CIS 335: Assignment 1

Stella Sterling

Part 1 & 2

Data downloaded from Kaggle: https://www.kaggle.com/rajyellow46/wine-quality

Part 3

There are 13 variables:

Name	Attribute
`type`	nominal
`fixed acidity`	numeric (ratio)
`volatile acidity`	numeric (ratio)
`citric acid`	numeric (ratio)
`residual sugar`	numeric (ratio)
`chlorides`	numeric (ratio)
`free sulfur dioxide`	numeric (ratio)
`total sulfur dioxide`	numeric (ratio)
`density`	numeric (ratio)
`pH`	numeric (interval)
`sulphates`	numeric (ratio)
`alcohol`	numeric (ratio)
`quality`	numeric (ratio), discrete could also be considered categorical

Table 2: Summary Stats for Quality Points

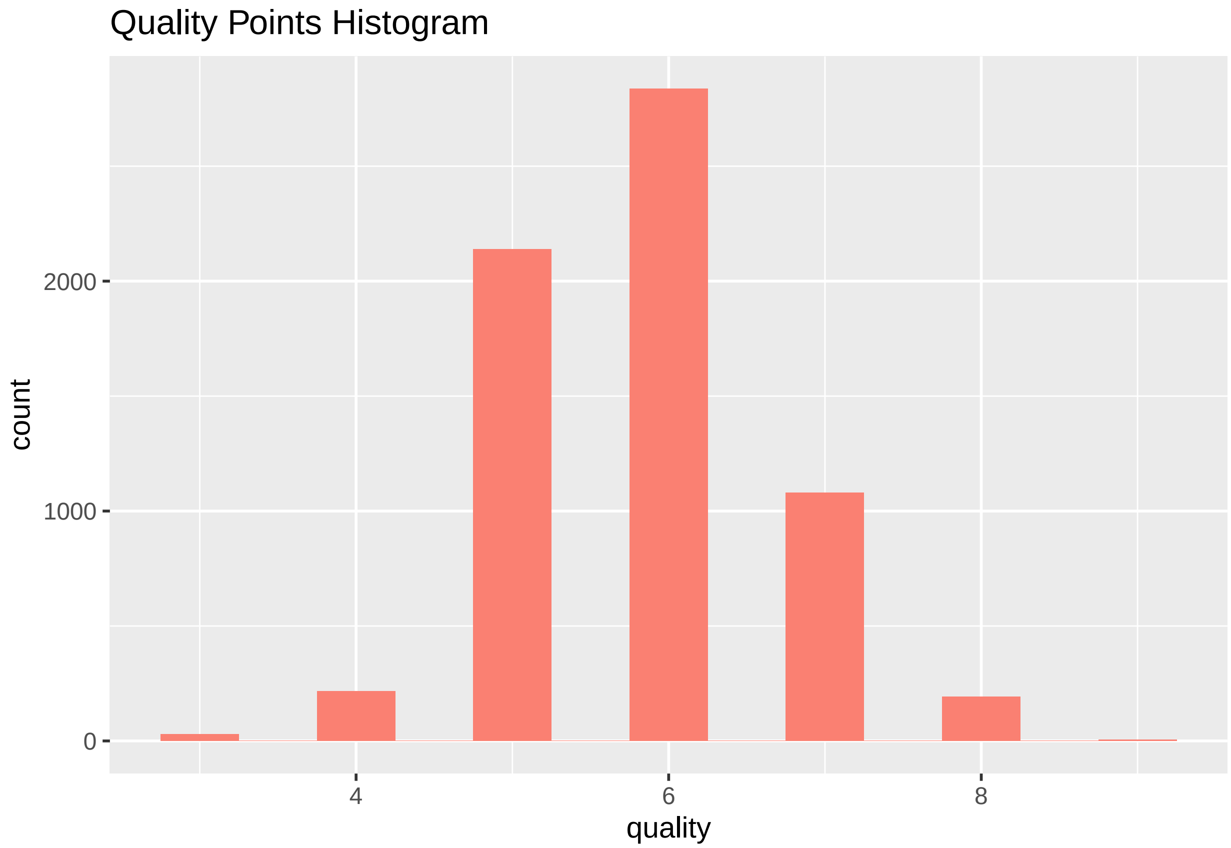
mean	median	mode	IQR
5.818378	6	6	1

Table 3: Summary Stats for Fixed Acidity

mean	median	mode	IQR
7.216579	7	6.8	1.3

Part 4

Looking at quality points, `quality`:

