EDA of White Wine Quality Data

Sohan Aryal

7/15/2022

```
library(ggplot2)
library(dplyr)
```

Loading Libraries

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
wine <- read.csv2("winequality-white.csv")
glimpse(wine)
```

Loading Data

```
## Rows: 4,898
## Columns: 12
## $ fixed.acidity        <chr> "7", "6.3", "8.1", "7.2", "7.2", "8.1", "6.2", "7~
## $ volatile.acidity     <chr> "0.27", "0.3", "0.28", "0.23", "0.23", "0.28", "0~
## $ citric.acid          <chr> "0.36", "0.34", "0.4", "0.32", "0.32", "0.4", "0.~
## $ residual.sugar       <chr> "20.7", "1.6", "6.9", "8.5", "8.5", "6.9", "7", "~
## $ chlorides            <chr> "0.045", "0.049", "0.05", "0.058", "0.058", "0.05~
## $ free.sulfur.dioxide  <chr> "45", "14", "30", "47", "47", "30", "30", "45", "~
## $ total.sulfur.dioxide <chr> "170", "132", "97", "186", "186", "97", "136", "1~
## $ density              <chr> "1.001", "0.994", "0.9951", "0.9956", "0.9956", "~
## $ pH                   <chr> "3", "3.3", "3.26", "3.19", "3.19", "3.26", "3.18~
## $ sulphates            <chr> "0.45", "0.49", "0.44", "0.4", "0.4", "0.44", "0.~
## $ alcohol              <chr> "8.8", "9.5", "10.1", "9.9", "9.9", "10.1", "9.6"~
## $ quality              <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 5, 5, 7, 5, 7, 6~
```

Since all variable except quality are as character we need to convert them from character into numerical.

```
wine <- data.frame(lapply(wine, function(x) as.numeric(as.character(x))))
```

Let's see the summary of each variable in dataset.

```
summary(wine)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol          quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

Wine Quality Distribution Our main variable of interest in this dataset is wine quality. Let's look at the summary of wine quality in the following table and histogram.

