Exploratory Data Analysis on Red Wine

by Bibin Jose

Contents

Introduction

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). The dataset is related to red variant of the Portuguese wine "Vinho Verde".

- Guiding Question: Which chemical properties influence the quality of red wines?

Data - Column names, description and units

- Number of Attributes: 11 + output attribute
- Input variables (based on physicochemical tests):
  1. fixed acidity (tartaric acid - g/dm^3)
     - most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
  2. volatile acidity (acetic acid - g/dm^3)
     - the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
  3. citric acid (g/dm^3)
     - found in small quantities, citric acid can add 'freshness' and flavor to wines
  4. residual sugar (g/dm^3)
     - the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
  5. chlorides (sodium chloride - g/dm^3)
     - the amount of salt in the wine
  6. free sulfur dioxide (mg/dm^3)
     - the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
  7. total sulfur dioxide (mg/dm^3)
     - amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
  8. density (g/cm^3)
     - the density of water is close to that of water depending on the percent alcohol and sugar content
  9. pH
     - describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
  10. sulphates (potassium sulphate - g/dm^3)
     - a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
  11. alcohol (% by volume)
     - the percent alcohol content of the wine
- Output variable (based on sensory data):
  12. quality (score between 0 and 10)

Data - Structure

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ 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 ...
```

```
## [1] "fixed.acidity -> 0"
```

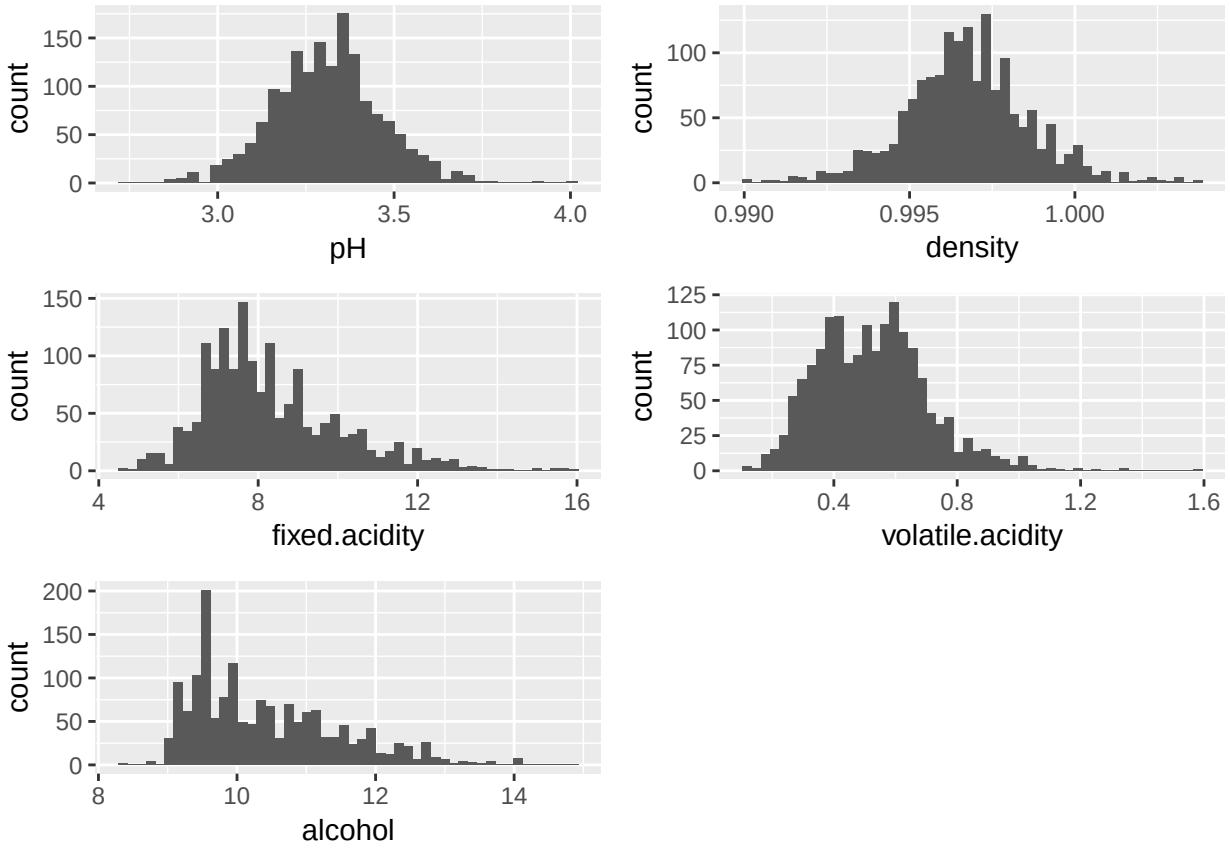
```
## [1] "volatile.acidity -> 0"
## [1] "citric.acid -> 0"
## [1] "residual.sugar -> 0"
## [1] "chlorides -> 0"
## [1] "free.sulfur.dioxide -> 0"
## [1] "total.sulfur.dioxide -> 0"
## [1] "density -> 0"
## [1] "pH -> 0"
## [1] "sulphates -> 0"
## [1] "alcohol -> 0"
## [1] "quality -> 0"
```

- No NA values were present in the dataset.

Single Variable Analysis

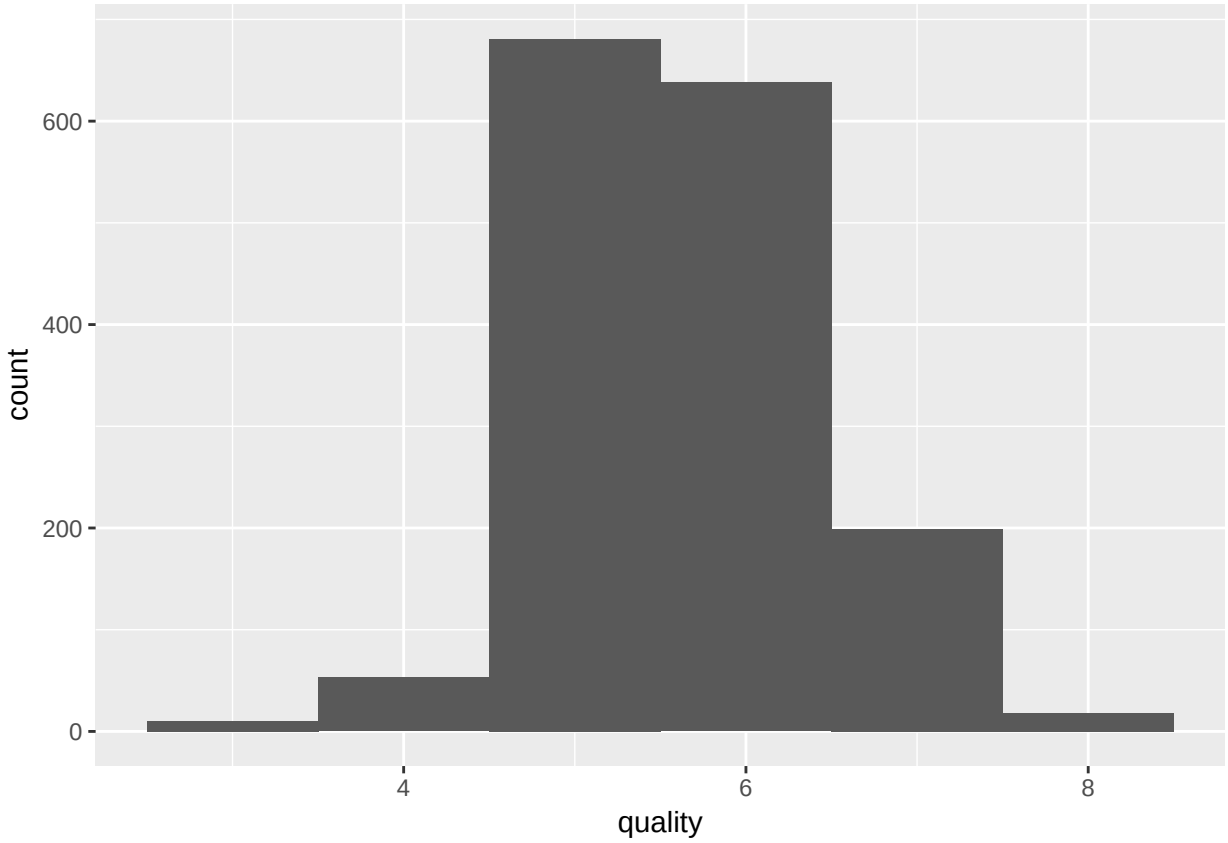
Each variable in the dataset is explored individually by plotting histograms to visualize the distribution of the data.

density, fixed & volatile acidity, alchol and quality




```
##        pH           density       fixed.acidity   volatile.acidity
##  Min.   :2.740   Min.   :0.9901   Min.   : 4.60   Min.   :0.1200
```

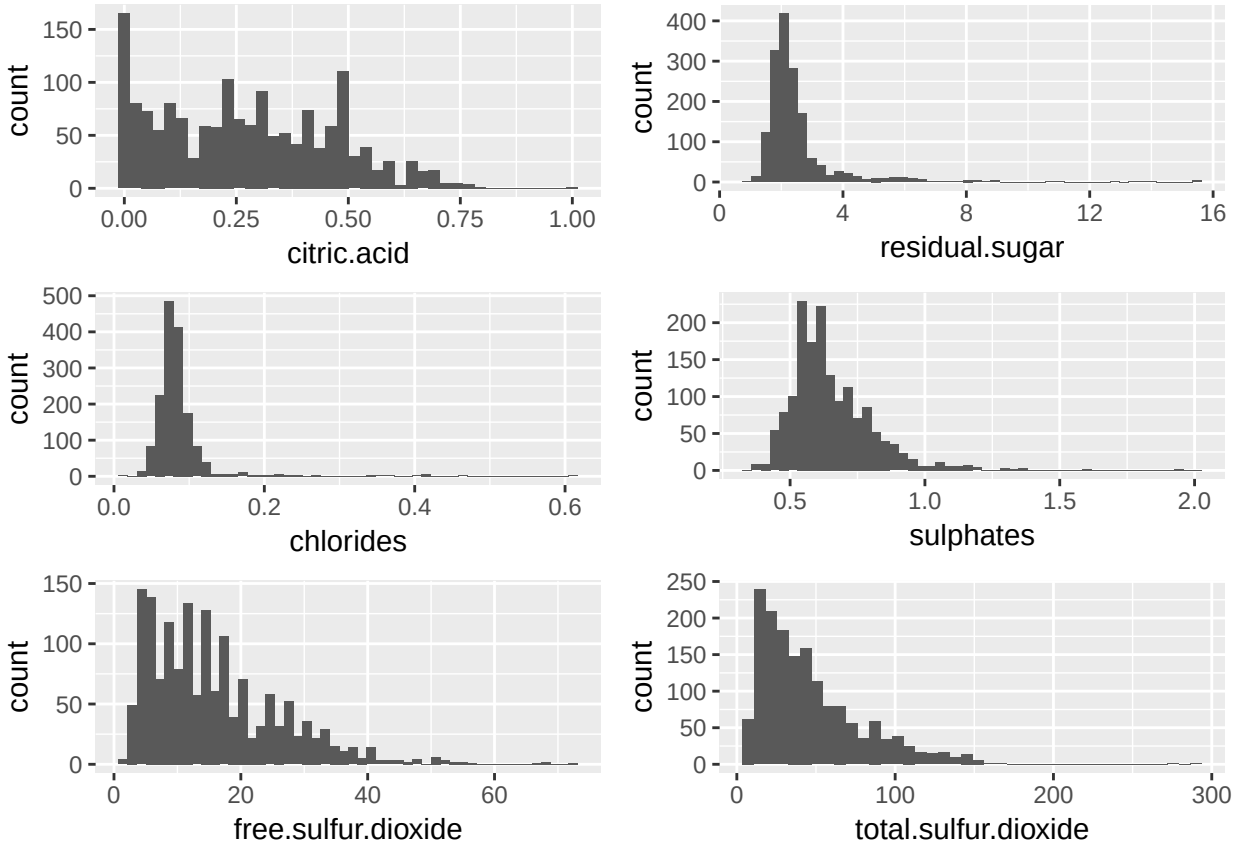
```
##  1st Qu.:3.210   1st Qu.:0.9956   1st Qu.: 7.10   1st Qu.:0.3900
##  Median :3.310   Median :0.9968   Median : 7.90   Median :0.5200
##  Mean   :3.311   Mean   :0.9967   Mean   : 8.32   Mean   :0.5278
##  3rd Qu.:3.400   3rd Qu.:0.9978   3rd Qu.: 9.20   3rd Qu.:0.6400
##  Max.   :4.010   Max.   :1.0037   Max.   :15.90   Max.   :1.5800
##     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
```




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

- Most of the wines in the dataset have a quality 5 or 6.

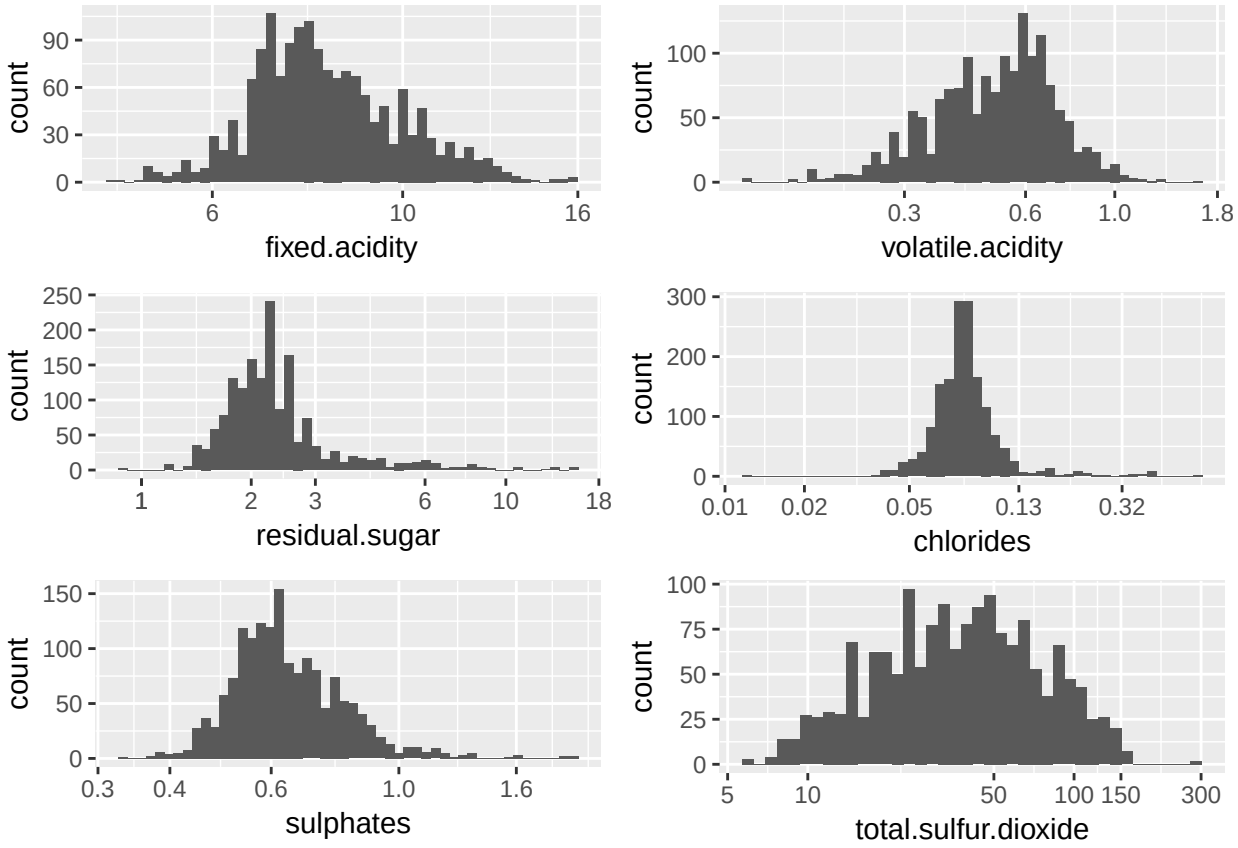
citric acid, residual sugar, chlorides, sulphates, free and total sulfur dioxide