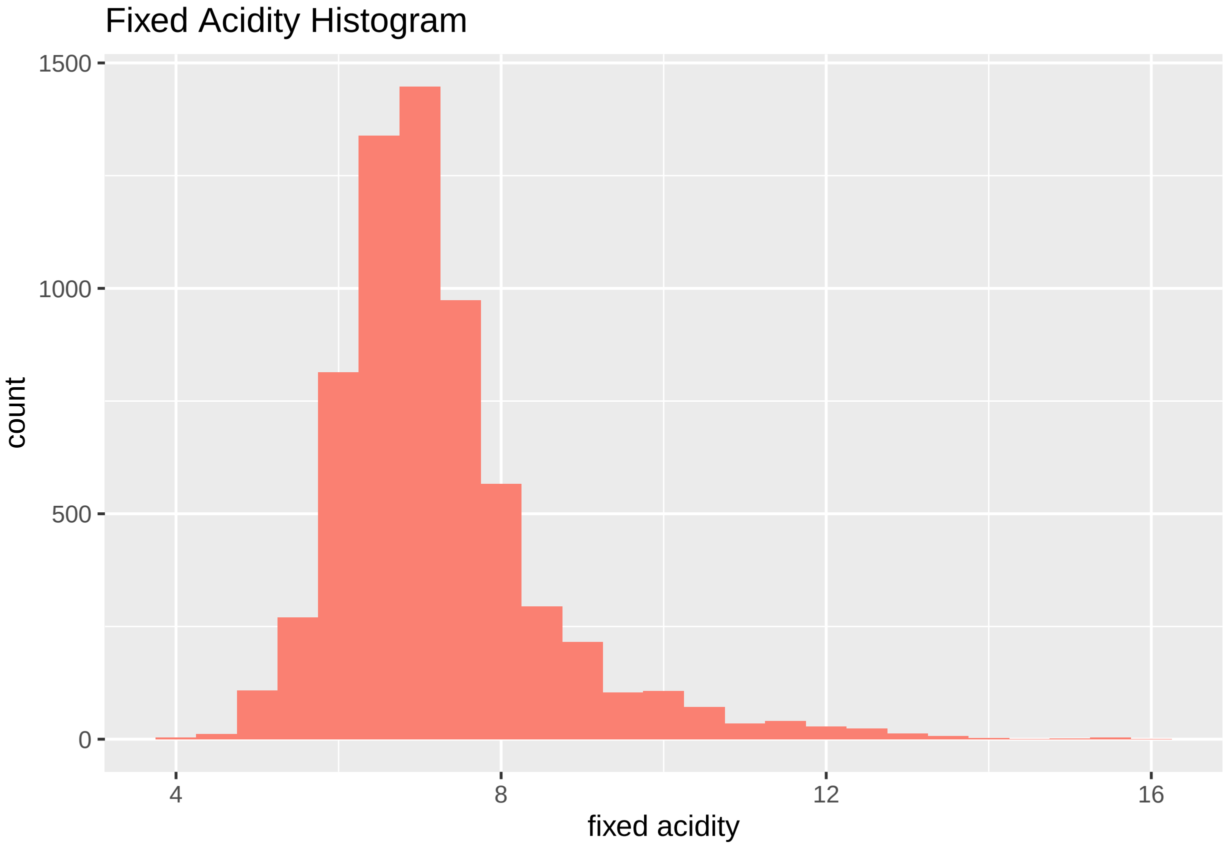


Appears normal; maybe a slight right skew. Will likely treat this as nominal.

