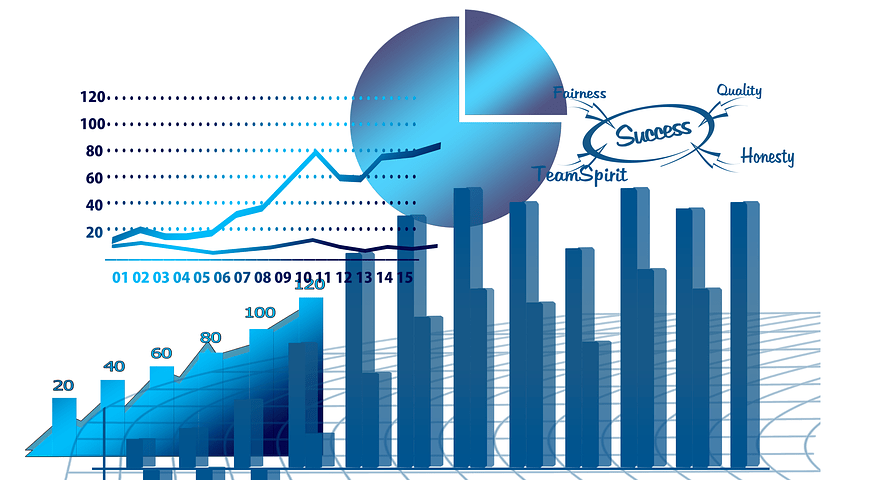
[Data Science Demonstrated](http://datasciencedemonstrated.com/)

Exploring data using the tools of data science



**Data Analysis Nanodegree 05: Exploratory Data Analysis**

by [Simon Thornewill von Essen](http://datasciencedemonstrated.com/author/admin/) on March 8, 2018 in [Udacity Projects](http://datasciencedemonstrated.com/category/udacity-projects/)

**Introduction**

Exploratory Data Analysis (EDA) is the process of investigating a data-set for insights. The analysis can be conducted on a uni-variate, bi-variate and multi-variate scale to generate an increasingly complex understanding for the secrets that this data-set holds. It is also useful when conducting machine learning as it helps with gaining a general understanding of the data and to quickly detect outliers. The authors of “Elements of Statistical Learning”, Hastie and Tibshirani, note the use of EDA in their first [online lecture](https://youtu.be/2wLfFB_6SKI?t=313) on machine learning. Understanding your data is paramount to performing good data science.

**R and the tidy-verse**

I mentioned in a previous [post](http://datasciencedemonstrated.com/2017/10/25/data-analysis-nanodegree-02-us-2016-bike-share-analysis/) how I would like to do an analysis in R to show how it works. There is often some debate between which tool to use when performing data science/analysis regarding Python or R. The general consensus is that both packages are very good and worth learning, although I would start with learning Python since it depends less on [vectorised](https://stackoverflow.com/questions/1422149/what-is-vectorization) functions than R does, teaches you more about the basics of programming and has an increased scope of uses. On the other hand, understanding how the tidy-verse within R works can be immensely useful in handling and visualising data efficiently and practically, python cannot compete on this specialized field as easily. The tidy-verse is a compilation of R libraries such as ggplot2, purrr, dplyr, tidyr and others. A good Book over how to perform analyses in R can be found [here](http://r4ds.had.co.nz/).

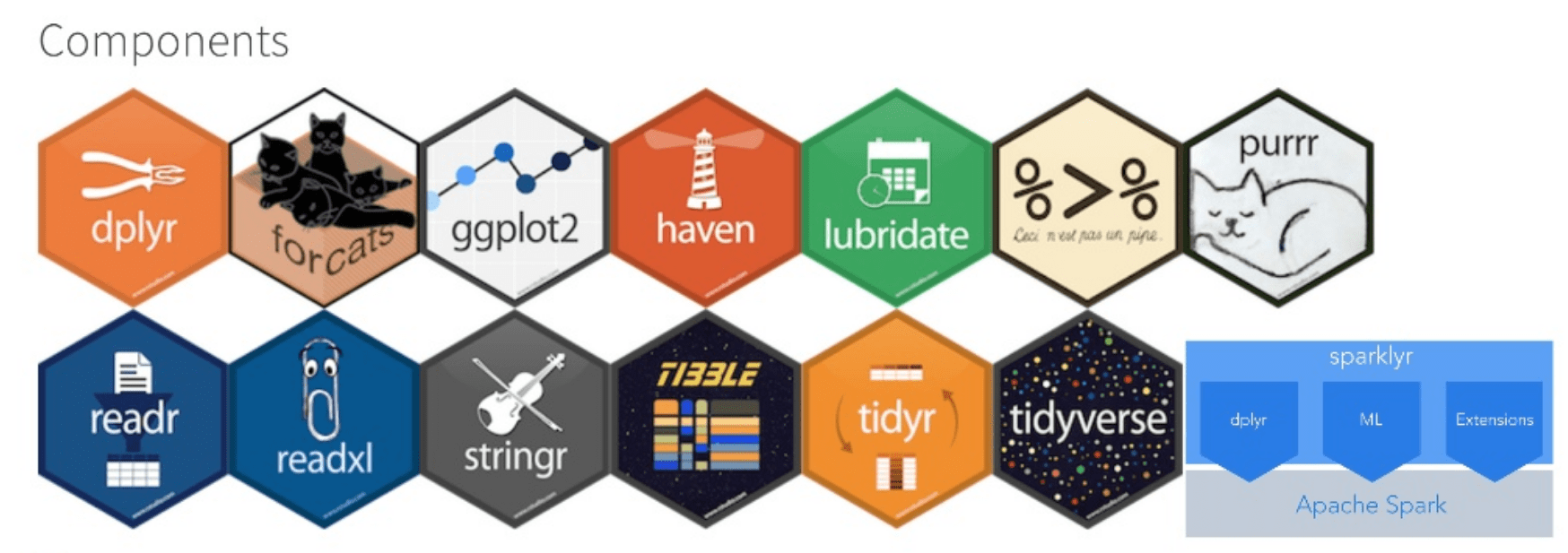
[](https://www.slideshare.net/databricks/extending-the-r-api-for-spark-with-sparklyr-and-microsoft-r-server-with-ali-zaidi)

Figure 1. R packages within the tidy-verse. These packages are written with a consistent logic which helps to transform, visualise and analyze data and is the back-bone of using R for data science.

One package that I will be using extensively in this example is ggplot2, which is one of the most powerful visualisation packages available. What makes ggplot2 so powerful is that it gives precise control different aspects of a visualisation to create the exact desired output in a way that might be difficult using another language such as python.

**Method – Exploratory Data Analysis**

**Uni-variate Analysis**

As mentioned above, there are various stages to data analysis. The first stage is uni-variate analysis, which concerns itself with understanding the distributions and typical values (mean, interquartile range, etc.) of a dataset. This is important to understand because it will help create initial questions and general insights which will help to discover where to add extra attention for analysis. The typical graphs that are used in this analysis are histograms (continuous variables) and bar-graphs (qualitative variables) because they can quickly capture the distribution of a feature.

**Bi-variate Analysis**

After the uni-variate analysis is complete, some insights and questions will have been derived. Normally when doing an analysis we want to measure an output variable. For instance, if we were doing analysis within a business context then *revenue* would be a usual output variable. Each of the variables measures in the uni-variate analysis are now measured against this one metric using scatter-plots, box-plots and violin-plots. Correlations between input features can also be investigated.

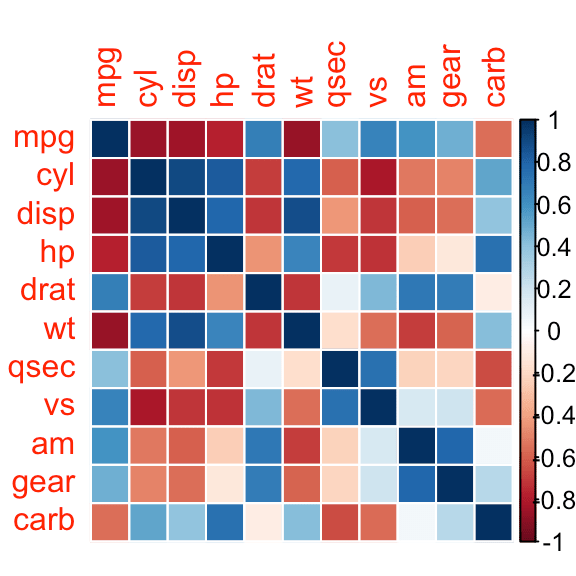


Fig 2. Example of a heat-map. Features that have strong correlations are highlighted with blue and strong negative correlations are outlined in red. This allows an analyst to quickly judge the features in a dataset to see which features are worth investigating further within the bi-variate analysis.

Lots of analysts often use matrix visualisations such as heat-maps and correlation-matrices in order to quickly discern interesting patterns to investigate. This can be done in R using the [ggally](https://cran.r-project.org/web/packages/GGally/index.html) package. There are two disadtanges to this approach:

1. It can require a lot of computational resources to create a matrix, weaker computers can take up to an hour to make a plot.
2. The graphical outputs are at a lower resolution and will sometimes hide trends in the data.
3. There is a lot of information within one graph making it difficult to digest

I would only correlation-matrices after performing some prior investigation. (I do not use correlation-matrices in this investigation, I just thought they were worth mentioning for their use and popularity.)

