Final Project

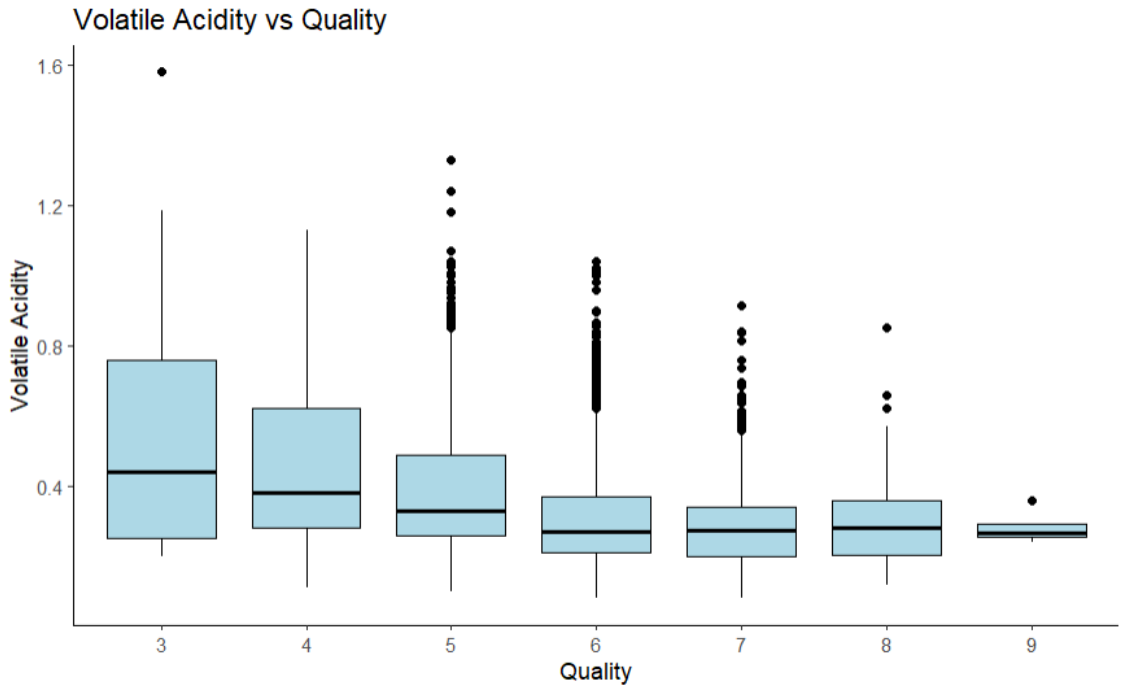
LOGAN BELL AND BOBBY CARSTENS

Data Cleaning

Fixing Typos	Replacing instances of “Californnia” being replaced with “California” in the location field
Imputing Missing Values	Missing values of type are mode imputed • We felt this was appropriate as the missing values followed a similar distribution to that of the white wine