```
##   citric.acid    residual.sugar     chlorides         sulphates
##  Min.   :0.000   Min.   : 0.900   Min.   :0.01200   Min.   :0.3300
##  1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.:0.5500
##  Median :0.260   Median : 2.200   Median :0.07900   Median :0.6200
##  Mean   :0.271   Mean   : 2.539   Mean   :0.08747   Mean   :0.6581
##  3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:0.7300
##  Max.   :1.000   Max.   :15.500   Max.   :0.61100   Max.   :2.0000
##  free.sulfur.dioxide total.sulfur.dioxide
##  Min.   : 1.00       Min.   :  6.00
##  1st Qu.: 7.00       1st Qu.: 22.00
##  Median :14.00       Median : 38.00
##  Mean   :15.87       Mean   : 46.47
##  3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :72.00       Max.   :289.00
```

- Each column in the dataset is checked for NA values and none was found. No operations were performed on the dataset to clean or tidy data.
- Cirtic.acid and free.sulfur.dioxide are skewed towards right.

Histograms with log

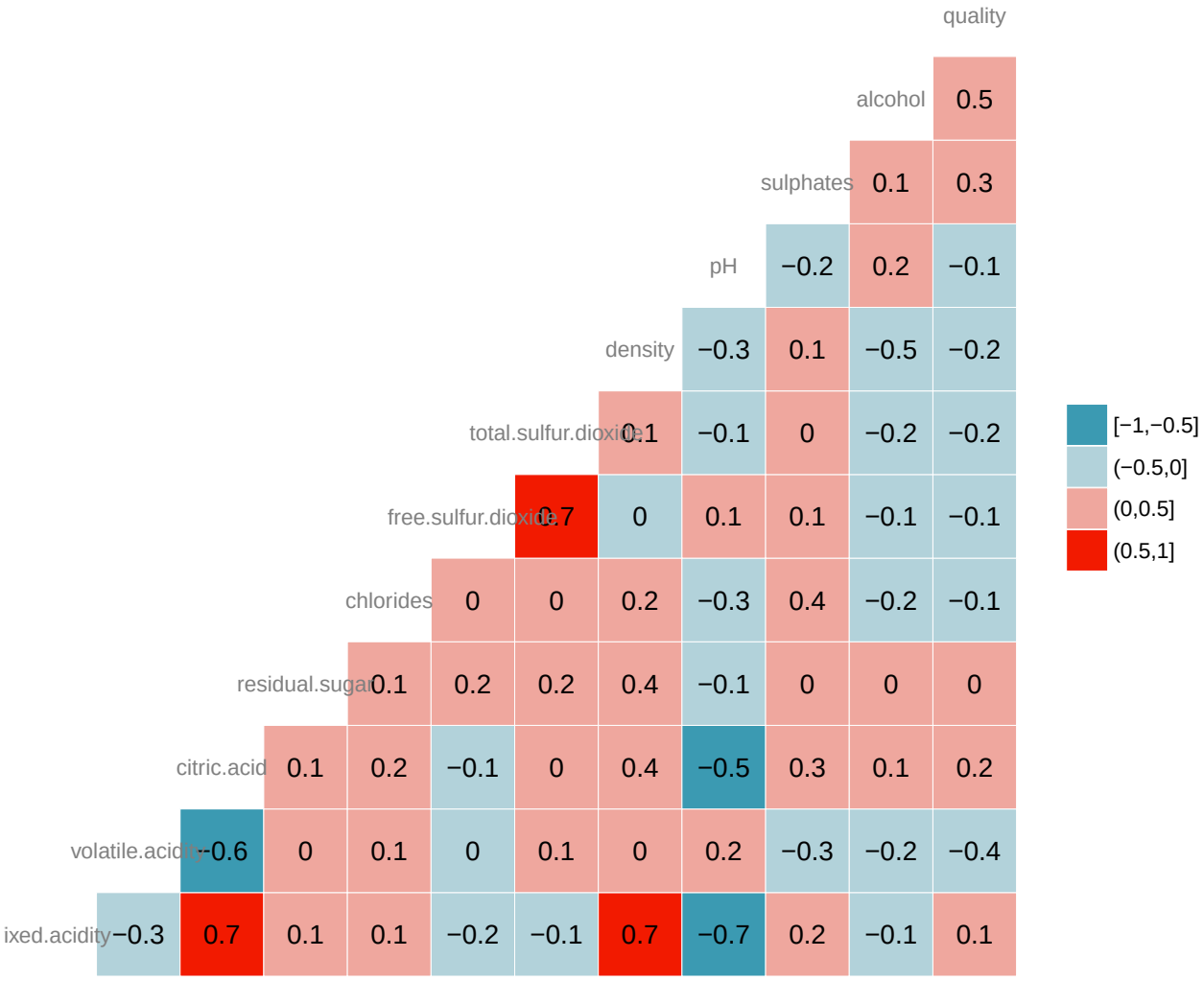


Most of the variables are normally distributed except quality, which shows 6 levels. These variables mentioned below are transformed to log scale for a normally distributed dataset:

- fixed acidity: Most data points between 6 - 10
- volatile acidity: Most data between 0.3 - 1.0
- residual sugar: Most data between 1 - 3
- chlorides: Most data in the range 0.05 - 0.13
- sulphates: Less data in the range 1-1.7
- total sulfur oxide: There are very few data in the range 150-300

Two Variable Analysis

Correlation Matrix



These sets of variables seems to have more than meaningful correlation (>0.5)

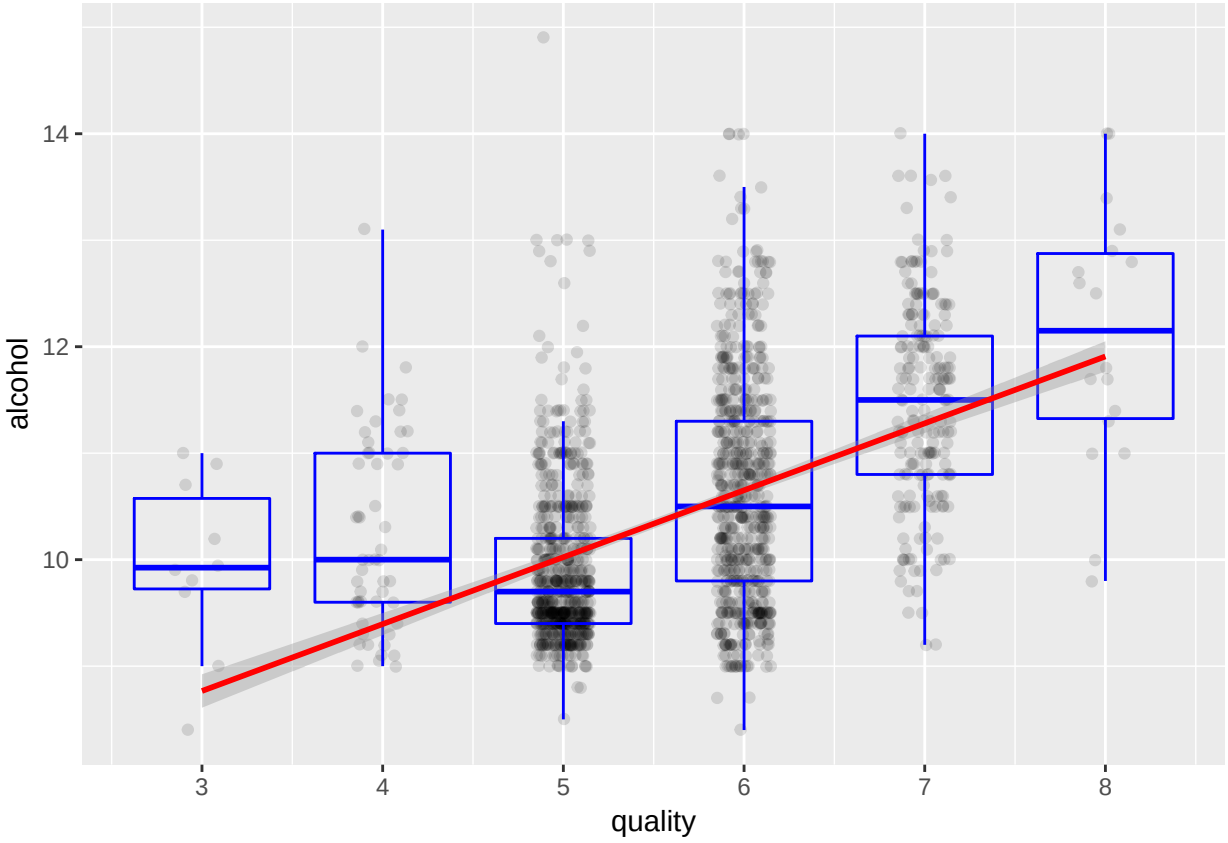
Meaningful Correlations

1. fixed.acidity vs. citric.acid (0.672)
2. fixed.acidity vs. density (-0.688)
3. fixed.acidity vs. pH (-0.683)
4. citic acid vs pH (-0.542)
5. volatile.acidity vs. citric acid (-0.552)
6. free sulfur dioxide vs total sulfur dioxide (0.667)

The strongest correlation is found to be between **pH-fixed.acidity** (-0.683) followed by **fixed acidity-density** (0.668) and **citric acid-fixed acidity** (0.672). Since pH is a measure of acidity these correlations are expected.

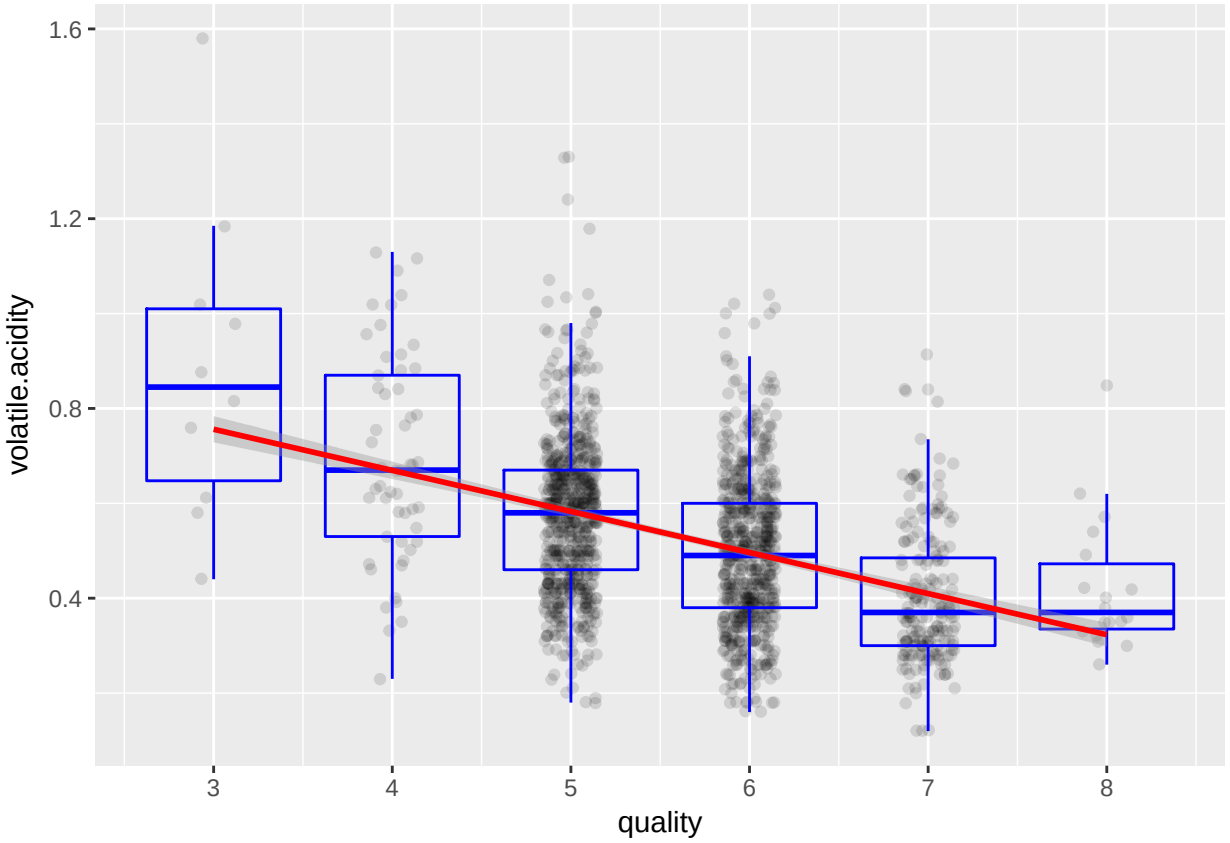
Quality vs. others

Quality vs. Alcohol



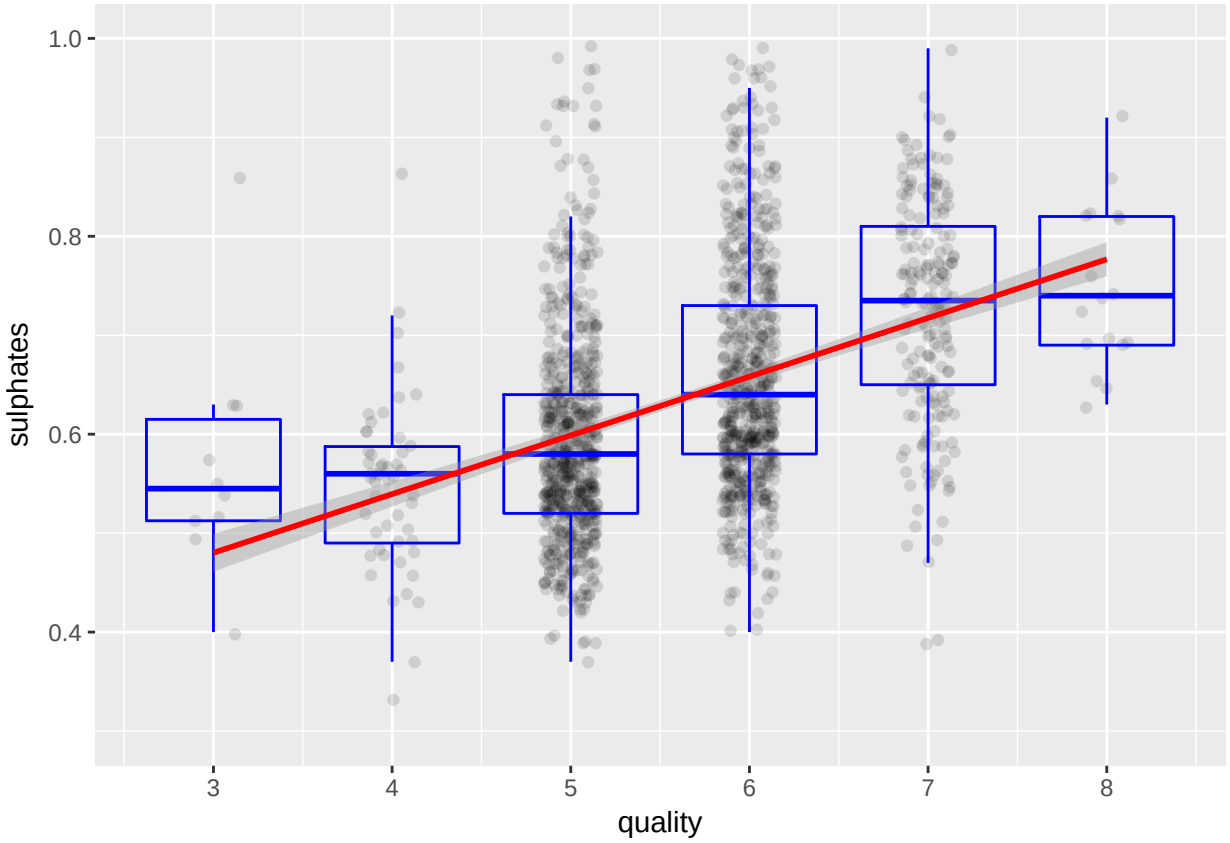
- Median values for alcohol also showed an upward trend but with a dip for intermediate quality (quality = 5).

Quality vs. Volatile Acidity



- Median values ofvolatile acidity showed a decreasing trend with quality.
- Since volatile acidity is the measure of acetic acid in wine, which at higher level increases unpleasant vinegar like taste, it can be infered that volatile acidity generally decreases quality.
- Low quality wines which are not rigourously monitored generally contains more acetic acid, hence higher volatility.

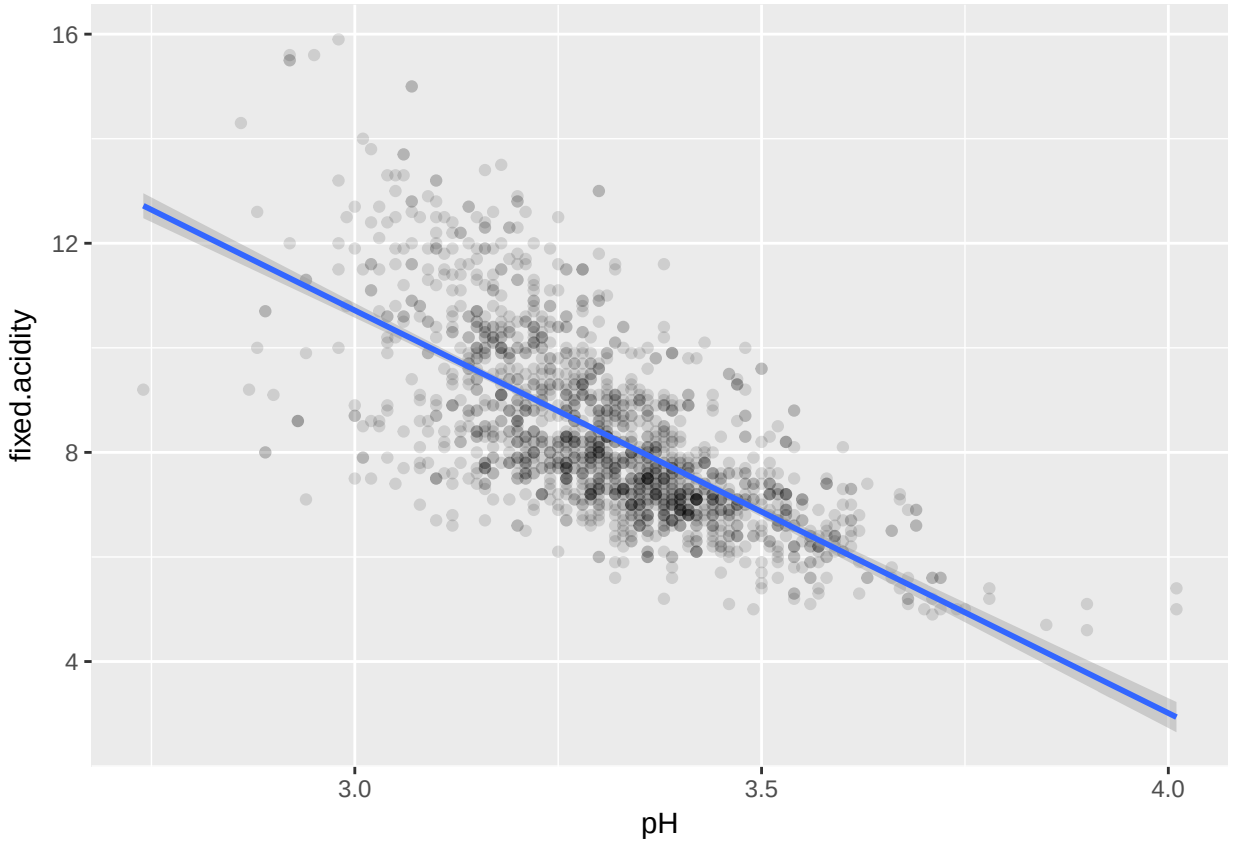
Quality vs. Sulphates



- Median values of citric acid and sulphates showed an increasing trend with quality.

pH vs. others

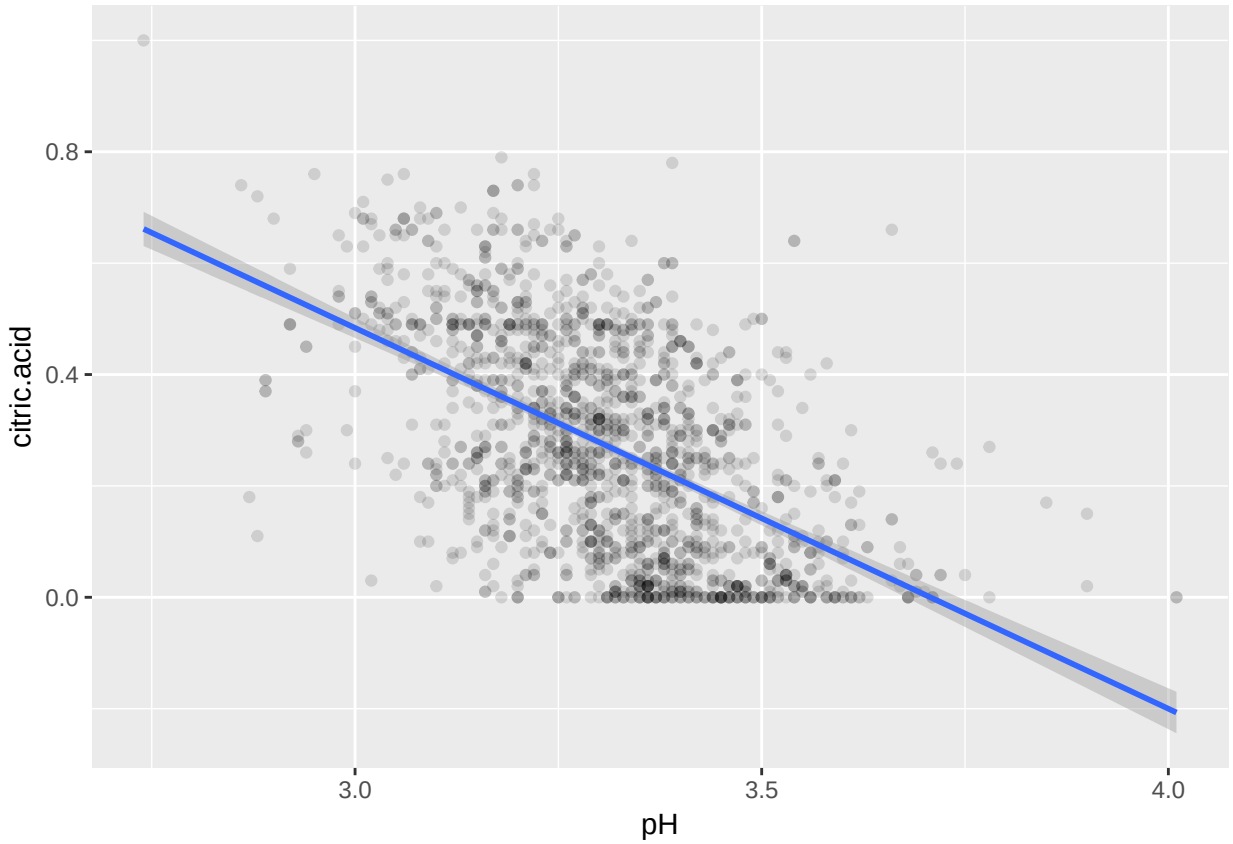
pH vs. Fixed Acidity



- Since pH and fixed acidity are both measure of acidity, there is a meaningful correlation between both. pH decreases with increasing fixed acidity.