**Multi-variate Analysis**

After completing the bi-variate analysis, the analyst should have yet more questions and insights from which to take the full plunge into multi-variate analysis. This means taking a bi-variate visualization and adding extra details so that we can see how the effects of multiple features interact with one another. During this phase we try to get as much information as possible from the data, and can help us to arrive at conclusions that would be incredibly difficult to find by just looking at the data including hypothesis on causal relationships which would need to be confirmed later.

**Explanatory Data Analysis**

Finally, when the analysis is complete, the most important plots are taken and optimised based on what findings should be communicated to the target audience. If you cannot communicate your findings simply and clearly then that will hurt your future prospects and the possibility that the stakeholders will be able to benefit from your hard work. There are great books on this topic and if you would like to learn more I reviewed a book from Cole Nussbaumer Knaflic named “[Storytelling with Data](http://datasciencedemonstrated.com/2018/01/21/storytelling/)” which I strongly encourage anyone interesting in this field reads.

**Project Objective**

For this post, I will be outlining the analysis I performed on data-set describing the chemical composition of [white wine quality](https://archive.ics.uci.edu/ml/datasets/wine+quality) from The UCI Machine Learning Library, which refers to work done by [Cortez, *et. al*.](http://dx.doi.org/10.1016/j.dss.2009.05.016)

As  usual, I won’t show all analysis and code in this blog post. I my purpose in this blog post is to quickly run through the EDA process, and so that is what I am going to showcase here. To see the full analysis, see the kernel at the bottom of this page or visit the corresponding [GitHub repository](https://github.com/SThornewillvE/Udacity-Project---Exploring-Wine-Data).

**Analysis**

**Uni-variate Example**

Since the purpose of this machine learning dataset is to predict quality using various features such as various acid, salt-chloride and sulfate concentrations, I think that it best to show a bar-plot of quality first in figure 3.

|  |  |
| --- | --- |
| 1  2  3  4  5  6 | ggplot(aes(x=quality), data=df) +   geom\_bar(fill='#FD7373', alpha=0.7) +   scale\_x\_continuous(breaks=seq(0, 10, 1)) +   ggtitle(&quot;Historam of Quality&quot;) +   xlab(&quot;Quality of Wine&quot;) +   ylab(&quot;Count&quot;) |

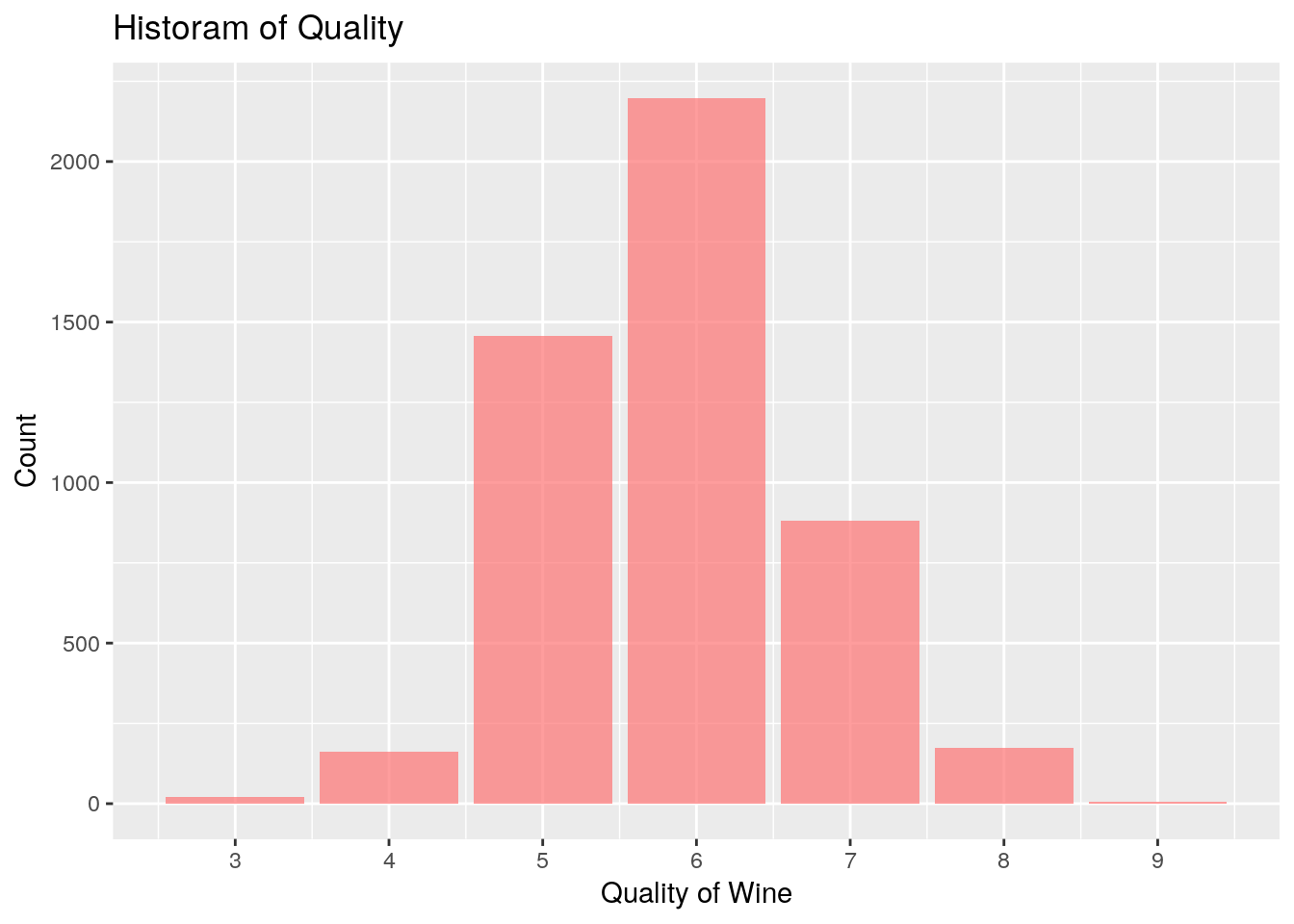


Fig 3. Bar-graph of wine quality rating. The Average rating should be around.5 but it is clearly one point higher than it should be. No wines are given the grade of 1 or 10 either. This shows bias in how the wine was examined.

The first question I asked myself is, how was the quality of the wine measured? This was done by taking the median grade of three experts who tasted the wine. The fact that this metric was measured using a median instead of an average means that wines have a difficult time being graded with an exceptionally high or low grade. This means that the full range between 0 and 10 is not utilised as much as you would expect by a random normal distribution. Furthermore, the average wine quality is a touch higher than the expected average of 5, which means that if one assumes that the wines tastes are representative of all wines then the experts tend to grade wine slightly higher than they should. This is important because if one would launch straight into doing machine learning then these biases might go ignored.

For this analysis, I categorised the “low”, “medium” and “high” quality wines together with bands from 1-4, 5-6 and then 7-10, which made things simpler for further analysis using the code below.

|  |  |
| --- | --- |
| 1  2  3  4  5 | df$quality &lt;- ifelse(df$quality &lt;= 4, 'average', ifelse(df$quality &gt;= 7, 'low',                       'high'))      df$quality &lt;- factor(df$quality, labels=c(&quot;low&quot;, &quot;average&quot;, &quot;high&quot;)) |

**Bi-variate Example**

After performing the uni-variate analysis, questions then needed to be posed. How does alcohol content effect wine quality? This was done by creating various bar/violin plots, one of which can be seen in Figure 4.

|  |  |
| --- | --- |
| 1  2  3  4  5 | ggplot(aes(y=alcohol, x=quality), data=df) +   geom\_violin(aes(fill=quality), alpha=0.7) +   ggtitle(&quot;Quality vs Alcohol Content&quot;) +   xlab(&quot;Quality&quot;) +   ylab(&quot;Alcohol (%)&quot;) |

