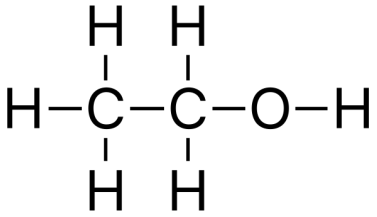
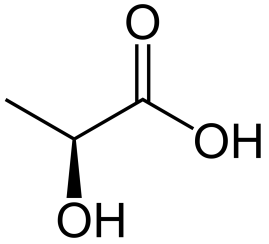
Regression Analysis for White Wine

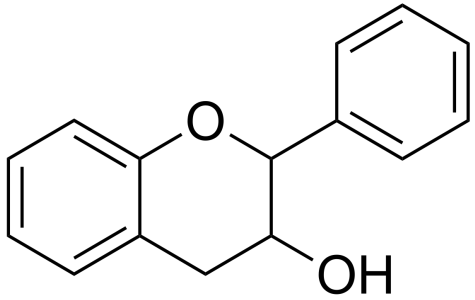
Elizabeth Do

Problem Statement

Can the chemical properties of wine predict its quality?







Key Questions

What variables contribute the most to the quality of a white wine?

Why might these variables contribute to the quality and in what way (positive or negative)?

Can these variables be used to predict the quality of future wine datasets?

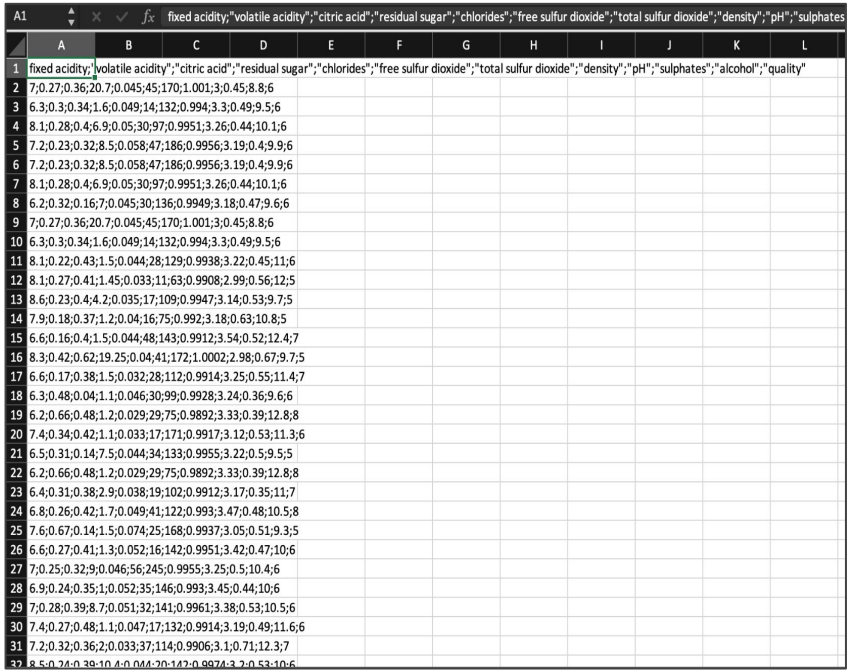
Data Sources

- .csv file from UCI Machine Learning Repository
- Dataset was created using red and white wine samples
- White wine variants of the Portuguese "Vinho Verde" wine
- 4,898 Observation & 12 Variables
- 0 missing attribute values
- All variables contain doubles/integers

