An investigation on the Quality of red wine

# Introduction

This analysis is about the quality of red wine, the varieties and involving their chemical properties as well as the ranking by tasters. The pricing of wine depends on a rather abstract concept of wine appreciation by wine testers, opinion among whom may have a high degree of variablitly. Another key factor is in wine certification and quality assessment are physiochemical tests, laboratory based. It takes into account for example the ph-score and chlorides amongst others. One interesting question in this context is, if there is a correlation between chemical properties and the human taste.

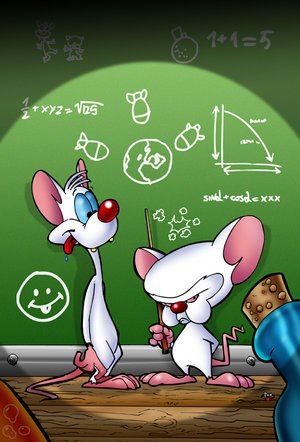
(source: Penn State Eberly College of Science)

##### What are we going to do tonight, Brain?

##### Same thing we do every night, Pinky.

###### oh, veyh Pinky and Brain are analysing the red wine data and plot

how to rule the world with red wine???



pbbottle.

Objective of the analysis will be to predict the Quality Ranking from the chemical properties of the red wine using explanatory data analysis (EDA) to explore the relationships between the variables: visualisation, distributions, outliers and potential anomalies. This project is prepared with R Studio.

The dataset contains 13 variables and 1599 observations.

Description of the variables (based on pysiochemical testing)

fixed acidity: most acids involved with wine volatile acidity: the amount of acetic acid, too high levels can lead to unpleasnt taste (vinegar like ) citric acid: can add ‘freshness’ to the wine residual sugar: amount of sugar after fermentation stops chlorides: amount of salt free sulfur dioxide: free from of SO2, prevents microbiological growth and oxidation total sulfur dioxied: amount of free and bound SO2, becomes evident in nose and taste density: water depending on the percent of alcohol ph: describes how acid or basic the wine is sulphates: additive, antimicrobial and antioxidant alcohol: percent of alcohol quality: output variable, sensoric data

## X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides  
## 1 1 7.4 0.70 0.00 1.9 0.076  
## 2 2 7.8 0.88 0.00 2.6 0.098  
## 3 3 7.8 0.76 0.04 2.3 0.092  
## 4 4 11.2 0.28 0.56 1.9 0.075  
## 5 5 7.4 0.70 0.00 1.9 0.076  
## 6 6 7.4 0.66 0.00 1.8 0.075  
## free.sulfur.dioxide total.sulfur.dioxide density pH sulphates alcohol  
## 1 11 34 0.9978 3.51 0.56 9.4  
## 2 25 67 0.9968 3.20 0.68 9.8  
## 3 15 54 0.9970 3.26 0.65 9.8  
## 4 17 60 0.9980 3.16 0.58 9.8  
## 5 11 34 0.9978 3.51 0.56 9.4  
## 6 13 40 0.9978 3.51 0.56 9.4  
## quality  
## 1 5  
## 2 5  
## 3 5  
## 4 6  
## 5 5  
## 6 5

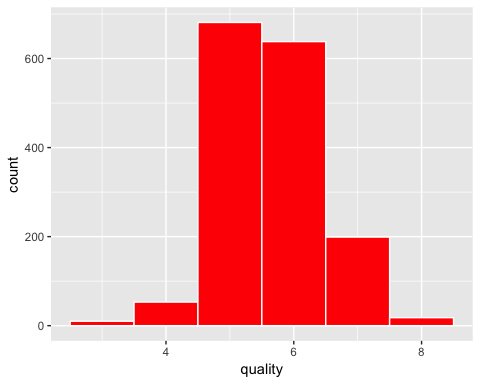
library(ggplot2)

This first summary gives us a first insight. The range of possible socres is 0 to ten. The min in our data set is 3, the max is 8, the median is 6 and the mean is 5.64.

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

The summary function is giving us as well some good information on the diffrent variable data.

