Red Wine Analysis by Fahad Khan

```
##  [1] "DataWranglingWithR.pdf"
##  [2] "demystifying.R"
##  [3] "demystifyingR2.Rmd"
##  [4] "Diamonds.Rmd"
##  [5] "Mitchell Soil Temperature.Rmd"
##  [6] "peakpriceHistogram.jpeg"
##  [7] "pseudo_facebook.tsv"
##  [8] "pseudofacebook.Rmd"
##  [9] "Red Wine Analysis by Fahad Khan.zip"
## [10] "reddit.csv"
## [11] "redwine.html"
## [12] "redwine.rmd"
## [13] "wineQualityReds.csv"
## [14] "yogurt.csv"
## [15] "yogurt.Rmd"
```

```
## [1] "C:/Users/Fahad/OneDrive/Desktop/Udacity - Data Analyst Nanodegree/P6 - EDA"
```

```
##  [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"
```

Red Wine Dataset

This tidy data set contains 1,599 red wines with 11 variables on the chemical properties of the wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent). I am trying to assess which chemical properties influence the quality of red wines.

Structure and summary of data

```
## 'data.frame':    1599 obs. of  13 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 ...
```

```
##        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
```

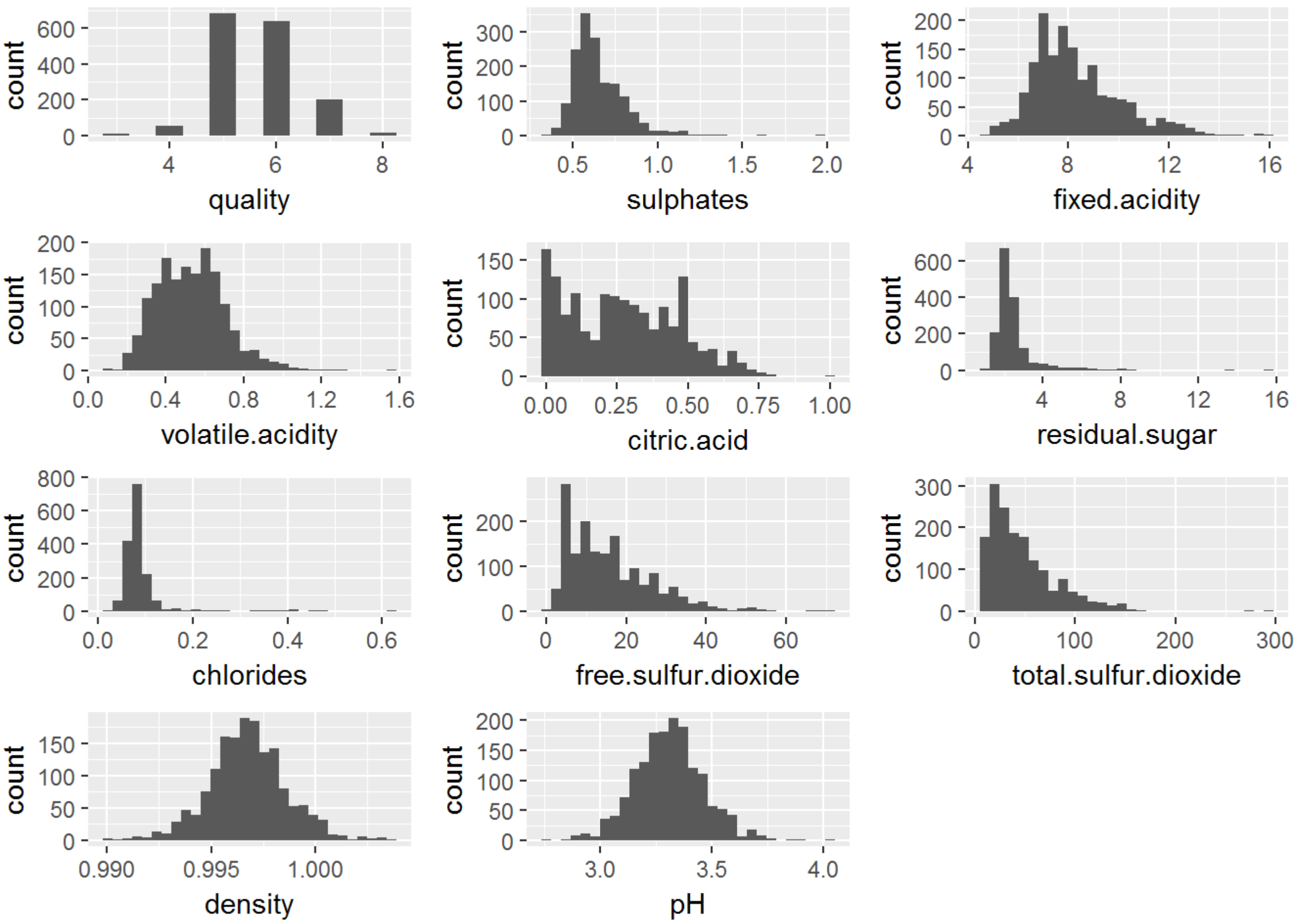
```
## 
##          3          4          5          6          7          8 
##  0.6253909  3.3145716 42.5891182 39.8999375 12.4452783  1.1257036
```

The outcome variable (quality) is an integer with a min value of 3 and max value of 8. The other attributes are of the type numeric. Among these, only the free and total sulfur-dioxide seem to have values which are whole numbers, while the remaining are decimal

Around 95% of the wines are rated between 5-7

Univariate Plots Section

Histogram for each variable



Quality - Most wines are rated 5-6

Sulphates - Distribution seems normal in IQR wih peak at around 0.6

Fixed Acidity - Distribution seems bimodal in IQR with peaks at around 7 and 8

Volatile Acidity - Distribution seems bimodal in IQR with peaks at around 0.4 and 0.6

Citric Acid - Distributions seems slightly positively skewed. A lot of wines have almost no citric acid. There seem to be two peaks at 0 & 0.50

Residual sugar and chlorides - The distribution looks very similar. Residual sugar peak is around 2 and chlorides peak is around 0.08

Free suphur dioxide - Distributions seems slightly positively skewed with multiple peaks. The tallest peak is around 5

Total suphur dioxide - Distributions seems slightly positively skewed with the peak at around 20

