# Red Wine Quality Analysis

**Author:** Shambhavi Danayak
**Student Id:** 012654513
**Contact:** sdanayak@csuchico.edu


CSCI 608 - Data Science

California State University – Chico

Chico, CA 95928

# PROBLEM FORMULATION

This project uses the Wine Quality dataset from the UCI website to examine how factors such as acidity, alcohol percentage, sugar, and pH are related to red wine quality.

## Descriptive Question

***Question:*** What is the distribution of observed quality ratings for red wine in the dataset?
***Goal:*** summarize how the observed wine quality scores are distributed across all samples and most common quality scores.

## Exploratory Question

***Question:*** how does alcohol content affect wine quality? Determine whether higher alcohol content is associated with higher quality ratings. Which variables show the strongest relationship with wine quality ratings?

## Predictive Question

***Question:*** Can a predictive model accurately estimate wine quality based on measurable chemical attributes as alcohol content, acidity, sulphates, and pH?

## Inferential Question

***Question:*** Is there a meaningful difference in the average alcohol content between high-quality and low-quality red wines, and what is the 95% confidence interval for this difference using bootstrap sampling?

---

# DATASET COLLECTION

The dataset used in this project is obtained from the UCI Machine Learning Repository, a publicly available and widely used source.

- Dataset used: UCI Wine Quality Dataset
- Format of Dataset: csv
- Number of observations: 1599
- Number of variables: 12
- Variables:

- **Input variables:**
    1. fixed acidity
    2. volatile acidity
    3. citric acid
    4. residual sugar
    5. Chlorides
    6. free sulfur dioxide
    7. total sulfur dioxide
    8. Density
    9. pH
    10. Sulphates
    11. alcohol
- **Output variable:** 12 - quality (score between 0 and 10)

[7]:

| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |
| 1 | 7.8 | 0.88 | 0.00 | 2.6 | 0.098 | 25.0 | 67.0 | 0.9968 | 3.20 | 0.68 | 9.8 | 5 |
| 2 | 7.8 | 0.76 | 0.04 | 2.3 | 0.092 | 15.0 | 54.0 | 0.9970 | 3.26 | 0.65 | 9.8 | 5 |
| 3 | 11.2 | 0.28 | 0.56 | 1.9 | 0.075 | 17.0 | 60.0 | 0.9980 | 3.16 | 0.58 | 9.8 | 6 |
| 4 | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |

```
4    7.4    0.70    0.00    1.9    0.076    11.0

[8]: df.shape

[8]: (1599, 12)

[9]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB

[10]: df.isna().sum()

[10]: fixed acidity           0
      volatile acidity        0
      citric acid             0
      residual sugar          0
      chlorides               0
      free sulfur dioxide     0
      total sulfur dioxide    0
      density                 0
      pH                      0
      sulphates               0
      alcohol                 0
      quality                 0
      dtype: int64
```

# DATA PREPARATION

Dataset is examined to ensure it is clean, if there are any missing values, number to entries, number of variables, data types and structure. Since no missing or inconsistent values are found, the dataset is ready for descriptive, exploratory, predictive, and inferential analysis.

- Shape of the dataset: (1599, 12)

```
[9]: df.shape

[9]: (1599, 12)
```

- Data types and structure

```
[10]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

- Missing value check

```
memory usage: 150.0 KB

[11]: df.isna().sum()

[11]: fixed acidity           0
      volatile acidity        0
      citric acid             0
      residual sugar          0
      chlorides               0
      free sulfur dioxide     0
      total sulfur dioxide    0
      density                 0
      pH                      0
      sulphates               0
      alcohol                 0
      quality                 0
      dtype: int64
```

- Working copy of the dataset created for future use

```
[12]: array([5, 6, 7, 4, 8, 3])

