Wine Quality Analysis

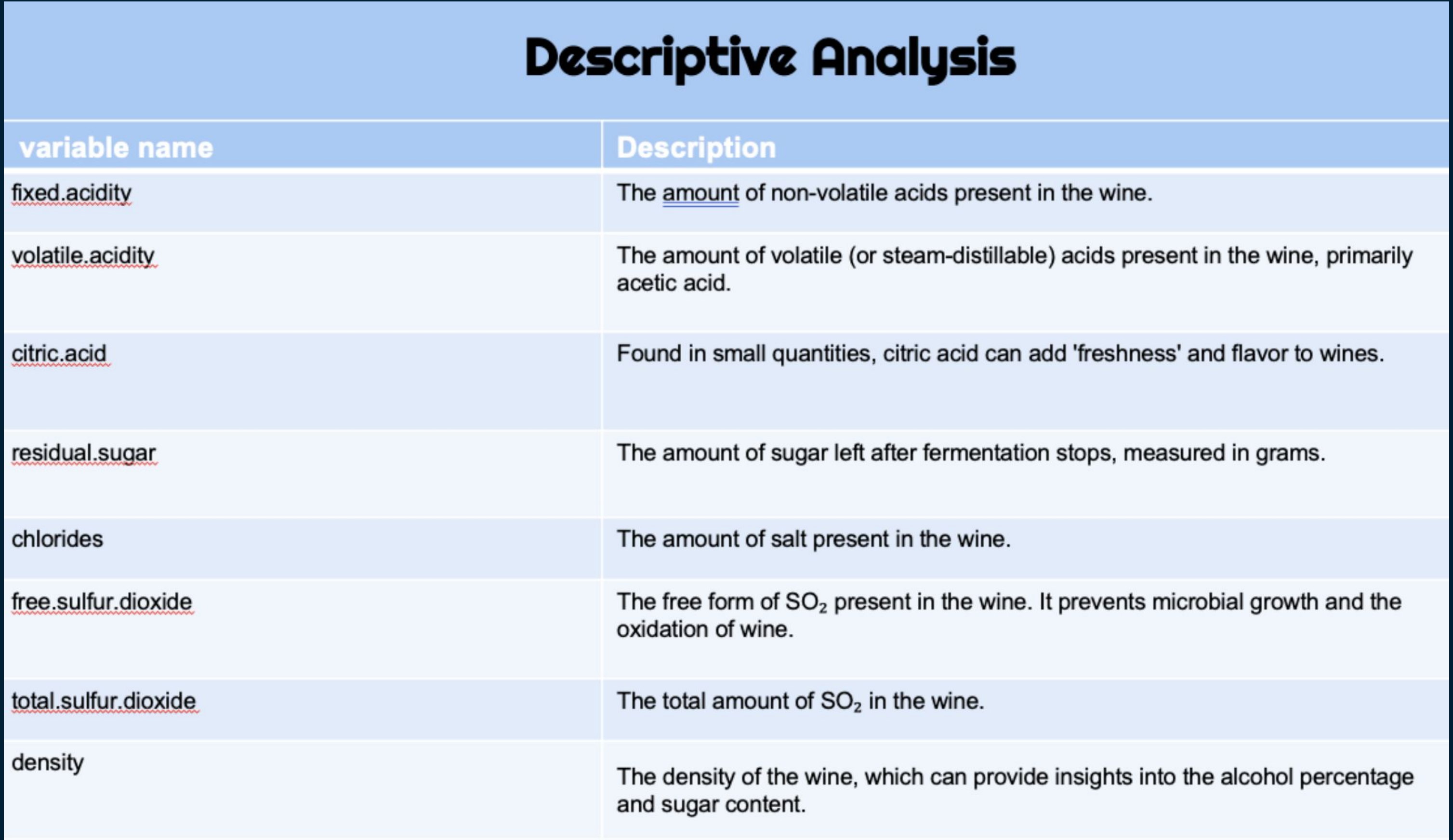
Executive Summary

Project Overview

Predicting the Quality of Wine

- Business Objectives
  - Better understand the factors related to identifying which attributes best determine the quality of the wine (the response variable is “Quality ”)
- Data Sources
  - **Wine quality dataset:** an excel file that contains a dataset of wine; it contains 15 columns of information with 1599 rows of data

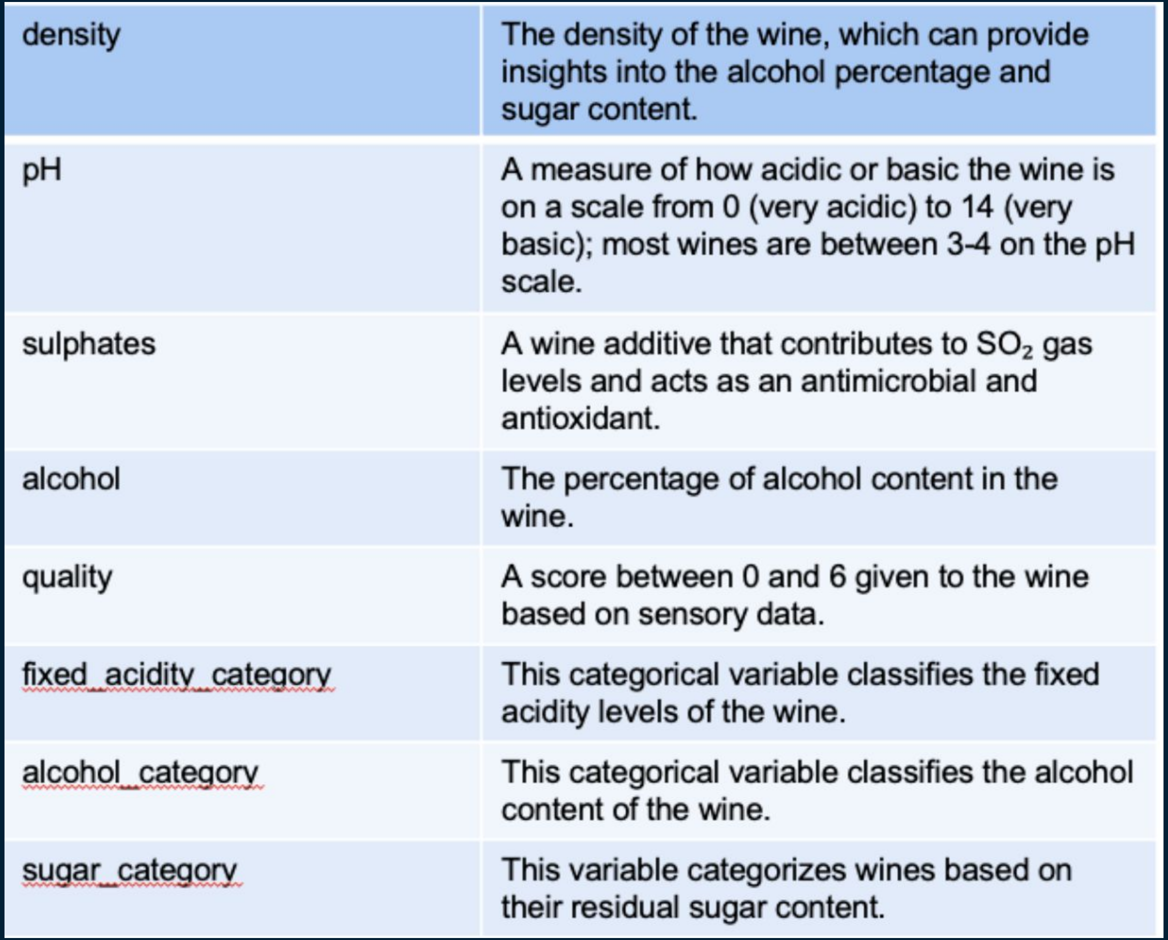
Data Dictionary



Descriptive Analysis

variable name	Description
fixed.acidity	The amount of non-volatile acids present in the wine.
volatile.acidity	The amount of volatile (or steam-distillable) acids present in the wine, primarily acetic acid.
citric.acid	Found in small quantities, citric acid can add 'freshness' and flavor to wines.
residual.sugar	The amount of sugar left after fermentation stops, measured in grams.
chlorides	The amount of salt present in the wine.
free.sulfur.dioxide	The free form of SO_2 present in the wine. It prevents microbial growth and the oxidation of wine.
total.sulfur.dioxide	The total amount of SO_2 in the wine.
density	The density of the wine, which can provide insights into the alcohol percentage and sugar content.

Data Dictionary

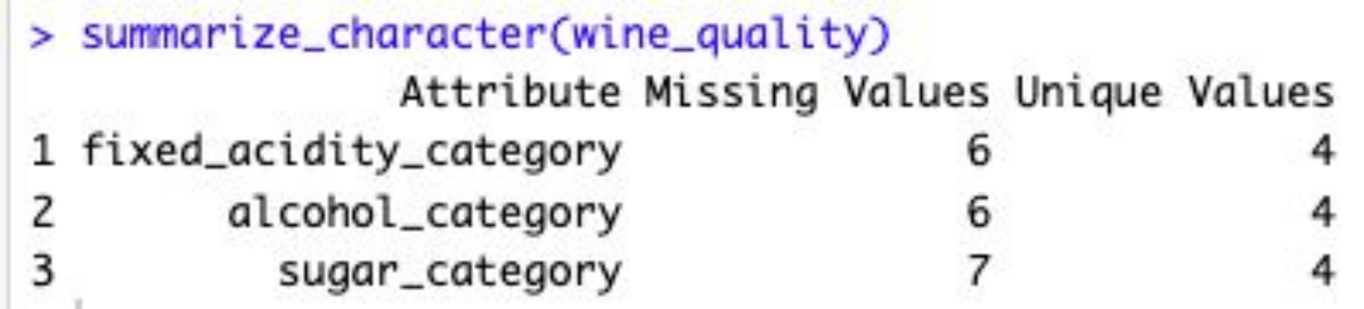


density	The density of the wine, which can provide insights into the alcohol percentage and sugar content.
pH	A measure of how acidic or basic the wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale.
sulphates	A wine additive that contributes to SO_2 gas levels and acts as an antimicrobial and antioxidant.
alcohol	The percentage of alcohol content in the wine.
quality	A score between 0 and 6 given to the wine based on sensory data.
fixed_acidity_category	This categorical variable classifies the fixed acidity levels of the wine.
alcohol_category	This categorical variable classifies the alcohol content of the wine.
sugar_category	This variable categorizes wines based on their residual sugar content.

Initial Data Review and Cleanup

Exploratory Data Analysis - Summary

Character Attributes

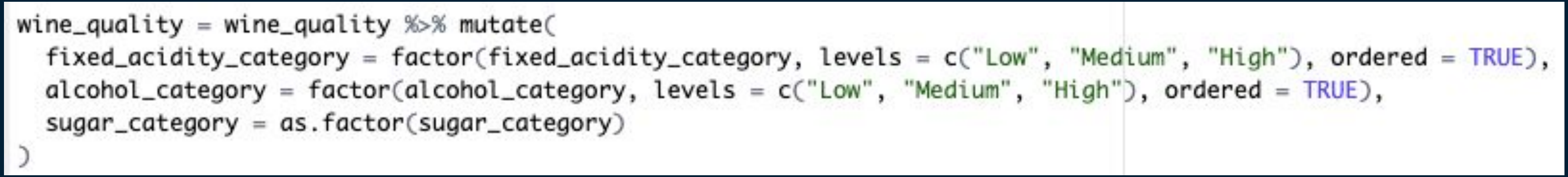


```
> summarize_character(wine_quality)
               Attribute Missing Values Unique Values
1 fixed_acidity_category              6             4
2       alcohol_category              6             4
3         sugar_category              7             4
```

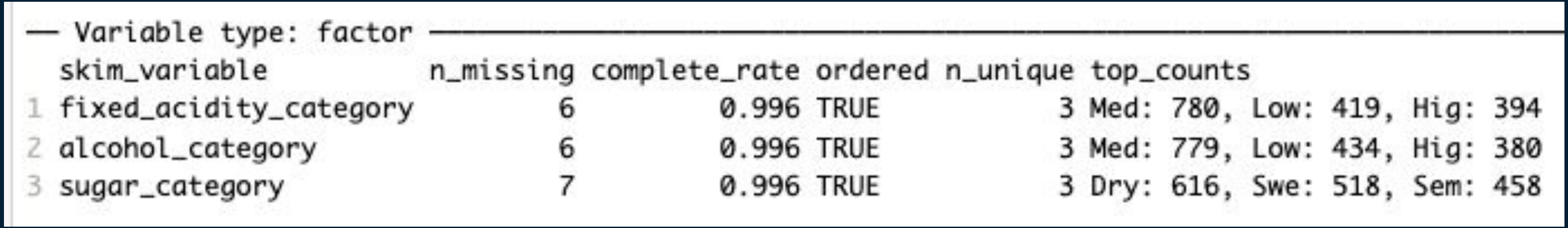
Convert the 3 categories to factors

Exploratory Data Analysis - Summary

Character Attributes to Factor Attributes



```
wine_quality = wine_quality %>% mutate(
  fixed_acidity_category = factor(fixed_acidity_category, levels = c("Low", "Medium", "High"), ordered = TRUE),
  alcohol_category = factor(alcohol_category, levels = c("Low", "Medium", "High"), ordered = TRUE),
  sugar_category = as.factor(sugar_category)
)
```

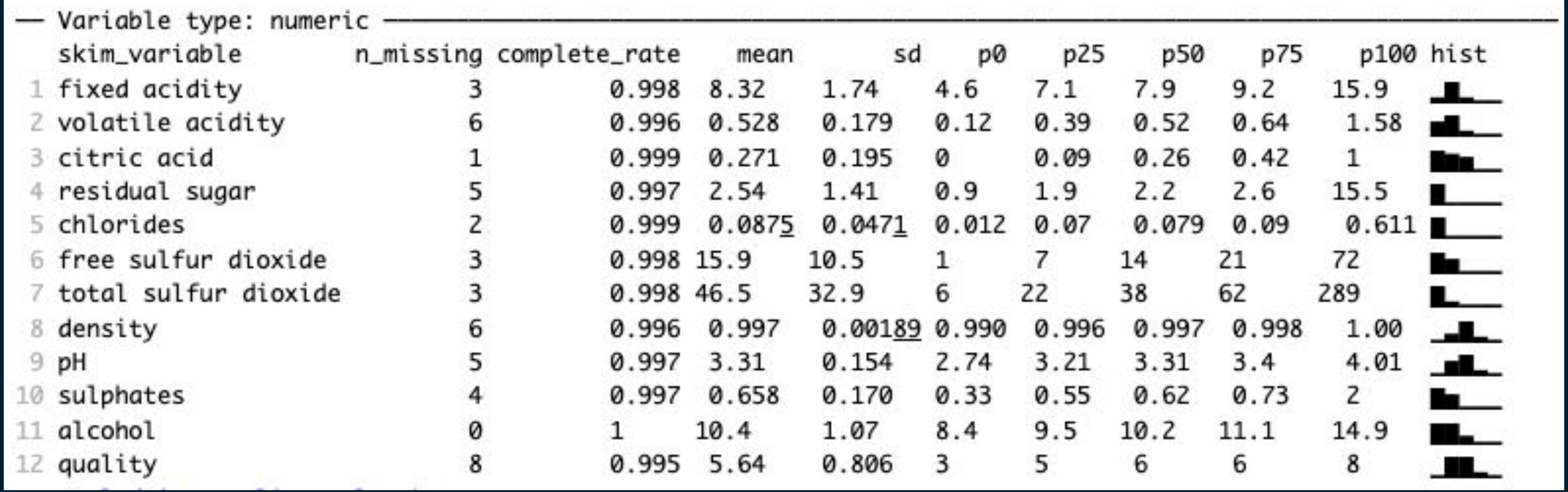


```
── Variable type: factor ──────────────────────────────────────────────────────
  skim_variable          n_missing complete_rate ordered n_unique top_counts
1 fixed_acidity_category         6         0.996 TRUE           3 Med: 780, Low: 419, Hig: 394
2 alcohol_category               6         0.996 TRUE           3 Med: 779, Low: 434, Hig: 380
3 sugar_category                 7         0.996 TRUE           3 Dry: 616, Swe: 518, Sem: 458
```

All 3 categories have values missing

Exploratory Data Analysis - Summary

Numeric Attributes



```
── Variable type: numeric ──────────────────────────────────────────────────────────────────────────────
   skim_variable        n_missing complete_rate   mean      sd      p0     p25    p50    p75    p100 hist
 1 fixed acidity                3         0.998  8.32     1.74    4.6    7.1    7.9    9.2    15.9
 2 volatile acidity             6         0.996  0.528    0.179   0.12   0.39   0.52   0.64    1.58
 3 citric acid                  1         0.999  0.271    0.195   0      0.09   0.26   0.42    1
 4 residual sugar               5         0.997  2.54     1.41    0.9    1.9    2.2    2.6    15.5
 5 chlorides                    2         0.999  0.0875   0.0471  0.012  0.07   0.079  0.09    0.611
 6 free sulfur dioxide          3         0.998 15.9     10.5     1      7     14     21      72
 7 total sulfur dioxide         3         0.998 46.5     32.9     6     22     38     62     289
 8 density                      6         0.996  0.997    0.00189 0.990  0.996  0.997  0.998   1.00
 9 pH                           5         0.997  3.31     0.154   2.74   3.21   3.31   3.4     4.01
10 sulphates                    4         0.997  0.658    0.170   0.33   0.55   0.62   0.73    2
11 alcohol                      0         1     10.4      1.07    8.4    9.5   10.2   11.1    14.9
12 quality                      8         0.995  5.64     0.806   3      5      6      6       8
```

11 out of 12 variables have missing values

Exploratory Data Analysis - Summary

Initial Observations

- Data quality overview
  - There are missing values for almost every variable except for alcohol
  - Data appears to contain suitable “response variable”: Quality but it is missing 8 values
  - Data may contain outliers in free sulfur dioxide and total sulfur dioxide
- Composition
  - – 12 numeric variables
    - 11 out of 12 variables having missing values (alcohol doesn’t have missing values)
  - – 3 factor variables
    - All are categories

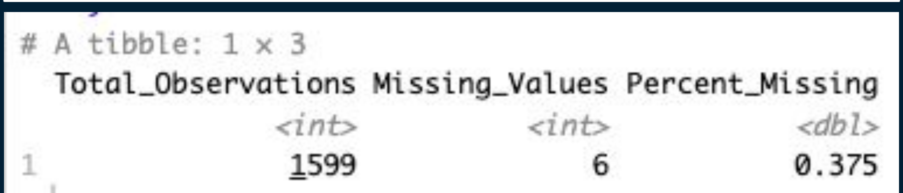
Data Cleaning

Factors

Data Cleansing

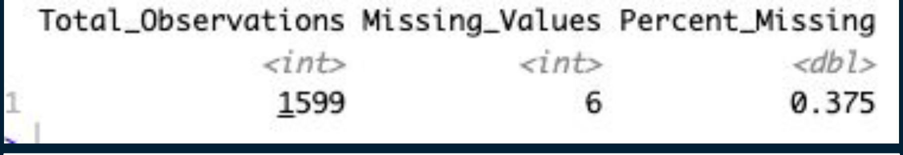
Investigate factors missing values

Fixed Acidity Category



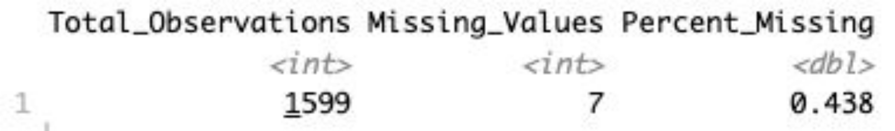
```
# A tibble: 1 × 3
  Total_Observations Missing_Values Percent_Missing
               <int>          <int>           <dbl>
1               1599              6           0.375
```

Alcohol Category



```
  Total_Observations Missing_Values Percent_Missing
               <int>          <int>           <dbl>
1               1599              6           0.375
```

Sugar Category



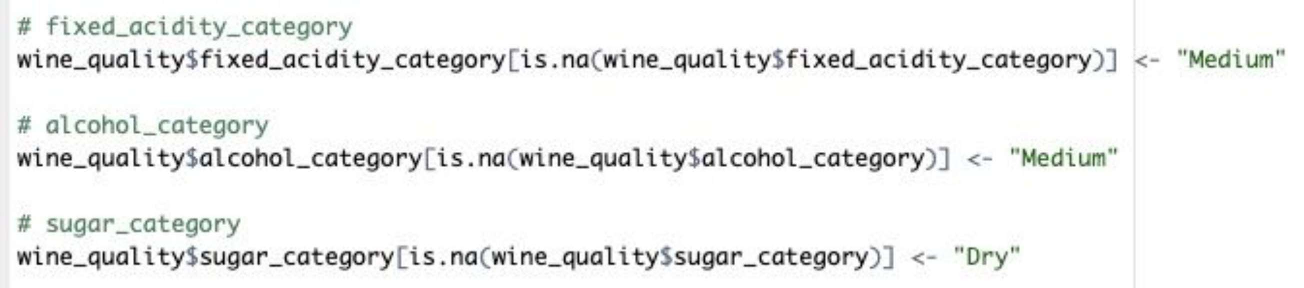
```
  Total_Observations Missing_Values Percent_Missing
               <int>          <int>           <dbl>
1               1599              7           0.438
```

Based on this analysis, the percentage of missing values for the 3 factors don’t appear to have a significant impact on the dataset but we don’t want to delete rows since the data could be meaningful