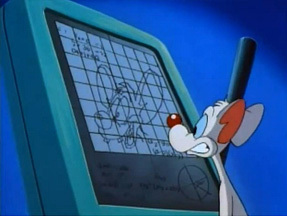
## X fixed.acidity volatile.acidity citric.acid   
## Min. : 1.0 Min. : 4.60 Min. :0.1200 Min. :0.000   
## 1st Qu.: 400.5 1st Qu.: 7.10 1st Qu.:0.3900 1st Qu.:0.090   
## Median : 800.0 Median : 7.90 Median :0.5200 Median :0.260   
## Mean : 800.0 Mean : 8.32 Mean :0.5278 Mean :0.271   
## 3rd Qu.:1199.5 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

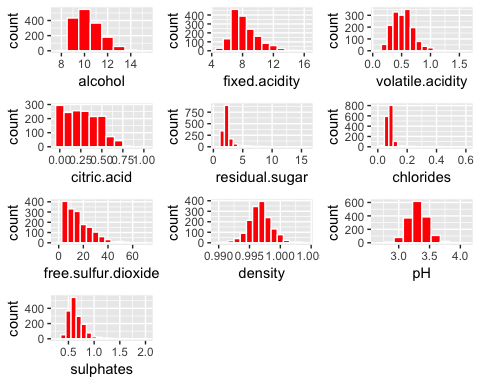


The histogramm shows normal distributed data for the variable quality. As we have seen before the median is 6 and the mean is 5.64 which is really near to the median and the graph perfectly reflects this analysis.

In the next steps I am going to investigate more deeply on the different variables and will check if there are outliers.

# Univariate Plotting

 Pinky is beginning the analysis on the different variables. What will be the outcome?



#### What is the structure of your dataset?

The dataset contains 13 variables and 1599 observations.

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

I think the most obvious features are quality, alcohol rate and sugar level. The more advanced wine consumer might be also interested in ph socre and the other chemical features.

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

The more advanced wine consumer might be also interested in ph socre and he other chemical features.

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

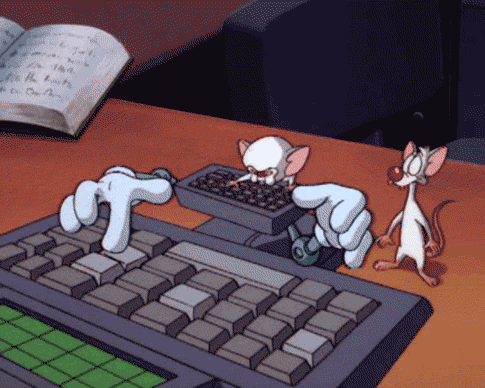
yes, I created a rating variable

##### 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?

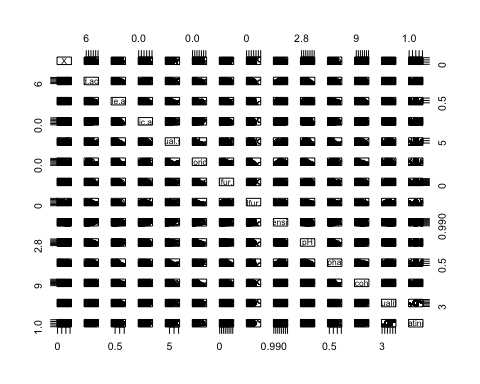
The alcohol rate and free sulfur dioxide are left skewed. Density and pH score are normally distributed. The alcohol content seems to vary from 8 to 14 with major peaks around 10 with a lower count between 13 and 14. The pH value seems to dispaly a normal distribution with major samples exhibiting values between 3.0 and 3.5 We find a normal distribution on the quality and the worst wine and the great one might be outliers. Most of the wines can be considered as average ones.

# Bivariate Plots Section

As my/ Brain’s main interest is in the quality, it will be interesting to check the correlation of the variables espcially with quality. But it might be also be interesting to check the chemical variables against each other.

