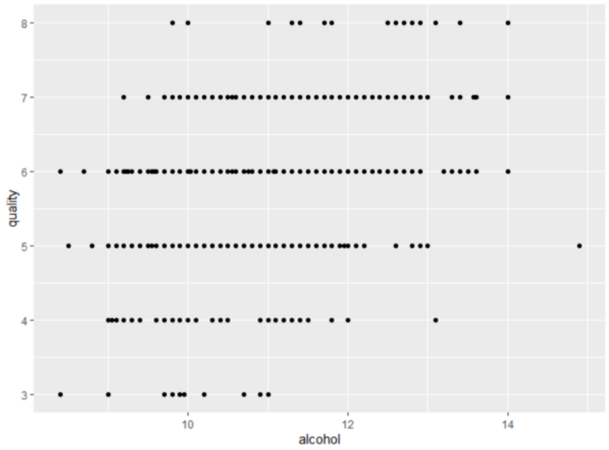
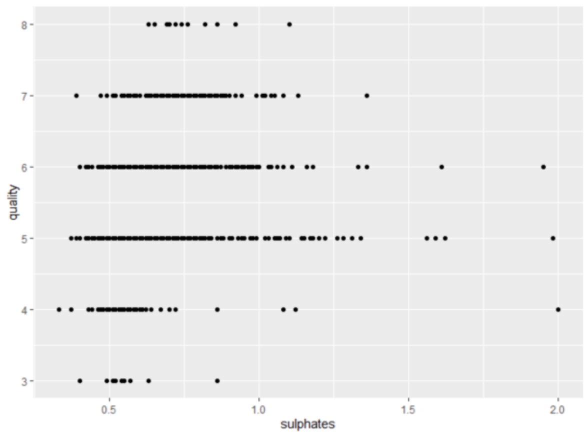
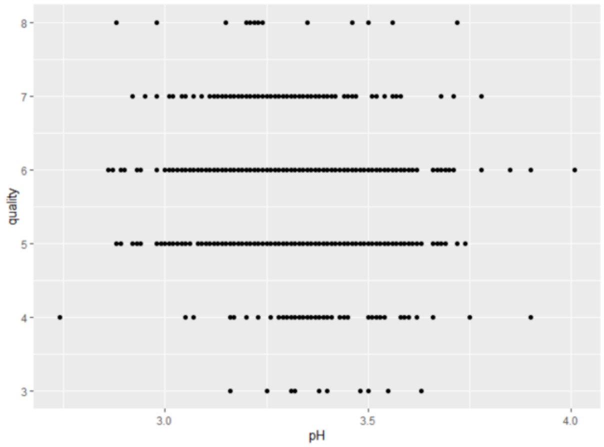


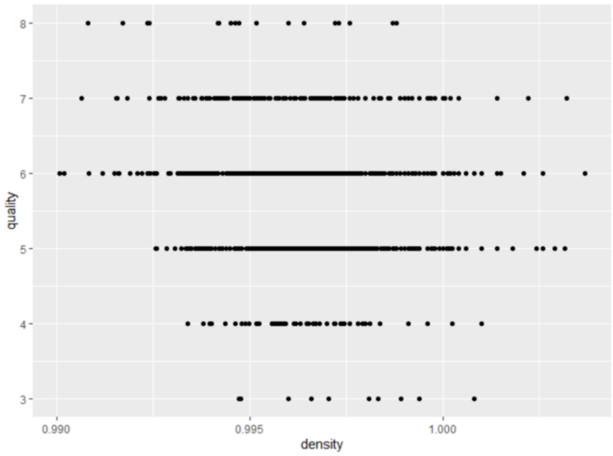
Figure 5

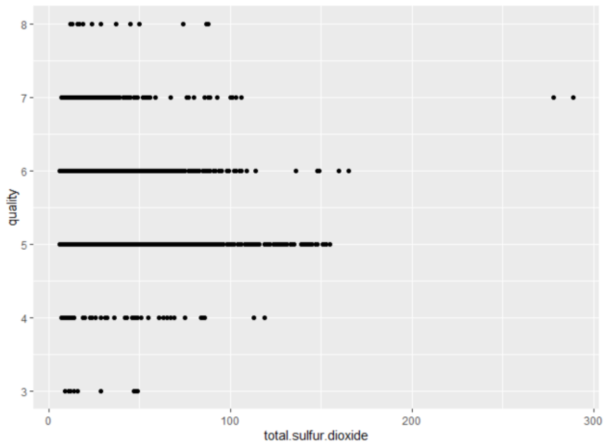
Next, I want to explore and visualize the relationship of the input variables and output variables. So I made 11 plots to visualize the relationship shown below. From these plots, we can see that for most attributes, there does not exists a clear pattern that the quality will increase or decrease as a specific attribute increases. Therefore we can conclude that the quality of the wine depends on most of these attributes.