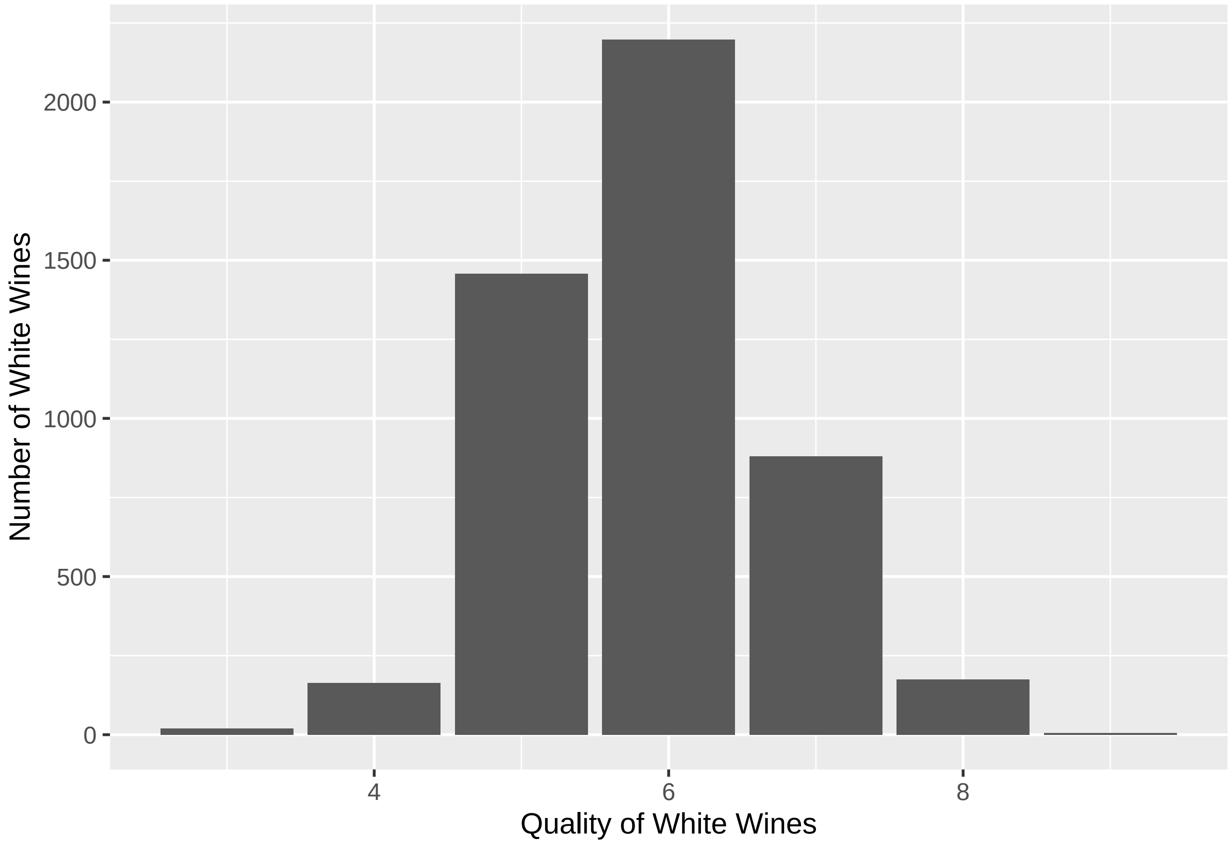
```
print("Number of wines with particular rating:")
```

```
## [1] "Number of wines with particular rating:"
```

```
table(wine$quality)
```

```
##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

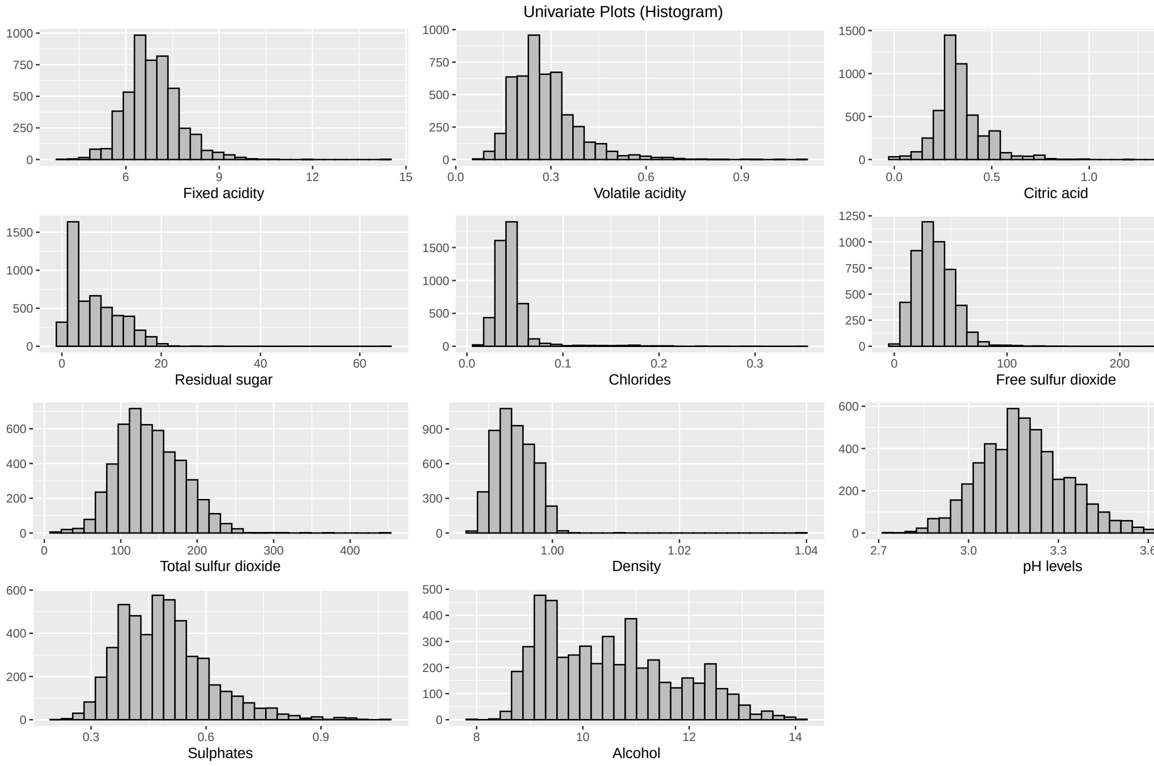
```
ggplot(aes(x = quality), data = wine) +
  geom_histogram(stat = "count") +
  xlab("Quality of White Wines") +