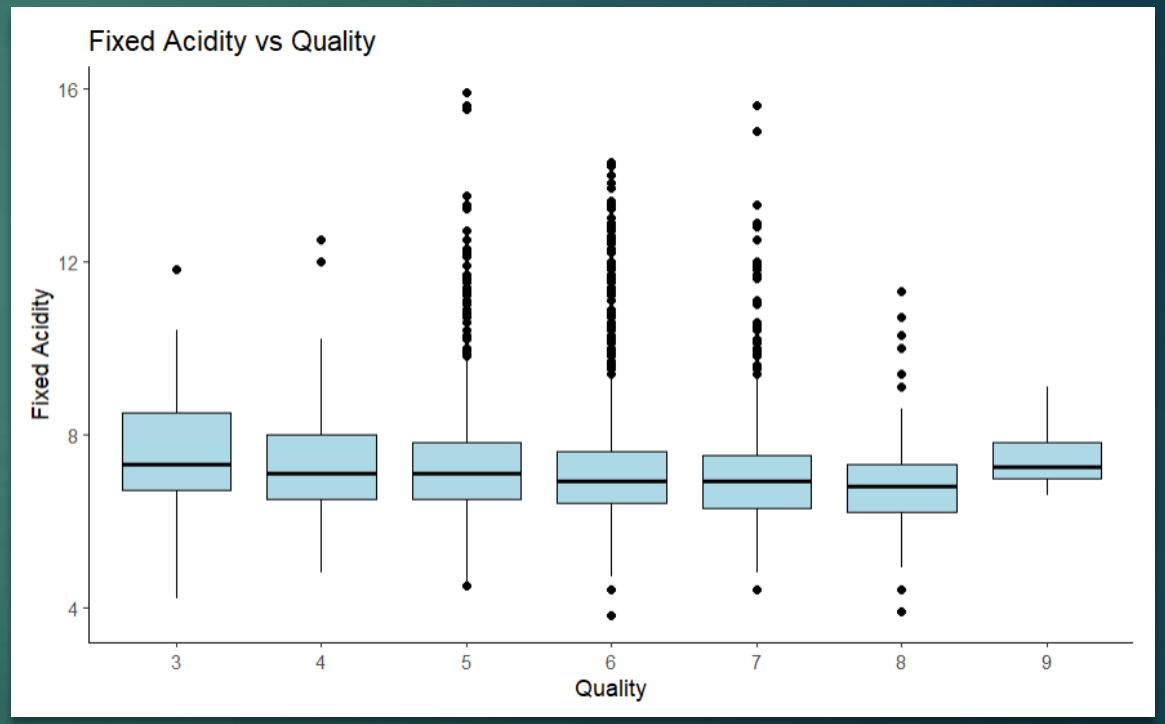
Volatile Acidity vs. Quality

- For the most part, a lower volatile acidity is associated with an average quality
- Higher volatile acidity is associated with both low and high quality wines
  - Highest average values as well as outliers are associated with lowest quality ratings