[13]: wine_df = df.copy()
      wine_df.head()
```

[13]:

| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |
| 1 | 7.8 | 0.88 | 0.00 | 2.6 | 0.098 | 25.0 | 67.0 | 0.9968 | 3.20 | 0.68 | 9.8 | 5 |
| 2 | 7.8 | 0.76 | 0.04 | 2.3 | 0.092 | 15.0 | 54.0 | 0.9970 | 3.26 | 0.65 | 9.8 | 5 |
| 3 | 11.2 | 0.28 | 0.56 | 1.9 | 0.075 | 17.0 | 60.0 | 0.9980 | 3.16 | 0.58 | 9.8 | 6 |
| 4 | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |

- Different quality scores in the dataset

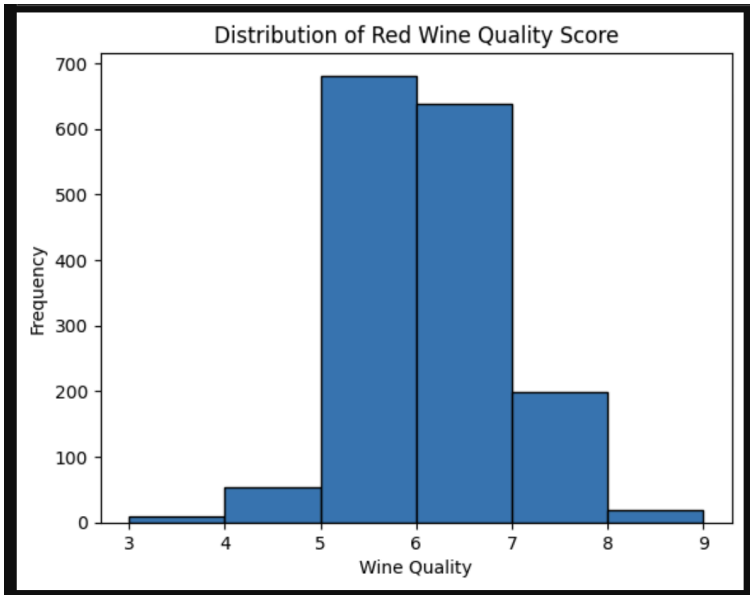
```
        dtype: int64

[12]: df['quality'].unique()