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

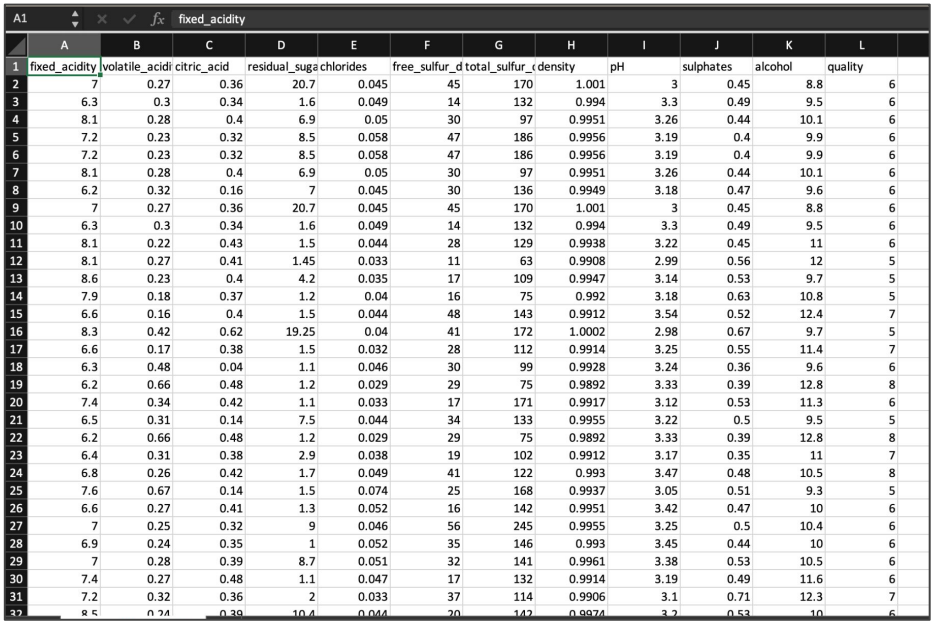
Data Organization/Wrangling

Original .csv file



	A
1	fixed acidity;"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide";"density";"pH";"sulphates";"alcohol";"quality"
2	7;0.27;0.36;20.7;0.045;45;170;1.001;3;0.45;8.8;6
3	6.3;0.3;0.34;1.6;0.049;14;132;0.994;3.3;0.49;9.5;6
4	8.1;0.28;0.4;6.9;0.05;30;97;0.9951;3.26;0.44;10.1;6
5	7.2;0.23;0.32;8.5;0.058;47;186;0.9956;3.19;0.4;9.9;6
6	7.2;0.23;0.32;8.5;0.058;47;186;0.9956;3.19;0.4;9.9;6
7	8.1;0.28;0.4;6.9;0.05;30;97;0.9951;3.26;0.44;10.1;6
8	6.2;0.32;0.16;7;0.045;30;136;0.9949;3.18;0.47;9.6;6
9	7;0.27;0.36;20.7;0.045;45;170;1.001;3;0.45;8.8;6
10	6.3;0.3;0.34;1.6;0.049;14;132;0.994;3.3;0.49;9.5;6
11	8.1;0.22;0.43;1.5;0.044;28;129;0.9938;3.22;0.45;11;6
12	8.1;0.27;0.41;1.45;0.033;11;63;0.9908;2.99;0.56;12;5
13	8.6;0.23;0.4;4.2;0.035;17;109;0.9947;3.14;0.53;9.7;5
14	7.9;0.18;0.37;1.2;0.04;16;75;0.992;3.18;0.63;10.8;5
15	6.6;0.16;0.4;1.5;0.044;48;143;0.9912;3.54;0.52;12.4;7
16	8.3;0.42;0.62;19.25;0.04;41;172;1.0002;2.98;0.67;9.7;5
17	6.6;0.17;0.38;1.5;0.032;28;112;0.9914;3.25;0.55;11.4;7
18	6.3;0.48;0.04;1.1;0.046;30;99;0.9928;3.24;0.36;9.6;6
19	6.2;0.66;0.48;1.2;0.029;29;75;0.9892;3.33;0.39;12.8;8
20	7.4;0.34;0.42;1.1;0.033;17;171;0.9917;3.12;0.53;11.3;6
21	6.5;0.31;0.14;7.5;0.044;34;133;0.9955;3.22;0.5;9.5;5
22	6.2;0.66;0.48;1.2;0.029;29;75;0.9892;3.33;0.39;12.8;8
23	6.4;0.31;0.38;2.9;0.038;19;102;0.9912;3.17;0.35;11;7
24	6.8;0.26;0.42;1.7;0.049;41;122;0.993;3.47;0.48;10.5;8
25	7.6;0.67;0.14;1.5;0.074;25;168;0.9937;3.05;0.51;9.3;5
26	6.6;0.27;0.41;1.3;0.052;16;142;0.9951;3.42;0.47;10;6
27	7;0.25;0.32;9;0.046;56;245;0.9955;3.25;0.5;10.4;6
28	6.9;0.24;0.35;1;0.052;35;146;0.993;3.45;0.44;10;6
29	7;0.28;0.39;8.7;0.051;32;141;0.9961;3.38;0.53;10.5;6
30	7.4;0.27;0.48;1.1;0.047;17;132;0.9914;3.19;0.49;11.6;6
31	7.2;0.32;0.36;2;0.033;37;114;0.9906;3.1;0.71;12.3;7
32	8.5;0.24;0.39;10.4;0.044;20;142;0.9974;3.2;0.53;10;6

