Red_Wine_Quality_Analysis

Flora Li

What factors will influence the quality of red wine?

1. Introduction

The Red Wine Quality dataset was created by using red and white wine samples. The inputs include objective tests (e.g. PH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 3 (very bad) and 8 (very excellent). This dataset included 1599 observations and effects of 11 different chemical properties.

```
## [1] 1599   13
```

```
##  [1] "X"                   "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"         "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide" "total.sulfur.dioxide" "density"
## [10] "pH"                  "sulphates"            "alcohol"
## [13] "quality"
```

X is a unique identifier with a integer value. Quality is also an integer value. All other values are numeric value. In this dataset, we mainly focus on factors impacting wine quality, so the quality is dependent variable.

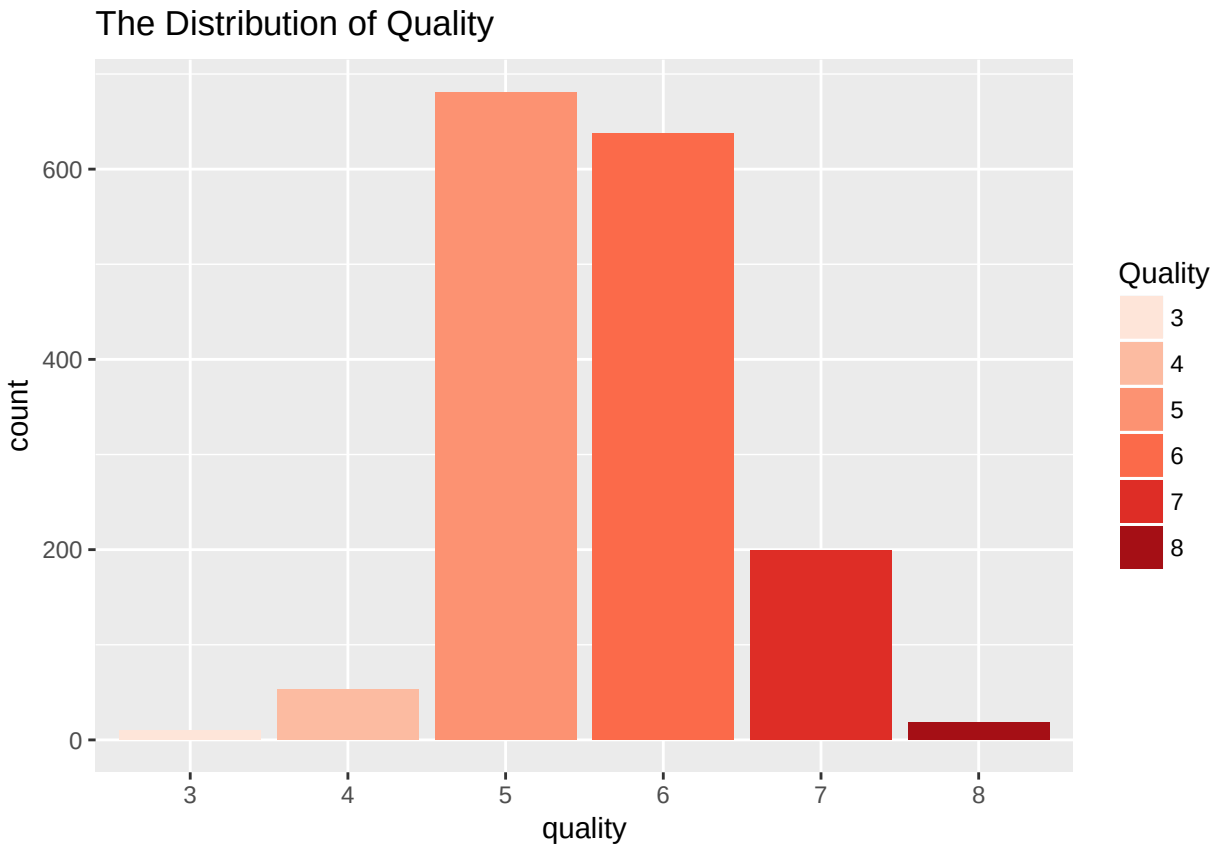
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol         quality
##  Min.   : 8.40   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.20   Median :6.000
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

From the summary, we can find out the average quality is 5.6 and the median quality is 6.

