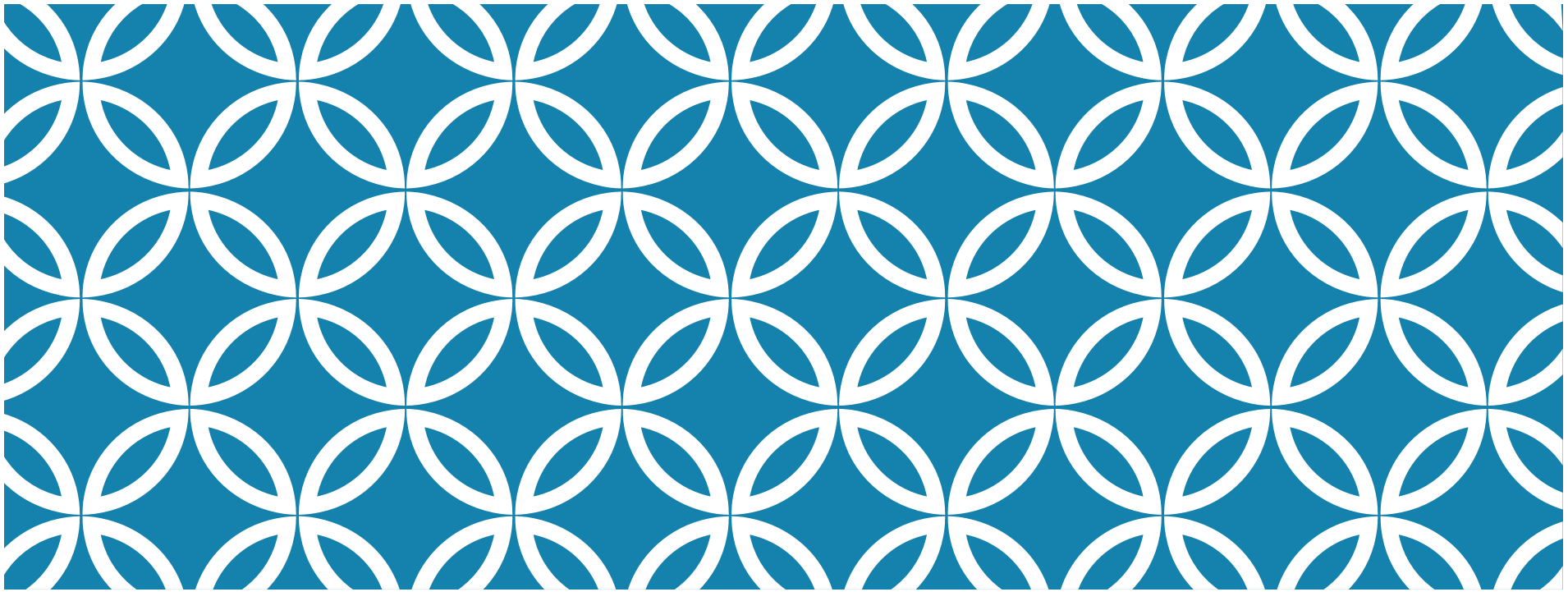


PREDICTING WINE QUALITY

DATASET

The dataset used is Wine Quality Data set from UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
12 - quality (0 - 10)

DATASET

2 different datasets

- White Wine Dataset – 4898 instances 12 columns
- Red Wine Dataset – 1599 instances 12 columns

```r
summary(winewhite)
```

```
 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
```

```r
summary(winered)
```

```
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100
 free.sulfur.dioxide total.sulfur.dioxide    density             pH
 Min.   : 1.00       Min.   :  6.00       Min.   :0.9901   Min.   :2.740
 1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210
 Median :14.00       Median : 38.00       Median :0.9968   Median :3.310
 Mean   :15.87       Mean   : 46.47       Mean   :0.9967   Mean   :3.311
 3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400
 Max.   :72.00       Max.   :289.00       Max.   :1.0037   Max.   :4.010
   sulphates         alcohol         quality
 Min.   :0.3300   Min.   : 8.40   Min.   :3.000
 1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
 Median :0.6200   Median :10.20   Median :6.000
 Mean   :0.6581   Mean   :10.42   Mean   :5.636
 3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
 Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

DATASET

Added wine type column to indicate wine type. (1- White Wine, 0- Red Wine)

6497 instances 13 columns