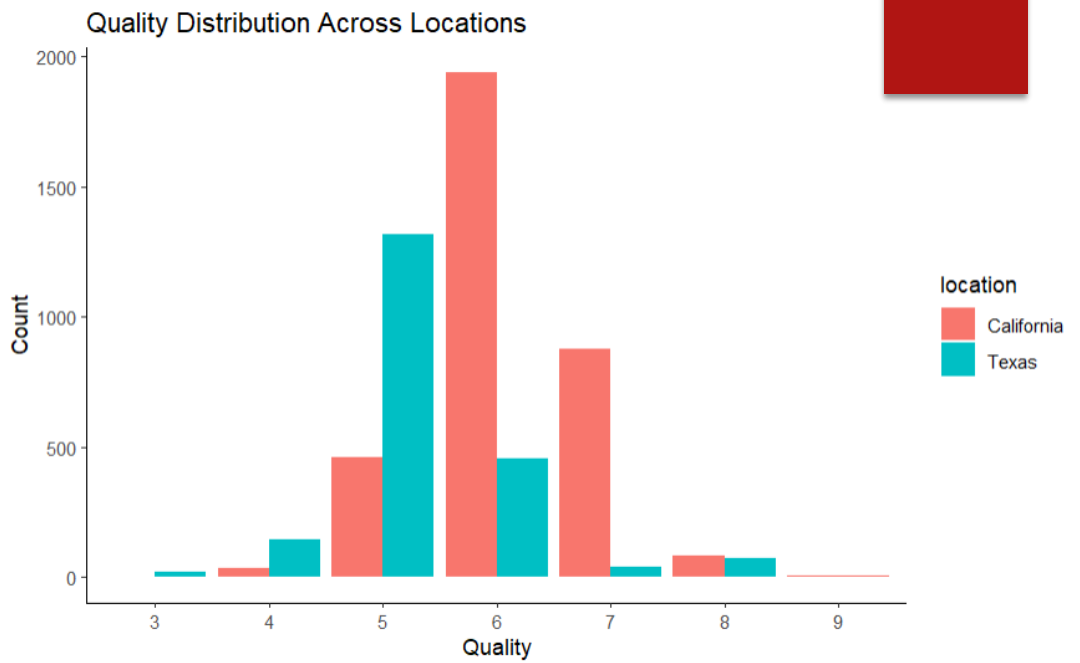
Fixed Acidity vs. Quality

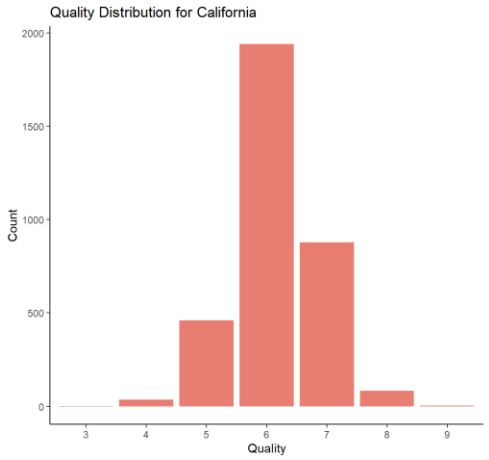
- For the most part, a lower fixed acidity is associated with a higher quality
- However, from 8 to 9, there is a jump in fixed acidity

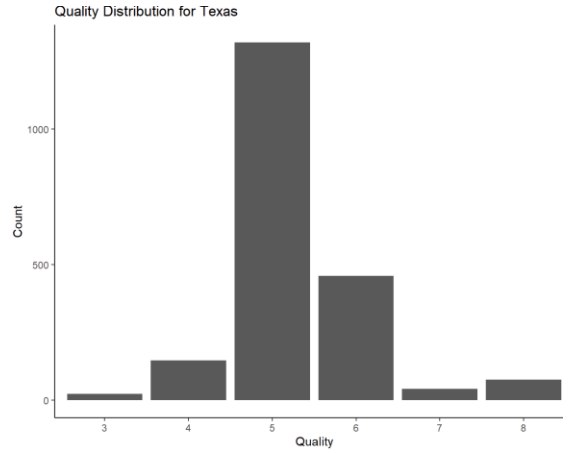


Location vs. Quality

- Californian wines seem to have a higher quality than Texas wines on average
- Both locations have a normal distribution of quality, with California's peak being 1 point higher than that of Texas.