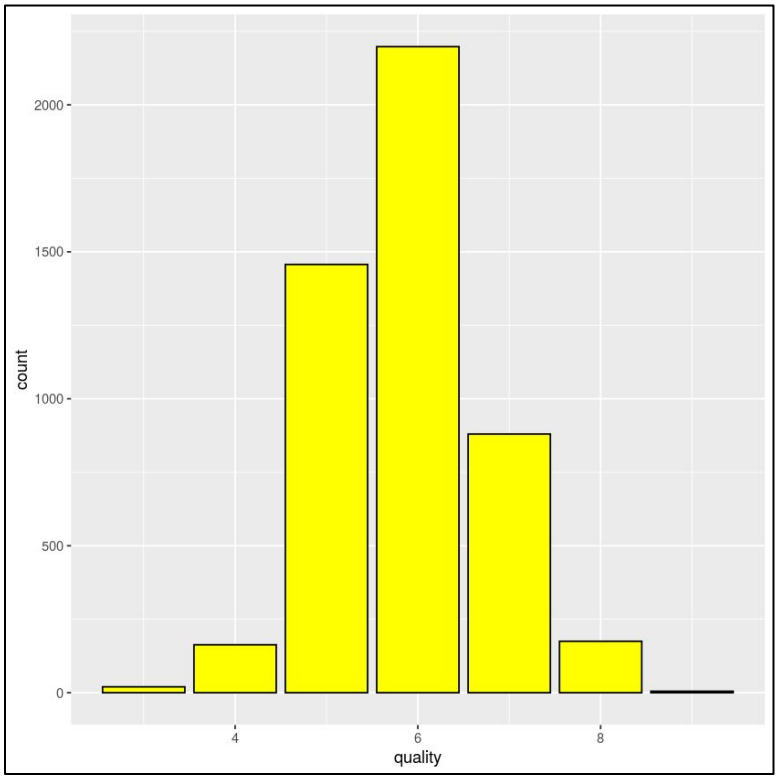
Clean .csv file



	A	B	C	D	E	F	G	H	I	J	K	L
1	fixed_acidity	volatile_acidi	citric_acid	residual_suga	chlorides	free_sulfur_d	total_sulfur_(	density	pH	sulphates	alcohol	quality
2	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
3	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
4	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6
5	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6
6	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6
7	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6
8	6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18	0.47	9.6	6
9	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
10	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
11	8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22	0.45	11	6
12	8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99	0.56	12	5
13	8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14	0.53	9.7	5
14	7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18	0.63	10.8	5
15	6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54	0.52	12.4	7
16	8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98	0.67	9.7	5
17	6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25	0.55	11.4	7
18	6.3	0.48	0.04	1.1	0.046	30	99	0.9928	3.24	0.36	9.6	6
19	6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	8
20	7.4	0.34	0.42	1.1	0.033	17	171	0.9917	3.12	0.53	11.3	6
21	6.5	0.31	0.14	7.5	0.044	34	133	0.9955	3.22	0.5	9.5	5
22	6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	8
23	6.4	0.31	0.38	2.9	0.038	19	102	0.9912	3.17	0.35	11	7
24	6.8	0.26	0.42	1.7	0.049	41	122	0.993	3.47	0.48	10.5	8
25	7.6	0.67	0.14	1.5	0.074	25	168	0.9937	3.05	0.51	9.3	5
26	6.6	0.27	0.41	1.3	0.052	16	142	0.9951	3.42	0.47	10	6
27	7	0.25	0.32	9	0.046	56	245	0.9955	3.25	0.5	10.4	6
28	6.9	0.24	0.35	1	0.052	35	146	0.993	3.45	0.44	10	6
29	7	0.28	0.39	8.7	0.051	32	141	0.9961	3.38	0.53	10.5	6
30	7.4	0.27	0.48	1.1	0.047	17	132	0.9914	3.19	0.49	11.6	6
31	7.2	0.32	0.36	2	0.033	37	114	0.9906	3.1	0.71	12.3	7
32	8.5	0.24	0.39	10.4	0.044	20	142	0.9974	3.2	0.53	10	6