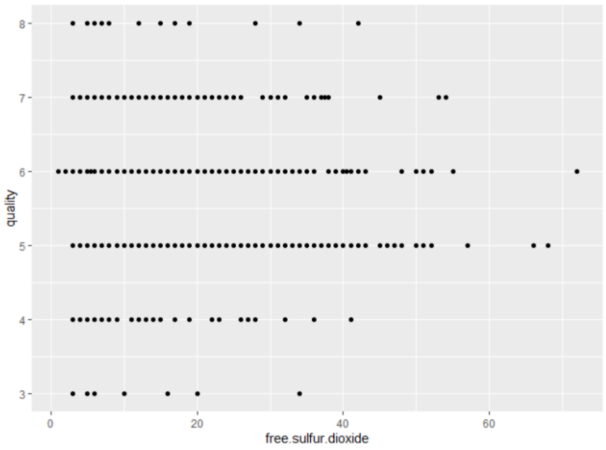

quality
8
7
6
5
4
3
10
12
14
alcohol

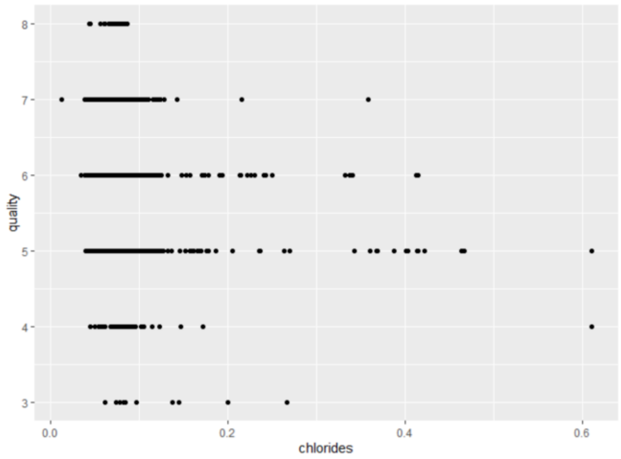

quality
8
7
6
5
4
3
0.5
1.0
1.5
2.0
sulphates

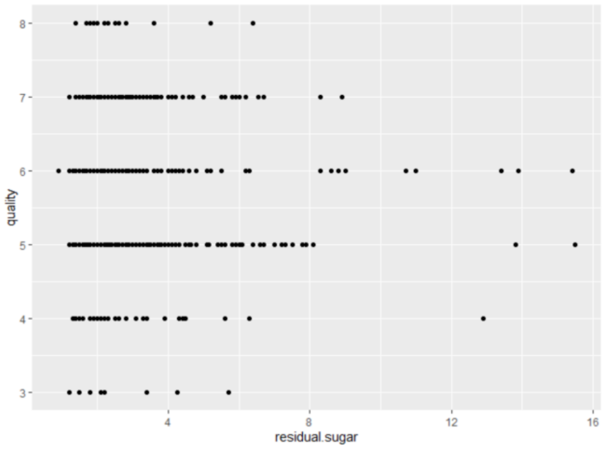

quality
8
7
6
5
4
3
3.0
3.5
4.0
pH

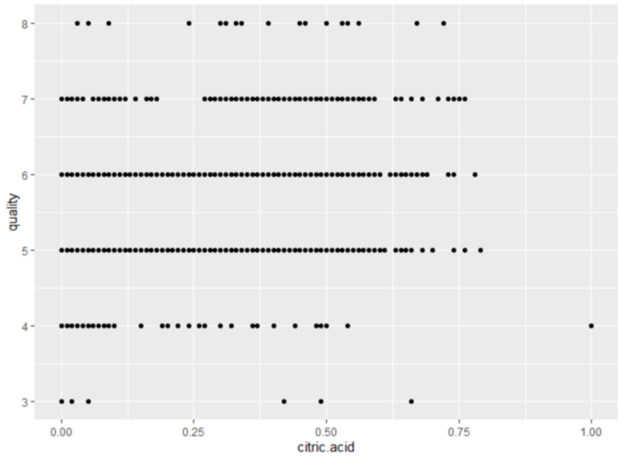

quality
8
7
6
5
4
3
0.990
0.995
1.000
density

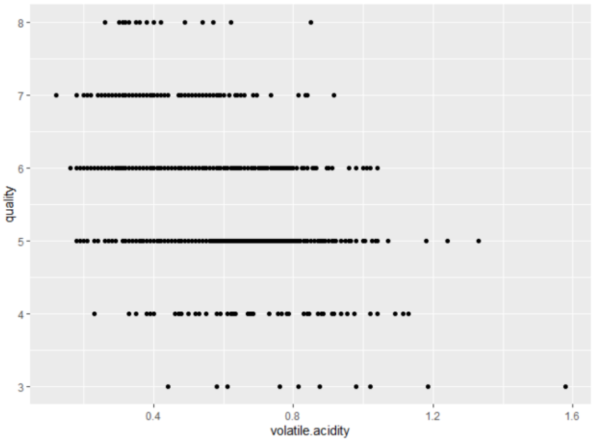

quality
8
7
6
5
4
3
0
100
200
300
total.sulfur.dioxide

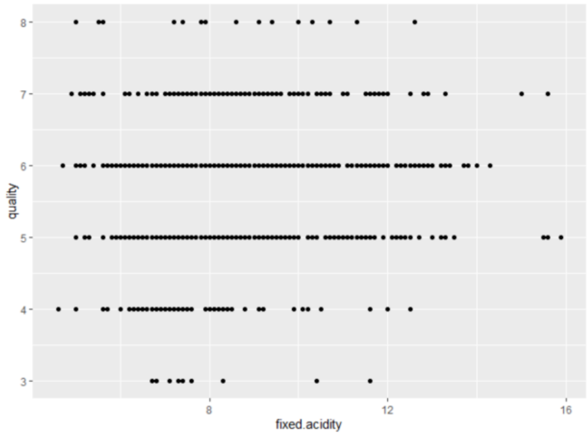

quality
8
7
6
5
4
3
0
20
40
60
free.sulfur.dioxide



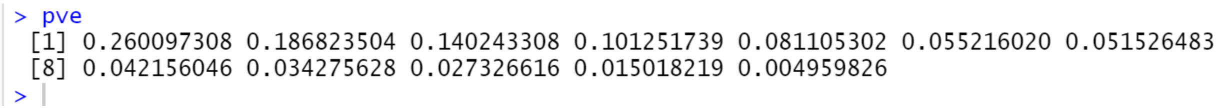








I also want to run a PCA analysis to turn my original variables into a smaller number of “Principle Components” . I calculate the variance of each principle component as shown in Figure 6, we can see the first principle component explains 26% of the variance. I also visualize the PCA (Figure 7)



```
> pve
 [1] 0.260097308 0.186823504 0.140243308 0.101251739 0.081105302 0.055216020 0.051526483
 [8] 0.042156046 0.034275628 0.027326616 0.015018219 0.004959826
> 
```