Density - Distribution seems normal in IQR wih peak at around 0.997

pH - Distribution seems normal in IQR wih peak at around 3.35

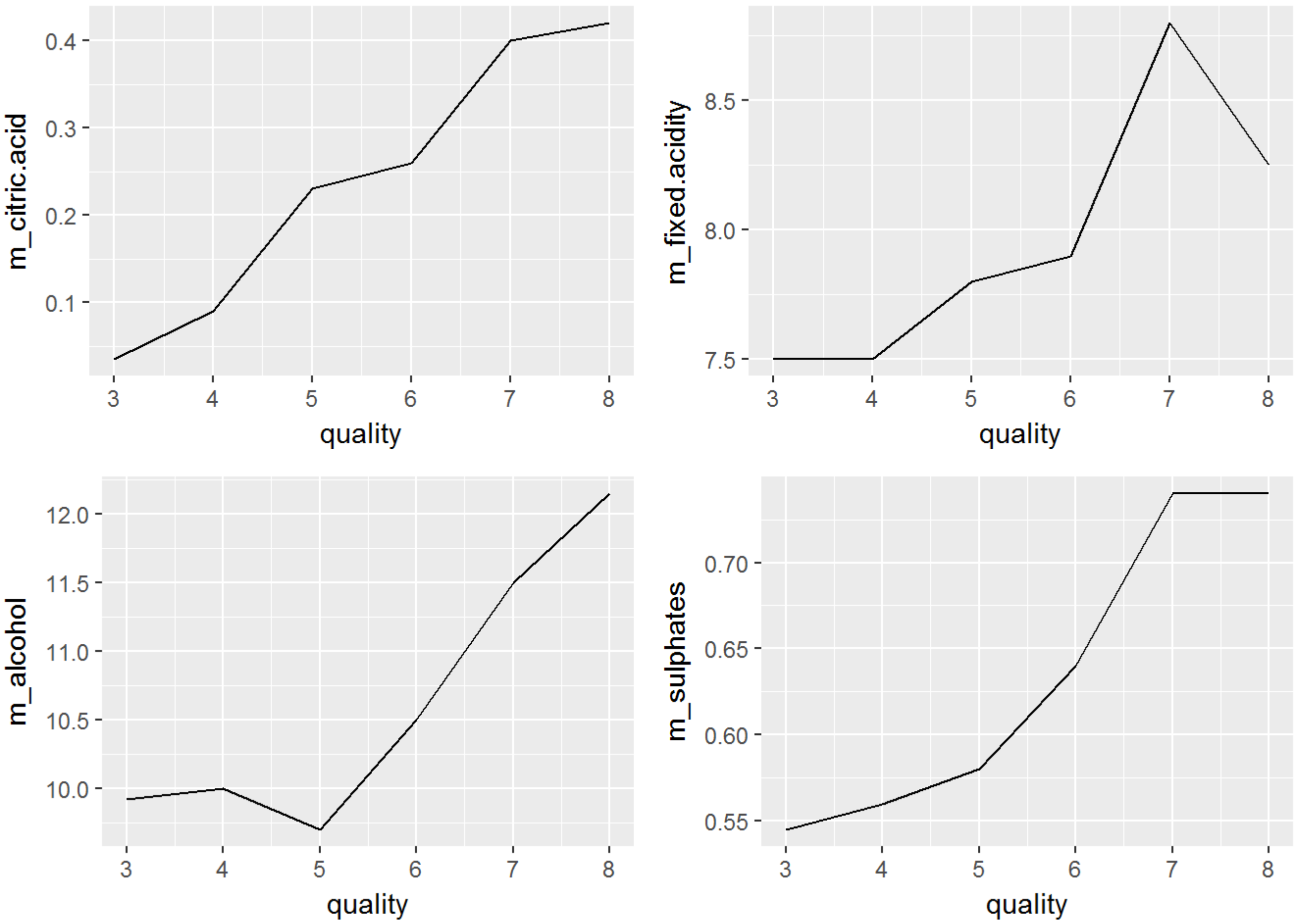
> Bivariate Plots Section

> Since we are primarily concerned about the affect of variables on quality, let us find out how does quality change with the change in attribute values. To do that we will group the data by quality and plot median attribute values against quality for all variables. This is a primary analysis, before we dig deeper into the relevant relationships

Median chemical attributes by quality

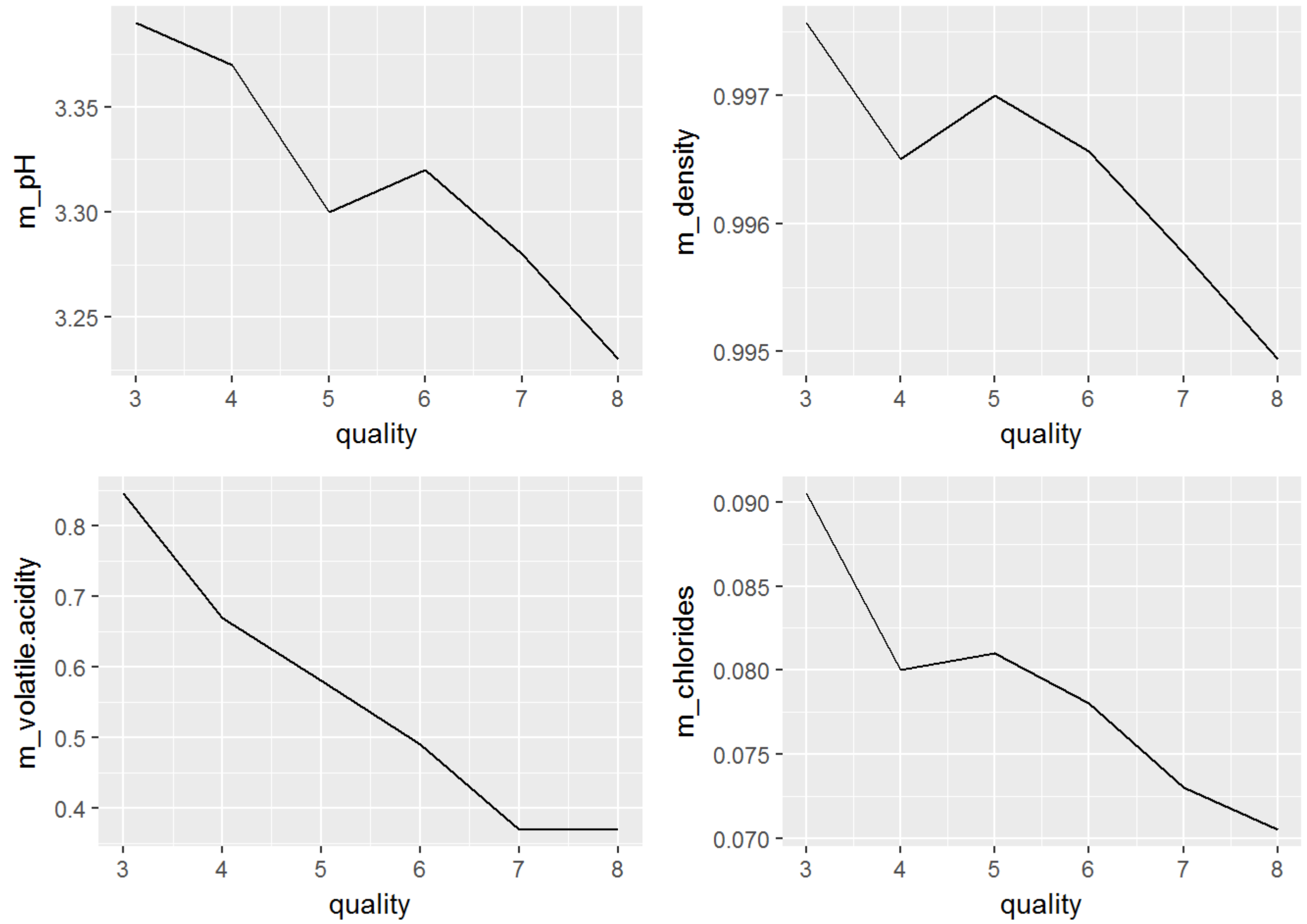
```
## # A tibble: 6 x 13
##   quality m_fixed.acidity m_sulphates m_residual.sugar m_chlorides
##     <int>           <dbl>       <dbl>            <dbl>       <dbl>
## 1       3            7.50       0.545              2.1      0.0905
## 2       4            7.50       0.560              2.1      0.0800
## 3       5            7.80       0.580              2.2      0.0810
## 4       6            7.90       0.640              2.2      0.0780
## 5       7            8.80       0.740              2.3      0.0730
## 6       8            8.25       0.740              2.1      0.0705
## # ... with 8 more variables: m_free.sulfur.dioxide <dbl>,
## #   m_total.sulfur.dioxide <dbl>, m_density <dbl>, m_pH <dbl>,
## #   m_alcohol <dbl>, m_volatile.acidity <dbl>, m_citric.acid <dbl>,
## #   n <int>
```