Data Cleansing

Investigate factors missing values

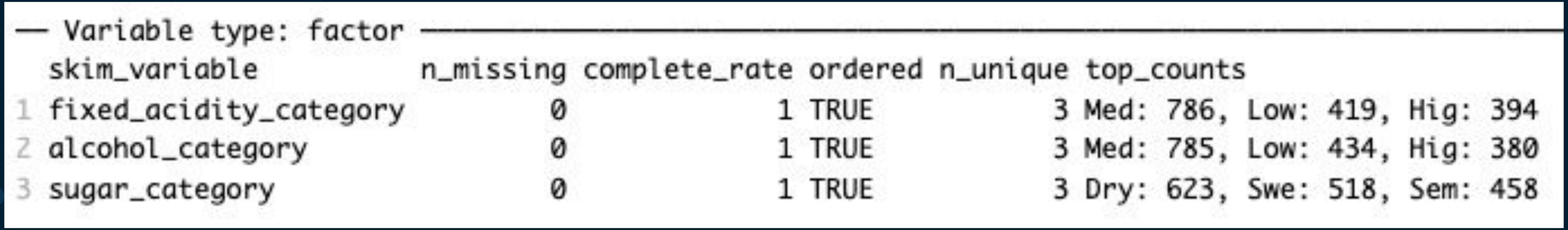
Approach: Impute most frequent level



```r
# fixed_acidity_category
wine_quality$fixed_acidity_category[is.na(wine_quality$fixed_acidity_category)] <- "Medium"

# alcohol_category
wine_quality$alcohol_category[is.na(wine_quality$alcohol_category)] <- "Medium"

# sugar_category
wine_quality$sugar_category[is.na(wine_quality$sugar_category)] <- "Dry"
```



```
── Variable type: factor ──────────────────────────────────────────────────────
  skim_variable          n_missing complete_rate ordered n_unique top_counts
1 fixed_acidity_category         0             1 TRUE           3 Med: 786, Low: 419, Hig: 394
2 alcohol_category               0             1 TRUE           3 Med: 785, Low: 434, Hig: 380
3 sugar_category                 0             1 TRUE           3 Dry: 623, Swe: 518, Sem: 458
```

No missing values for 3 factors

Data Cleaning

Numeric

Investigate numeric missing values

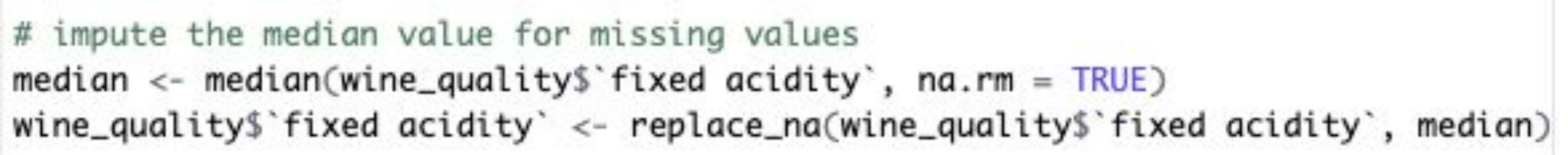
Variable	Missing Values	Percent Missing
Fixed acidity	3	0.188
volatile acidity	6	0.375
citric acid	1	0.0625
residual sugar	5	0.313
chlorides	2	0.125
free sulfur dioxide	3	0.188
total sulfur dioxide	3	0.188
density	6	0.375
pH	5	0.313
sulphates	4	0.250
quality	8	0.500
Total	46	2.88%

Based on this analysis, the percentage of missing values don’t appear to have a significant impact on the dataset but since there are missing values on 11 of the 12 numeric values, we don’t want to delete the rows that could have meaningful information

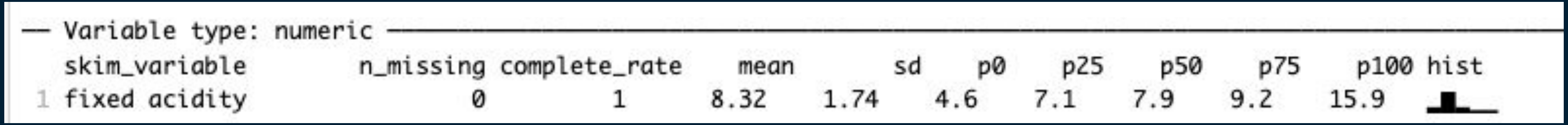
Data Cleaning

Investigate numeric missing values

Approach 2: Impute with median value for fixed acidity



```r
# impute the median value for missing values
median <- median(wine_quality$`fixed acidity`, na.rm = TRUE)
wine_quality$`fixed acidity` <- replace_na(wine_quality$`fixed acidity`, median)
```



```
── Variable type: numeric ──────────────────────────────────────────────────────────
  skim_variable        n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
1 fixed acidity                0             1  8.32  1.74   4.6   7.1   7.9   9.2  15.9 ▁▇▂▁▁
```

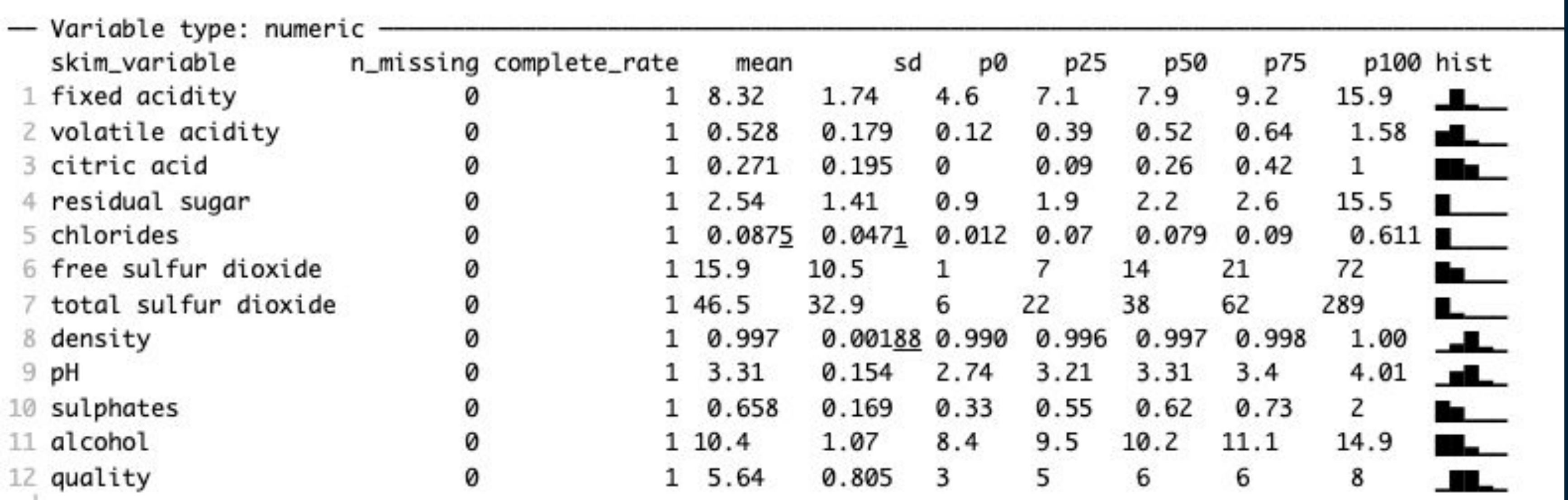
No missing values for variable

Data Cleaning

Investigate numeric missing values

Approach 2: Impute with median value for missing values

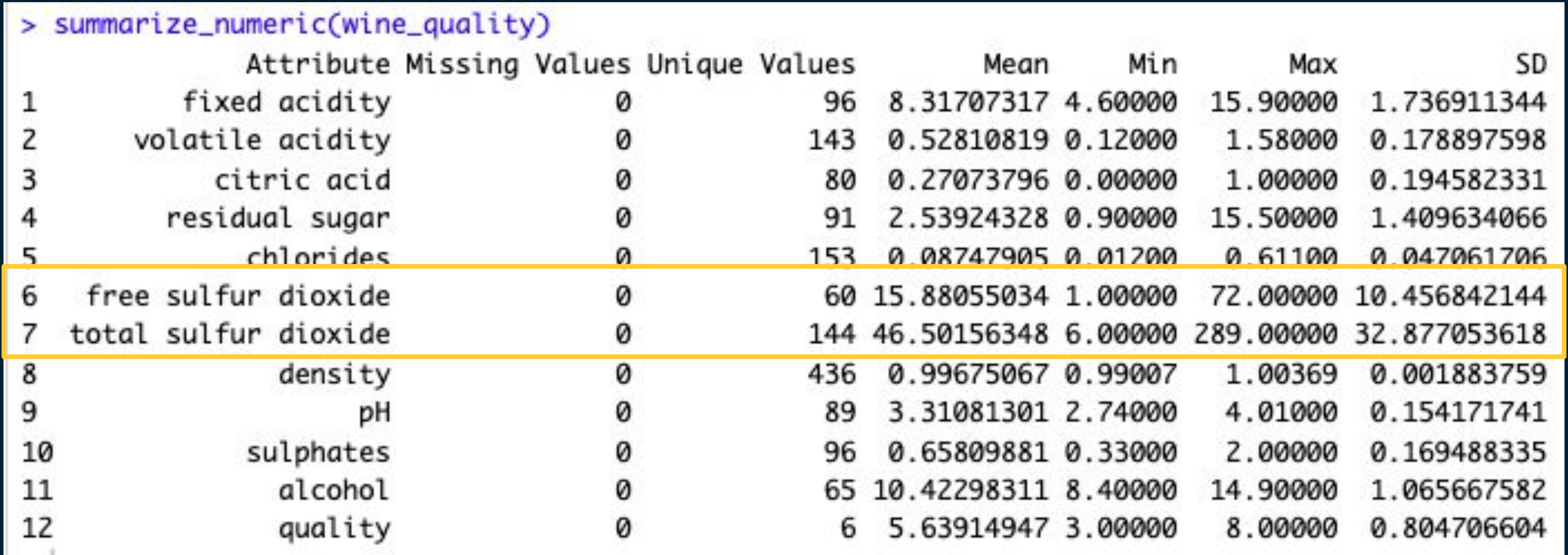
Repeat for rest of variables with missing values



```
── Variable type: numeric ──────────────────────────────────────────────────────────────────
   skim_variable        n_missing complete_rate    mean      sd     p0     p25    p50    p75    p100 hist
 1 fixed acidity                0             1  8.32    1.74    4.6    7.1    7.9    9.2    15.9  ▂▇▂▁▁
 2 volatile acidity             0             1  0.528   0.179   0.12   0.39   0.52   0.64    1.58 ▅▇▂▁▁
 3 citric acid                  0             1  0.271   0.195   0      0.09   0.26   0.42   1     ▇▇▅▁▁
 4 residual sugar               0             1  2.54    1.41    0.9    1.9    2.2    2.6    15.5  ▇▁▁▁▁
 5 chlorides                    0             1  0.0875  0.0471  0.012  0.07   0.079  0.09    0.611 ▇▁▁▁▁
 6 free sulfur dioxide          0             1 15.9    10.5     1      7     14     21      72     ▇▅▁▁▁
 7 total sulfur dioxide         0             1 46.5    32.9     6     22     38     62     289     ▇▂▁▁▁
 8 density                      0             1  0.997   0.00188 0.990  0.996  0.997  0.998   1.00 ▁▃▇▂▁
 9 pH                           0             1  3.31    0.154   2.74   3.21   3.31   3.4     4.01 ▁▅▇▂▁
10 sulphates                    0             1  0.658   0.169   0.33   0.55   0.62   0.73   2     ▇▅▁▁▁
11 alcohol                      0             1 10.4     1.07    8.4    9.5   10.2   11.1    14.9   ▇▇▃▁▁
12 quality                      0             1  5.64    0.805   3      5      6      6      8     ▁▇▇▂▁
```

Data Cleaning

Investigate numeric values for outliers



```
> summarize_numeric(wine_quality)
              Attribute Missing Values Unique Values        Mean     Min       Max           SD
1         fixed acidity              0            96  8.31707317 4.60000  15.90000  1.736911344
2      volatile acidity              0           143  0.52810819 0.12000   1.58000  0.178897598
3           citric acid              0            80  0.27073796 0.00000   1.00000  0.194582331
4        residual sugar              0            91  2.53924328 0.90000  15.50000  1.409634066
5             chlorides              0           153  0.08747905 0.01200   0.61100  0.047061706
6   free sulfur dioxide              0            60 15.88055034 1.00000  72.00000 10.456842144
7  total sulfur dioxide              0           144 46.50156348 6.00000 289.00000 32.877053618
8               density              0           436  0.99675067 0.99007   1.00369  0.001883759
9                    pH              0            89  3.31081301 2.74000   4.01000  0.154171741
10            sulphates              0            96  0.65809881 0.33000   2.00000  0.169488335
11              alcohol              0            65 10.42298311 8.40000  14.90000  1.065667582
12              quality              0             6  5.63914947 3.00000   8.00000  0.804706604
```

Free sulfur dioxide and total sulfur dioxide seem to have outliers but we didn’t remove/impute them because we still need to analyze the dataset. Other variables seem to have potential outliers but we will leave them to analyze the whole dataset first

Exploratory Data Analysis - Summary

Logical Groupings of Attributes

Numeric (12)

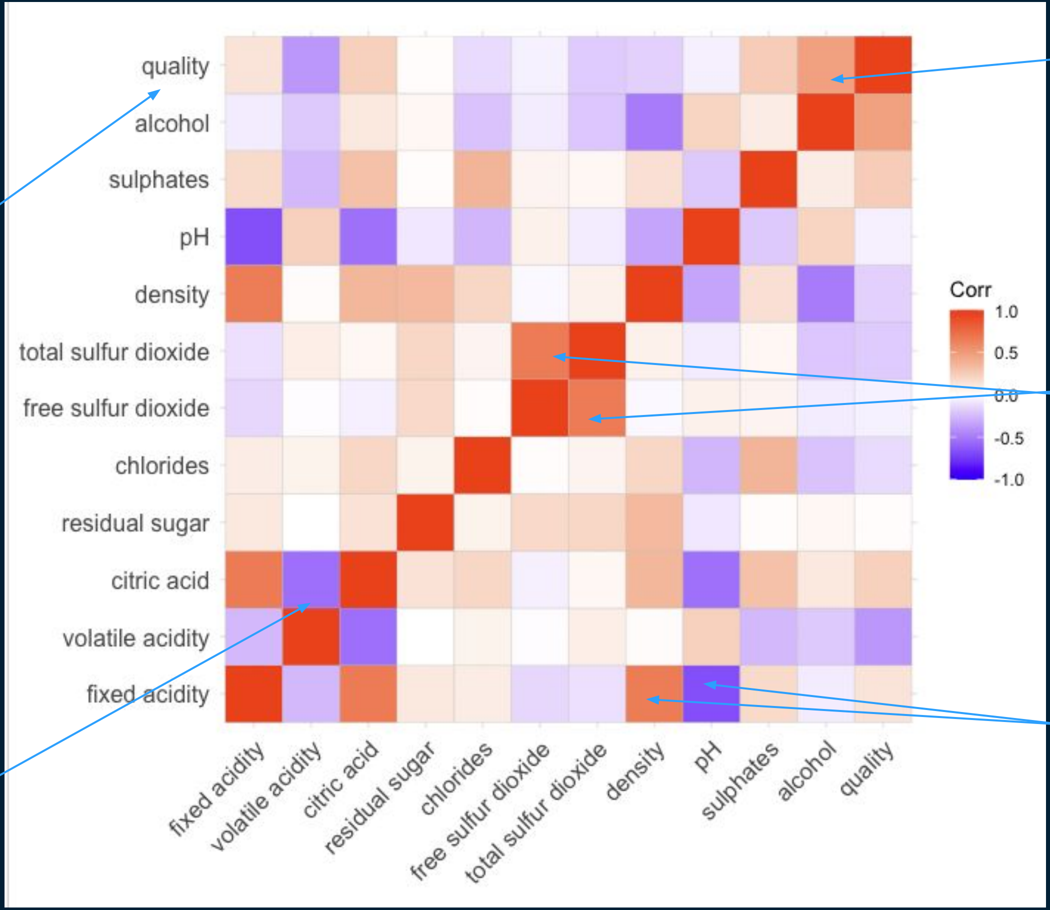
- Wine Composition: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, density, pH, sulphates, alcohol
- Sulfur Dioxide Levels: free sulfur dioxide, total sulfur dioxide

Factors (3)

- Flavor Content: Fixed acidity, alcohol, sugar
- Quality Metrics: Quality

Analysis and Initial Observations

All Numeric Attributes



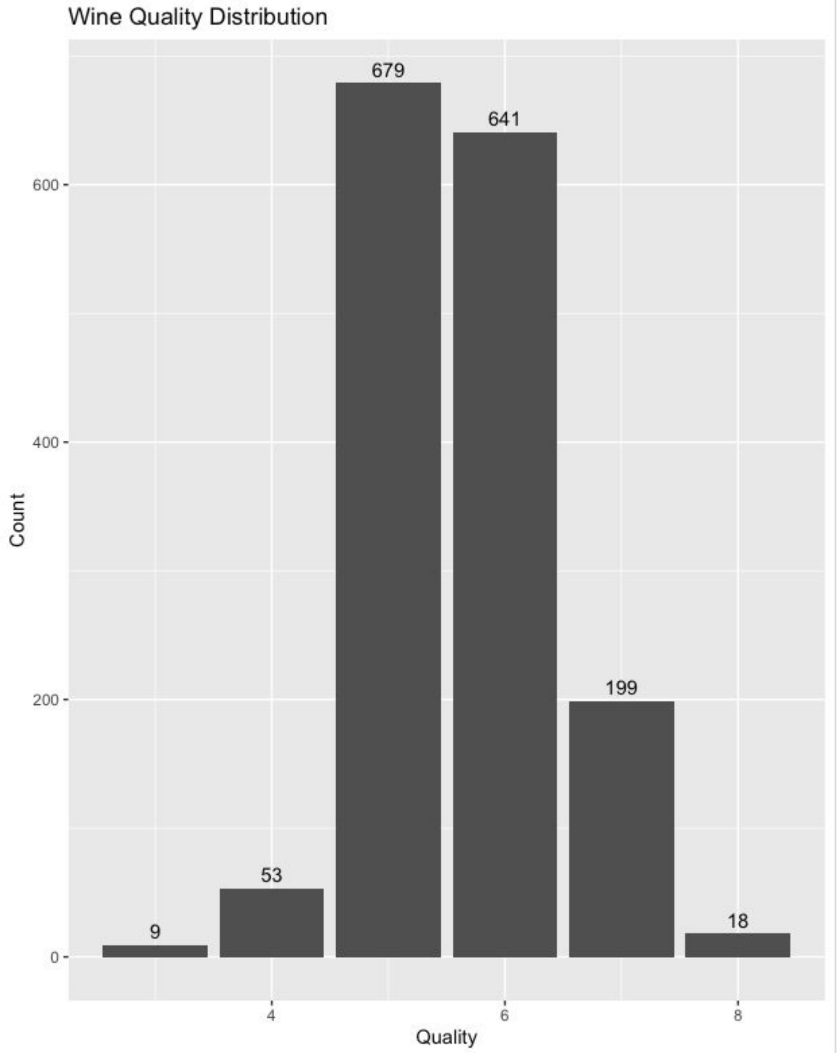
Quality doesn't seem to have a strong correlation with many of the variables; it seems to have some correlation with alcohol

There seems to be some correlation between alcohol and quality

There seems to be a strong correlation between total sulfur dioxide and free sulfur dioxide

There are multiple correlations in this area of the map: citric and fixed acidity, citric acid and volatile acid

There seems to be a strong correlation between fixed acidity and density and fixed acidity and pH

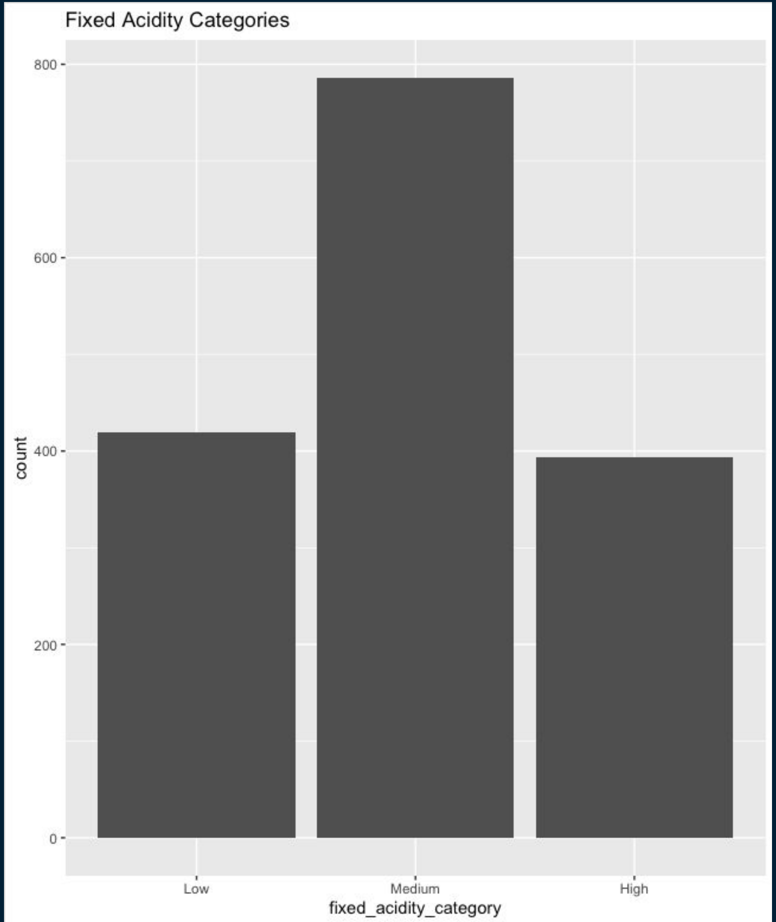


Category	Frequency	Percentage
3	9	0.56 %
4	53	3.31 %
5	679	**42.46%**
6	641	**40.09%**
7	199	12.45%
8	18	1.13%
Total	1599	100%