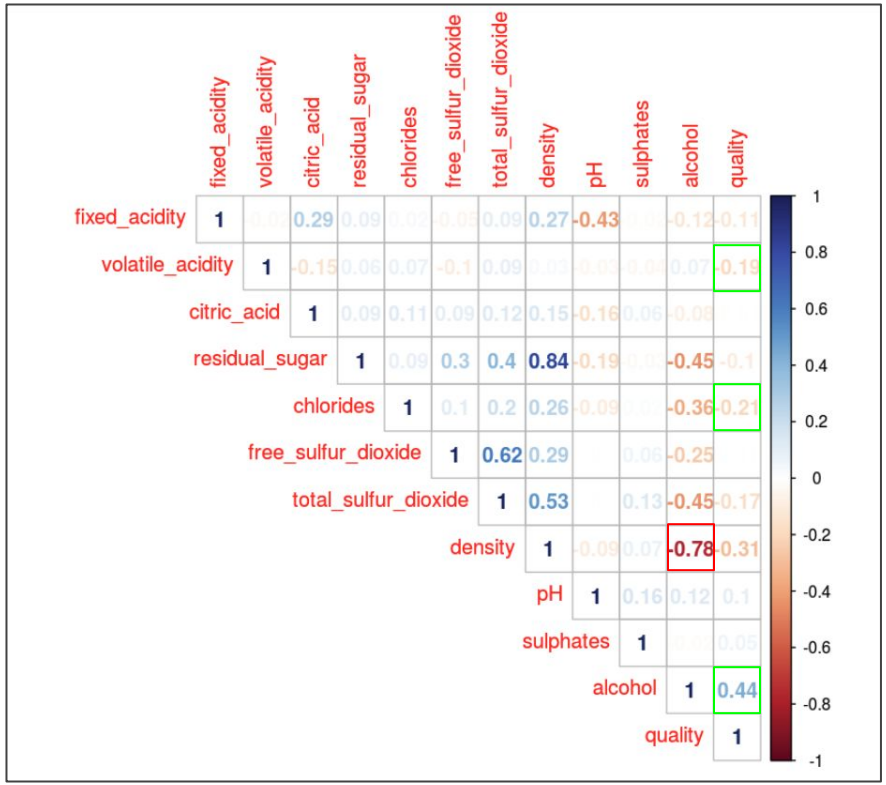
Data Exploratoration: Response Variable

- **Quality**: An integer score between 1-10 assigned to a wine.
- Quality can be subjective and is usually determined by four key indicators
  - Complexity
  - Balance
  - Typicity
  - Finish
- In our dataset, the distribution of quality ranges from 3 (worst) to 9 (best) and is relatively normal



Data Exploratoration: Explanatory Variables

- We created a correlation plot to view the relationships between all variables and select the explanatory variables for our models
- Found the highest correlations with Quality to be
  - Alcohol
  - Chlorides
  - Volatile Acidity
  - *Density*
- Due to Density's strong negative correlation with Alcohol, it was dropped from consideration

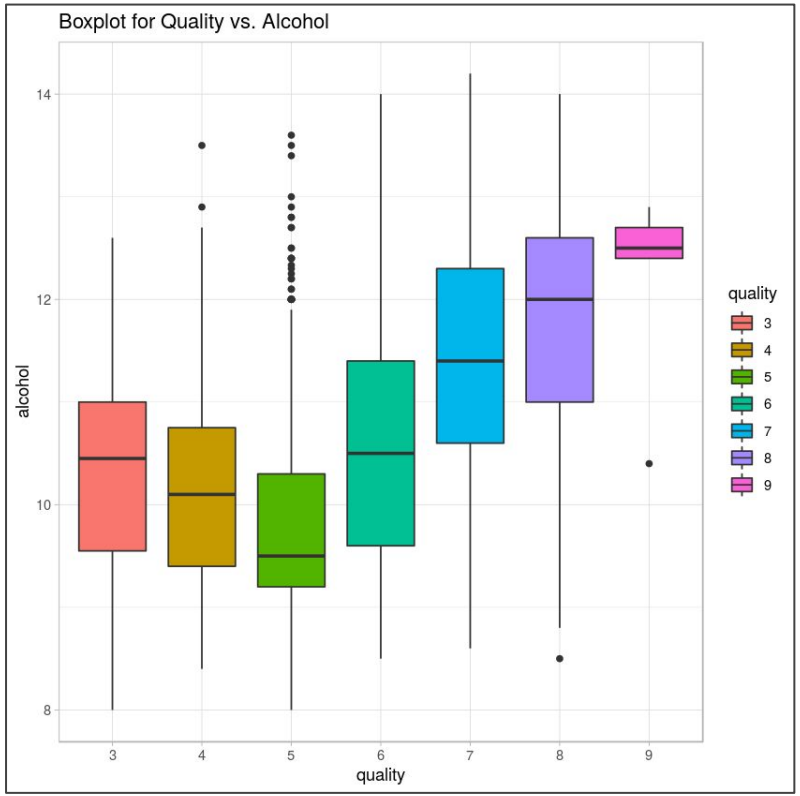


Data Exploratoration: Explanatory Variables

- **Alcohol**: The percentage of alcohol present in the wine. Wines with higher alcohol percentage tend to be more favorable.
  - *Min: 8%*
  - *Median: 10.4%*
  - *Max: 14.2%*
- **Chlorides**: The concentration of chlorides in the wine. Wines with higher concentrations of chlorides tend to be more salty.
  - *Min: 0.009 g/L*
  - *Median: 0.043 g/L*
  - *Max: 0.346 g/L*
- **Volatile Acidity**: The presence of acetic acid in the wine. High concentrations of acetic acid can contribute to a vinegar-like aroma.
  - *Min: 0.08 g/L*
  - *Median: 0.26 g/L*
  - *Max: 1.10 g/L*

