Week 1 Delivery Summary: White Wine Quality Dataset

1. Data Loading and Cleaning

- The dataset contains physicochemical features of white wines along with a quality score (ranging from 0 to 10).
- Data was loaded using pandas and explored using `.head()`, `.info()`, `.describe()`, and `.isnull().sum()`.
- No missing values were found in the dataset.
- Feature normalization was applied using `StandardScaler` to standardize the range of all numerical features.
- Normalization is especially important when working with machine learning algorithms that are sensitive to scale.

2. Exploratory Data Analysis (EDA)

- The target variable `quality` is categorical (integer-valued) but can be used in both regression and classification settings.
- Most wines have a quality rating between 5 and 7, with fewer samples at the extremes (3 or 9), indicating class imbalance.
- Histograms were plotted to observe the distribution of each numerical feature. Some features, like `residual sugar` and `free sulfur dioxide`, are right-skewed.
- Box plots showed how the distribution of each feature varies with wine quality. For example, higher alcohol content is often associated with better wine quality.
- Value counts and unique values were explored to understand the range and balance of the target variable.
- Outliers were observed in several features (e.g., sulfur dioxide), which may influence model performance and might require further handling.

3. Correlation Analysis

- Pearson correlation coefficients were computed between all pairs of features.
- The correlation matrix revealed that `alcohol` has the highest positive correlation with wine quality.

- Some features were moderately correlated with each other, such as `density` and `residual sugar`, which can indicate multicollinearity.
- Features with an absolute correlation greater than 0.1 with the target were considered for selection.

4. Feature Selection

- Variance Threshold was used to remove features with very low variance (below 0.01), as they add little to model performance.
- Mutual Information was calculated to measure the dependency between each feature and the target. The top 5 features based on mutual information were selected.
- Forward Feature Selection was applied using a Random Forest Classifier to iteratively add the most predictive features.
- Backward Feature Elimination was used to remove the least predictive features step-by-step until the optimal subset was achieved.
- All these techniques were used to compare different feature sets for modeling.

5. Visualizations and Patterns

- A countplot was used to display the frequency distribution of wine quality scores.
- Box plots were generated for each feature against the target to examine variability and detect patterns.
- Histograms helped visualize skewness and the range of numerical values.
- A heatmap of the correlation matrix helped identify strong linear relationships between features and potential multicollinearity.

6. Key Insights

- Higher alcohol content tends to be associated with higher quality wines.
- Density and residual sugar are negatively correlated with wine quality.
- Class imbalance in the target variable may require handling via resampling or weighted models in future modeling stages.
- Feature selection methods consistently highlighted `alcohol`, `density`, `volatile acidity`, and `sulphates` as important predictors of wine quality.