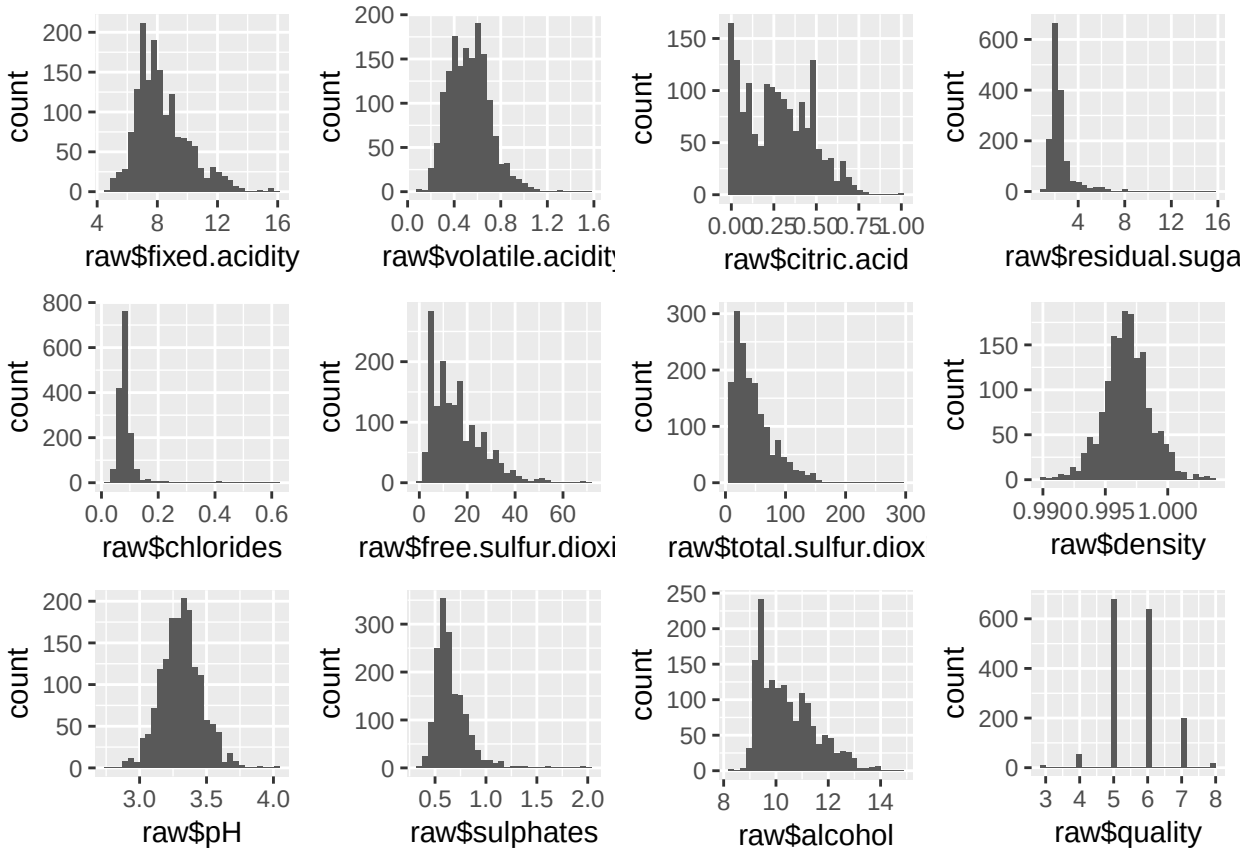
2. Univariate Plots Section

Analyzing variable : Quality

```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```



The plot shows the wine quality basically has a normal distribution. Most of the quality are around 5 and 6. Then we can plot the distribution of all the variables.



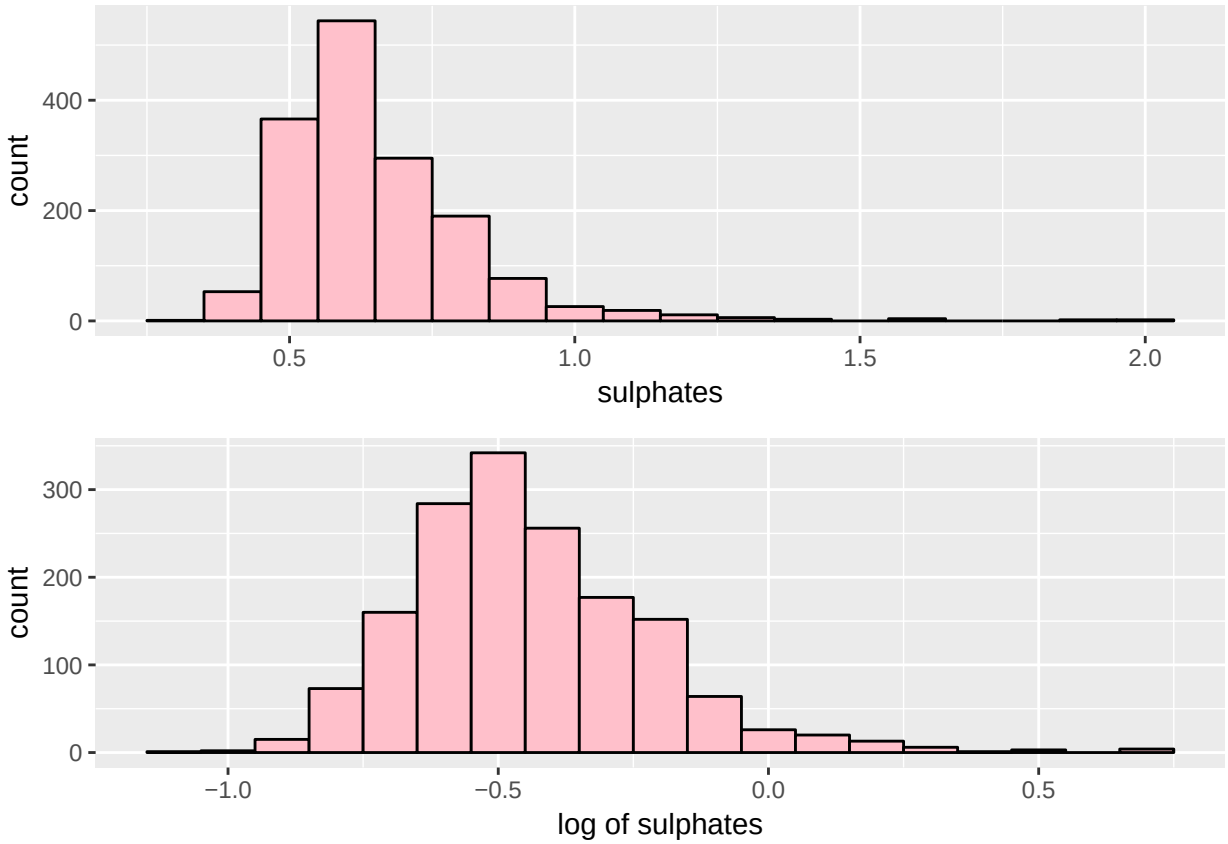

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

From the plots above, we can find out: 1.Density and ph are basically normal distributed. 2.Fixed and volatile acidity, sulfur dioxides, sulphates, and alcohol are right skewed distributions, with means their means are larger than median. 3.Residual sugar and chlorides are also right skewed and they have extreme outliers.

Rescale features

Skewed and long tail data can be transformed toward more normally distribution by taking square root or log function. Since residual sugar have lower correlation with quality (see Bivariate Plots Section), I just log-transformed the sulphates to compare the differences.



For the first plot, we can see feature Sulphates have really large numbers and is right skewed. After log transformation, the sulphates is more normal distributed.

Univariate Analysis

What is the structure of your dataset?