Both the plot and the table shows that the vast majority of the data points in the dataset are concentrated within the set of average wines (5-6) with a decrease of data points for the higher quality (7-8) and lower quality (3-4) wines

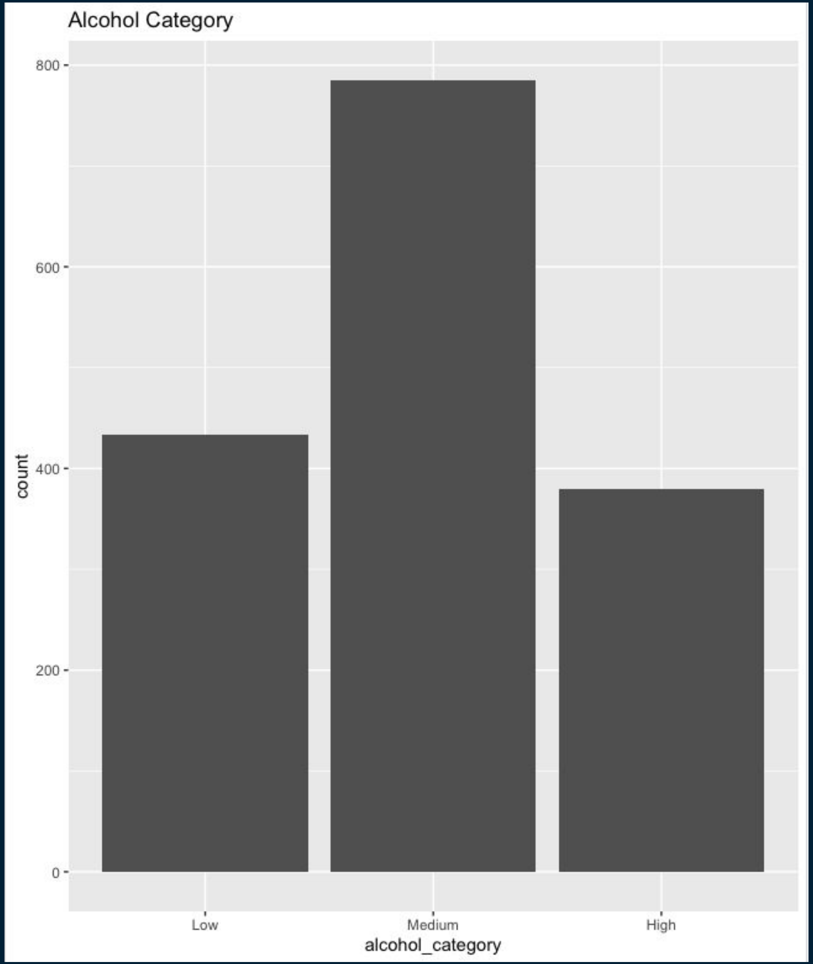
Full univariate analysis: Factors

Univariate Summary of Factors – Fixed Acidity Category



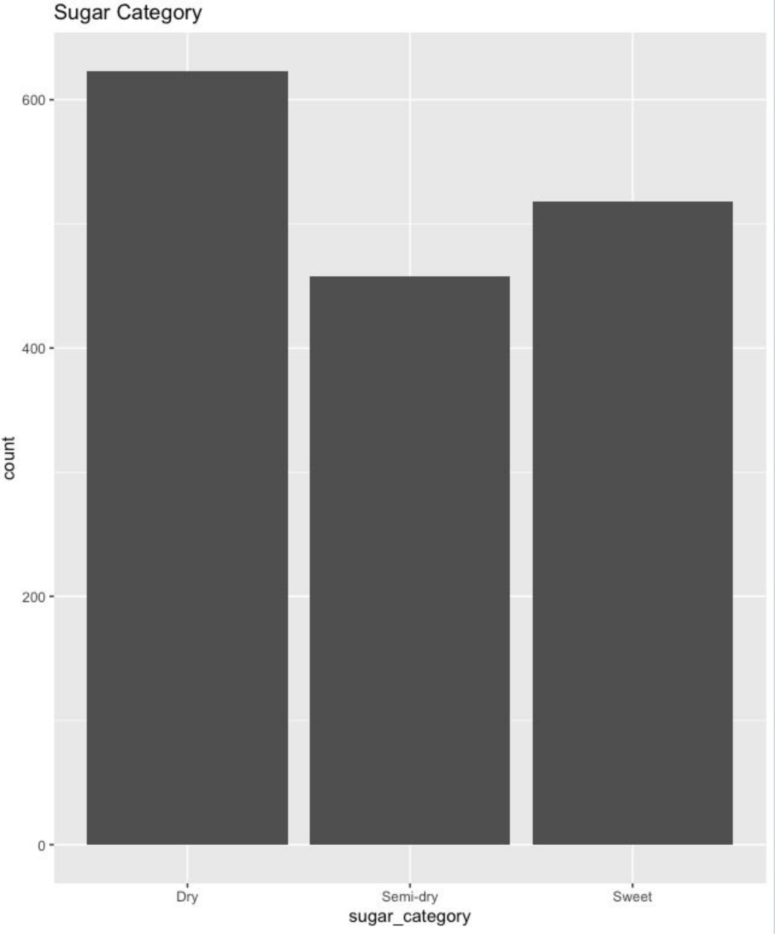
It is observed that the majority of wine variety falls under medium acidity. Low and high fixed acidity roughly have the same count

Univariate Summary of Factors – Alcohol Category



It is observed that majority of the wine has medium alcohol content. Low and high fixed alcohol categories roughly have the same count

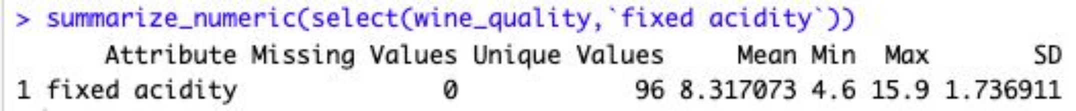
Univariate Summary of Factors – Sugar Category



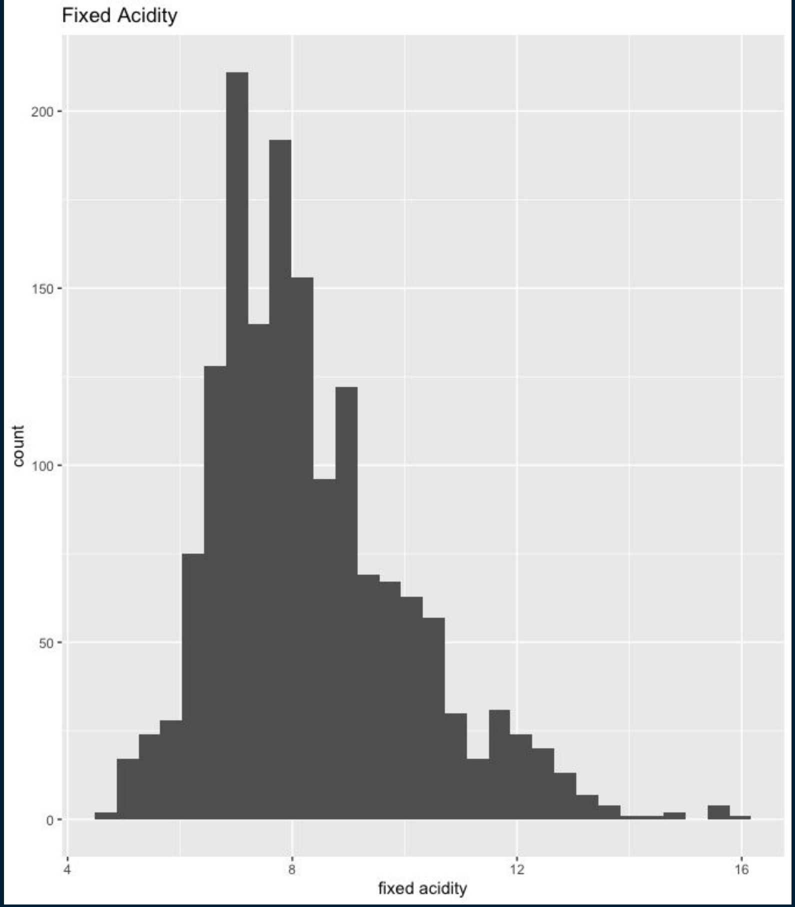
It is observed that majority of the wine falls under the dry sugar category followed by sweet then semi-dry.

Full univariate analysis: Numeric

Univariate Summary of Numeric – Fixed Acidity

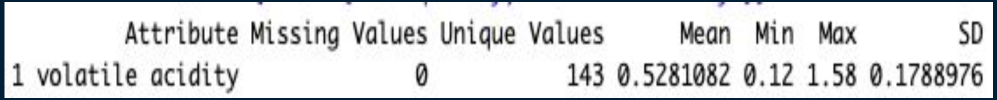


```
> summarize_numeric(select(wine_quality,`fixed acidity`))
      Attribute Missing Values Unique Values     Mean Min  Max       SD
1 fixed acidity              0            96 8.317073 4.6 15.9 1.736911
```



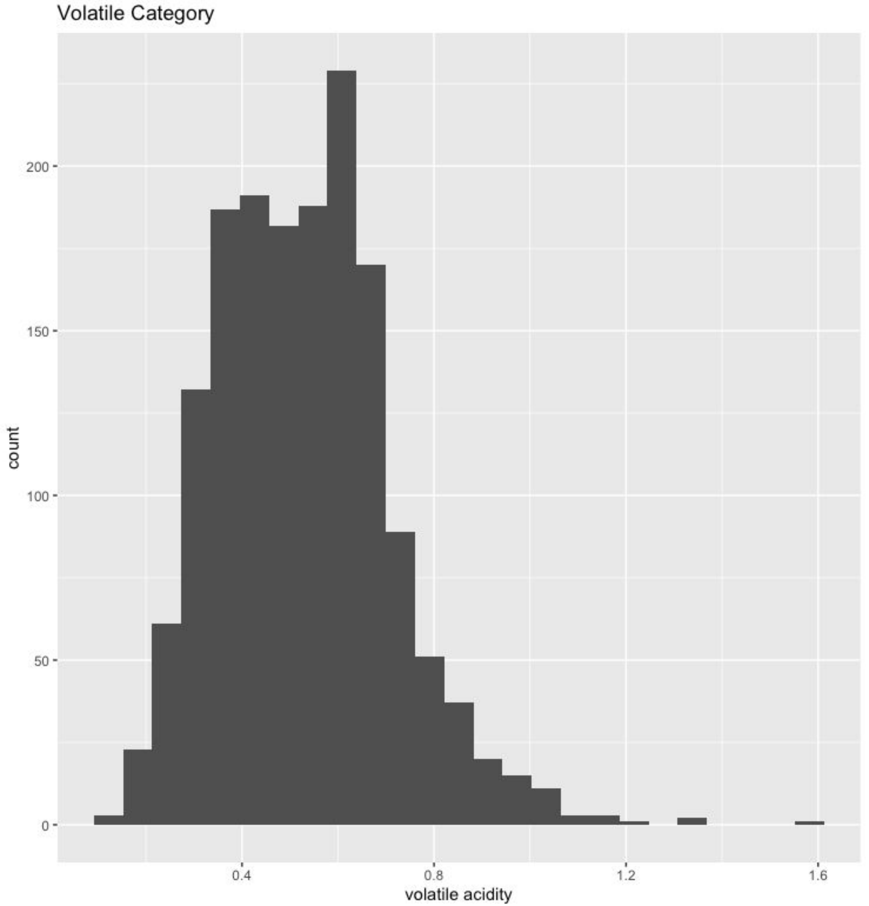
Fixed acidity appears to have a positive skew (mean > median) and the average is roughly 8.31.

Univariate Summary of Numeric – Volatile Acidity



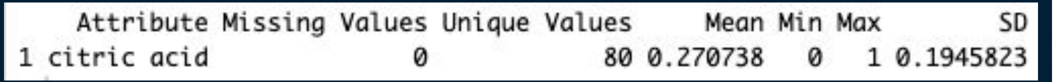
```
          Attribute Missing Values Unique Values      Mean  Min  Max        SD
1 volatile acidity              0           143 0.5281082 0.12 1.58 0.1788976
```

Volatile Acidity appears to have a positive skew and the average is about 0.53. Also some outliers are observed here that have high volatile acidity (it could be a true or false outlier but we don’t know yet).



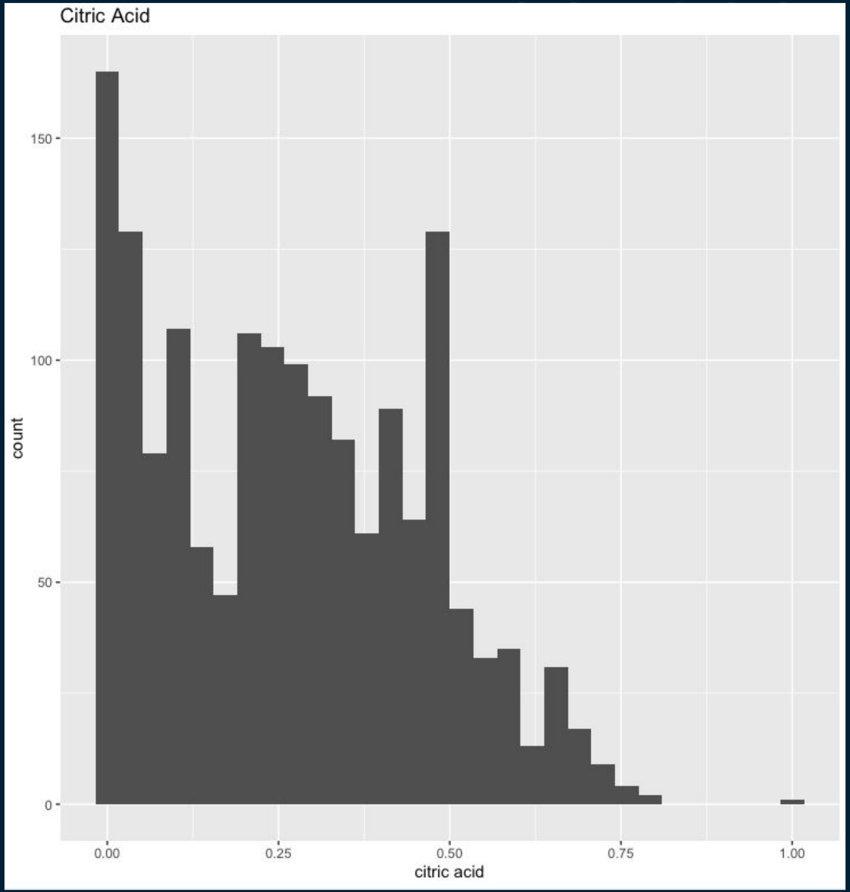
Volatile Category

Univariate Summary of Numeric – Citric acid

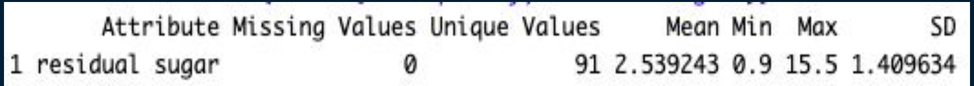


```
      Attribute Missing Values Unique Values       Mean Min Max        SD
1 citric acid              0            80 0.270738   0   1 0.1945823
```

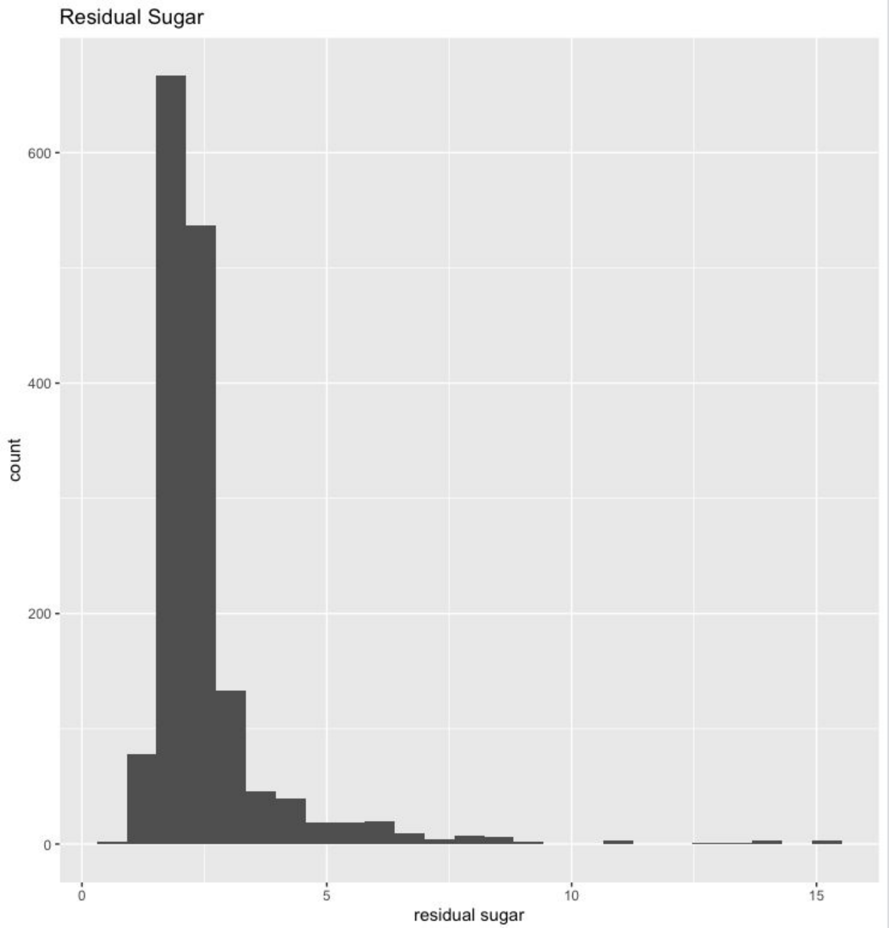
Citric acid appears to have a positively skewed; citric acid is found in small quantities which adds to the freshness and flavor of the wines. There seems to be an outlier towards the end of the tail.



Univariate Summary of Numeric – Residual Sugar

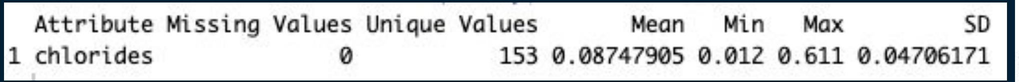


```
       Attribute Missing Values Unique Values     Mean Min  Max       SD
1 residual sugar              0            91 2.539243 0.9 15.5 1.409634
```

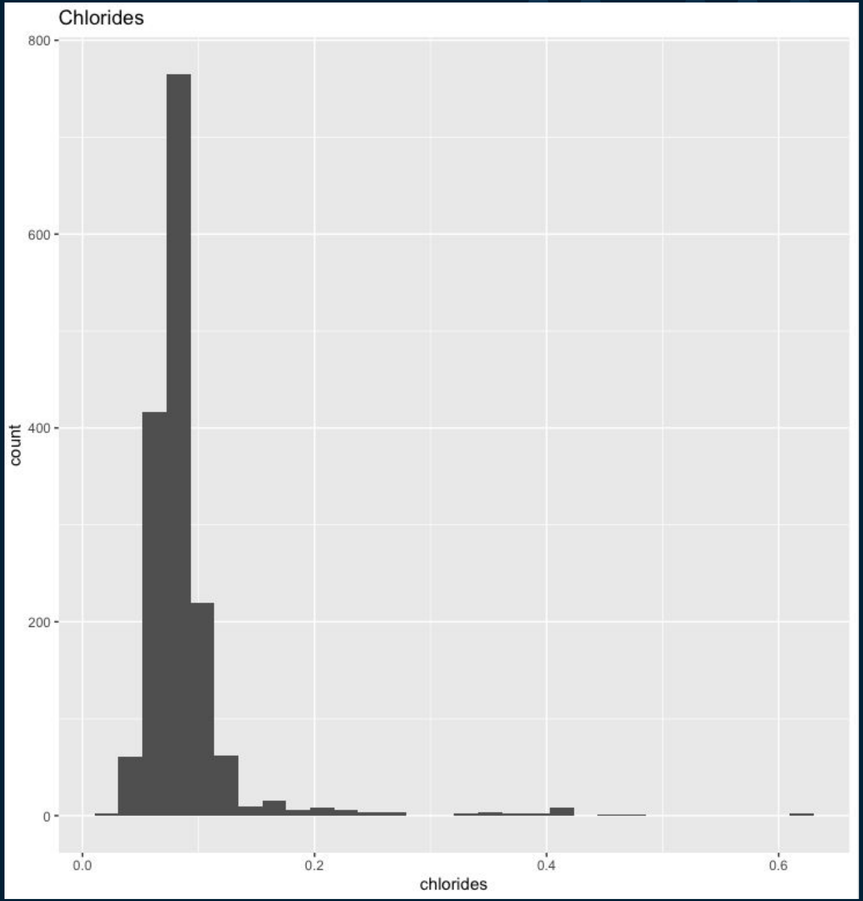


It appears that the Residual sugar has a positive skew with majority of the wines having less amount of sugar post fermentation, while some wines with very large quantities of residual sugar can be observed. There seems to be a few outliers towards the tail.

