**1.Introduction**

The next step of my Advanced Data Mining project is to work on Deliverable 1: Data Collection, Cleaning, and Exploration. In this step, I chose the UCI Wine Quality data set that used 4,898 samples of red-wine labeled with 12 physicochemical features and one quality rating provided by experts. My intention is to come up with a sound basis of modeling undertaken through preparing and comprehending data. To begin with, I imported the CSV to a Pandas DataFrame, checked its structure, making sure that all columns were parsed correctly. I then identified and filled in missing values, eliminated 240 literal duplicates and standardized column names to snake\_case. And then I performed some exploratory visualizations histograms, KDE plots, heatmaps, as well as boxplots, to discover distributions forms, outliers, and the correlations with the target. The above steps enable my data to be clean and consistent after understanding my data in a better manner before proceeding to predictive modeling.

**2. Choice of Data Set and Explanation**

I selected the Wine Quality dataset of the UCI Machine Learning Repository. This data is characterized by 4898 wines of red type with twelve physicochemical variables or characteristics - such as acidity measures, content of sugar, pH, sulphates, alcohol percentage - and a quality score assigned by experts (scale of 0-10).

Size and Scope: It has had about 5,000 records and some 12 features, which exceed the project minimum of 500 records and 8 attributes.

Relevance: It is an up to date regression problem, realizable in the context of food science and quality control as a need to predict wine quality.

Complexity: The combination of continuous variables, opportunity of outliers and imbalanced target distribution provide a good chance to do cleaning, feature transformation and robust exploration.

**3. Data Loading/Checking First Look**

I am reading the CSV into a Pandas DataFrame and checking its shape and format at once:

df = pd.read\_csv('winequality-red.csv'; sep=';')

print(df.shape) # features 4898, 12

df.head()

I checked the accuracy of parsing and found that no text fields exist, and the quality score is recorded as an integer value.

**4. Data Cleaning**

4.1 Management of Missing Values

I used df.isna().sum() first to see whether there were any null values anywhere in the twelve columns and there were no missing values.

4.2 De-Duplication

I found out that there were 240 duplicate rows that were exact duplicates. In order not to skew my analysis, I dropped every duplicate:

df = df.drop\_duplicates()

New shape = (4658, 12)

4.3 Normalization of Columns Names

I changed all column names to snake\_case so that it was more easily referenced:

df.columns = [c.strip().lower().replace(' ','\_') for c in df.columns]

4.4 Type of Data

I ensured that all the features were in their similar form as numeric dtypes and no additional conversion was to be made.

**5. Exploratory Data Analysis (EDA)**

5.1 Univariate Distributions

I created the histograms with kernel density estimates of some main variables, alcohol, volatile\\_acidity, residual\\_sugar, and citric\\_acid, to check the skew and identify outliers.

Observation: Alcohol volume is approximating Gaussian distribution with positive slight skew; residual sugar and volatile acidity are having longer flanks.

5.2 Correlation analysis

I calculated Pearson correlation matrix among all numeric features and plotted it in the form of a heatmap.

Key findings:

Alcohol ↔ Quality: 0.48 (strong positive)

Volatile acidity Quality: -0.39 (strong negative)

Citric acid 23 [reamos+0.23]

5.3 Feature to Target Relationships

I checked that median percent alcohol is a steadily rising function of quality score and volatile acidity is a steadily decreasing function of quality score using boxplots of each against the quality score.

**6. Insights & Applications to Modeling**

Predictor Prioritization:

When it comes to regression, I will give more emphasis on the alcohol and volatile\_acidity variables since they have the best linear relationship with quality.

Outlier Treatment:

Extreme values (e.g. alcohol > 14% or volatile acidity > 1.2) could be enough to confound model estimates. I will either stem these outliers or use robust regression methods.

Skewed Features:

Other variables such as residual sugar and sulphates are skewed. Under modeling I will take log-transformations or Box-Cox transforms.

Target Imbalance:

Most of the wines lie at the range of quality of 5-6. In future, when I require to perform classification tasks (e.g. “good” vs. “bad” wine), a stratified sampling or class-weighting to alleviate imbalance will be applied.

These results precondition Deliverable 2, where I will construct and test regression models, use feature engineering, and optimize my preprocessing, using this EDA results.

**7. Advantages/Disadvantages & Solutions**

Duplicate Rows: I have deleted 240 duplicates so that my effective sample size has observations that are really independent.

Imputation Demo: I wrote and used median-imputation code, even though this dataset did not contain any missing values, so that best practices would be present in case it is used in the future.

**8. Conclusion**

At the time of submitting Deliverable 1, I was able to make a good clean analysis-ready dataset out of ugly wine-quality data and derive actionable insights I can use to shape my further course of action. The EDA that I did confirmed that alcohol content and volatile acidity are those parameters that are strongly correlated with quality and as such, other parameters like residual sugar and sulphates exhibited skewness which could be transformed. I observed the presence of a few extreme outliers to deal with and the imbalance in the target score of 5 6, which will be used in my sampling and appraisal plans. These results will have a direct effect on the regression and classification pipeline created by me in Deliverable 2, such as feature engineering, outliers, and cross-validation planning. Given a strong, clear dataset at my disposal, I am ready to create and verify predictive models that reflect some of the most prevailing factors that influence wine quality.

Github\_Link:

<https://github.com/sromika27019/MSCS_634_AdvancedDataMining_Deliverable1>

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A white rectangular object with black text

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.