```{r}
summary(wine)
```

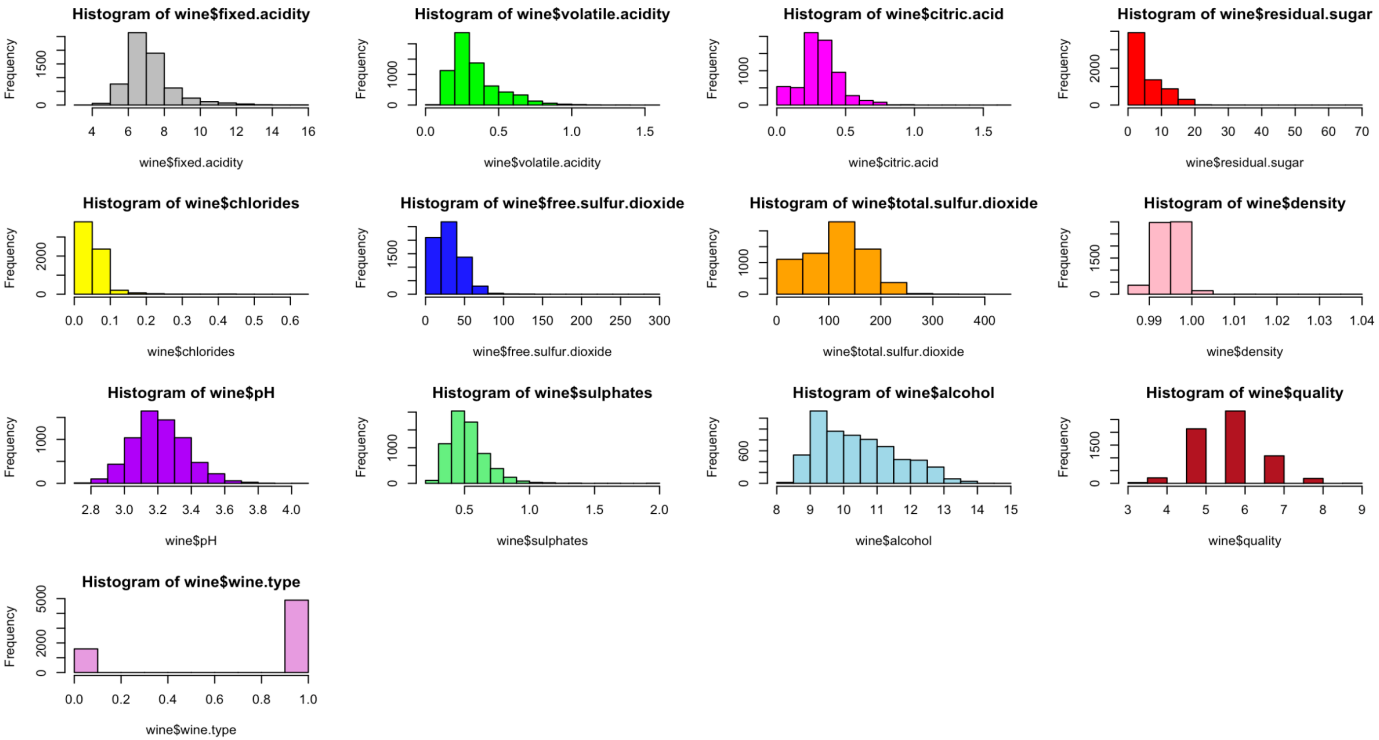
```
fixed.acidity    volatile.acidity  citric.acid     residual.sugar
Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
  chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :0.9871
1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923
Median :0.04700   Median : 29.00      Median :118.0        Median :0.9949
Mean   :0.05603   Mean   : 30.53      Mean   :115.7        Mean   :0.9947
3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970
Max.   :0.61100   Max.   :289.00      Max.   :440.0        Max.   :1.0390
      pH          sulphates         alcohol         quality        wine.type
Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000   Min.   :0.0000
1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000   1st Qu.:1.0000
Median :3.210   Median :0.5100   Median :10.30   Median :6.000   Median :1.0000
Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818   Mean   :0.7539
3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000   3rd Qu.:1.0000
Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :9.000   Max.   :1.0000
```

DESCRIPTIVE STATISTICS

- Pastecs Library – stat.desc function

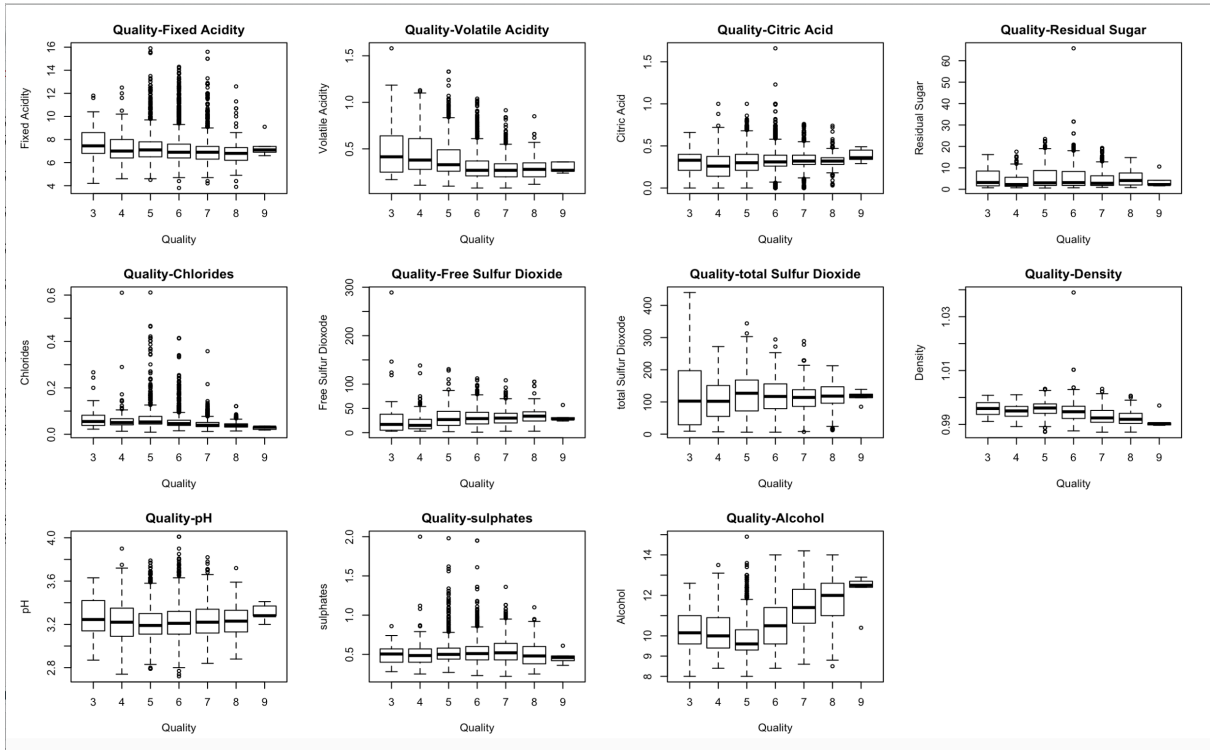