```
##
##  Pearson's product-moment correlation
##
## data:  fixed.acidity and pH
## t = -37.366, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7082857 -0.6559174
## sample estimates:
##       cor
## -0.6829782
```

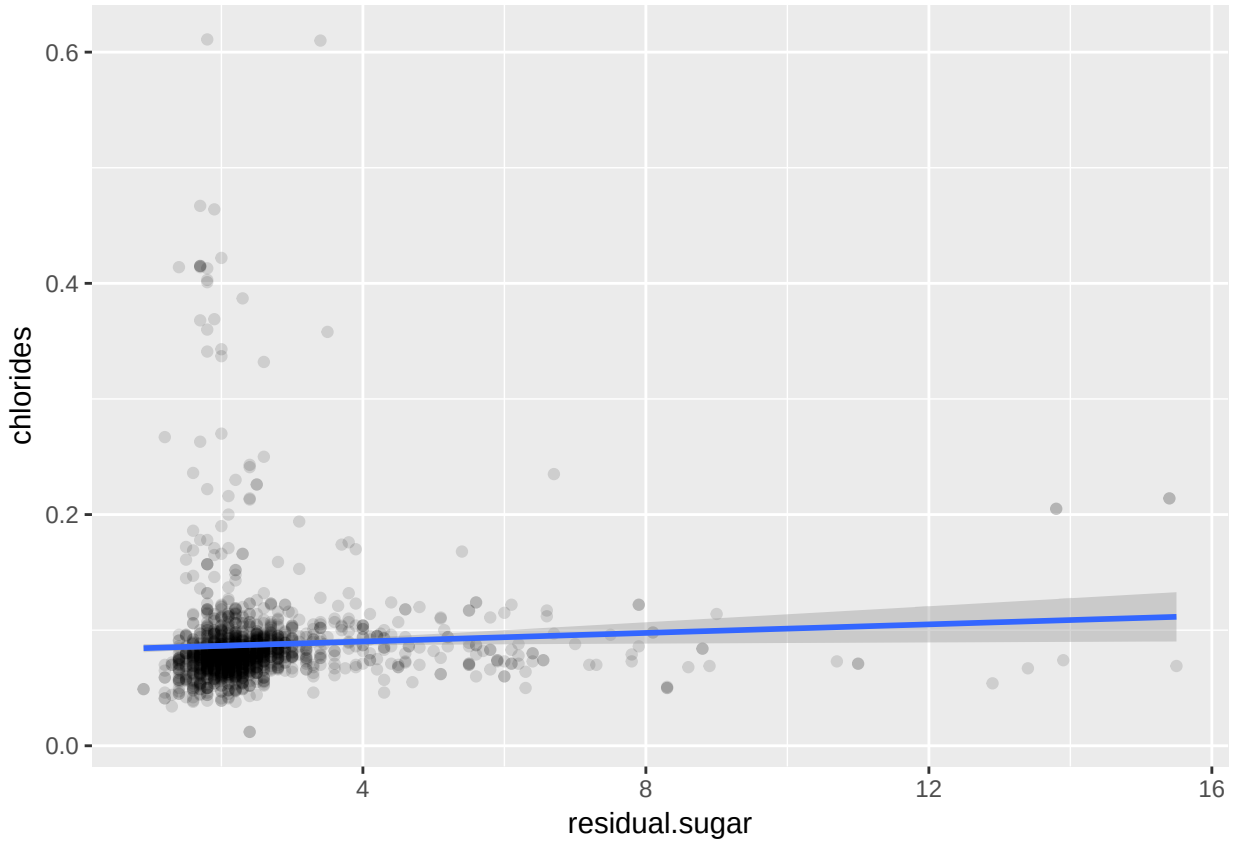
pH vs. Citric Acid

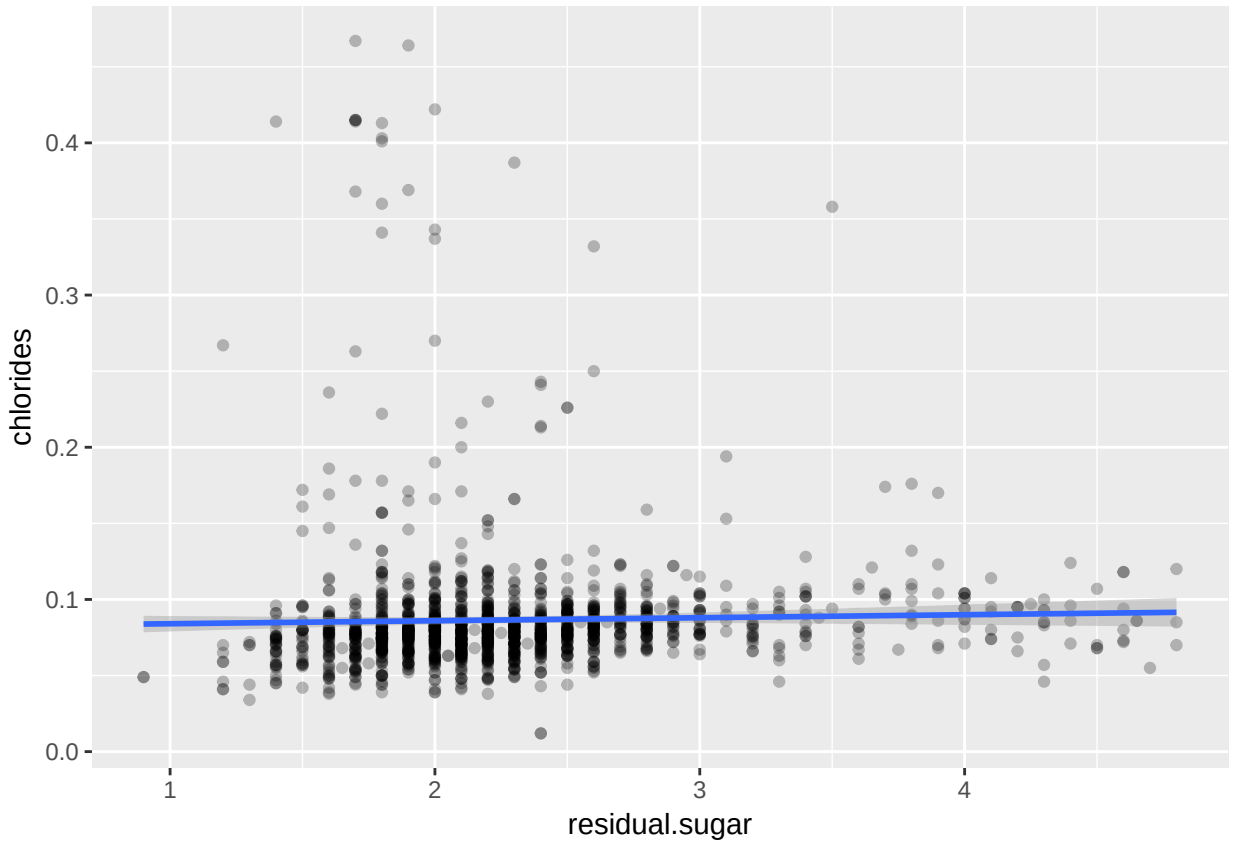


- Since citric acid is an acid, it is increases acidity which inturn decrease with pH. Citric acid is therefor expected to follow a trend simillar to that of fixed acidity.

```
##
##  Pearson's product-moment correlation
##
## data:  citric.acid and pH
## t = -25.767, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5756337 -0.5063336
## sample estimates:
##        cor
## -0.5419041
```

Residual Sugar vs. Chlorides

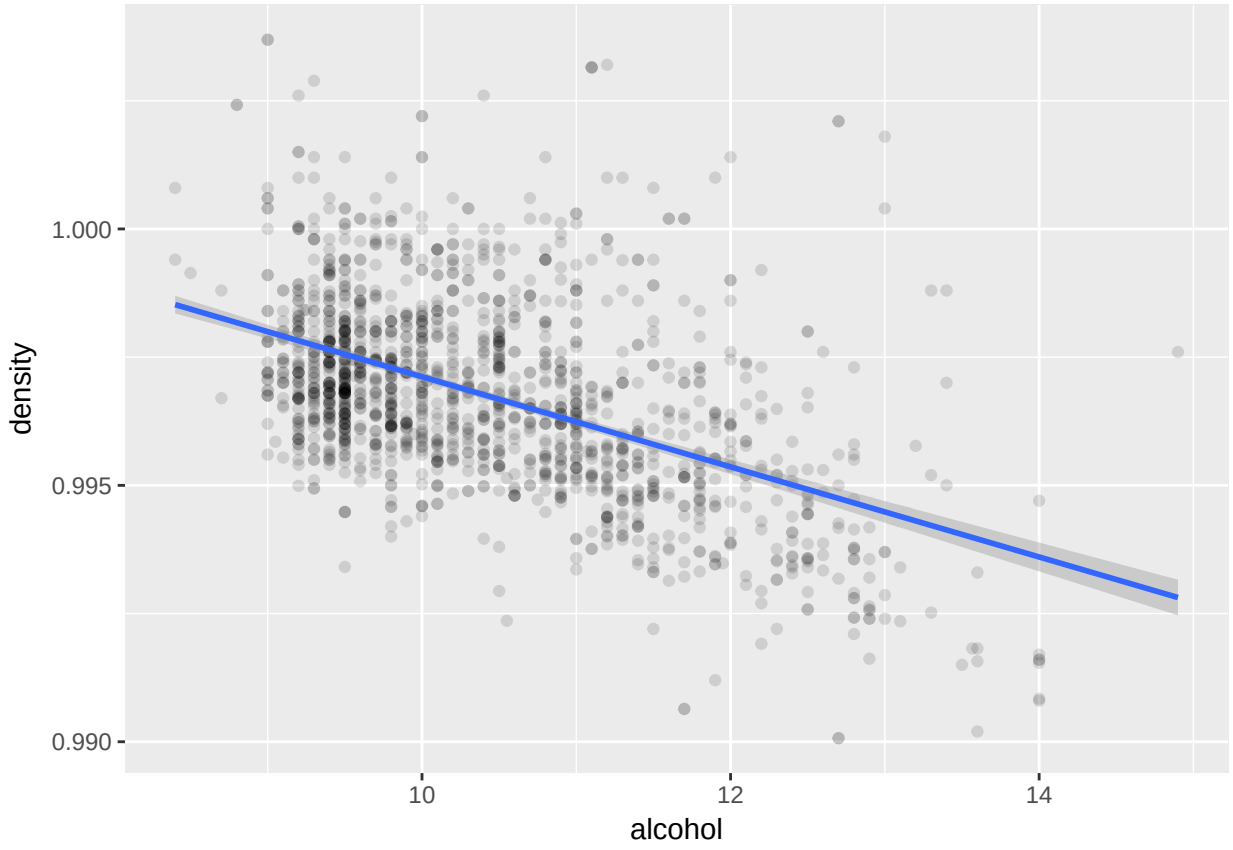





```
##
##  Pearson's product-moment correlation
##
## data:  residual.sugar and chlorides
## t = 9.8152, df = 1451, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2006699 0.2971338
## sample estimates:
##       cor
## 0.2495207
```

- Rresidual sugar seems to vary independent of chlorides even after removing the outliers.

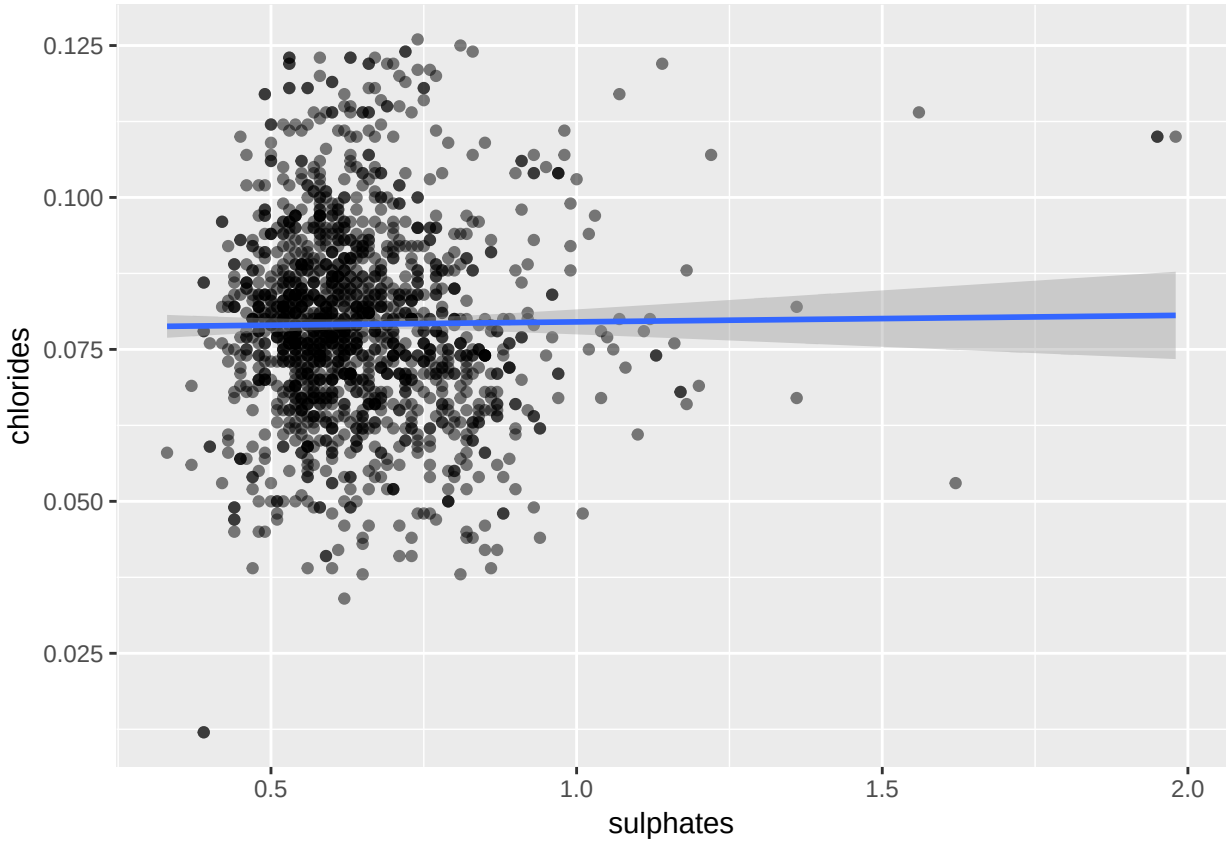
Alcohol vs. Density




```
##
##  Pearson's product-moment correlation
##
## data:  density and alcohol
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##       cor
## -0.4961798
```

- Density decreases with alcohol
- Reason for this correlation : Alcohol is lighter than water, hence density decreases with increased alcohol presence.

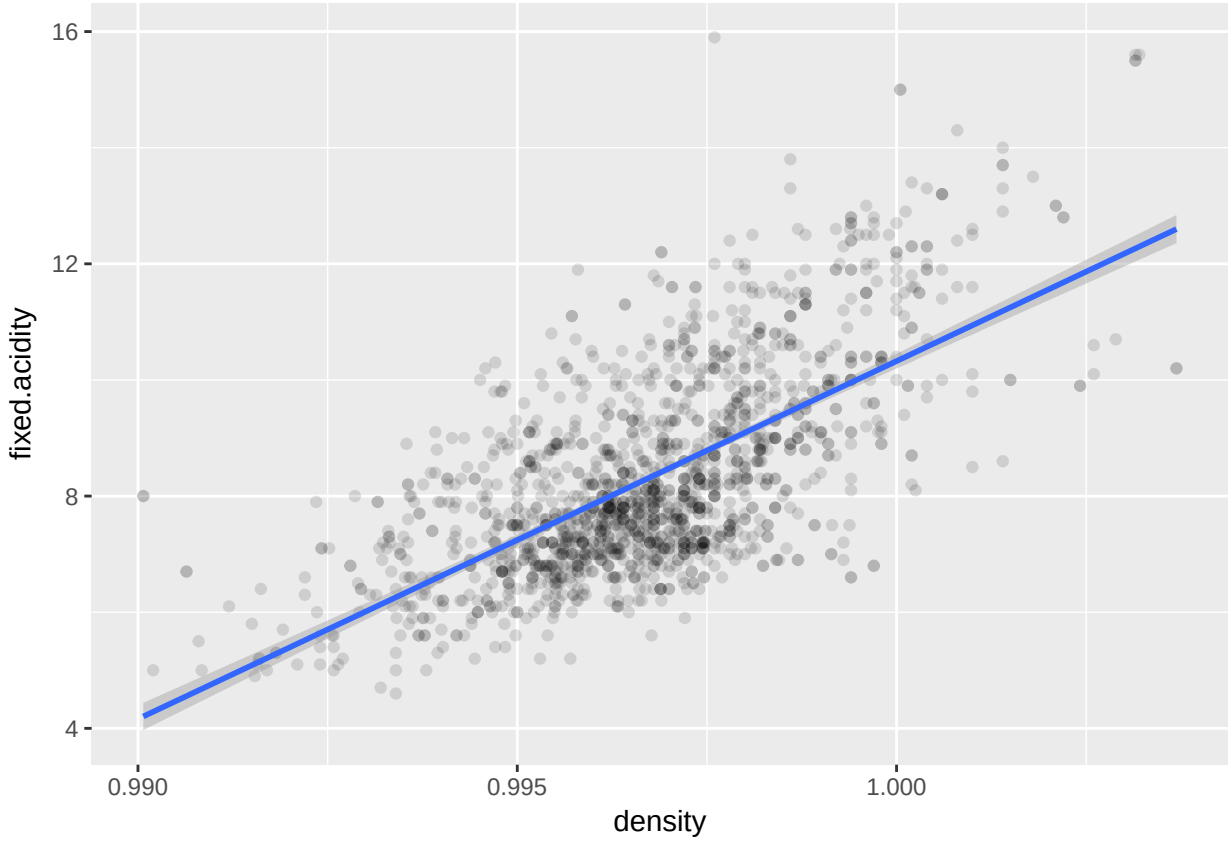
Sulphates vs. Chlorides



- Sulphates have the highest correlation with chlorides, but not not high enough for any meaningful trend.
- 95% of the data points for chlorides are in the range : chlorides < 0.125 and seems to vary independent of sulphates.

Density vs. others

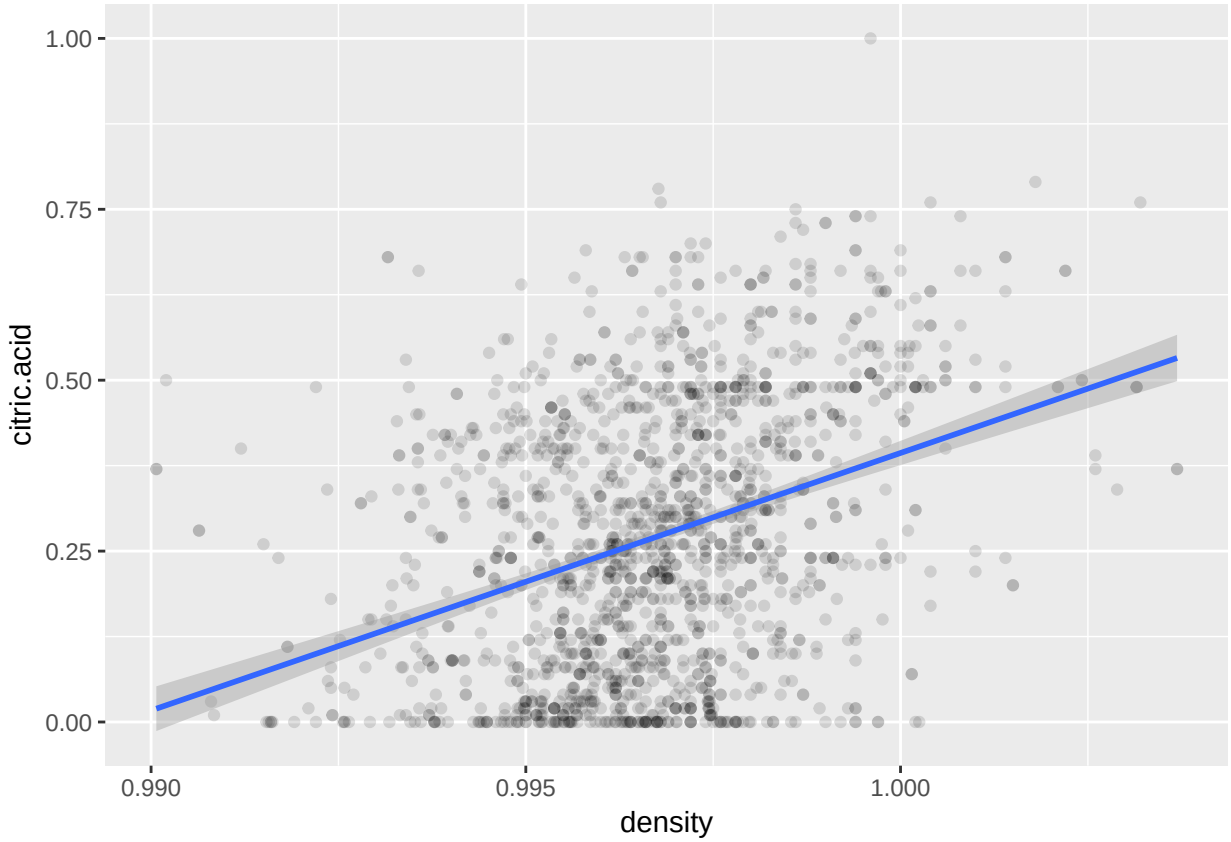
Density vs. Fixed Acidity