Univariate Summary of Numeric – Chlorides

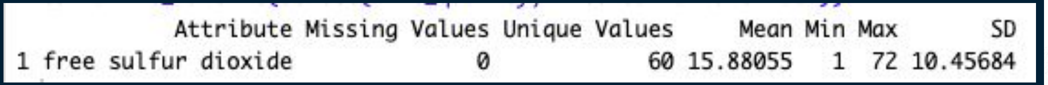


```
  Attribute Missing Values Unique Values       Mean   Min   Max         SD
1 chlorides              0           153 0.08747905 0.012 0.611 0.04706171
```



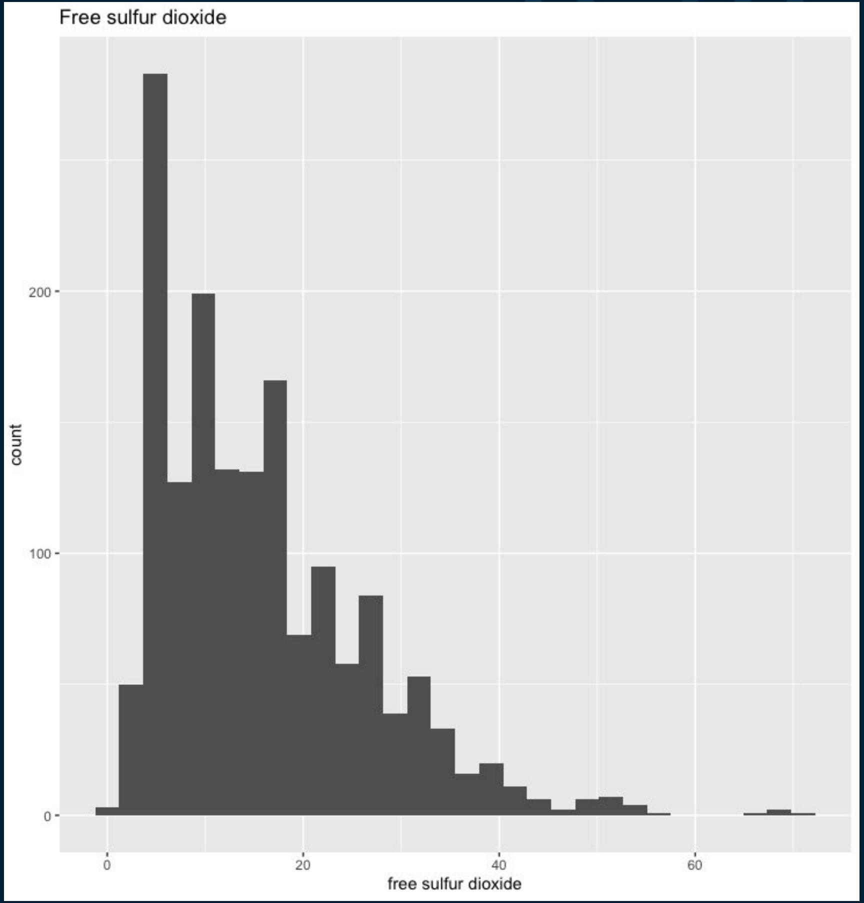
It appears that Chlorides has a positive skew with majority of the wines having smaller levels of salt present. There’s a few outliers towards the tail.

Univariate Summary of Numeric – Free sulfur dioxide

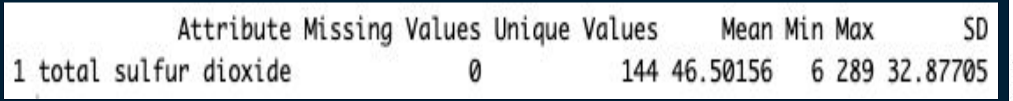


```
           Attribute Missing Values Unique Values     Mean Min Max       SD
1 free sulfur dioxide              0            60 15.88055   1  72 10.45684
```

It appears that free sulfur dioxide has a positive skew with majority of the wines having on average 15.88 which helps with preventing microbial growth and oxidation of wine. There’s a few outliers towards the tail.

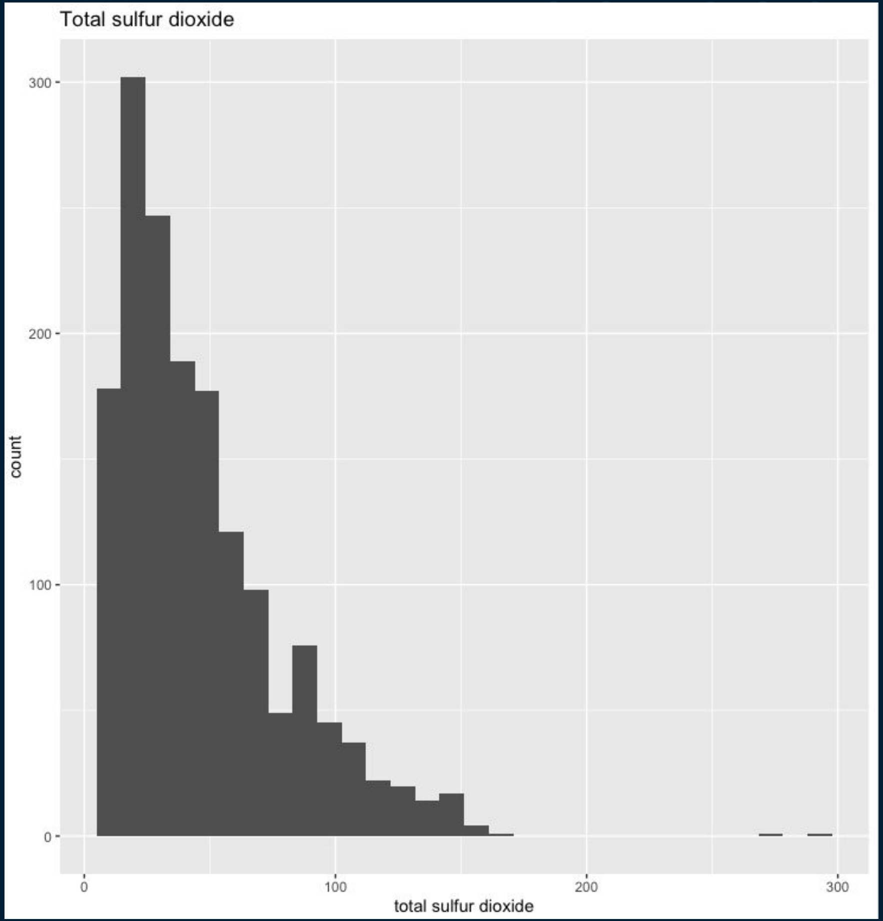


Univariate Summary of Numeric – Total sulfur dioxide

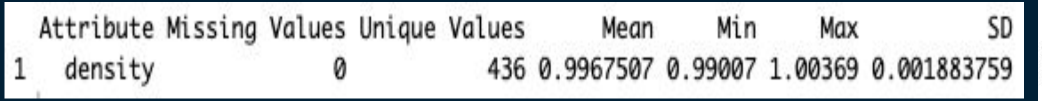


```
            Attribute Missing Values Unique Values     Mean Min Max       SD
1 total sulfur dioxide              0           144 46.50156   6 289 32.87705
```

It appears that total sulfur dioxide has a positive skew with majority of the wines having on average 46.5. There’s a few outliers towards the tail.

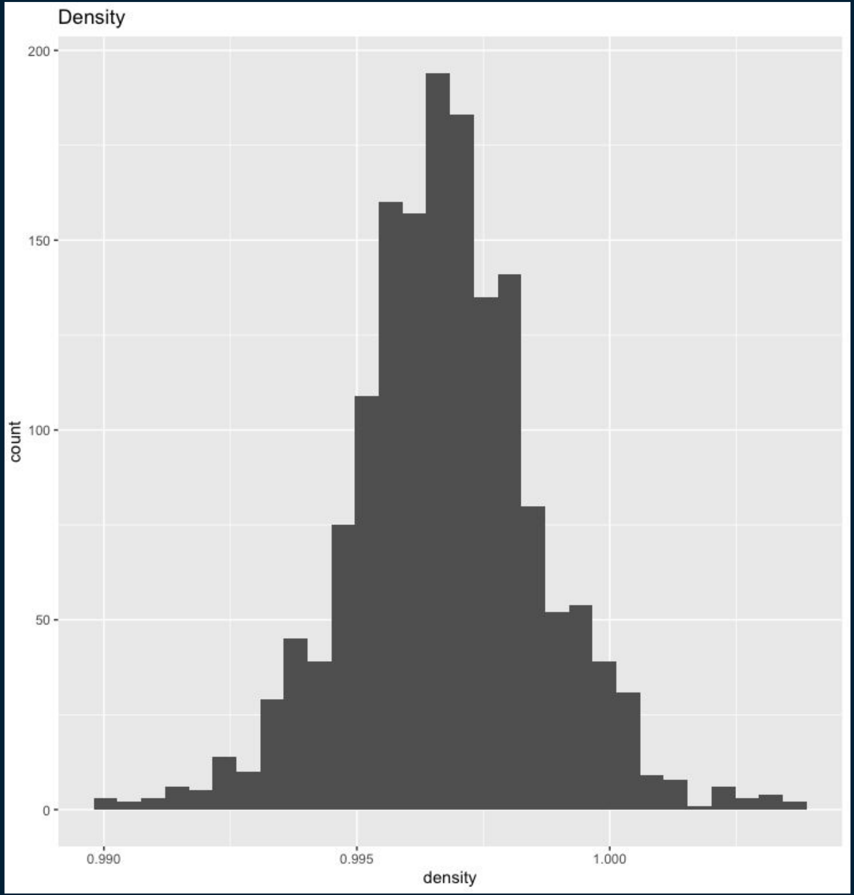


Univariate Summary of Numeric – Density

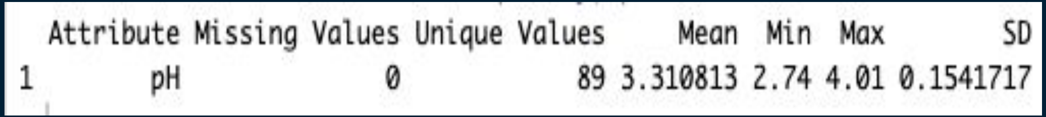


```
  Attribute Missing Values Unique Values      Mean     Min     Max          SD
1   density              0           436 0.9967507 0.99007 1.00369 0.001883759
```

It appears that density has a semi normal distribution with an average of 0.997 which provides insights into the alcohol percentage and sugar content.

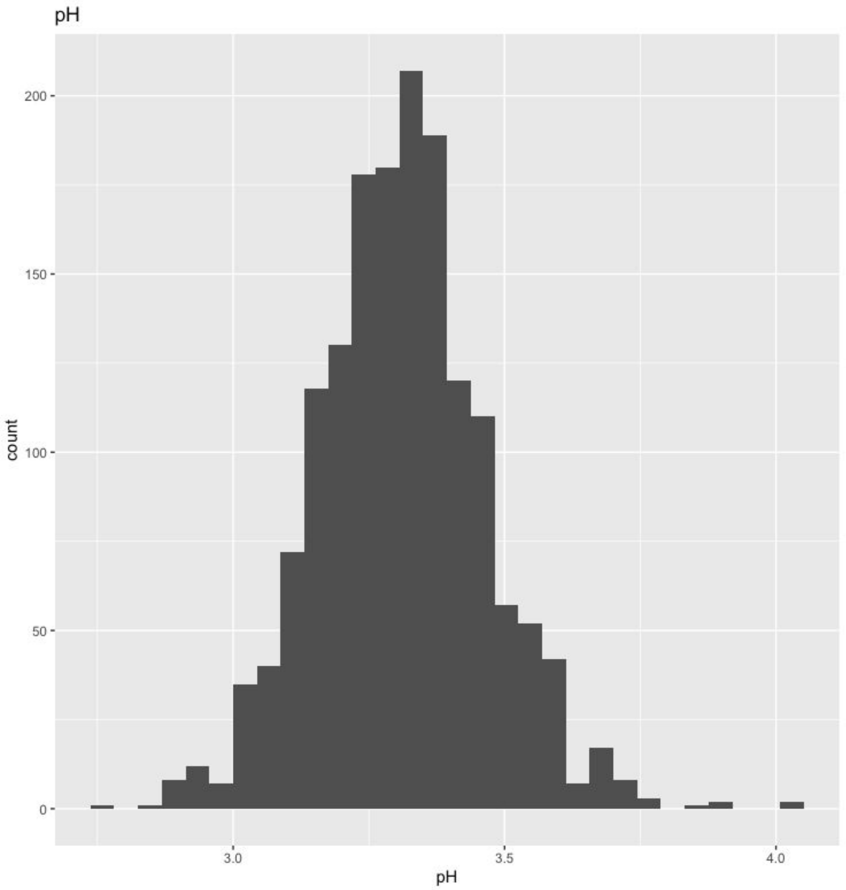


Univariate Summary of Numeric – pH

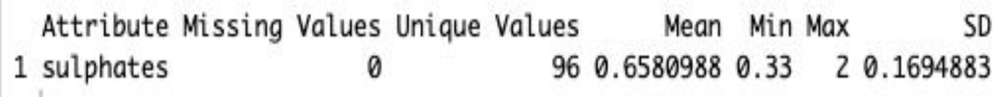


```
  Attribute Missing Values Unique Values     Mean  Min  Max        SD
1        pH              0            89 3.310813 2.74 4.01 0.1541717
```

It appears that pH has a semi normal distribution with an average of 3.13 which provides insights on the acidity of the wine so most of the wine falls under the the average pH of 3-4.

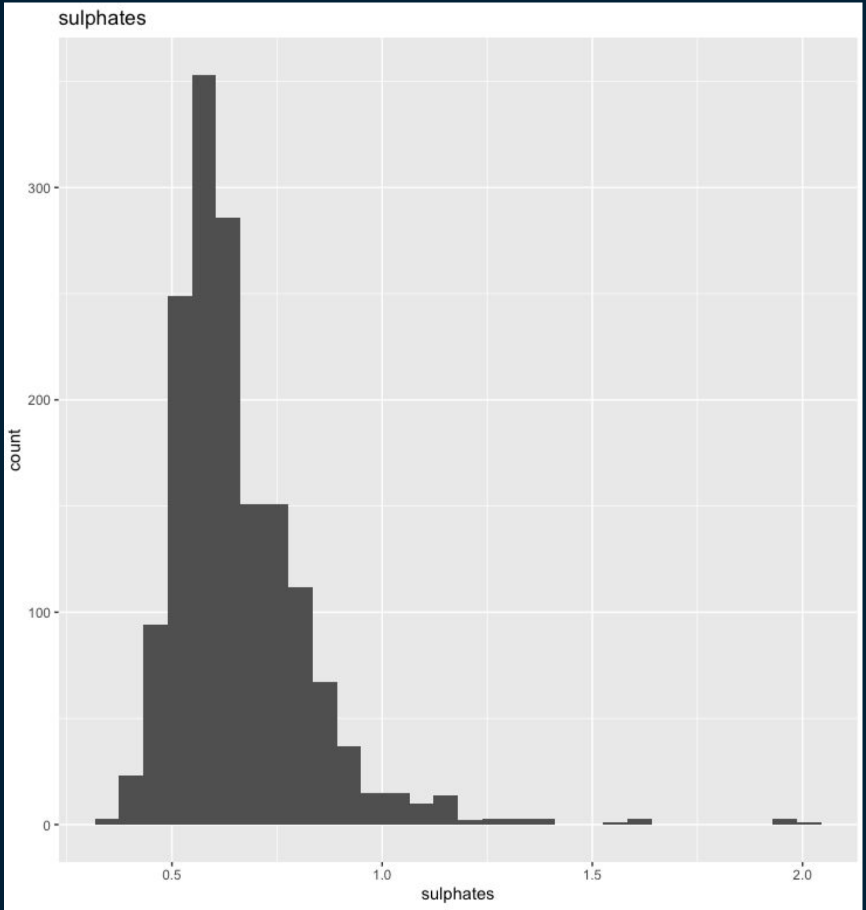


Univariate Summary of Numeric – Sulphates

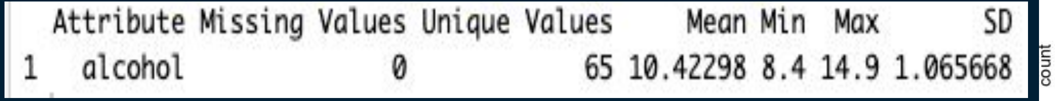


```
  Attribute Missing Values Unique Values      Mean  Min Max        SD
1 sulphates              0            96 0.6580988 0.33   2 0.1694883
```

It appears that sulphates has a positive skew with majority of the wines with an average of 0.66 which acts as a antimicrobial and antioxidant. There’s a few outliers towards the tail.

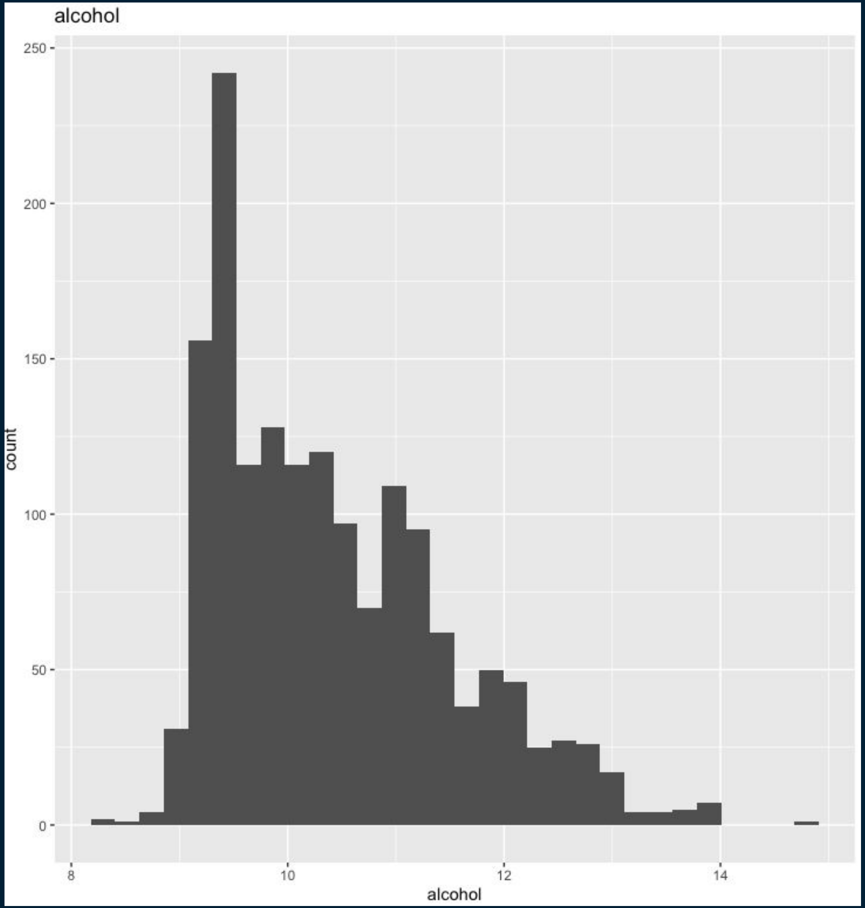


Univariate Summary of Numeric – Alcohol

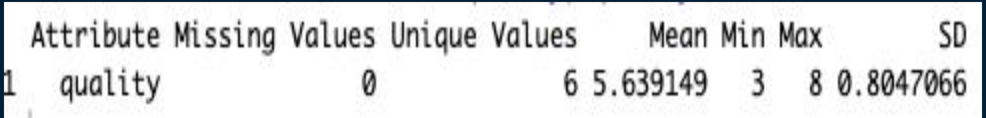


```
  Attribute Missing Values Unique Values     Mean Min  Max       SD
1   alcohol              0            65 10.42298 8.4 14.9 1.065668
```

It appears that alcohol has a positive skew with majority of the wines with an average of 10.42 which represents the alcohol content in the wine.