	fixed.acidity <dbl>	volatile.acidity <dbl>	citric.acid <dbl>	residual.sugar <dbl>	chlorides <dbl>	free.sulfur.dioxide <dbl>	total.sulfur.dioxide <dbl>	density <dbl>	pH <dbl>	sulphates <dbl>	alcohol <dbl>	quality <dbl>	wine.type <dbl>
nbr.val	6.497000e+03	6.497000e+03	6.497000e+03	6.497000e+03	6.497000e+03	6.497000e+03	6.497000e+03	6.497000e+03	6.497000e+03	6.497000e+03	6.497000e+03	6.497000e+03	6.497000e+03
nbr.null	0.000000e+00	0.000000e+00	1.510000e+02	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.599000e+03
nbr.na	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
min	3.800000e+00	8.000000e-02	0.000000e+00	6.000000e-01	9.000000e-03	1.000000e+00	6.000000e+00	9.871100e-01	2.720000e+00	2.200000e-01	8.000000e+00	3.000000e+00	0.000000e+00
max	1.590000e+01	1.580000e+00	1.660000e+00	6.580000e+01	6.110000e-01	2.890000e+02	4.400000e+02	1.038980e+00	4.010000e+00	2.000000e+00	1.490000e+01	9.000000e+00	1.000000e+00
range	1.210000e+01	1.500000e+00	1.660000e+00	6.520000e+01	6.020000e-01	2.880000e+02	4.340000e+02	5.187000e-02	1.290000e+00	1.780000e+00	6.900000e+00	6.000000e+00	1.000000e+00
sum	4.687785e+04	2.206810e+03	2.070160e+03	3.536470e+04	3.640520e+02	1.983230e+05	7.519925e+05	6.462544e+03	2.091060e+04	3.451650e+03	6.816523e+04	3.780200e+04	4.898000e+03
median	7.000000e+00	2.900000e-01	3.100000e-01	3.000000e+00	4.700000e-02	2.900000e+01	1.180000e+02	9.948900e-01	3.210000e+00	5.100000e-01	1.030000e+01	6.000000e+00	1.000000e+00
