# Red Wine Quality Exploration by Samuel Duan

> **Tip**: You will see quoted sections like this throughout the template to help you construct your report. Make sure that you remove these notes before you finish and submit your project!

> **Tip**: One of the requirements of this project is that your code follows good formatting techniques, including limiting your lines to 80 characters or less. If you're using RStudio, go into Preferences > Code > Display to set up a margin line to help you keep track of this guideline!

> **Tip**: Before you create any plots, it is a good idea to provide a short introduction into the dataset that you are planning to explore. Replace this quoted text with that general information!

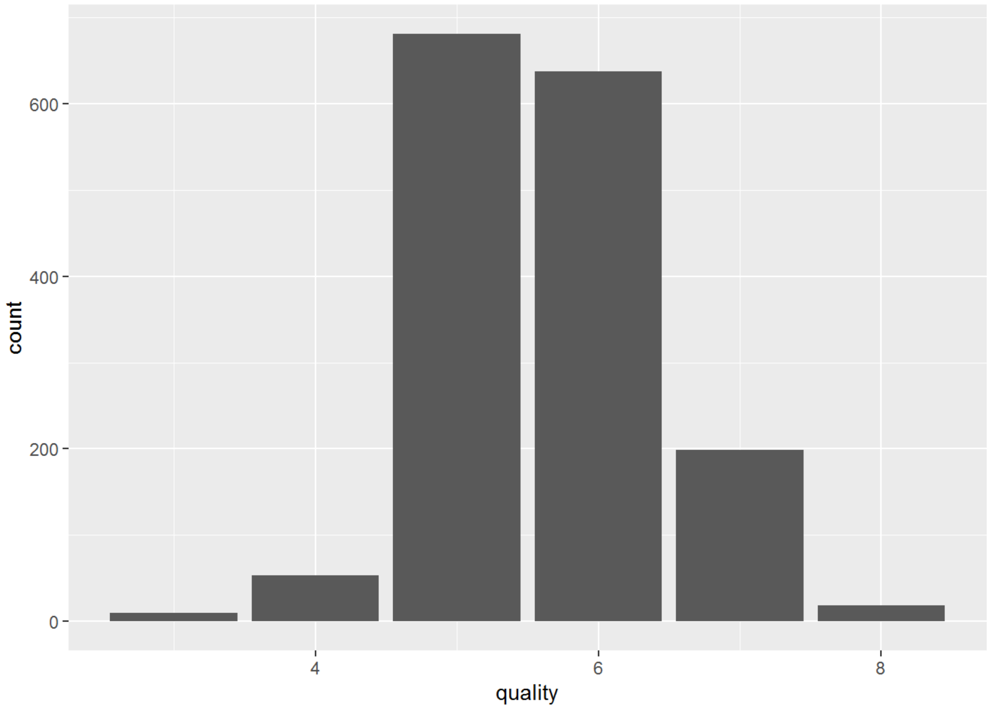
# Univariate Plots Section

> **Tip**: In this section, you should perform some preliminary exploration of your dataset. Run some summaries of the data and create univariate plots to understand the structure of the individual variables in your dataset. Don't forget to add a comment after each plot or closely-related group of plots! There should be multiple code chunks and text sections; the first one below is just to help you get started.

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

This data set contains 1599 observations and 11 chemical qualities that could affect the wine quality.

### Histogram of wine quality

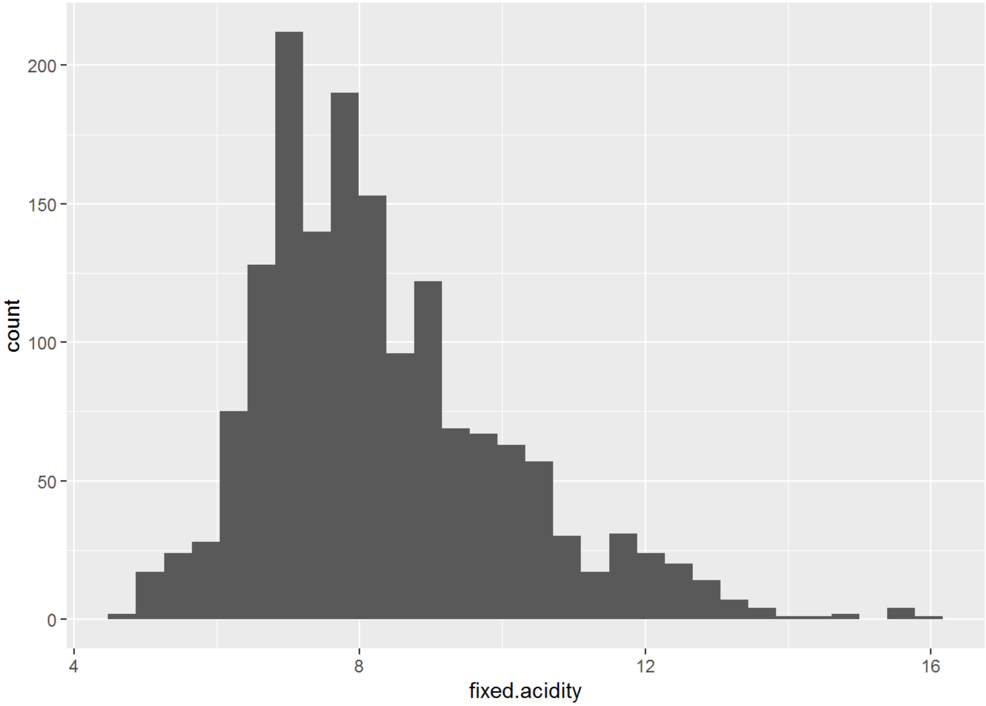


```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.000   5.000   6.000   5.636   6.000   8.000
```

It is visible from the plot above that the mode of the wine quality is 5.

### Histogram of fixed acidity

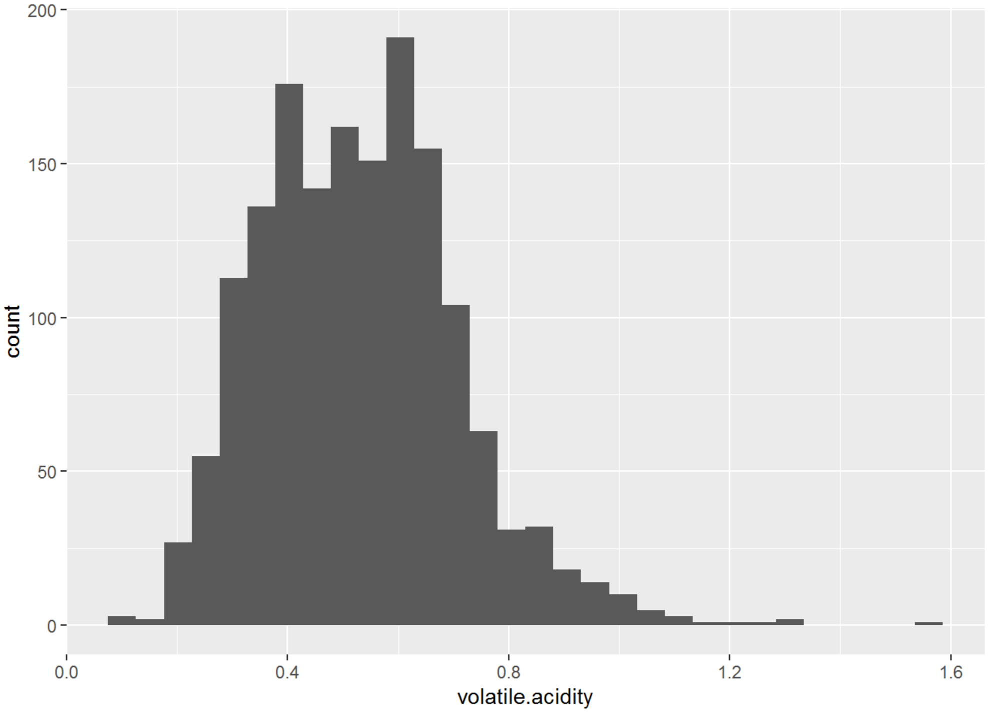
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