  ylab("Number of White Wines")
```



Quality of white wine has near-normal distribution as seen from the plot. Most of the wines are rated 5 to 7, while the very few are rated [3,4] and [8,9]. There is no wine with rating 1, 2 and 10.

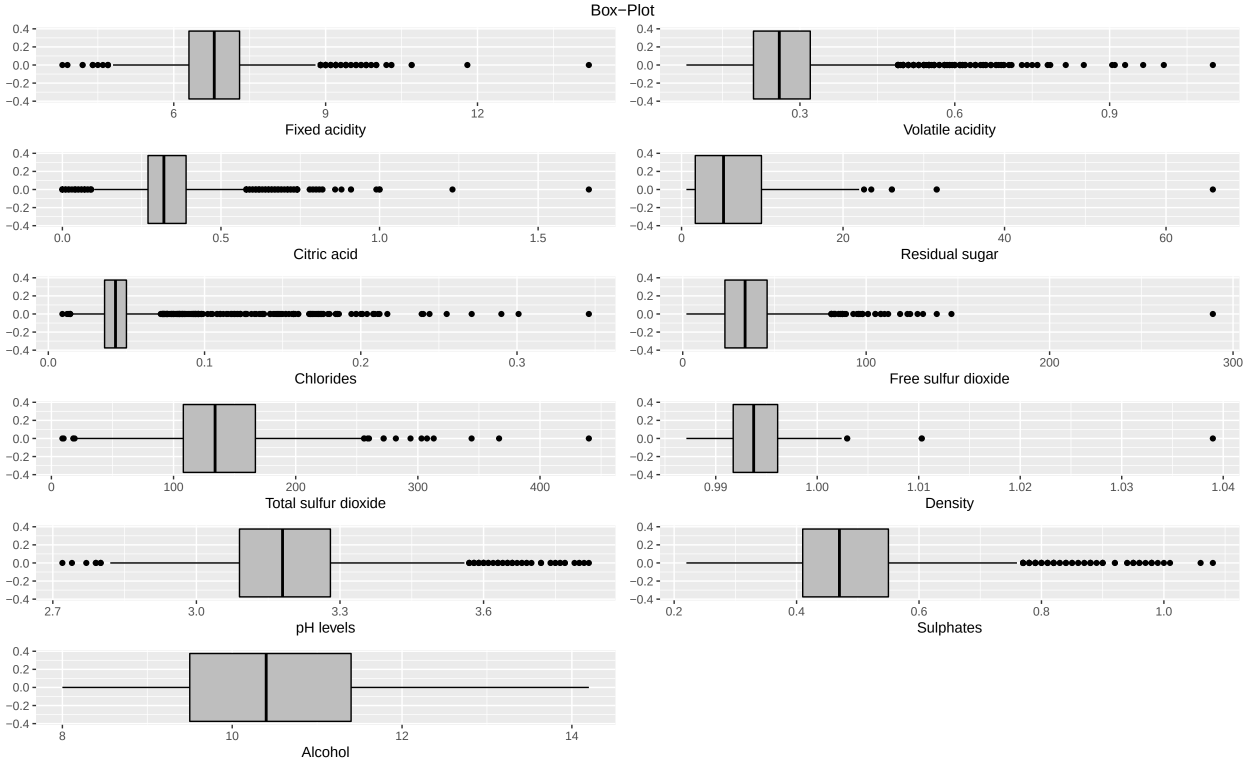
```r
p1 <- ggplot(aes(x=fixed.acidity),data = wine) +
  geom_histogram(color = "black", fill = "gray") +
  labs(x = 'Fixed acidity', y = NULL)