mean	7.215307e+00	3.396660e-01	3.186332e-01	5.443235e+00	5.603386e-02	3.052532e+01	1.157446e+02	9.946966e-01	3.218501e+00	5.312683e-01	1.049180e+01	5.818378e+00	7.538864e-01
SE.mean	1.608399e-02	2.042536e-03	1.802862e-03	5.902692e-02	4.346387e-04	2.202050e-01	7.012292e-01	3.720255e-05	1.994780e-03	1.846136e-03	1.479718e-02	1.083390e-02	5.344385e-03
CI.mean.0.95	3.152992e-02	4.004042e-03	3.534204e-03	1.157122e-01	8.520349e-04	4.316744e-01	1.374640e+00	7.292924e-05	3.910426e-03	3.619034e-03	2.900735e-02	2.123801e-02	1.047675e-02
var	1.680740e+00	2.710517e-02	2.111728e-02	2.263670e+01	1.227353e-03	3.150412e+02	3.194720e+03	8.992040e-06	2.585252e-02	2.214319e-02	1.422561e+00	7.625748e-01	1.855703e-01
std.dev	1.296434e+00	1.646365e-01	1.453179e-01	4.757804e+00	3.503360e-02	1.774940e+01	5.652185e+01	2.998673e-03	1.607872e-01	1.488059e-01	1.192712e+00	8.732553e-01	4.307787e-01
coef.var	1.796783e-01	4.847011e-01	4.560663e-01	8.740764e-01	6.252220e-01	5.814648e-01	4.883326e-01	3.014661e-03	4.995717e-02	2.800955e-01	1.136804e-01	1.500857e-01	5.714106e-01

DESCRIPTIVE STATISTICS



DESCRIPTIVE ANALYSIS

BOX-PLOT ANALYSIS

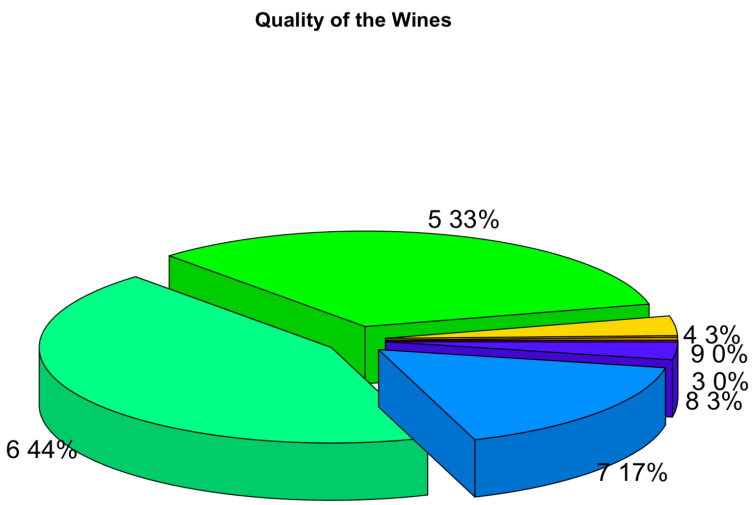


DIAGNOSTIC ANALYSIS

- Distribution with respect to Wine Quality