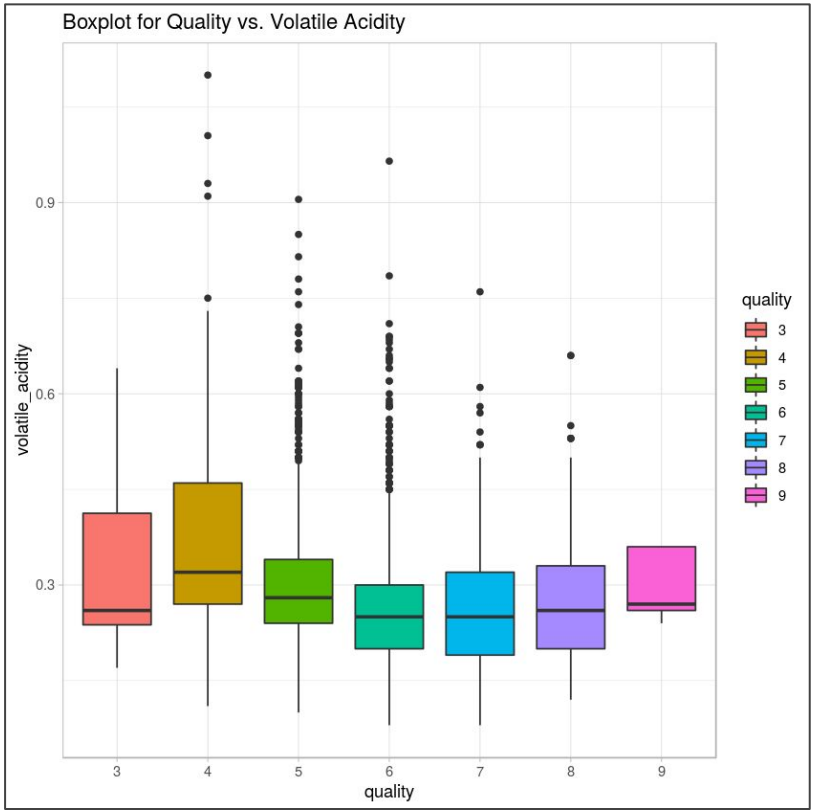
Data Visuals (Alcohol)

- Distribution between wine quality and alcohol level
- Wine quality 9 has the highest average alcohol level while wine quality 5 has the lowest average alcohol level
- Wine quality 5 has the most outliers compared to all the other wine quality categories
- Can assume that the higher the alcohol level, the higher the wine quality



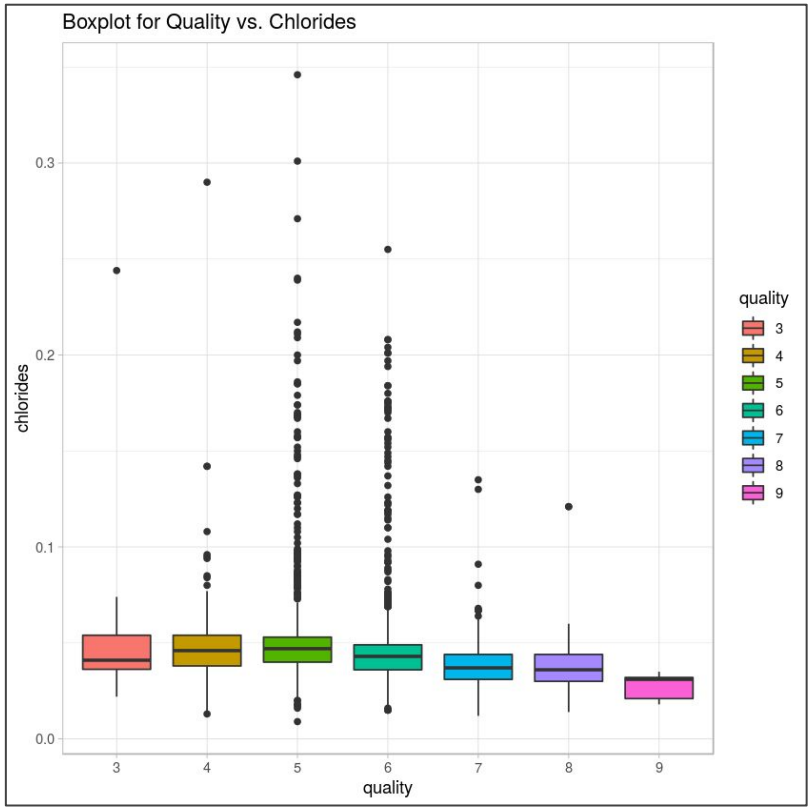
Data Visuals (Volatile Acid)

- Boxplot to show distributions between wine quality and volatile acidity
- Wine quality 5 & 6 shows more outliers compared to the other wine quality categories
- Volatile acidity ranges more closely within the different wine quality ranges