Median attributes by quality - Positive Relationship



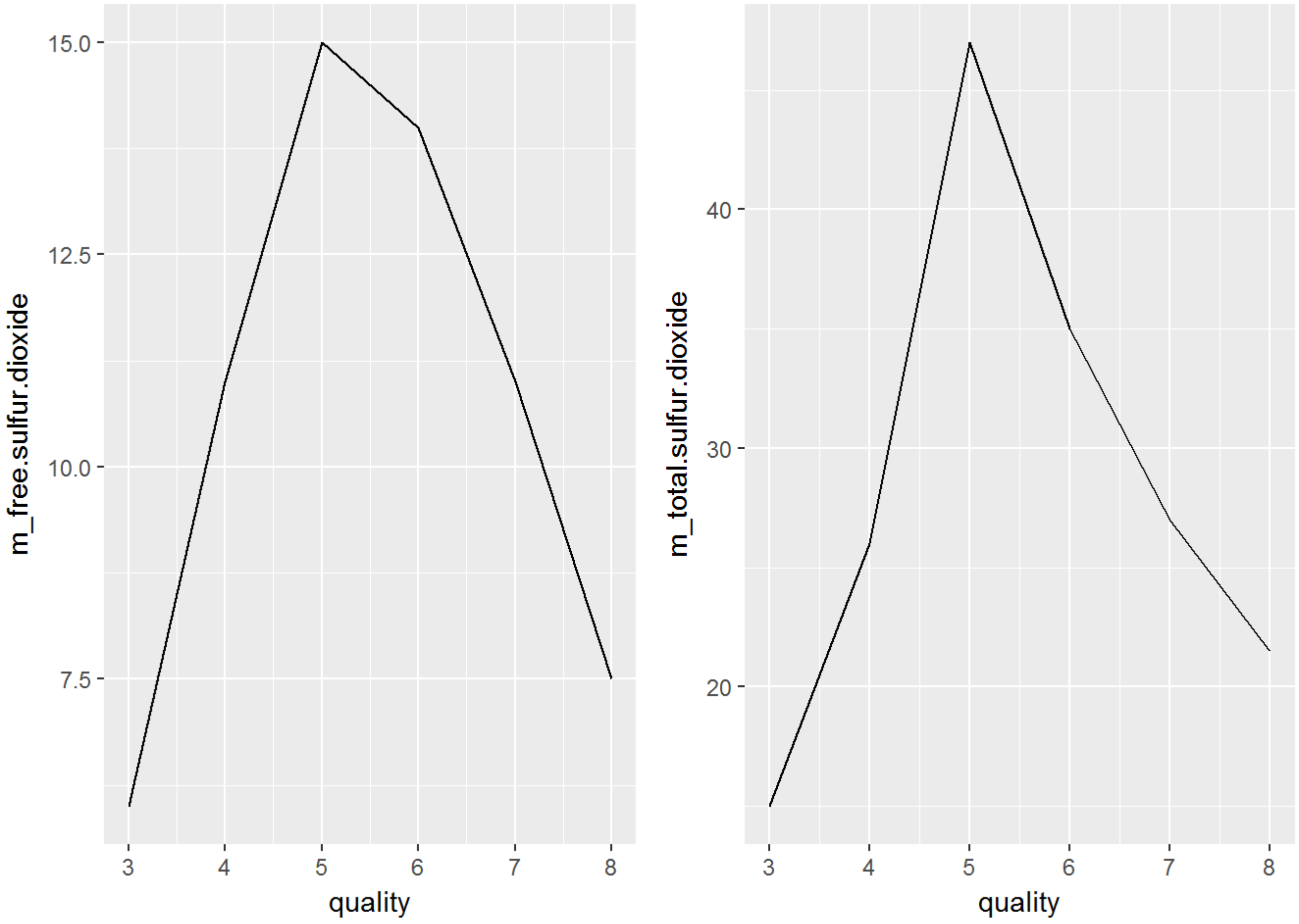
Overall, median citric acid, median fixed acidity, median alcohol percentage and median sulphates increase with the increase in the quality of the red wine

Median attributes by quality - Negative Relationship



Median pH, median volatile acidity, median density and median chlorides decrease with the increase in the quality of the red wine

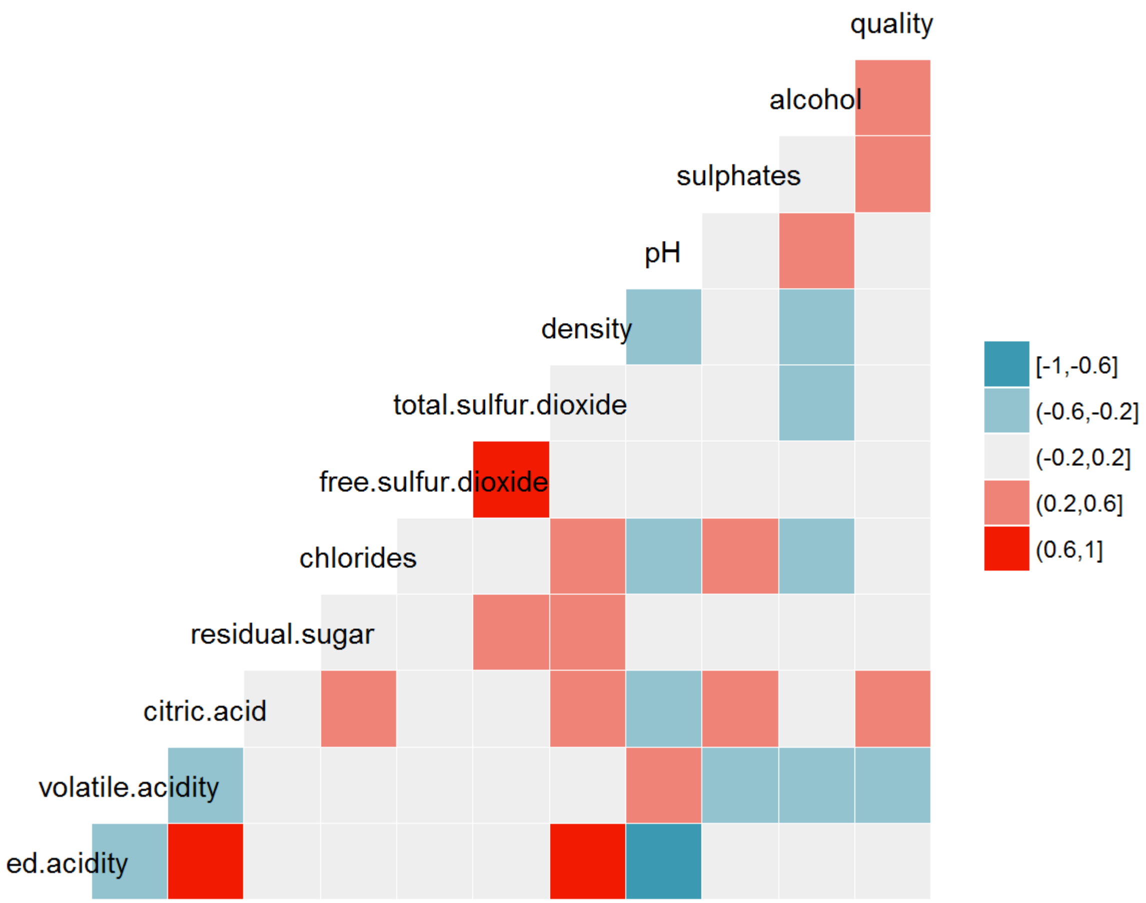
Median attributes by quality - Sulfur Dioxide