[12]: array([5, 6, 7, 4, 8, 3])
```

---

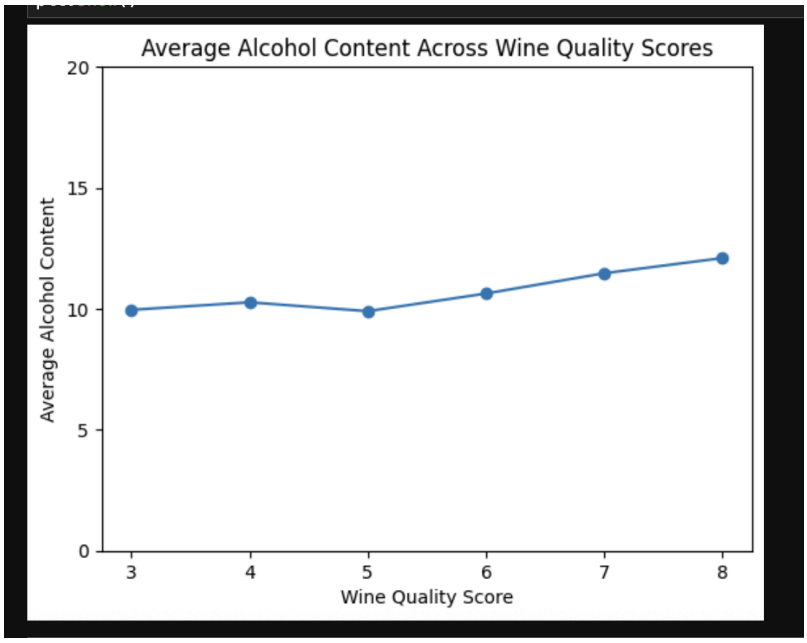
# DESCRIPTIVE ANALYSIS

In this analysis the distribution of wine quality ratings is examined to understand how quality scores are spread across the dataset. Multiple Visualizations were made to examine the distribution.

- **Histogram Showing Quality score distribution**
  - The histogram shows the distribution of red wine quality ratings in the dataset.
  - X-axis = quality ratings; Y-axis = frequency
  - Key Observations:
    - Wine quality ranges between 3 to 9
    - Most wines are rated 5 or 6, indicating most wines in the dataset are average of average quality ratings
    - Very low (3–4) and very high (8–9) quality ratings are rare
    - The distribution is centered around average quality levels

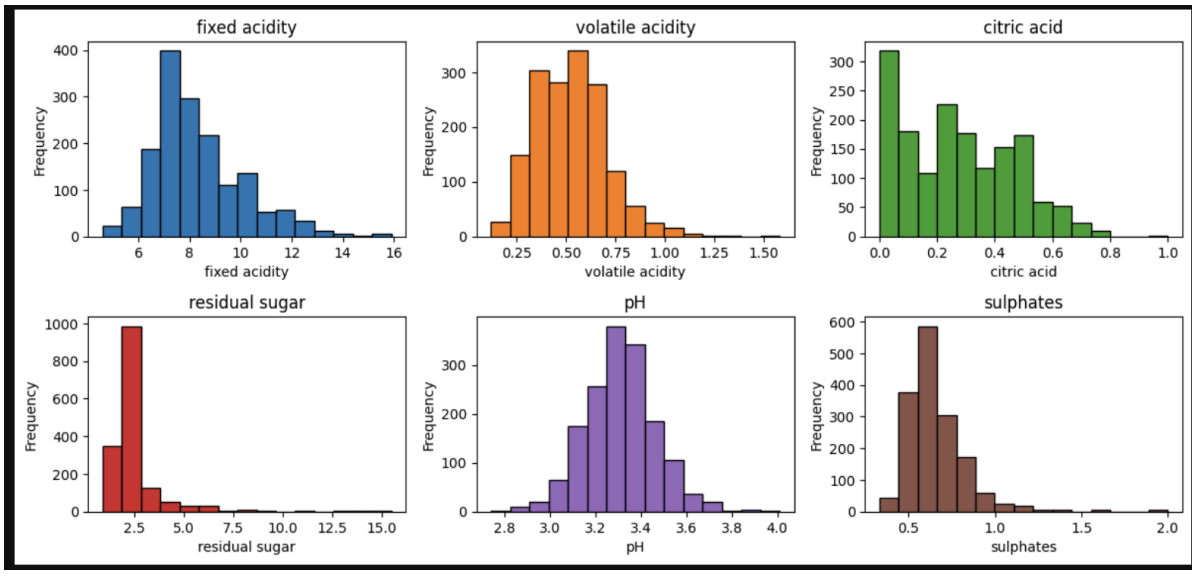


- ## Line Plot Showing Average Alcohol Content Per Quality Score
    - This plot shows the average alcohol content for each wine quality score. This plot gives real insight into the relationship between wine quality score and how much alcohol content it has.
    - X-axis = quality ratings; Y-axis = average alcohol content
    - Key Observations:
        - Average alcohol content generally increases as wine quality scores increase
        - Wines with quality scores 5 and 6 have similar average alcohol levels
        - High quality wines (7 and 8) have higher average alcohol content
        - Low quality wines (3–4) have lower average alcohol levels




- ## Histograms descriptive of key physicochemical variables
    - Histograms below summarize the distributions of key physicochemical variables in the dataset.
    - **Observations from fixed acidity vs frequency:**
        - Most red wines have fixed acidity values concentrated between approximately 6 and 9.
        - The distribution is right-skewed, with fewer wines having very high fixed acidity levels.

        - The majority of observations cluster around a moderate fixed acidity range, indicating limited variability for most wines.
    - **Observations from volatile acidity vs frequency:**
        - Most red wines have volatile acidity values concentrated between approximately 0.3 and 0.7.
        - The distribution is right-skewed, with a small number of wines showing higher volatile acidity levels.
        - The majority of wines cluster around moderate volatile acidity levels, indicating limited variability for most samples.
    - **Observations from citric acid vs frequency:**
        - Citric acid values are concentrated at lower levels, with many wines having values close to zero.
        - The distribution is right-skewed
    - **Observations from residual sugar vs frequency:**
        - Most wines have low residual sugar values ~2.5
        - The distribution is strongly right-skewed, with a small number of wines having high residual sugar levels.