```
##
##  Pearson's product-moment correlation
##
## data:  density and fixed.acidity
## t = 35.877, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6399847 0.6943302
## sample estimates:
##       cor
## 0.6680473
```

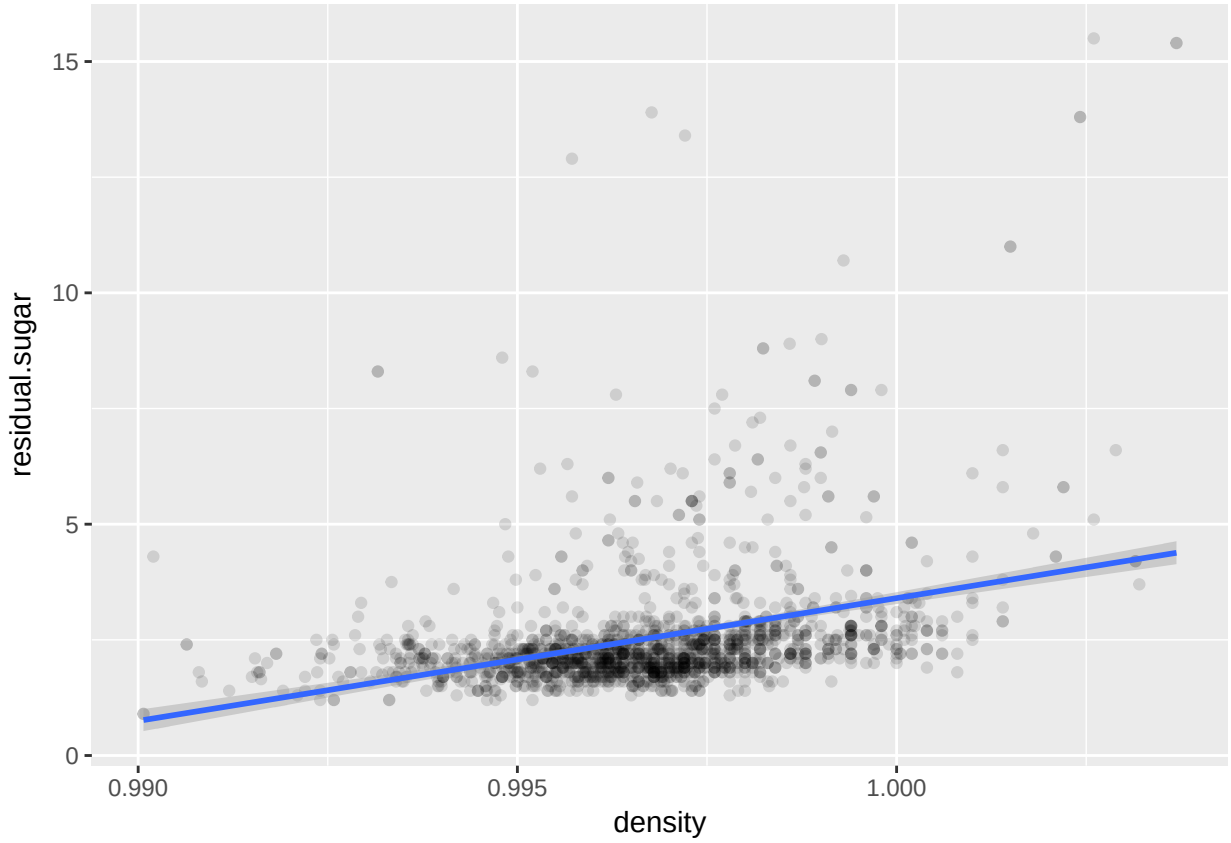
- Density is highly correlated with fixed acidity and increases with fixed acidity.
- Citric acid has a density of 1.66 gm/cm^3 . Since citric acid is mostly the cause of fixed acidity, it is plausible that density increases with citric acid content.

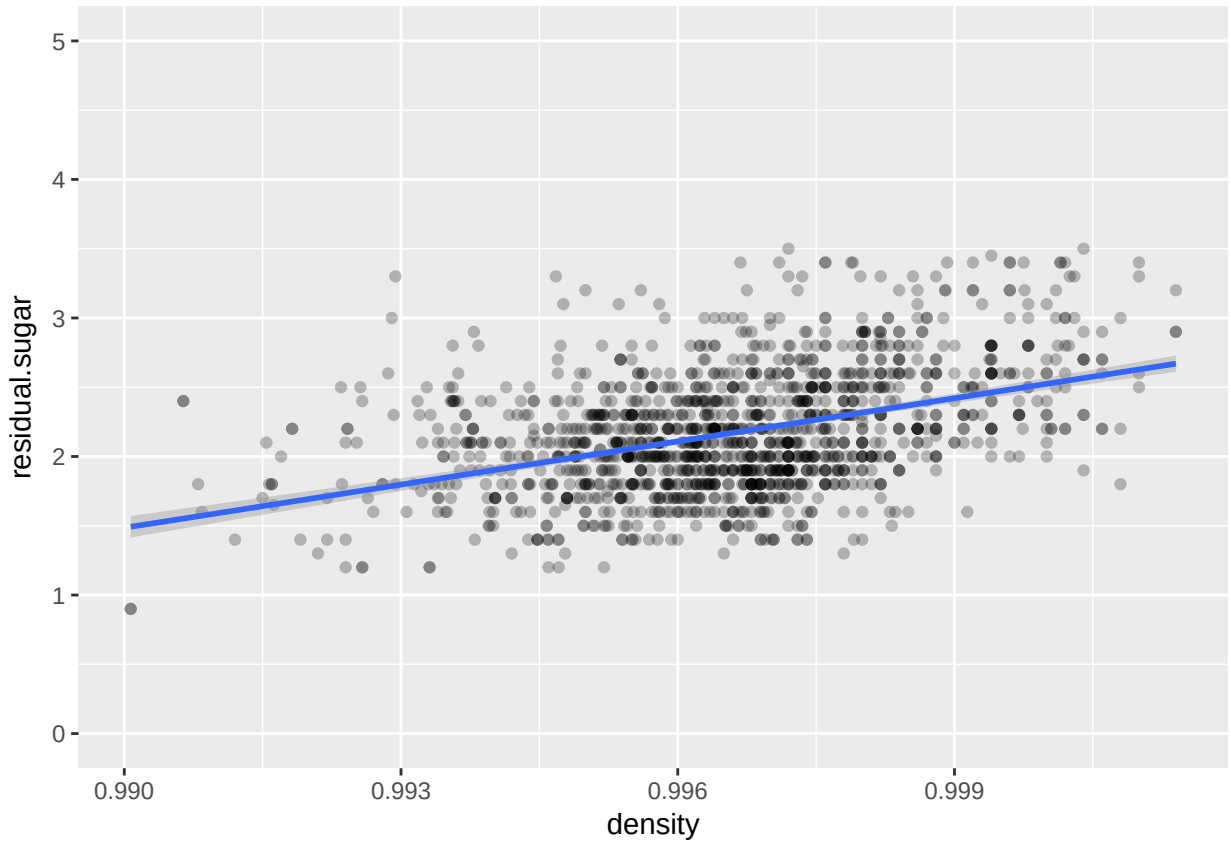
Density vs. Citric Acid



- The correlation between citric acid and density is not strong enough.

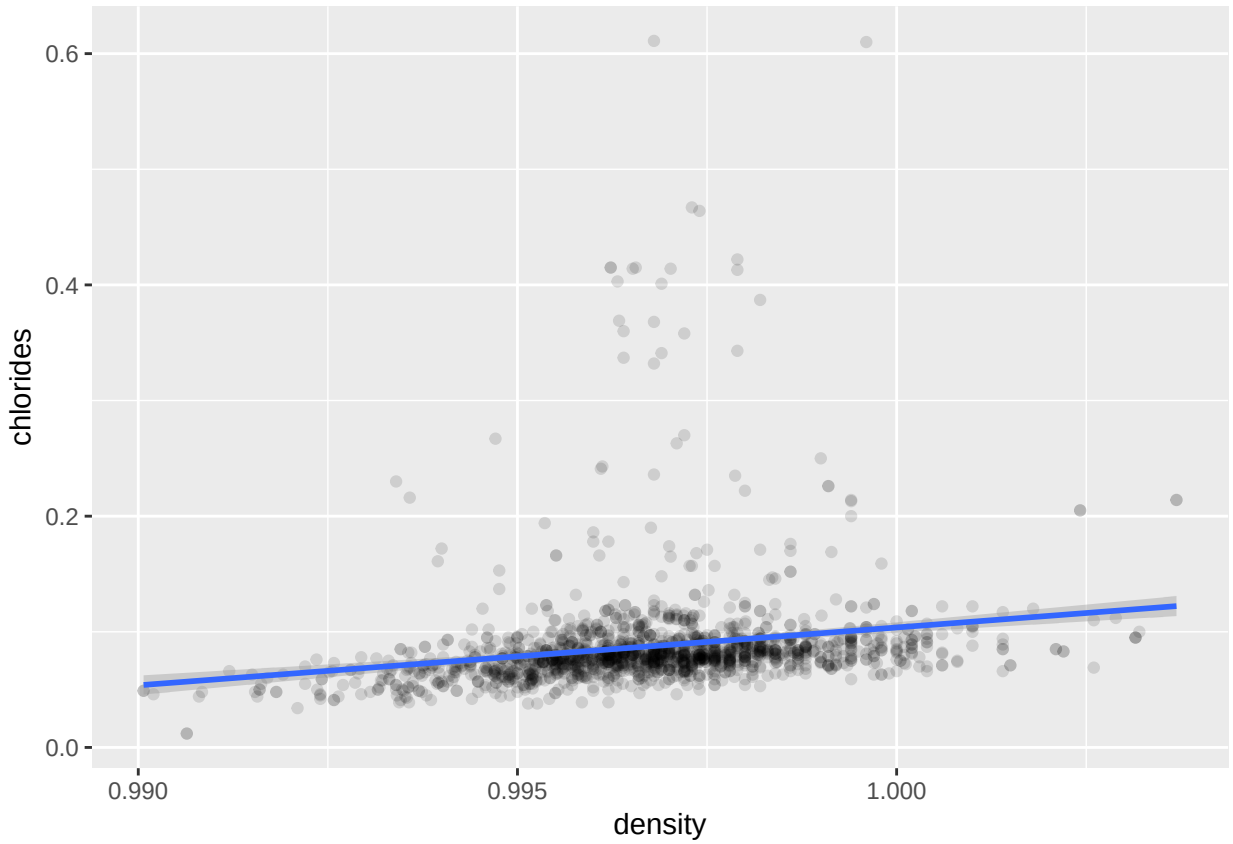
Density vs. Residual Sugar

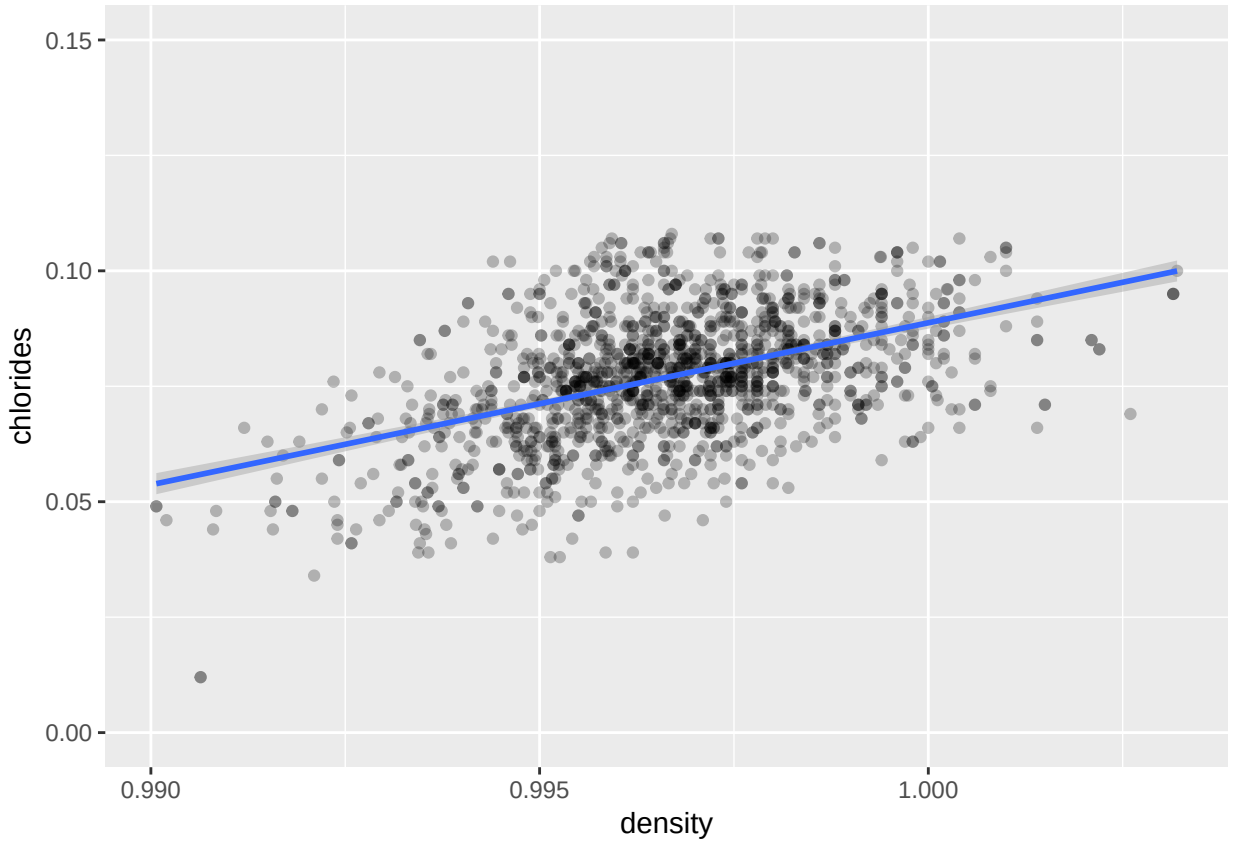





```
##
##  Pearson's product-moment correlation
##
## data:  density and fixed.acidity
## t = 34.697, df = 1433, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6465499 0.7028622
## sample estimates:
##       cor
## 0.6756906
```

- After removing the upper 90% of the residual sugar data (outliers), density seems to have an upward trend with residual sugar with a correlation factor = 0.67. It is plausible as increasing sugar content increases density as mentioned in the description of density.





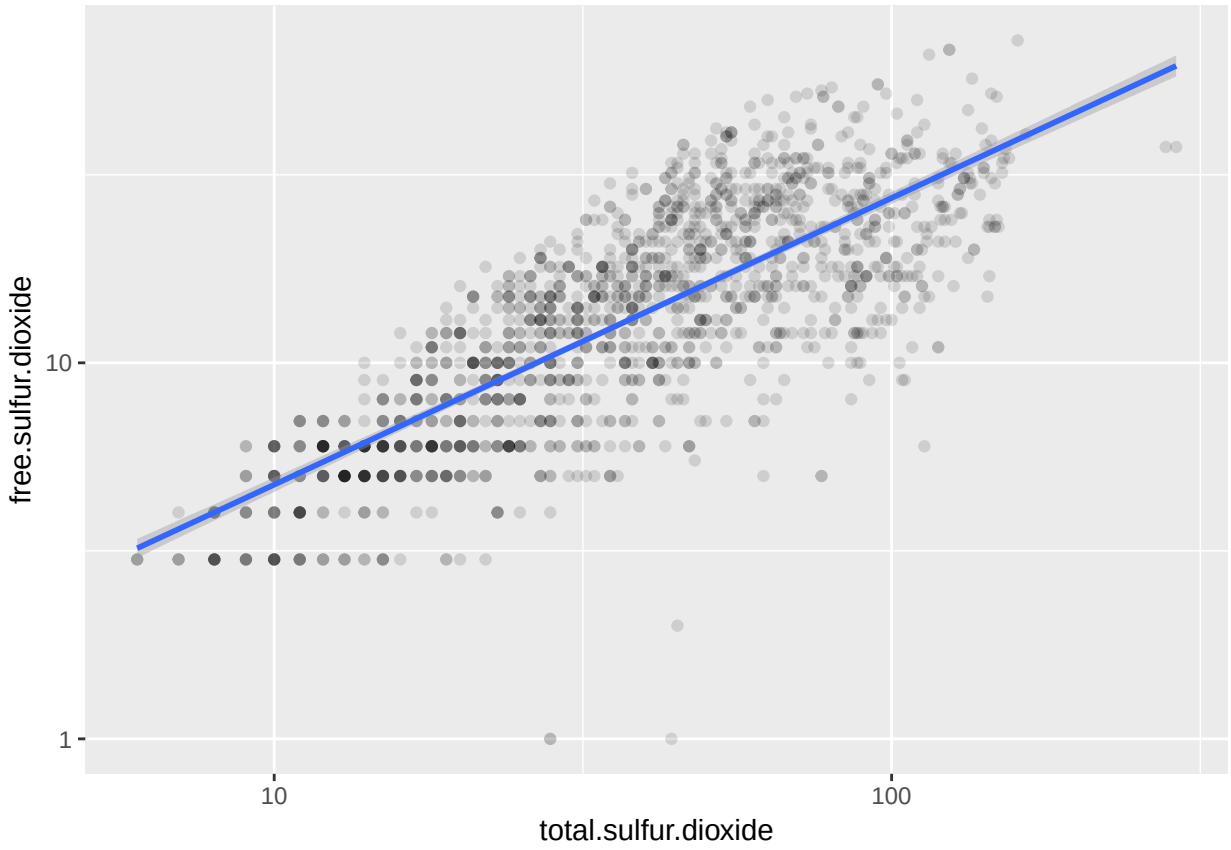
- After removing the upper 90% of the chlorides data (outliers), density seems to have an upward trend

with chlorides with a correlation factor = 0.67.