The free and total sulfur dioxide follow a similar pattern, with the peak at average wine quality (5)

> Now let us narrow down the variables with strong correlation, so that they can be studied in more detail

Relationship between all pairs of variables



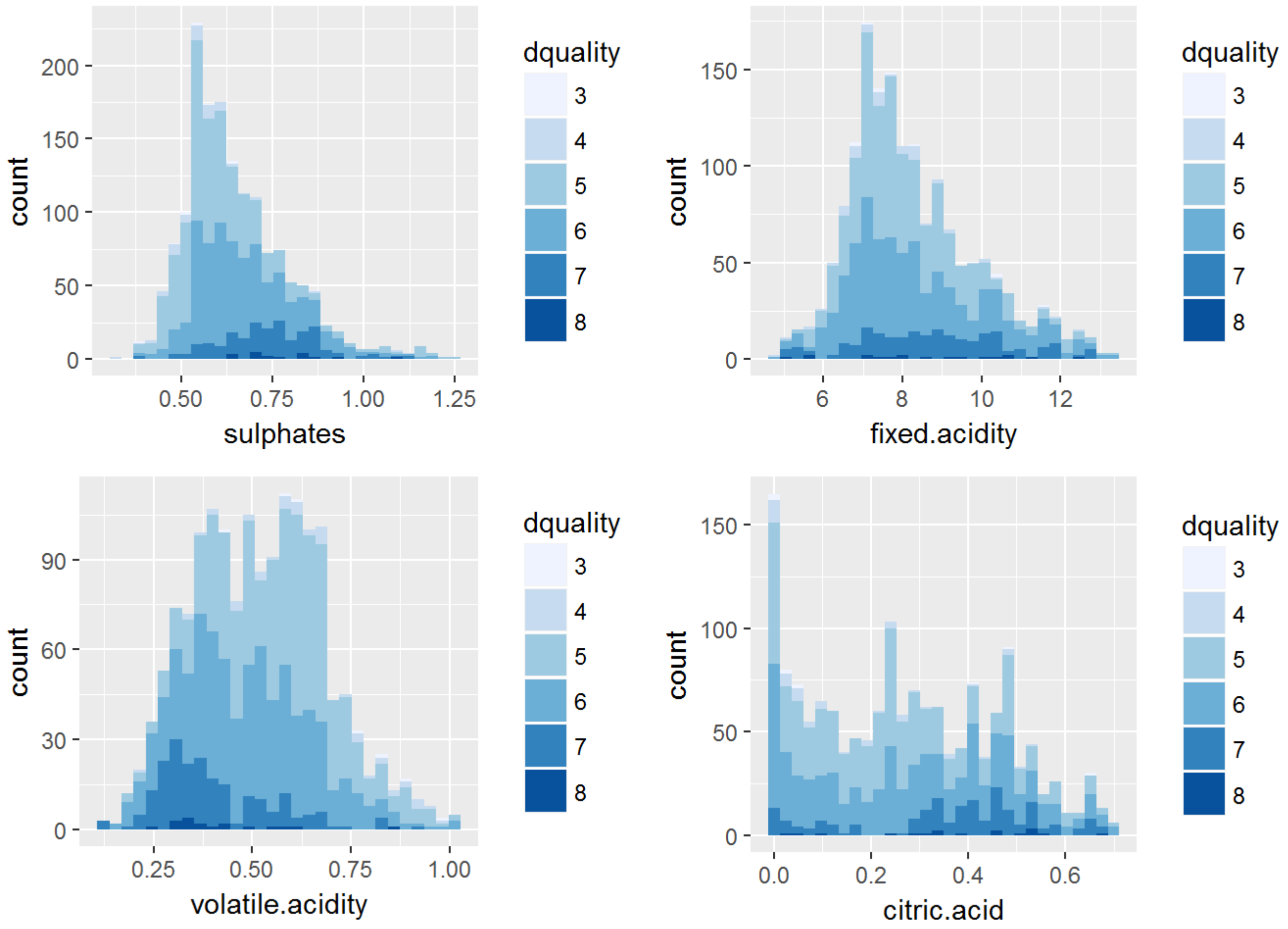
Strong correlation between: a) Citric acid and Fixed acidity b) Free and Total Sulfur dioxide c) Density and Fixed acidity