p12 <- ggplot(aes(x=fixed.acidity),data = wine) +
  geom_boxplot(color = "black", fill = "gray") +
  labs(x = 'Fixed acidity', y = NULL)

p2 <- ggplot(aes(x=volatile.acidity),data = wine) +
  geom_histogram(color = "black", fill = "gray") +
  labs(x = 'Volatile acidity', y = NULL)

p22 <- ggplot(aes(x=volatile.acidity),data = wine) +
  geom_boxplot(color = "black", fill = "gray") +
  labs(x = 'Volatile acidity', y = NULL)

p3 <- ggplot(aes(x=citric.acid),data = wine) +
  geom_histogram(color = "black", fill = "gray") +
  labs(x = 'Citric acid', y = NULL)

p32 <- ggplot(aes(x=citric.acid),data = wine) +
  geom_boxplot(color = "black", fill = "gray") +
  labs(x = 'Citric acid', y = NULL)

p4 <- ggplot(aes(x=residual.sugar),data = wine) +
  geom_histogram(color = "black", fill = "gray") +
  labs(x = 'Residual sugar', y = NULL)

p42 <- ggplot(aes(x=residual.sugar),data = wine) +
  geom_boxplot(color = "black", fill = "gray") +
  labs(x = 'Residual sugar', y = NULL)

p5 <- ggplot(aes(x=chlorides),data = wine) +
  geom_histogram(color = "black", fill = "gray") +
  labs(x = 'Chlorides', y = NULL)

p52 <- ggplot(aes(x=chlorides),data = wine) +
  geom_boxplot(color = "black", fill = "gray") +
  labs(x = 'Chlorides', y = NULL)

p6 <- ggplot(aes(x=free.sulfur.dioxide),data = wine) +
  geom_histogram(color = "black", fill = "gray") +
  labs(x = 'Free sulfur dioxide', y = NULL)

p62 <- ggplot(aes(x=free.sulfur.dioxide),data = wine) +
  geom_boxplot(color = "black", fill = "gray") +
  labs(x = 'Free sulfur dioxide', y = NULL)

p7 <- ggplot(aes(x=total.sulfur.dioxide),data = wine) +
  geom_histogram(color = "black", fill = "gray") +
  labs(x = 'Total sulfur dioxide', y = NULL)
```

```r
p72 <- ggplot(aes(x=total.sulfur.dioxide),data = wine) +
  geom_boxplot(color = "black", fill = "gray") +
  labs(x = 'Total sulfur dioxide', y = NULL)

p8 <- ggplot(aes(x=density),data = wine) +
  geom_histogram(color = "black", fill = "gray") +
  labs(x = 'Density', y = NULL)

p82 <- ggplot(aes(x=density),data = wine) +
  geom_boxplot(color = "black", fill = "gray") +
  labs(x = 'Density', y = NULL)

p9 <- ggplot(aes(x=pH),data = wine) +
  geom_histogram(color = "black", fill = "gray") +
  labs(x = 'pH levels', y = NULL)

p92 <- ggplot(aes(x=pH),data = wine) +
  geom_boxplot(color = "black", fill = "gray") +
  labs(x = 'pH levels', y = NULL)

p10 <- ggplot(aes(x=sulphates),data = wine) +
  geom_histogram(color = "black", fill = "gray") +
  labs(x = 'Sulphates', y = NULL)

p102 <- ggplot(aes(x=sulphates),data = wine) +
  geom_boxplot(color = "black", fill = "gray") +
  labs(x = 'Sulphates', y = NULL)

p11 <- ggplot(aes(x=alcohol),data = wine) +
  geom_histogram(color = "black", fill = "gray") +
  labs(x = 'Alcohol', y = NULL)

p112 <- ggplot(aes(x=alcohol),data = wine) +
  geom_boxplot(color = "black", fill = "gray") +
  labs(x = 'Alcohol', y = NULL)

grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, p9, p10, p11, ncol = 3, top = "Univariate Plots (Histogram
```



Univariate Analysis

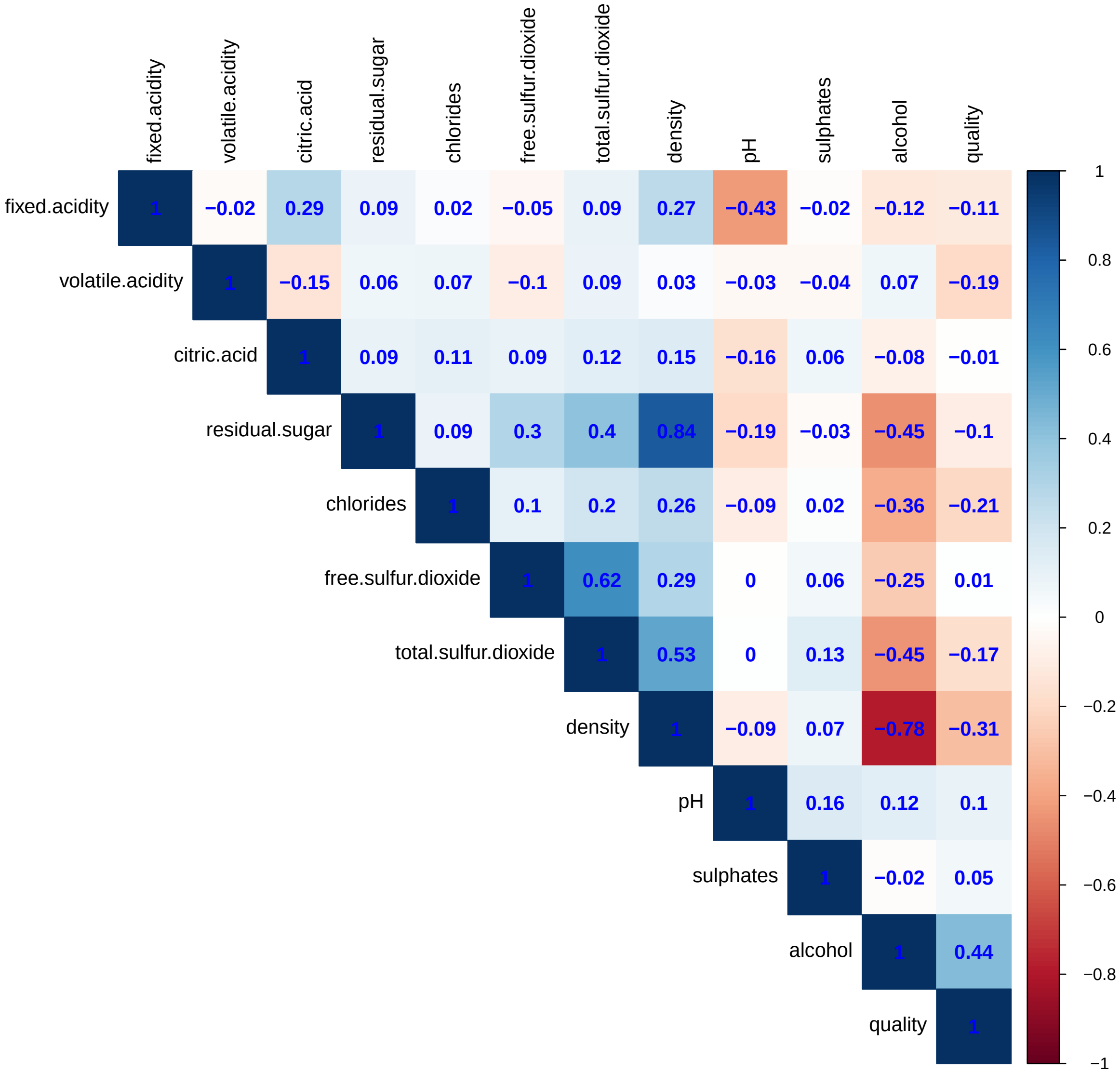
```
grid.arrange(p12, p22, p32, p42, p52, p62, p72, p82, p92, p102, p112, ncol = 2, top ="Box-Plot")
```



Variables like fixed acidity, volatile acidity and citric acid have outliers; which when removed their distribution looks symmetric. Some variables like residual sugar doesn't become symmetric even when their outliers are removed. Mostly outliers of these variables are on the larger values. Alcohol doesn't have outliers but have some irregular type of distribution.

Bivariate Plots Let's do the correlation plot to see how associated variables are.