Looking at fixed acidity, `fixed acidity`:

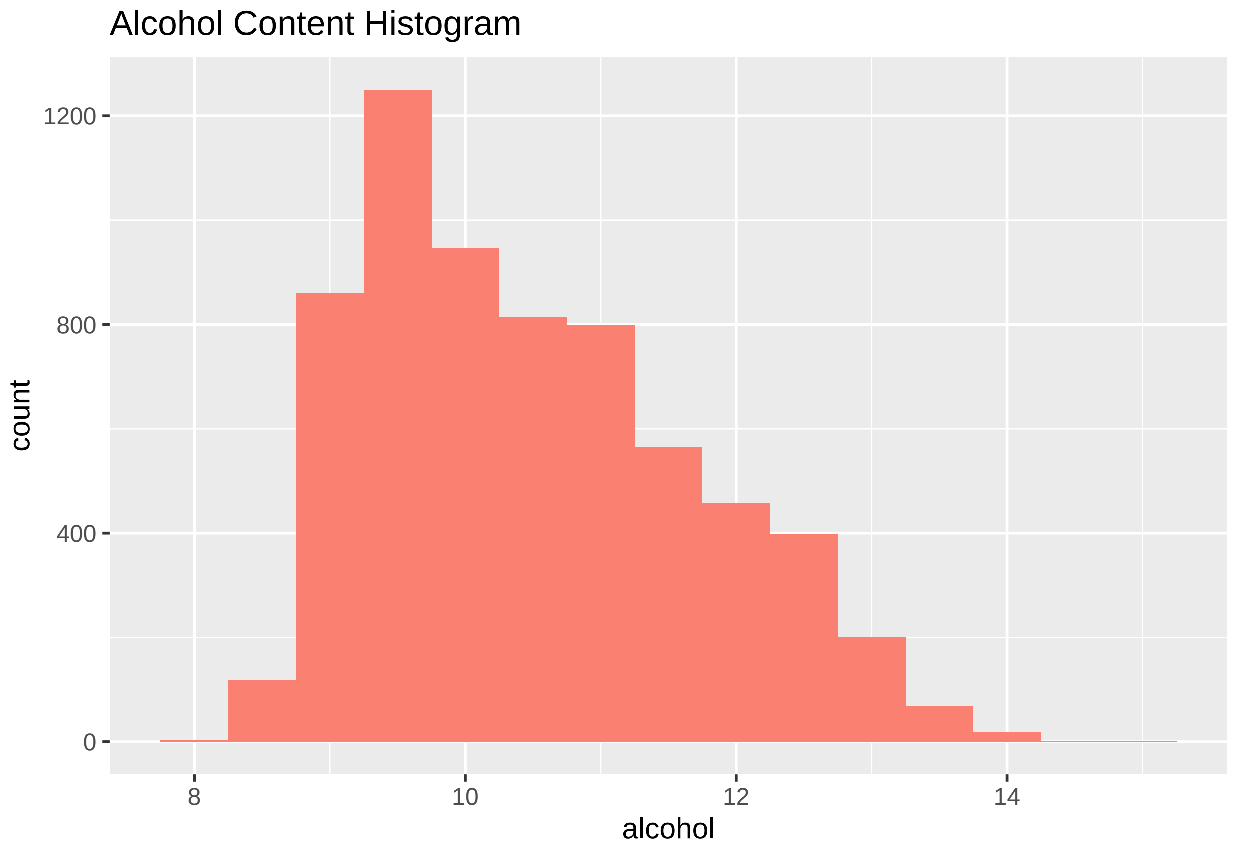
Table 4: Summary Stats for Alcohol Content

mean	median	mode	IQR
10.4918	10.3	9.5	1.8



Moderate-high right skew.

Looking at alcohol content, `alcohol`:



Highly right-skewed.

Part 5

There appears to be outliers certainly looking at the histograms of the fixed acidity and alcohol content. I computed the IQR for those two variables and quality points in the summary statistics tables above.

```
# minor outlier = + or - IQR*1.5; 1(1.5) = 1.5
5.82-1.5
```

For `quality` the mean = ~5.82 and the IQR = 1.

```
## [1] 4.32
```

```
5.82+1.5
```

```
## [1] 7.32
```

This means that values less than 4.32 and greater than 7.32 are considered outliers. Quality points are only in whole numbers, so I would consider values less than 4 and greater than 8 minor outliers. The minimum value is 3 and the maximum is 9 for quality points; these values are few and barely minor outliers so I wouldn't remove these observations.

```
7.22 - 1.5*1.3
```