Moderate correlation between quality a) Alcohol b) Sulphates c) Citric Acid

Strong negative correlation between pH and fixed acidity

> Let us look at these relationships more closely

Histogram colored by quality



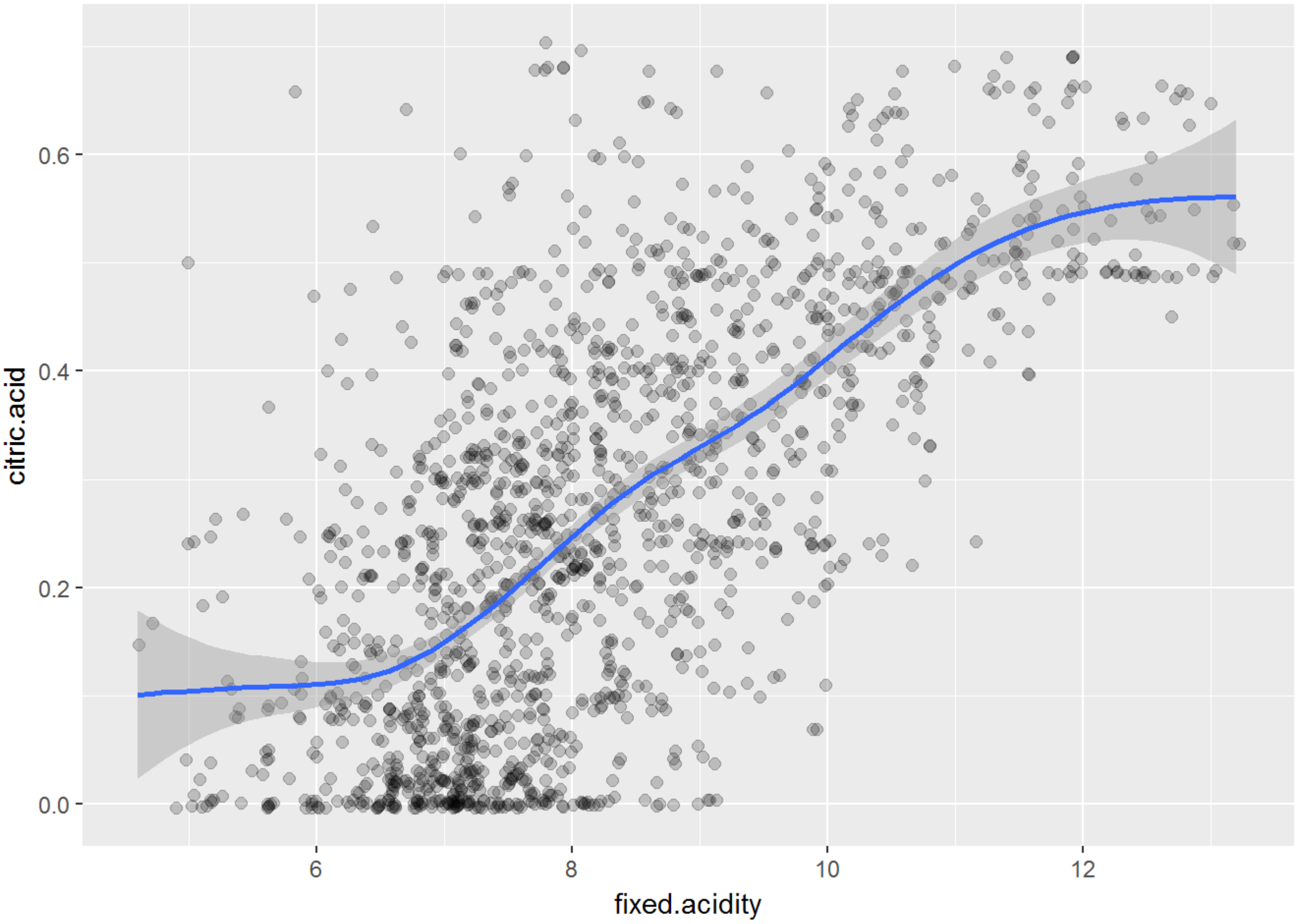
Excluded top 1% of the values to remove outlier values

Citric Acid histogram is skewed because of several wines having citric acid value as 0

In the histogram for sulphates, fixed acidity and citric acid, the higher quality wines are mostly observed towards the higher values of these variables, while the lower quality wines are mostly observed towards the lower value of these variables

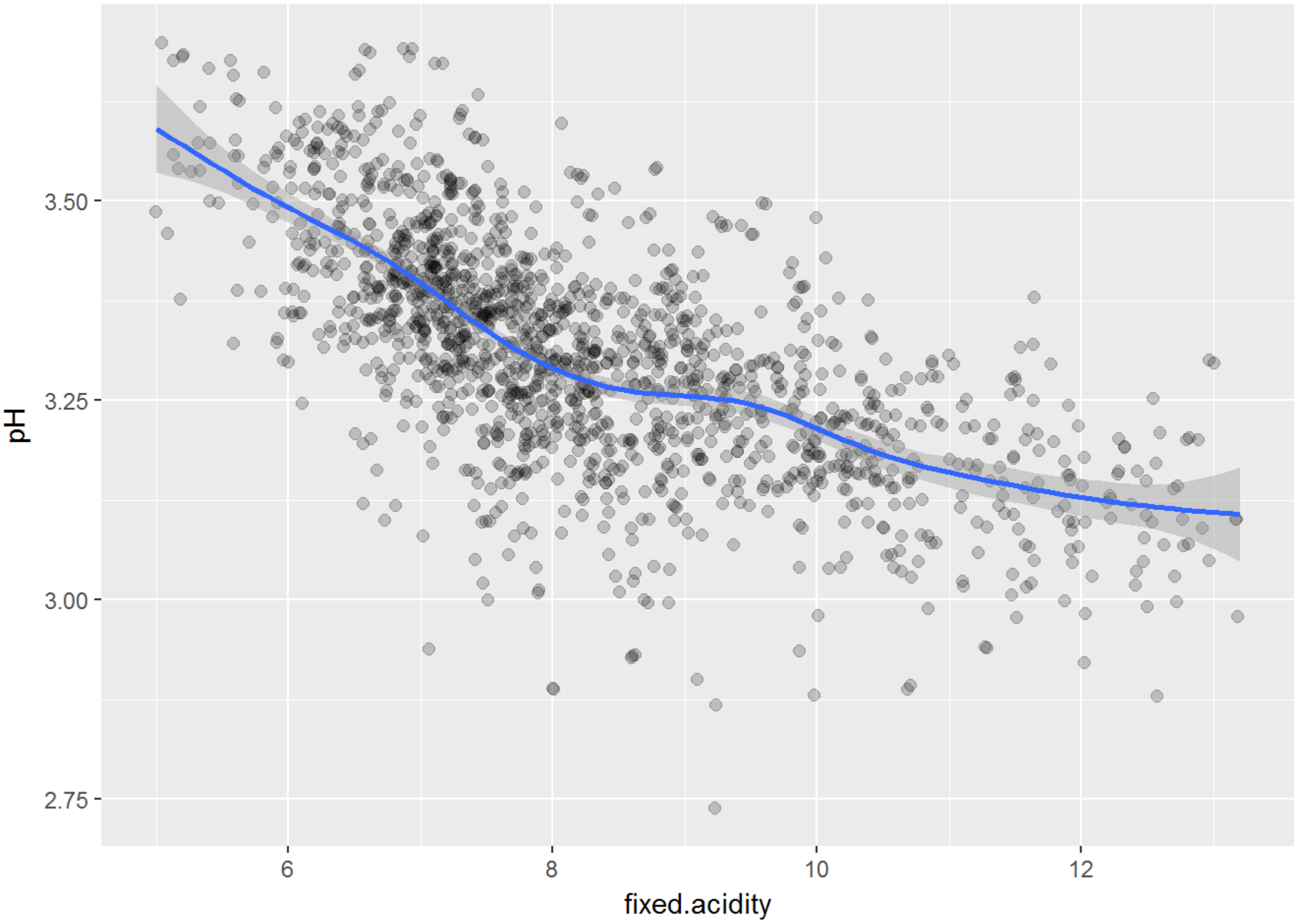
In the histogram of volatile acidity, the opposite is observed. Higher quality wines observed towards lower values of volatile acidity

Fixed Acidity vs Citric Acid



There is almost a linear relationship between fixed acidity and the citric acid in the range of 7-11 of fixed acidity