### Histogram of volatile acidity

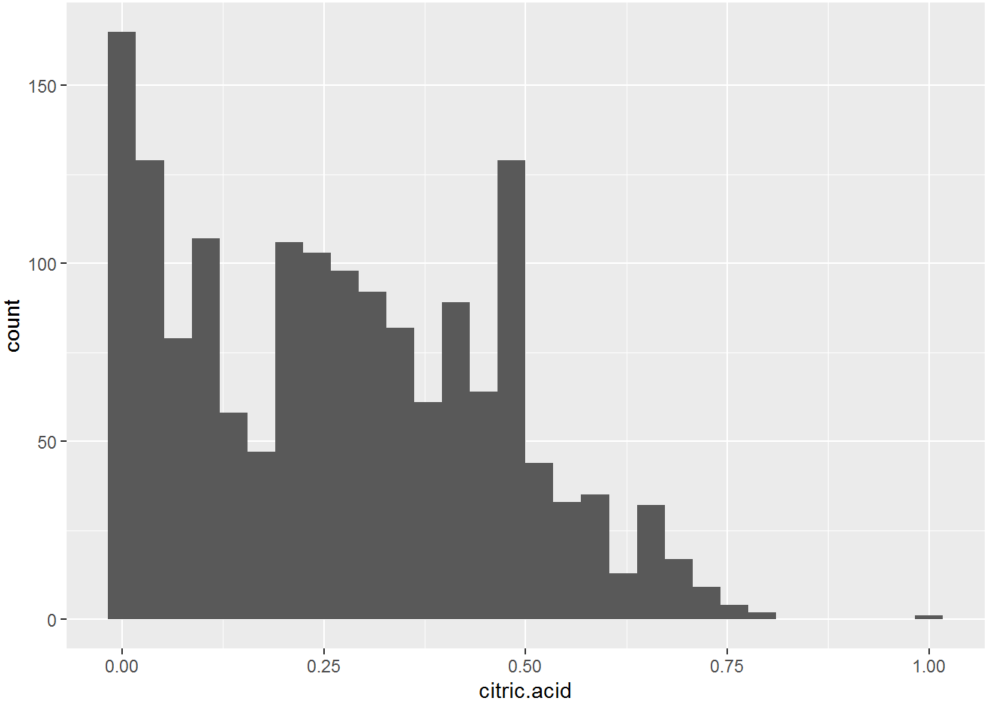
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

### Histogram of citric acid

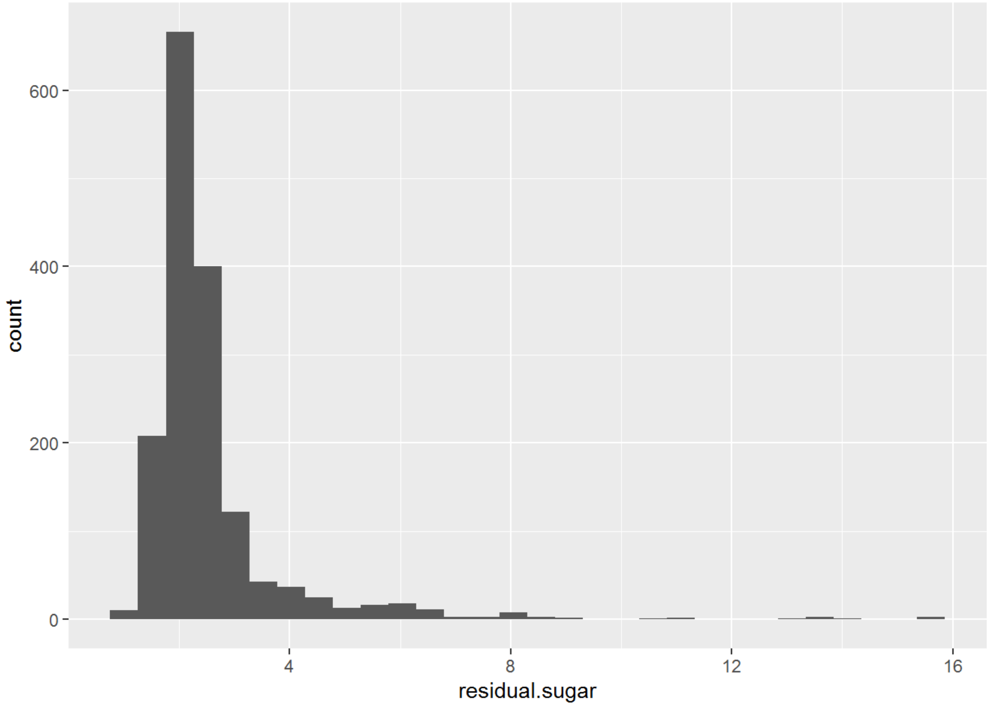
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

### Histogram of residual sugar

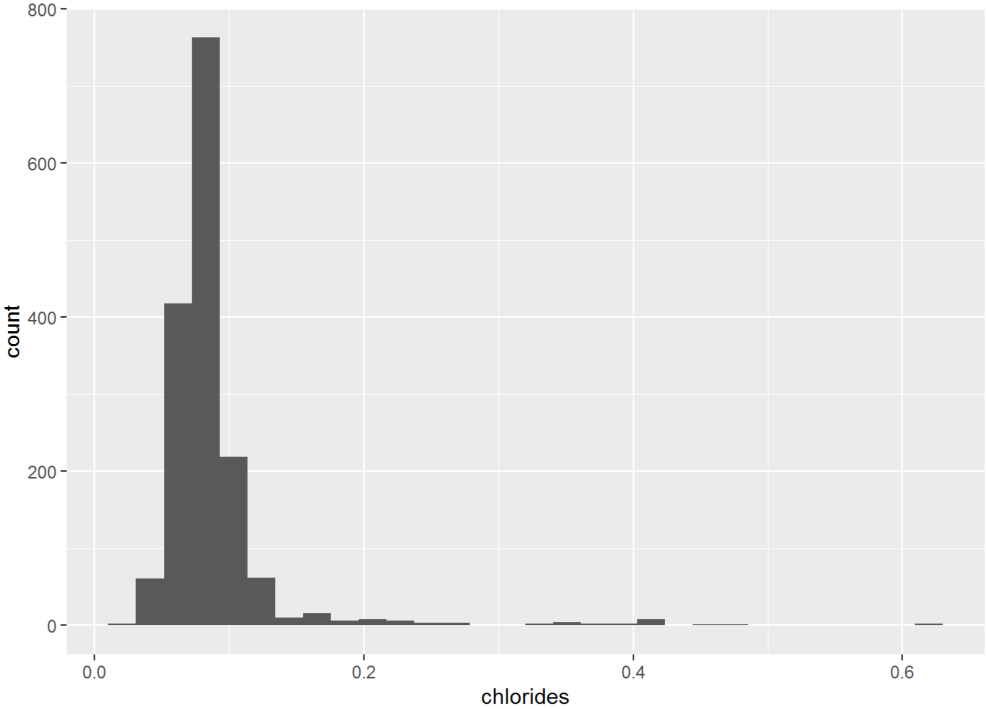
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

