# Task 3: Feature Engineering

Through the EDA performed on our dataset, we came to the conclusion that:

- Target Variable: Quality

- Features not affecting Quality: Residual Sugar, Density

- Residual sugar and density do not appear to have a significant observable correlation with quality, so we are removing them from the dataset to perform the next task.

- Volatile acidity, Chlorides and pH have a strong negative correlation with quality.

- Citric acid, Sulphates and Alcohol have a positive correlation with quality.

- Detailed explanations are given.

- The observed positive association between alcohol content and wine quality suggests a preference for higher alcohol levels among consumers, a trend visually evident from the graph where the median quality score for wines of good quality surpasses that of those deemed bad or average.
- The quality demonstrates a positive correlation with citric acid and fixed acidity, indicating that wines with elevated acidity tend to garner better quality ratings, likely due to the favored "fresh" taste attributed to higher citric acid levels. Additionally, wines with increased levels of sulfates exhibit a slight improvement in quality.
- While excessive volatile acidity is undesirable in wines, a moderate presence can be tolerated, as corroborated by the graph findings.
- A negative correlation exists between quality and pH levels, with lower pH levels associated with higher quality scores. Notably, wines of superior quality possess lower density, a characteristic consistent with higher alcohol content.
- Variables such as free sulfur dioxide, total sulfur dioxide, residual sugar, and chlorides do not directly influence wine quality in this dataset. Surprisingly, sweetness (residual sugar), a defining wine characteristic, does not appear to significantly impact quality within this dataset.
- The inverse relationship between acidity and pH levels is expected, given that pH is a measure of acidity. However, the positive correlation observed between pH and volatile acidity appears counterintuitive and could potentially be attributed to a lurking variable not accounted for in the analysis.
- The negative correlation between density and alcohol content can be rationalized by the fact that wines with higher alcohol content tend to have lower density.
- The prevalence of citric acid as a primary constituent within fixed acidity elucidates the robust positive correlation observed between these two variables.
- The positive correlation between free sulfur dioxide and total sulfur dioxide can be explained by the latter being a composite measure encompassing both free and bound forms of sulfate.

Original Dataset with all variables:

| fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulfates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **7.4** | 0.7 | 0 | 1.9 | 0.076 | 11 | 34 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |
| **7.8** | 0.88 | 0 | 2.6 | 0.098 | 25 | 67 | 0.9968 | 3.2 | 0.68 | 9.8 | 5 |
| **7.8** | 0.76 | 0.04 | 2.3 | 0.092 | 15 | 54 | 0.997 | 3.26 | 0.65 | 9.8 | 5 |
| **11.2** | 0.28 | 0.56 | 1.9 | 0.075 | 17 | 60 | 0.998 | 3.16 | 0.58 | 9.8 | 6 |
| **7.4** | 0.7 | 0 | 1.9 | 0.076 | 11 | 34 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |

Variables with no meaningful trends, removed from original dataset:

| Residual sugar | density |
|---|---|
| 1.9 | 0.9978 |
| 2.6 | 0.9968 |
| 2.3 | 0.997 |
| 1.9 | 0.998 |
| 1.9 | 0.9978 |

New Dataset with all the remaining variables:

| fixed acidity | volatile acidity | citric acid | chlorides | free sulfur dioxide | total sulfur dioxide | pH | sulfates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|
| **7.4** | 0.7 | 0 | 0.076 | 11 | 34 | 3.51 | 0.56 | 9.4 | 5 |
| **7.8** | 0.88 | 0 | 0.098 | 25 | 67 | 3.2 | 0.68 | 9.8 | 5 |
| **7.8** | 0.76 | 0.04 | 0.092 | 15 | 54 | 3.26 | 0.65 | 9.8 | 5 |
| **11.2** | 0.28 | 0.56 | 0.075 | 17 | 60 | 3.16 | 0.58 | 9.8 | 6 |
| **7.4** | 0.7 | 0 | 0.076 | 11 | 34 | 3.51 | 0.56 | 9.4 | 5 |