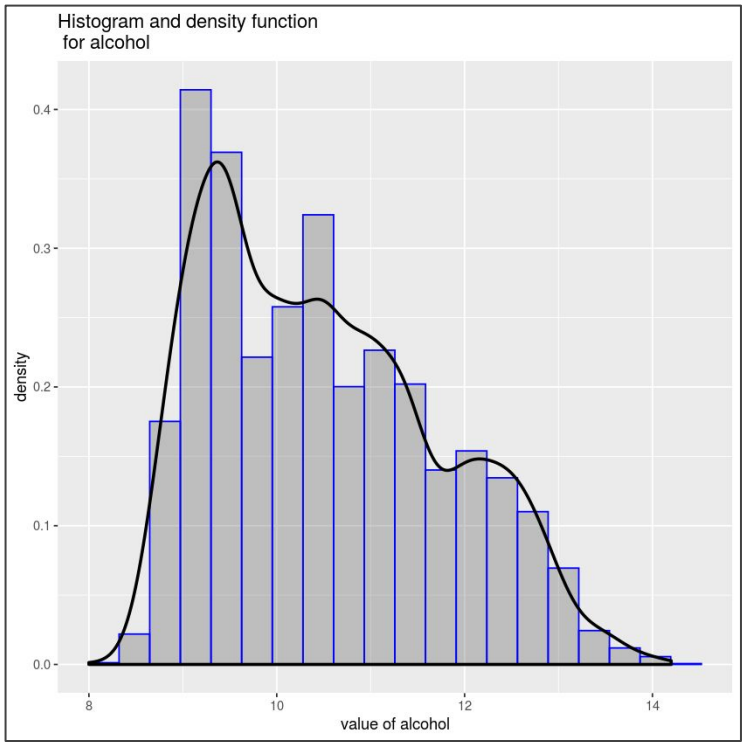


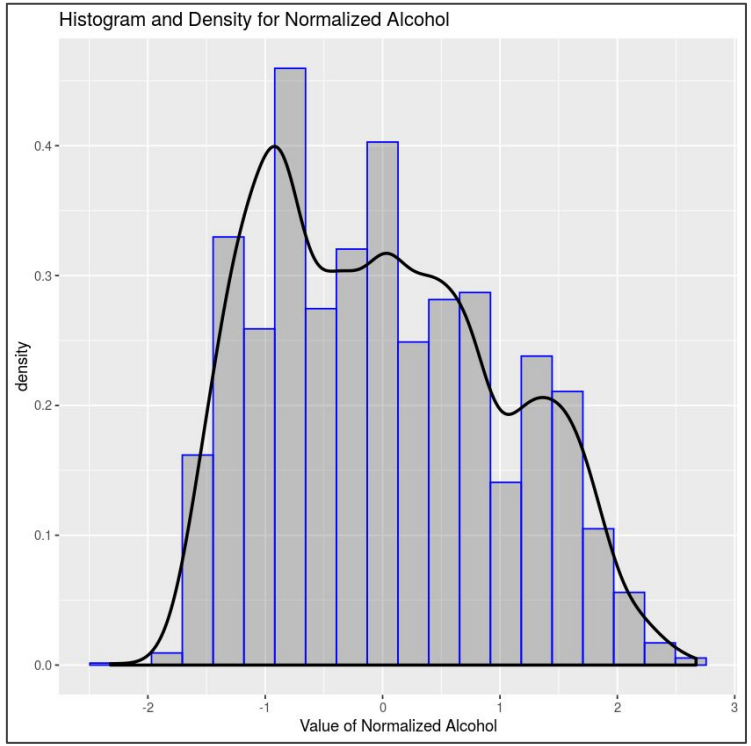
Data Visuals (Chlorides)

- Boxplot to show distribution between wine quality and chloride levels
- Wine quality 5 & 6 shows more outliers compared to the other wine quality categories
- Higher wine quality seems to show lower levels of chlorides