### Histogram of chlorides

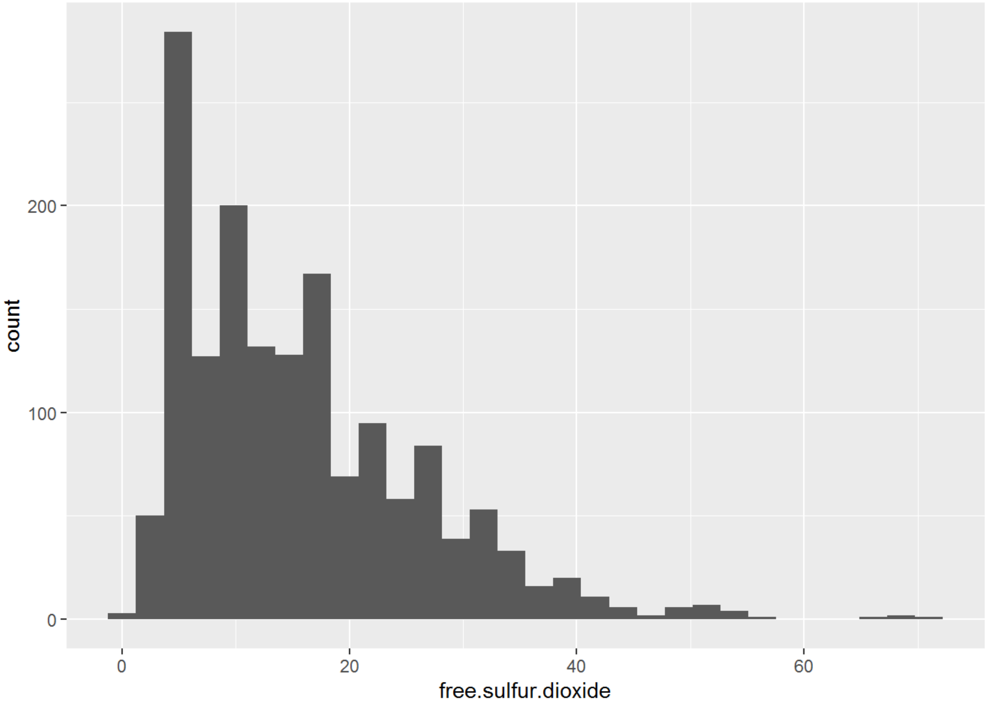
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

### Histogram of free sulfur dioxide

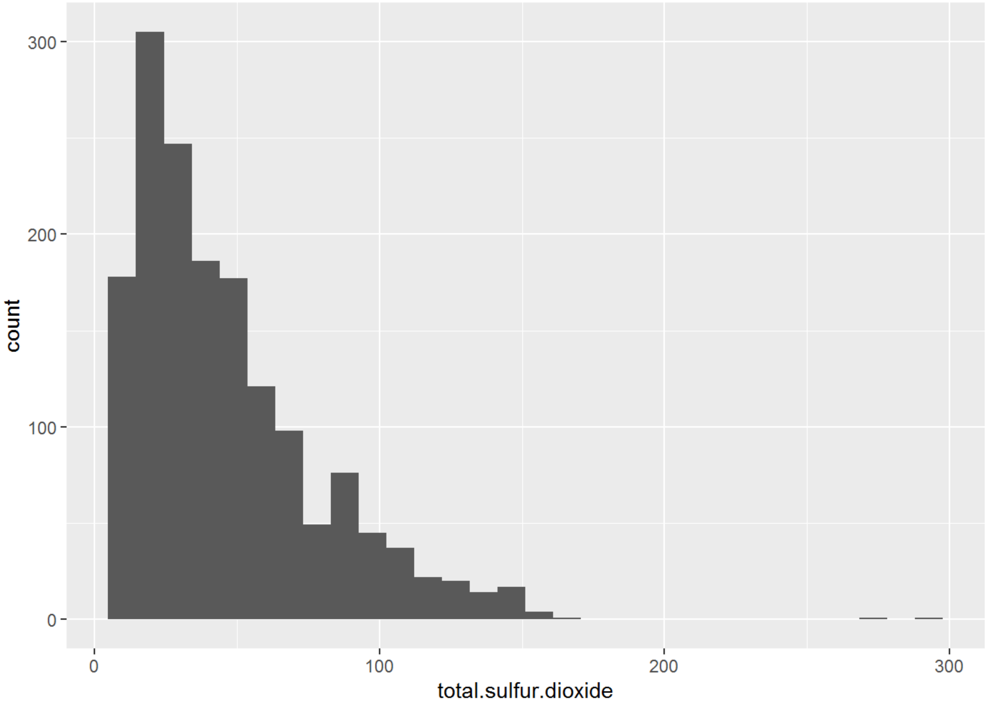
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

### Histogram of total sulfur dioxide

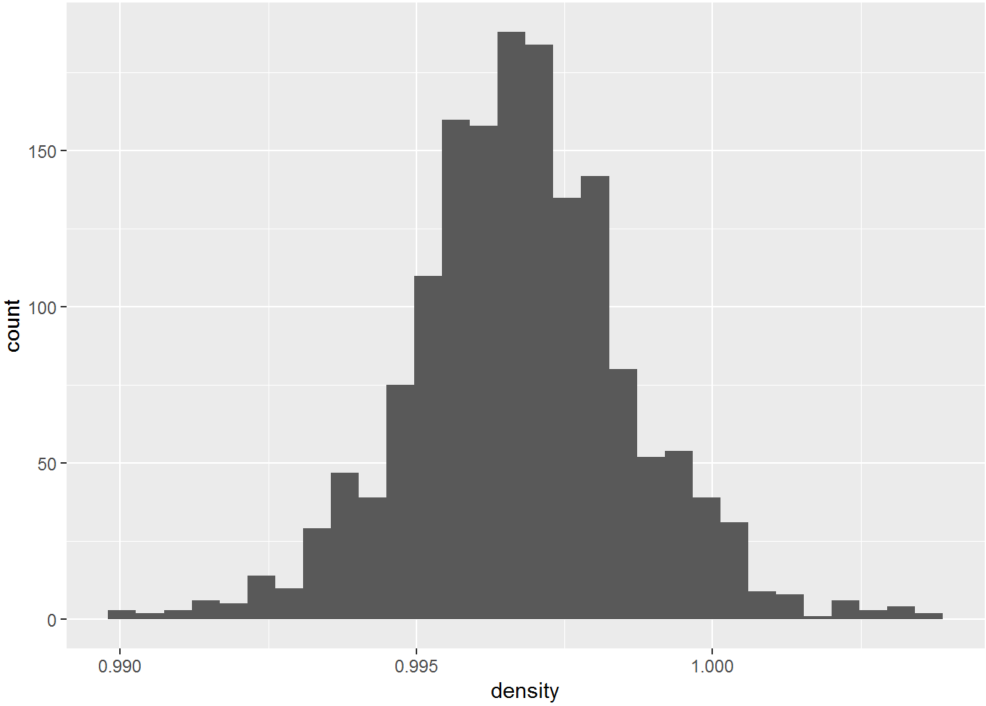
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

### Histogram of density

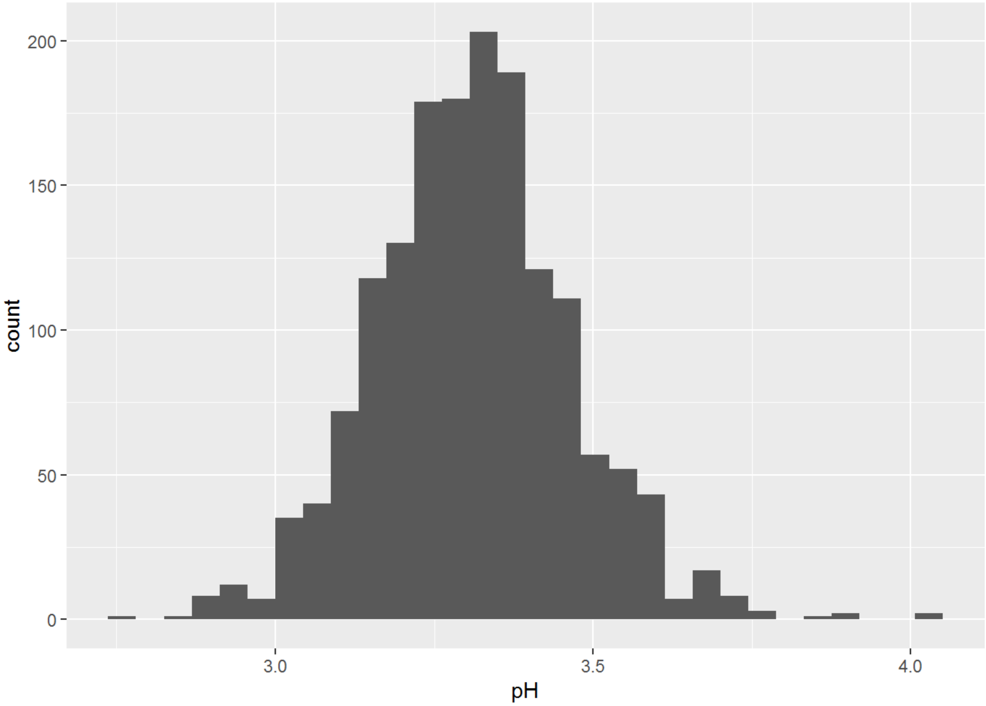
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