```
##
##  Pearson's product-moment correlation
##
## data:  density and fixed.acidity
## t = 34.989, df = 1436, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6494617 0.7053386
## sample estimates:
##       cor
## 0.6783799
```

Total Sulfur dioxide vs. all other variables

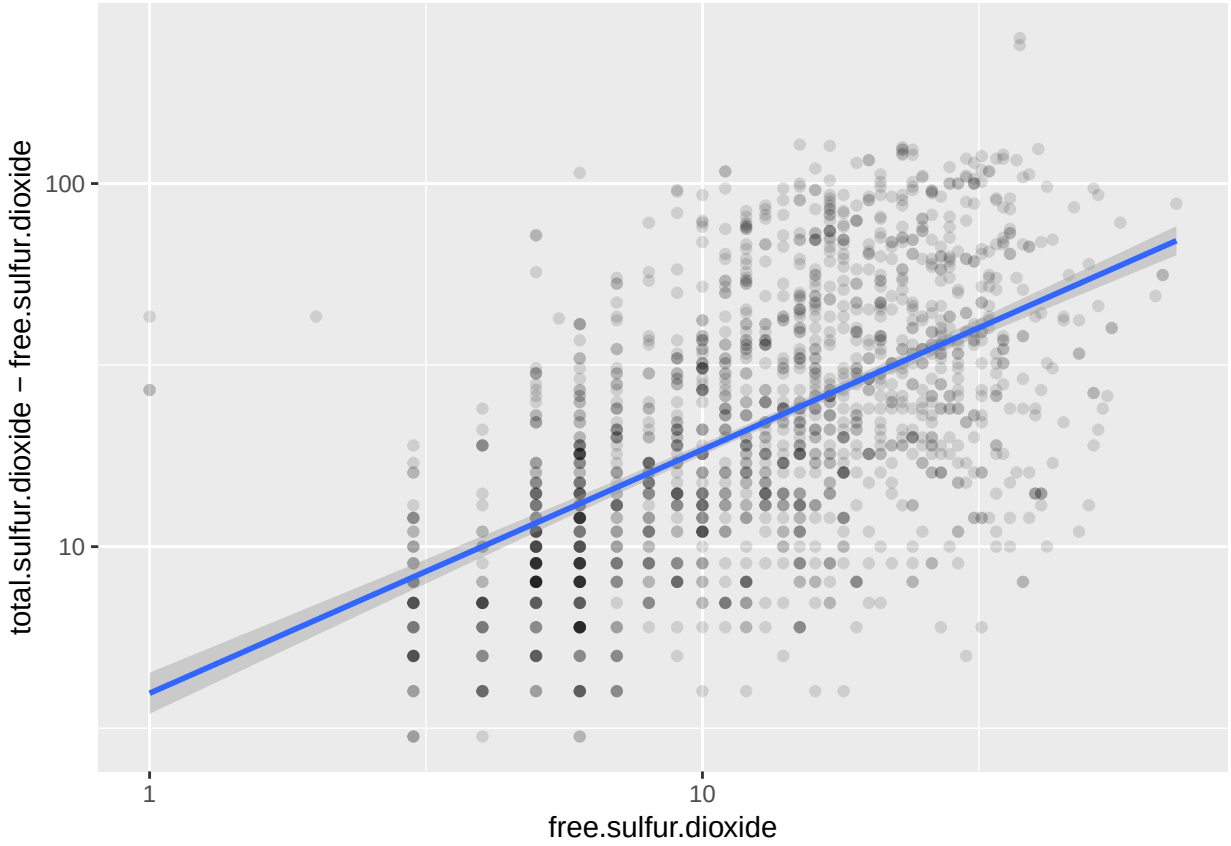
Total Sulfur dioxide vs. Free Sulfur dioxide




```
##
##  Pearson's product-moment correlation
##
## data:  free.sulfur.dioxide and total.sulfur.dioxide
## t = 35.84, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6395786 0.6939740
## sample estimates:
```

```
##       cor
## 0.6676665
```

- Free sulfur dioxide and Total sulfur dioxide seems to be highly correlated since free sulfur dioxide is a part of the total sulfur dioxide

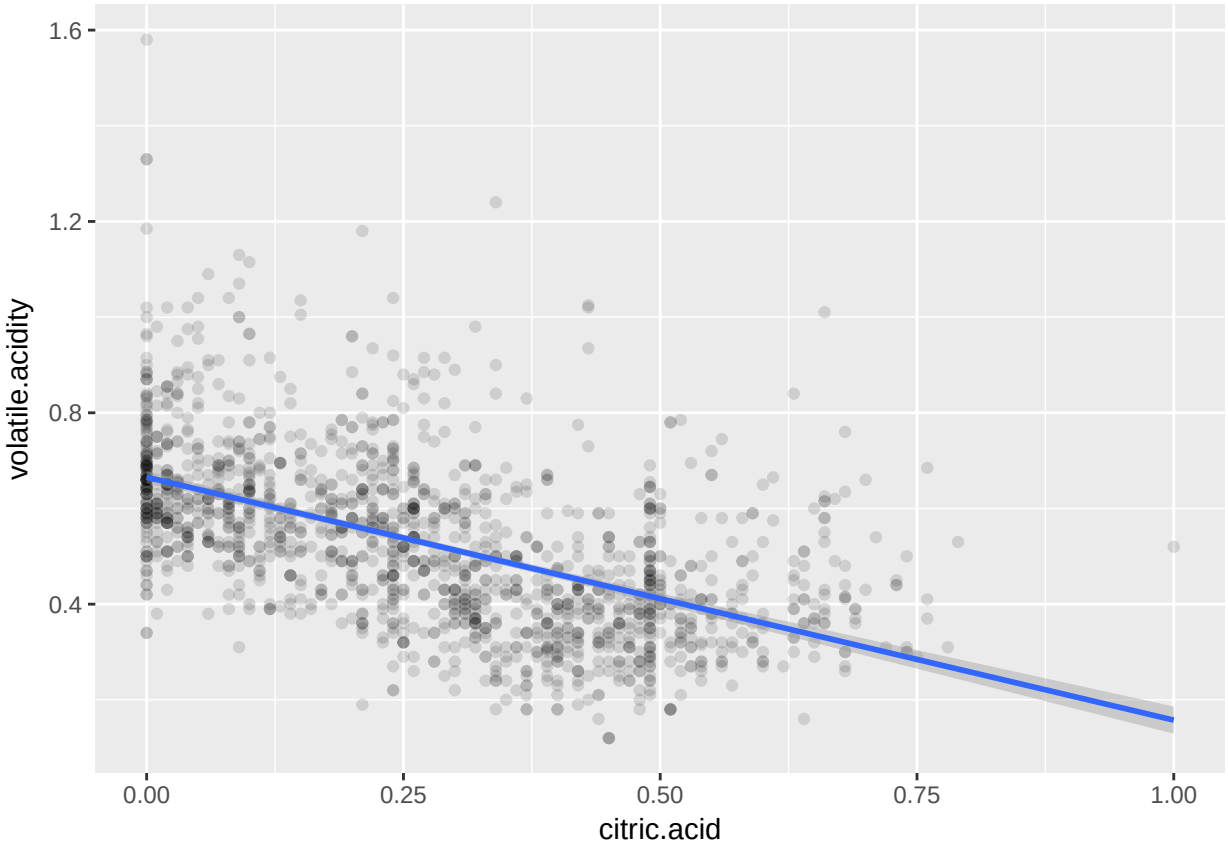



```
##
##  Pearson's product-moment correlation
##
## data:  free.sulfur.dioxide and total.sulfur.dioxide - free.sulfur.dioxide
## t = 18.771, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3841336 0.4644895
## sample estimates:
##       cor
## 0.4251489
```

- The difference in total and free sulfur dioxide is less correlated aganist each other.

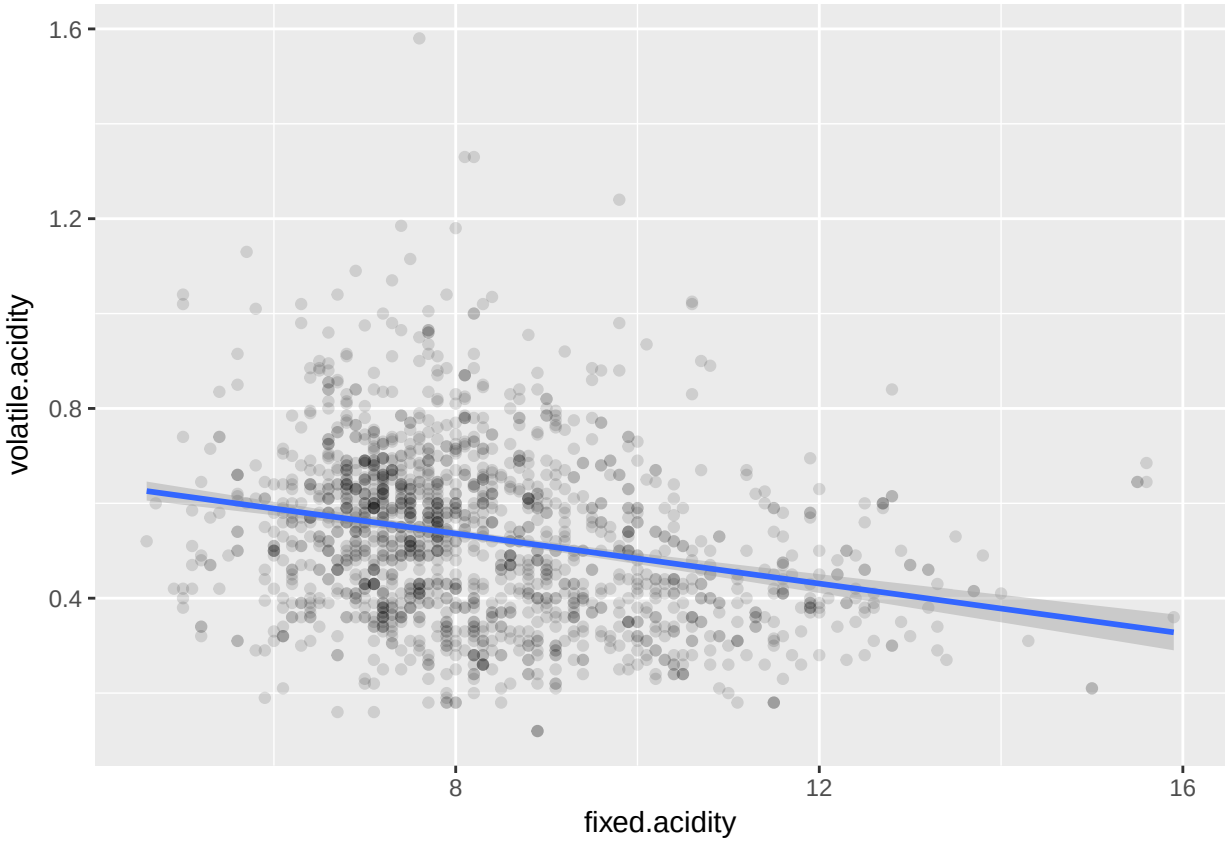
Citric Acid vs. others

Citric Acid vs. Volatile Acidity



- Volatile acidity is found to be decreasing with citric acid.

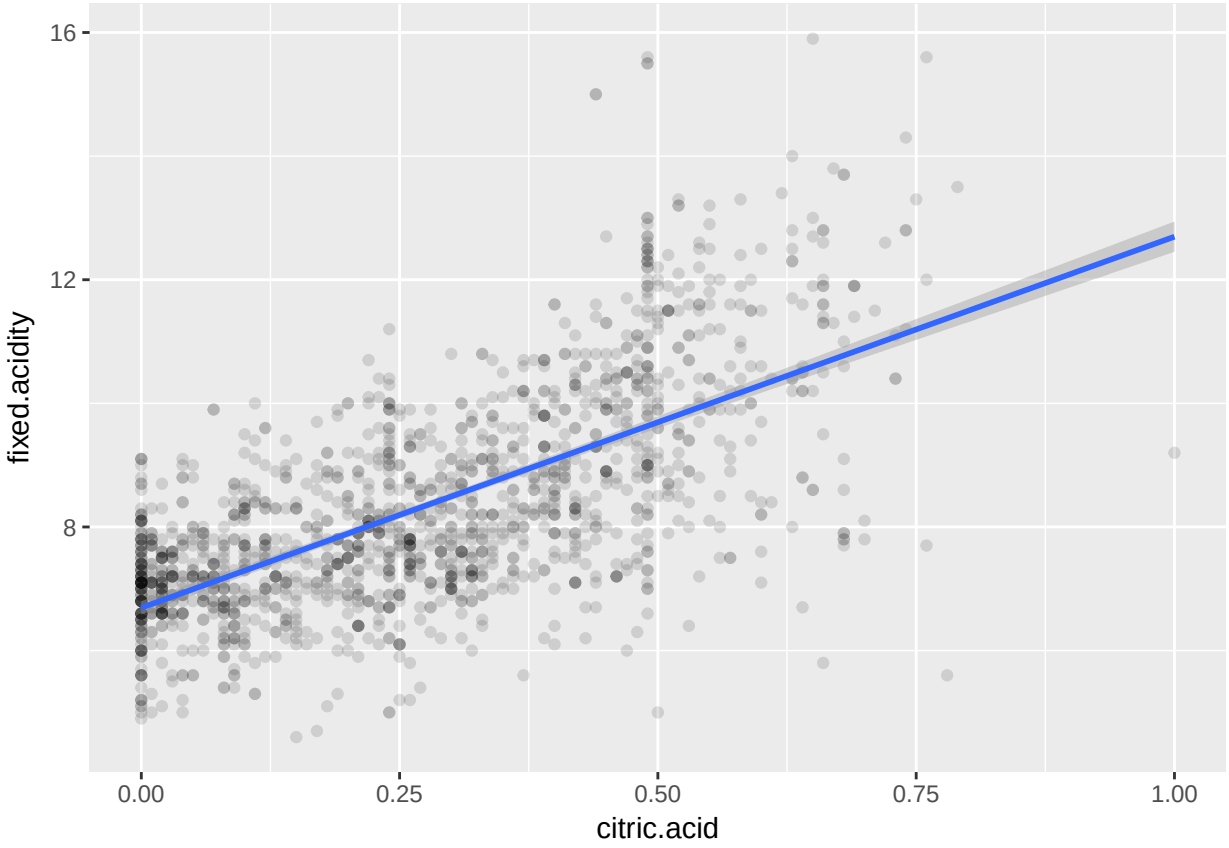
Volatile Acidity vs. Fixed Acidity




```
##
##  Pearson's product-moment correlation
##
## data:  fixed.acidity and volatile.acidity
## t = -10.589, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3013681 -0.2097433
## sample estimates:
##        cor
## -0.2561309
```

- Fixed acidity does not seem to be highly correlated with volatile acidity. This seems to agree with prior observations. Since fixed acidity is highly correlated with citric acid and pH, it can be deduced that volatile acidity(acetic acid) does not significantly affect pH or fixed acidity of wine.

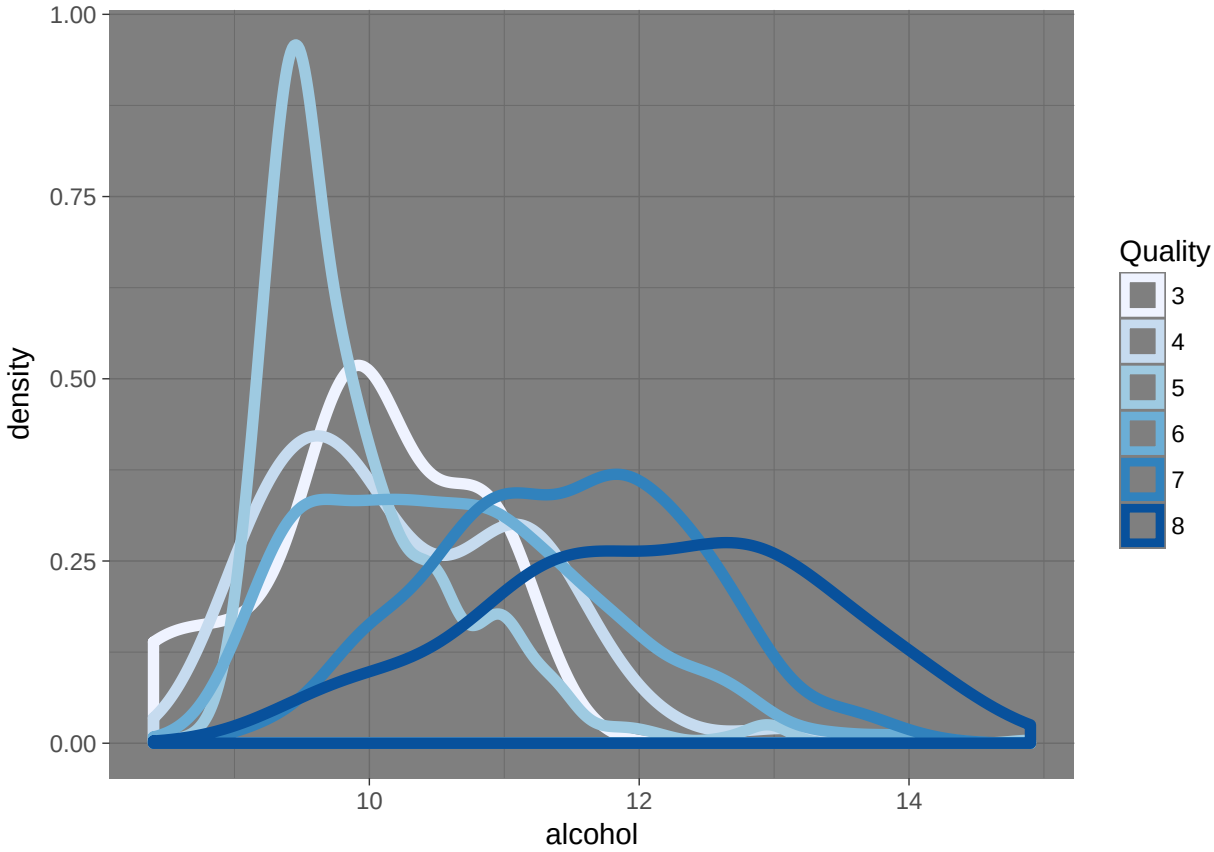
Citric Acid vs. Fixed Acidity



- It is suprising to find that fixed acidity (a measure of tartaric acid), is correlated with citric acid.

Density distribution

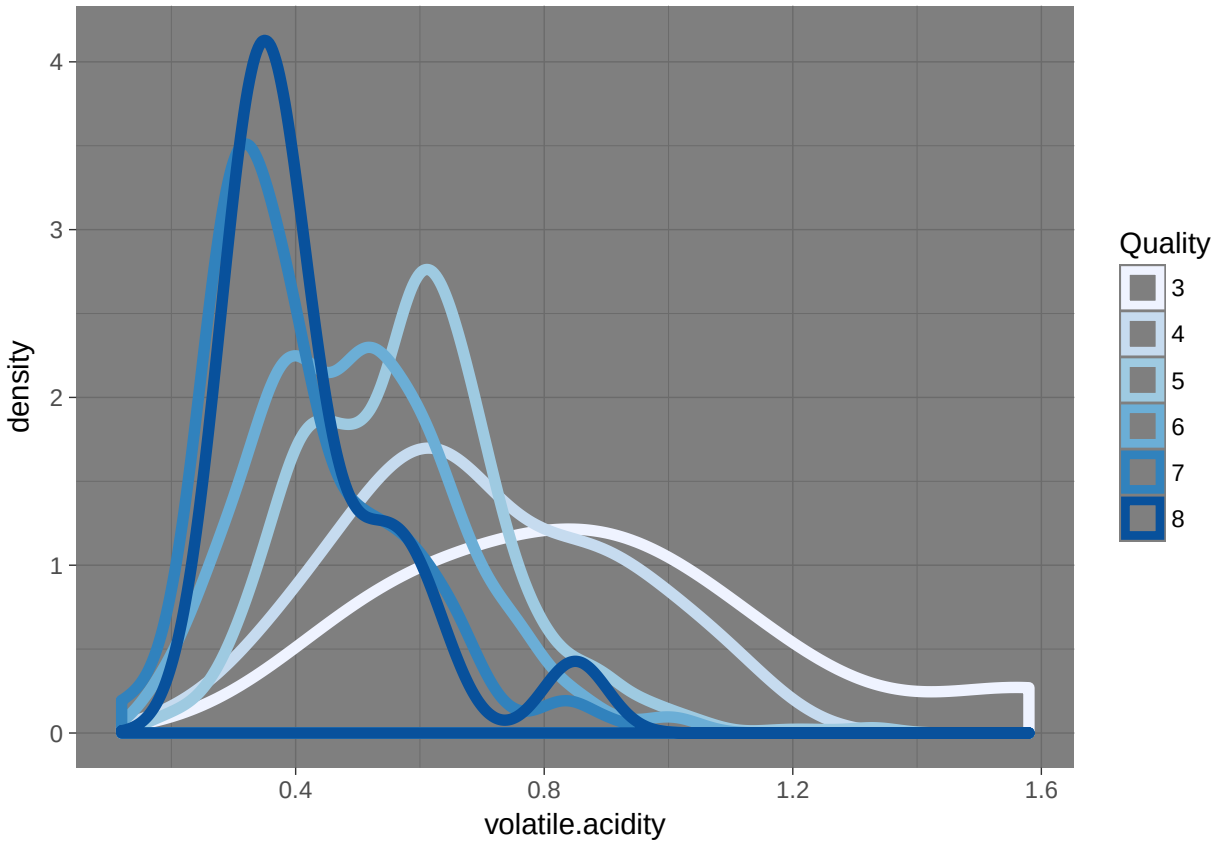
Alcohol