```{r}
table(wine$quality)
```

```
  3   4   5   6   7   8   9 
 30 216 2138 2836 1079 193   5 
```



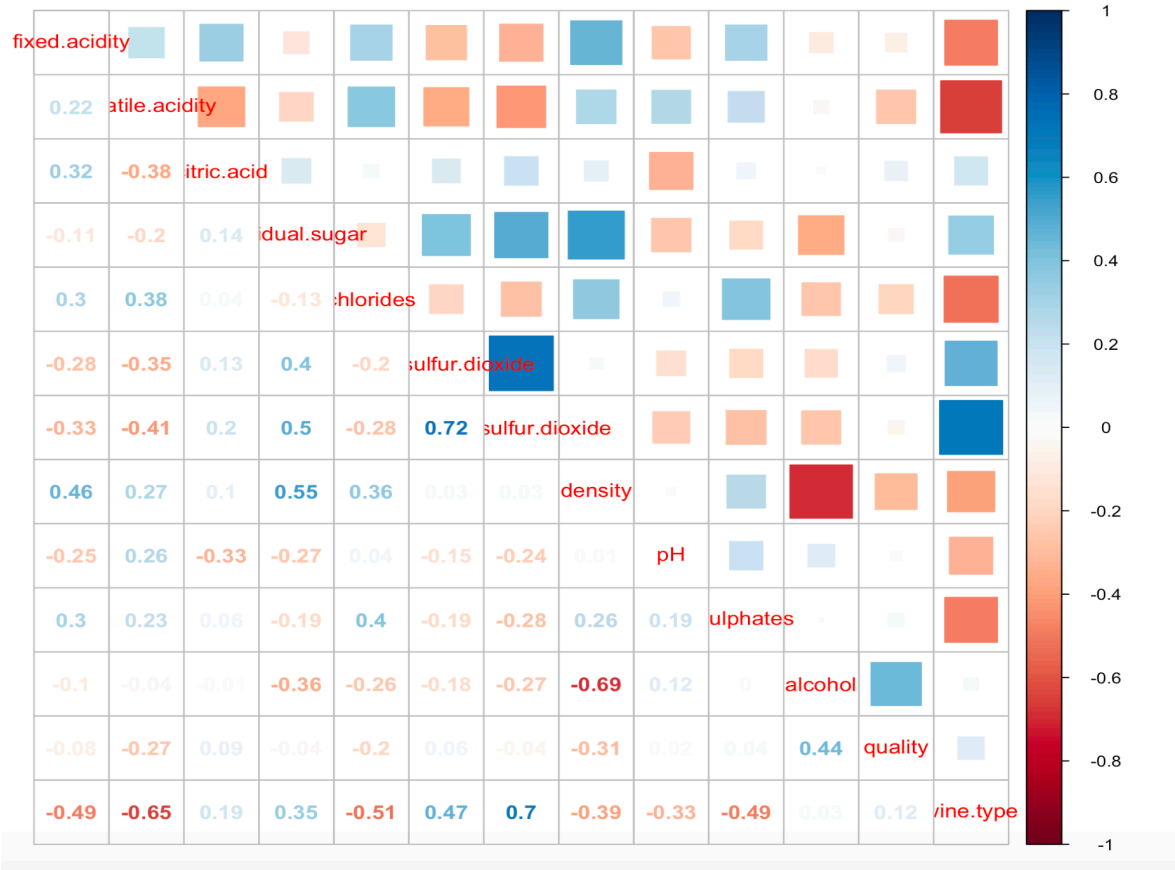
Quality	Fixed Acidity	Volatile Acidity	Citric acid	residual sugar	chlorides	free Sulphur dioxide	total sulphur dioxide	density	pH	sulphates	alcohol	quality	wine type
9	7.42	0.298	0.386	4.12	0.0274	33.4	116	0.9915	3.308	0.466	12.18	9	1
8	6.835	0.291	0.3325	5.383	0.04112	34.53	117.5	0.9925	3.223	0.5125	11.68	8	0.9067
7	7.129	0.2888	0.3348	4.732	0.04527	30.42	108.5	0.9931	3.228	0.547	11.39	7	0.85156
6	7.177	0.3139	0.3236	5.55	0.05416	31.17	115.4	0.9946	3.218	0.5325	10.59	6	0.775
5	7.327	0.3896	0.3077	5.804	0.066467	30.24	120.8	0.9958	3.212	0.52664	9.838	5	0.6815
4	7.289	0.458	0.2723	4.154	0.660006	20.64	103.4	0.9948	3.232	0.5056	10.18	4	0.7546
3	7.853	0.517	0.281	5.14	0.07703	39.22	122	0.9957	3.219	3.258	0.5063	4	0.6667
Overall Mean	7.215	0.3397	0.3186	5.443	0.05603	30.53	115.7	0.9947	3.219	0.5313	10.49	5.818	0.7539

DIAGNOSTIC ANALYSIS

Mean Values Respect to Wine Quality

PREDICTIVE MODELING

Correlation Matrix



TRAINING DATA (%80)