### Histogram of pH value

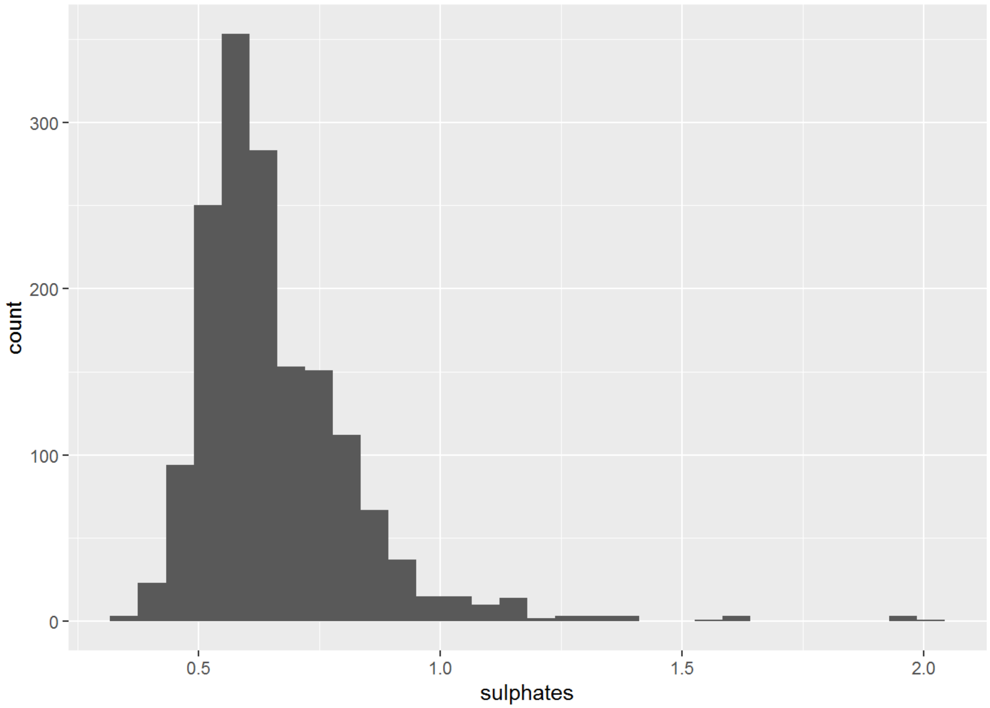
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

### Histogram of sulphates

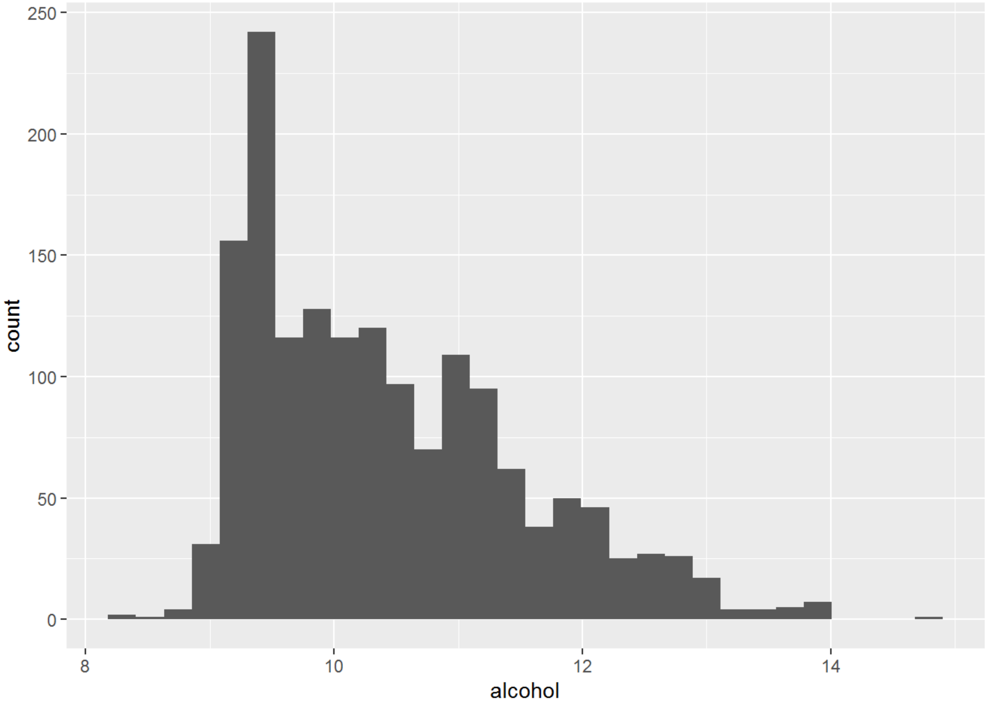
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

### Histogram of alcohol

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

> **Tip**: Make sure that you leave a blank line between the start / end of each code block and the end / start of your Markdown text so that it is formatted nicely in the knitted text. Note as well that text on consecutive lines is treated as a single space. Make sure you have a blank line between your paragraphs so that they too are formatted for easy readability.

# Univariate Analysis

> **Tip**: Now that you've completed your univariate explorations, it's time to reflect on and summarize what you've found. Use the questions below to help you gather your observations and add your own if you have other thoughts!

## What is the structure of your dataset?

This data set contains 1599 observations and 11 chemical qualities that could affect the wine quality. All of 11 chemical qualities are numerical variables.
Other observations:

1. The distribution of Citric acid is trimodal, while the others are unimodal.
2. The distributions of free sulfur dioxide, total sulfur dioxide, sulphates and alcohol are right skewed.
3.

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

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

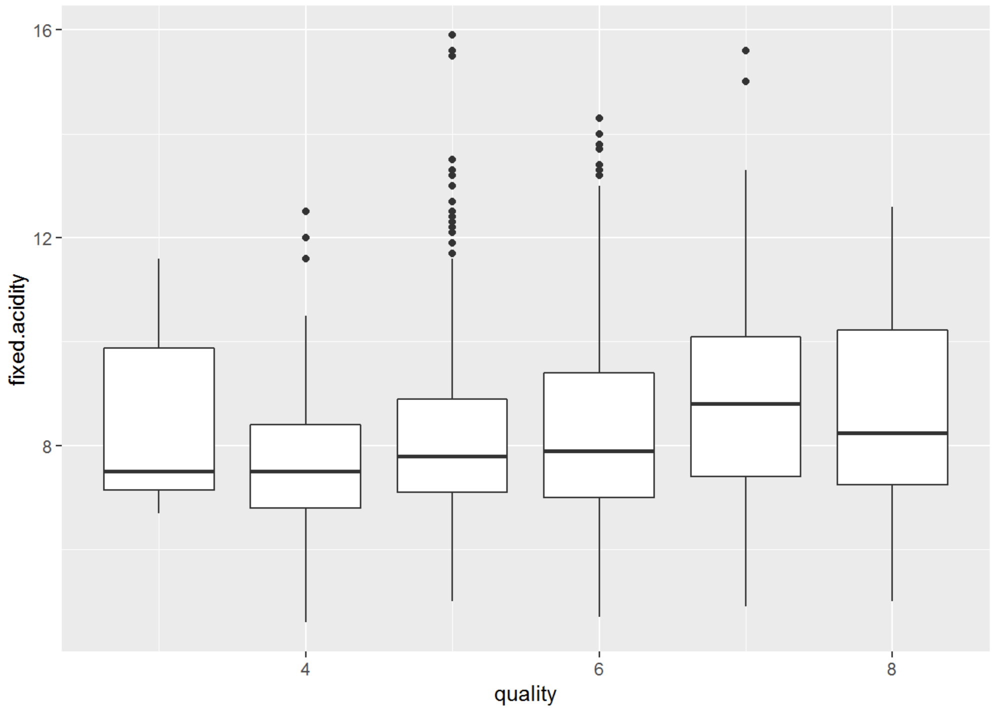
### Did you create any new variables from existing variables in the dataset?