```
## # A tibble: 6 × 3
##   quality        Skew         Ku
##     <int>       <dbl>      <dbl>
## 1       3 -0.40890877 -0.9932596
## 2       4  0.61327279 -0.2320630
## 3       5  1.83042317  5.2544218
## 4       6  0.54442666 -0.1579873
## 5       7  0.01033155 -0.4677010
## 6       8 -0.19693316 -0.9776767
```

- Distributuion of alcohol is more skewed towards right for low quality wines but more spread out with high quality wine.

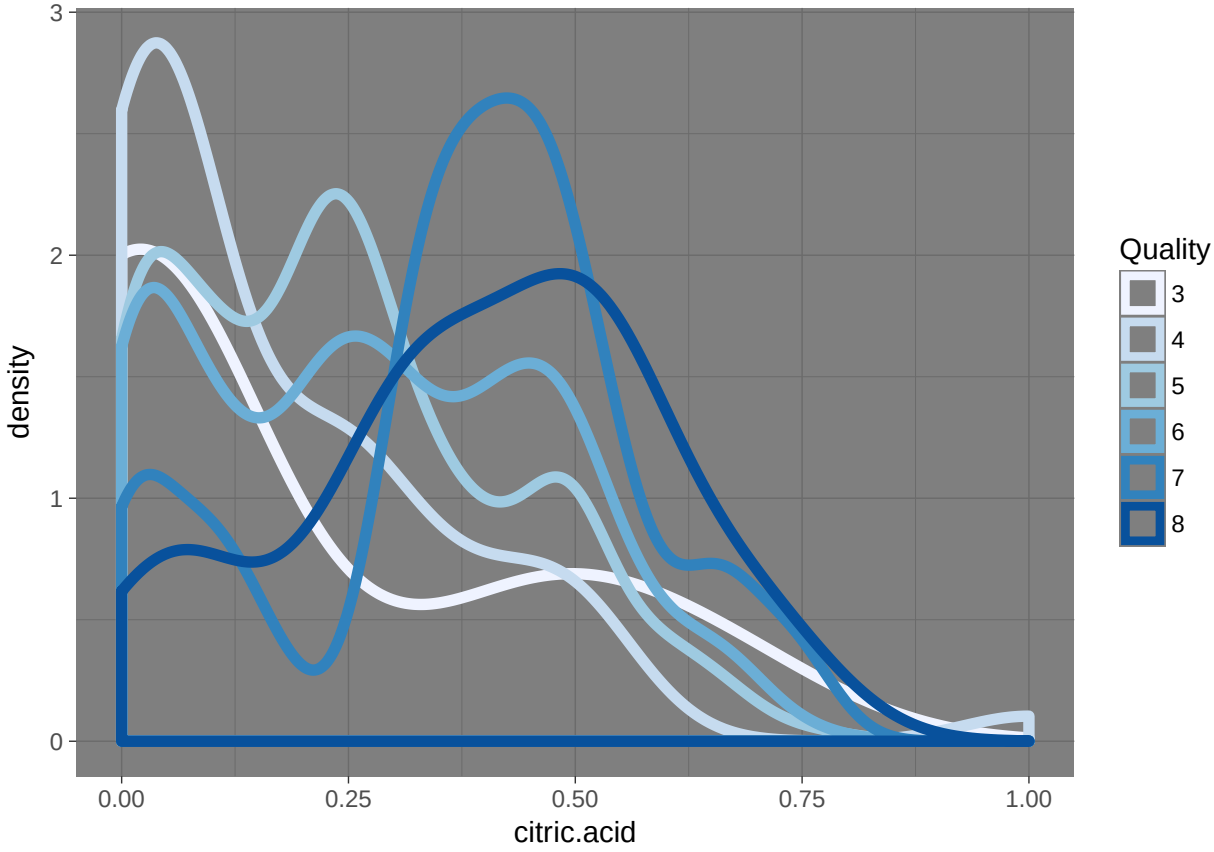
Volatile Acidity




```
## # A tibble: 6 × 3
##   quality      Skew          Ku
##     <int>     <dbl>       <dbl>
## 1       3 0.6270191 -0.54054899
## 2       4 0.1488264 -0.81529662
## 3       5 0.5911007  1.36339622
## 4       6 0.4327494  0.08843808
## 5       7 0.9428877  0.76022793
## 6       8 1.4455109  1.62440497
```

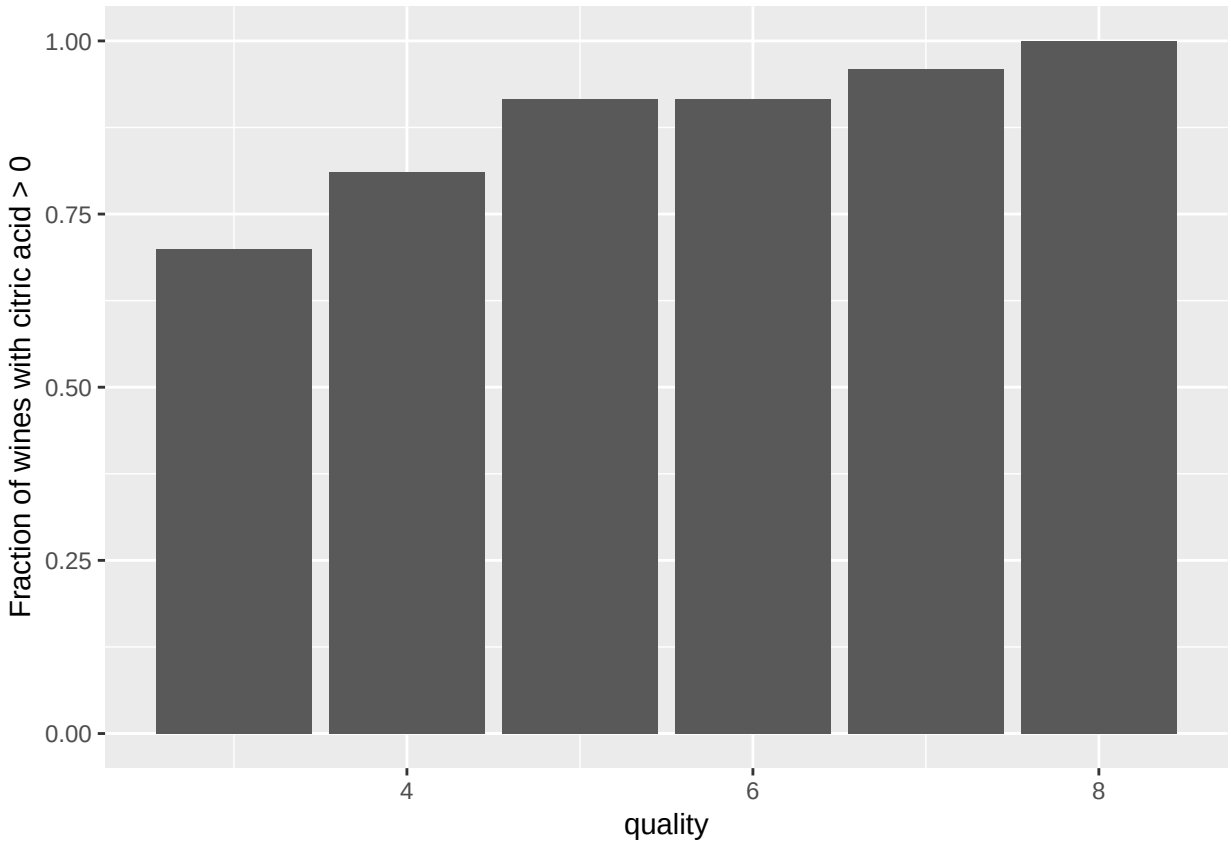
- Volatile acidity of high quality wines have narrow distribution compared to more broadly distributed low quality wines.
- Mode of the volatile distribution is decreasing with quality and the distribution is getting narrower with increasing quality as well.

Citric Acid



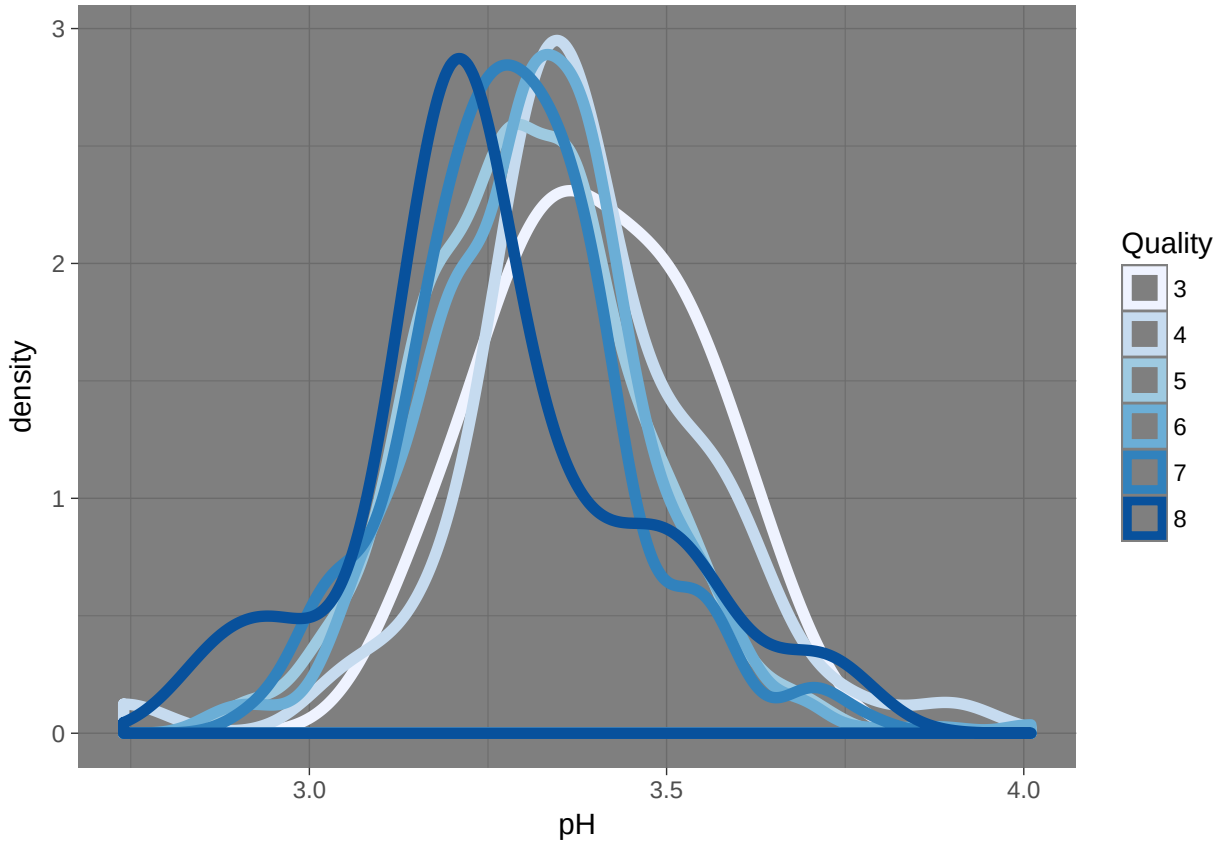

```
## # A tibble: 6 × 3
##   quality       Skew         Ku
##     <int>      <dbl>      <dbl>
## 1       3  0.8850968 -1.0898173
## 2       4  1.6198326  3.3427853
## 3       5  0.5222718 -0.5163235
## 4       6  0.2213311 -0.9880063
## 5       7 -0.3755223 -0.4595533
## 6       8 -0.3260505 -0.9180710
```

- This plot shows that citric acid = 0 decreases with increasing quality.



- Citric acid adds freshness to wine. Higher the quality of wine, more fraction of wines contain some amount of citric acid. This plot explains the freshness of high quality wines.

pH



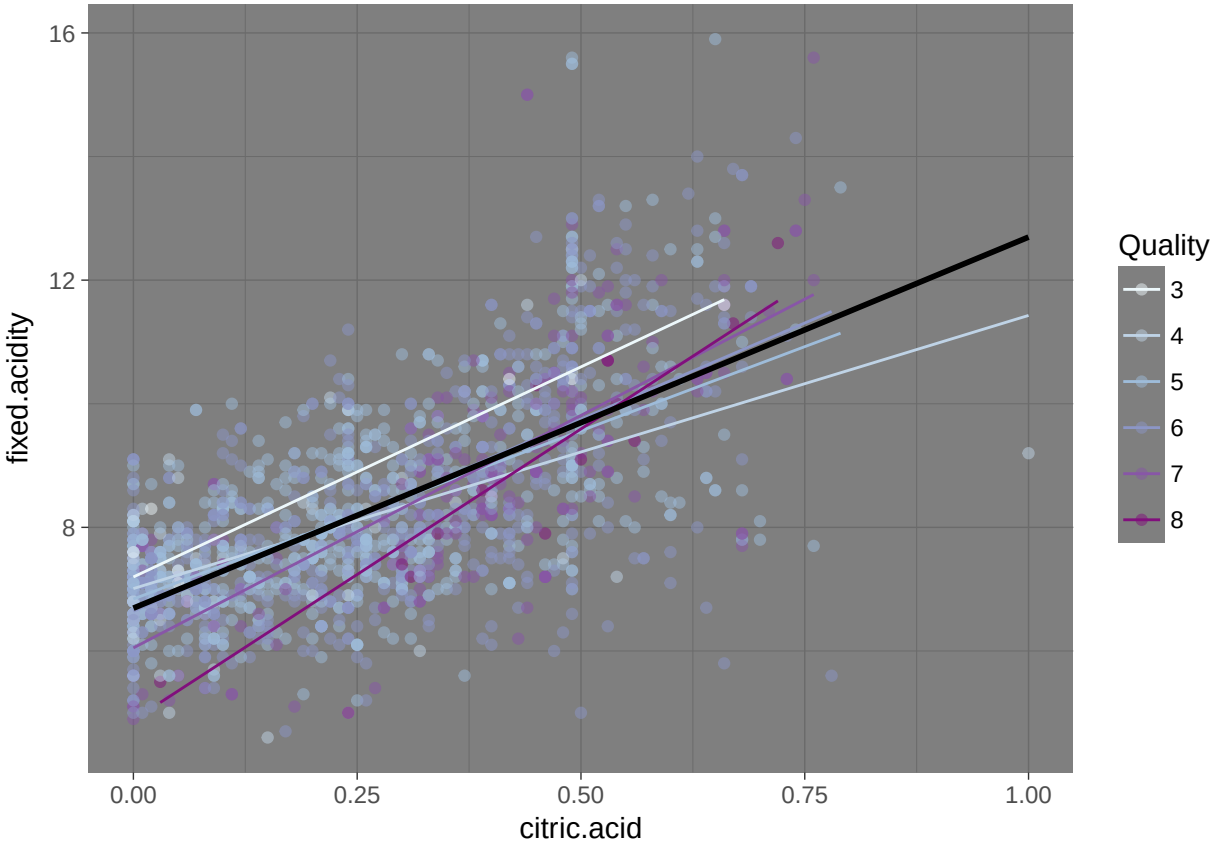