```
## 'data.frame':    1599 obs. of  14 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
```

```
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
##  $ log_sulphates       : num  -0.58 -0.386 -0.431 -0.545 -0.58 ...
```

What is/are the main feature(s) of interest in your dataset?

The quality of red wine is the main feature of interest in this dataset.

What other features in the dataset do you think will help support your ### investigation into your feature(s) of interest?

From the plot above, I think the variables related to acidity (fixed, volatile, citric.acid and pH) will influence the taste of red wine. Citric acid provides freshness taste to wines. Thus, better red wine would contain higher citric acid rate. Volatile Acidity provides unpleasant and vinegar taste to wines so the lower the volatile acidity, the better the red wine is. Besides, the percentage of alcohol will also influence the quality of wine. With higher percent of alcohol, the red wine would have better quality. Residual.sugar dictates how sweet a red wine is and will also have an influence in taste.

Did you create any new variables from existing variables in the dataset?

No. I didn'd create any new variables.

Of the features you investigated, were there any unusual distributions? ### Did you perform any operations on the data to tidy, adjust, or change the form ### of the data? If so, why did you do this?

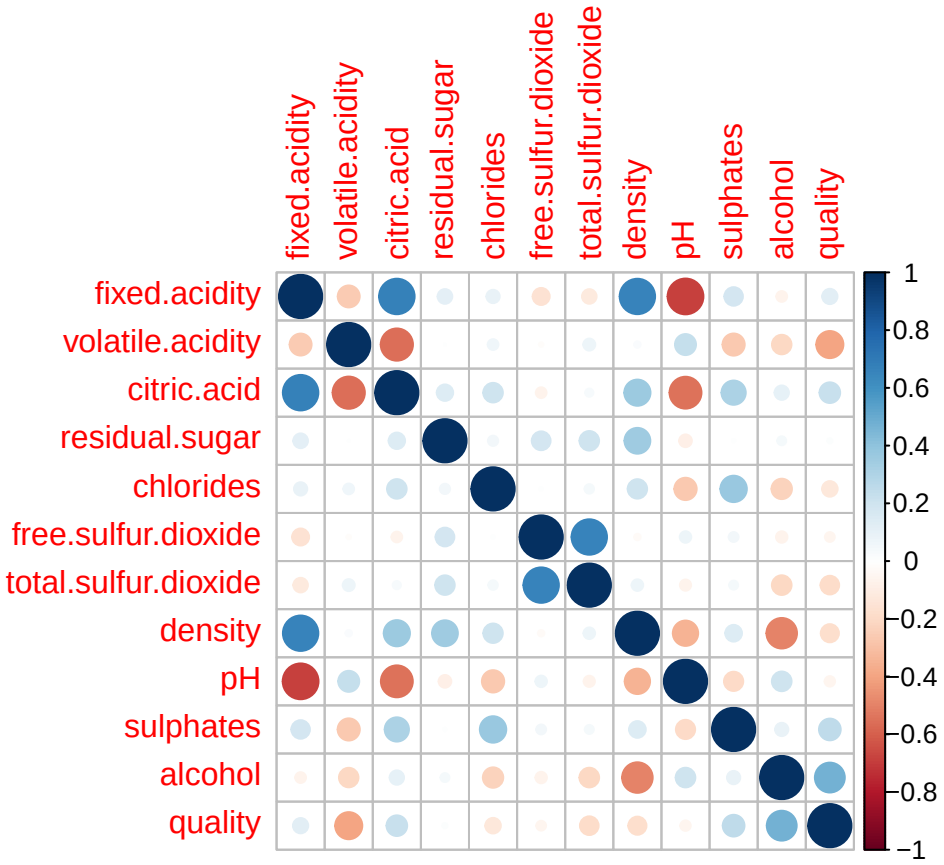
For the variable "Quality", there are no unusual distributions. It's basically normal distributed. For the other features such as sulphates and acidity,they are long-tailed.I tried log transformation because the it will be more normally distributed after log transformation.

3. Bivariate Plots Section

Correlation Matrix

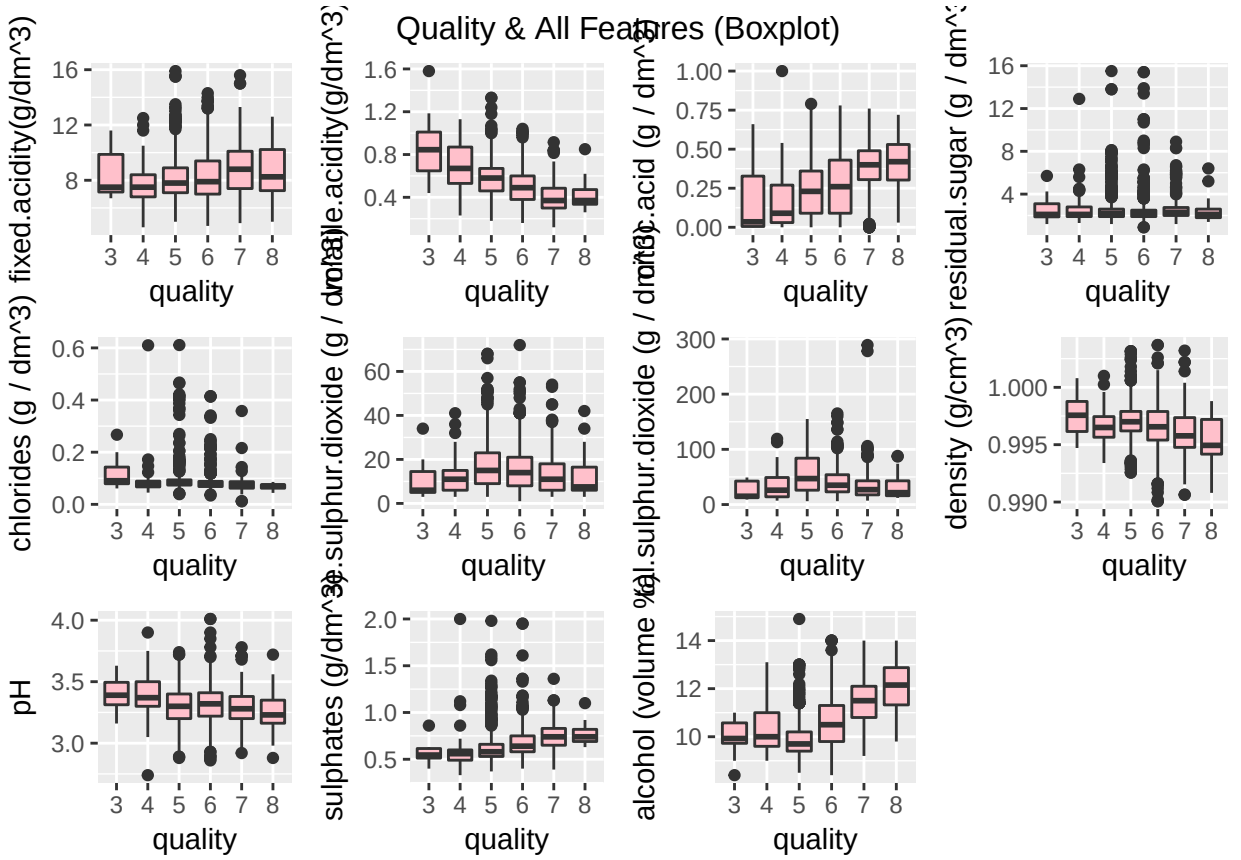
```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                 1.000          -0.256       0.672
## volatile.acidity             -0.256           1.000      -0.552
## citric.acid                   0.672          -0.552       1.000
## residual.sugar                0.115           0.002       0.144
## chlorides                     0.094           0.061       0.204
## free.sulfur.dioxide          -0.154          -0.011      -0.061
## total.sulfur.dioxide         -0.113           0.076       0.036
## density                       0.668           0.022       0.365
## pH                           -0.683           0.235      -0.542
```