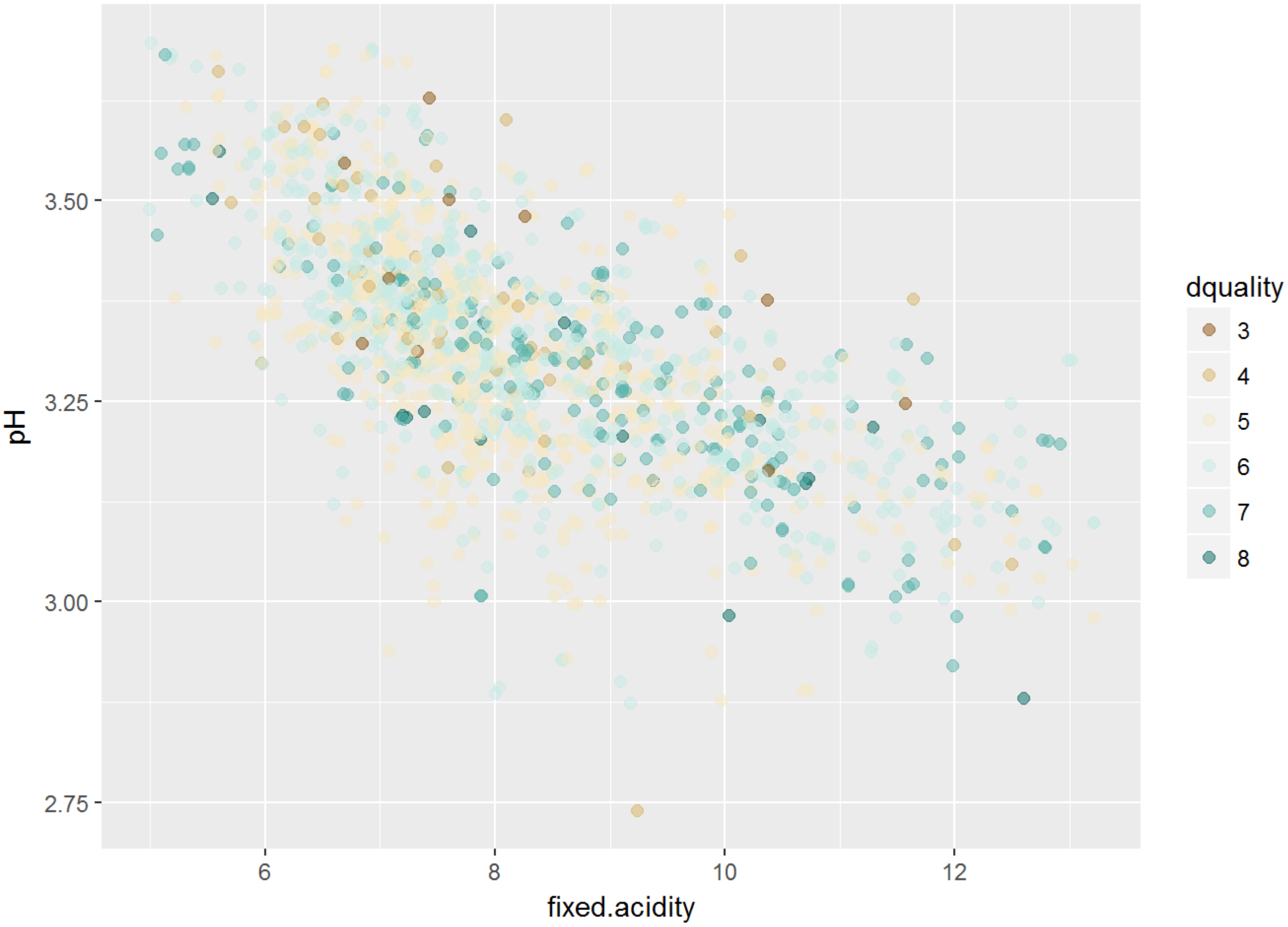
Fixed Acidity vs pH



The negative correlation between fixed acidity and pH can be clearly observed in the above plot

> Multivariate Plots Section

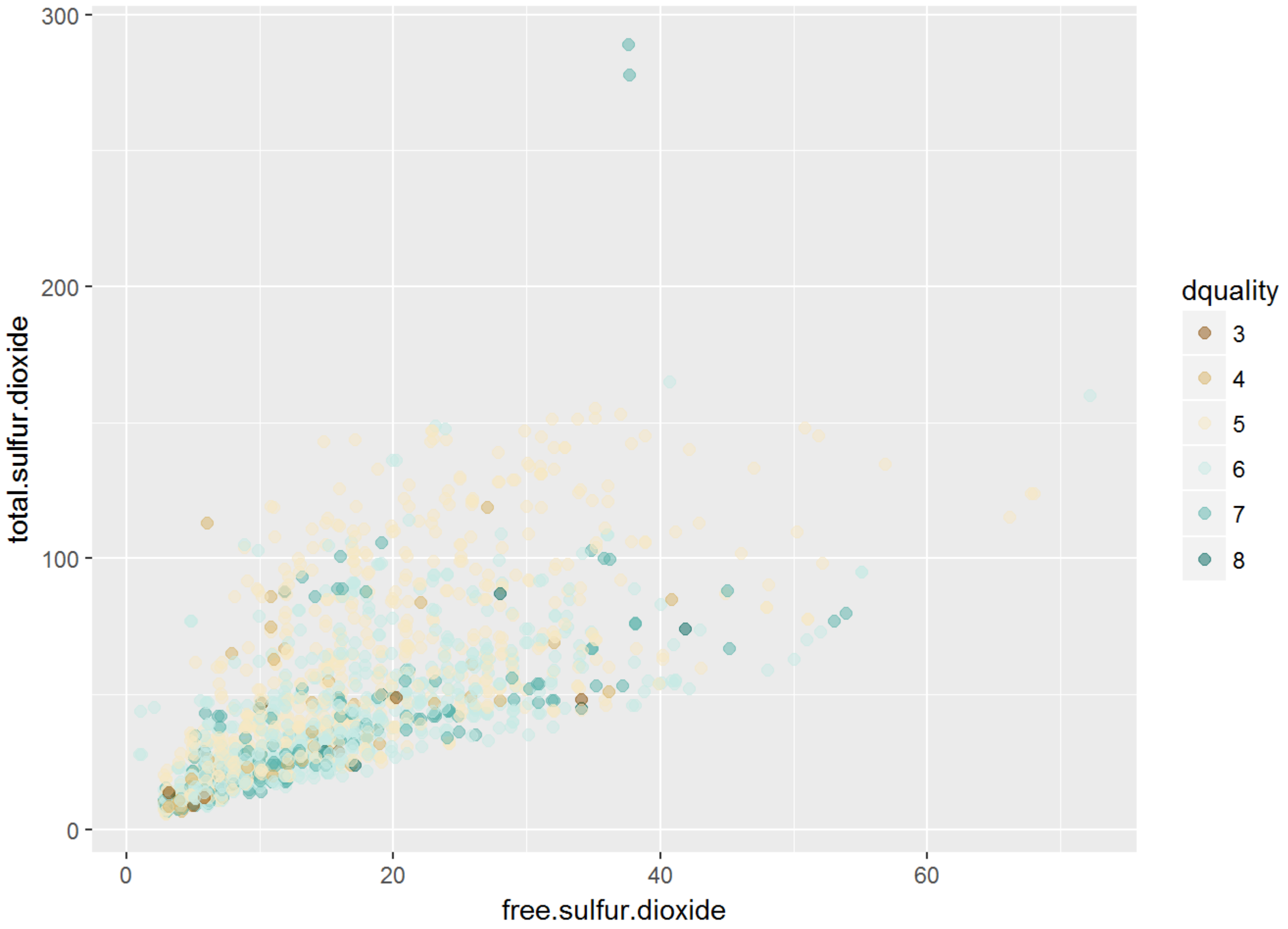
Scatter plot of fixed acidity vs pH colored by quality