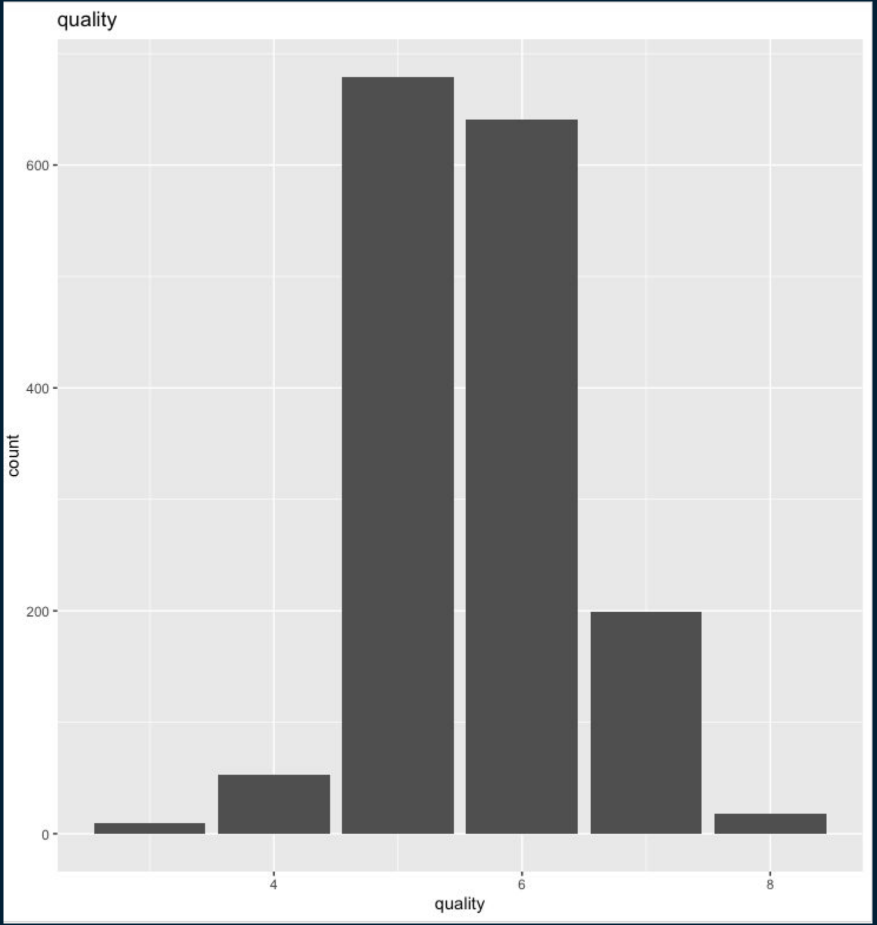


Univariate Summary of Numeric – Quality



```
  Attribute Missing Values Unique Values     Mean Min Max        SD
1   quality              0             6 5.639149   3   8 0.8047066
```

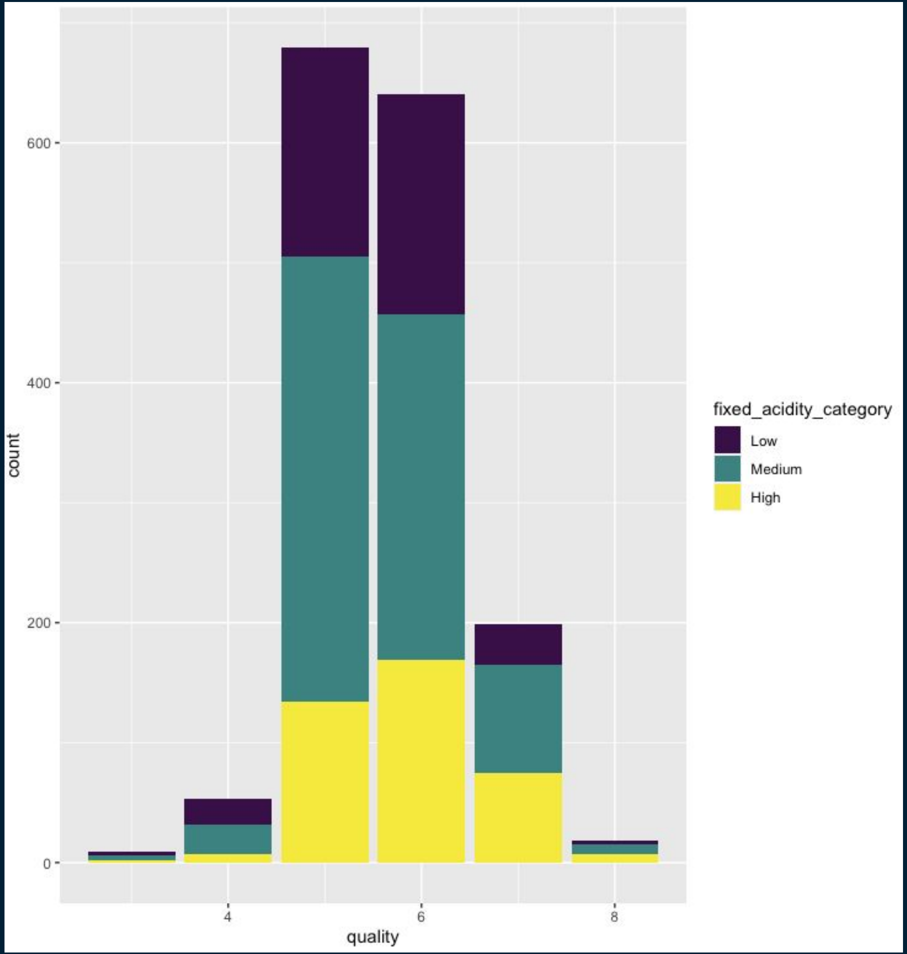
It appears that quality has the majority of wines in the 5 to 6 category at an average of 5.63.



Full bivariate analysis: Factors

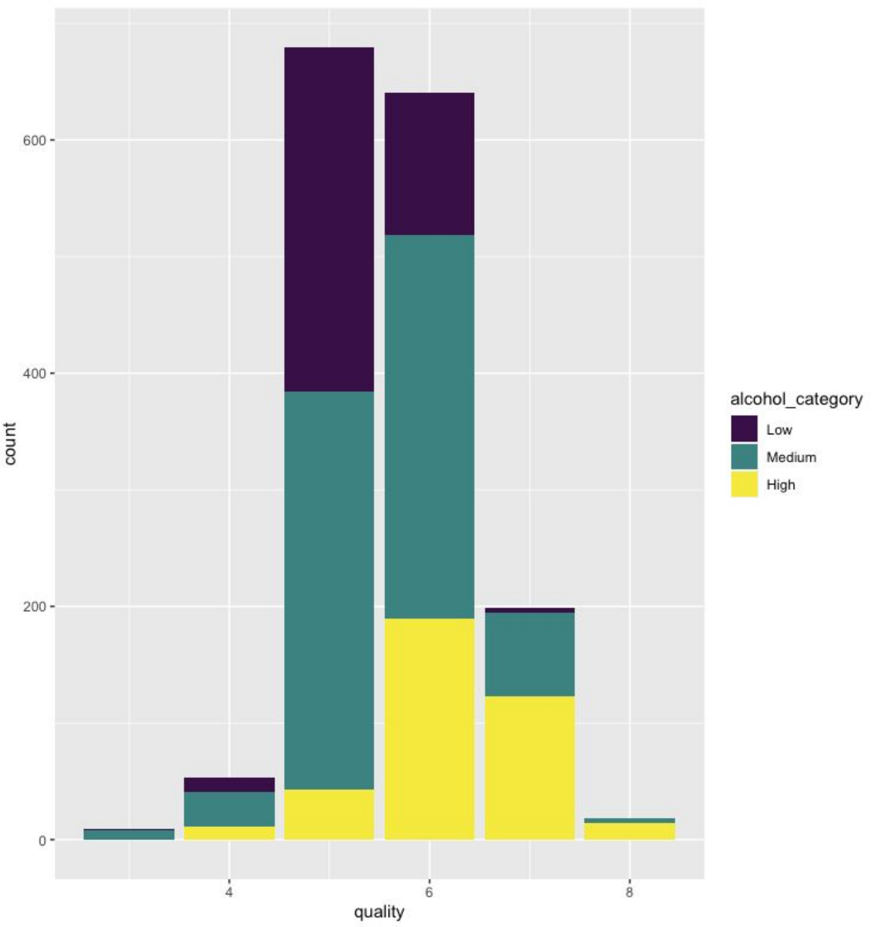
Each variable vs the response variable

Bivariate Summary of Factors – Fixed Acidity Category



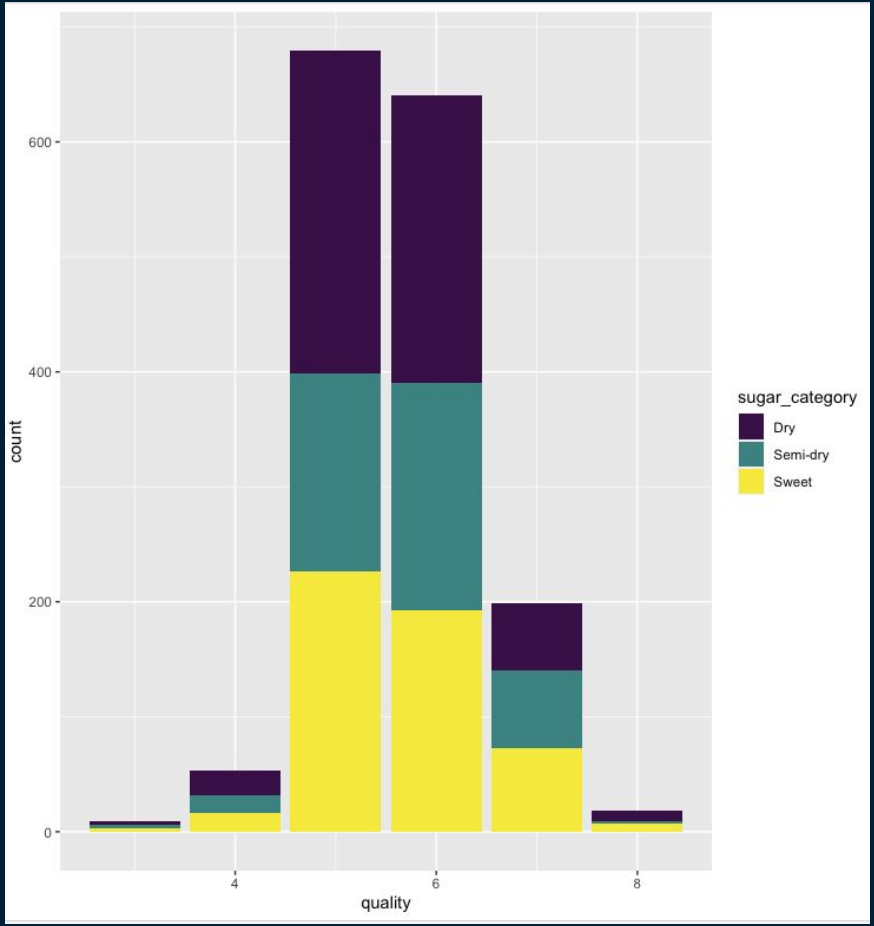
The wines tend to fall under the medium fixed acidity category with the majority being at 5-6 quality.

Bivariate Summary of Factors – Alcohol Category



The wines tend to fall under the medium alcohol category with the majority being at 5-6 quality.

Bivariate Summary of Factors – Sugar Category

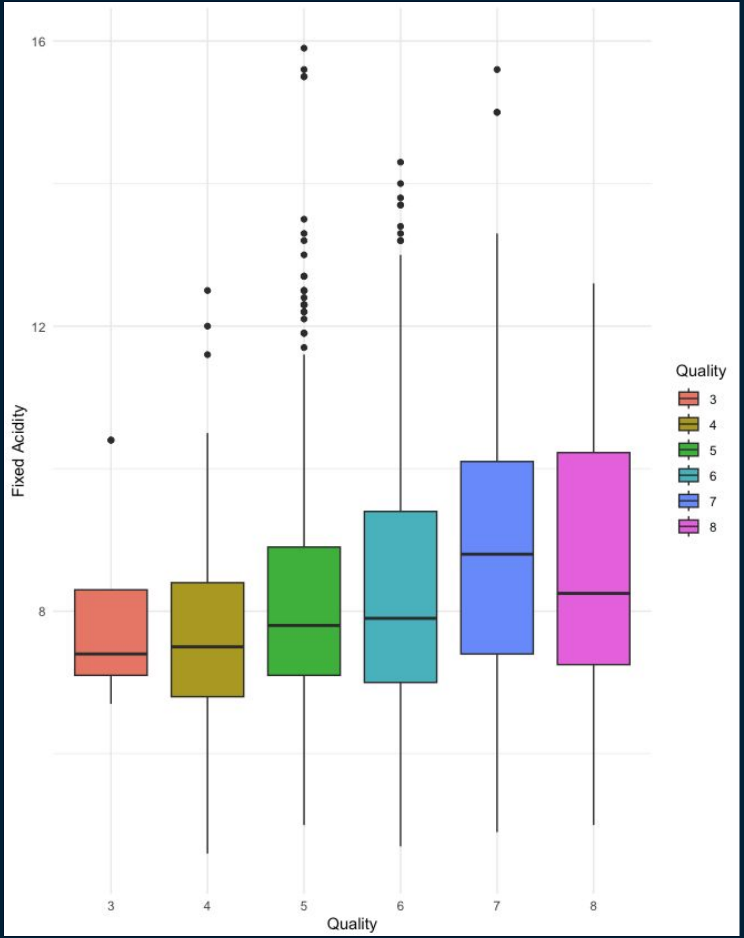


The wines tend to fall under the dry sugar category with the majority being at 5-6 quality.

Full bivariate analysis: Numerics

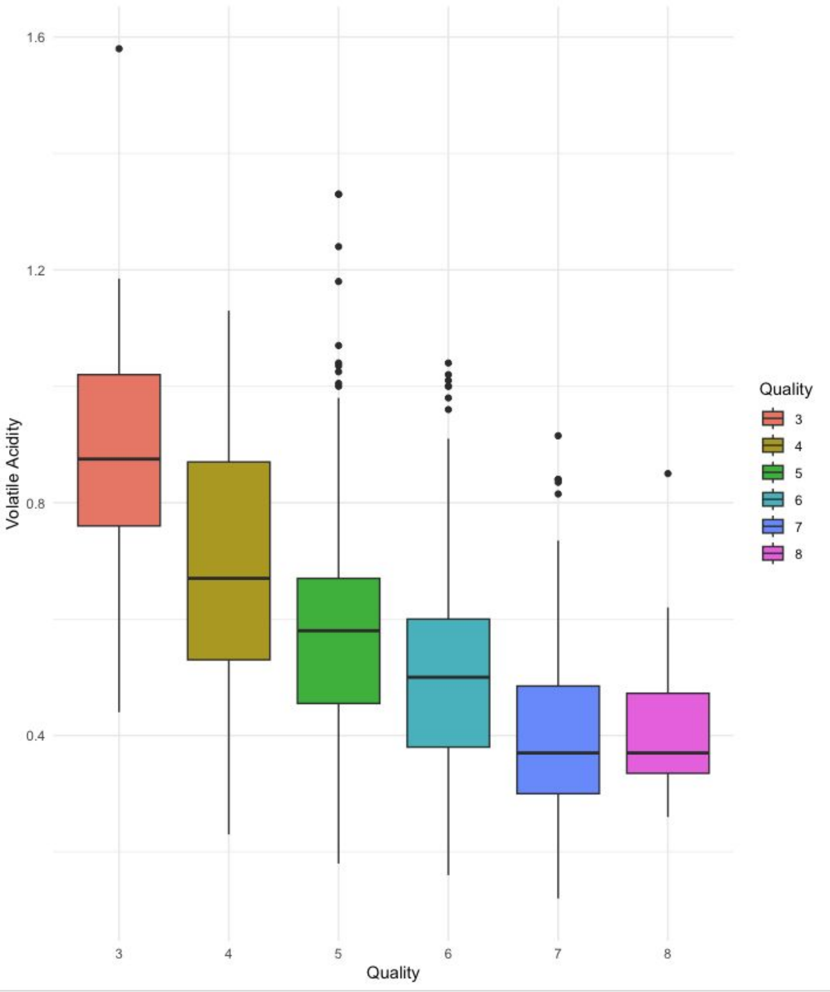
Each variable vs the response variable

Fixed Acidity



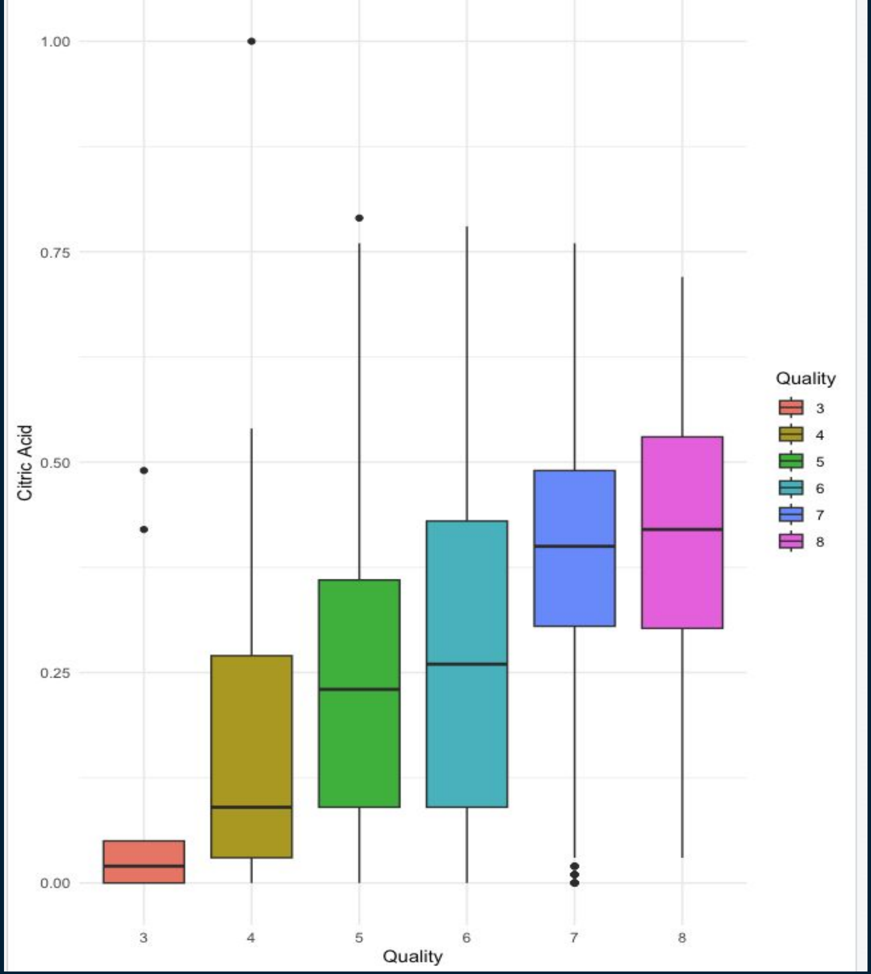
It seems that the fixed acidity for each factor is around the same for each category for quality. Quality 8 seems to have the largest range for fixed acidity. There also seems to be some outliers in some of the qualities.

Volatile Acidity - Potential Correlation



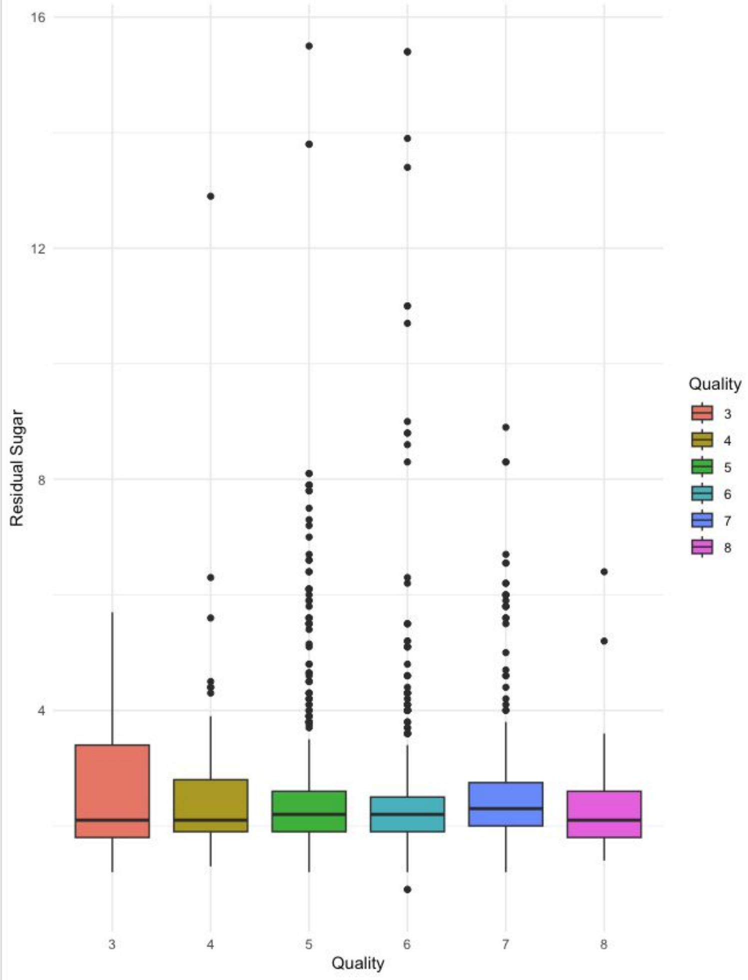
As quality increases, volatile acidity decreases so there could be a negative linear correlation. This could indicate that people prefer wines with lower volatile acid.

Citric Acid - Potential Correlation



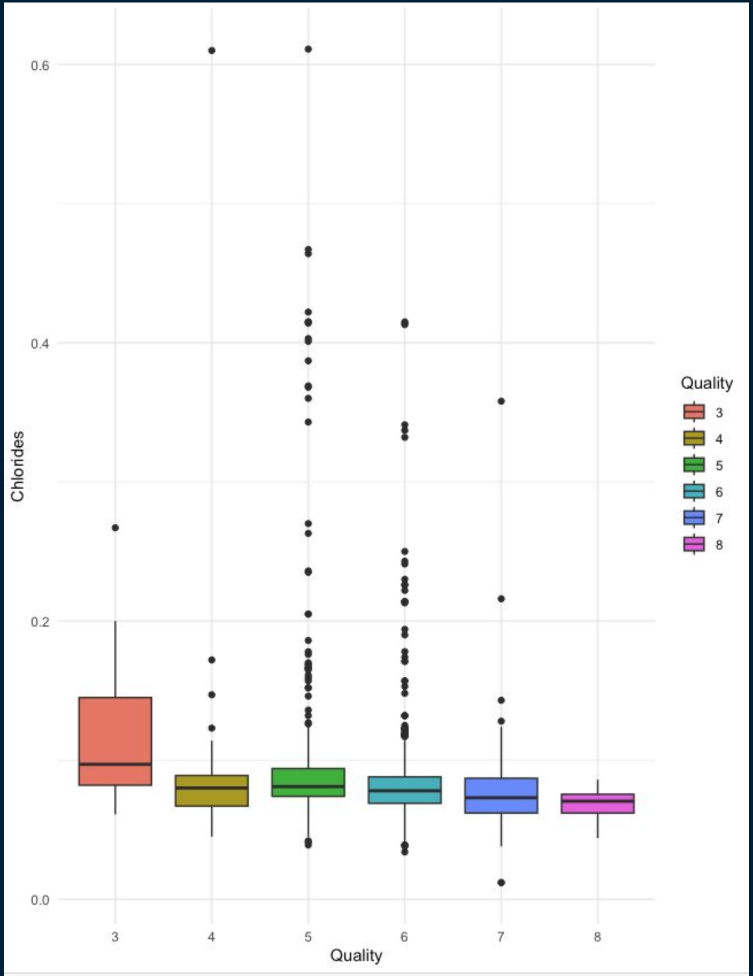
As quality increases, citric acid increases so there could be a positive linear correlation. Citric acid adds to the freshness and flavor of the wines which could explain why higher quality wines have more citric acid.