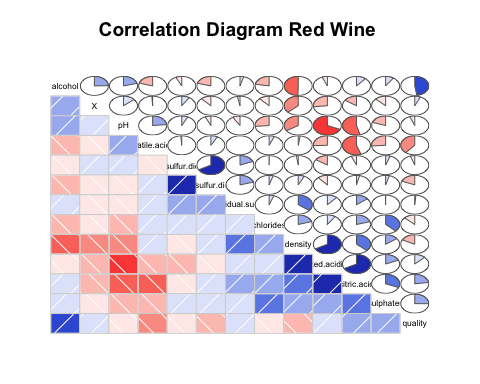


type.



Pearson’s correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. (source <https://www.statisticssolutions.com/pearsons-correlation-coefficient/>)

A graphic solution:



testing example for correlation with the pearson method

##   
## Pearson's product-moment correlation  
##   
## data: rw$quality and rw$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: rw$quality and rw$residual.sugar  
## t = 0.5488, df = 1597, p-value = 0.5832  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.03531327 0.06271056  
## sample estimates:  
## cor   
## 0.01373164

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':  
##   
## combine

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

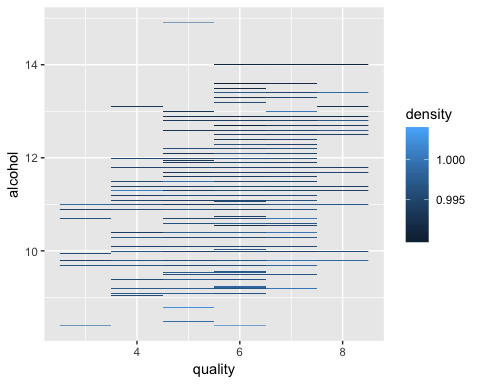
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