```
## sulphates                    0.183           -0.261       0.313
## alcohol                     -0.062           -0.202       0.110
## quality                      0.124           -0.391       0.226
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                 0.115     0.094              -0.154
## volatile.acidity              0.002     0.061              -0.011
## citric.acid                   0.144     0.204              -0.061
## residual.sugar                1.000     0.056               0.187
## chlorides                     0.056     1.000               0.006
## free.sulfur.dioxide           0.187     0.006               1.000
## total.sulfur.dioxide          0.203     0.047               0.668
## density                       0.355     0.201              -0.022
## pH                           -0.086    -0.265               0.070
## sulphates                     0.006     0.371               0.052
## alcohol                       0.042    -0.221              -0.069
## quality                       0.014    -0.129              -0.051
##                       total.sulfur.dioxide density     pH sulphates alcohol
## fixed.acidity                       -0.113   0.668 -0.683     0.183  -0.062
## volatile.acidity                     0.076   0.022  0.235    -0.261  -0.202
## citric.acid                          0.036   0.365 -0.542     0.313   0.110
## residual.sugar                       0.203   0.355 -0.086     0.006   0.042
## chlorides                            0.047   0.201 -0.265     0.371  -0.221
## free.sulfur.dioxide                  0.668  -0.022  0.070     0.052  -0.069
## total.sulfur.dioxide                 1.000   0.071 -0.066     0.043  -0.206
## density                              0.071   1.000 -0.342     0.149  -0.496
## pH                                  -0.066  -0.342  1.000    -0.197   0.206
## sulphates                            0.043   0.149 -0.197     1.000   0.094
## alcohol                             -0.206  -0.496  0.206     0.094   1.000
## quality                             -0.185  -0.175 -0.058     0.251   0.476
##                       quality
## fixed.acidity           0.124
## volatile.acidity       -0.391
## citric.acid             0.226
## residual.sugar          0.014
## chlorides              -0.129
## free.sulfur.dioxide    -0.051
## total.sulfur.dioxide   -0.185
## density                -0.175
## pH                     -0.058
## sulphates               0.251
## alcohol                 0.476
## quality                 1.000
```



From the result, we can find out that quality has higher positive correlation with alcohol(0.47),sulphates(0.25),citric acid(0.22) and fixed acid(0.124). It also has higher negative correlation with valatile acidity(-0.39),total sulfur dioxide(-0.185) and density(-0.175).

Boxplot of Quality



Based on the plots above, we can infer good wines have the following attribures: 1.Lower fixed acidity, volatile acidity, density and ph.All these features would make the wine taset bad so 2.Higher alcohol,sulphates and citric acidity.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the ### investigation. How did the feature(s) of interest vary with other features in ### the dataset?

The top 5 features correlated with red wine quality: 1. Density (-0.496) 2. Alcohol (0.476) 3. Volatile acidity (-0.39) 4. Sulphates (0.251) 5. Citric acid (0.22) From this result we can find out, good red wine tend to have lower density and valatile acidity and higher alcohol, sulphates and citric acid.

