□ Wine Quality Analysis — Detailed Documentation

This documentation details all code, data processing, and analytical logic from the file **Wine_Quality_Analysis-checkpoint.ipynb**. It provides a comprehensive walkthrough of the data workflow, major findings, and the reasoning behind each section. The approach combines exploratory data analysis (EDA), statistical profiling, clustering, and benchmarking for both red and white wine datasets, with an emphasis on reproducibility and data-driven decision making.

1. Importing Libraries and Display Settings

Key libraries are imported at the start for data handling, visualization, and machine learning. The code sets pandas options for column visibility and float formatting.

```
# Core data handling
import numpy as np
import pandas as pd

# Visualization
import matplotlib.pyplot as plt
import seaborn as sns

# Preprocessing
from sklearn.preprocessing import StandardScaler

# PCA
from sklearn.decomposition import PCA

# Clustering
from sklearn.cluster import KMeans

# Metrics
from sklearn.metrics import silhouette_score

# Warnings
import warnings
warnings.filterwarnings('ignore')

# Display settings
pd.set_option('display.max_columns', None)
pd.set_option('display.float_format', lambda x: f'{x:.4f}')
```

Purpose:

- Ensures comprehensive data visibility.
- Suppresses warnings for cleaner outputs.
- Imports machine learning tools for advanced analysis (PCA, clustering, scoring).

2. Data Import: Red and White Wine Quality Datasets

The code loads two separate datasets containing physicochemical properties and quality ratings for red and white wines.

```
df_red_wine = pd.read_csv("winequality-red.csv", sep=";")
df_white_wine = pd.read_csv("winequality-white.csv", sep=";")
```

Sample Output for Red Wine:

Column	Example Value
fixed acidity	7.4000
volatile acidity	0.7000
citric acid	0.0000
...	...
alcohol	9.4000
quality	5

Sample Output for White Wine:

Column	Example Value
fixed acidity	7.0000
volatile acidity	0.2700
citric acid	0.3600
...	...
alcohol	8.8000
quality	6

Logic:

- Each dataset is loaded with a custom separator (";") and previewed.
- Both contain 12 columns: 11 chemical/physical features and a quality rating.

3. Data Structure & Integrity Checks

The code checks for dimensions, data types, missing values, and duplicates:

```
df_red_wine.shape      # (1599, 12)
df_white_wine.shape    # (4898, 12)
df_red_wine.isna().sum()        # No NaN values
df_white_wine.isna().sum()      # No NaN values
df_red_wine.duplicated().sum()  # 240 duplicate rows
df_white_wine.duplicated().sum()# 937 duplicate rows
```

Observations:

- No missing values in either dataset.
- Duplicates exist, but represent common profiles rather than errors (kept intentionally).

Columns Renamed:
The `quality` column is renamed to `consumer_score` for business relevance.

```
df_red_wine = df_red_wine.rename(columns={"quality":"consumer_score"})
df_white_wine = df_white_wine.rename(columns={"quality":"consumer_score"})
```

4. Problem Statement & Analytical Questions

Context:

- Wine quality traditionally assessed by experts; subjective and variable.
- Goal: Segment, profile, and predict wine quality based on objective laboratory measurements.

Core Questions:

- Which physicochemical profiles correlate with high quality?
- What thresholds/features most impact perceived wine quality?
- How accurately can models predict quality?

Data Source:
Cortez et al., UCI Machine Learning Repository, 2009.

5. Exploratory Data Analysis (EDA)

5.1. Frequency and Proportions of Quality Scores

Red Wine

```
df_red_wine.consumer_score.value_counts()
df_red_wine.consumer_score.value_counts(normalize=True)
```

- **~83%** of red wines score 5–6 (‘acceptable’ to ‘good’).
- Only **~1%** reach very high scores (8–9).