##   
## ---------------------------------------------------------------------------  
## &nbsp; fixed.acidity volatile.acidity citric.acid   
## -------------------------- --------------- ------------------ -------------  
## \*\*fixed.acidity\*\* 1 -0.2561 \*\*0.6717\*\*   
##   
## \*\*volatile.acidity\*\* -0.2561 1 \*\*-0.5525\*\*   
##   
## \*\*citric.acid\*\* \*\*0.6717\*\* \*\*-0.5525\*\* 1   
##   
## \*\*residual.sugar\*\* 0.1148 0.001918 0.1436   
##   
## \*\*chlorides\*\* 0.09371 0.0613 0.2038   
##   
## \*\*free.sulfur.dioxide\*\* -0.1538 -0.0105 -0.06098   
##   
## \*\*total.sulfur.dioxide\*\* -0.1132 0.07647 0.03553   
##   
## \*\*density\*\* \*\*0.668\*\* 0.02203 \*\*0.3649\*\*   
##   
## \*\*pH\*\* \*\*-0.683\*\* 0.2349 \*\*-0.5419\*\*   
##   
## \*\*sulphates\*\* 0.183 -0.261 \*\*0.3128\*\*   
##   
## \*\*alcohol\*\* -0.06167 -0.2023 0.1099   
##   
## \*\*quality\*\* 0.1241 \*\*-0.3906\*\* 0.2264   
## ---------------------------------------------------------------------------  
##   
## Table: Table continues below  
##   
##   
## ------------------------------------------------------------------------------  
## &nbsp; residual.sugar chlorides free.sulfur.dioxide   
## -------------------------- ---------------- ------------ ---------------------  
## \*\*fixed.acidity\*\* 0.1148 0.09371 -0.1538   
##   
## \*\*volatile.acidity\*\* 0.001918 0.0613 -0.0105   
##   
## \*\*citric.acid\*\* 0.1436 0.2038 -0.06098   
##   
## \*\*residual.sugar\*\* 1 0.05561 0.187   
##   
## \*\*chlorides\*\* 0.05561 1 0.005562   
##   
## \*\*free.sulfur.dioxide\*\* 0.187 0.005562 1   
##   
## \*\*total.sulfur.dioxide\*\* 0.203 0.0474 \*\*0.6677\*\*   
##   
## \*\*density\*\* \*\*0.3553\*\* 0.2006 -0.02195   
##   
## \*\*pH\*\* -0.08565 -0.265 0.07038   
##   
## \*\*sulphates\*\* 0.005527 \*\*0.3713\*\* 0.05166   
##   
## \*\*alcohol\*\* 0.04208 -0.2211 -0.06941   
##   
## \*\*quality\*\* 0.01373 -0.1289 -0.05066   
## ------------------------------------------------------------------------------  
##   
## Table: Table continues below  
##   
##   
## -----------------------------------------------------------------------------  
## &nbsp; total.sulfur.dioxide density pH   
## -------------------------- ---------------------- ------------- -------------  
## \*\*fixed.acidity\*\* -0.1132 \*\*0.668\*\* \*\*-0.683\*\*   
##   
## \*\*volatile.acidity\*\* 0.07647 0.02203 0.2349   
##   
## \*\*citric.acid\*\* 0.03553 \*\*0.3649\*\* \*\*-0.5419\*\*   
##   
## \*\*residual.sugar\*\* 0.203 \*\*0.3553\*\* -0.08565   
##   
## \*\*chlorides\*\* 0.0474 0.2006 -0.265   
##   
## \*\*free.sulfur.dioxide\*\* \*\*0.6677\*\* -0.02195 0.07038   
##   
## \*\*total.sulfur.dioxide\*\* 1 0.07127 -0.06649   
##   
## \*\*density\*\* 0.07127 1 \*\*-0.3417\*\*   
##   
## \*\*pH\*\* -0.06649 \*\*-0.3417\*\* 1   
##   
## \*\*sulphates\*\* 0.04295 0.1485 -0.1966   
##   
## \*\*alcohol\*\* -0.2057 \*\*-0.4962\*\* 0.2056   
##   
## \*\*quality\*\* -0.1851 -0.1749 -0.05773   
## -----------------------------------------------------------------------------  
##   
## Table: Table continues below  
##   
##   
## -------------------------------------------------------------------  
## &nbsp; sulphates alcohol quality   
## -------------------------- ------------ ------------- -------------  
## \*\*fixed.acidity\*\* 0.183 -0.06167 0.1241   
##   
## \*\*volatile.acidity\*\* -0.261 -0.2023 \*\*-0.3906\*\*   
##   
## \*\*citric.acid\*\* \*\*0.3128\*\* 0.1099 0.2264   
##   
## \*\*residual.sugar\*\* 0.005527 0.04208 0.01373   
##   
## \*\*chlorides\*\* \*\*0.3713\*\* -0.2211 -0.1289   
##   
## \*\*free.sulfur.dioxide\*\* 0.05166 -0.06941 -0.05066   
##   
## \*\*total.sulfur.dioxide\*\* 0.04295 -0.2057 -0.1851   
##   
## \*\*density\*\* 0.1485 \*\*-0.4962\*\* -0.1749   
##   
## \*\*pH\*\* -0.1966 0.2056 -0.05773   
##   
## \*\*sulphates\*\* 1 0.09359 0.2514   
##   
## \*\*alcohol\*\* 0.09359 1 \*\*0.4762\*\*   
##   
## \*\*quality\*\* 0.2514 \*\*0.4762\*\* 1   
## -------------------------------------------------------------------



sorry.

after all this numbers, time for a little bit of fun :-)



# 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?

Alcohol has negative correlation with density. This is expected as alcohol is less dense than water.

Residual.sugar does not show correlation with quality. Free.sulfur.dioxide and total.sulfur.dioxide are highly correlated as expected.

Density has a very strong correlation with fixed.acidity.

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

Volatile.acidity has a positive correlation with pH. This is unexpected as pH is a direct measure of acidity.

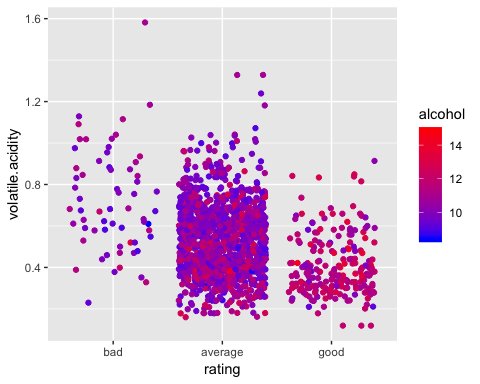
### What was the strongest relationship you found?

The variables that have the strongest correlations to quality are volatile.acidity and alcohol.