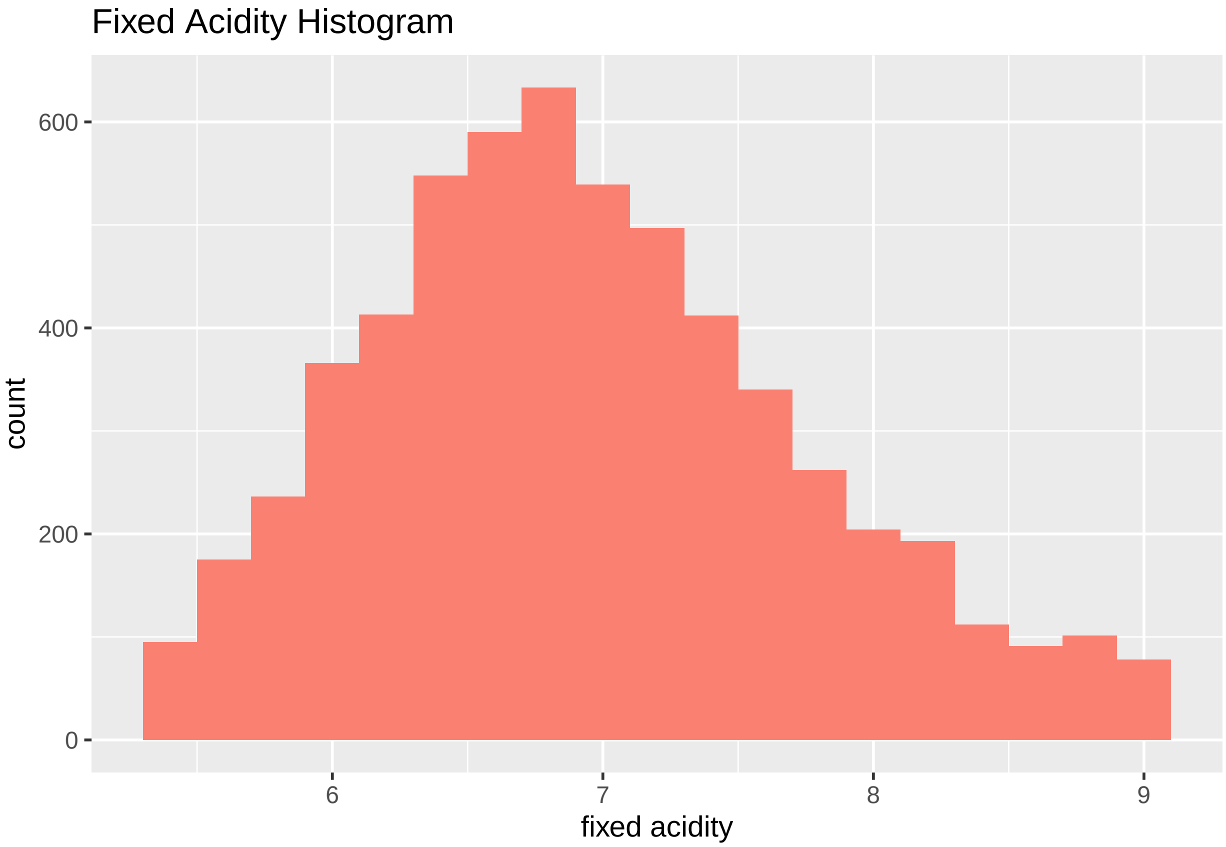
For `fixed acidity` the mean = ~7.22 and the IQR = 1.3.

```
## [1] 5.27
```

```
7.22 + 1.5*1.3
```

```
## [1] 9.17
```

Values less than 5.27 and greater than 9.17 considered minor outliers. Lots of observations would be removed (~1000). Maybe I would increase my multiplier to 3 and remove major outliers if any? What would the data look like with the minor outliers removed?