Figure 6

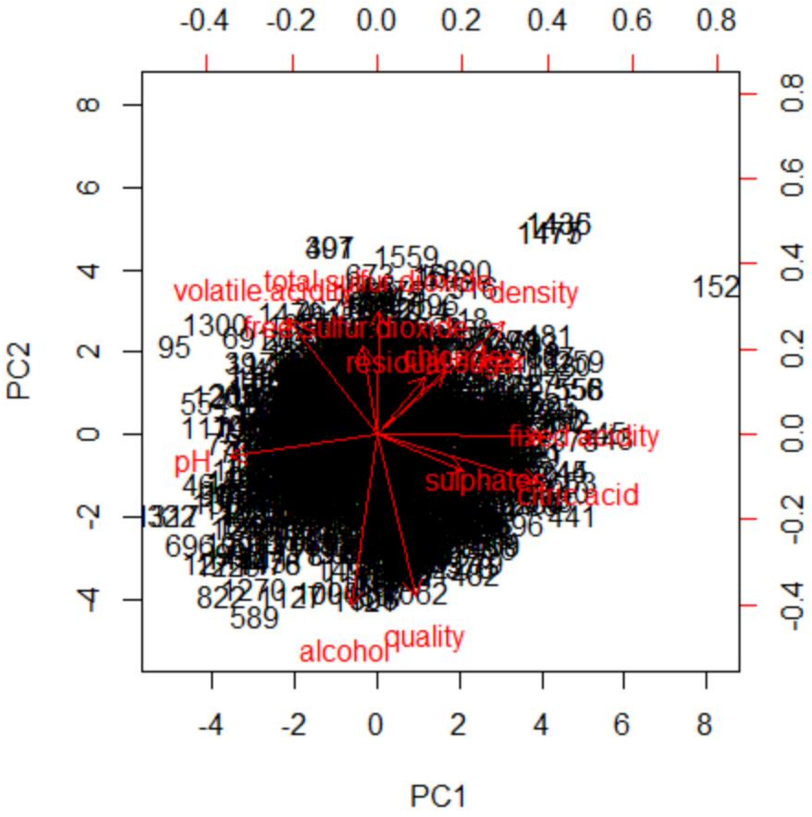


Figure 7

2.Analysis:

Crime and Communities Dataset:

Regression model:

I use this dataset is a multivariate dataset and regression model is very useful in the multivariate dataset. In addition, there are a lot of attributes in this dataset and I have selected the most important retributes, this will give my model higher accuracy.

From the exploratory analysis, I decided to choose violent crime per population as dependent variable, and population, householdsize, agePct12t21, agePct12t29, agePct16t24, agePct65up, medIncome, MedRent, PolicOperBudg as independent variable.

For the accuracy of the model, I remove those data frames which has NA in my chosen variables. The fitted line generated by model is shown below as Figure 2.1,2.2 2.3 2.4.

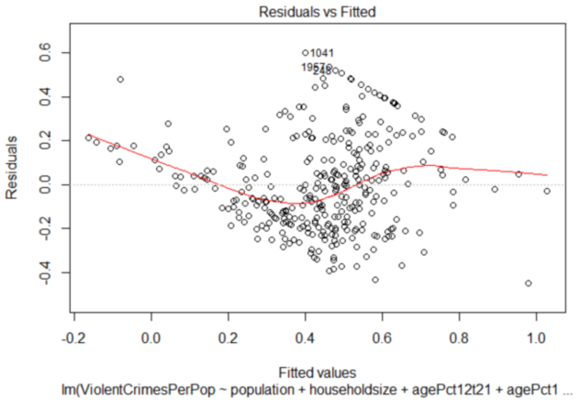


Figure 2.1

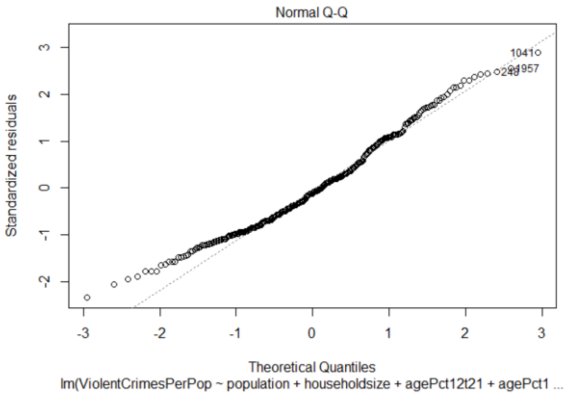


Figure 2.2

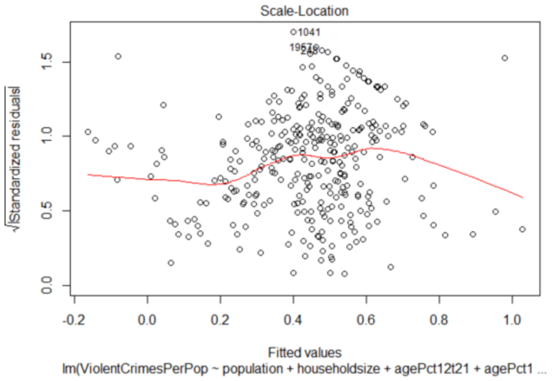


Figure 2.3

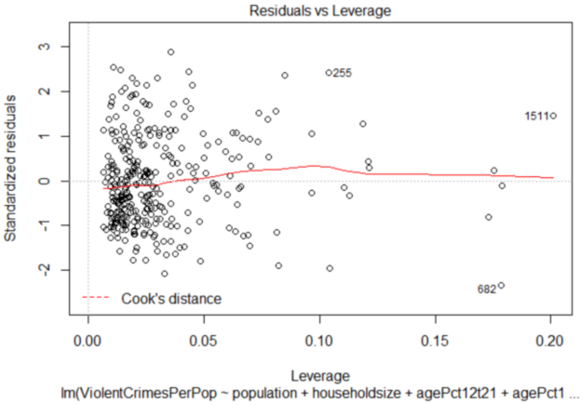
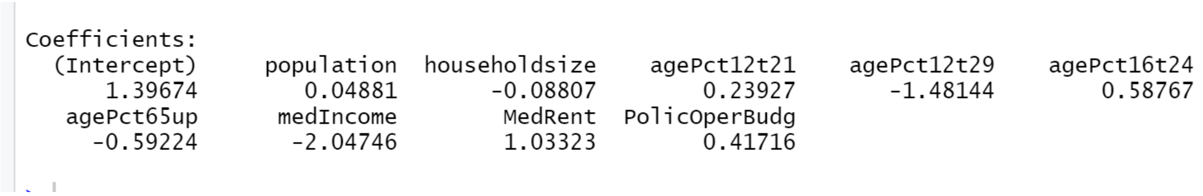


Figure 2.4

The coefficient of the model are listed as the following figure shows:



```
Coefficients:
  (Intercept)     population  householdsize    agePct12t21    agePct12t29    agePct16t24
      1.39674        0.04881       -0.08807        0.23927       -1.48144        0.58767
    agePct65up      medIncome        MedRent  PolicOperBudg
     -0.59224       -2.04746        1.03323        0.41716
```

The line of the fitted value against the true value is shown as figure 2.5:

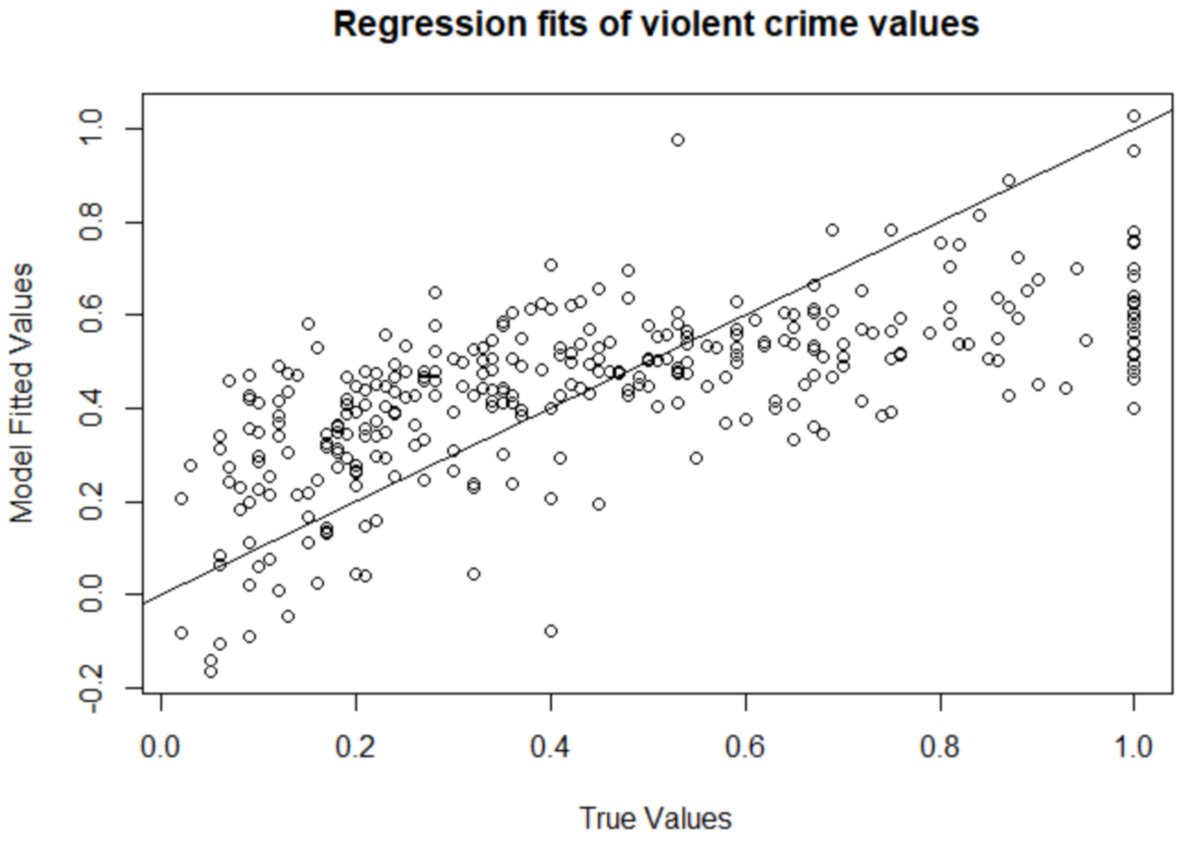


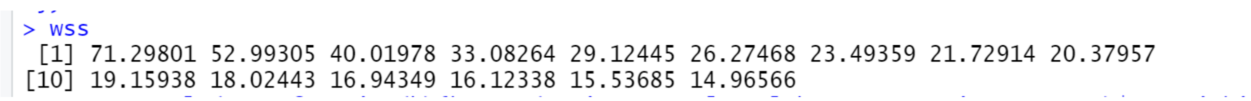
Figure 2.5

Kmeans:

I use this model because the dataset I used is unlabeled dataset, and K-means clustering is a type of unsupervised learning. I can use the K-means to find groups in data.

For the Kmeans, I still use the same independent variables and dependent variables as what I use in regression model. I also remove those data frames which has NA in their values since Kmeans function does not allow missing values in the dataset.

Figure 2.6 shows the first 12 values of total within-cluster squares and Figure 2.7 shows the visualization of the total within-cluster squares :



```
> wss
 [1] 71.29801 52.99305 40.01978 33.08264 29.12445 26.27468 23.49359 21.72914 20.37957
[10] 19.15938 18.02443 16.94349 16.12338 15.53685 14.96566
```

Figure 2.6

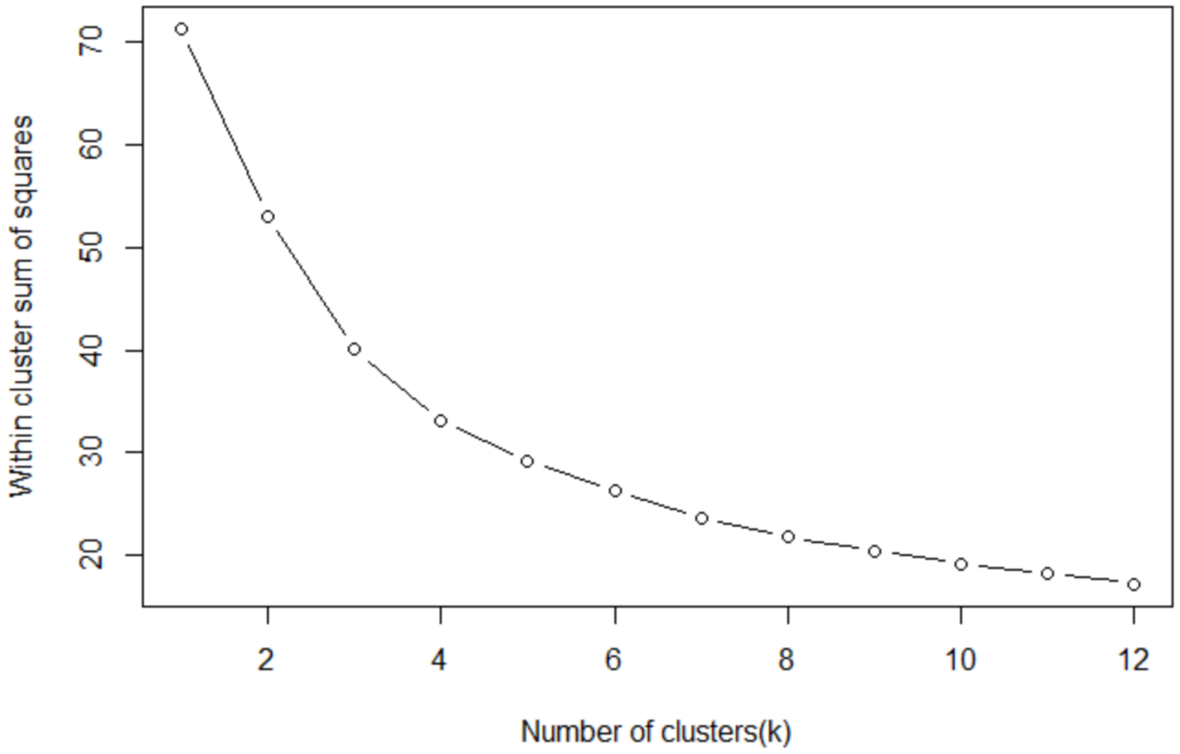


Figure 2.7

I also use the fviz_cluster function to plot the kmeans plot of 3 centers:

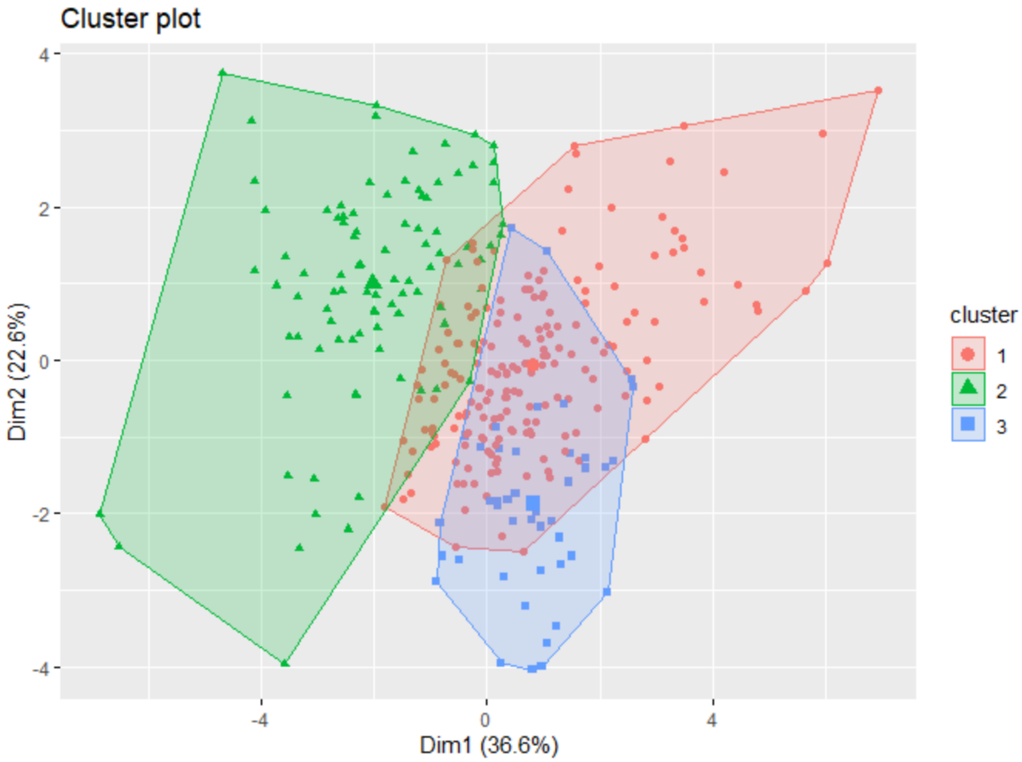


Figure 2.8

The plot of 4 centers(figure 2.9):

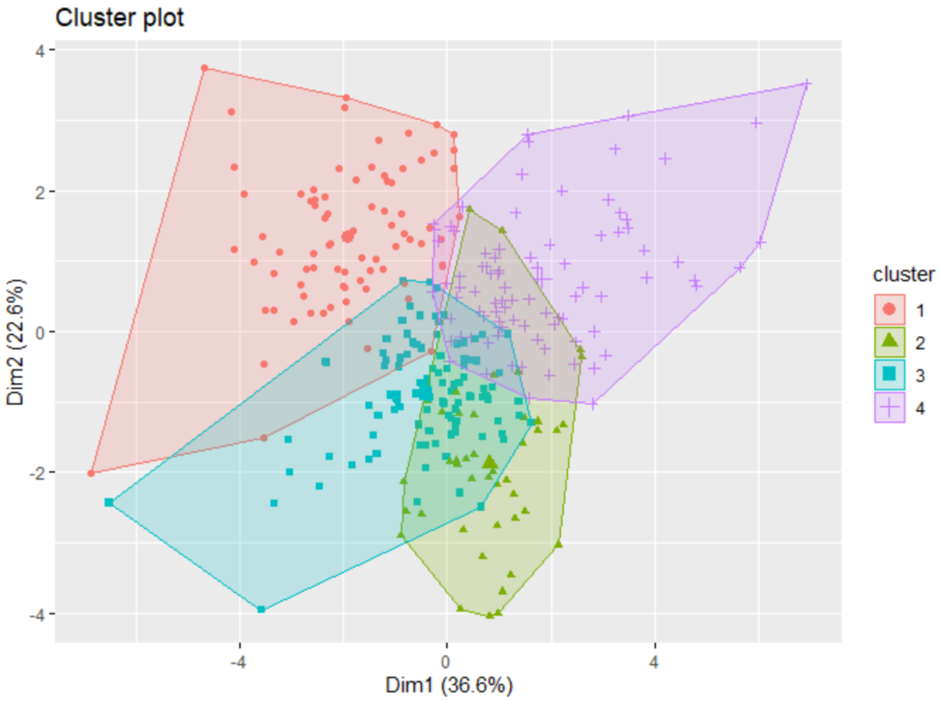
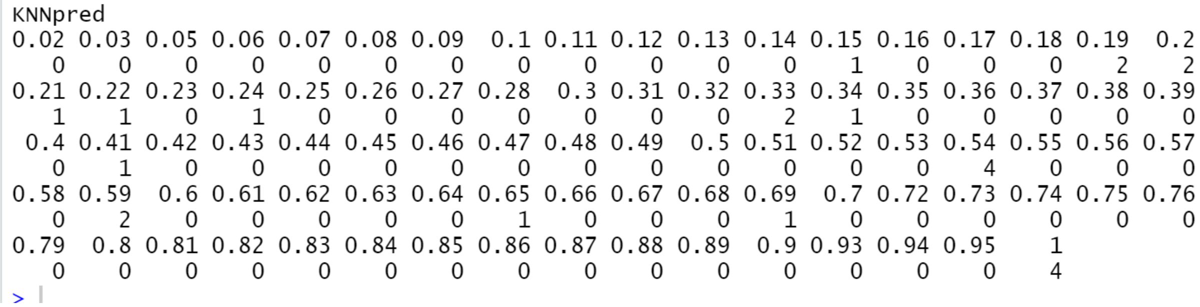


Figure 2.9

KNN:

For the K nearest neighbors, I still use the same independent variables and dependent variables as what I use in regression model. I also remove those data frames which has NA in their values since KNN function does not allow missing values in the training dataset. I split the model into the KNNtrain dataset and KNNtest dataset. Since there are approximately 300 rows in the dataset, I use 17 which is the square root of 300 as the number of K.