    - **Observations from pH vs Frequency:**
        - pH values are clustered primarily between 3.0 and 3.5
        - The distribution is approximately symmetric, indicating limited variability in pH levels.
        - Extreme pH values are uncommon, suggesting consistent acidity levels across most wines.
    - **Observations from Sulphates vs Frequency:**
        - Sulphate levels are concentrated within 0.5 and 1.0
        - Distribution is right-skewed.
        - Very few wines exhibit high sulphate values



## Descriptive Question

***Question:*** What is the distribution of observed quality ratings for red wine in the dataset?
***Answer:*** The distribution of observed quality ratings for red wine shows that most wines in the dataset are rated at mid-range quality levels, with ratings of 5 and 6 occurring most frequently. Overall, the distribution is centered around average quality scores, indicating that the majority of red wines in the dataset are of moderate quality.

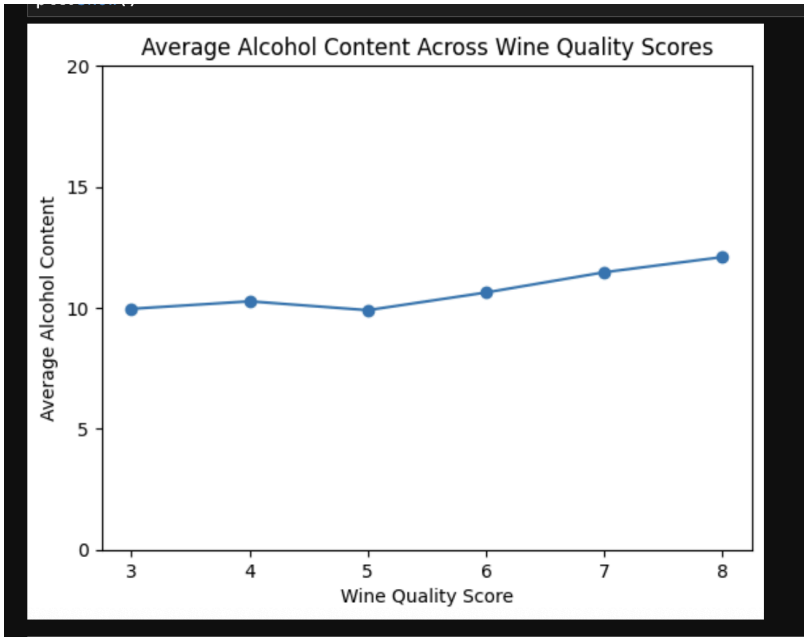
---

# EXPLORATORY ANALYSIS

In this analysis relationships between wine quality and selected physicochemical properties are examined. Visualizations are used to explore potential associations between alcohol content and wine quality ratings.

- **Average alcohol content  across wine quality scores**

The line plot suggests a positive association between alcohol content and wine quality ratings. As quality scores increase, average alcohol levels tend to increase, particularly for higher quality wines.

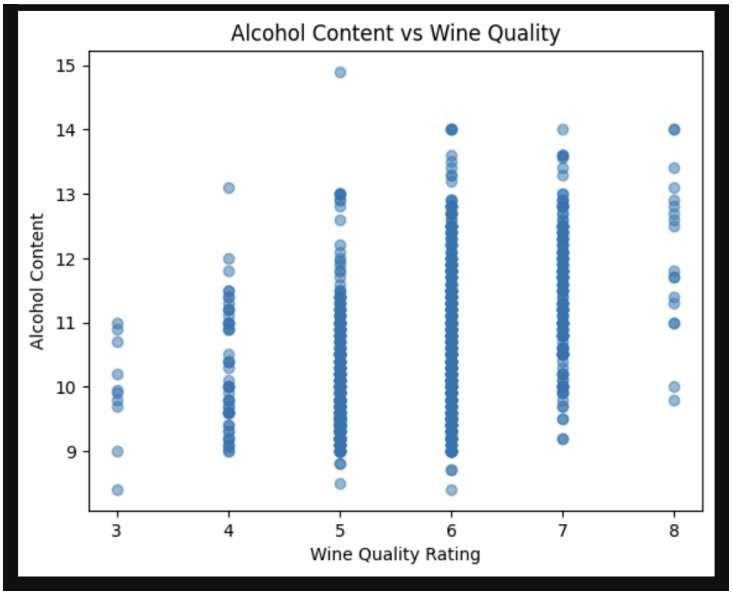



- **Scatter Plot: Alcohol content vs Quality score**

The scatter plot shows the relationship between alcohol content for each quality score. The following observations can be made:

  - The highest observed alcohol value occurs at a mid-range quality score (5) rather than the highest quality scores.

- Wines with higher quality ratings (7–8) generally have moderately high but not extreme alcohol values, i.e highest alcohol content it reaches is 14.
- This indicates that very high alcohol content alone does not guarantee higher wine quality, and other factors likely contribute.




- **Correlation Analysis**

  This analysis was done to explore which attributes are more strongly related to quality ratings.

  The following observations were made:
  - Alcohol content shows the strongest positive correlation with wine quality (≈ 0.48), suggesting that wines with higher alcohol levels tend to receive higher quality ratings.
  - Sulphates and citric acid also exhibit moderate positive correlations
  - Volatile acidity has a strong negative correlation with quality (≈ −0.39), suggesting that higher volatile acidity is generally associated with lower quality ratings.

- Residual sugar shows little to no correlation with wine quality, indicating minimal association in this dataset.