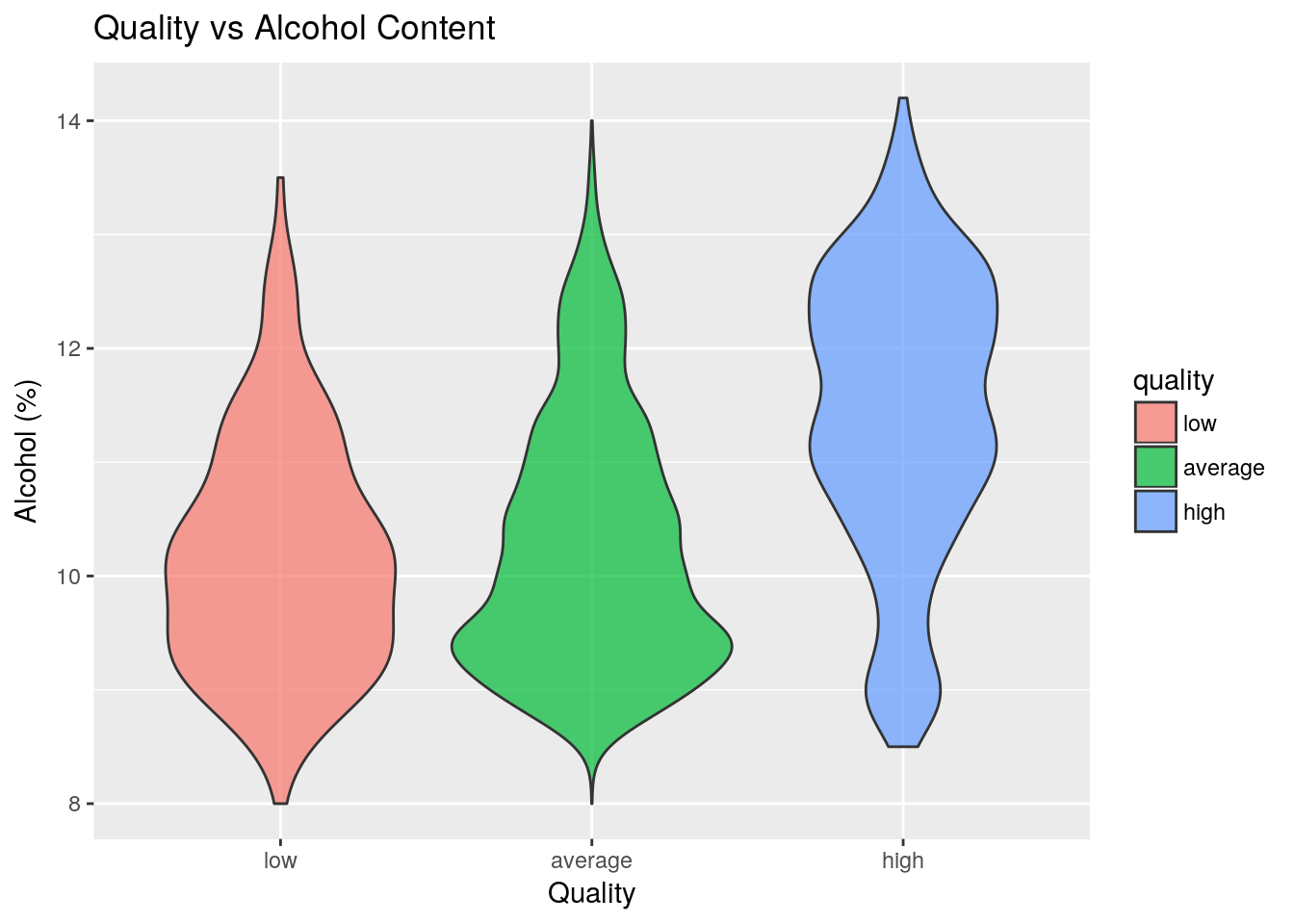


Fig 4. Violin plot of Alcohol content vs quality. The width of the plot shows the frequency of items have the same percentage of alcohol concentration. Note how highly rates wines also tend to have high alcohol content.

We can see that wine with strong alcoholic content tends to be rated higher than wine with low alcoholic content. This is interesting because one might think that a high concentration of alcohol might be seen as “trashy” or otherwise uncouth. It is not uncommon for humans to say one thing and to then do another, this discrepancy between our thoughts and actions is something that is well docuemented in [psychology](https://en.wikipedia.org/wiki/Value-action_gap). On the other hand, it is also known that wine that has been fermenting for a long time is also perceived as being high in quality. So this might be a strong deciding factor between the two that was not measured.

It should be noted, however, that this is only a correlation and hence there are likely to be other factors that come into play. When one ignores the high quality wine then it seems that low quality wine does actually have higher alcohol content. This reinforces the point that there is more going on here that cannot be easily understood by simply making two plots and would be a “point of attack” for further investigation.

**Multi-Variate Example**

Another feature of this data-set is the residual sugar content, the concentration of sugar remaining after fermentation. Alcohol is produced by the anaerobic fermentation of sugars in environments containing yeast. I would like to investigate the correlation between residual sugar, quality and alcohol content and see how they all affect each other. This is when the multi-variate analysis phase begins.

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7 | ggplot(data=df, aes(x=alcohol, y=residual.sugar)) +   geom\_point(aes(colour=quality), alpha=0.35) +   geom\_smooth(method=&quot;lm&quot;, aes(color=quality)) +   ggtitle(&quot;Residual Sugar vs Alcohol Content&quot;) +   ylab(&quot;Residual Sugar (g/L)&quot;) +   xlab(&quot;Alcohol Concentration (%)&quot;) +   ylim(0,25) |

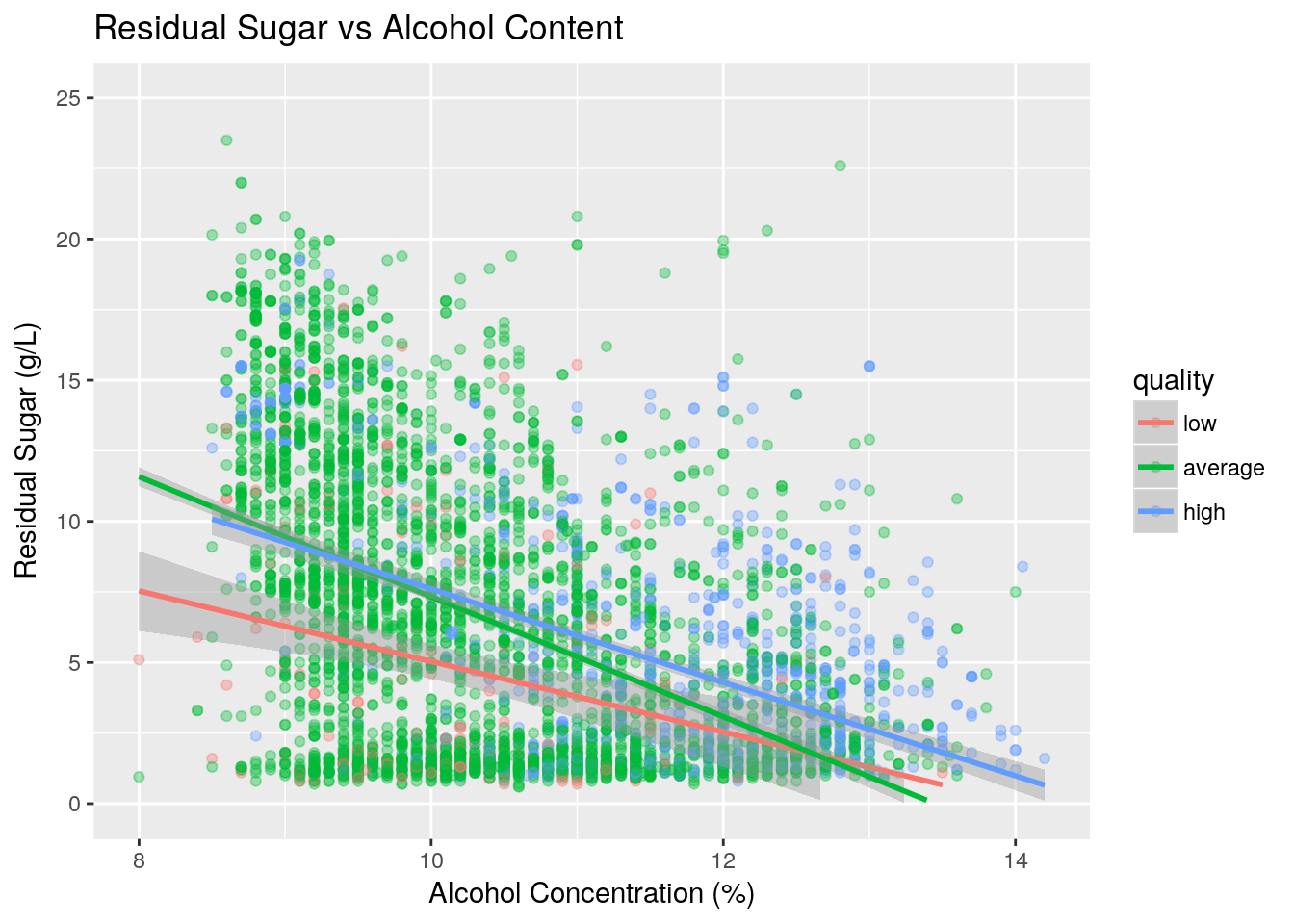


Fig 5. Multi-variate plot showing a scatter plot of wines colored by their rating and summarised using a trend line. Note how residual sugar decreases with alcohol concentration and how high quality wines retain higher levels of sugar.

We can see that residual sugar concentration and alcohol concentration are inversely correlated as seen by the trend lines in Figure 5. When alcohol concentration is high, highly rated wine will have a high residual sugar content. The bias introduced by how the ratings were measured becomes very apparent here, as most of the points on this graph are green although the wines rated as “average” only covered 2/10 ratings. This means that it is difficult to draw conclusions about “low” quality wine because the sample size is relatively small.

Note how in this section the transparency (alpha) of the points have been decreased in order to prevent “over-plotting”, which is what happens when too many points are plotted on a single graph making it difficult to interpret. The colors were chosen in order to give equal “visual weighting”.

**Challenges and Implemented Solutions**

I find the initial steps of exploratory data analysis to be a little overwhelming because because of the number of plots you need to make, understand and synthesise together to find deeper findings. There is a combinatorial explosion that takes place during the bi-varaite or multi-variate analysis, which can make it overwhelming if you do not focus on *why* you are doing the analysis. Hence, it is worth having a specific question that you would like to have answered about your data-set will help direct your attention to only create graphs that contain information and patterns that are relevant for you. Earlier stages of the analysis require some patience but can be achieved by doing this analysis in short bursts and keeping a record of your findings. It is amazing how much progress you make when you concentrate your effort on a single step at a time. (Much like how I wrote this blog post!)