pH decreases with increase in fixed acidity

Wine quality increases with fixed acidity and pH value

Scatter plot of free vs total sulfur dioxide colored by quality



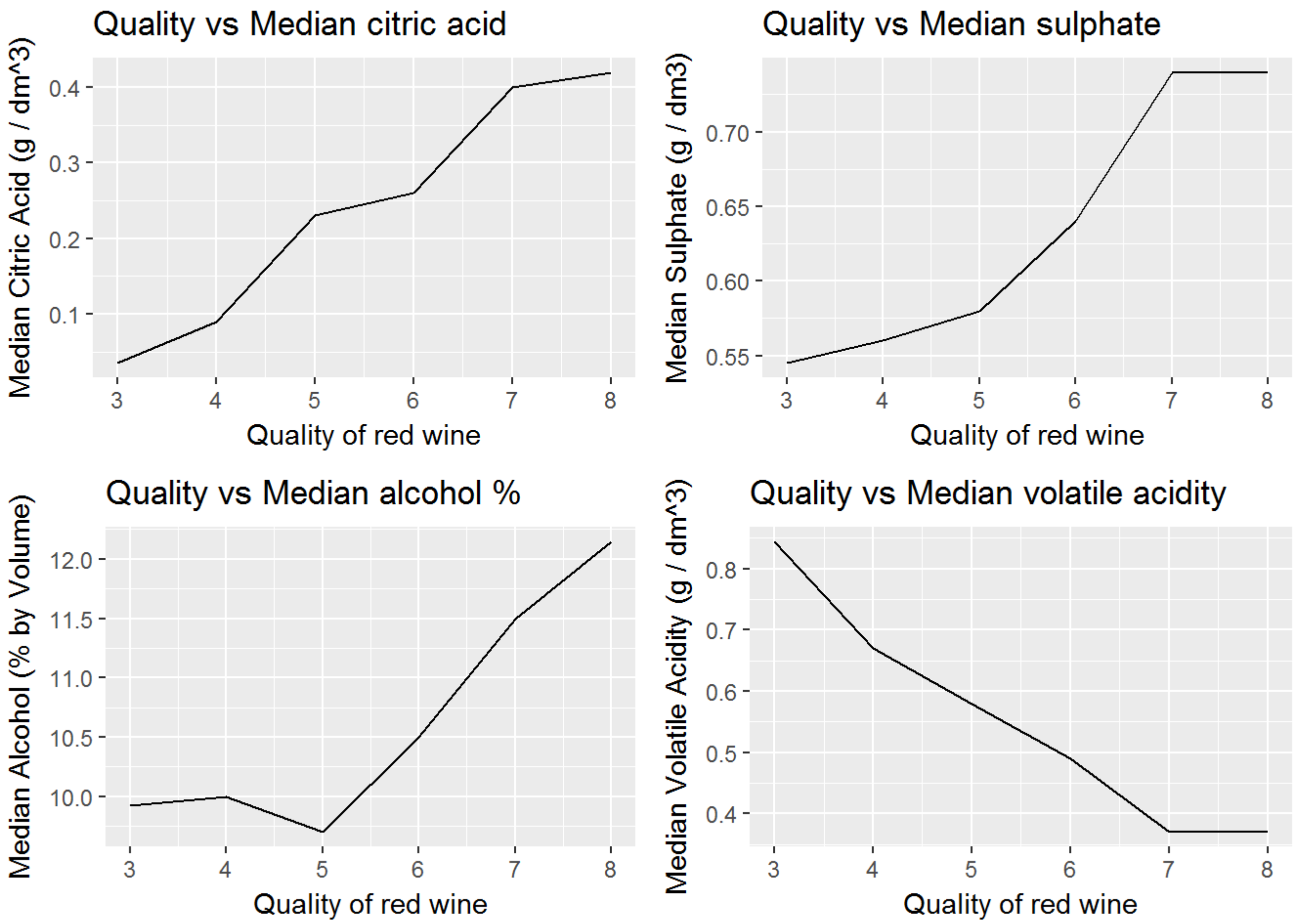
Total sulfur dioxide increases with increase in free sulfur dioxide

At any given value of free sulfur dioxide, quality seems higher at lower total sulfur dioxide