# Multivariate Plots Section / 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?

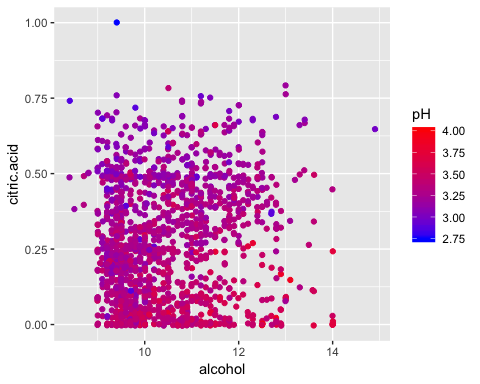
I think it makes sense to check first on the strongest correlation, which is the one between alcohol and volatile.acidity.

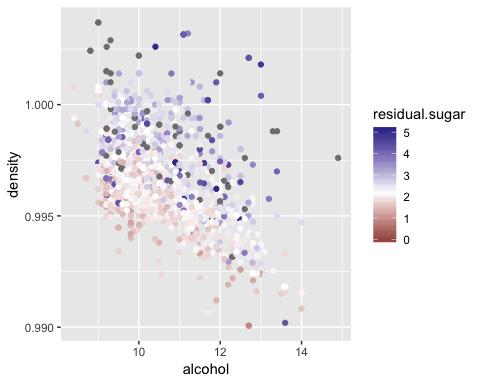


Hm, it looks as if alcohol rate is higher in good wines. But not yet any real pattern visible despite that.

## 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?

Next step to try is using the ph-score as this is realted with acidity.

 I googled the relation between the ph- socre and citirc acid: H3Citrate citiric acid, C6H8O7 (citirc acid) 3.24, I found that citric acid is used to regulate the ph- score (in cosmetic products) and is an anti- oxidant. (source: wikipedia) From my opinion this graph is underlining the correlation between the ph score and the citric acid addition to a wine.

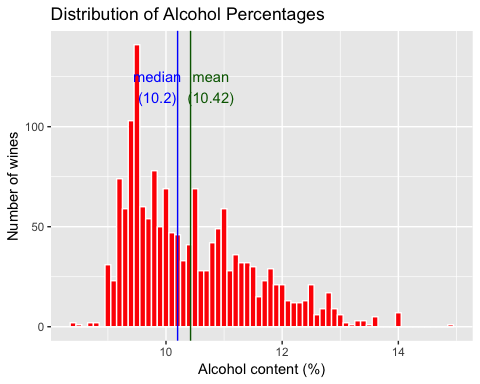


### Were there any interesting or surprising interactions between features?

The density decreases as the alcohol increases. What may not be as obvious is that the density increases as sugar increases, but it’s along the opposite direction. The median of the residual sugar lies parallel to the alcohol vs density trendline. Interesting as I had no knowledge about wine before doing this analyis.

# Final Plots and Summary

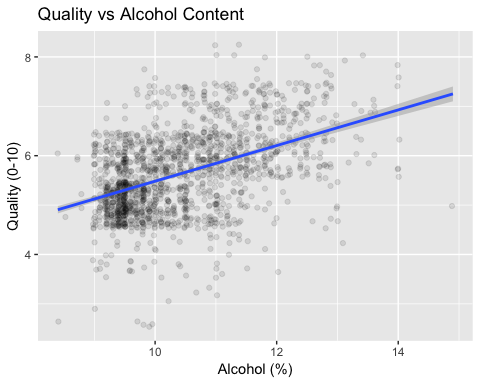
### Plot One

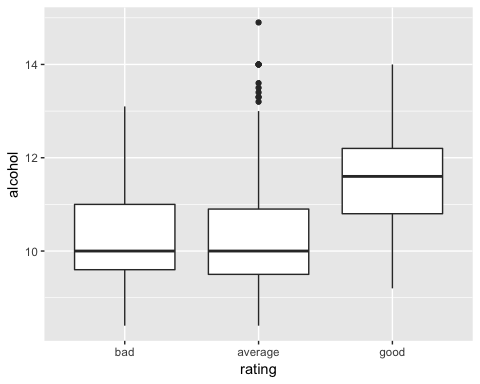


### Description One

We see here the right skewed distribution of the alcohol and the number of wines, the mean (10,42%) and the median (10.2%) are realtively close.There are not many wines which have a percentage higher than 12%.