The most important thing is to allow your natural curiosity to guide you. The reason why it is difficult for computers to perform analysis like this is because human intuition and outside knowledge that computers often do not have access to is what makes it easy for us to perform. That is not to say that computers will never be able to do rudimentary analysis or find unintuitive findings otherwise machine learning would not be as popular as it is. The point I want to make is that humans have the ability to understand a wider context that computers will not have access to or even be able to synthesise for the foreseeable future. But beware, there is no dichotomy here! A good data scientist should be able to do both things, but a bad data scientist is surely one that lets the computer do all of the work instead of letting it *augment* what they are capable of.

**Conclusion**

From this blog post I quickly outlined some important packages that one ought to know in R when performing exploratory data analysis. To showcase this, I took one feature and added extra features depending on questions that I had in order to find increasingly complex findings. These findings are as follows:

* The way how the dataset measured “quality” is slightly dubious, lots of wine gets classified as average.
* High concentrations of alcohol and residual sugar tends to help wine quality

I hope this sheds some light onto this process, which can be a lot of fun once you go through the process a number of times. These skills are not overly complex to master but can help make unexpected insights which can help to turn a profit.

Thanks for Reading.

**Kernel**

| **title** | **author** | **date** | **output** |
| --- | --- | --- | --- |
| Udacity-Project---Exploring-Wine-Data | Simon Thornewill von Essen | Janurary 17, 2018 | html\_document |

# Set working directory

setwd("~/coding/git-repos/udacity/Udacity-Project---Exploring-Wine-Data")

# Import Libraries

library(ggplot2)

library(dplyr)

library(tidyr)

library(gridExtra)

# Import data

df <- read.csv("data/wineQualityWhites.csv")

**General notes about the dataset**

Inputs are objective physical/chemical attributes, output (quality) is a qualitative measurement based on the median of 3 seperate taste tasters (0 - Very Bad, 10 - Very Excelent).

There is no data on grape types, wine brand or selling price.

Before doing any serious work on the data I should try and get a feel for it first:

* Read the wineQualityinfo.txt file to understand the features
* Create a dataframe with both red and white wine

**Description of features**

* Fixed acidity: non-volatile acids (tartaric acid) which account for the sour taste of wine - g/L
* Volatile acidity: volatile acids (acetic acid/vinegar) which give wine its nose - g/L
* Citric acid: account for the "fresh" taste - g/L
* Residual Sugar: Remaining sugar after fermentation
* Chlorides: Saltiness of wine - g/L
* Free sulfur dioxide: Prevents alcohol oxidation, $SO\_2$ reacts with oxygen instead of the alcohol to prevent formation of acetic acid (forms 2 eq. of $SO\_3$) - mg/L
* Total sulfur dioxide: - mg/L Bound and free sulfur dioxide
* density: mass/volume of wine - g/mL
* pH: concentration of $H^+$ ions in solution (log scale)
* sulphates: potassium sulphate additive, contributes to Free sulfur dioxide - g/L
* alcohol: % alcohol content
* quality: median measurement between 1 and 10

**What are some questions I have of the data?**

* What variables are correlated with quality?
  + Special interest in "residual sugar", "pH", the various "acidity" features and "alcohol" content
* What does the average glass of white wine look like?
* How varied are tastes in wine? i.e. Are there wines that are clearly superior in quality or is it completely subjective? (random)

**Possible biases**

* There is no information on how intoxicated the testers became. Drunkeness probably plays a role in deciding the quality of the wine.
* Most, but not all compounds present in wine has been listed. Phenols and other [aromatic](https://en.wikipedia.org/wiki/Aromaticity) molecules which would contrubute to a wine's color/taste/smell have not been measured
* This data comes from the taste of connoisseurs, who have might different or more sentisive tastes than that of the general population, there is no way to account for this difference (if it exists) using the data at hand.

**Notes about outliers**

Although there are some outliers that can be clearly seen during the first part of this investigation (Uni-variate plots), these outliers will not be removed because they should not have a large impact upon statistics.

I am aware that there are statistical tests that can be used to remove outliers such as the g-test. However, this test can only remove a single datapoint from a dataset.

Outliers should only really be removed when there is fault with the measurement, investigation of the dataset up until this point does not suggest that this is an issue and as such these points will be left alone.

**Quick Wrangling to fix the data**

Given that "fixed.acidity", "volatile.acidity", "chlorides", are simply concentrations of tartaric acid, acetic acid and salt chorlides respectively, I'll be renameing them. (As a chemist, this makes much more sene to me.)

Density is also interesting because it measures a similar sort of thing to concentration. (i.e. How much stuff is there in an area of space?) Hence, I will also convert the value from g/cm to g/L so that it can be compared more easily to the other values of concentration.

sulfur dioxide concentrations are also measured in mg/L instead of g/L, so I'll change that to make sure the measurements are consistent for comparison across variables.

# drop X from dataframes as it contains redundant information

drops <- c("X")

df <- df[, !names(df) %in% drops]

rm(drops)

# renaming df

df <- rename(df, "tartaric.acid" = "fixed.acidity")

df <- rename(df, "acetic.acid" = "volatile.acidity")

df <- rename(df, "salt.chlorides" = "chlorides")

# correcting sulfur df

df$free.sulfur.dioxide <- df$free.sulfur.dioxide/1000

df$total.sulfur.dioxide <- df$total.sulfur.dioxide/1000

**Univariate Plots**

**PLot 1.1 - Residual Sugar**

# Univariate plott of residual sugar

ggplot(aes(x=residual.sugar), data=df) +

geom\_histogram(fill='#FD7373', alpha=0.7, bins=100) +

scale\_x\_continuous(breaks=seq(from=0, to=100, by=2)) +

ggtitle("Frequency of residual sugar content") +

xlab("Residual Sugar Conc. (g/L)") +

ylab("Count")

We can see in this historam of residual sugar that there are roughly two types of wine, one type that has very little sugar in it (between 1 and 2 g/L) with a very small standard deviation and those type of wine which have more sugar whose concentration varies more with a rough mean of 7.5 g/L.

**Plot 1.2 - pH**

# Univariate plott of residual sugar

ggplot(aes(x=pH), data=df) +

geom\_histogram(fill='#FD7373', alpha=0.7, bins=100) +

xlim(0, 7) +

ggtitle("Frequency of pH") +

xlab("pH") +

ylab("Count")

The distribution for pH is roughly normally distributed with a mean of 3.1 (acidic)and a standard deviation of 0.151. In relation to other food substances, wine is as acidic as soda or orange juice but not as strong as vinegar or lemon juice (pH = 2) and certainly not as strong as sulphuric acid (ph = 1)

mean(df$pH)

sd(df$pH)

**Plot 1.3 - Acetic Acid**

ggplot(data=df) +

geom\_histogram(aes(x=acetic.acid), fill='#FD7373', alpha=0.6, bins=40,

labels="Acetic Acid") +

ggtitle("Histogram of Acetic Acid Concentrations") +

xlab("Acetic Acid Concentration (g/mL, log-transformed)") +

ylab("Count")

We can also see that the Acetic Acid concentration follows a similar log-distribution to the pH concentration with the long tail. This makes sense since the two variables are related due to acetic acid being acidic and pH being a measure of acidity.

The x-axis of the figure has not been transformed to a log scale because the values only cover one order of magnitude.

However, it is not expected that acetic acid and pH will coincide very strongly because any molecule that is acidic (such as tartaric acid, originally named fixed acidity) will also affect the pH.

cor(df$pH, df$acetic.acid)

Interestingly, although acetic acid concentration and pH have a similar shape/distribution on a histogram, they do not correlate with each other very strongly.

I would like to explore the relationship between these two variables further in the bivariate-graph section.

It should be noted that pH is already on a log scale which is why it does not need to be transformed in order to create the bell curve as above.

**Plot 1.4 - Tartaric Acid**

ggplot(aes(x=tartaric.acid), data=df) +

geom\_histogram(fill='#FD7373', alpha=0.6,

labels="Tartaric Acid") +

ggtitle("Histogram of Tartaric Acid Concentration") +

xlab("Tartaric Acid Concentration (g/mL, log-transformed)") +

ylab("Count")

Tartaric Acid Concentration adheres more to a normal distribution and seems to have roughly one order of magnitude higher concentrations tan that of acetic acid.

This makes sense because Tartaric Acid (originally named fixed acidity) has a boiling point (275°C) almost 3 times as high as Acetic acid (118.1°C), hence the name volatile acidity. This is reflected in the higher concentration of Tartaric acid present within the wine.

**Plot 1.5 - Citric Acid**

ggplot(aes(x=citric.acid), data=df) +

geom\_histogram(fill='#FD7373', alpha=0.7) +

scale\_x\_log10() +

ggtitle("Histogram of Citric Acid Concentration") +

xlab("Citric Acid Concentration (g/mL, log-transformed)") +

ylab("Count")

It is also rather interesting that citric acid concentration does not follow a log-distributions such as tartaric acid and acetic acid. It also exists in similar concentrations as acetic acid.

This leads me to wonder why wine has such high concentrations of tartaric acid, it's clear that acetic acid is volatile and that citric acid does not likely appear in large concentrations in grapes, which are not citrus fruits. The [wikipedia](https://en.wikipedia.org/wiki/Tartaric_acid) article on tartaric acid mentions that it exists in high concentrations within grapes which is why the concentration is also high in wine.