Looks better with these removed, however may consider checking out transformations before removing outliers. This is quite a large data set so I'm not too worried about removing so many observations…

```
10.49 - 1.5*1.8
```

For `alcohol` the mean = ~10.49 and the IQR = 1.8.

```
## [1] 7.79
```

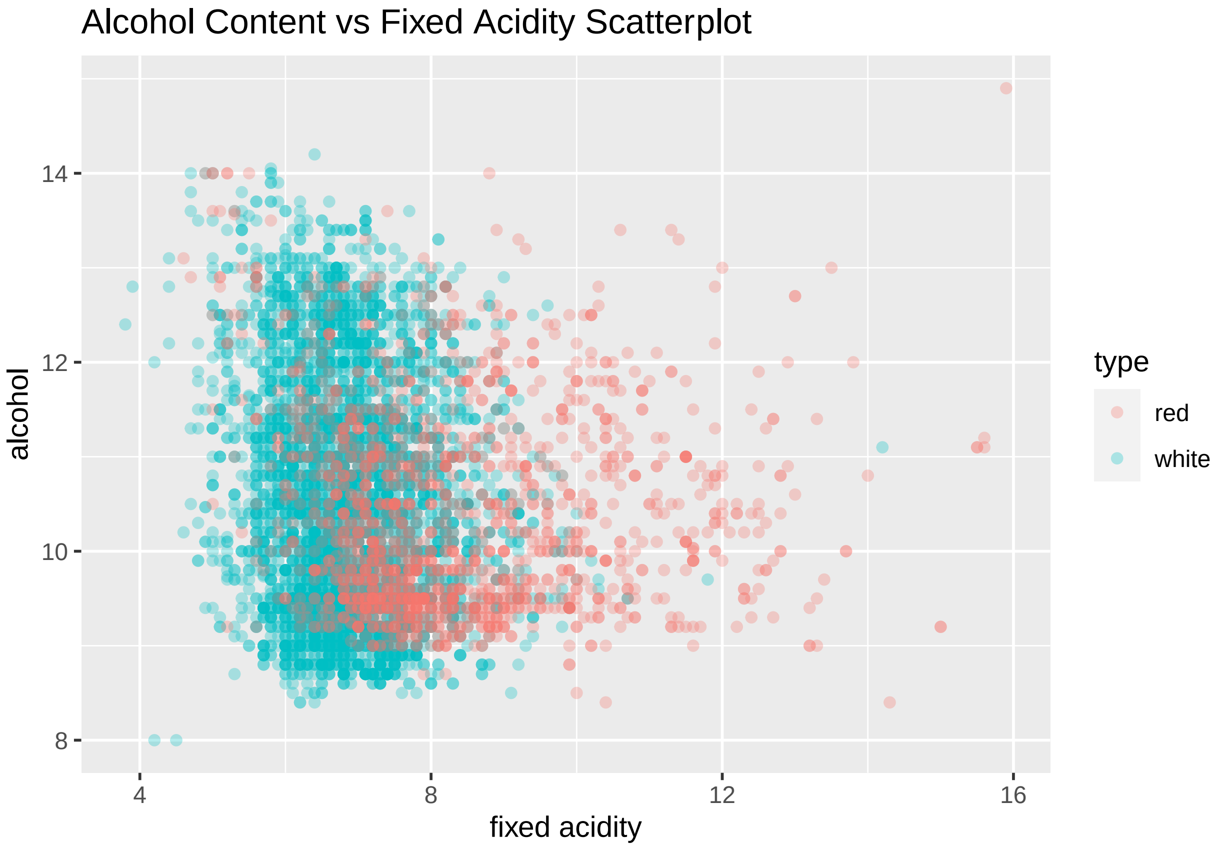
```
10.49 + 1.5*1.8
```

```
## [1] 13.19
```

Minor outliers would be less than 7.79 and greater than 13.19. This is only about 100 observations and they are pretty close to the edge so I would keep them.