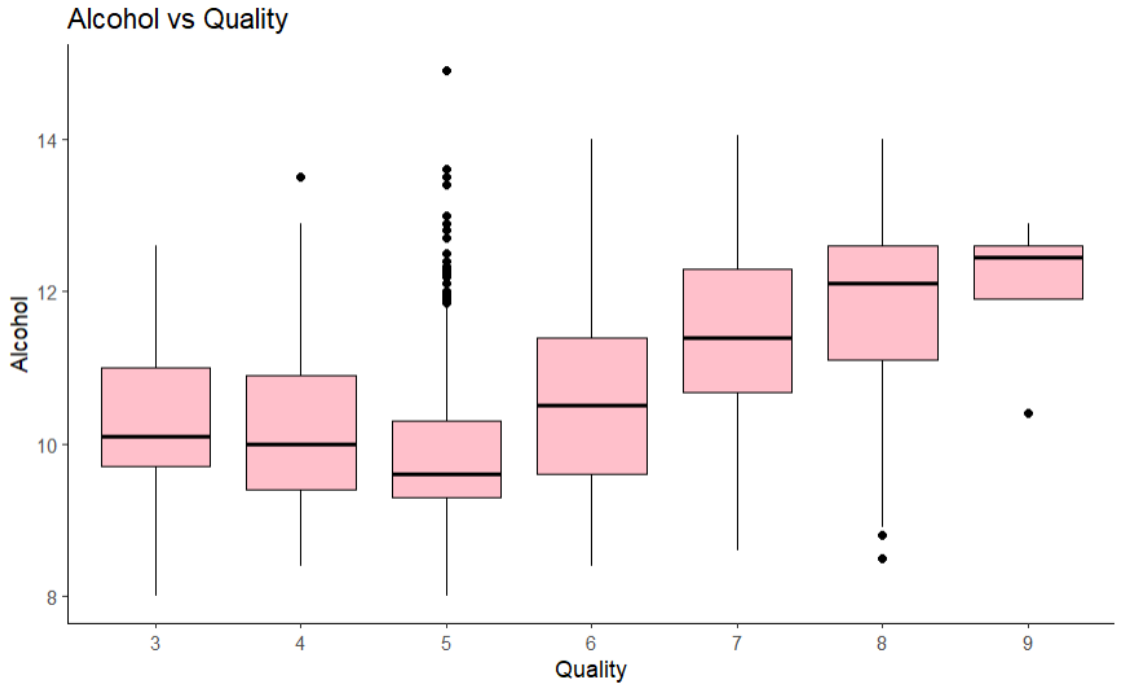






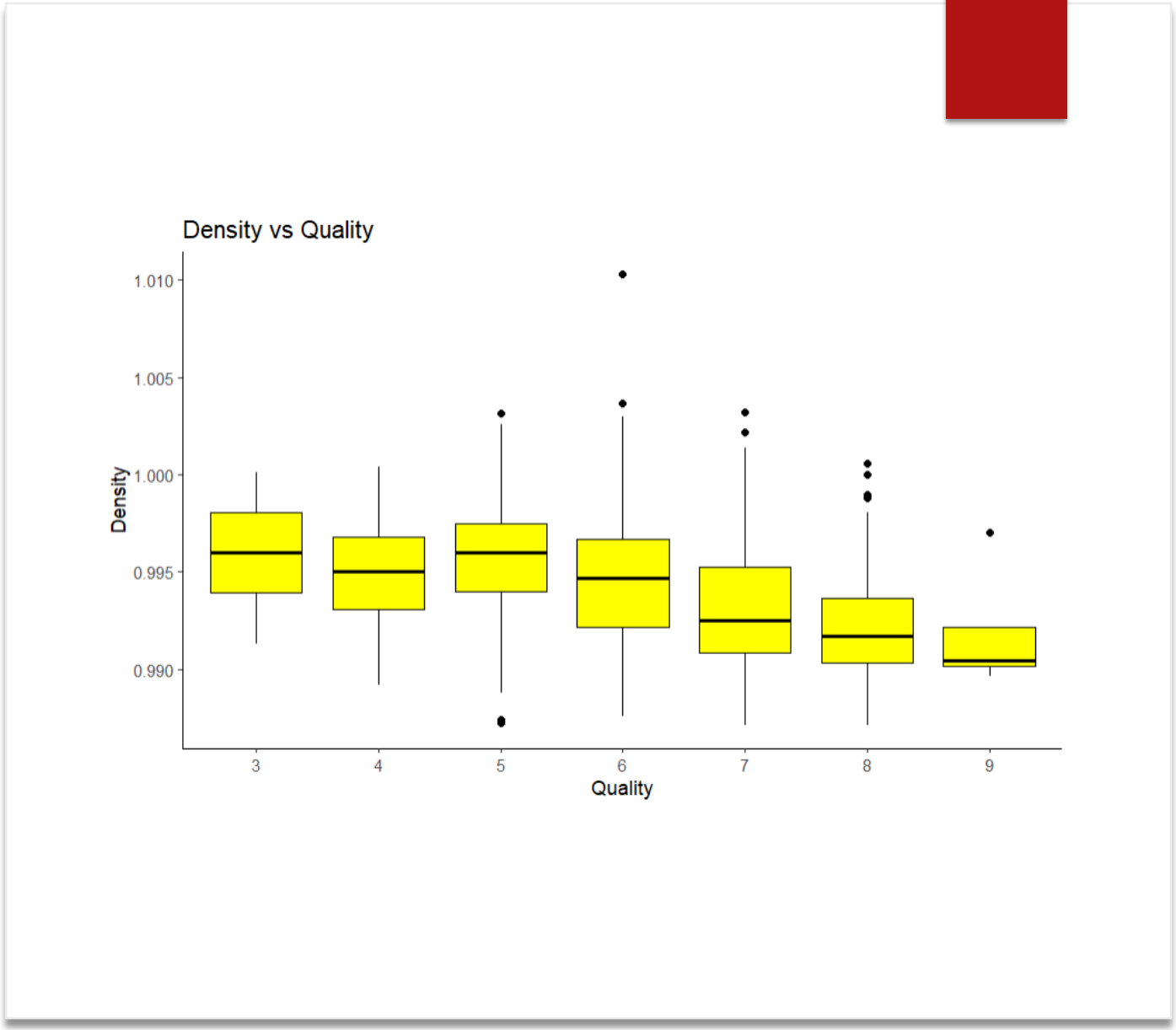
Alcohol vs. Quality

- Lower quality wines appear to be associated with lower average alcohol content as well as broad interquartile ranges.
- Quality Level 5 appears to have many outliers
- Average alcohol content steadily rises from quality levels 5-9.