These acids likely follow similar distributions because of natural laws which decide which accounts for the randomness. As to why it would be log-distributed instead of normally distributed... it would normally happen because after a grape reaches a certain threshold of concentration of a certain acid then it tends to skyrocket.

This could be the result of human breeding/genetically-altering grapes which have really high concentrations of certain compounds within them to gain certain novel tastes?

**Plot 1.6 - Alcohol Content**

ggplot(aes(x=alcohol), data=df) +

geom\_histogram(fill='#FD7373', alpha=0.7, bins=70) +

ggtitle("Histogram of Alcohol Content") +

xlab("Alcohol Content (%)")

ylab("Count")

Of the features that I wanted to analyze, alcohol content is the only feature that does not follow a usual type of distribution. Alcohol content seems to range between 8% and 14.5%, whether or not this kind of difference is significant likely depends on each individual person's alcohol tolerance.

If this metric does not create any easily digestable graphs in the next two sections, then I can break the continuous nature of this quantitative data and convert it to categorical data (low, medium, high) using the natural breaks in the bins as seen above.

I am not entirely sure as to what these breaks are, it is likely that they represent some sort of rounding taking place. I am inferring this from the observation that these breaks come at regular intervals and have a constant width (i.e. One bin long). The producers of these wines might be rounding the numbers because humans have an easier time mentally digesting round numbers.

**Plot 1.7 - Salt Chloride Concentration**

ggplot(aes(x=salt.chlorides), data=df) +

geom\_histogram(fill='#FD7373', alpha=0.7, bins=70) +

ggtitle("Histogram of Salt Chloride concentration") +

scale\_x\_log10() +

xlab("Salt Chloride Concentration (g/L)")

ylab("Count")

Just like with other cocentrations, salt chloride also follows a log- distribution.

It's interesting that a lot of the chemicals dissolved in wine follow this distribution. What would be interesting would be to investigate the quality of wines that have high concentrations of one or more chemical compound. Is this novelty good or bad? What about is wine has multiple substances in high concentrations?

These questions are a little beyond the scope of this investigation because you would need to define what an outlier would be and then you would need to write some code to automatically identify these if you don't want to do it by hand, but it would be worth investigating.

**Plot 1.8 - Density**

ggplot(aes(x=density), data=df) +

geom\_histogram(fill='#FD7373', alpha=0.7, bins=70) +

ggtitle("Histogram of Wine Density") +

xlab("Density (g/L)") +

ylab("Count")

The values for density look rather spread in the graph above, but this is misleading as the x-ticks would reveal. One tick on this graph contains a density range of 0.02, which is not very much at all. Hence, there is not very much variation in white wines when it comes to density. This can be backed up using a standard deviation measurement of the data

sd(df$density)

The Standard deviation is calculated to be 0.030 g/L which is a difference that is not easily discernible by human senses. Hence, density will not be investigated further in this analysis.

**Plot 1.9 - Total Sulfur Dioxide**

ggplot(aes(x=total.sulfur.dioxide), data=df) +

geom\_histogram(fill='#FD7373', alpha=0.7, bins=70) +

ggtitle("Histogram of Total Sulfur Dioxide") +

xlab("Concentration of Total Sulfur Dioxide (g/L)") +

ylab("Count")

We can see that total sulfur dioxide has a rough concentration of 0.138 with a standard deviation of roughly 0.04 g/L, which is actually relatively narrow, once again, compared to one might visually gather from the histogram itself.

The concentration of total sulfur dioxide tends to be around the same sort of order of magnitude as citric and acetic acid in solution.

mean(df$total.sulfur.dioxide)

sd(df$total.sulfur.dioxide)

**Plot 1.10 - Free Sulfur Dioxide**

ggplot(aes(x=free.sulfur.dioxide), data=df) +

geom\_histogram(fill='#FD7373', alpha=0.7, bins=70) +

ggtitle("Histogram of Free Sulfur Dioxide") +

xlab("Concentration of Free Sulfur Dioxide (g/L)") +

ylab("Count")

Here, we can see that the concentration of free sulfur dioxide is normally about one half the concentration of the total concentration of sulfur dioxide. This could be due to equilibrium, there free sulfur dioxide is disfavored or because of the reaction between free sulfur dioxide and oxygen to prevent the formation of vinegar. This difference can be calculated by taking the dividend of the two means.

mean(df$free.sulfur.dioxide)/mean(df$total.sulfur.dioxide)

Variation in concentration would also be explained by the total age of the wine which is another metric that I don't have access to.

**Plot 1.11 - Sulphates**

ggplot(aes(x=sulphates), data=df) +

geom\_histogram(fill='#FD7373', alpha=0.7, bins=40) +

ggtitle("Histogram of Sulfates") +

xlab("Concentration of Sulfates (g/L)") +

ylab("Count")

Sulfates are added as a salt additive to help increase the concentration of free sulfur dioxide in solution. Depending on what kind of salt this is, it might play a role in the salt.chloride concentration of the wine. Sulfates seem to exist in higher concentration of total sulfur dioxide so it is likely not very active in solution. (i.e. It doesnt dissolve particularly well, otherwise one might not need add so much per liter)

As with many of the other metrics that have been investigated, this distribution also has a right skew.

**Plot 1.12 - Quality**

ggplot(aes(x=quality), data=df) +

geom\_bar(fill='#FD7373', alpha=0.7) +

scale\_x\_continuous(breaks=seq(0, 10, 1)) +

ggtitle("Historam of Quality") +

xlab("Quality of Wine") +

ylab("Count")

nrow(subset(x=df, quality<5))

nrow(subset(x=df, quality>=5))

Above we can see that Quality is roughly binomially distributed with most wines being rated as average (5 or 6) and with few wines being exceptionally expediant or poor in quality.

The range of the qualities lies between 3 and 9, perhaps this rating scale is too wide since it is difficult to find wines with a rating of 1, 2, and 10.

It is noteworthy that most wines are rated with a 6, slightly above average and wines are about twice as likely to be given a grade above 6 as seen below. Either most of the wine procured was of higher than average quality, the wine given during the test was not truely random in terms of quality, or even connoseurs have a hard time distinguishing between different wines.

It should be observed that since the quality of a the wine has been measured as a median, which means that true outliers of 1 or 10 would be exceptionally rare since all three of the ratings would have to match, which is unlikely.

For future plots, it would probably be a good idea to split the quality of the wine from low, medium and high quality. This will likely help to make trends in the bi-variate and multi-variate analyses more clear.

**Quick adjustments and wrangling**

It was found during the investigation that Quality is displayed at too high a resolution to be useful and as such the wine quality will be changed from a scale between 1 and 10 to values describing "low", "average" and "high" quality;

* "low": 1-4
* "average": 5-6
* "high": 6-10

# Encode values

df$quality <- ifelse(df$quality <= 4, 'average', ifelse(

df$quality >= 7, 'low', 'high'))

df$quality <- factor(df$quality, labels=c("low", "average", "high"))

**Plot 1.13 - Quality Histogram 2**

ggplot(aes(x=quality), data=df) +

geom\_bar(fill='#FD7373', alpha=0.7) +

ggtitle("Bar chart of Quality") +

xlab("Quality of Wine") +

ylab("Count")

Allocating the scores of wines into the bins outlined above creates this bar chart. Immediately we can see that there is a large bias to rate wine as either high quality or average. Very few wines recieve a score of less than 4. (n=183, from chunk 20)

It might be the case that it will be difficult to find trends for low quality wine because the sample size is not as large as one might want. Normally a decent sample size is between 500 and 1000 examples, although this might vary depending on what you are measuring. This needs to be considered when carrying out this analysis.

In any case, I think this makes a strong case for why ranking wine with the median score of 1-10 of 3 people is not the best way to measure quality.

**Univariate Analysis**

In this section, I found that most of the chemical concentrations within Wine have a log-distribution, where most wines certain a certain amount of a chemical with a few examples that become outliers becuase of how much higher the concentration of that chemical is. This is less true for sugar and alcohol content, which can vary substantially depending on the initial concentration of sugar and the fermentation time. It would be interesting to investigate this relationship further in the bi-variate plots section coming up next.

I was impressed to find that there is regular rounding that takes place when reporting the concentration of the wine, this is likely because it is possible to measure the concentration of wine to a higher degree of acuracy than a human would normally care to measure when drinking it.

I also found that the scale of 1-10 was too high a resolution for a qualitative evaluation. The resolution was too high because the very far extremes for wine quality (1 and 10) were not used at all, most wines tended to be slightly above average (6) as the result of human bias. To account for this, I will change the scale from a 1-10 metric to a qualitative metric. This should help to show the trends in the data much better, although it is likely that the sample size for low quality wine is probably too small.

It was found that pH and acid concentrations tended to have similar distributions. However, this did not translate into a strong correlation between any of the two features as it stands. This is something I would like to investigate further.

Wine has a much higher concentration of tartaric acid than citric and acetic acid, which results in wine's sour taste. This can be explained by acetic acid's low boiling point, citric acid likely simply exists in lower concentrations in grapes, which do not taste like citric fruits. Graphes are also known to have high concentrations of tartaric acid.