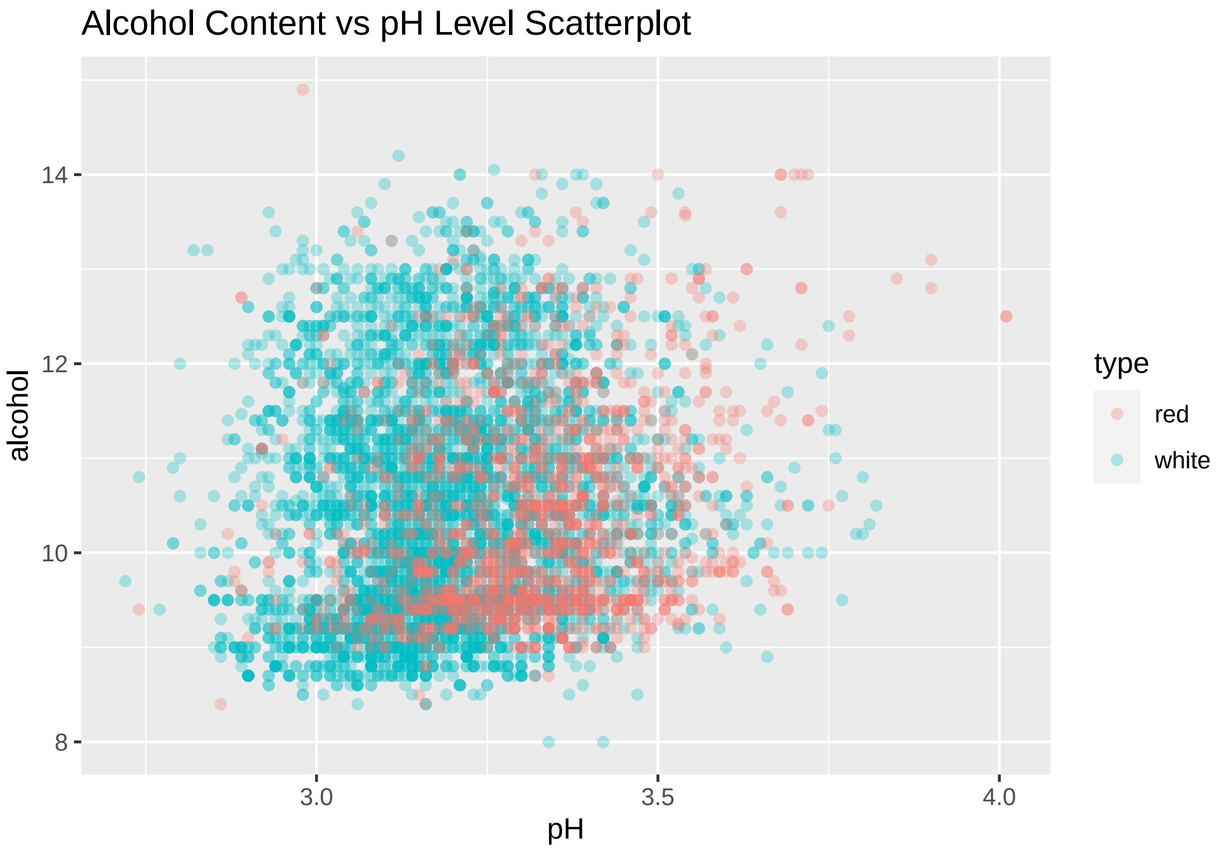
Part 6

Comparing alcohol content (`alcohol`) and `fixed acidity`:



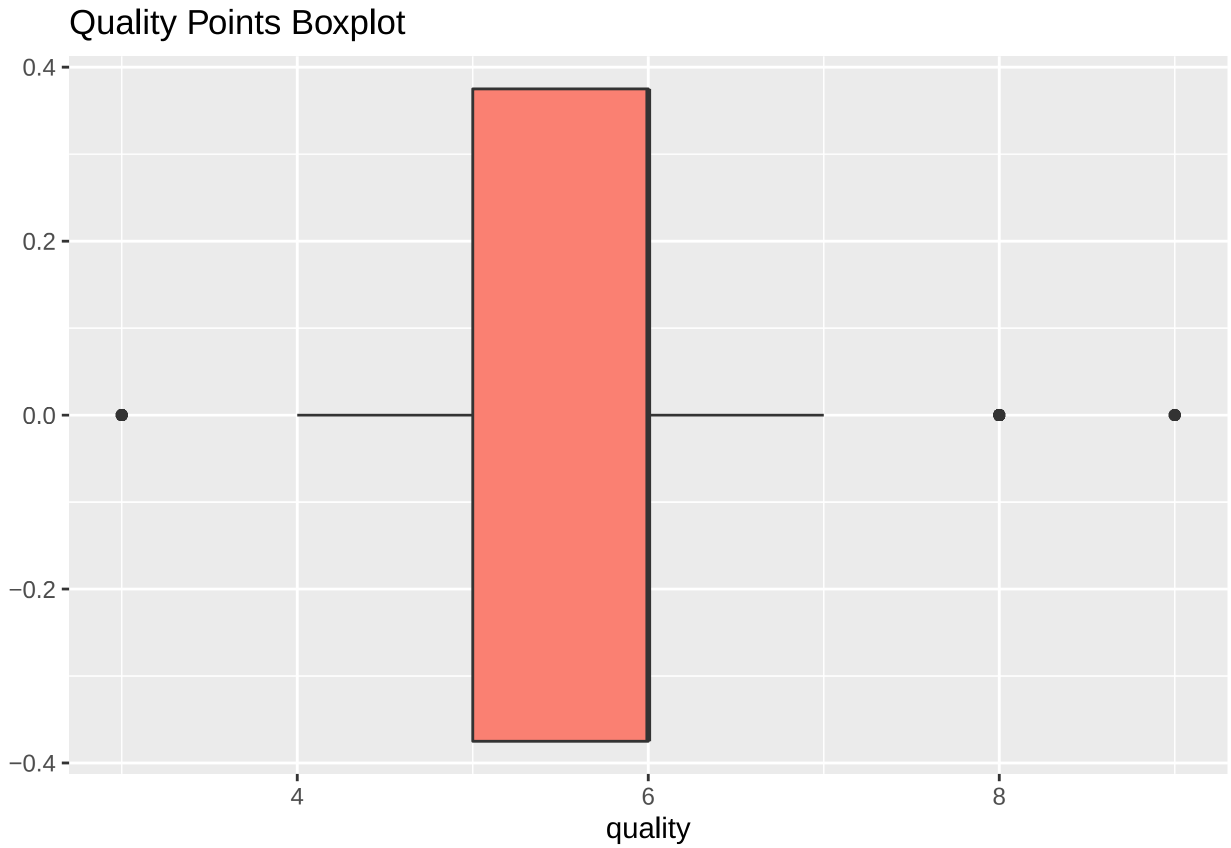
A low-moderate negative correlation. Outliers are definitely an issue. I've colored the points for fun because I think in the future it would be cool to see the differences between red and white wines.

Comparing alcohol content (`alcohol content`) and ph level (`pH`):