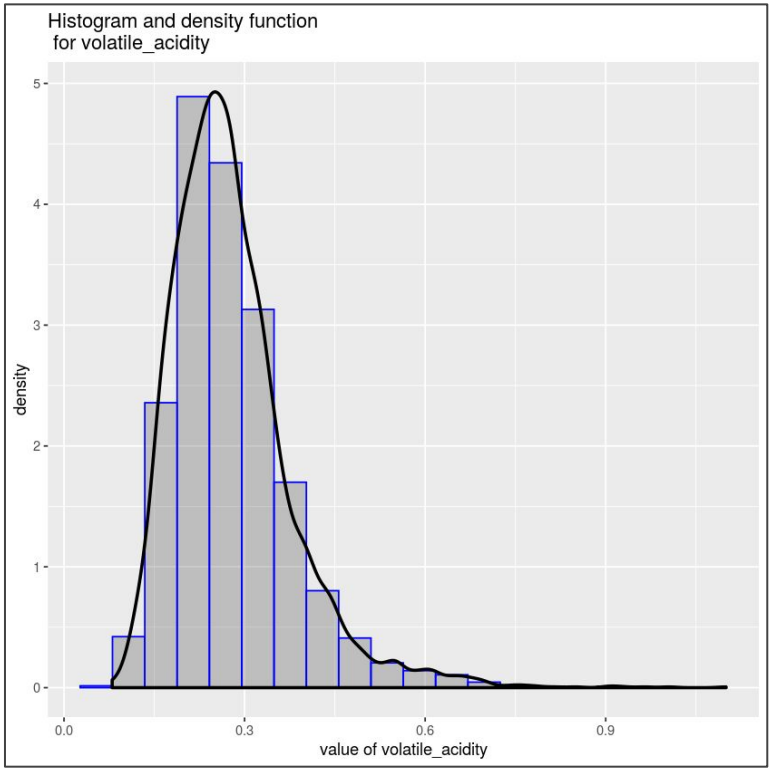


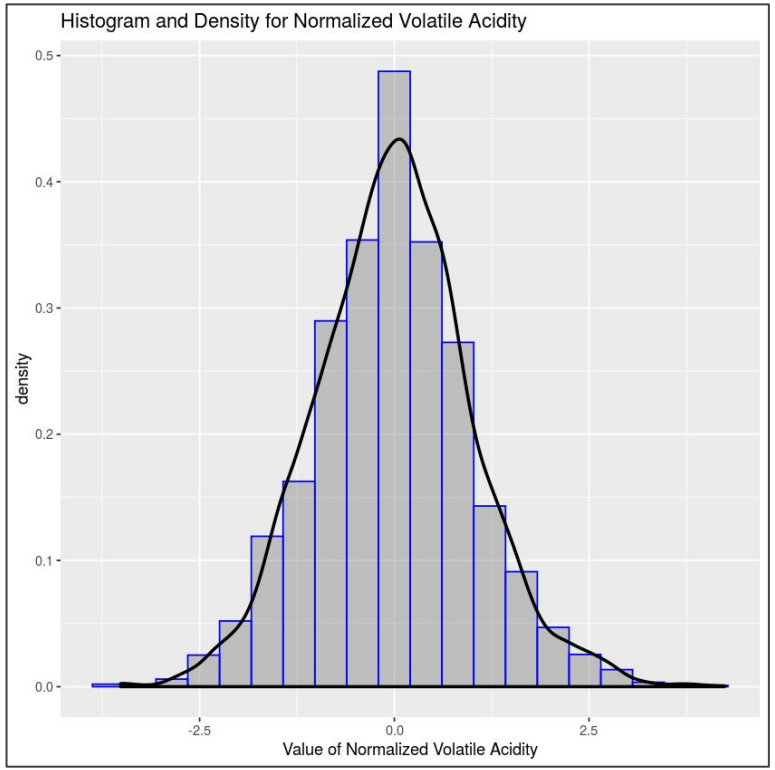
Data Normalization (Alcohol)