```
##
##  Pearson's product-moment correlation
##
## data:  raw$quality and raw$density
## t = -7.0997, df = 1597, p-value = 1.875e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2220365 -0.1269870
## sample estimates:
```

```
##        cor
## -0.1749192
```

```
##
##  Pearson's product-moment correlation
##
## data:  raw$quality and raw$alcohol
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663
```

```
##
##  Pearson's product-moment correlation
##
## data:  raw$quality and raw$volatile.acidity
## t = -16.954, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4313210 -0.3482032
## sample estimates:
##        cor
## -0.3905578
```

```
##
##  Pearson's product-moment correlation
##
## data:  raw$quality and raw$sulphates
## t = 10.38, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2049011 0.2967610
## sample estimates:
##       cor
## 0.2513971
```

```
##
##  Pearson's product-moment correlation
##
## data:  raw$quality and raw$citric.acid
## t = 9.2875, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1793415 0.2723711
## sample estimates:
##       cor
## 0.2263725
```

From the Pearson's correlation test, we can find out density is less correlated with quality compared to other features.

It's also observed free sulfur dioxide and residual sugar don't have much effect on the quality of red wine.

Did you observe any interesting relationships between the other features ### (not the main feature(s) of interest)?

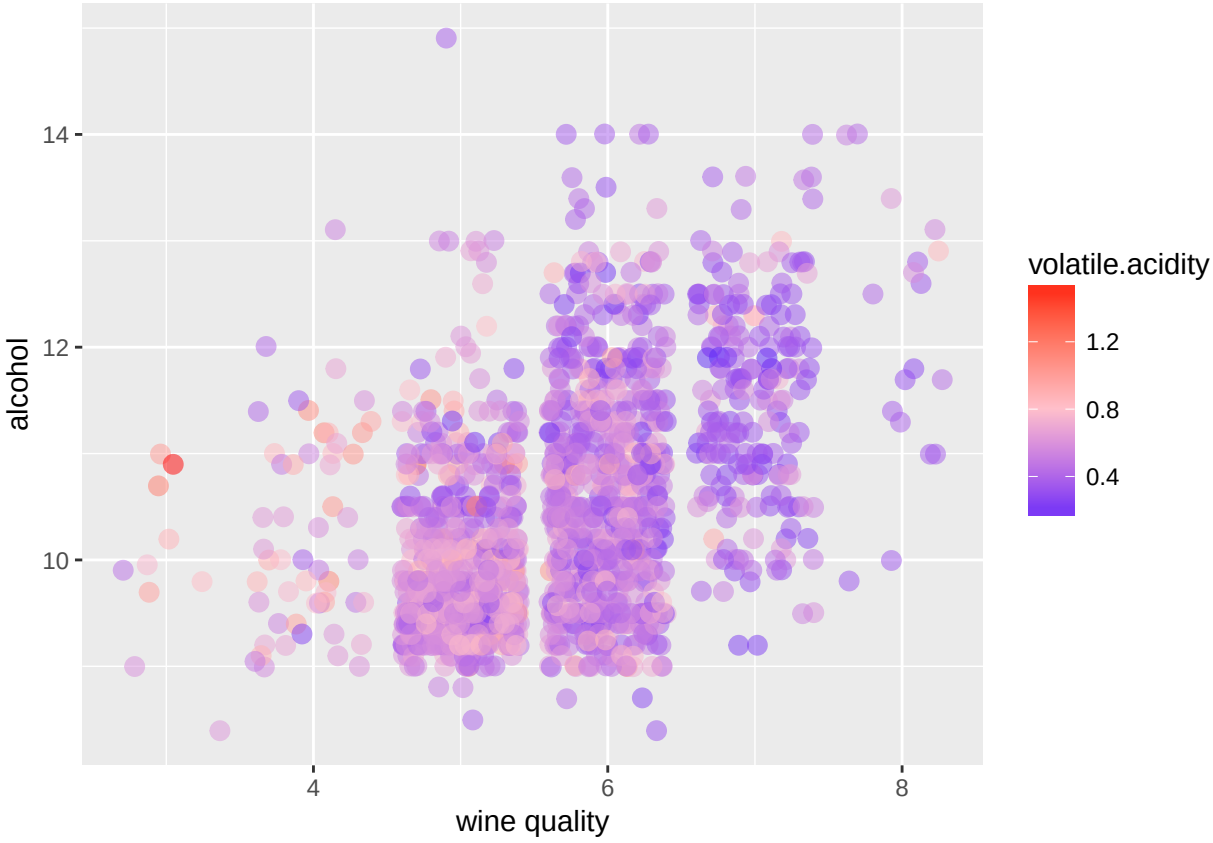
Observations about relationship between supporting variables:

1.All the acid features(fixed acidity, volatile acidity and citric acidity) are highly correlated.Because volatile Acidity provides unpleasant and vinegar taste to wines, the lower the volatile acidity, the better the wine is.While citric acid provides freshness taste to wines. Thus, better red wine would contain higher citric acid rate. In this dataset, I mainly focus on volatile acidity and citric acid because compared to fixed acidity, they seems have higher correlation with wine quality. 2.Acid features have higher correlation with ph.This makes sense because any ph number less than 7 is considered an acid. 3.Acid feature have higher correlation with sulphates and density. 4.Density has higher positive correlation with residual sugar and chlorides.

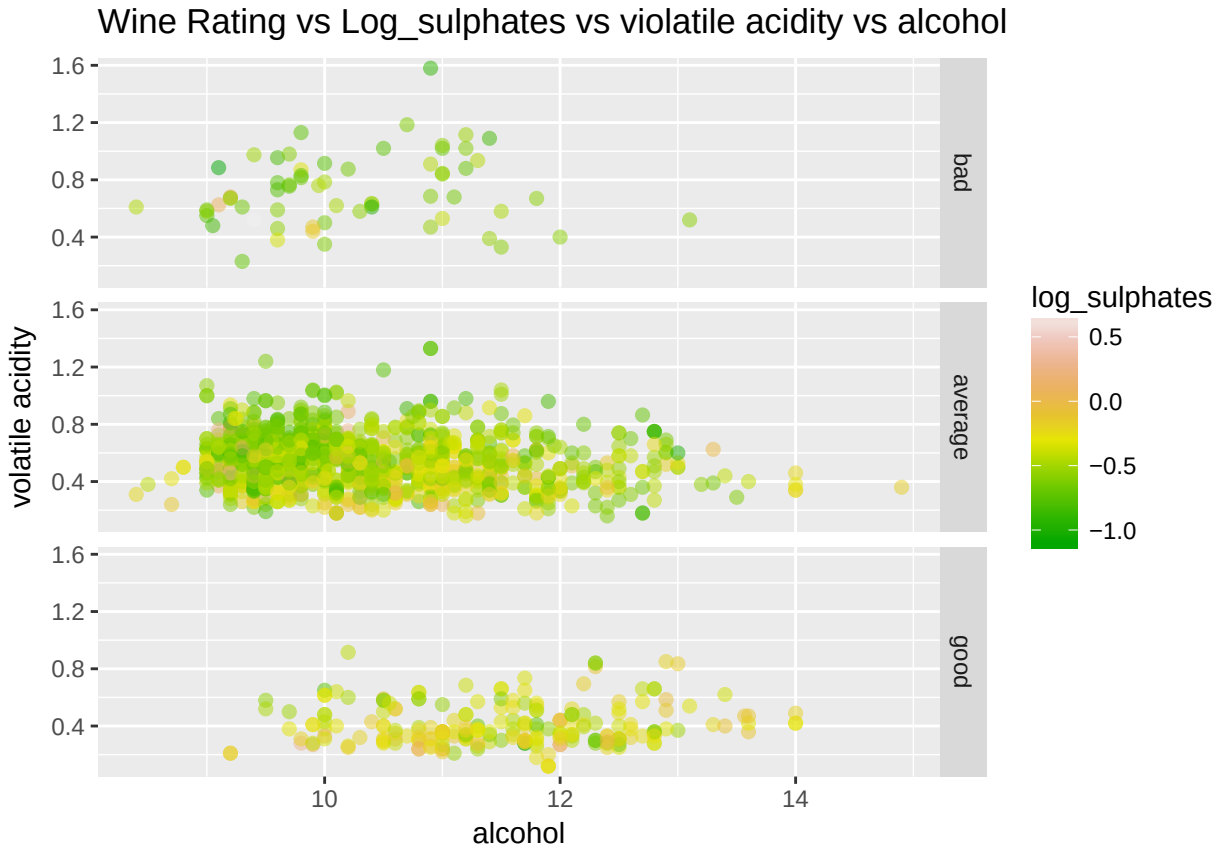
What was the strongest relationship you found?

The relationship between total sulfur dioxide and free sulfur dioxide.The correlation is 0.668.

4. Multivariate Plots Section



We can add another feature, the log scale of sulphates.In order to visualize the result, we can classify the quality into 3 categories, good, average and bad, and name this new feature as rating.



Multivariate Analysis