### Plot Two

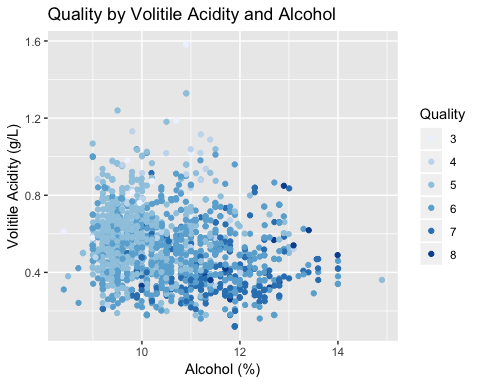


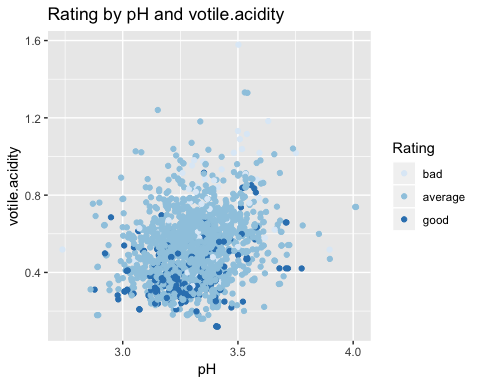


### Description Two

The boxplot is devlivering less information than the scatterplot above, but it is more compact and they can be easily compared. Here we can see that good wines have a higher level of alcohol. Both pictures are to be seen complementary to underline the correlation between quality / rating and alcohol content.

### Plot Three





### Description Three

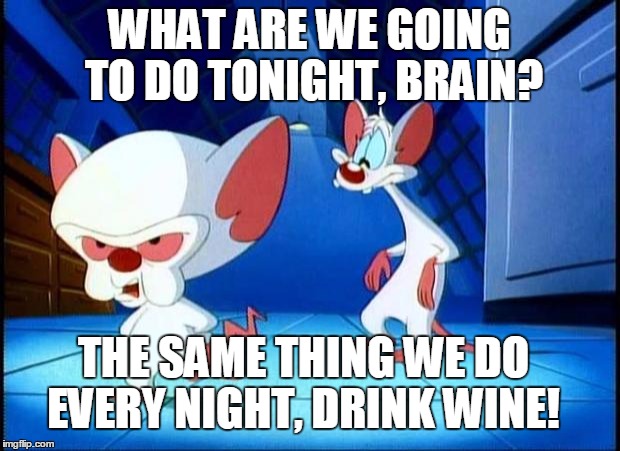
This chart shows how quality improves as the alcohol content increases and the volitile acidity decreases.The overall trend of the colors getting darker as they go to the bottom right. The second plot shows the relation of the rating with votile acidity and pH score. I was interested in the result because normally acidity and pH have a strong correlation.

# Reflection

As I do not know much about chemistry, my analysis is clearly limited and there is potential to take more care of the correlations between them. My analysis is concentrated clearly on quality / ranking - a variable I built, and the density and level of alcohol. It shows there is a strong link between them. According to my analyis a good wine has a level of alcohol around 10% and 10.5%. From the the final plot 1 we can say a good wine has a votile acidity 0.4. The additional checking of votile acidity and pH together with quality did not give much more insight.

I had a struggle with finding a good method of correlating the variables whereas the first steps went pretty ok. I did not pay much attention on outliers. All in all it was fun to step into something really new as I never had any touch with R. I learned that R is a very powerful tool in explanatory data analysis and give pretty surprising insights of an anknown topic.

The topic of building a model could be one of further investigation. I am not sure how to choose the best variables for a predictive model. On top my knowledge of wine is so limited that I would like to know which acids could explain the variation in the pH score as I think the three listed ones do not explain well enough.



night.

And on top I hope you enjoy the little side story of Pinky and Brain which is my favourite comic show. When I am working together at work with a programmer we always make the joke that we are Pinky and Brain, never sure, who is who.