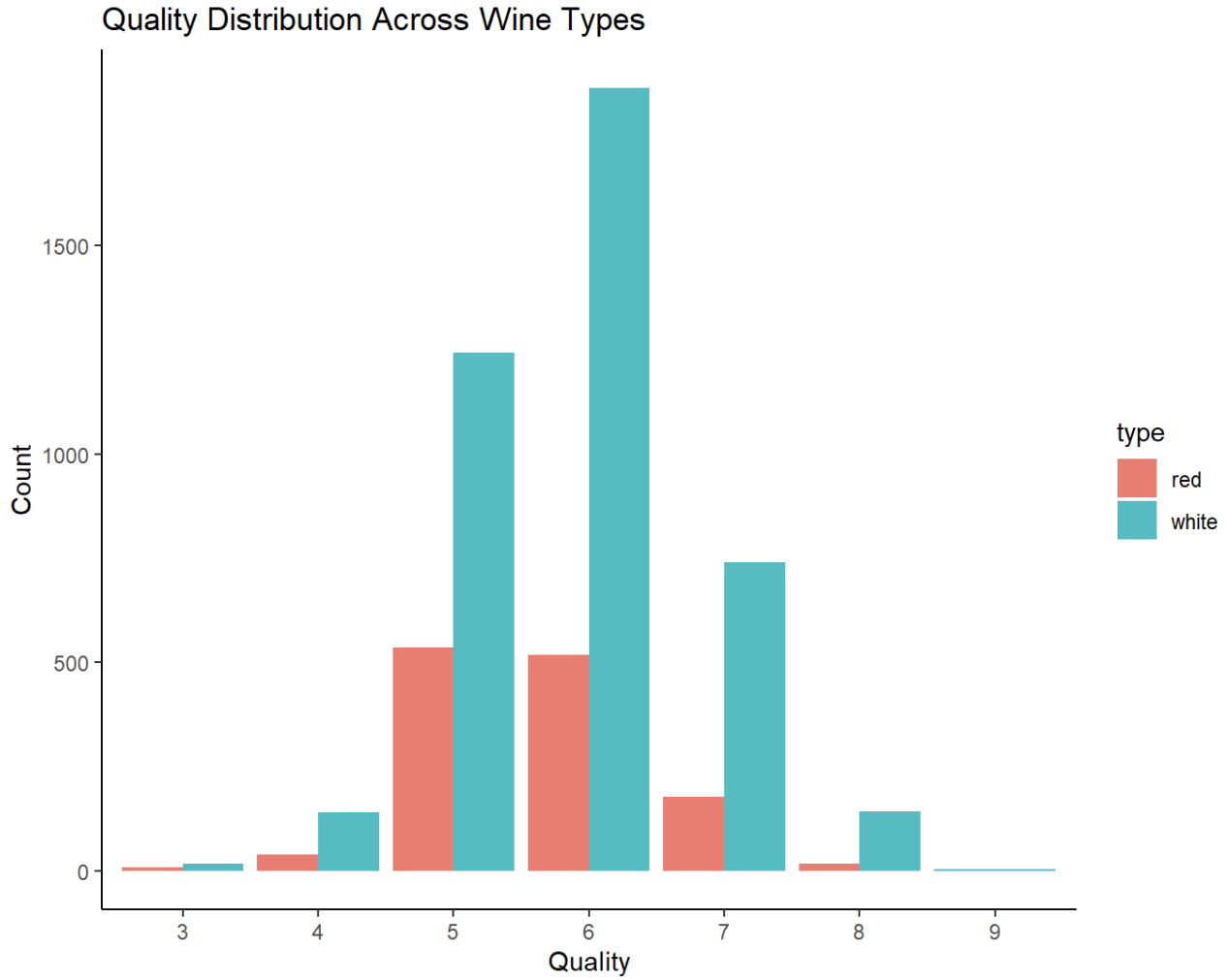
Density vs. Quality

- General trend that a lower density equals a higher quality
- Lower density is correlated with a lower average quality from quality levels 5-9.