Residual Sugar



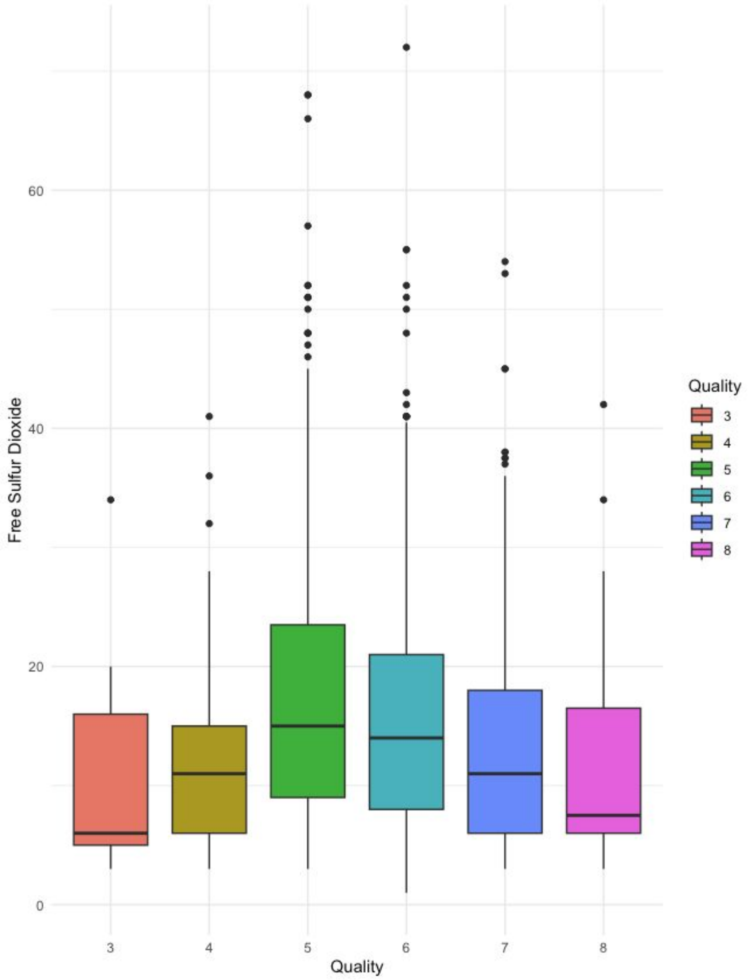
It seems that the residual sugar for each factor is around the same for each category for quality. Quality 1 seems to have the largest range for residual sugar. There also seems to be some outliers in some of the qualities.

Chlorides



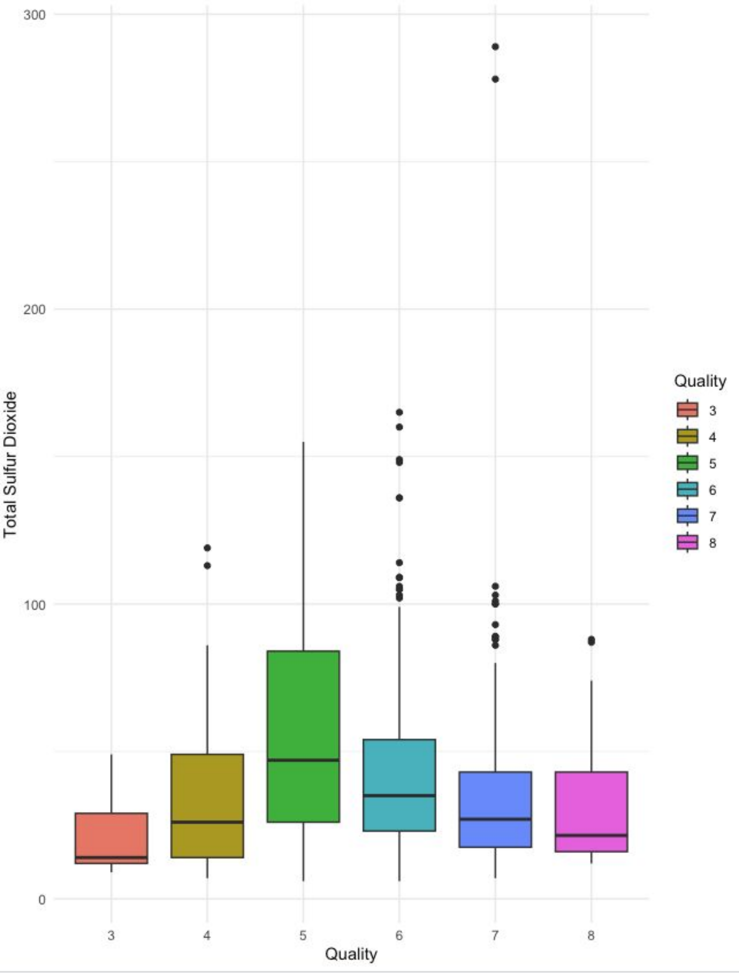
It seems that the chlorides for each factor is around the same for each category for quality. Quality 1 seems to have the largest range for residual sugar. There also seems to be some outliers in some of the qualities. This could indicate that wines contains roughly the same amount of salt.

Free Sulfur Dioxide



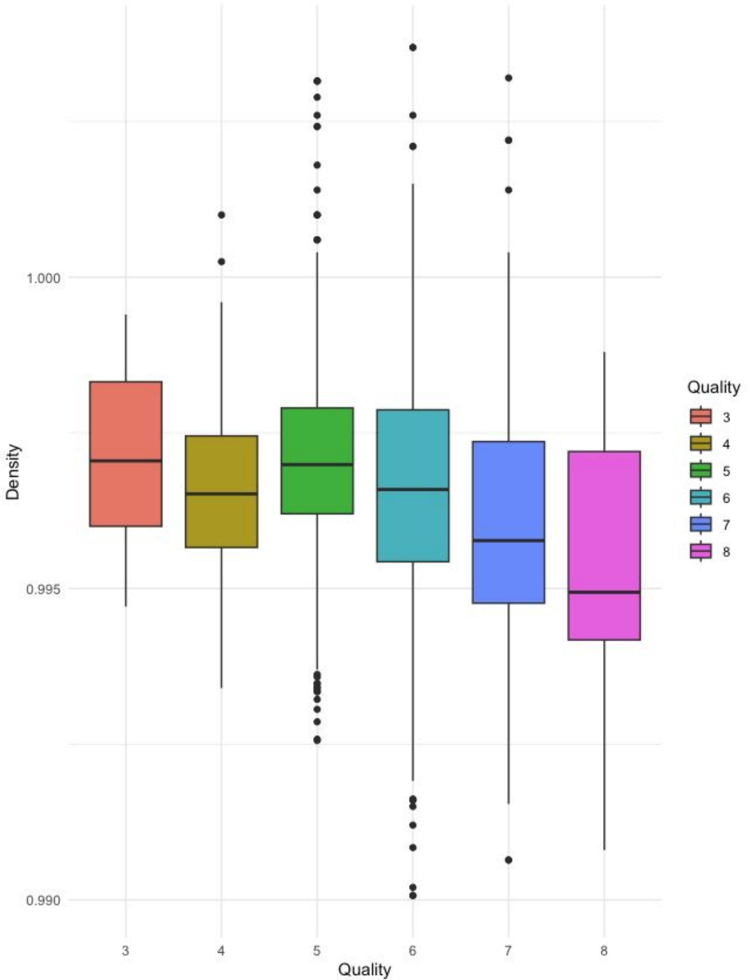
It seems that the free sulfur dioxide increases up until quantity 5 then it starts to slowly decrease. Free sulfur dioxide prevents microbial growth and oxidation of wine which could explain why wines in quality 5-6 have more of it.

Total Sulfur Dioxide



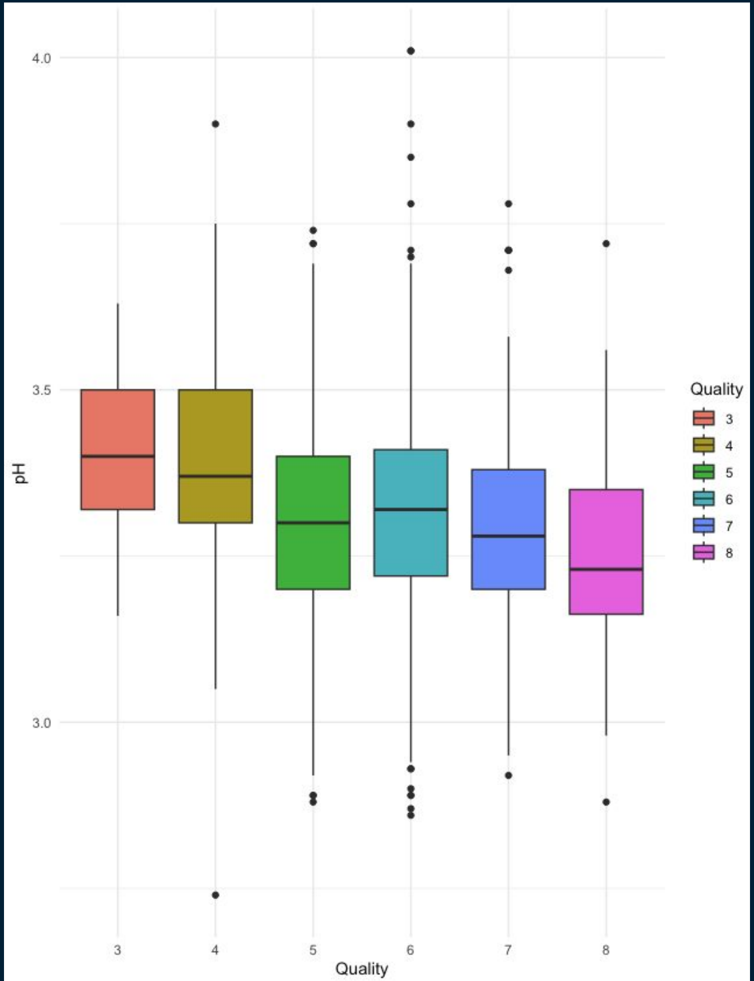
It seems that the total sulfur dioxide increases up until quantity 5 then it starts to slowly decrease.

Density



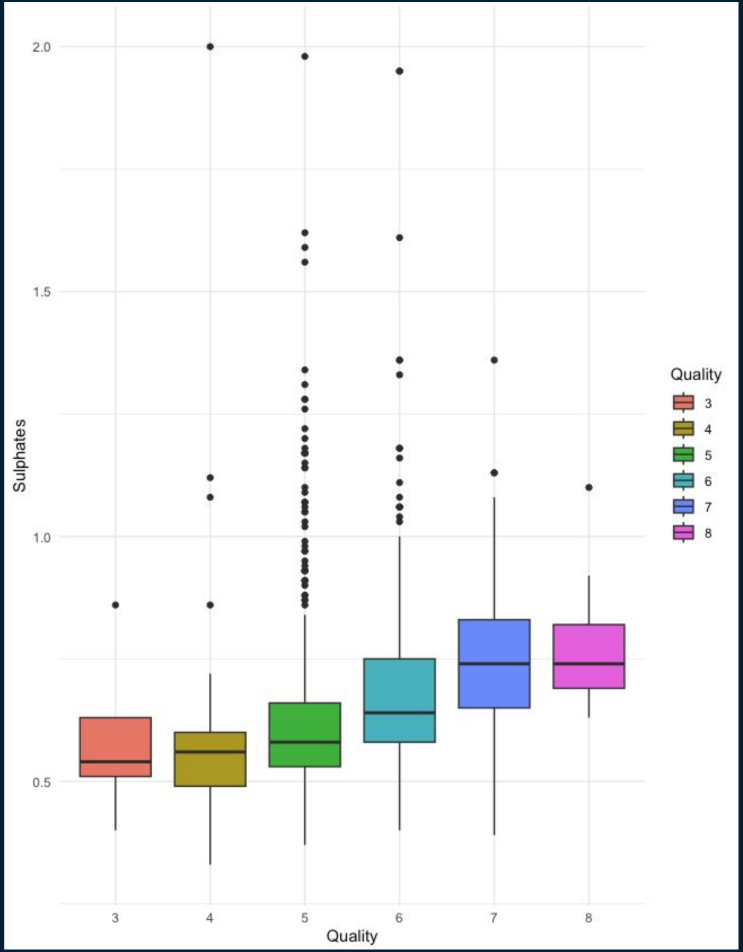
It seems that the density is the same for quality 0 to 4 then it slightly decreases for quality 5 and 6. Density provides insight into the alcohol percentage and sugar content so it could be inferred that people potentially prefer wines with lower density.

pH



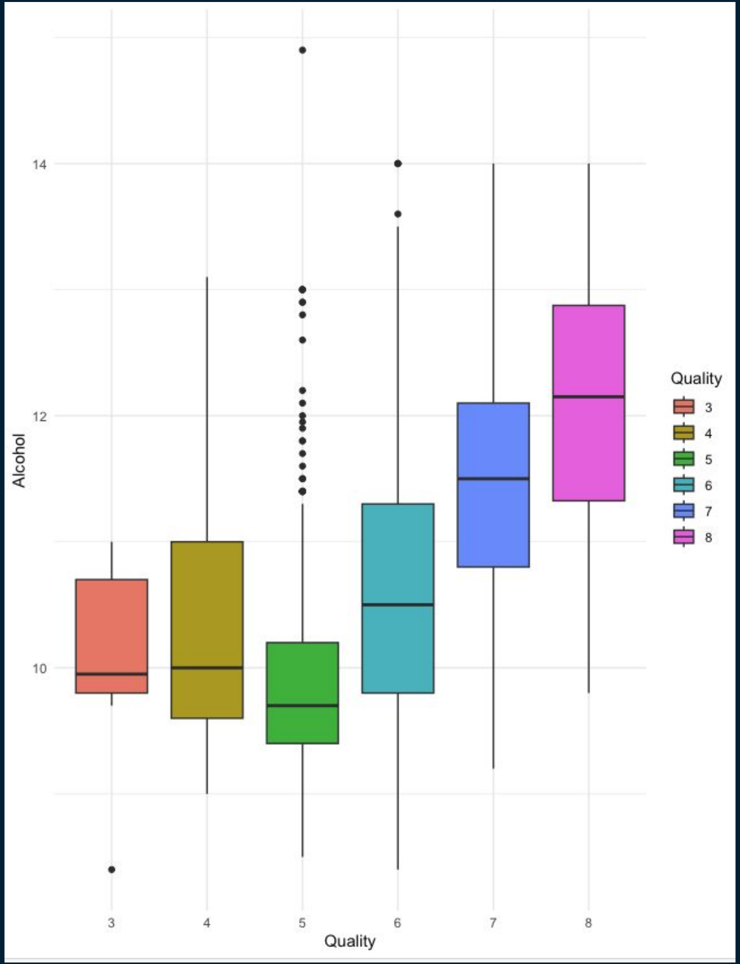
Wine with lower PH value tend to have better quality. pH indicates how acid the wine is so people might prefer wines with lower pH

Sulphates- Potential Correlation



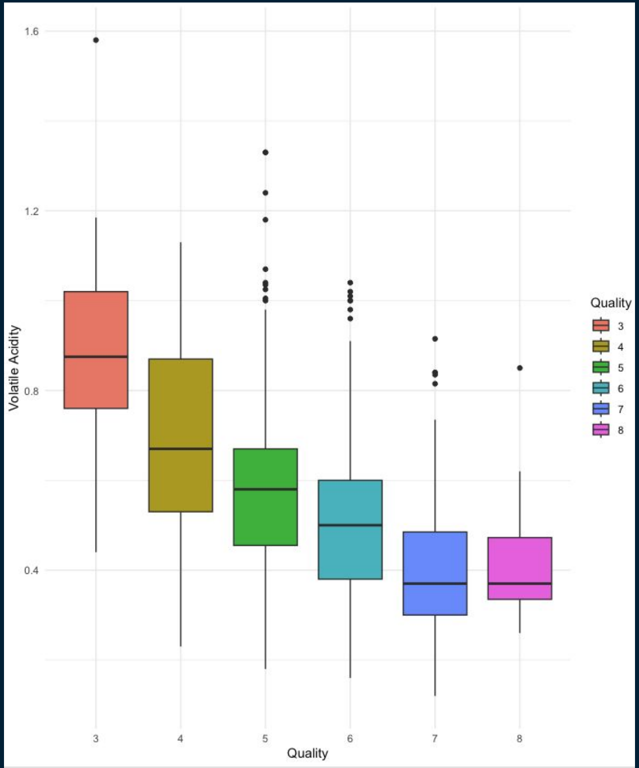
As quality increases, sulphates slightly increase so there could be a positive linear correlation. This could be because people might prefer wines that have higher levels of sulphates since it acts as a antimicrobial and antioxidant.

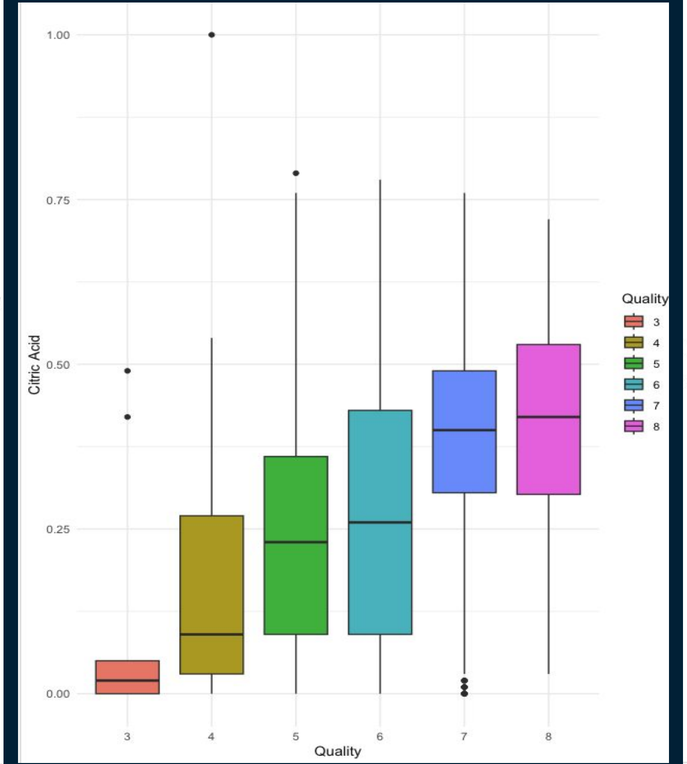
Alcohol - Potential Correlation

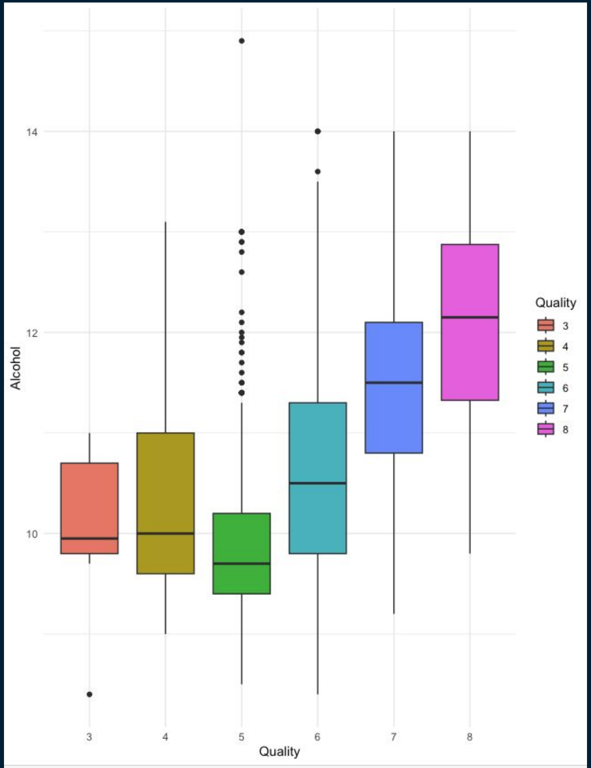


As quality increases, alcohol seems constant for quality 3 to 5 then it increases from 6 to 8 which could indicate that people prefer wines with higher levels of alcohol.