### 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?

# Bivariate Plots Section

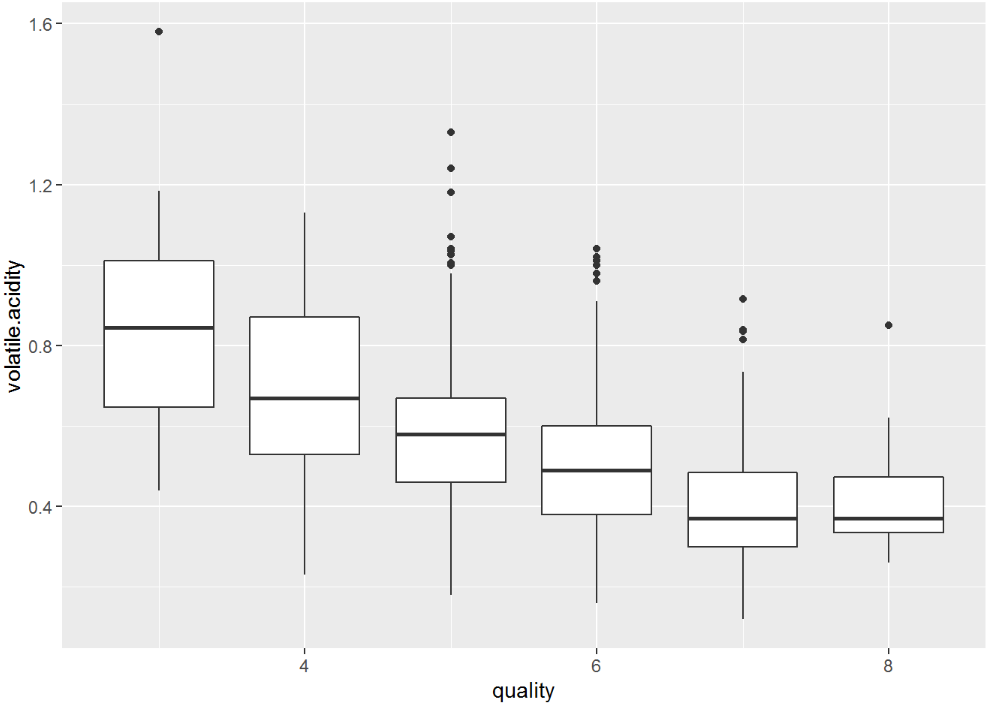
> **Tip**: Based on what you saw in the univariate plots, what relationships between variables might be interesting to look at in this section? Don’t limit yourself to relationships between a main output feature and one of the supporting variables. Try to look at relationships between supporting variables as well.

Boxplot: fixed.acidity vs quality



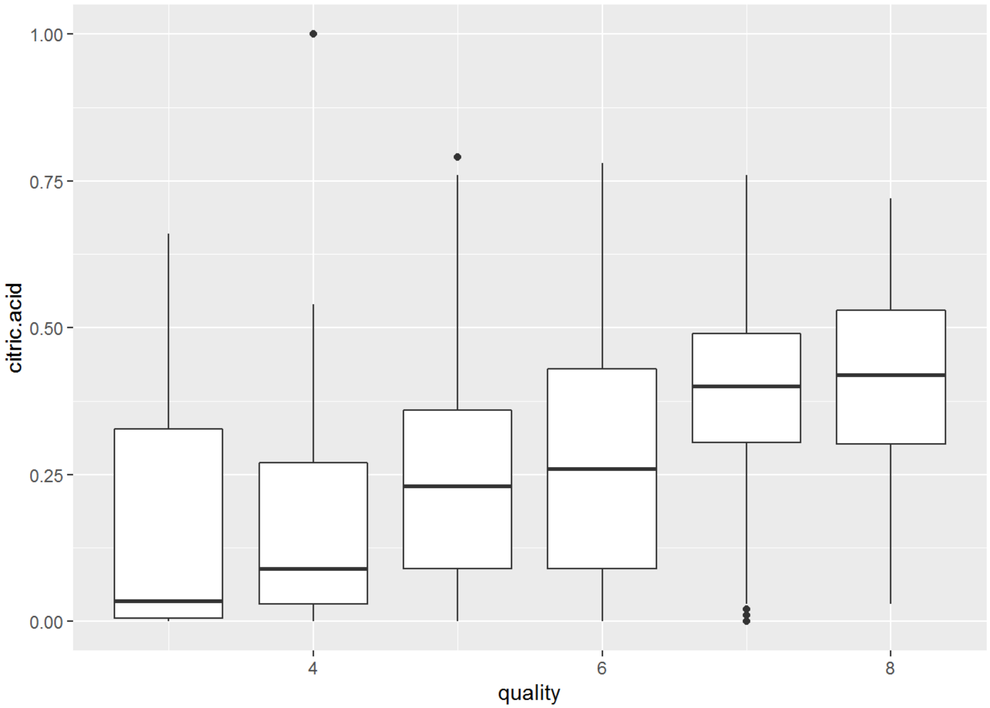
```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   6.700   7.150   7.500   8.360   9.875  11.600 
## 
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.600   6.800   7.500   7.779   8.400  12.500 
## 
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   5.000   7.100   7.800   8.167   8.900  15.900 
## 
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.700   7.000   7.900   8.347   9.400  14.300 
## 
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.900   7.400   8.800   8.872  10.100  15.600 
## 
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   5.000   7.250   8.250   8.567  10.230  12.600
```