```
'data.frame':   5199 obs. of  13 variables:
 $ fixed.acidity       : num  7.4 7.8 11.2 7.4 7.4 7.8 7.5 6.7 7.5 7.8 ...
 $ volatile.acidity    : num  0.7 0.76 0.28 0.7 0.66 0.58 0.5 0.58 0.5 0.61 ...
 $ citric.acid         : num  0 0.04 0.56 0 0 0.02 0.36 0.08 0.36 0.29 ...
 $ residual.sugar      : num  1.9 2.3 1.9 1.9 1.8 2 6.1 1.8 6.1 1.6 ...
 $ chlorides           : num  0.076 0.092 0.075 0.076 0.075 0.073 0.071 0.097 0.071
0.114 ...
 $ free.sulfur.dioxide : num  11 15 17 11 13 9 17 15 17 9 ...
 $ total.sulfur.dioxide: num  34 54 60 34 40 18 102 65 102 29 ...
 $ density             : num  0.998 0.997 0.998 0.998 0.998 ...
 $ pH                  : num  3.51 3.26 3.16 3.51 3.51 3.36 3.35 3.28 3.35 3.26 ...
 $ sulphates           : num  0.56 0.65 0.58 0.56 0.56 0.57 0.8 0.54 0.8 1.56 ...
 $ alcohol             : num  9.4 9.8 9.8 9.4 9.4 9.5 10.5 9.2 10.5 9.1 ...
 $ quality             : int  5 5 6 5 5 7 5 5 5 5 ...
 $ wine.type           : num  0 0 0 0 0 0 0 0 0 0 ...
```

TESTING DATA (%20)

```
'data.frame':   1298 obs. of  13 variables:
 $ fixed.acidity       : num  7.8 7.9 7.3 5.6 7.9 7.6 8.3 7.8 7.8 5.7 ...
 $ volatile.acidity    : num  0.88 0.6 0.65 0.615 0.32 0.41 0.655 0.645 0.6 1.13 ...
 $ citric.acid         : num  0 0.06 0 0 0.51 0.24 0.12 0 0.14 0.09 ...
 $ residual.sugar      : num  2.6 1.6 1.2 1.6 1.8 1.8 2.3 5.5 2.4 1.5 ...
 $ chlorides           : num  0.098 0.069 0.065 0.089 0.341 0.08 0.083 0.086 0.086 0.172
...
 $ free.sulfur.dioxide : num  25 15 15 16 17 4 15 5 3 7 ...
 $ total.sulfur.dioxide: num  67 59 21 59 56 11 113 18 15 19 ...
 $ density             : num  0.997 0.996 0.995 0.994 0.997 ...
 $ pH                  : num  3.2 3.3 3.39 3.58 3.04 3.28 3.17 3.4 3.42 3.5 ...
 $ sulphates           : num  0.68 0.46 0.47 0.52 1.08 0.59 0.66 0.55 0.6 0.48 ...
 $ alcohol             : num  9.8 9.4 10 9.9 9.2 9.5 9.8 9.6 10.8 9.8 ...
 $ quality             : int  5 5 7 5 6 5 5 6 6 4 ...
 $ wine.type           : num  0 0 0 0 0 0 0 0 0 0 ...
```

PREDICTIVE MODELING

Splitted Data for Training and Testing

PREDICTIVE MODELING

LINEAR REGRESSION

First Model (All variables excluding wine type)

```
Call:
lm(formula = quality ~ . - wine.type, data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6383 -0.4545 -0.0407  0.4583  3.0070