```
## # A tibble: 6 × 3
##   quality         Skew          Ku
##     <int>        <dbl>       <dbl>
## 1       3 -0.009144773 -1.28854080
## 2       4 -0.293059304  2.35443949
## 3       5  0.053054942 -0.01995589
## 4       6  0.301382195  1.44202134
## 5       7  0.377166503  0.58981835
## 6       8  0.354035873 -0.13122625
```

- pH of wines seems to have a simillar distribution curve accross all wine qualities. This explains why pH might not be a good indicating factor of wine quality.

Multi-Variable Analysis

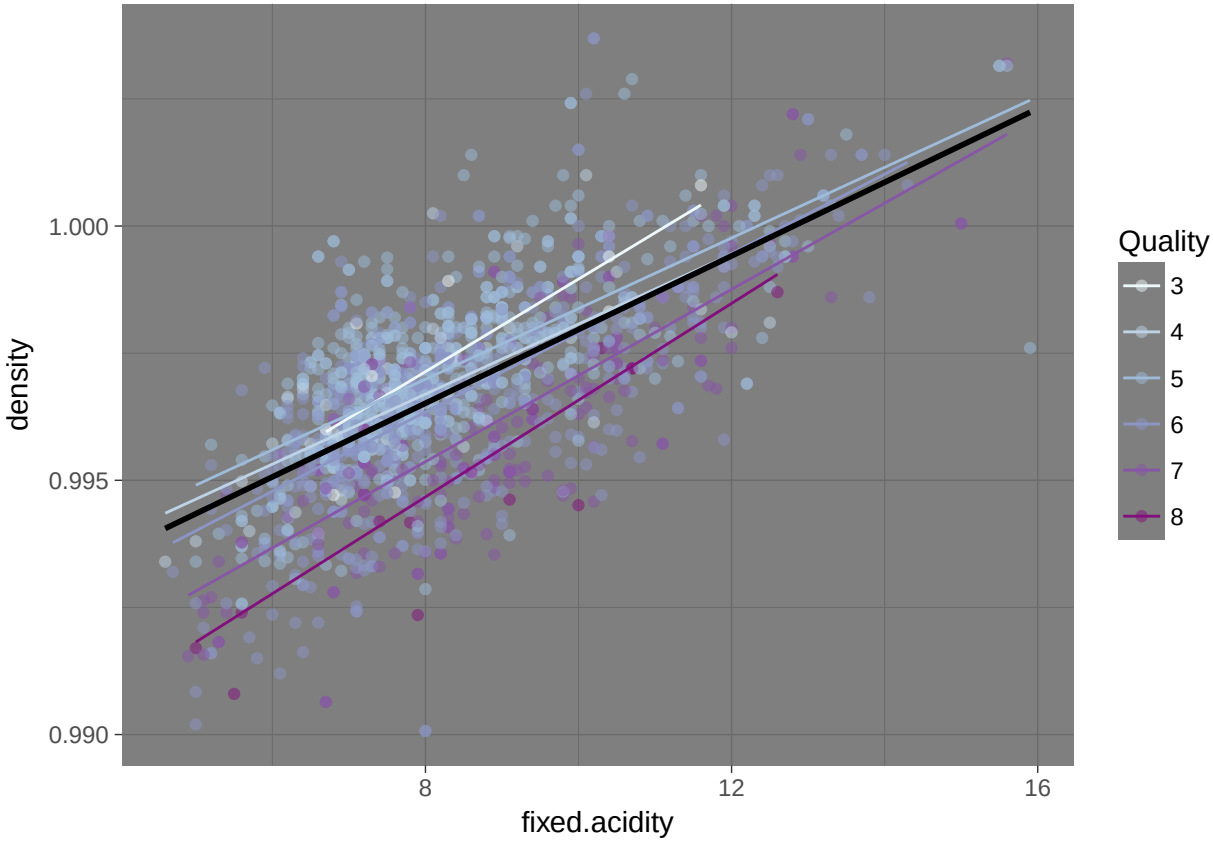
quality vs.

quality, fixed.acidity vs. citric.acid



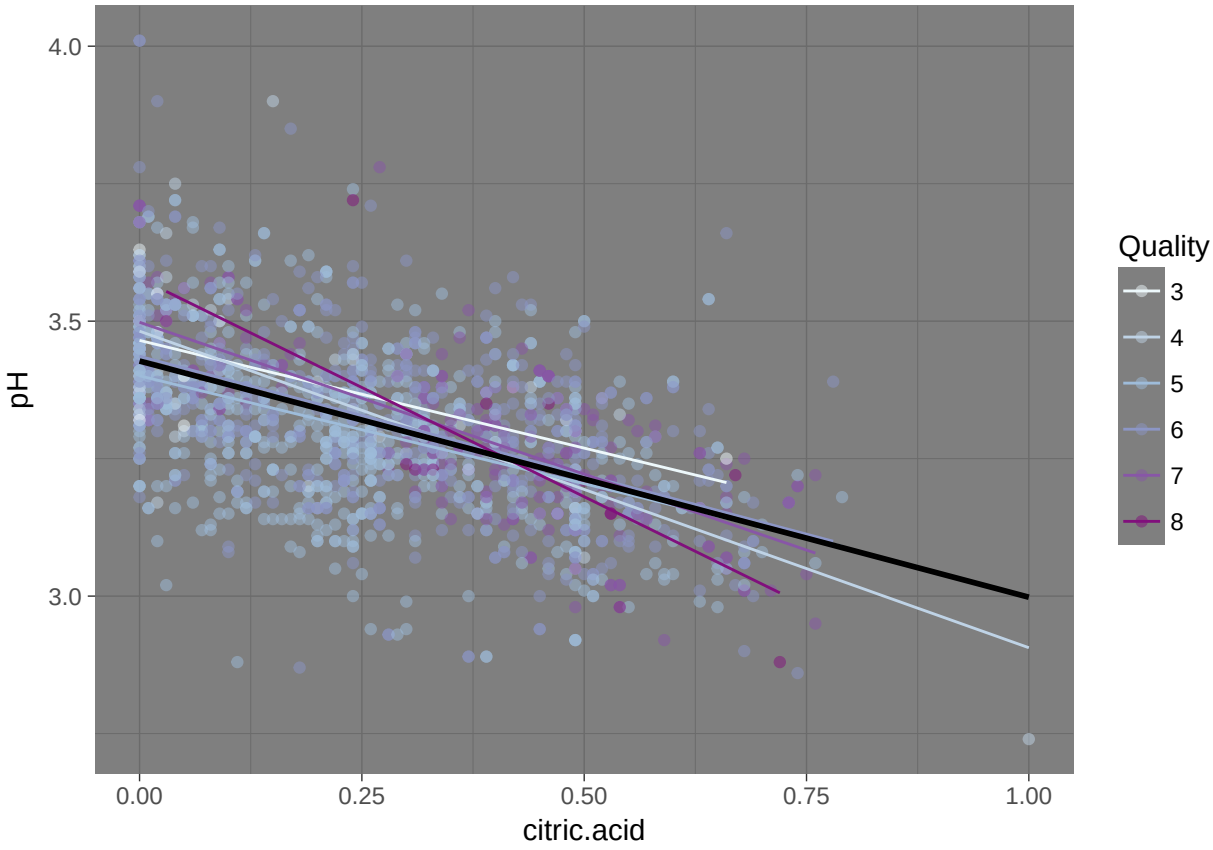
- Fixed acidity increases with citric acid accross all quality wines. It can be infered that citic acid has a significant influence on fixed acidity accross different quality wines.

quality, fixed.acidity vs. density



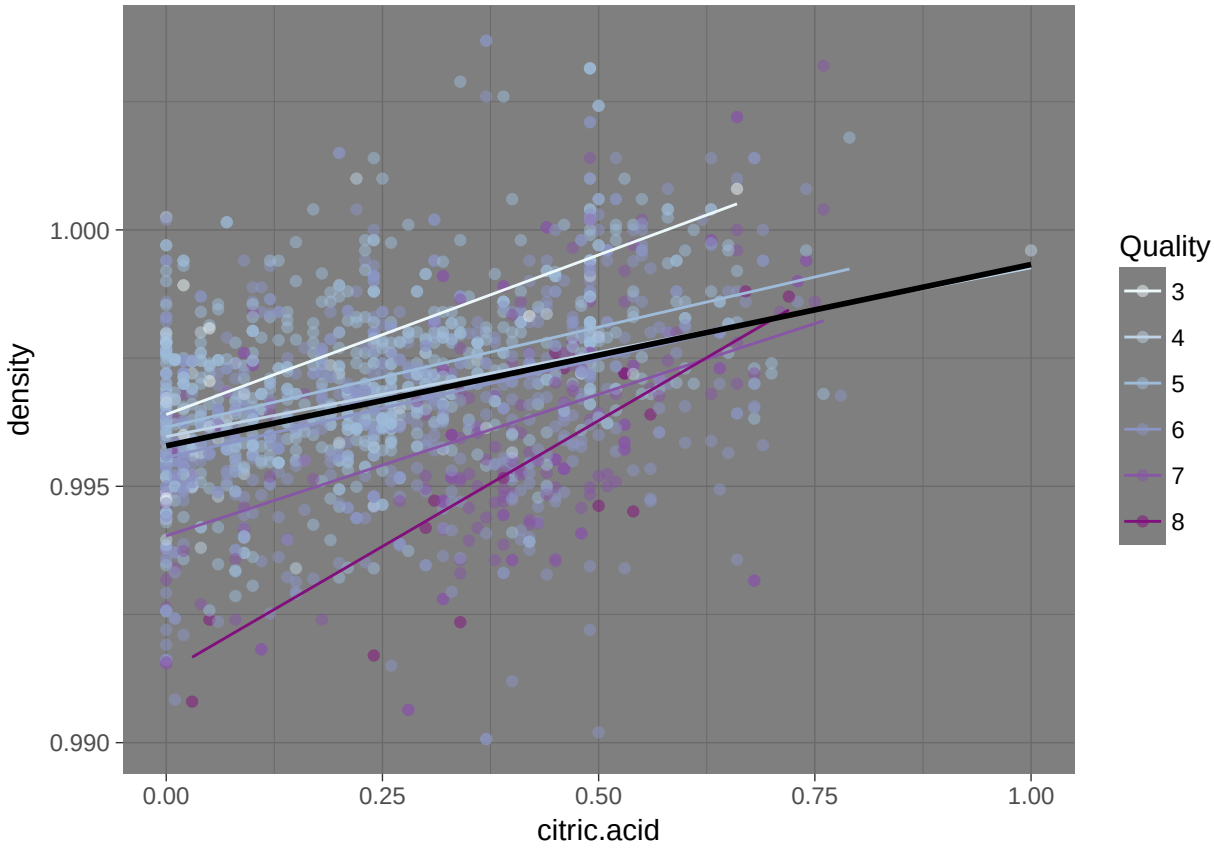
- Fixed acidity and density are positively correlated accross all quality wines. Shows tartaric acid plays a vital role in controlling the wine density.

quality, fixed.acidity vs. pH



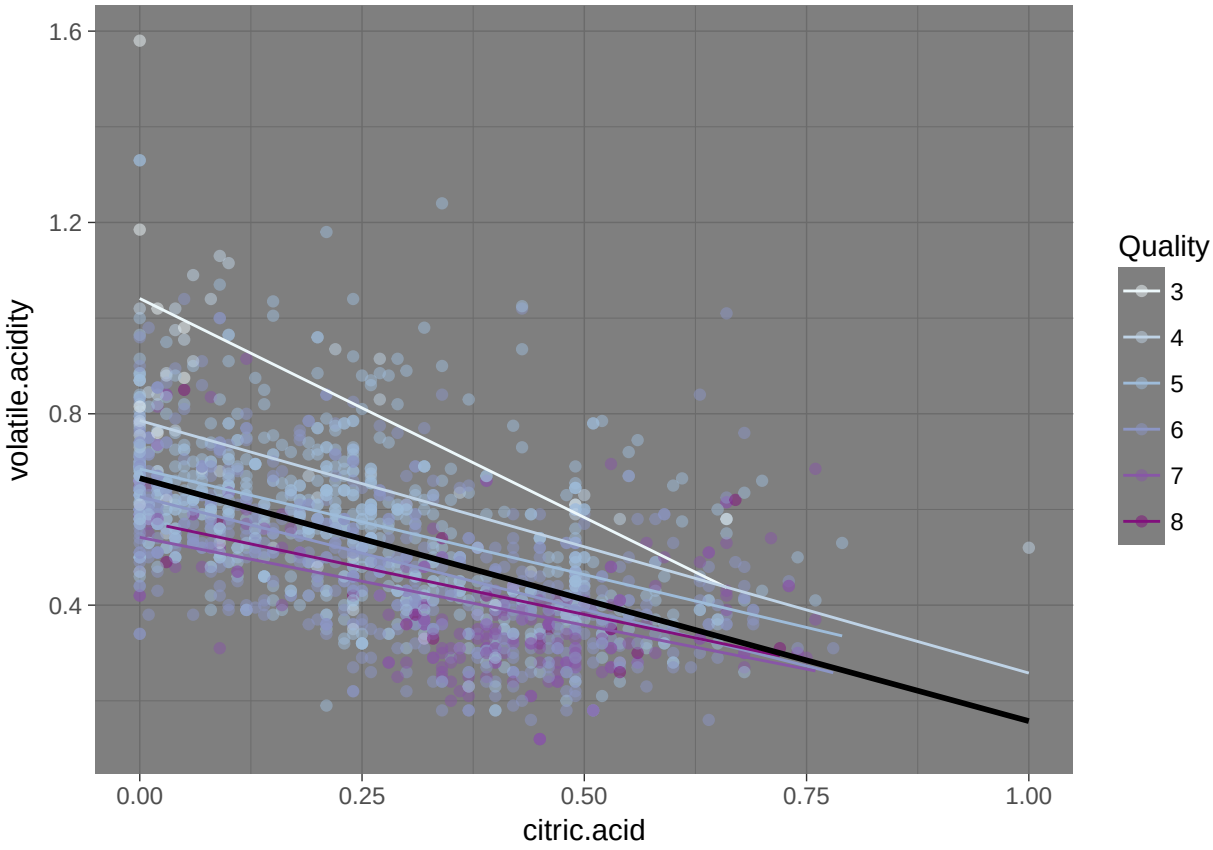
- In general fixed acidity decreases with pH accross all quality wines.

quality, citric.acid vs. pH



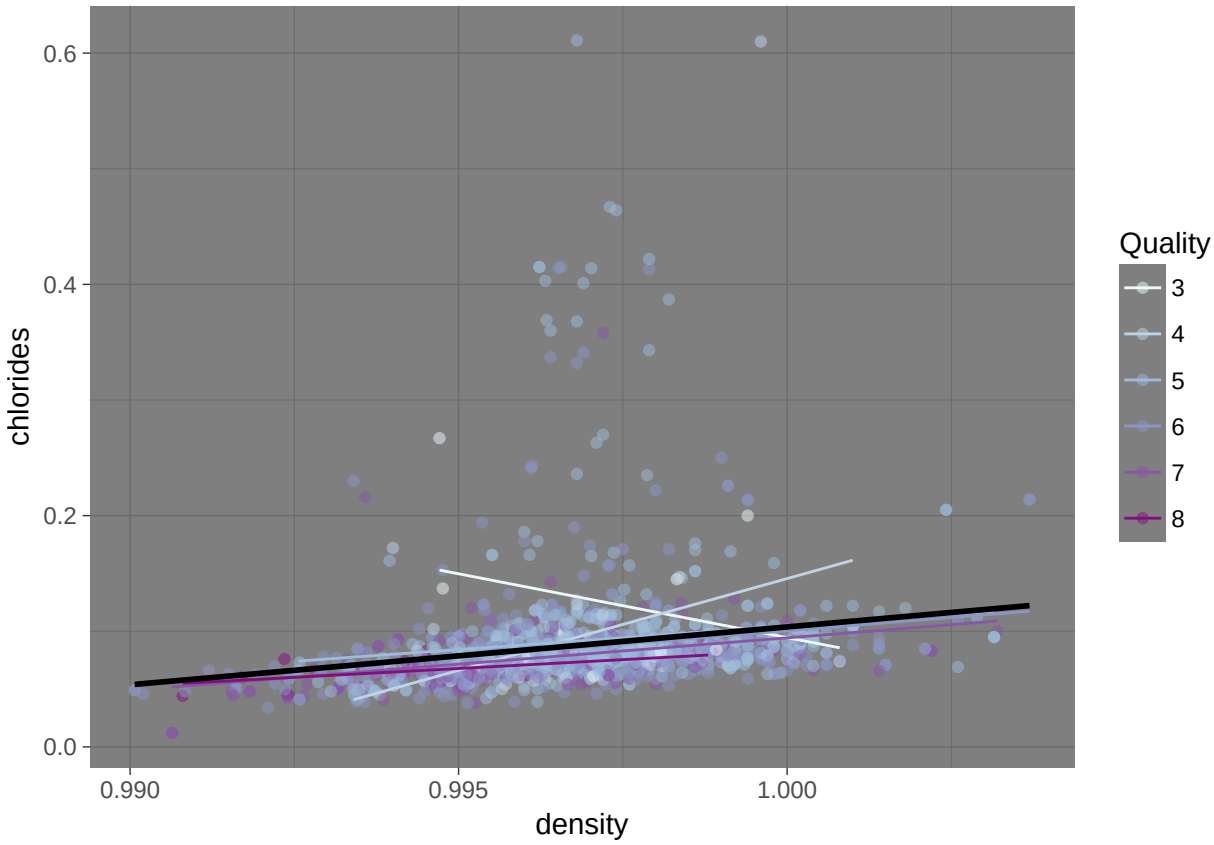
- Density seems to increase with citric acid across all quality wines. This clearly indicates citric acid presence significantly contributes to the density of wine.

quality, citric.acid vs. volatile.acidity



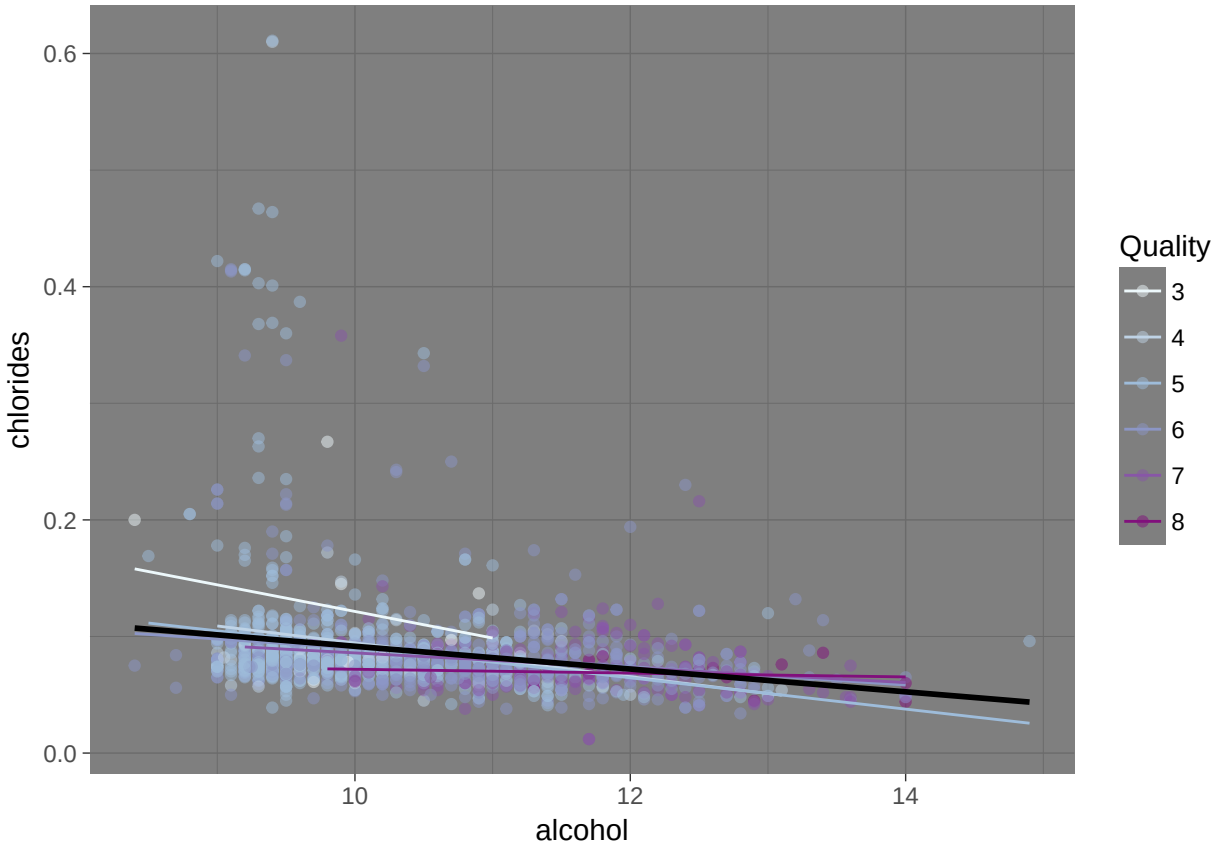
- Volatile acidity seems to decrease with citric acid irrespective of quality. The lowest quality wine 3 seems to show the sharpest decline.

quality, chlorides vs. density



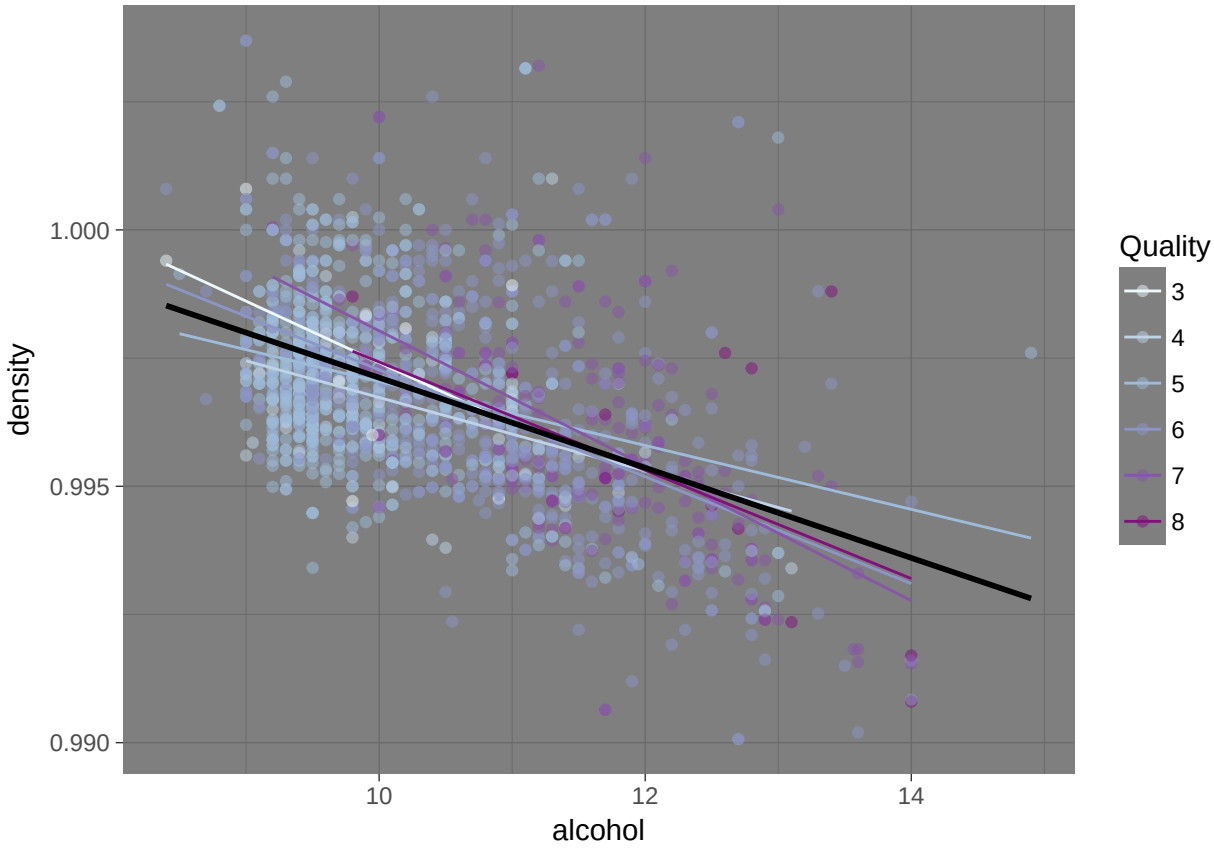
- Chlorides doesn't seem to be highly correlated with density. It also shows varying trend across different quality wines.

quality, chlorides vs. alcohol



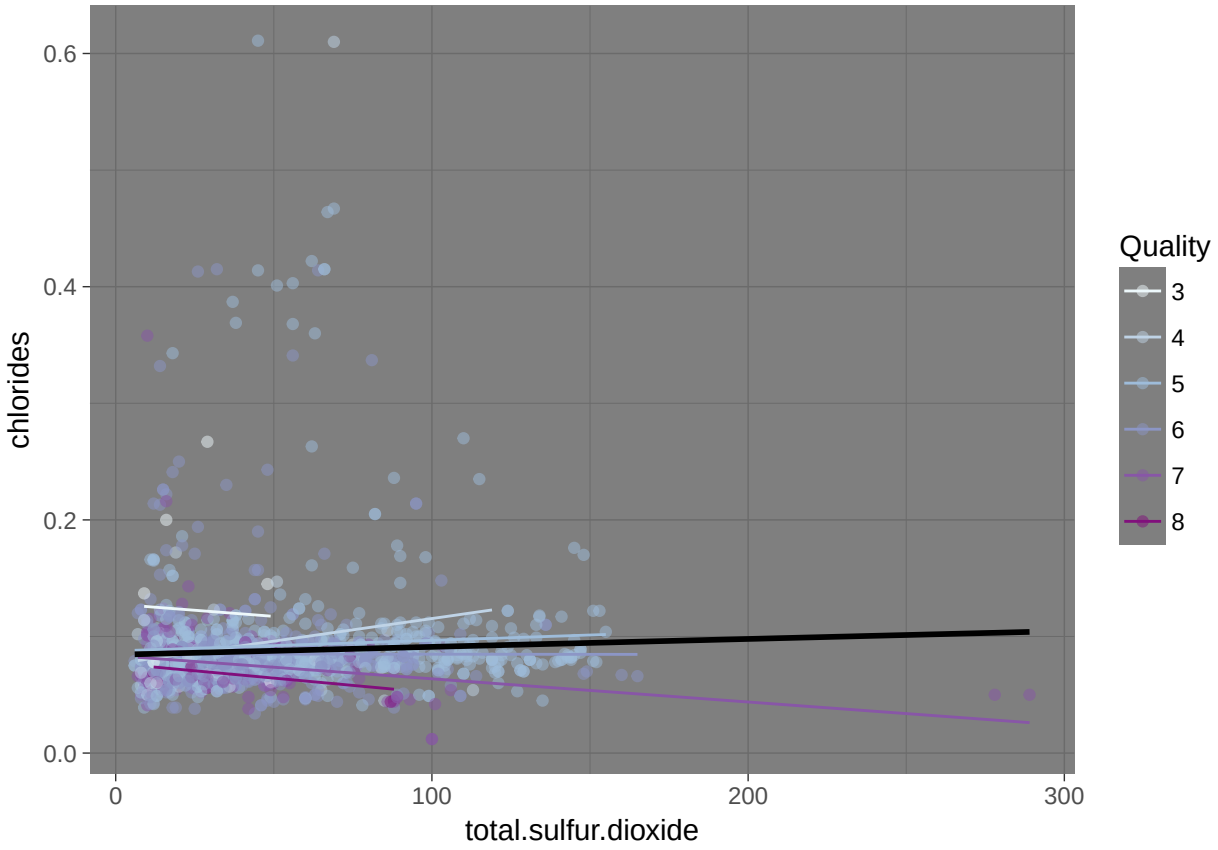
- Variation of chlorides seems to be independent of alcohol content irrespective of the quality of wines.

quality,density vs. alcohol



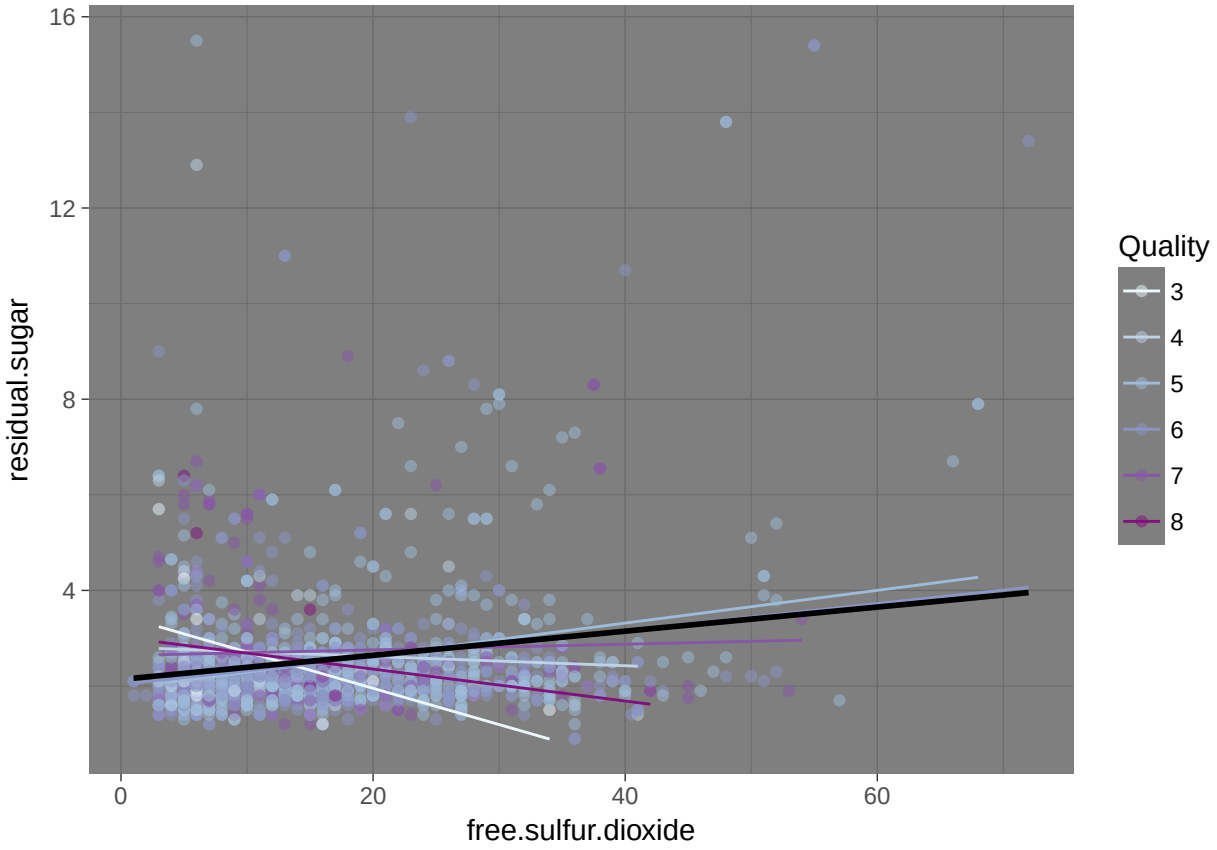
- Desitiy seems to decrease with alcohol content accross all quality wines. This can be attributted to the lower density of alcohol compared to water.
- Data points with density > 1.0 seems a bit suspicious and has to be rechecked.

quality, chlorides vs. total sulphur dioxide



- Chlorides vary independently of total sulphur dioxide across all the wine qualities and can be used as an independent parameter in the predictive model.
- Most of the outliers in chlorides vs. density / alcohol / total sulphur dioxide plot seems to be from low and medium quality wines (5,6). This could be due to the higher sampling rate / availabilty of the medium quality wines as well as the incosistency of prepartion process. Higher quality wines go thrugh more stringent preparation protocol resulting in the consistency. Moreover, chlorides seems to be almost absent in high qulaity wines (7, 8).

quality, residual sugar vs. free sulphur dioxide



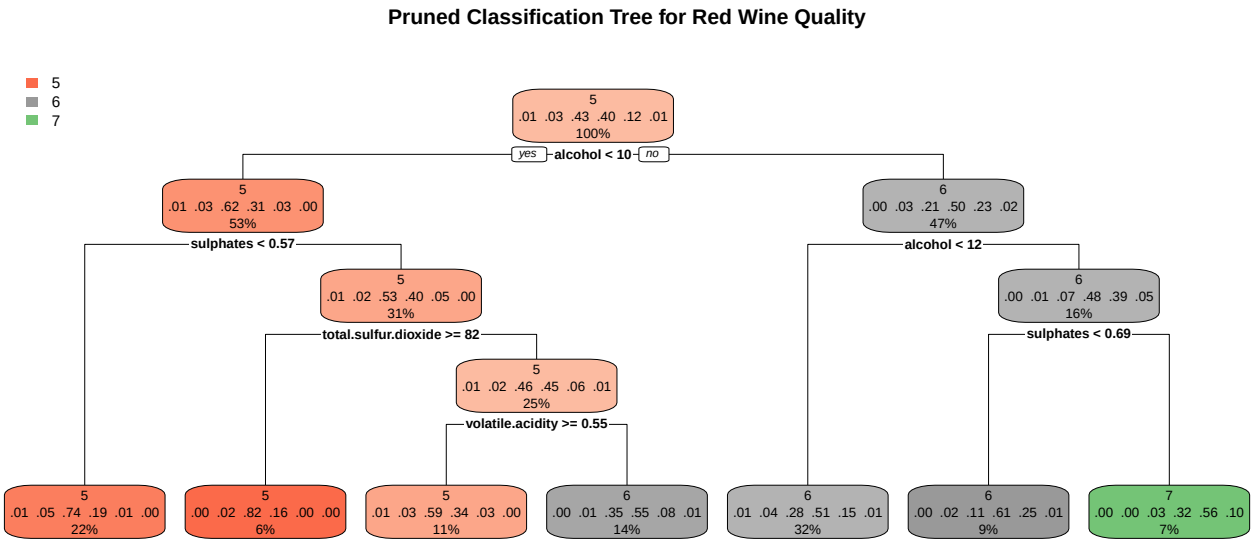
- Residual sugar doesn't seem to be correlated with free sulfur dioxide for any given quality.

Predictive Model

Multicolinearity

- Since the variables citric acid, fixed acidity and pH are highly correlated with each other, fixed acidity alone can be used in the predictive model to reduce redundancy.
- Simillarly, free sulphur dioxide can be dropped as well since it is highly correlated with total sulphur dioxide.
- Another correlation is between alcohol and density. DEnsity can be avoided while making a predictive model.

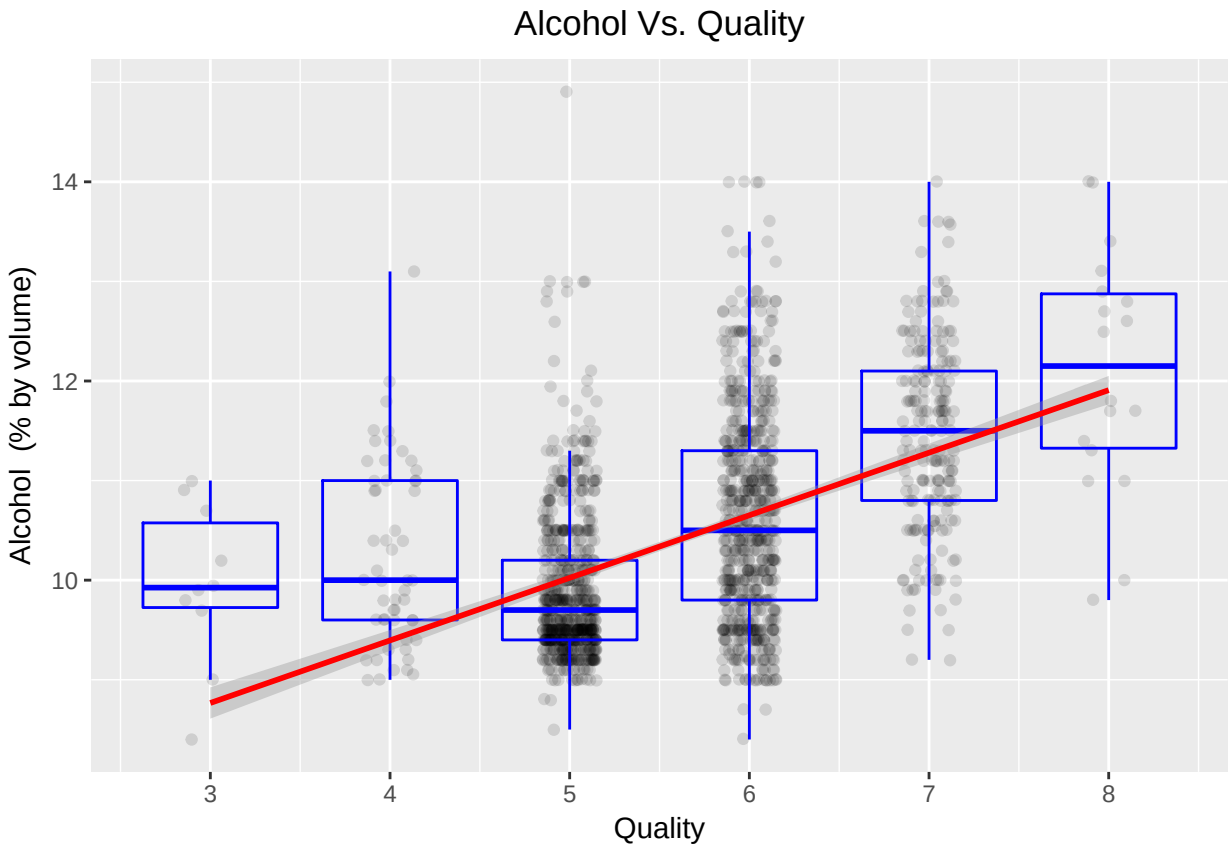
Tree based model



Although the predictive model includes the outcome for all quality 1-10, we have only shown a prunned version due to space constraints. The model clearly shows the significance of alcohol content in the initial partitioning of the tree diagram.

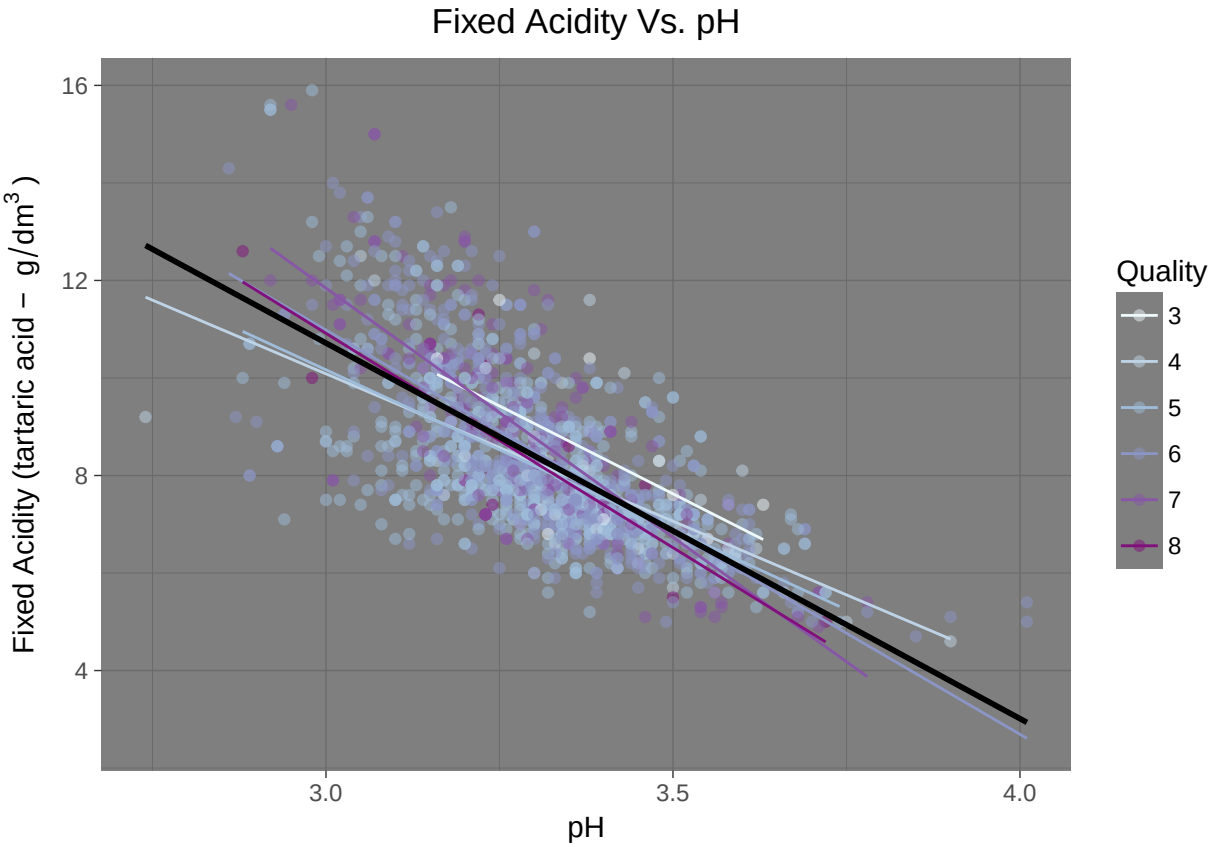
Final Plots and Summary

Plot One



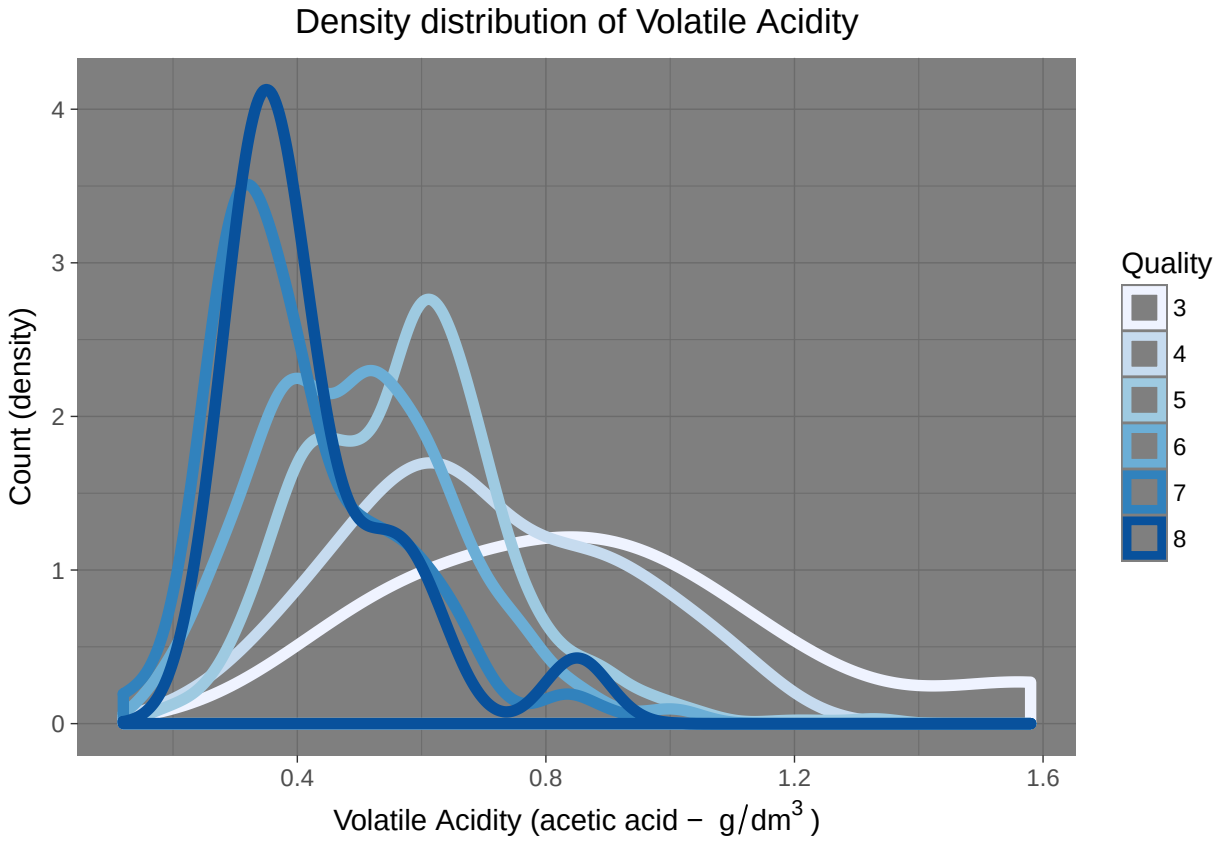
- Median values for alcohol also showed an upward trend but with a dip for intermediate quality (quality = 5).
- We found a strong correlation between alcohol content and quality. Alcohol perhaps may be the most likely indicator of the quality of the wine followed by sulphates. The influence of alcohol content was visibly pronounced in the predictive model.

Plot Two



- The negative correlation of fixed acidity with pH is due to the scale with which pH is defined.
- Fixed acidity is has strong negative correlation with pH which expected since pH is a measure of acidity. But more intriguiging factor is the extent to which fixed acidity (tartaric acid) influences pH when other acids are also present in wine. One of the probably reasons for this high correlation is because fixed acidity is also a measure of non volatile acidic content.

Plot Three