### Boxplot: volatile.acidity vs quality



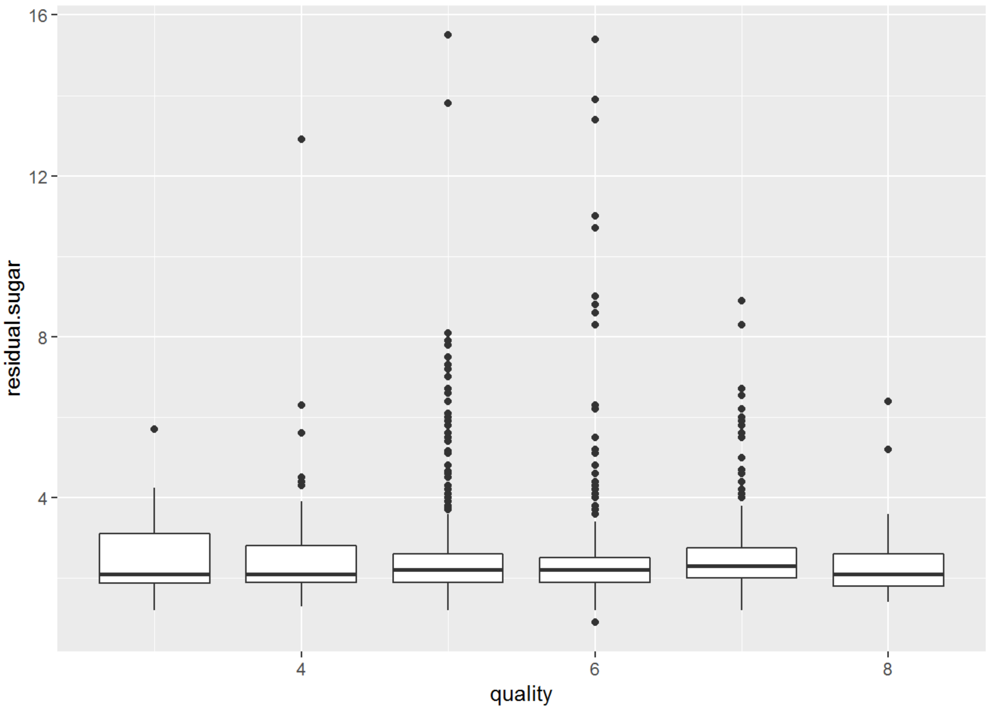
```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.4400  0.6475  0.8450  0.8845  1.0100  1.5800 
## 
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.230   0.530   0.670   0.694   0.870   1.130 
## 
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.180   0.460   0.580   0.577   0.670   1.330 
## 
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1600  0.3800  0.4900  0.4975  0.6000  1.0400 
## 
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1200  0.3000  0.3700  0.4039  0.4850  0.9150 
## 
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.2600  0.3350  0.3700  0.4233  0.4725  0.8500
```

### Boxplot: citric.acid vs quality



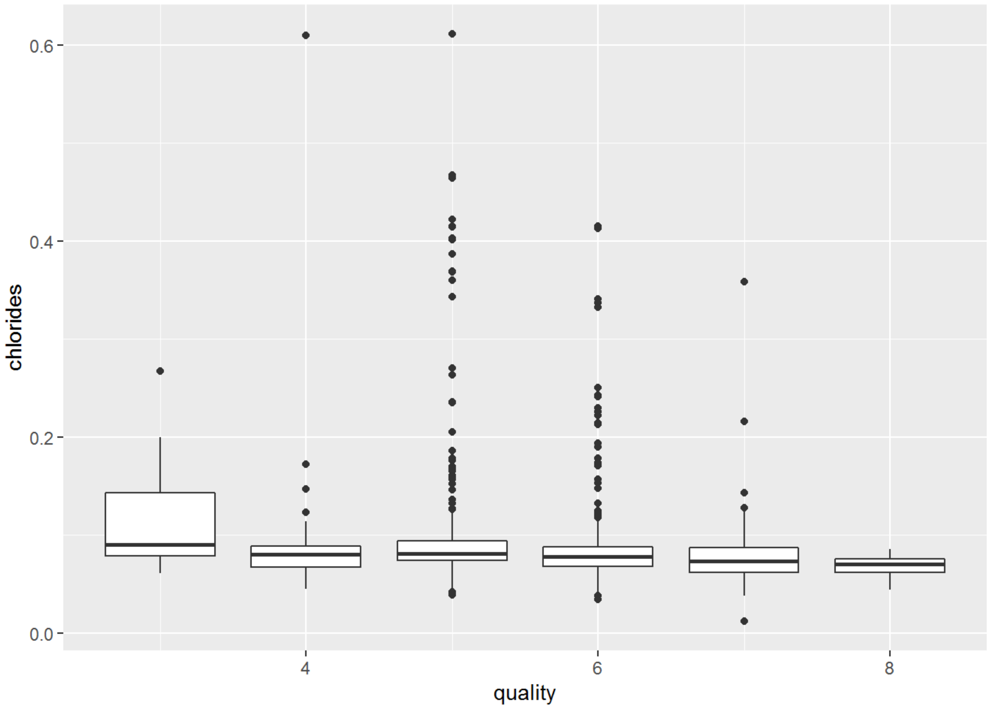
```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.0050  0.0350  0.1710  0.3275  0.6600 
## 
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.0300  0.0900  0.1742  0.2700  1.0000 
## 
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.0900  0.2300  0.2437  0.3600  0.7900 
## 
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.0900  0.2600  0.2738  0.4300  0.7800 
## 
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.3050  0.4000  0.3752  0.4900  0.7600 
## 
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0300  0.3025  0.4200  0.3911  0.5300  0.7200
```

### Boxplot: residual.sugar vs quality



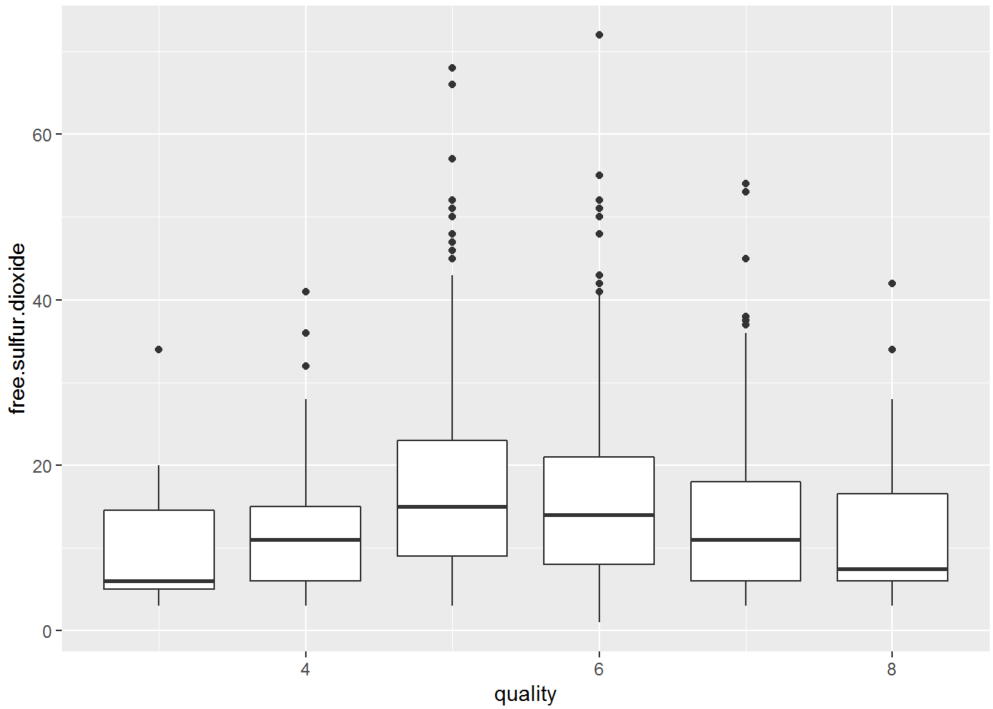
```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.200   1.875   2.100   2.635   3.100   5.700 
## 
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.300   1.900   2.100   2.694   2.800  12.900 
## 
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.200   1.900   2.200   2.529   2.600  15.500 
## 
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.900   1.900   2.200   2.477   2.500  15.400 
## 
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.200   2.000   2.300   2.721   2.750   8.900 
## 
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.400   1.800   2.100   2.578   2.600   6.400
```