```
[36]: corr = df.corr()

      corr[['quality']].sort_values(by='quality', ascending=False)
```

[36]:

| | quality |
|---|---|
| quality | 1.000000 |
| alcohol | 0.476166 |
| sulphates | 0.251397 |
| citric acid | 0.226373 |
| fixed acidity | 0.124052 |
| residual sugar | 0.013732 |
| free sulfur dioxide | -0.050656 |
| pH | -0.057731 |
| chlorides | -0.128907 |
| density | -0.174919 |
| total sulfur dioxide | -0.185100 |
| volatile acidity | -0.390558 |

## Exploratory Question

***Question:*** how does alcohol content affect wine quality? Determine whether higher alcohol content is associated with higher quality ratings. Which variables show the strongest relationship with wine quality ratings?

***Answer:*** Exploratory analysis indicates that higher alcohol content is generally associated with higher wine quality ratings. Visualizations above show that wine with higher quality scores tend to have a high average alcohol content but there is a considerable overlap therefore alcohol alone does not determine quality.

Scatter plot further shows that the highest observed alcohol value (around 15%) occurs at a moderate quality rating (quality 5), while higher quality wines (ratings 7 and 8) generally exhibit high but slightly lower alcohol values, typically not exceeding about 14%. This suggests that extremely high alcohol content is not exclusive to the highest-quality wines.

Additionally, Correlation analysis supports these observations, showing that alcohol has the strongest positive correlation with wine quality, followed by sulphates and citric acid, while volatile acidity shows the strongest negative relationship with quality.

---

# PREDICTIVE ANALYSIS

In this analysis, a Linear regression model is used as the predictive model because wine quality is a numeric variable and the model provides an interpretable baseline for understanding how chemical attributes contribute to predicted quality scores.

- The dataset is split into training and testing sets to evaluate how well the model generalizes. After training, the model's intercept and coefficients are printed to understand how each feature contributes to predicted quality.
  Inference from model coefficients:
  - In this model, alcohol and sulphates have positive coefficients, indicating they increase predicted wine quality, while volatile acidity and pH have negative coefficients, indicating they reduce predicted quality.