The results of the KNN is shown as Figure 2.6.



```
KNNpred
0.02 0.03 0.05 0.06 0.07 0.08 0.09  0.1 0.11 0.12 0.13 0.14 0.15 0.16 0.17 0.18 0.19  0.2 
   0    0    0    0    0    0    0    0    0    0    0    0    1    0    0    0    2    2 
0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28  0.3 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 
   1    1    0    1    0    0    0    0    0    0    0    2    1    0    0    0    0    0 
 0.4 0.41 0.42 0.43 0.44 0.45 0.46 0.47 0.48 0.49  0.5 0.51 0.52 0.53 0.54 0.55 0.56 0.57 
   0    1    0    0    0    0    0    0    0    0    0    0    0    0    4    0    0    0 
0.58 0.59  0.6 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69  0.7 0.72 0.73 0.74 0.75 0.76 
   0    2    0    0    0    0    0    1    0    0    0    1    0    0    0    0    0    0 
0.79  0.8 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89  0.9 0.93 0.94 0.95    1 
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    4 
> 
```

Figure 2.6

Red wine quality Dataset:

Regression model:

I use this dataset is a multivariate dataset and regression model is very useful in the multivariate dataset. We can use the regression model to help us find which input variable matters most and how these variables affect each other . For this regression model, I want to use all the input variables in the dataset as independent variables.

The fitted line generated by model is shown below as Figure 2.7,2.8 2.9 2.10.

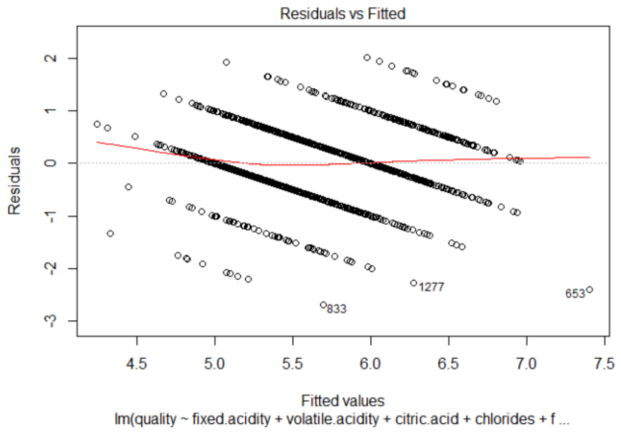


Figure 2.7

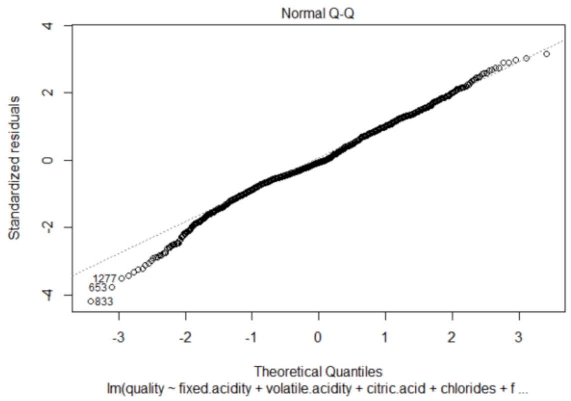


Figure 2.8

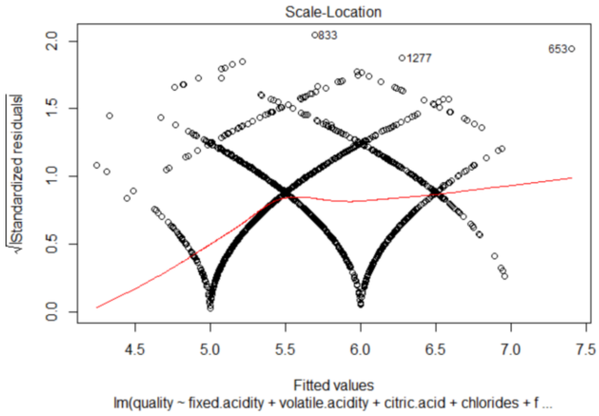


Figure 2.9

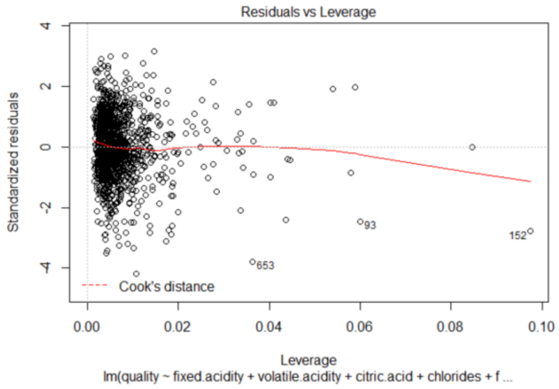
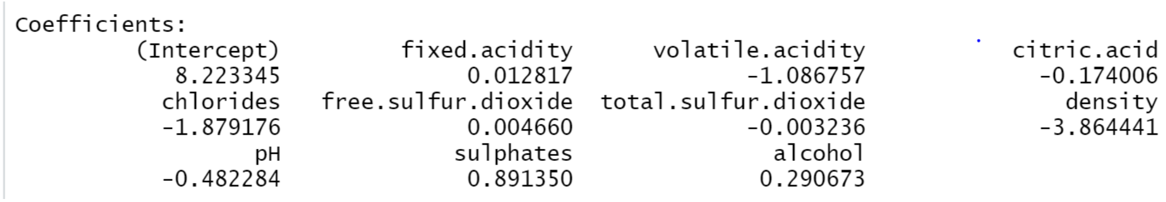