Attributes with highest correlation (based on graphs)





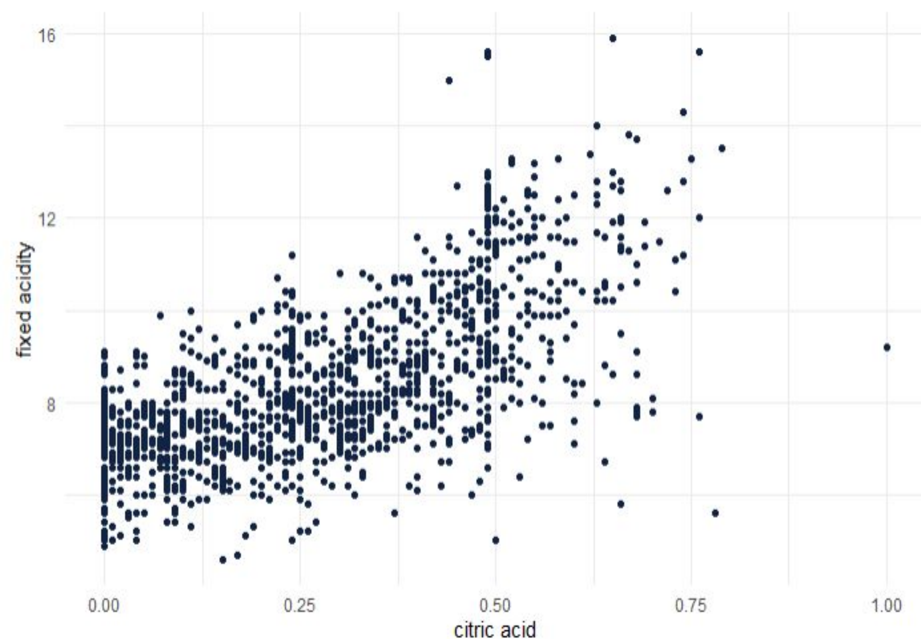


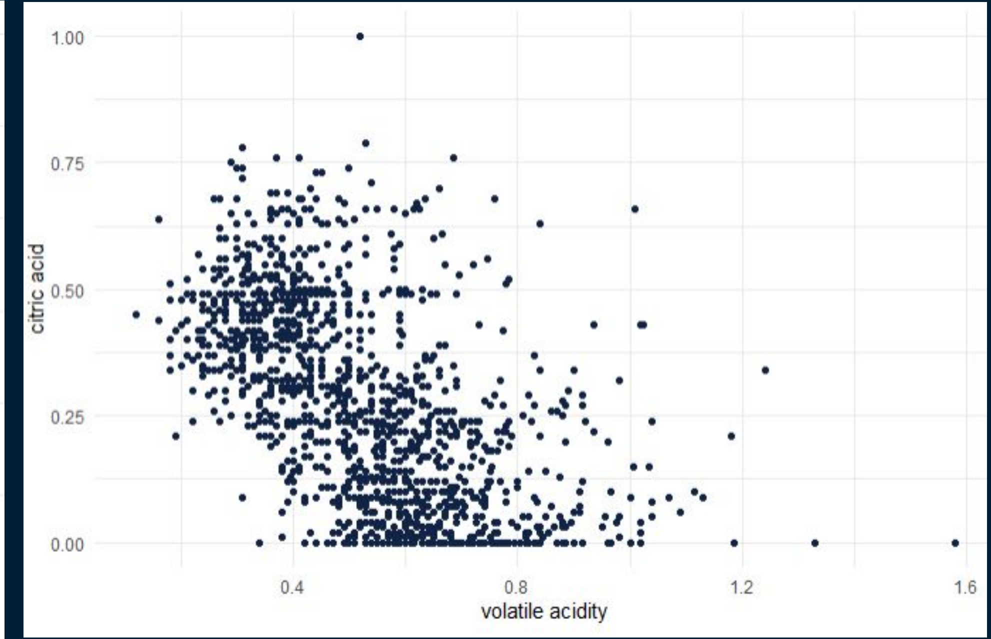
Volatile Acidity:
negative linear correlation

Citric Acid:
positive linear correlation

Alcohol:
constant then positive linear correlation

Initial Observations for Numeric vs Numeric Variables





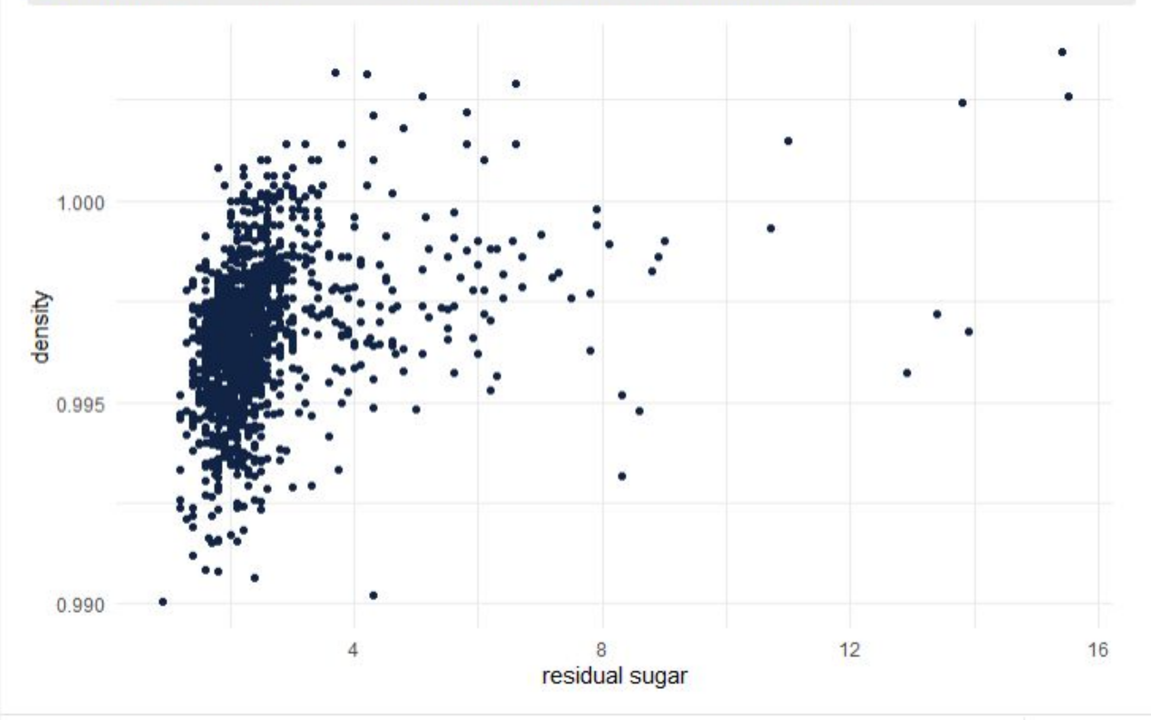
Fix acidity VS Citric Acid (Positive Correlation)

Could potentially indicate that the more citric acid the more fresh and flavorful the wine is which increased the amount of non-volatile acids present in the wine.

Citric Acid VS Volatile Acidity (Negative Relation)

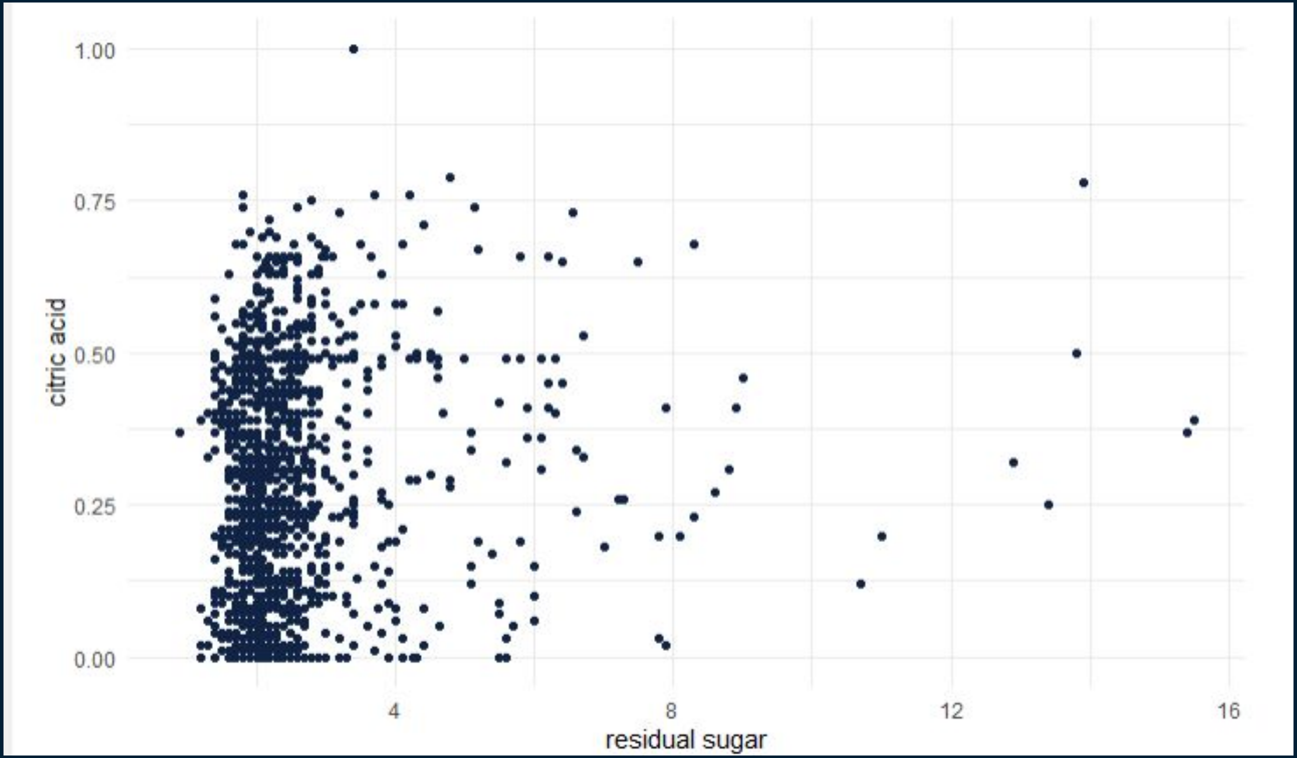
Could potentially indicate that the less citric acid there is in the wine which could indicate that the wine isn’t as fresh increases the steam-distillable acids present on wine.

Density VS Residual Sugar (Positive Correlation)



Could potentially indicate that as density increases the amount of sugar slightly increases so potentially the alcohol percentage and sugar content increase as the amount of sugar left after fermentation increases.

Citric Acid VS Volatile Acidity (No Relation)



Analysis and Initial Observations

Analysis and Initial Observations

Further Investigation Indicated

- Why there is more data for medium fixed acidity category ?
- Why there is a tendency for medium alcohol category?

Analysis and Initial Observations

Potential Feature Generation

- We convert wine with quality from 3 to 6 as “Medium Quality Wine” and convert wine with quality 7 or 8 as “Premium Wine”
- We can make a logistical regression to find out what variables could influence the these two categories

Analysis and Initial Observations

Potentially influence factors in Wine Quality are observed in the following situations:

- Fix Acidity appears to have positive correlation with quality of wine
- Citric Acid has positive correlation with fix acidity, thus it has positive correlation with quality as well
- Sulphates appears to have an obvious and linear positive correlation with quality of wine
- Alcohol appears to have a positive correlation with quality of wine
- PH value as an indicator of acidity, thus wine smaller PH value appears to have better quality
- Volatile Acidity appear to have negative correlation with quality
- Total Sulfur dioxide and Free Sulfur dioxide has a trend of increasing from quality 3 to 5 and a trend of decreasing from 6 to 8
- Residual Sugar and Chlorides have no obvious correlation with quality
- Density appears to have positive correlation with sugar, thus it has no relation to quality.

Next Steps

Bivariate Analysis

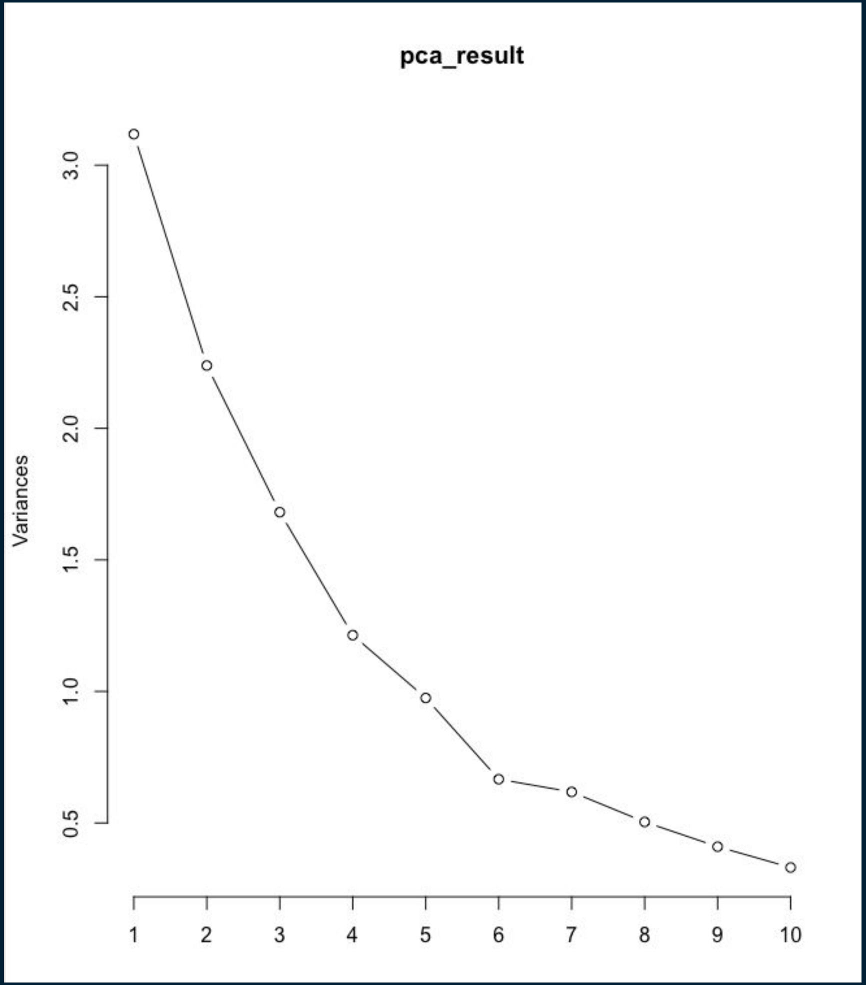
- Investigate the relationship between quality and the numeric attributes including volatile acidity, citric acid, and alcohol
- Investigate the relationship between some pairs of numeric attributes

Purpose: determine which variables have a correlation with quality

Next Steps: statistical analysis

Feature Engineering

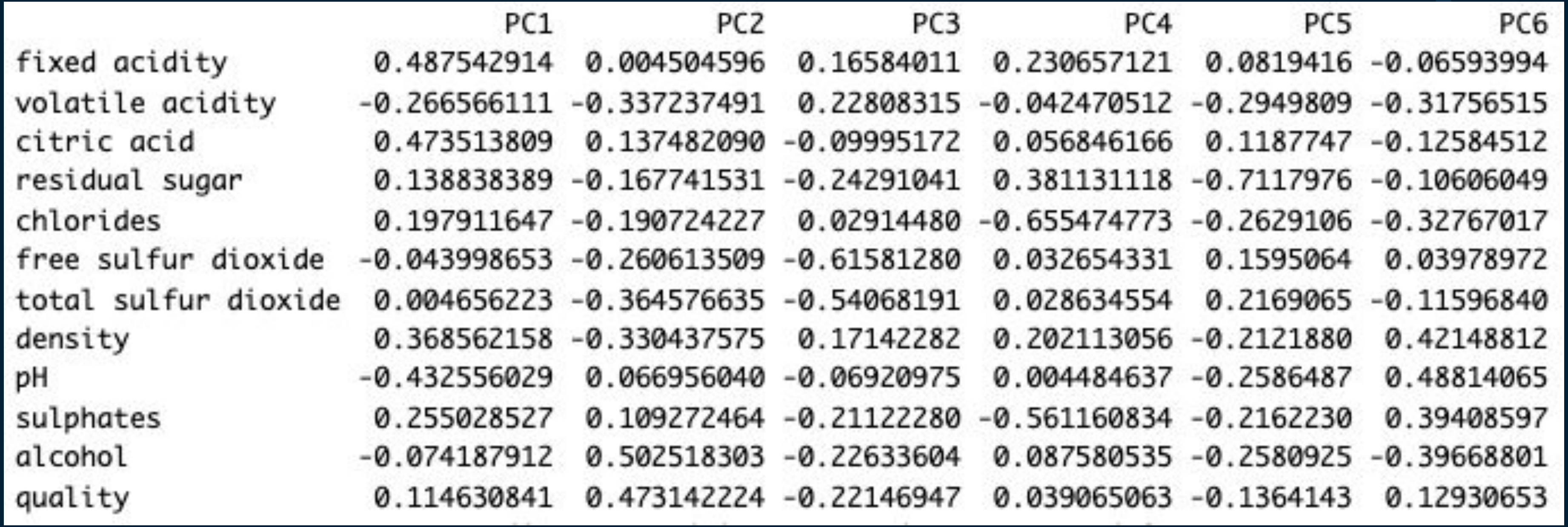
PCA



6 principal components should be used for subsequent analyses

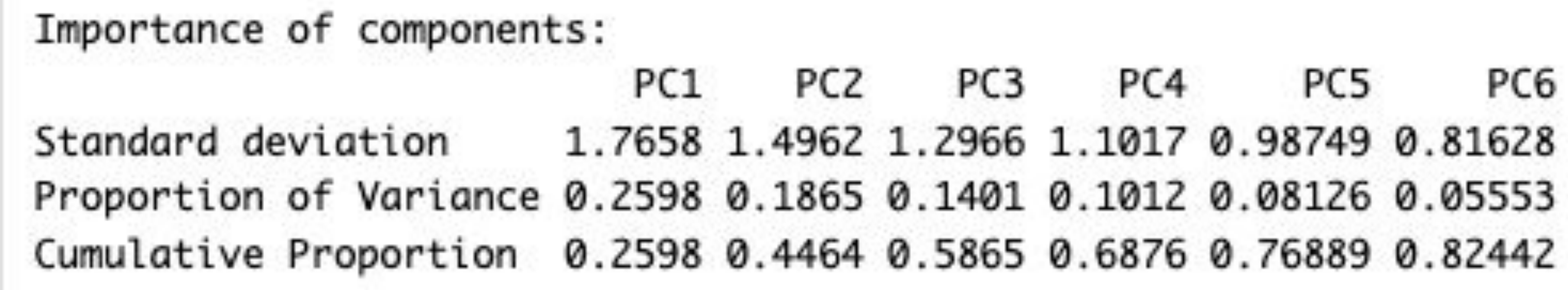
This means that the first 6 principal components capture a significant amount of variance and the loadings for these components have a clear pattern of strong influences from specific variables

PCA



```
                              PC1          PC2         PC3          PC4        PC5         PC6
fixed acidity          0.487542914  0.004504596  0.16584011  0.230657121  0.0819416 -0.06593994
volatile acidity      -0.266566111 -0.337237491  0.22808315 -0.042470512 -0.2949809 -0.31756515
citric acid            0.473513809  0.137482090 -0.09995172  0.056846166  0.1187747 -0.12584512
residual sugar         0.138838389 -0.167741531 -0.24291041  0.381131118 -0.7117976 -0.10606049
chlorides              0.197911647 -0.190724227  0.02914480 -0.655474773 -0.2629106 -0.32767017
free sulfur dioxide   -0.043998653 -0.260613509 -0.61581280  0.032654331  0.1595064  0.03978972
total sulfur dioxide   0.004656223 -0.364576635 -0.54068191  0.028634554  0.2169065 -0.11596840
density                0.368562158 -0.330437575  0.17142282  0.202113056 -0.2121880  0.42148812
pH                    -0.432556029  0.066956040 -0.06920975  0.004484637 -0.2586487  0.48814065
sulphates              0.255028527  0.109272464 -0.21122280 -0.561160834 -0.2162230  0.39408597
alcohol               -0.074187912  0.502518303 -0.22633604  0.087580535 -0.2580925 -0.39668801
quality                0.114630841  0.473142224 -0.22146947  0.039065063 -0.1364143  0.12930653
```

PCA



```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.7658 1.4962 1.2966 1.1017 0.98749 0.81628
Proportion of Variance 0.2598 0.1865 0.1401 0.1012 0.08126 0.05553
Cumulative Proportion  0.2598 0.4464 0.5865 0.6876 0.76889 0.82442
```