**Bivariate Plots**

**Plot 2.1 - Alcohol vs Quality**

ggplot(aes(y=alcohol, x=quality), data=df) +

geom\_violin(aes(fill=quality), alpha=0.7) +

ggtitle("Quality vs Alcohol Content") +

xlab("Quality") +

ylab("Alcohol (%)")

From the above plot we can see that high alcohol content markedly improves the quality of the wine. This is likely due to the longer fementation time for high quality wine which has a very noticable bulge towards the higher end of the distribution of the violin plot.

It is a shame that fermentation time is not included among the features, this would be an interesting dimention to analyse alongside this graph.

To back up this finding, I should compare this plot with a plot of residual sugar to see if there is a relationship.

**Plot 2.2 Residual Sugar vs Alcohol Content**

ggplot(aes(y=residual.sugar, x=alcohol), data=df) +

geom\_point(alpha=0.1) +

geom\_smooth(method='lm', formula=y~x) +

ggtitle("Residual Sugar vs Alcohol Content") +

xlab("Alcohol Content (%)") +

ylab("Residual Sugar (g/L)") +

ylim(0, 25)

Here we can at least confirm my hypothesis that alcohol content leads to lower residual sugar concentration. Although it should be noted that low alcohol content does not necessarily mean high concentrations of residual sugar, since we do not know the initial sugar concentration.

If we knew the rate of alcohol fermentation (which would depend on fermentation time and temperature) then this would be something that we could calculate.

**Plot 2.3 - Residual Sugar vs Quality**

ggplot(aes(y=residual.sugar, x=quality), data=df) +

geom\_jitter(aes(fill=quality), alpha=0.1) +

geom\_boxplot(aes(fill=quality), alpha=0.3) +

stat\_summary(fun.y = "mean", geom="point", color="red", shape=8, size=4) +

ggtitle("Residual Sugar vs Quality") +

xlab("Quality") +

ylab("Residual Sugar (g/L)")

From the above boxplot we can see that there is a strong "baseline" for residual sugar. This can be seen with the large overlap of points when approaching 0 g/L residual sugar. This is likely because the fermentation time for these examples was long enough to ensure that very little residual sugar remained within solution. We can see that high residual sugar produces markedly average wine but high quality wine also has its fair share of sugar as well. although this average is not quite as high.

This is somewhat consistent with my previous plot in regards to alcohol content if high quality wine has a higher concentration of alcohol then it must stand to reason that assuming the initial sugar concentration does not vary much that there will be less residual sugar within high quality wine as a result. Although it should probably be noted that initial sugar concentration (as well as fermentation time and temperature) can vary and this assumption likely does not hold entirely fast and true producing the graph above.

In the multivariate section of this analysis we will take the plot from 2.2 and color the points to highlight them according to quality this should place all of the information in one graph and should tell the story between alcohol content, residual sugar and quality much better.

**Plot 2.4 - Citric Acid Concentration vs Quality**

ggplot(aes(y=citric.acid, x=quality), data=df) +

geom\_jitter(aes(fill=quality), alpha=0.1) +

geom\_boxplot(aes(fill=quality), alpha=0.3) +

stat\_summary(fun.y = "mean", geom="point", color="red", shape=8, size=4) +

ggtitle("Citric Acid Concentration vs Quality") +

xlab("Quality") +

ylab("Citric Acid Concentration (g/L)")

The plot above shows that the citric acid concentration converges onto an ideal value. The quality of wine increases as the wine reaches this value. It should be noted that even low quality wine that has this concentration of citric acid can still be ruined by other means.

**Plot 2.5 - Acetic Acid Concentration vs Quality**

ggplot(aes(y=acetic.acid, x=quality), data=df) +

geom\_jitter(aes(fill=quality), alpha=0.1) +

geom\_boxplot(aes(fill=quality), alpha=0.3) +

stat\_summary(fun.y = "mean", geom="point", color="red", shape=8, size=4) +

ggtitle("Acetic Acid Concentration vs Quality") +

xlab("Quality") +

ylab("Acetic Acid Concentration (g/L)")

This finding is more in-line with expectations, namely that high levels of vinegar (acetic acid) causes the quality of the wine to decrease. This is to be expected because great lengths are taken to make sure that alcohol does not oxidise to vinegar by making sure the wine remains under conditions which are not oxidising as well as additives which prevent this from taking place.

There does not seem to be much difference between average and high quality wine in terms of distribution, which says to me that after basic precautions are taken to prevent wine oxidation that some amount will occur and that the quality of wine will be determined by other factors.

**Plot 2.6 - Tartaric Acid Concentration vs Quality**

ggplot(aes(y=tartaric.acid, x=quality), data=df) +

geom\_jitter(aes(fill=quality), alpha=0.1) +

geom\_boxplot(aes(fill=quality), alpha=0.3) +

stat\_summary(fun.y = "mean", geom="point", color="red", shape=8, size=4) +

ggtitle("Tartaric Acid Concentration vs Quality") +

xlab("Quality") +

ylab("Tartaric Acid Concentration (g/L)")

Here we can see that tartaric acid decreases the quality of wine, i.e. high quality wine tends to have lower concentrations of tartaric acid. I think this finding is relatively interesting because tartaric acid exists in relatively concentrations within grapes, the major ingredient of wine.

I think that this is likely because tartaric acid exists in such high concentrations that it likely obscures the more interesting tastes within wine by overpowering the taster with the sour taste that it has.

**Plot 2.7 - Salt Chloride Concentration vs Quality**

ggplot(aes(y=salt.chlorides, x=quality), data=df) +

geom\_jitter(aes(fill=quality), alpha=0.05) +

geom\_boxplot(aes(fill=quality), alpha=0.3) +

stat\_summary(fun.y = "mean", geom="point", color="red", shape=8, size=4) +

ggtitle("Salt Chloride Concentration vs Quality") +

xlab("Quality") +

ylab("Salt Chloride Concentration (g/L)")

Here we can see a slight decrease in the concentration of salt chloride with quality. This is because both the means and the medians decrease in salt chloride concentration. This difference is relatively small since wines tend to have a consistent level of salt within them but the trend if noticible for sure.

**Plot 2.8 - Sulphates Concentration vs Quality**

ggplot(aes(y=sulphates, x=quality), data=df) +

geom\_jitter(aes(fill=quality), alpha=0.1) +

geom\_boxplot(aes(fill=quality), alpha=0.3) +

stat\_summary(fun.y = "mean", geom="point", color="red", shape=8, size=4) +

ggtitle("Concentration of Sulphates vs Quality") +

xlab("Quality") +

ylab("Concentration of Sulphates (g/L)")

There is also a similarly consistent increase of quality with sulphate concentration. This is to be expected because sulphates prevent the formation of acetic acid which would inhibit the quality of the wine. So it is good that this is used as an additive.

**Plot 2.9 - Free & Total Sulfur Dioxide Concentration vs Quality**

p1 <- ggplot(aes(y=free.sulfur.dioxide, x=quality), data=df) +

geom\_jitter(aes(fill=quality), alpha=0.05) +

geom\_boxplot(aes(fill=quality), alpha=0.3) +

stat\_summary(fun.y = "mean", geom="point", color="red", shape=8, size=4) +

ggtitle("Free Sulfur Dioxide Concentration vs Quality") +

xlab("Quality") +

ylab("Free Sulfur Dioxide Concentration (g/L)")

p2 <- ggplot(aes(y=total.sulfur.dioxide, x=quality), data=df) +

geom\_jitter(aes(fill=quality), alpha=0.05) +

geom\_boxplot(aes(fill=quality), alpha=0.3) +

stat\_summary(fun.y = "mean", geom="point", color="red", shape=8, size=4) +

ggtitle("Total Sulfur Dioxide Concentration vs Quality") +

xlab("Quality") +

ylab("Total Sulfur Dioxide Concentration (g/L)")

grid.arrange(p1, p2, ncol=2)

rm(p1, p2)

Finally, we can see that free sulfur dixide concentration tends to increase the quality of the wine. It is worth noting that sulfur dioxide is the "eggy" smell of flatulence so it makes sense that you would not want this chemical to exist in high concentrations within wine. This would almost certainly hold back a good wine from being a great wine.

Although having low sulfur dioxide concentrations would mean that the wine is at higher risk of oxidation causing the formation of acetic acid, so a balance should really be struck here.

**Plot 2.10 - pH vs Acid Concentration**

# calculate total acid concentration

df$acid.conc <- df$tartaric.acid + df$acetic.acid + df$citric.acid

ggplot(aes(y= pH, x=acid.conc), data=df) +

geom\_point(alpha=0.25) +

ggtitle("pH vs Acid Concentration") +

xlab("Acid Concentration (g/L)")

Above, we can see the change in pH with acid concentration calculated as a sum of tartaric, acetic and citric acids. We can see that although there is a general correlation between total acid concentration and pH, that relationship is not very strong for various reasons.

Firstly, tartaric, citric and acetic acid all have different pKa values. This is basically the way chemists measure the acidic strength of certain molecules and is central to understanding organic chemistry.

The relationship between pH and pKa can be related with this equation:

$$pH = pK\_a + log(\frac{[A-]}{[HA]})$$

As we can see, since we did not do a log transformation nor did we take pKa values into account, there is no way we will create any sort of resolved line in this chart. Though we can still see the general trend that pH decreases with acid concentration.

If we were to take all of these details into account, we might be able to find a difference between the predicted pH and the real pH and see if there are any trace acids that were not considered in this analysis. Although, we would likely need to report pH to a higher precision as the acids that lead to the discrepancy are probably trace compounds that do not contribute very much to a difference in taste unless it is very noticable in small concentrations, which is possible but unlikely.