```
88]: from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LinearRegression
     from sklearn.metrics import mean_squared_error, r2_score

     X = df[['alcohol', 'volatile acidity', 'fixed acidity', 'sulphates', 'pH']]
     y = df['quality']

     X_train, X_test, y_train, y_test = train_test_split(
         X, y, test_size=0.2, random_state=42
     )

     # train
     model = LinearRegression()
     model.fit(X_train, y_train)

     # Print model intercept
     print("Intercept:", model.intercept_)

     # Print model coefficients
     for feature, coef in zip(X.columns, model.coef_):
         print(f"{feature}: {coef}")

     Intercept: 2.6539471878754215
     alcohol: 0.3203472696273629
     volatile acidity: -1.0799479324136712
     fixed acidity: 0.02463603927930521
     sulphates: 0.5739210013748581
     pH: -0.1142391709417869
```

- The trained model is used to predict wine quality score for the test set. Model performance is evaluated using Mean Squared Error (MSE) and $R^2$. MSE summarizes the average squared prediction error, while $R^2$ indicates how much of the variation in wine quality is explained by the model.
  - MSE of ~0.399 indicates that the model's predictions are, on average, reasonably close to the true wine quality values, with some prediction error remaining.
  - $R^2$ of ~0.388 shows that approximately 38% of the variability in wine quality is explained by the chemical attributes included in the model.

Overall, the model performs moderately well in predicting wine quality scores.

```
[39]: y_pred = model.predict(X_test)

[40]: mse = mean_squared_error(y_test, y_pred)
      r2 = r2_score(y_test, y_pred)

      mse, r2

[40]: (0.39976092465919083, 0.3882825701952938)
```

- **Actual vs Predicted values**
  - An Actual vs Predicted table is created for the test set. This clearly shows how close the model's predicted quality scores are to the true ratings.

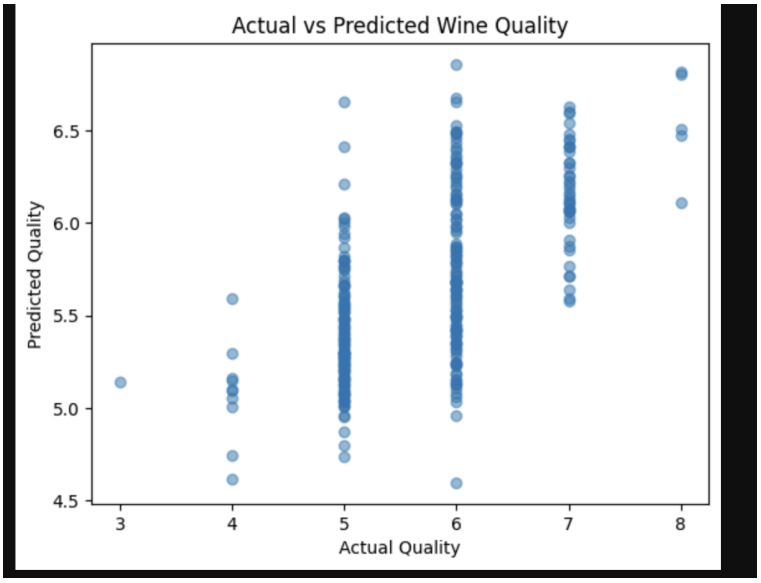
```
[46]: results_df = pd.DataFrame({
          'Actual Quality': y_test.values,
          'Predicted Quality': y_pred
      })