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.263e+01  1.301e+01   4.045 5.32e-05 ***
fixed.acidity         5.793e-02  1.714e-02   3.380 0.000731 ***
volatile.acidity     -1.407e+00  8.753e-02 -16.079  < 2e-16 ***
citric.acid          -7.200e-02  8.864e-02  -0.812 0.416692
residual.sugar        4.072e-02  5.693e-03   7.153 9.70e-13 ***
chlorides            -3.989e-01  3.691e-01  -1.081 0.279885
free.sulfur.dioxide   5.446e-03  8.371e-04   6.506 8.45e-11 ***
total.sulfur.dioxide -2.439e-03  3.090e-04  -7.893 3.56e-15 ***
density              -5.160e+01  1.328e+01  -3.885 0.000104 ***
pH                    4.021e-01  1.001e-01   4.018 5.95e-05 ***
sulphates             7.537e-01  8.340e-02   9.037  < 2e-16 ***
alcohol               2.685e-01  1.829e-02  14.685  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7345 on 5187 degrees of freedom
Multiple R-squared:  0.2942,    Adjusted R-squared:  0.2927
F-statistic: 196.5 on 11 and 5187 DF,  p-value: < 2.2e-16
```

Model Coefficients

```
Call:
lm(formula = quality ~ . - wine.type, data = trainData)

Coefficients:
         (Intercept)         fixed.acidity      volatile.acidity
           52.631433              0.057931             -1.407340
         citric.acid        residual.sugar             chlorides
           -0.072000              0.040723             -0.398893
 free.sulfur.dioxide  total.sulfur.dioxide               density
            0.005446             -0.002439            -51.602005
                  pH             sulphates               alcohol
            0.402082              0.753722              0.268539
```

Accuracy

	ME	RMSE	MAE	MPE	MAPE
Test set	5.810139	5.830144	5.810139	101.3736	101.3736

PREDICTIVE MODELING
LINEAR REGRESSION

Second Model (All variables –(wine type, citric acid, residual sugar)

```
Call:
lm(formula = quality ~ . - wine.type - citric.acid - chlorides,
    data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6231 -0.4607 -0.0364  0.4556  3.0074

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.584e+01  1.277e+01   4.371 1.26e-05 ***
fixed.acidity         5.717e-02  1.654e-02   3.456 0.000552 ***
volatile.acidity     -1.395e+00  7.965e-02 -17.512  < 2e-16 ***
residual.sugar        4.213e-02  5.563e-03   7.573 4.29e-14 ***
free.sulfur.dioxide   5.414e-03  8.355e-04   6.480 1.00e-10 ***
total.sulfur.dioxide -2.449e-03  3.030e-04  -8.085 7.68e-16 ***
density              -5.494e+01  1.303e+01  -4.217 2.51e-05 ***
pH                    4.308e-01  9.806e-02   4.394 1.14e-05 ***
sulphates             7.289e-01  8.142e-02   8.953  < 2e-16 ***
alcohol               2.672e-01  1.820e-02  14.681  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7345 on 5189 degrees of freedom
Multiple R-squared:  0.2939,    Adjusted R-squared:  0.2927
F-statistic:   240 on 9 and 5189 DF,  p-value: < 2.2e-16
```

Model Coefficients

```
Call:
lm(formula = quality ~ . - wine.type - citric.acid - chlorides,
    data = trainData)

Coefficients:
        (Intercept)         fixed.acidity      volatile.acidity
          55.835378              0.057165             -1.394742
     residual.sugar   free.sulfur.dioxide  total.sulfur.dioxide
           0.042128              0.005414             -0.002449
            density                    pH             sulphates
          -54.938843              0.430848              0.728894
            alcohol
           0.267235
```

Accuracy

```
                ME     RMSE      MAE      MPE     MAPE
Test set 5.810087 5.830086 5.810087 101.3745 101.3745
```

CONCLUSION

Model 2 works better than Model1.

If you want to buy a good quality wine you should look for these values.

Quality	Fixed Acidity	Volatile Acidity	Citric acid	residual sugar	chlorides	free Sulphur dioxide	total sulphur dioxide	density	pH	sulphates	alcohol	quality	wine type
9	7.42	0.298	0.386	4.12	0.0274	33.4	116	0.9915	3.308	0.466	12.18	9	1