---

> Final Plots Section

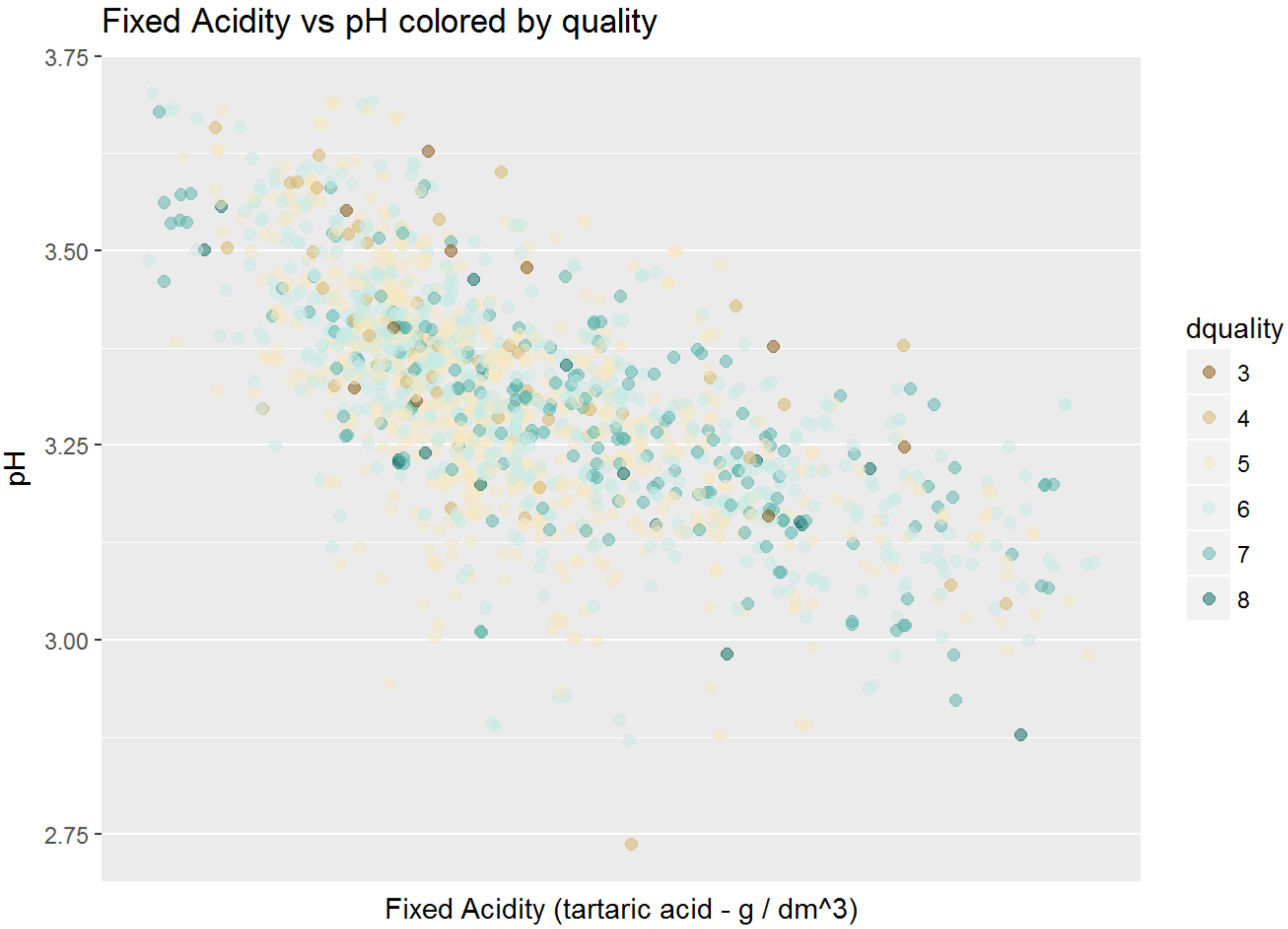
Quality vs Median Attribute Values



Out of all attributes, the above 4 attributes have the most clear affect on quality, which is why I selected them for the final plot

Quality increases with increase in median citric acid, sulphates and alcohol percentage, and it decreases with volatile acidity

Scatter plot - Fixed Acidity vs pH colored by quality

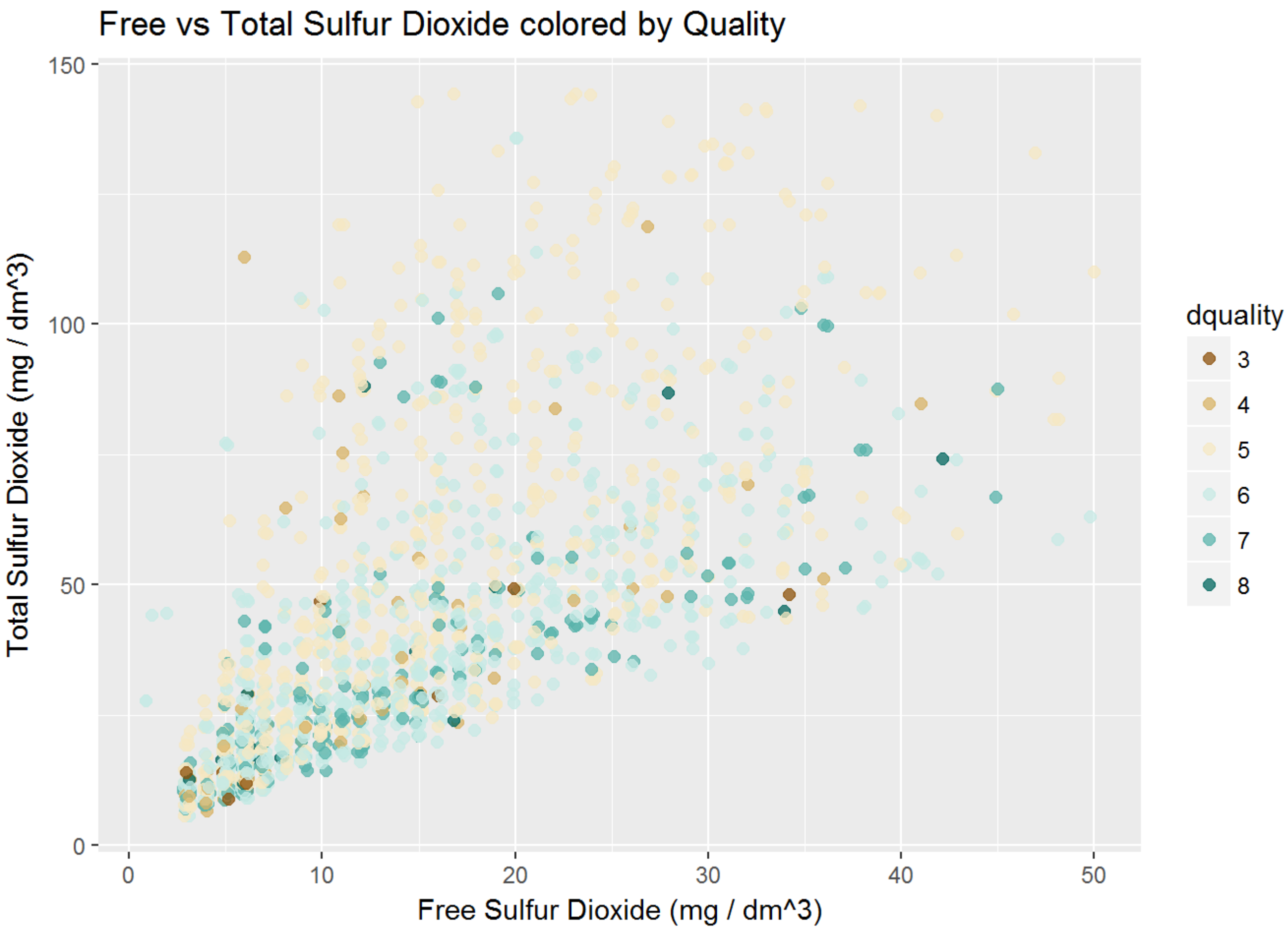


Selected this scatter plot to highlight the inverse relationship between fixed acidity and pH value

pH decreases with increase in fixed acidity

Wine quality increases with fixed acidity and pH value

Scatter plot of free vs total sulfur dioxide colored by quality



Selected this plot to see the affect of free and total sulfur dioxide on quality

Total sulfur dioxide increases with increase in free sulfur dioxide

At any given value of free sulfur dioxide, quality seems higher at lower total sulfur dioxide

Reflection

Observation - 95% of the wines are rated between 5-7 quality. So most of the observations and findings will be applicable to this quality range. We don’t have enough observations for the wines with other quality ratings

Insight - Based on the correlation coefficient and the line charts (median attributes vs quality), citric acid, alcohol and sulphates seem to have the most positive affect on quality whereas volatile acidity seems to have the most negative affect on quality

Struggle - Only the output variable, quality ahd limited values and could be converted into a categorical variable. That limited the kind of graphs and charts I could create, since all other variables are continuous

Surprising - It was surprising to observe that the median sulphates and median sulfur dioxide have a different relationship with quality. While quality increases consistently with increase in median sulphates, quality would first increase and then decrease with increase in sulfur dioxide

Future work - Having more data points for wines rated between 1-4 and wines rated greater than 7, might illuminate some other trends.