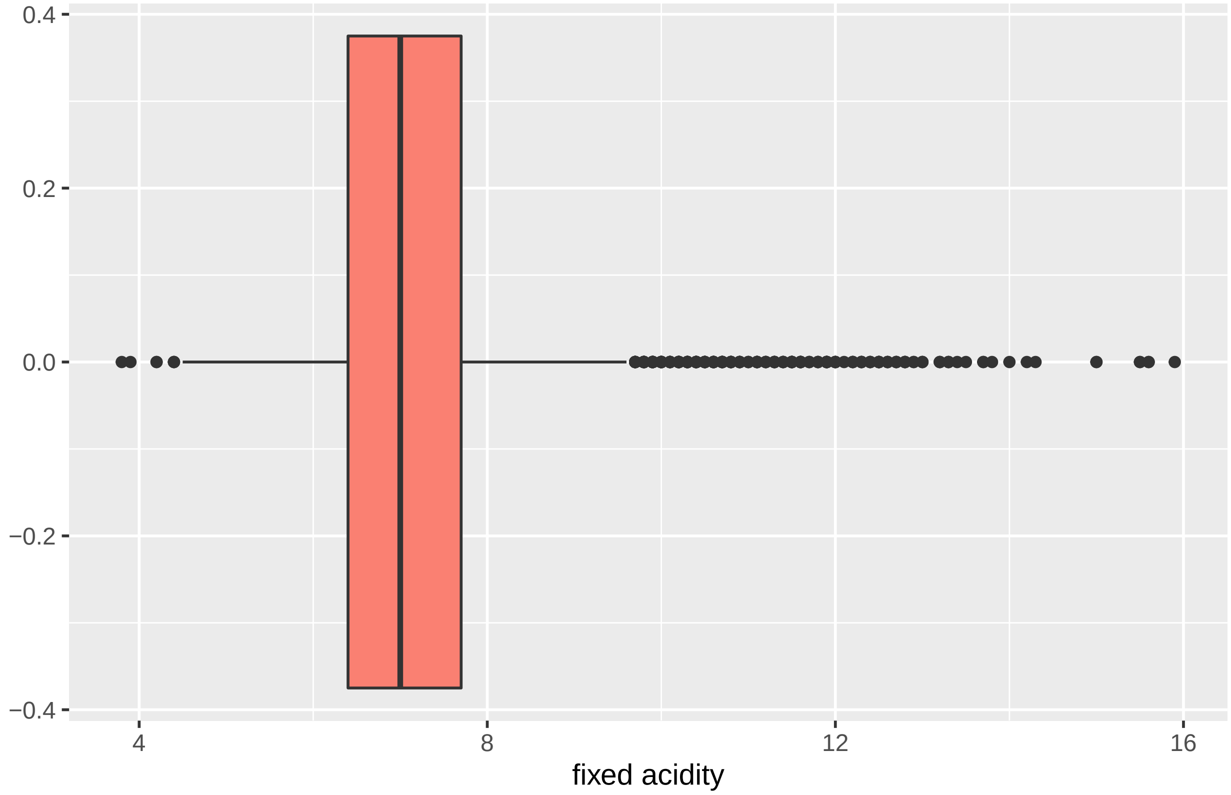


Not seeing much correlation here.

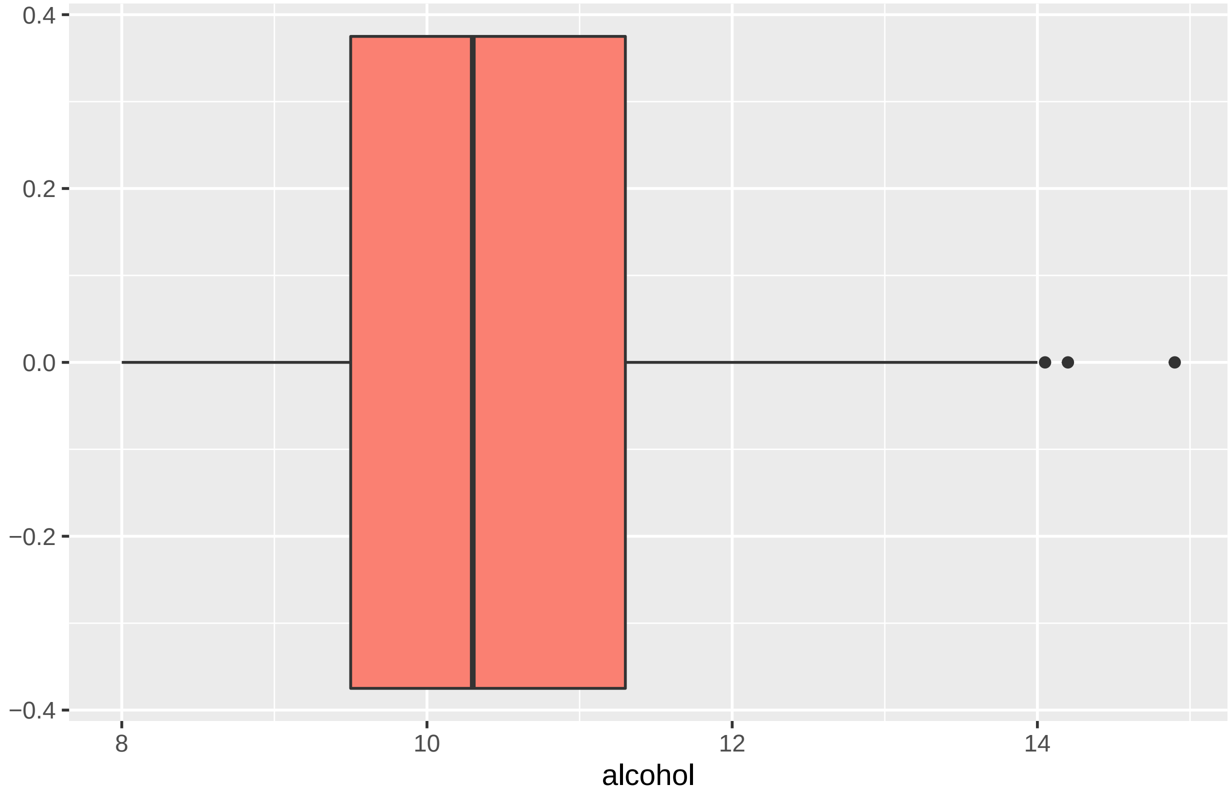
Part 7



Fixed Acidity Boxplot



Alcohol Content Boxplot


alcohol