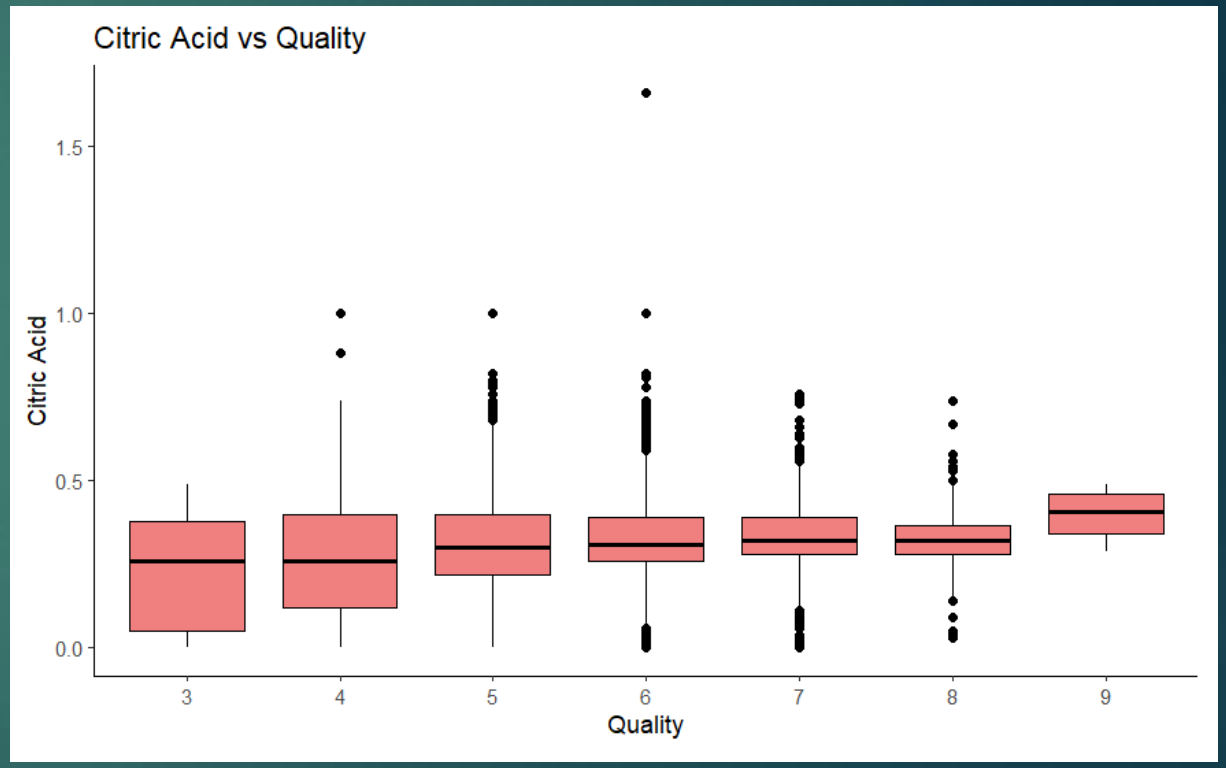
Wine Type vs. Quality

- With imputed values, there are much more white wines than red
  - Mode imputation
- White wines have a higher quality distribution than red wines
  - Correlation matrix shows a moderate correlation between type and location



Citric Acid vs. Quality

- There does not seem to be much of a relationship between citric acid and quality
- That being said, the highest quality wines look to have more citric acid on average



Conclusion

- There are many variables which can be used to predict the quality of wine
  - Location and type seem to have a moderate correlation, which can show some collinearity
- More precise evaluation needs to be done to determine which variables are significant in a multiple linear regression model