Talk about some of the relationships you observed in this part of the ### investigation. Were there features that strengthened each other in terms of ### looking at your feature(s) of interest?

For the first plot, we can observe higher quality wine have higher alcohol and lower volatile acidity. For the second plot, we can see higher quality wine have higher alcohol (x-axis),lower volatile acidity (y-axis) and higher sulphates (hue).

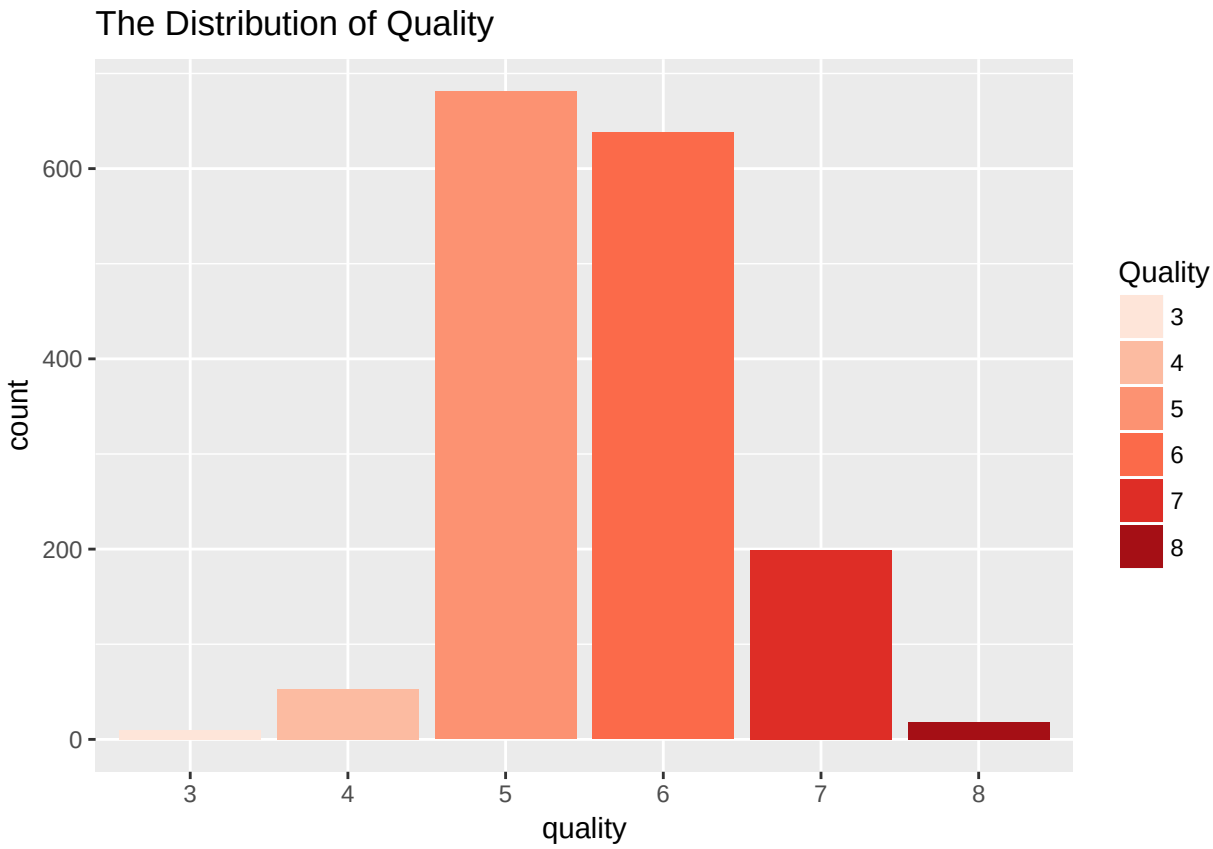
Were there any interesting or surprising interactions between features?

It's interesting to find out that in second plot if volatile acidity is lower, sulphates will be higher.

---

5. Final Plots and Summary

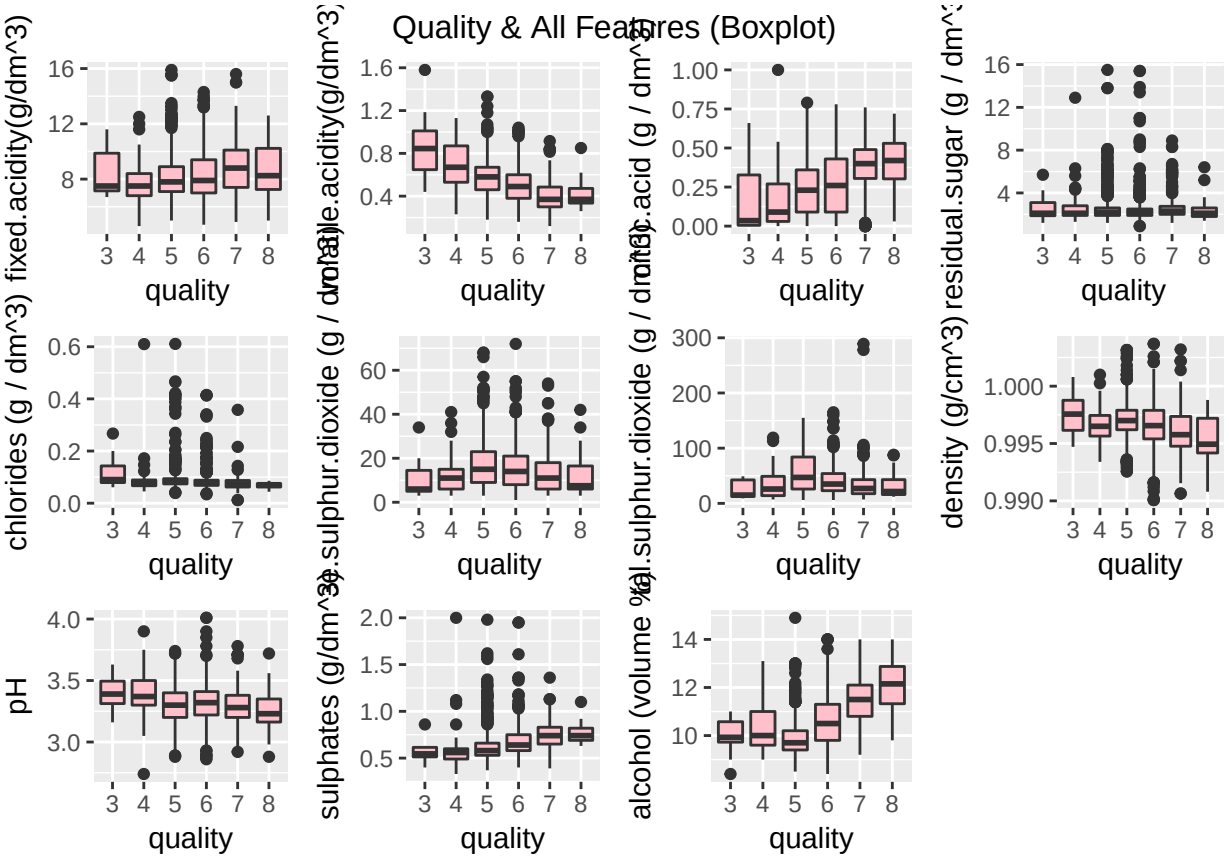
Plot One: The histogram of "Quality"



Description One

The wine quality basically has a normal distribution. Most of the quality are 5 and 6. Followed by quality 7 and quality 4. Quality 8 followed by quality 3 are the least available.

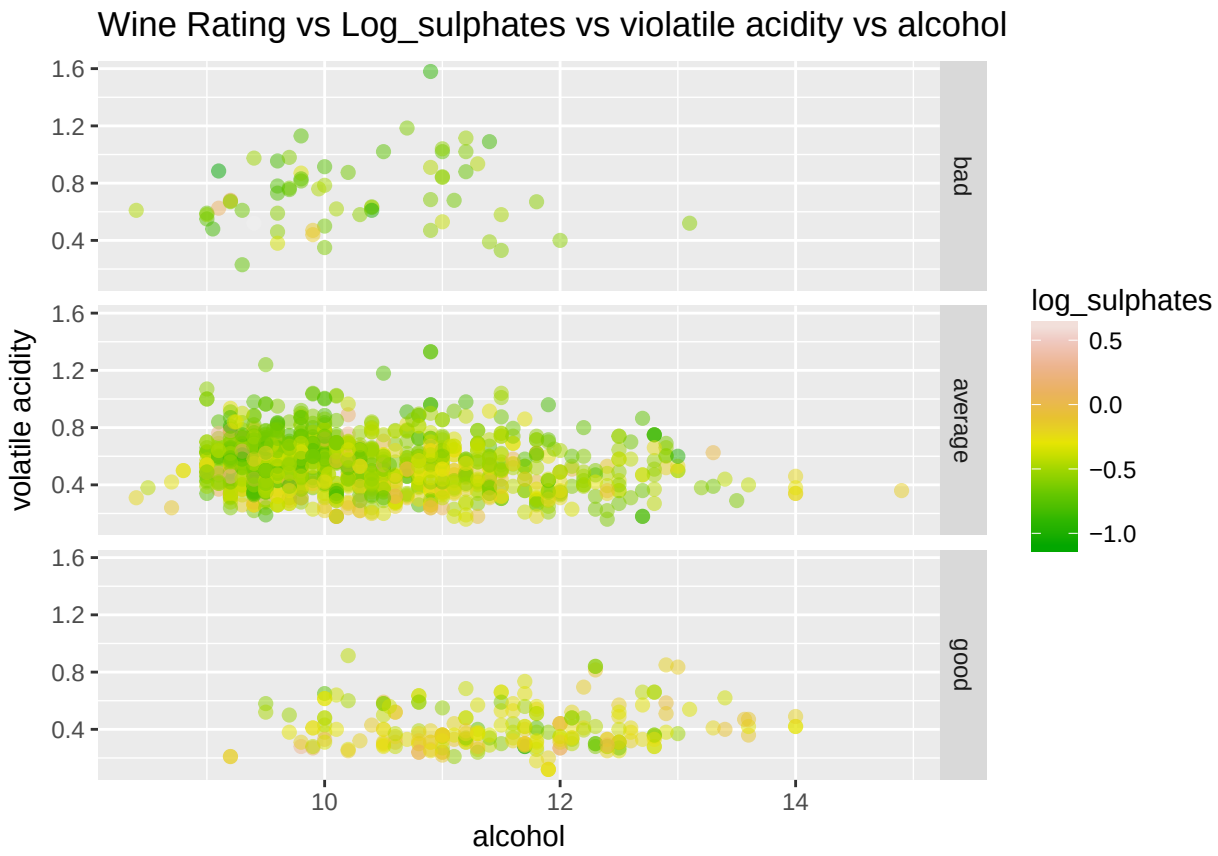
Plot Two



Description Two

Based on the plots above, we can infer good wines have the following attribures: 1.Lower fixed acidity, volatile acidity, density and ph.All these features would make the wine taset bad so 2.Higher alcohol,sulphates and citric acidity.

Plot Three



Description Three

Higher quality wine have higher alcohol (x-axis),lower volatile acidity (y-axis) and higher sulphates (hue).

---

6. Reflection