**Bivariate Analysis**

In this section of the analysis we found that alcohol content tends to increase the quality of the wine. High alcohol content tends to coincide with longer fermentation times and hence lower concentrations of residual sugar.

It would see that the various concentrations of acids within wine contribute differently towards the quality as well. Citric acid seems to have an ideal value which is best converged upon, to make the wine taste fresh but not too fresh. Acetic acid (i.e. vinegar) is negatively correlated with quality and is a chemical that is best avoided. Tartaric acid's sour taste also exists in high concentrations and can obscure the tastes created by other chemical compounds and so high concentrations of this acid is negatively correlated with quality.

It also seems that salt negatively impacts the quality of wine, I am not sure as to why this is because salt stimulates the taste glands and so is usually good when it comes to making things taste better. Hence, saltiness must disturb the ideal taste of white wine in some way. Increasing concentration of sulphates tends to help wine quality as it prevents acetic acid formation but high concentrations of free sulfur dioxide are not preferable either because it does not necesarily taste very nice. (The gaseous verion smell quite bad actually.)

**Multivariate Plots**

**Plot 3.1 - Residual Sugar vs Alcohol Content, Colored by Quality**

ggplot(data=df, aes(x=alcohol, y=residual.sugar)) +

geom\_point(aes(colour=quality), alpha=0.35) +

geom\_smooth(method="lm", aes(color=quality)) +

ggtitle("Residual Sugar vs Alcohol Content") +

ylab("Residual Sugar (g/L)") +

xlab("Alcohol Concentration (%)") +

ylim(0,25)

The decrease in alcohol concentration with residual sugar is very clear to see across wine qualities. We can see that high quality wine tends to cluster towards higher alcohol concentration, which is consistent with findings shown earlier.

One interesting finding is that the residual sugar of average and high quality wine is higher than low quality wine. This suggests that having a good remainder of sugar after fementation will increase the quality of the wine even when taking the error of these linear regression models into account. (Notice the broad grey line surrounding the low quality regression.)

**Plot 3.2 - Concentration of Sulfur Dioxide Concentrations vs Salt**

Cocentration

# Calculate bound sulfure dioxide

df$bound.sulfur.dioxide <- df$total.sulfur.dioxide - df$free.sulfur.dioxide

# Create melted df

melt\_vector <- c("free.sulfur.dioxide",

"bound.sulfur.dioxide",

"salt.chlorides")

df\_melt <- subset(df, select=melt\_vector)

df\_melt <- gather(df\_melt, variable, value, -salt.chlorides)

# Plot data

ggplot(data=df\_melt) +

geom\_point(aes(y=value, x=salt.chlorides, color=variable), alpha=0.25) +

scale\_x\_log10() +

ggtitle("Concentration of Sulfur Dioxide Concentrations vs Salt Concentration") +

xlab("Salt Chloride Concentration (g/L)") +

ylab("Sulfur Dioxide Concentration (g/L)")

# Free-up memory

rm(df\_melt, melt\_vector)

This plot disproves the hypothesis I had in section 2.5. My hypothesis was that increasing salt concentration would lead to a decrease in free sulfur concentration and increase the concentration of bound sulfur dioxide, which I calculated as being the concentration left over when substracting the total sulfur dioxide concentration from the free sulfur concentration.

Here, we can see that salt concentration tends to follow a log-distribution where most concentrations are crowded around one value with a couple of notable outliers. In general, it seems that more sulfur dioxide is bound rather than free in solution, which aligns with the description given in the text file accompanying this dataset.

Plot 3.3 - Acid Constributions

# Create melted df

melt\_vector <- c("tartaric.acid", "acetic.acid", "citric.acid", "acid.conc", "pH")

# Create melted df

df\_melt <- subset(df, select=melt\_vector)

df\_melt <- gather(df\_melt, variable, concentration, -c("acid.conc", "pH"))

# Plot Data

ggplot(aes(y= pH, x=acid.conc), data=df\_melt) +

geom\_point(aes(color=concentration)) +

facet\_grid(~variable) +

ggtitle("pH vs Acid Concentration") +

xlab("Total Acid Concentration (g/L)")

# Clean up memory

rm(df\_melt, melt\_vector)

Above, we can see a graph showing the same graph as pH vs Total Acid Concentration as seen in plot 2.6, except this time the points have been colored according to the concentration of each acid within that total value. We can see that tartaric acid contributes the most in terms of mass over volume (g/L) and with acetic and citric acid contributing much less.

To use the Henderson-Hasselbach equation shown above we would need to know the concentration of $[A^-]$ (the unbound acid) in solution, right now we only know the bound concentration.

However, drawbacks with this diagram are that it does not control for molecular mass and it does not account for the actual strength of the acids. This means that a molecule could be heavier and thus contrubute more to the concentration per molecule, if a molecule also was able to dissociate more strongly from its protons then we would not be able to see it from a graph like this.

**Multivariate Analysis**

In this final stage of the analysis, it was discovered that some hypotheses that I drew earlier in this analysis turned out to likely be false.

I hypothesised that salt likely played a role within the bound and unbound concentration of bound and unbound sulfur dioxide in solution, but it turns out that this concentration is independent of salt concentration.

More evidence was found that suggests that high alcohol content is favorable for the quality of wine. There also seems to be an added bonus if there is also a high concentration of residual sugar after the fementation has finished.

Finally, another way was found to communicate the concentration of tartaric acid within wine, I find this to be interesting becuase before this analysis I did not have any notion about what acids are present within wine, this plot gives a feel for that using color to express concentration.

**Final Plots**

**Plot 4.1 - Histogram of acid concentrations**

# Create melted df

melt\_vector <- c("tartaric.acid", "acetic.acid", "citric.acid")

df\_melt <- subset(df, select=melt\_vector)

df\_melt <- gather(df\_melt, variable, value)

ggplot(aes(x=value), data=df\_melt) +

geom\_histogram(aes(fill=variable, color=I("grey")), alpha=0.95, bins=50) +

ggtitle("Histogram of acid concentrations") +

ylab("Count") +

xlab("Total Acid Concentration (g/L)")

rm(df\_melt, melt\_vector)

With this plot I wanted to quickly summarise the acid concentrations because I spent a lot of time investigating this. Although I do like plot 3.3 from an aesthetic perspective, using a color graident is not always the best way to compare differences in an precise way. The human eye is better at understanding vertical and horizontal lengths and so I have chosen to use that instead.

Grouping the acid concentrations of acetic and citric acid together creates a connection in the readers mind that makes it separate from tartaric acid. The reader will be able to intuitively understand that the values of citric and tartaric acid are rather consistent whereas tartaric acid has higher variation and concentration.

By melting the data in this way, I was able to condense multiple histograms into a single graph. Where it is clear to see what kind of concentrations of the various acids are within wine.

**Plot 4.2**

corr\_coeff = cor(x=df$residual.sugar, y=df$alcohol)

ggplot(aes(y=residual.sugar, x=alcohol), data=df) +

geom\_point(color=I("grey40"), alpha=0.25) +

geom\_smooth(method='lm', formula=y~x, aes(colour="Linear Model")) +

ggtitle("Residual Sugar vs Alcohol Content") +

xlab("Alcohol Content (%)") +

ylab("Residual Sugar (g/L)") +

scale\_colour\_manual(name="Legend", values=c("blue")) +

ylim(0, 32) +

annotate("text", x=14, y=9.5, label="R² = -0.451", colour="royalblue1")

rm(corr\_coeff)

I also really liked plot 2.2 which tells a story about how residual sugar content decreases with alcohol content.

I changed the graph so that instead of using a really low alpha level for the points, I changed the color to grey so that they go into the background and used a blue line to catch the attention of the viewer and draw their attention to the clear decrease in trend.

I added a legend to make it clear that a linear regression was used and I added some statistics such as the $R²$ value in order to express that this is not an indredibly strong correlation but is worth noting nonetheless.

The readers attention is usually drawn from left to right, and so the first thing that a reader will notice is the line and will follow it. Their eye will likely be drawn next to the legend and will cross the correlation coefficient on the way there. After this information is understood they will be able to see the original data points which lead to this model. The axes are clearly labeled so that it is clear what these data points are.

**Plot 4.3**

# Create Values

bins <- 10

cols <- c("#F5ED1C","#86C246")

colGradient <- colorRampPalette(cols)

cut.cols <- colGradient(bins)

cuts <- cut(df$pH,bins)

names(cuts) <- sapply(cuts,function(t) cut.cols[which(as.character(t) == levels(cuts))])

# Univariate plott of residual sugar

ggplot(aes(x=pH, fill=cut(pH, bins)), data=df) +

geom\_histogram(alpha=0.7, bins=bins, show.legend=FALSE) +

scale\_color\_manual(values=cut.cols,labels=levels(cuts))+

scale\_fill\_manual(values=cut.cols,labels=levels(cuts)) +

xlim(0, 14) +

geom\_vline(xintercept=2, color="#F7C611") +

geom\_vline(xintercept=6.5, color="green") +

geom\_vline(xintercept=1, color="red") +

geom\_vline(xintercept=5, color="#4DB749") +

scale\_x\_continuous(breaks=seq(0,14,1)) +

annotate("text",

x=c(1, 2, 5, 6.5),

y=3000,label=c("Battery Acid", "Vinegar", "Banana","Water"),

hjust=0,

color="grey35") +

ggtitle("pH of Wine") +

xlab("pH") +

ylab("Count")

rm (bins, cols, colGradient, cut.cols, cuts)

I also really liked this histogram which shows how the pH of wine ranks compared to other chemicals and objects.