```
## # A tibble: 6 × 3
##   quality       Skew         Ku
##     <int>      <dbl>      <dbl>
## 1       3  0.8850968 -1.0898173
## 2       4  1.6198326  3.3427853
## 3       5  0.5222718 -0.5163235
## 4       6  0.2213311 -0.9880063
## 5       7 -0.3755223 -0.4595533
## 6       8 -0.3260505 -0.9180710
```

- Volatile acidity of high quality wines have narrow distribution compared to more broadly distributed low quality wines.
- Mode of the volatile distribution is decreasing with quality and the distribution is getting narrower with increasing quality as well.
- Higher quality wines go thrugh more stringent preparation protocol resulting in the consistency and hence the narrow distribution of the volatile acidity(acetic acid), a major factor controlling the quality.

Reflections

The redwine dataset containes 1599 observations by 12 variables. I started digging the dataset by looking at the distribution of individual variables and then by probing the relation through 2 variable plots and multivariable plots.

Struggles and successes

The relationships between most of the variables are not apparent until it is filtered based on another independent variable. It was also not obvious to me that quality was an output variable based on other continuous variables in the dataset.

The lack of 'NA' values in the dataset made it much easier to hadle data without using many filters.

There are multiple variables which are related among each other and showed multicolinearity. these variables ahave to dropped while building a predictive model to reduce the redendancy.

Another struggle was to visualize the huge data set based on three variables other than the variable quality.

Future work

A visualization based on three continuous variables could identify interestings trends. Another idea would be to cut off the data in both tails of the distribution and look at the correlation between variables in the middle values.

References

1. http://ggplot2.tidyverse.org/
2. http://ggplot2.org/
3. http://rmarkdown.rstudio.com/
4. https://briatte.github.io/ggcorr/
5. https://onlinecourses.science.psu.edu/stat501/node/343
6. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html
7. https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt
8. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html
9. http://ggplot2.tidyverse.org/reference/geom_density.html
10. https://stats.stackexchange.com/questions/18844/when-and-why-should-you-take-the-log-of-a-distribution-of-numbers
11. http://stackoverflow.com/questions/15736370/special-characters-and-superscripts-on-plot-axis-titles
12. https://onlinecourses.science.psu.edu/stat501/node/343
13. http://www.statmethods.net/advstats/cart.html
14. http://www.milbo.org/rpart-plot/prp.pdf
15. http://blog.revolutionanalytics.com/2013/06/plotting-classification-and-regression-trees-with-plotrpart.html

---