EDA Summary: Red Wine Quality Dataset

# Dataset Overview

- The dataset contains 11 physicochemical features and 1 target variable (`quality`) for red wines.  
- Total number of entries: 1,599

# Statistical Summary & Data Info

- No missing values were found.  
- Features like `fixed acidity`, `citric acid`, `alcohol`, and `density` are numeric and continuous.  
- `.describe()` showed significant variance in `residual sugar` and `alcohol`.

# Univariate Analysis

- Histograms revealed:  
 - `alcohol` is right-skewed, mostly around 9–10%.  
 - `volatile acidity` is slightly left-skewed.  
 - `quality` is mostly between 5 and 6.  
- Boxplots highlighted:  
 - Outliers present in `residual sugar`, `chlorides`, and `free sulfur dioxide`.

# Bivariate Analysis

- Heatmap showed:  
 - `alcohol` has the strongest positive correlation with quality.  
 - `volatile acidity` has a moderate negative correlation with quality.  
 - `density` and `pH` are negatively correlated.  
- Pairplots and scatter plots confirmed:  
 - Higher `alcohol` content generally indicates better quality.  
 - High `volatile acidity` tends to reduce quality.

# Categorical Analysis

- `quality` value counts:  
 - Most wines are rated 5 or 6.  
 - Very few wines are rated above 7, indicating class imbalance.

# Key Insights

- `Alcohol` is the most influential feature on wine quality.  
- Wines with lower `volatile acidity` have higher quality.  
- Features like `residual sugar` and `chlorides` contain outliers and may need preprocessing.  
- Slight class imbalance in wine quality ratings.

By –

Pulkit Upreti