This graph does not calture that this scale is a logarhithmic one, which means that pH 1 is 10$\times$ as concentrated as pH2 is 0$\times$ as concentrated as pH3, etc.

Using a color gradient is normall when viewing a pH scale because it gives a qualitative measurement to see how this difference changes.

**Reflection**

**What insights were found?**

The major insights found in this investigation are as follows:

* Good for Quality:
  + Higher Alcohol Concentration is good for quality
  + Bonus points for having high concentration of residual sugar
  + Having high concentrations of sulfates to prevent the formation of acetic acid is also beneficial
* Bad for Quality:
  + High concentrations of tartaric acid obscures the finer tastes of wine
  + High concentratoins of acetic acid also makes wine too sour and so should be avoided
  + High concentrations of salt also is not very good for the quality of wine
  + High concentrations of free sulfur dioxide is also not very good for the taste of wine because sulfur dioxide is smelly.
* Balance Required:
  + Citric acid should exist but not in overly high concentrations

Other miscilaneous obersvations that I made were that rounding-up takes place when it comes to labeling wine concentrations to attain round numbers.

The method of rating the wine meant that wines were overrated on average, this is likely because a median on a scale of 1-10 is too high resolution for now accurate peoples evaluation of taste is. This would likely mean that a lower resolution of quality would lead to more consistent measurements.

It should also be noted that residual sugar is negatively correlated with alcohol concentration because the fermentation process that creates alcohol requires sugar.

I had a hypothesis that salt concentration might play a role in the equilibrium between free and fixed sulfur dioxide in solution, with higher concentrations of salt inhibiting the formation of free sulfur dioxide. However it was found that these two metrics are independent of each other.

**Difficulties and Solutions**

When I first started this analysis I used Ggally to create a scatter matrix of the features within the dataset. However, looking at the data in this way did not highlight clear correlations to investigate as I had hoped. Instead, it really obscured all of the patterns in the data and it discouraged me quite a bit.

I tried to counter this problem by plotting the features one-by-one. This helped a lot to help bring out the patterns in the data, I am continually impressed by how much changing the visualisation method and the ways to draw emphasis to different elments in a graph contributes toward how it is percieved!

Finally, I found half way through my analysis that I had made a mistake when it came to the ordering of wines when it changed the variables. This meant that I had to correct this problem and then rewrite most of my analysis that I had done which took up a lot of time. Next time I should try and do some more checks to make sure that the transformations that I perform are correct.

**Ideas for Further Analysis**

Although it would be interesting to train a deep learning model to this dataset and then use a random number generator to make plausible wines and perhaps use a kind of evolutionary algorithm to create wines that the deel learning network predicts to be of very high quality and strive to create this wine based on similar ones as a template. That might be interesting.

**Improvements for the analysis**

* Better features: fermentation time
* Better outcome recording: Controll for drunkness of assessers
* Removal of bias: Dont take median of quality

[view raw](https://gist.github.com/SThornewillvE/9b8e6ccb9bb7000b8a5b6d19cfe1f497/raw/d00e9dc4a1eecd86d3313474e7b13a08ef6ece25/investigating-a-dataset.Rmd) [investigating-a-dataset.Rmd](https://gist.github.com/SThornewillvE/9b8e6ccb9bb7000b8a5b6d19cfe1f497#file-investigating-a-dataset-rmd) hosted with ❤ by [GitHub](https://github.com)

**Share this:**

* [Click to share on Twitter (Opens in new window)](http://datasciencedemonstrated.com/2018/03/08/exploratory-data-analysis/?share=twitter&nb=1)
* [1Click to share on Facebook (Opens in new window)1](http://datasciencedemonstrated.com/2018/03/08/exploratory-data-analysis/?share=facebook&nb=1)
* [Click to share on Google+ (Opens in new window)](http://datasciencedemonstrated.com/2018/03/08/exploratory-data-analysis/?share=google-plus-1&nb=1)

**Like this:**

Loading...

***Related***

[Data Analysis Nanodegree 01: Weather Trend Analysis](http://datasciencedemonstrated.com/2017/10/15/data-analysis-nanodegree-01-weather-trend-analysis/)October 15, 2017In "Udacity Projects"

[Data Analysis Nanodegree 02: US 2016 Bike-share Analysis](http://datasciencedemonstrated.com/2017/10/25/data-analysis-nanodegree-02-us-2016-bike-share-analysis/)October 25, 2017In "Udacity Projects"

[Data Analysis Nanodegree 04: AB-Tests](http://datasciencedemonstrated.com/2018/02/22/ab-tests/)February 22, 2018In "Udacity Projects"

Previous article [Data Analysis Nanodegree 04: AB-Tests](http://datasciencedemonstrated.com/2018/02/22/ab-tests/)

**Leave a Reply**

Top of Form

Comment 

Name \* 

Email \*

Website

Bottom of Form

Top of Form

Notify me of follow-up comments by email.

Notify me of new posts by email.

Bottom of Form

Top of Form

Bottom of Form

**Categories**

Top of Form

Categories

Bottom of Form

**Social Media**

* [View sthornewillve’s profile on LinkedIn](https://www.linkedin.com/in/sthornewillve/)
* [View SThornewillvE’s profile on GitHub](https://github.com/SThornewillvE/)

**Pages**

* [Social Media](http://datasciencedemonstrated.com/social_media/)
* [Welcome to Data Science Demonstrated!](http://datasciencedemonstrated.com/welcome_to_data_science_demonstrated/)

**Recent Posts**

* [Data Analysis Nanodegree 05: Exploratory Data Analysis](http://datasciencedemonstrated.com/2018/03/08/exploratory-data-analysis/)
* [Data Analysis Nanodegree 04: AB-Tests](http://datasciencedemonstrated.com/2018/02/22/ab-tests/)
* [Book Review: Storytelling With Data by Cole Nussbaumer Knaflic](http://datasciencedemonstrated.com/2018/01/21/storytelling/)
* [Data Analysis Nanodegree 03: Exploring Gapminder Data](http://datasciencedemonstrated.com/2017/12/03/gapminder/)
* [Data Analysis Nanodegree 02: US 2016 Bike-share Analysis](http://datasciencedemonstrated.com/2017/10/25/data-analysis-nanodegree-02-us-2016-bike-share-analysis/)

**Goodreads**

[Animal Farm](https://www.goodreads.com/review/show/2316485838?utm_medium=api&utm_source=custom_widget)

[Animal Farm](https://www.goodreads.com/review/show/2316485838?utm_medium=api&utm_source=custom_widget)

by [George Orwell](https://www.goodreads.com/author/show/3706.George_Orwell)

[The Business Blockchain: Promise, Practice, and Application of the Next Internet Technology](https://www.goodreads.com/review/show/2309496195?utm_medium=api&utm_source=custom_widget)

[The Business Blockchain: Promise, Practice, and Application of the Next Internet Technology](https://www.goodreads.com/review/show/2309496195?utm_medium=api&utm_source=custom_widget)

by [William Mougayar](https://www.goodreads.com/author/show/15172416.William_Mougayar)

[Eat That Frog!: 21 Great Ways to Stop Procrastinating and Get More Done in Less Time](https://www.goodreads.com/review/show/2307134667?utm_medium=api&utm_source=custom_widget)

[Eat That Frog!: 21 Great Ways to Stop Procrastinating and Get More Done in Less Time](https://www.goodreads.com/review/show/2307134667?utm_medium=api&utm_source=custom_widget)

by [Brian Tracy](https://www.goodreads.com/author/show/22033.Brian_Tracy)

[Journey to Data Scientist: Interviews with More Than Twenty Amazing Data Scientists](https://www.goodreads.com/review/show/2306883452?utm_medium=api&utm_source=custom_widget)

[Journey to Data Scientist: Interviews with More Than Twenty Amazing Data Scientists](https://www.goodreads.com/review/show/2306883452?utm_medium=api&utm_source=custom_widget)

by [Kate Strachnyi](https://www.goodreads.com/author/show/17359628.Kate_Strachnyi)

[The Data Science Handbook](https://www.goodreads.com/review/show/2300026990?utm_medium=api&utm_source=custom_widget)

[The Data Science Handbook](https://www.goodreads.com/review/show/2300026990?utm_medium=api&utm_source=custom_widget)

by [Carl Shan](https://www.goodreads.com/author/show/7512321.Carl_Shan)

[goodreads.com](https://www.goodreads.com/)

**Archives**

* [March 2018](http://datasciencedemonstrated.com/2018/03/)
* [February 2018](http://datasciencedemonstrated.com/2018/02/)
* [January 2018](http://datasciencedemonstrated.com/2018/01/)
* [December 2017](http://datasciencedemonstrated.com/2017/12/)
* [October 2017](http://datasciencedemonstrated.com/2017/10/)

**Meta**

* [Log in](http://datasciencedemonstrated.com/wp-login.php)
* [Entries RSS](http://datasciencedemonstrated.com/feed/)
* [Comments RSS](http://datasciencedemonstrated.com/comments/feed/)
* [WordPress.org](https://wordpress.org/)

© 2018 ·[SingleBox](https://www.wpstud.io/) · by [WPStud.io](https://www.wpstud.io)