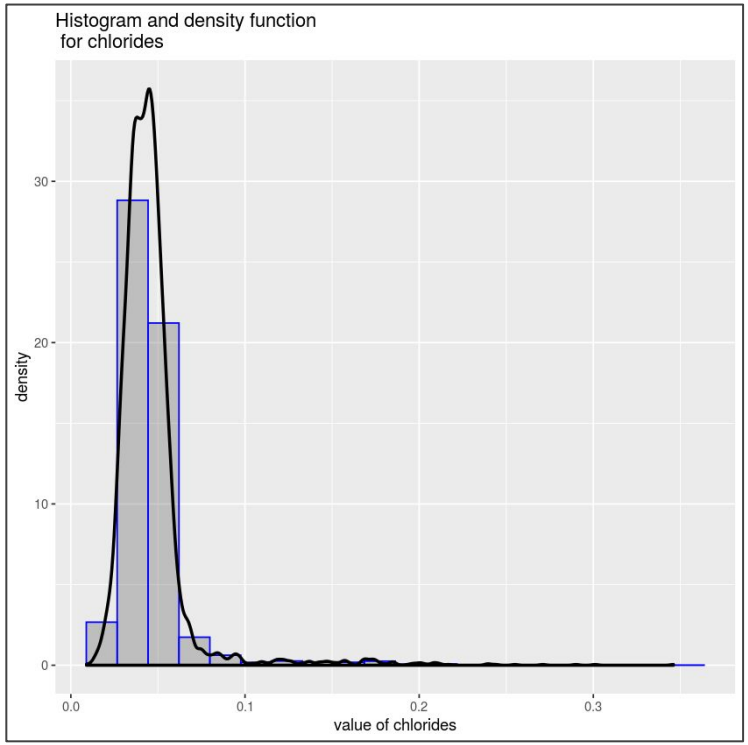


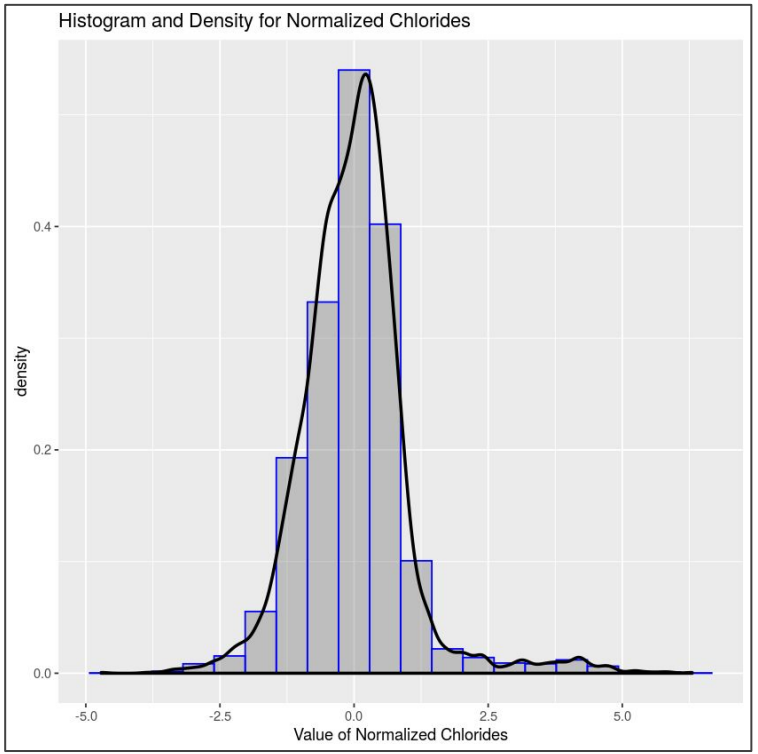
Data Normalization (Volatile Acidity)





Data Normalization (Chlorides)





Hypothesis Testing

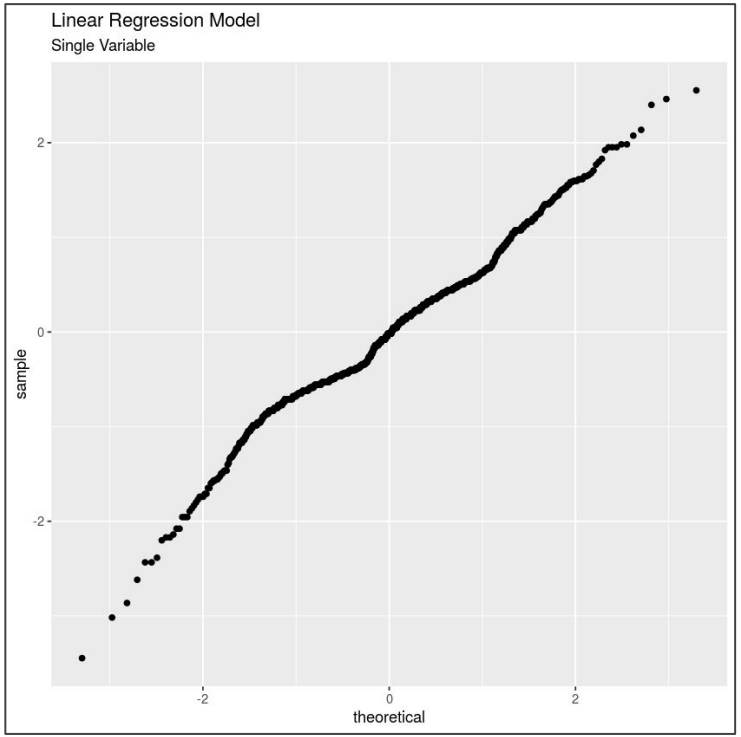
- Performed one sided and two sided t-test models for wine qualities 5 & 8 at 95% confidence level
- P-Value: 2.2e-16
- Confidence interval: -2.022977 ~ -1.631343
- Reject null hypothesis

```
        Welch Two Sample t-test

data:  df_5$alcohol and df_8$alcohol
t = -18.404, df = 192.72, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.022977 -1.631343
sample estimates:
mean of x mean of y
  9.80884  11.63600
```

Prediction Models (Single Variable)

- Q-Q plot for wine quality and alcohol level
- Predicting variable
  - Alcohol (positive effect)
- Split original dataset into train (70%) and test (30%) sets
- Alcohol was statistically significant
- Determined alcohol is a good predictor for wine quality



Prediction Models (Multivariate)

- Predicting variables
  - Alcohol (positive effect)
  - Volatile Acidity (negative effect)
  - Chlorides (negative effect)
- Split original dataset into train (70%) and test (30%) sets
- All variables were statistically significant
- Q-Q plot for wine quality and alcohol level
- Determined, when used together, alcohol, volatile acidity, and chlorides are good predictors of wine quality