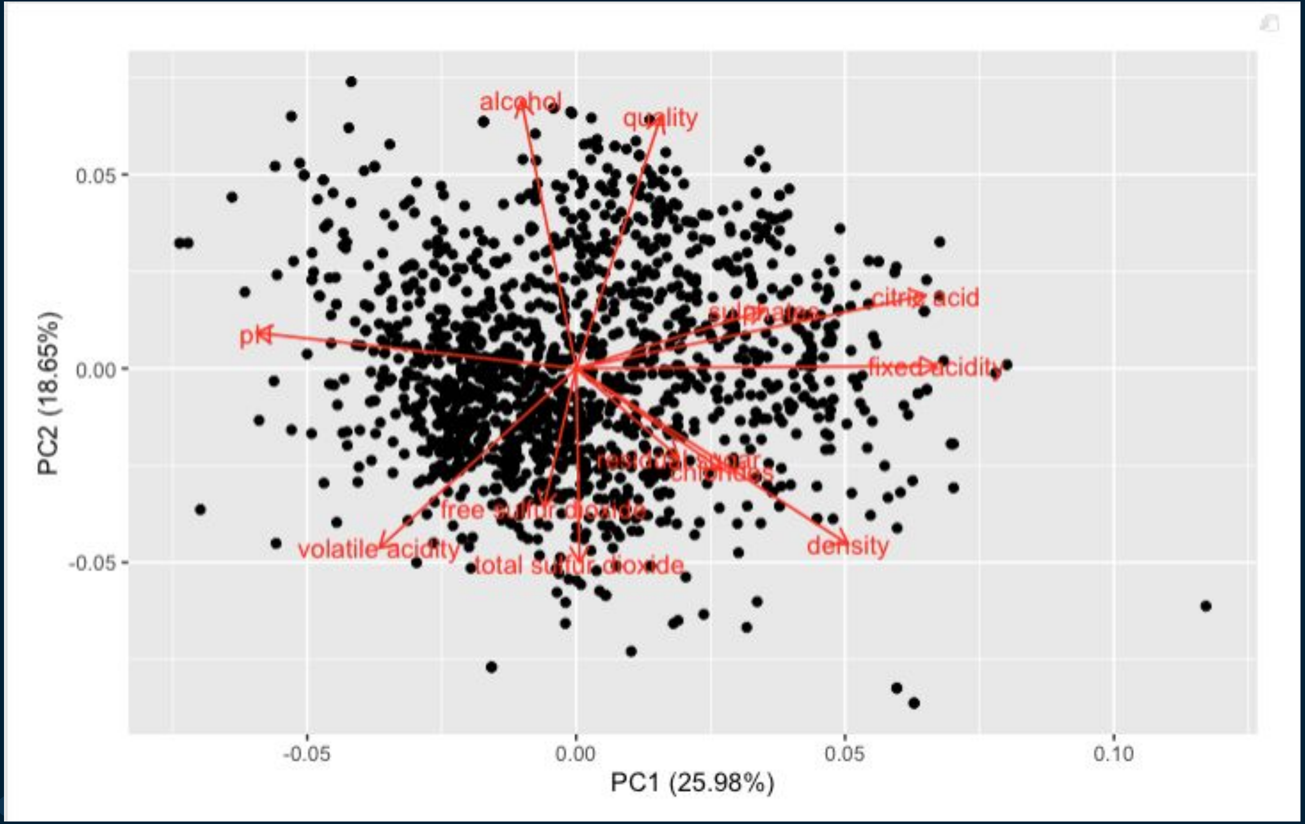
PC1 explains the majority of the total variance followed by PC2 and PC3 and so on

We only only analyze the first 4 PCAs

PCA

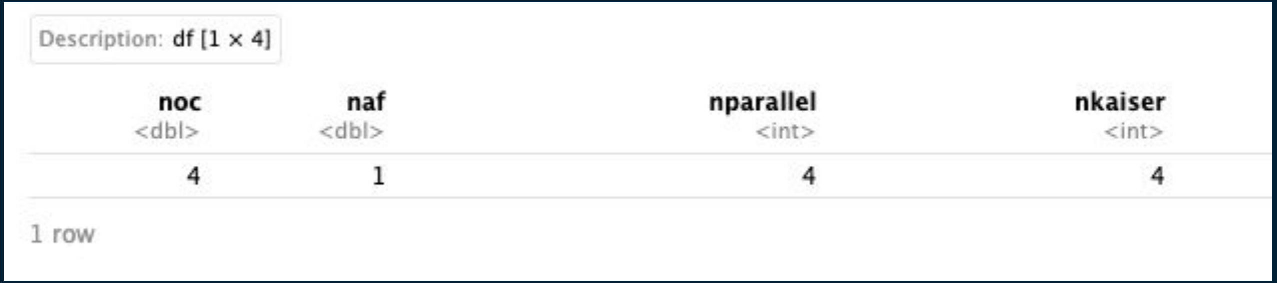
PCA	Positive values vs Negative values	Interpretation
PC1	Fixed acidity, citric acid, density vs pH	Could indicate a specific type, style or origin of wines
PC2	Alcohol, quality vs volatile acidity, total sulfur dioxide, density	Could indicate a specific type, style or origin of wines
PC3	Significantly negative values: total sulfur dioxide, free sulfur dioxide	Could indicate a certain style or type of wine
PC4	Significantly negative values: chlorides, sulphates	Could indicate a certain style or type of wine

PCA



Doesn’t tell us much

EFA



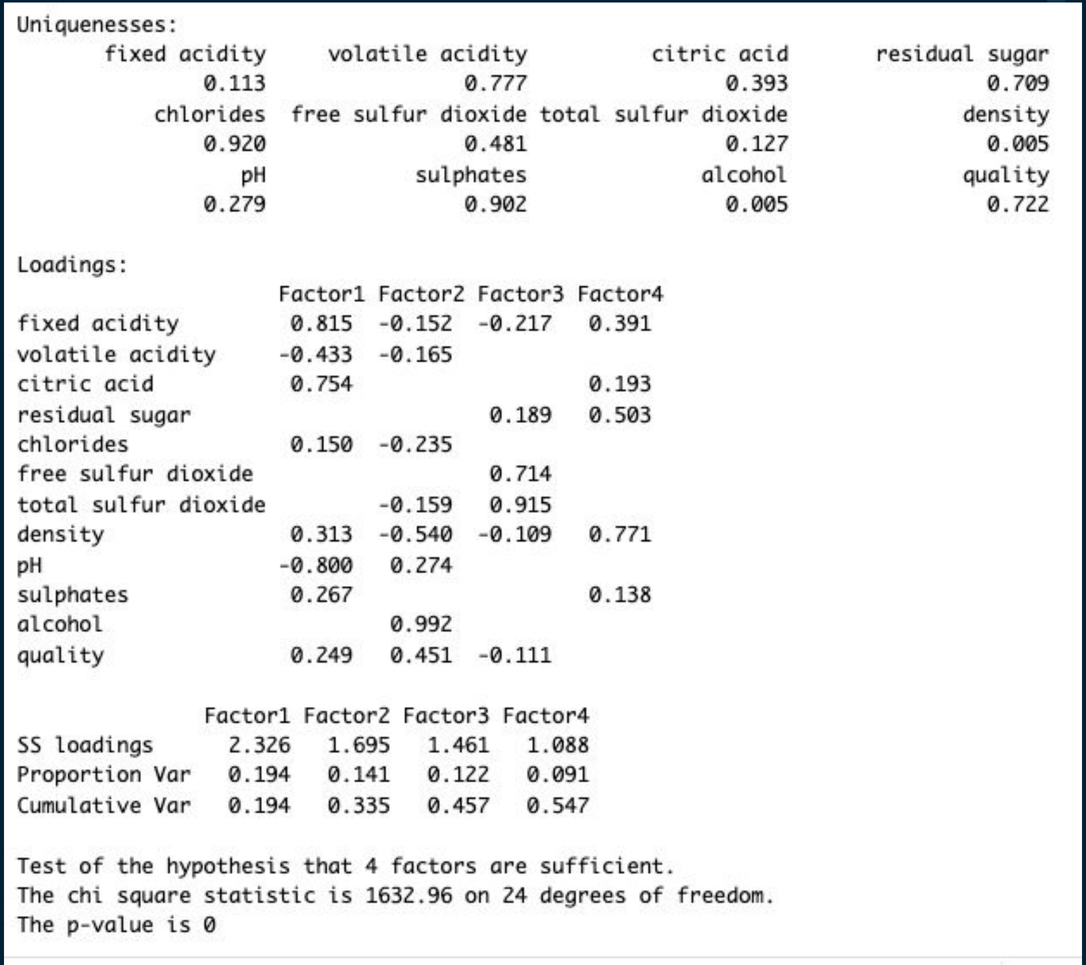
Description: df [1 × 4]

noc	naf	nparallel	nkaiser
<dbl>	<dbl>	<int>	<int>
4	1	4	4

1 row

4 factors to use

EFA



```
Uniquenesses:
       fixed acidity     volatile acidity          citric acid       residual sugar
               0.113                0.777                0.393                0.709
           chlorides  free sulfur dioxide total sulfur dioxide              density
               0.920                0.481                0.127                0.005
                  pH            sulphates              alcohol              quality
               0.279                0.902                0.005                0.722

Loadings:
                     Factor1 Factor2 Factor3 Factor4
fixed acidity         0.815  -0.152  -0.217   0.391
volatile acidity     -0.433  -0.165
citric acid           0.754                   0.193
residual sugar                        0.189   0.503
chlorides             0.150  -0.235
free sulfur dioxide                   0.714
total sulfur dioxide         -0.159   0.915
density               0.313  -0.540  -0.109   0.771
pH                   -0.800   0.274
sulphates             0.267                   0.138
alcohol                       0.992
quality               0.249   0.451  -0.111

               Factor1 Factor2 Factor3 Factor4
SS loadings      2.326   1.695   1.461   1.088
Proportion Var   0.194   0.141   0.122   0.091
Cumulative Var   0.194   0.335   0.457   0.547

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 1632.96 on 24 degrees of freedom.
The p-value is 0
```

EFA

Factors	Important values	Interpretation
1	Positive: Fixed acidity, citric acidity vs Negative: pH	Wine Quality and Composition
2	Positive: alcohol, quality vs negative: density	Types of wine
3	Total sulfur dioxide, free sulfur dioxide	Better preservation, reduced spoilage, could indicate white wines
4	Density, residual sugar	Sweet or dry wines, dessert wines, regional varations

Analysis from PCA & EFA

Potential variables to analyze for model

PCA	Positive values vs Negative values
PC1	Fixed acidity, citric acid, density vs pH
PC2	Alcohol, quality vs volatile acidity, total sulfur dioxide, density
PC3	Significantly negative values: total sulfur dioxide, free sulfur dioxide
PC4	Significantly negative values: chlorides, sulphates

EFA	
Factors	**Important values**
1	Positive: Fixed acidity, citric acidity vs Negative: pH
2	Positive: alcohol, quality vs negative: density
3	Total sulfur dioxide, free sulfur dioxide
4	Density, residual sugar

- Analyze the values in PC1 & Factor 1
- Analyze the values in PC2 & Factor 2
- Analyze the values in PC3 & Factor 3

Analysis from PCA & EFA

Potential variables to analyze for model

Positive values vs Negative values
Fixed acidity, citric acid, density vs pH
Alcohol, quality vs volatile acidity, density
total sulfur dioxide, free sulfur dioxide

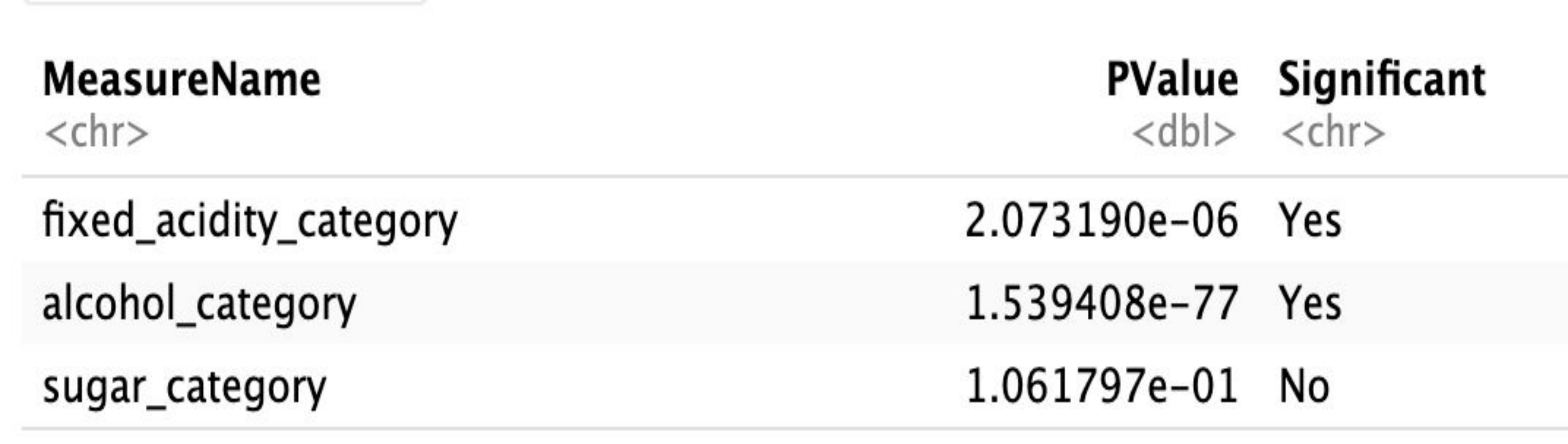
Statistical Analyses

Tests to Do

- T-Test for Measures Against Response
- Chi-square for Factors Against Response

Testing statistical significance of factors

Chi-Squared Test Results



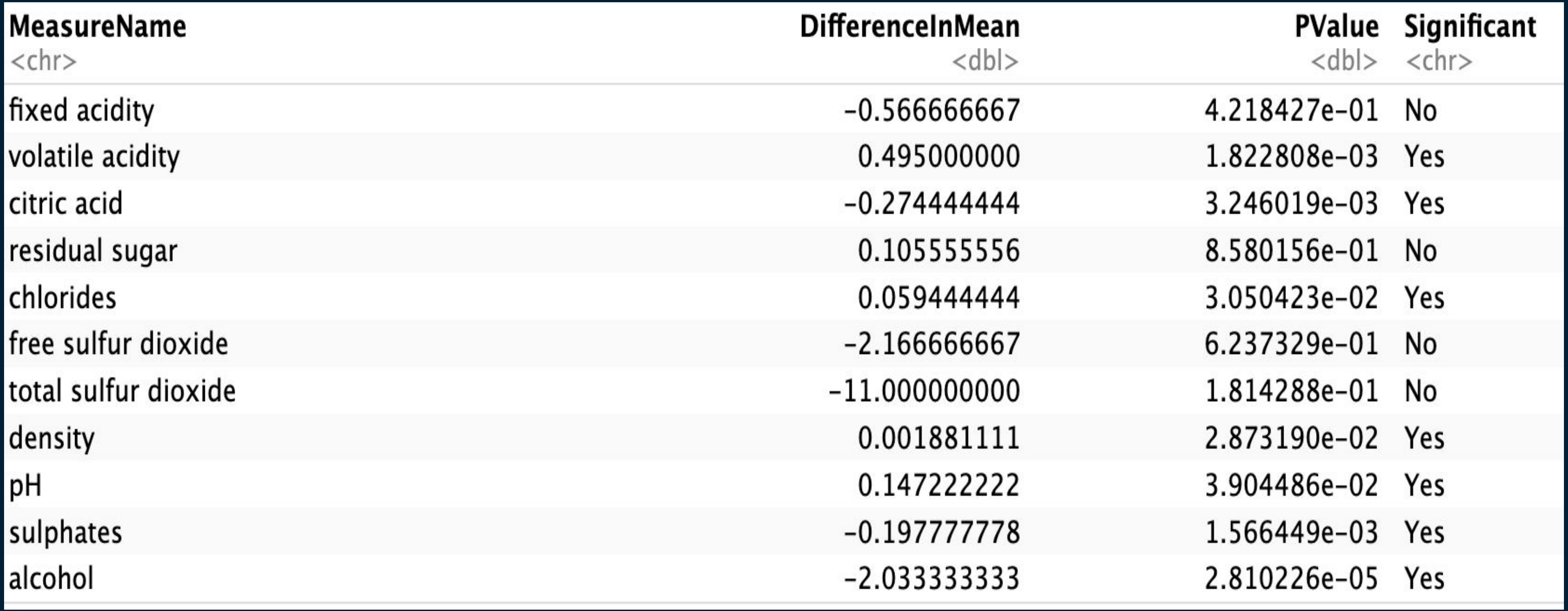
MeasureName	PValue	Significant
<chr>	<dbl>	<chr>
fixed_acidity_category	2.073190e-06	Yes
alcohol_category	1.539408e-77	Yes
sugar_category	1.061797e-01	No

Fixed Acidity and Alcohol Categories: Both show strong statistical evidence of an association with the quality

Sugar Category: Does not show a statistically significant association

Testing statistical significance of measures

T- Test Results



MeasureName <chr>	DifferenceInMean <dbl>	PValue <dbl>	Significant <chr>
fixed acidity	-0.566666667	4.218427e-01	No
volatile acidity	0.495000000	1.822808e-03	Yes
citric acid	-0.274444444	3.246019e-03	Yes
residual sugar	0.105555556	8.580156e-01	No
chlorides	0.059444444	3.050423e-02	Yes
free sulfur dioxide	-2.166666667	6.237329e-01	No
total sulfur dioxide	-11.000000000	1.814288e-01	No
density	0.001881111	2.873190e-02	Yes
pH	0.147222222	3.904486e-02	Yes
sulphates	-0.197777778	1.566449e-03	Yes
alcohol	-2.033333333	2.810226e-05	Yes

volatile acidity, citric acid, chlorides, density, pH, sulphates, and alcohol showing significant differences suggest that these properties vary substantially between different levels of wine quality

Analysis from PCA & EFA

Potential variables to analyze for model

Numeric
Values
Fixed acidity, citric acid, density vs pH
Alcohol, quality vs volatile acidity, density
Total sulfur dioxide, free sulfur dioxide
T-test & PCA: sulphates, alcohol

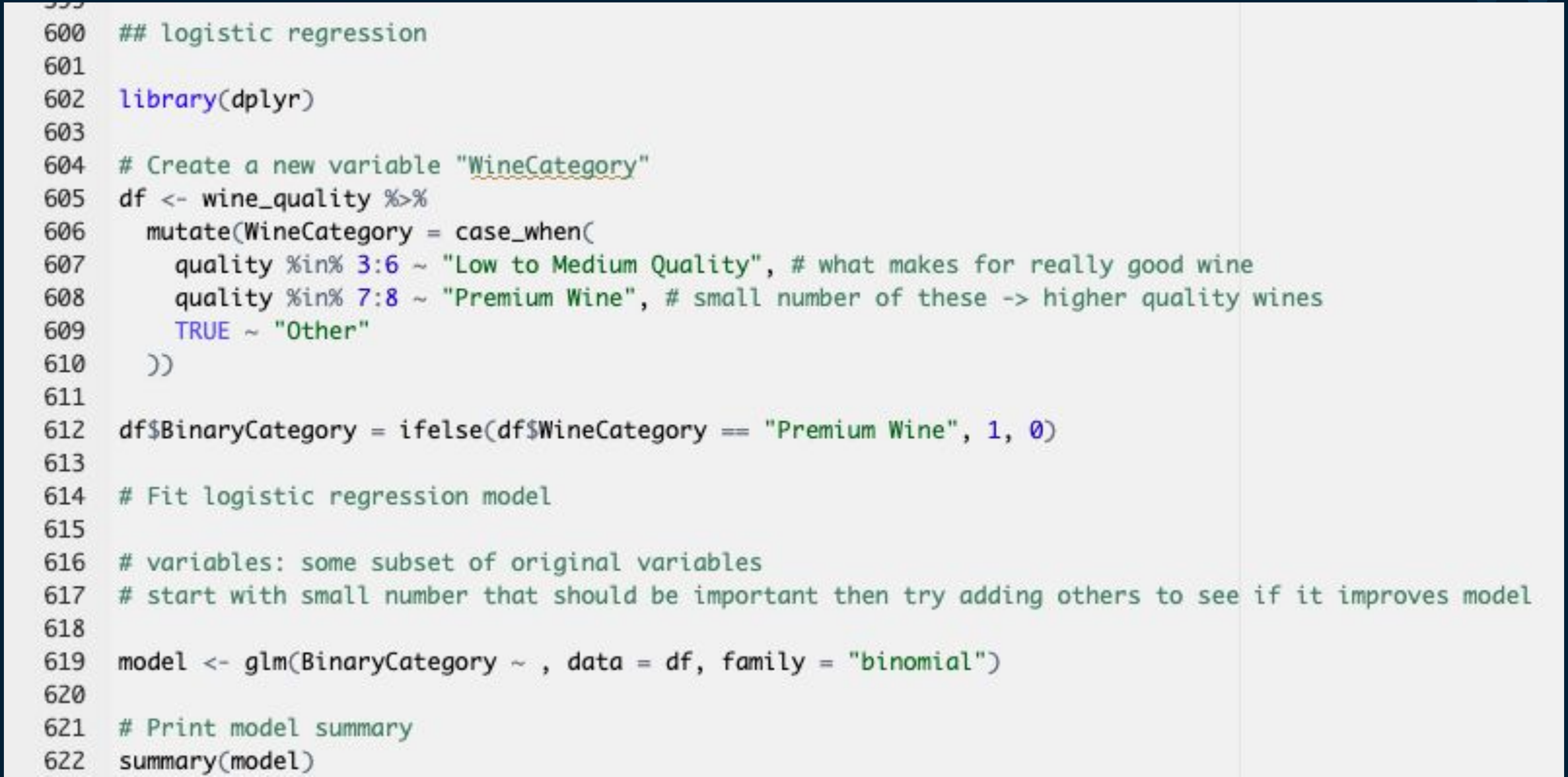
Factors
Category
Chi-square: Alcohol

Descriptive Modeling

Test to do

- one linear model
- one decision tree model
- will add a classification with 6 possible clusters
- multinomial regression
- Model 1, 2, 3 similar to sample EDA report

Binary Model



```r
## logistic regression

library(dplyr)

# Create a new variable "WineCategory"
df <- wine_quality %>%
  mutate(WineCategory = case_when(
    quality %in% 3:6 ~ "Low to Medium Quality", # what makes for really good wine
    quality %in% 7:8 ~ "Premium Wine", # small number of these -> higher quality wines
    TRUE ~ "Other"
  ))

df$BinaryCategory = ifelse(df$WineCategory == "Premium Wine", 1, 0)

# Fit logistic regression model

# variables: some subset of original variables
# start with small number that should be important then try adding others to see if it improves model

model <- glm(BinaryCategory ~ , data = df, family = "binomial")

# Print model summary
summary(model)
```

Logistic Regression Models

Model 1 - Limited Numerics

Based on results of EDA and statistical analyses, an initial logistic regression model was created with the following variables:

-

Clustering

Analyses Plan