```
wq.corr <- cor(wine)
corrplot(wq.corr, method='color', type = 'upper',
         tl.col = "black", addCoef.col='blue')
```



From correlation matrix, we see how quality of wine is related with other chemical properties of wine. * Quality and alcohol have moderate correlation (0.44), it is highest amongst all other chemical properties. * Quality and density has lower negative correlation (-0.31). * Quality has very low negative correlation with total sulfur dioxide, chlorides and volatile acidity. * Quality has very weak negative correlation with residual sugar(-0.1), citric acid(-0.01) and fixed acidity(-0.11). * Quality has very weak positive correlation with pH(0.1), sulphates(0.05) and free sulphur dioxide(0.01).

- Alcohol has strong negative correlation (-0.78) with the density of wine.
- Density is strongly correlated (0.84) with residual sugar quantity.
- Alcohol has moderate negative correlations with Residual sugar(-0.45), Total SO2 (-0.45) and Chlorides (-0.36).

Overall it was seen that main taste qualities of wine, such as acidity and sweetness have a low effect on quality. The highest positive correlation of quality rating appears to be with alcohol. Density is a variable which seems to be affected by other variables the most. In fact strongest relationship in the given sample is with Density and Alcohol.

Data Preparation From above plot, we have seen that most outliers are on higher end of values, so we consider those observation as outlier which are above

$$Q_3 + 1.5 * IQR$$

.

```
df <- wine

limit <- rep(0,11)

for (i in 1:11){
  q3 <- quantile(df[,i], 0.75)
  iqr <- IQR(df[,i], 0.75)

  limit[i] <- q3 + 1.5*iqr
}

df_index <- matrix(0, 4898,11)



for (i in 1:4898)
  for (j in 1:11)
    if (df[i,j] > limit[j])
      df_index[i,j] <- 1


dfIndex <- base::apply(df_index, 1, sum)


df2 <- df[dfIndex == 0, ]

indexes = sample(1:nrow(df2), size=0.7*nrow(df2))
train <- df2[indexes,]
test <- df2[-indexes,]
```

Above procedure reduces the data size from 4899 to 4074. Data is randomly divided into Training data and Test Data in ratio of 7:3.

Regression Model Now we run linear regression to find out which variables are significant. We go on improving our model by omitting insignificant ones by checking improvement in R-square value.

Now we will create the model to predict quality of wines. We use the train data set.

```
model1 <- lm(quality ~ ., data = train)
summary(model1)
```

```
##
## Call:
## lm(formula = quality ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2533 -0.5200 -0.0395  0.4722  2.4993
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.184e+02  3.278e+01   6.662 3.24e-11 ***
## fixed.acidity         1.718e-01  3.247e-02   5.290 1.31e-07 ***
## volatile.acidity     -1.826e+00  1.953e-01  -9.351  < 2e-16 ***
## citric.acid           9.873e-02  1.636e-01   0.604  0.54619
## residual.sugar        1.057e-01  1.224e-02   8.637  < 2e-16 ***
## chlorides            -2.712e+00  1.747e+00  -1.552  0.12083
## free.sulfur.dioxide   4.457e-03  1.257e-03   3.545  0.00040 ***
## total.sulfur.dioxide  1.550e-04  5.319e-04   0.291  0.77072
## density              -2.205e+02  3.321e+01  -6.641 3.73e-11 ***
## pH                    1.171e+00  1.523e-01   7.686 2.07e-14 ***
## sulphates             7.956e-01  1.480e-01   5.375 8.27e-08 ***
## alcohol               1.114e-01  4.116e-02   2.707  0.00682 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.738 on 2839 degrees of freedom
## Multiple R-squared:  0.2652, Adjusted R-squared:  0.2624
## F-statistic: 93.15 on 11 and 2839 DF,  p-value: < 2.2e-16
```

Citric acid and Total sulfur dioxide are not significant in this model with adjusted R-square of 0.2407. Now we create model without these variables.