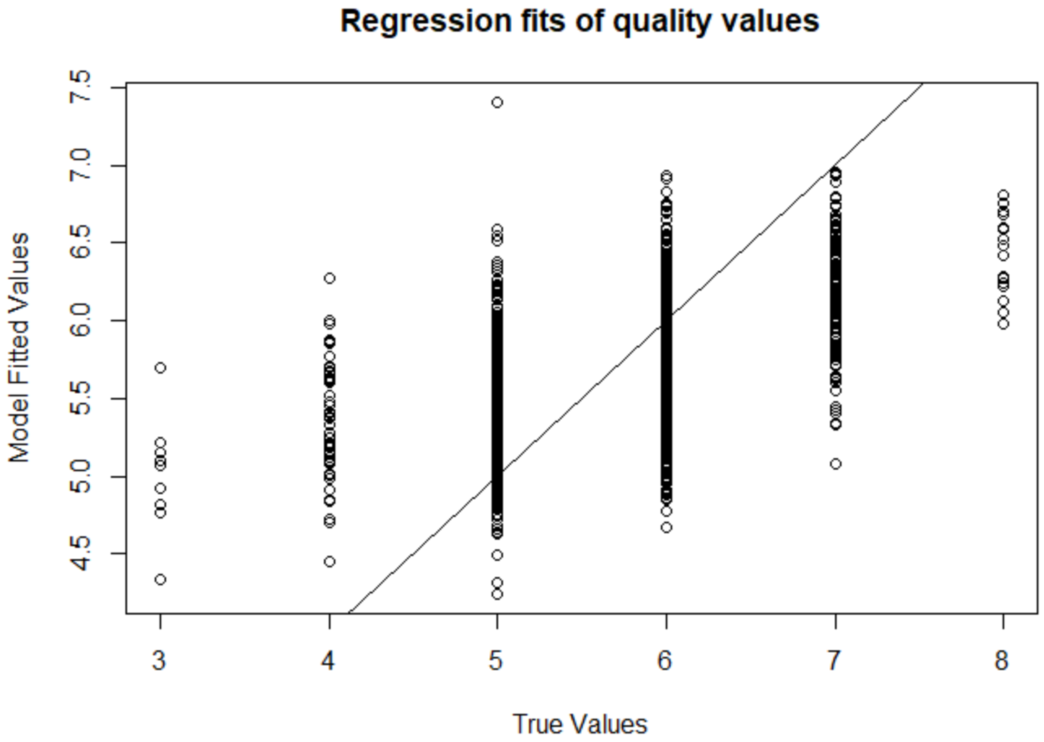
Figure 2.10

The coefficient of the model are listed as the following figure shows:



```
Coefficients:
         (Intercept)        fixed.acidity     volatile.acidity          citric.acid
            8.223345             0.012817            -1.086757            -0.174006
           chlorides  free.sulfur.dioxide total.sulfur.dioxide              density
           -1.879176             0.004660            -0.003236            -3.864441
                  pH            sulphates              alcohol
           -0.482284             0.891350             0.290673
```

The line of the fitted value against the true value is shown as figure 2.11



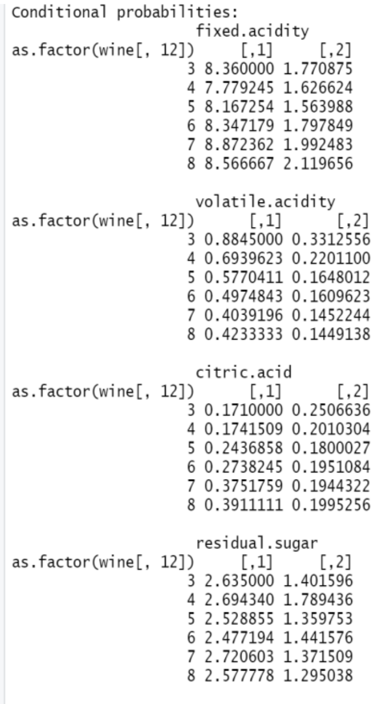
:

Figure 2.11

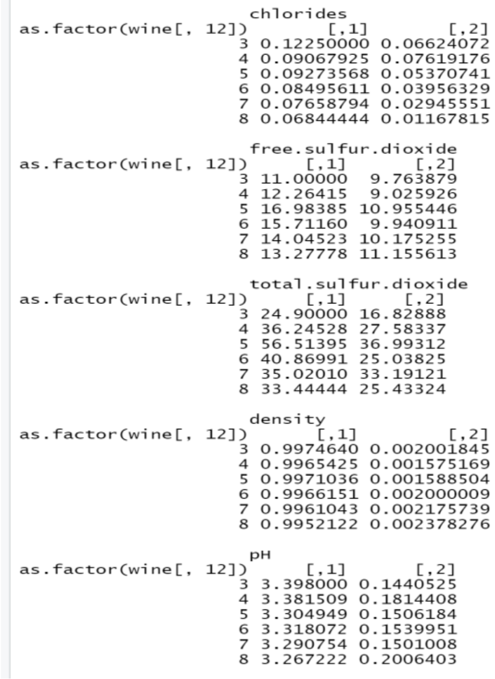
Naïve Bayes:

Naive Bayes is a classification algorithm that is suitable for multiclass classification. It is used to classify data frames by assigning them class labels. In our wine quality dataset, we can set our quality as class labels.

The conditional probability of the model is shown below:

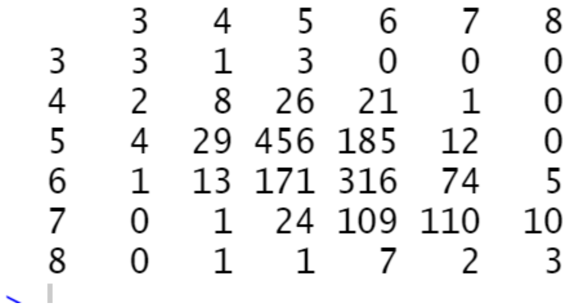


```
Conditional probabilities:
                       fixed.acidity
as.factor(wine[, 12])      [,1]     [,2]
                      3 8.360000 1.770875
                      4 7.779245 1.626624
                      5 8.167254 1.563988
                      6 8.347179 1.797849
                      7 8.872362 1.992483
                      8 8.566667 2.119656

                       volatile.acidity
as.factor(wine[, 12])      [,1]      [,2]
                      3 0.8845000 0.3312556
                      4 0.6939623 0.2201100
                      5 0.5770411 0.1648012
                      6 0.4974843 0.1609623
                      7 0.4039196 0.1452244
                      8 0.4233333 0.1449138

                       citric.acid
as.factor(wine[, 12])      [,1]      [,2]
                      3 0.1710000 0.2506636
                      4 0.1741509 0.2010304
                      5 0.2436858 0.1800027
                      6 0.2738245 0.1951084
                      7 0.3751759 0.1944322
                      8 0.3911111 0.1995256

                       residual.sugar
as.factor(wine[, 12])      [,1]     [,2]
                      3 2.635000 1.401596
                      4 2.694340 1.789436
                      5 2.528855 1.359753
                      6 2.477194 1.441576
                      7 2.720603 1.371509
                      8 2.577778 1.295038
```



```
                       chlorides
as.factor(wine[, 12])         [,1]       [,2]
                      3 0.12250000 0.06624072
                      4 0.09067925 0.07619176
                      5 0.09273568 0.05370741
                      6 0.08495611 0.03956329
                      7 0.07658794 0.02945551
                      8 0.06844444 0.01167815

                       free.sulfur.dioxide
as.factor(wine[, 12])      [,1]      [,2]
                      3 11.00000  9.763879
                      4 12.26415  9.025926
                      5 16.98385 10.955446
                      6 15.71160  9.940911
                      7 14.04523 10.175255
                      8 13.27778 11.155613

                       total.sulfur.dioxide
as.factor(wine[, 12])      [,1]     [,2]
                      3 24.90000 16.82888
                      4 36.24528 27.58337
                      5 56.51395 36.99312
                      6 40.86991 25.03825
                      7 35.02010 33.19121
                      8 33.44444 25.43324

                       density
as.factor(wine[, 12])       [,1]        [,2]
                      3 0.9974640 0.002001845
                      4 0.9965425 0.001575169
                      5 0.9971036 0.001588504
                      6 0.9966151 0.002000009
                      7 0.9961043 0.002175739
                      8 0.9952122 0.002378276

                       pH
as.factor(wine[, 12])      [,1]      [,2]
                      3 3.398000 0.1440525
                      4 3.381509 0.1814408
                      5 3.304949 0.1506184
                      6 3.318072 0.1539951
                      7 3.290754 0.1501008
                      8 3.267222 0.2006403
```

The following figure shows the confusion matrix of my model. We can see that most objects are assigned to the correct label, but we can also see that the wine with quality 6 and wine with quality 7 are assigned with most error labels

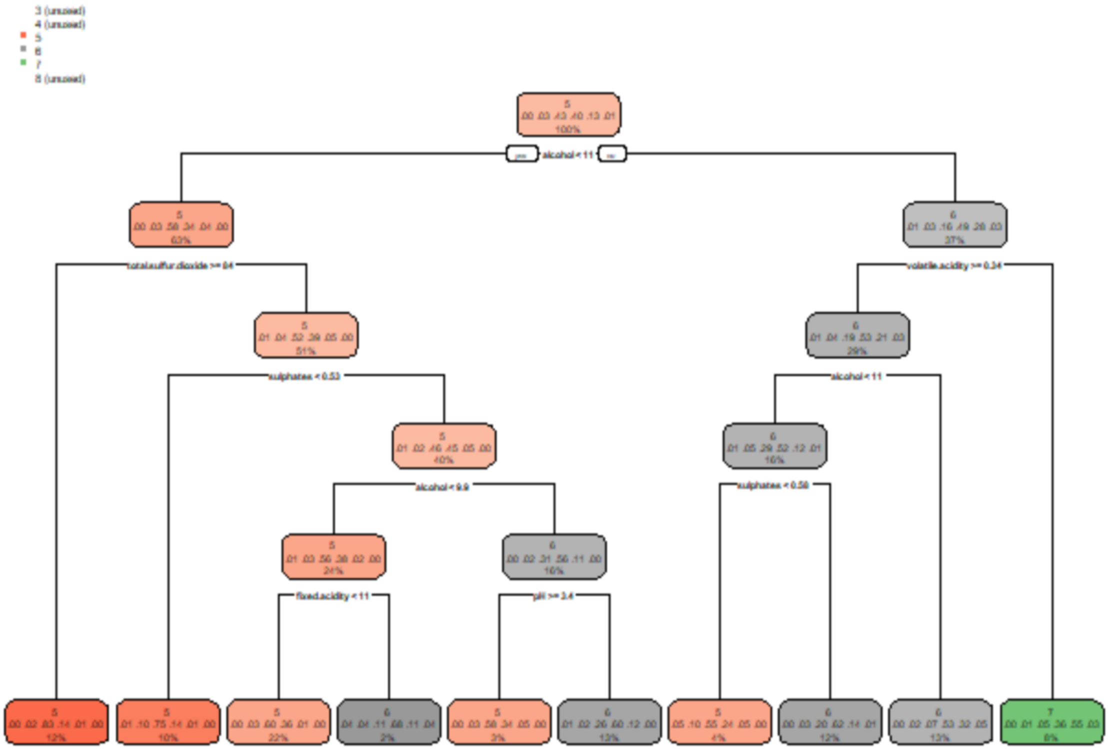


```
     3   4   5   6   7   8
  3  3   1   3   0   0   0
  4  2   8  26  21   1   0
  5  4  29 456 185  12   0
  6  1  13 171 316  74   5
  7  0   1  24 109 110  10
  8  0   1   1   7   2   3
```

Decision tree:

Since decision tree is a supervised learning technique to predict the label of the given input. The wine dataset is a labeled dataset, so decision tree can be helpful.

For the decision tree model, I split the dataset into train set and test set, train set contains the first 1400 rows and test set contains the remaining



3. **Conclusion:**

For the communities and crime dataset, I think both the regression model and Kmeans work well. Since the communities and crime dataset is a unlabeled dataset, and Kmeans works very well for the unlabeled dataset.

The drawback is that our datasets too many dimensions, it is very difficult to find a perfect number of clusters to separate these points clearly. The solution is that we can reduce the number of dimensions and remove some outliers. For example, the datasets contains the communities in foreign countries, and we can remove them when we train the models.

For the regression model, since we have many input variables, we can use regression model based on different input variables to get a model of highest accuracy. In the exploratory analysis, we have conducted the PCA analysis to find the most important factors for our output variable. Therefore regression model also works well on our datasets.

For the wine quality dataset, I think both the naïve Bayes model and regression model both work well. For the naïve Bayes model, since the output variable can act as the label for the dataset, we can treat the wine dataset as the labeled dataset. Naïve Bayes model can work quite well on the labeled dataset, and we can see from the confusion matrix that we shown in the above, the accuracy is considered to be high.

For the regression model, since we have some very detailed input variables and we know that they both impact our output variables from the exploratory analysis, we can also find the regression model works well for our model.