### Boxplot: chlorides vs quality



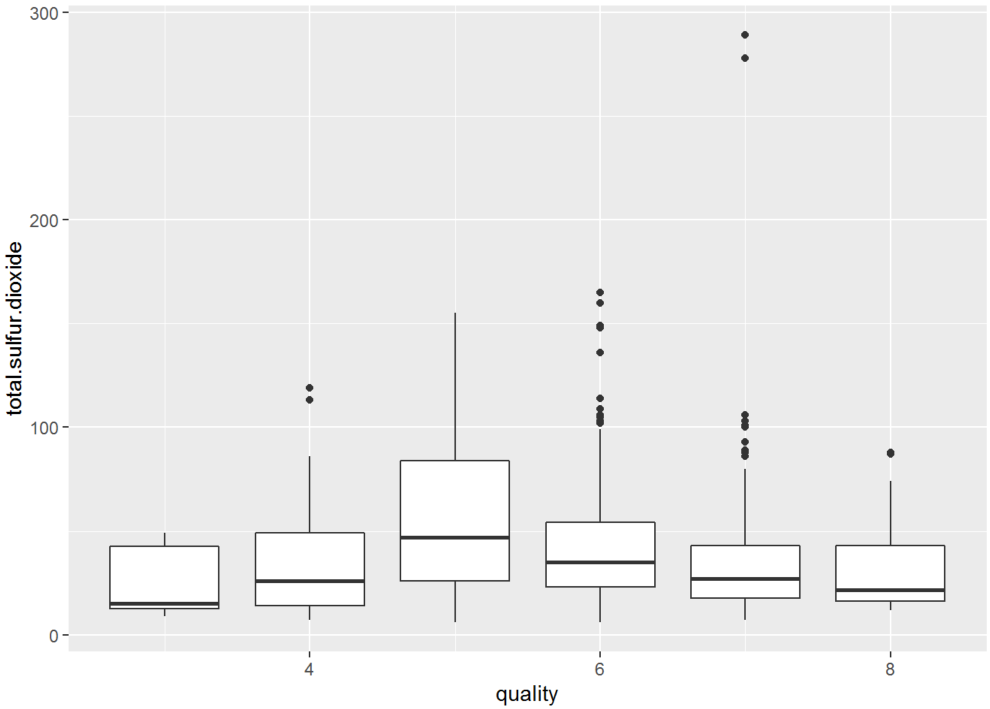
```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0610  0.0790  0.0905  0.1225  0.1430  0.2670 
## 
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.04500 0.06700 0.08000 0.09068 0.08900 0.61000 
## 
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.03900 0.07400 0.08100 0.09274 0.09400 0.61100 
## 
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.03400 0.06825 0.07800 0.08496 0.08800 0.41500 
## 
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.01200 0.06200 0.07300 0.07659 0.08700 0.35800 
## 
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.04400 0.06200 0.07050 0.06844 0.07550 0.08600
```

### Boxplot: free.sulfur.dioxide vs quality



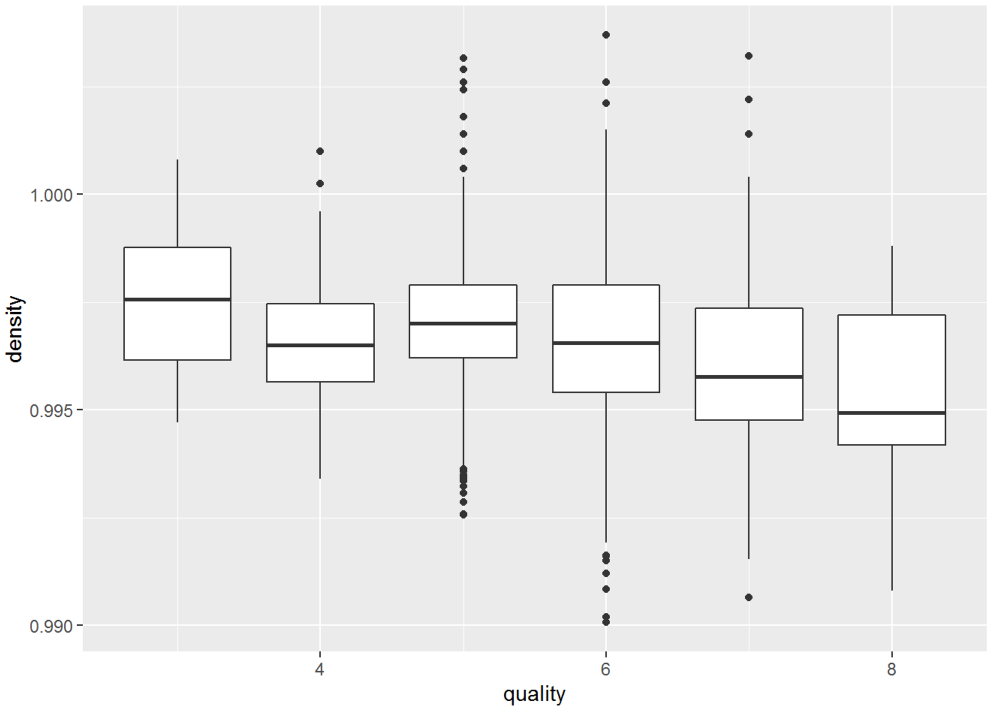
```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     3.0     5.0     6.0    11.0    14.5    34.0 
## 
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    3.00    6.00   11.00   12.26   15.00   41.00 
## 
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    3.00    9.00   15.00   16.98   23.00   68.00 
## 
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00    8.00   14.00   15.71   21.00   72.00 
## 
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    3.00    6.00   11.00   14.05   18.00   54.00 
## 
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    3.00    6.00    7.50   13.28   16.50   42.00
```

### Boxplot: free.sulfur.dioxide vs quality



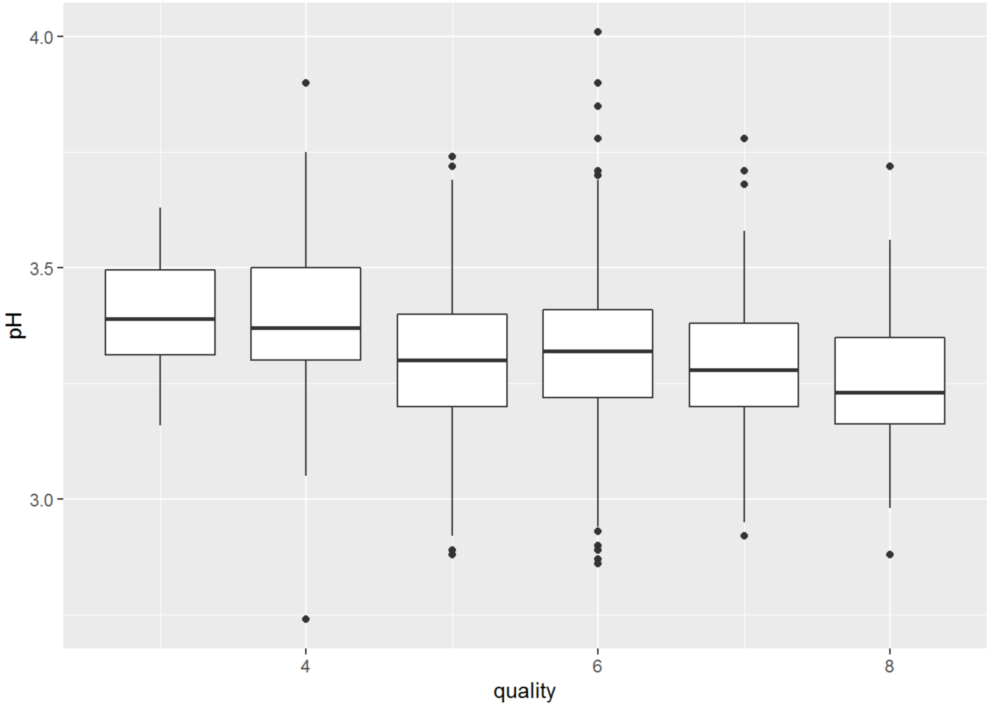
```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     9.0    12.5    15.0    24.9    42.5    49.0 
## 
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    7.00   14.00   26.00   36.25   49.00  119.00 
## 
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.00   26.00   47.00   56.51   84.00  155.00 
## 
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.00   23.00   35.00   40.87   54.00  165.00 
## 
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    7.00   17.50   27.00   35.02   43.00  289.00 
## 
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   12.00   16.00   21.50   33.44   43.00   88.00
```