```
model2 <- lm(quality ~ . -citric.acid -total.sulfur.dioxide, data = train)
summary(model2)
```

```
##
## Call:
## lm(formula = quality ~ . - citric.acid - total.sulfur.dioxide,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2514 -0.5154 -0.0451  0.4741  2.4905
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.136e+02  3.142e+01   6.799 1.28e-11 ***
## fixed.acidity        1.720e-01  3.213e-02   5.354 9.31e-08 ***
## volatile.acidity    -1.832e+00  1.860e-01  -9.852  < 2e-16 ***
## residual.sugar       1.041e-01  1.186e-02   8.780  < 2e-16 ***
## chlorides           -2.657e+00  1.738e+00  -1.528  0.12653
## free.sulfur.dioxide  4.751e-03  9.981e-04   4.760 2.03e-06 ***
## density             -2.157e+02  3.184e+01  -6.776 1.49e-11 ***
## pH                   1.157e+00  1.509e-01   7.670 2.34e-14 ***
## sulphates            7.993e-01  1.479e-01   5.405 7.00e-08 ***
## alcohol              1.166e-01  4.035e-02   2.891  0.00387 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7378 on 2841 degrees of freedom
## Multiple R-squared:  0.2651, Adjusted R-squared:  0.2628
## F-statistic: 113.9 on 9 and 2841 DF,  p-value: < 2.2e-16
```

It didn’t quite improve the model. But all the variables are significant now. We can see that predictors are continuous while quality is integer. Prediction made will not be integer, so it is difficult to explain the variation in quality using linear model. We can check the prediction by our model2.

```
predictTest <- predict.lm(model2, newdata = test)
summary(predictTest)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.627   5.624   5.927   5.946   6.284   7.171
```

The model2 didn’t predict quality over 7.165. So now we will use different method for prediction.

Logistic Regression

We will make a logistics regression model to predict the quality of wine using input variables. First we will create a subset of wines rated 8 or above effectively making quality having two factors. Quality rating > 7 : wine is “Good” Quality rating <=7 : wine is “Not good”

```
train$rating <- ifelse (as.integer(train$quality) > 7, 1, 0)

test$rating <- ifelse (as.integer(test$quality) > 7, 1, 0)

model3 = glm(rating ~.-quality ,data=train,family = "binomial")
summary(model3)
```

```
##
## Call:
## glm(formula = rating ~ . - quality, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9118  -0.2997  -0.1995  -0.1339   3.1460
##
```

```
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           5.573e+02  2.682e+02   2.078 0.037708 *
## fixed.acidity         6.246e-01  2.551e-01   2.448 0.014362 *
## volatile.acidity     -2.539e+00  1.460e+00  -1.739 0.081988 .
## citric.acid          -3.224e-01  1.417e+00  -0.227 0.820056
## residual.sugar        3.342e-01  1.012e-01   3.300 0.000965 ***
## chlorides             5.640e+00  1.313e+01   0.429 0.667574
## free.sulfur.dioxide   2.412e-02  9.832e-03   2.453 0.014172 *
## total.sulfur.dioxide -4.224e-03  4.499e-03  -0.939 0.347849
## density              -5.876e+02  2.717e+02  -2.162 0.030587 *
## pH                    4.304e+00  1.175e+00   3.662 0.000250 ***
## sulphates            -2.774e-02  1.028e+00  -0.027 0.978477
## alcohol               2.871e-01  3.182e-01   0.902 0.366878
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 925.35  on 2850  degrees of freedom
## Residual deviance: 808.91  on 2839  degrees of freedom
## AIC: 832.91
##
## Number of Fisher Scoring iterations: 7
```

Significant variables are: fixed.acidity, residual.sugar and pH. We will use this model to predict over testing set.

```
prediction = predict(model3, newdata=test, type="response")
table(test$rating, prediction > 0.5)
```

```
##
##     FALSE
##   0  1179
##   1    44
```

Our model predicts that no wine is in the best rated wine group, but actually 54 of the wines are rated 8 or 9 in the test set. Our prediction is same as the baseline prediction. So, there is not much benefit from this model. Let’s check what is the maximum value of prediction over test set.

```
print("The maximum value of prediction over testing set is ")
```

```
## [1] "The maximum value of prediction over testing set is "
```

```
round(max(prediction),3)
```

```
## [1] 0.378
```

As the maximum value is less than 0.5, our model will always predict that wine quality will be less than 8.

Conclusion

We used linear regression model and logistic regression model. Though we see some association from correlation plot and significant variables on these two models, it seems like wine quality is not well supported by its chemical properties. At each quality level variability of the predictors is high.