      results_df
```

[46]:

| | Actual Quality | Predicted Quality |
|---|---|---|
| 0 | 6 | 5.322861 |
| 1 | 5 | 5.237645 |
| 2 | 6 | 5.553164 |
| 3 | 5 | 5.435482 |
| 4 | 6 | 5.674995 |
| ... | ... | ... |
| 315 | 6 | 5.718817 |
| 316 | 5 | 5.072176 |
| 317 | 5 | 5.294186 |
| 318 | 6 | 6.203200 |
| 319 | 4 | 5.093414 |

320 rows × 2 columns

  - Scatterplot for actual vs predicted quality helps analyse how closely the model's predictions align with true values and shows patterns, errors, and variability in model performance.
    - The predicted values are clustered around mid-range quality scores, suggesting that the model performs best for average-quality wines. This aligns with the dataset distribution, which has a large number of moderate-quality wines compared to low- or high-quality wines.
    - The model tends to underestimate very high quality wines and overestimate very low quality wines.
    - Overall, the plot shows moderate predictive accuracy, consistent with the $R^2$ value obtained during model evaluation.




## Predictive Question

***Question:*** Can a predictive model accurately estimate wine quality based on measurable chemical attributes as alcohol content, acidity, sulphates, and pH?
***Answer:*** Yes a linear regression model was able to estimate wine quality with moderate accuracy using measurable chemical attributes such as alcohol content, acidity, sulphates, and pH. Although the model captures overall trends in wine quality and performs best for average-quality wines, it tends to underestimate very high-quality wines and overestimate very low-quality wines, reflecting limitations of a simple linear model.

---

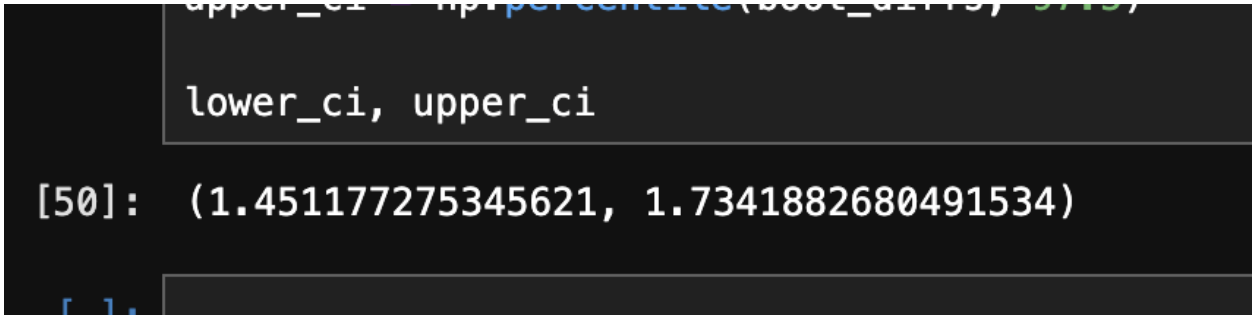
# INFERENTIAL ANALYSIS

After splitting the dataset into high(quality ≥ 7) and low quality(quality ≤ 5) wines, Bootstrap sampling was used to estimate the difference in mean alcohol content between the two.

- For bootstrap resampling with replacement and 10,000 bootstrap samples are used.
- Data split and initial mean difference (~1.59%)

```
[47]: low_quality = df[df['quality'] <= 5]['alcohol']
      high_quality = df[df['quality'] >= 7]['alcohol']

      observed_diff = high_quality.mean() - low_quality.mean()
      observed_diff

[47]: 1.591570660522276
```

- The 95% bootstrap confidence interval ranged from 1.45 to 1.73




## Inferential Question

***Question:*** Is there a meaningful difference in the average alcohol content between high-quality and low-quality red wines, and what is the 95% confidence interval for this difference using bootstrap sampling?
***Answer:*** Based on the analysis above, high-quality red wines have a higher average alcohol content than low-quality wines, with an observed mean difference of approximately 1.59%. The 95% bootstrap confidence interval for this difference ranges from 1.45 to 1.73. The small bootstrap confidence interval indicates low uncertainty in the estimate, meaning the observed difference in alcohol content between high- and low-quality wines is reliable given the data.

---

# PROBLEM REFORMULATION

Based on the analysis done, future analysis could do the following,

- Explore more complex predictive models to better capture non-linear relationships and improve performance at extreme quality levels.
- Improve dataset by to enhance model generalization,
    - adding more samples for very high quality and low quality wines,
    - adding more influential variables such as production time, production area, aged or not etc.

---

# VERSION CONTROL

GitHub was used as version control and the commit log is attached below,