Boxplot: density vs quality



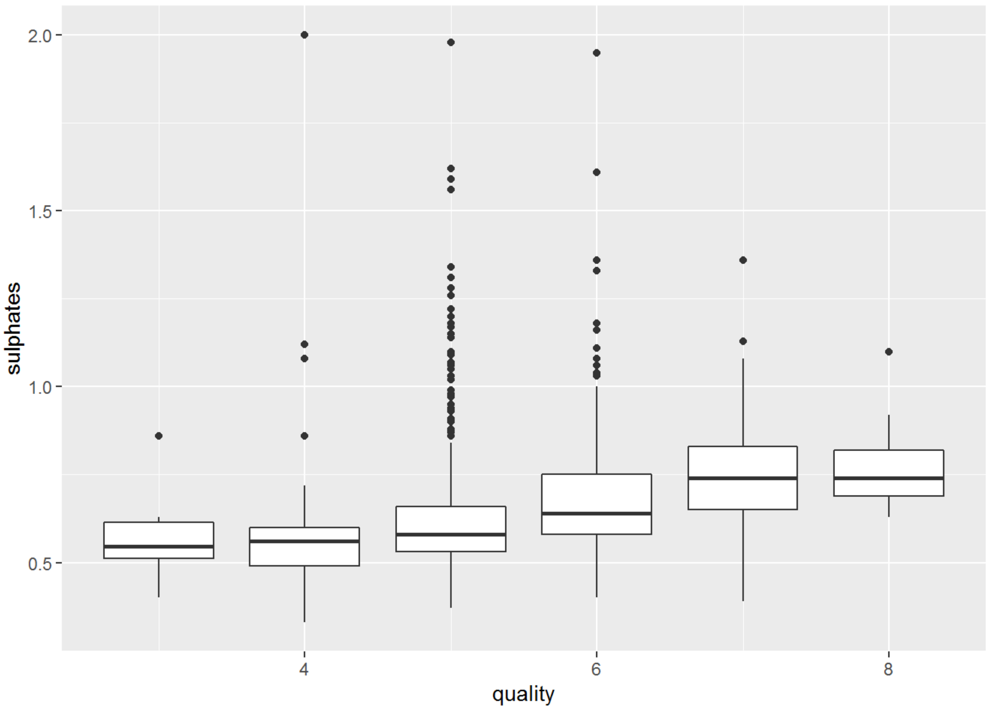
```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.9947  0.9962  0.9976  0.9975  0.9988  1.0010 
## 
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.9934  0.9956  0.9965  0.9965  0.9974  1.0010 
## 
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.9926  0.9962  0.9970  0.9971  0.9979  1.0030 
## 
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.9901  0.9954  0.9966  0.9966  0.9979  1.0040 
## 
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.9906  0.9948  0.9958  0.9961  0.9974  1.0030 
## 
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.9908  0.9942  0.9949  0.9952  0.9972  0.9988
```

### Boxplot: pH vs quality



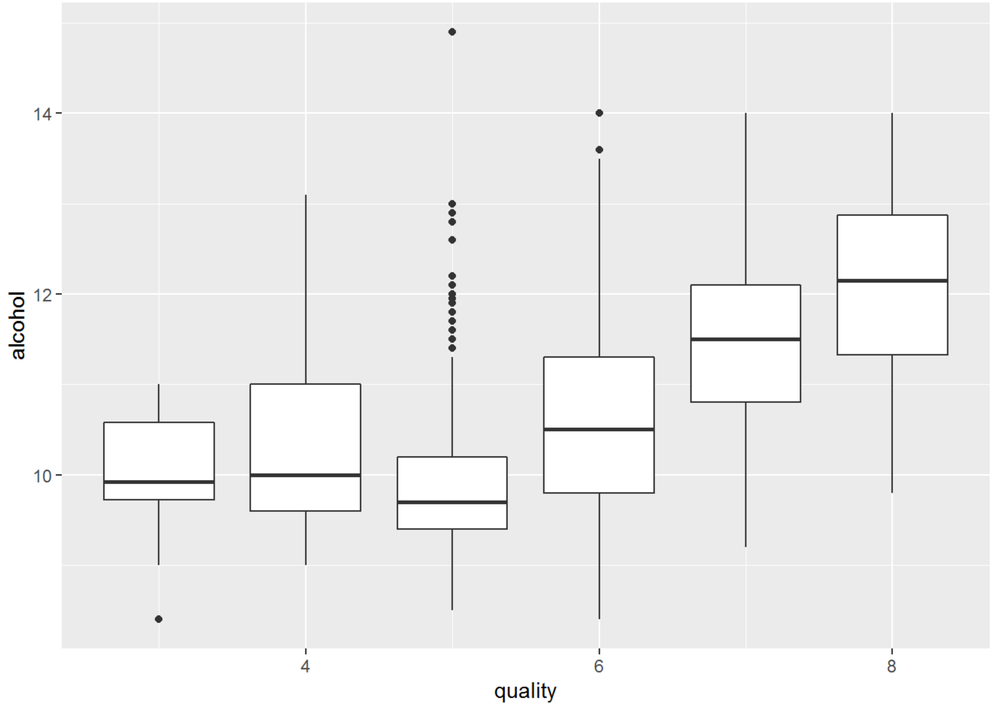
```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.160   3.312   3.390   3.398   3.495   3.630 
## 
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.740   3.300   3.370   3.382   3.500   3.900 
## 
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.880   3.200   3.300   3.305   3.400   3.740 
## 
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.860   3.220   3.320   3.318   3.410   4.010 
## 
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.920   3.200   3.280   3.291   3.380   3.780 
## 
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.880   3.162   3.230   3.267   3.350   3.720
```

## Boxplot: sulphates vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.4000  0.5125  0.5450  0.5700  0.6150  0.8600 
## 
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3300  0.4900  0.5600  0.5964  0.6000  2.0000 
## 
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.370   0.530   0.580   0.621   0.660   1.980 
## 
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.4000  0.5800  0.6400  0.6753  0.7500  1.9500 
## 
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3900  0.6500  0.7400  0.7413  0.8300  1.3600 
## 
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.6300  0.6900  0.7400  0.7678  0.8200  1.1000
```

### Boxplot: alcohol vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   8.400   9.725   9.925   9.955  10.580  11.000 
## 
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.00    9.60   10.00   10.27   11.00   13.10 
## 
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     8.5     9.4     9.7     9.9    10.2    14.9 
## 
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    8.40    9.80   10.50   10.63   11.30   14.00 
## 
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.20   10.80   11.50   11.47   12.10   14.00 
## 
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.80   11.32   12.15   12.09   12.88   14.00
```

# Bivariate Analysis

> **Tip**: As before, summarize what you found in your bivariate explorations here. Use the questions below to guide your discussion.

### 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?

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

### What was the strongest relationship you found?

# Multivariate Plots Section

> **Tip**: Now it's time to put everything together. Based on what you found in the bivariate plots section, create a few multivariate plots to investigate more complex interactions between variables. Make sure that the plots that you create here are justified by the plots you explored in the previous section. If you plan on creating any mathematical models, this is the section where you will do that.

# 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?

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

### OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

---

# Final Plots and Summary

> **Tip**: You've done a lot of exploration and have built up an understanding of the structure of and relationships between the variables in your dataset. Here, you will select three plots from all of your previous exploration to present here as a summary of some of your most interesting findings. Make sure that you have refined your selected plots for good titling, axis labels (with units), and good aesthetic choices (e.g. color, transparency). After each plot, make sure you justify why you chose each plot by describing what it shows.

### Plot One

### Description One

### Plot Two

### Description Two

### Plot Three

### Description Three

---

## Reflection

> **Tip**: Here’s the final step! Reflect on the exploration you performed and the insights you found. What were some of the struggles that you went through? What went well? What was surprising? Make sure you include an insight into future work that could be done with the dataset.

> **Tip**: Don’t forget to remove this, and the other **Tip** sections before saving your final work and knitting the final report!