From this project, I find out that higher quality red wine have higher alcohol ,lower volatile acidity and higher sulphates.I was surprising to found out that some chemical parameters such as residual sugar, chlorides, free sulfur dioxide and total sulfur dioxide did not have much impact on the quality of red wine.

For the third plot,at first, I didn't use log transformation to deal with the sulphates data. The result looked so confusing because it's basically all green dots in the plot.After I use the log transformation, the trend looks more obvious.

The analysis could have been performed better if there was more data on other qualities of red wines available. The results would be more accurate if more data was availabe. It's also possible the results would be totally different with more data.

For future explorations I can combine the red wine dataset with the dataset for white wine to compare analyze the differences.

References:

https://rstudio-pubs-static.s3.amazonaws.com/240657_5157ff98e8204c358b2118fa69162e18.html

http://genomicsclass.github.io/book/pages/dplyr_tutorial.html

https://github.com/Dalaska/Udacity-Red-Wine-Quality/blob/master/redwine_final.rmd

https://github.com/SThornewillvE/Udacity-Project---Exploring-Wine-Data/blob/master/investigating-a-dataset.Rmd

https://github.com/anilsai/Explorartory-data-Analytics_-R-_Udacity-_Wine-Data/blob/master/Wine_quality_Anil%20Bodepudi_new.Rmd

https://github.com/jeswingeorge/EDA-project---wineQualityReds/blob/master/projecttemplate.rmd