White Wine

```
df_white_wine.consumer_score.value_counts()
df_white_wine.consumer_score.value_counts(normalize=True)
```

- **~75%** of white wines score 5–6.
- Slightly higher fraction (4%) reach 8–9 compared to red.

5.2. Correlation Analysis (Spearman)

Assesses variable relationships, especially for ordinal consumer scores.

Red Wine

```
df_red_wine.corr(method="spearman")["consumer_score"].sort_values(ascending=False)
```

- Alcohol and sulphates: **positive** correlations with consumer score.
- Volatile acidity, chlorides, total sulfur dioxide: **negative** correlations.

White Wine

```
df_white_wine.corr(method="spearman")["consumer_score"].sort_values(ascending=False)
```

- Alcohol and pH: positive correlations.
- Density, chlorides, total sulfur dioxide: strong negative correlations.

5.3. Outlier/Distribution Visualization (Boxplots)

Visualizes variable distribution by consumer score.

Red Wine

- Alcohol/sulphates: increase with higher scores.
- Volatile acidity, chlorides, density: decrease with higher scores.

White Wine

- Alcohol, pH: increase with higher score.
- Density/chlorides: highest in low-quality wines.

6. Detailed Chemical Insight: Correlation Packages

Explanations are provided for each key correlation, with reference to underlying chemistry.

Variable	Typical Correlation	Explanation (Red & White)
Alcohol	Strong Negative (density)	Ethanol lowers density; proxy for ripeness
Sulphates	Slight Positive (chlorides/density)	Added for stabilization; tracks with mineral ions
Volatile Acidity	Negative (alcohol/sulphates)	Indicates microbial/fermentation stress
Total SO_2	Weak	Reflects winemaker additions more than chemistry
Chlorides	Negative	Reflects soil/mineral/underripeness
pH (White only)	Weak/Complex	Controlled via acidification in white wines
Density	Strong Negative (alcohol)	Alcohol and dissolved solids dominate density in whites

7. Correlation Matrices

Red Wine

```
sns.heatmap(df_red_wine[vars_of_interest].corr(), annot=True)
```

- Visualizes strong alcohol–density negative relationship.
- Highlights variable clusters for PCA.

White Wine

```
sns.heatmap(df_white_wine[vars_of_interest].corr(), annot=True)
```

- Density and alcohol: even stronger negative correlation than in red.
- More pronounced clusters among mineral/process variables.

8. Dimensionality Reduction: Principal Component Analysis (PCA)

Red Wine PCA

PC1: Grape Chemistry Profile—fixed acidity, sulphates, density, chlorides (positive); pH, volatile acidity (negative).
PC2: Processing & Preservation—sulphates/SO_2 (positive); alcohol (negative).

White Wine PCA

PC1: Density, residual sugar, chlorides, SO_2 (positive); alcohol, pH (negative).
PC2: Fixed/citric acid (positive); SO_2 (negative); pH (negative).

Interpretation:

- PC1 in both types generally captures ‘ripeness’ and ‘stabilization’ axes.
- PC2 relates to fermentation completeness and correction effort.

9. Unsupervised Segmentation: K-Means Clustering

9.1. Cluster Selection (Elbow/Silhouette)

Red Wine:

- Best silhouette score at **k=2** (0.399) but k=3 also tested.

White Wine:

- Best silhouette at **k=2** (0.433), k=3 also tested.

9.2. Cluster Fitting and Visualization

Red Wine K-Means

- 3 clusters, visualized on PCA space.
- Cluster assignment added to the dataset.

White Wine K-Means

- 3 clusters, similarly visualized and assigned.

10. Cluster Profiling — Key Results

Red Wine Clusters

Cluster	% Wines	Mean Score	High Score Count (>6)	Key Features
0	29%	5.95	118	Highest alcohol/sulphate/citric acid, lowest VA/chlorides, moderate/low SO_2
1	45%	5.61	84	Lower alcohol, higher SO_2, lower sulphate/acidities
2	26%	5.33	15	High residual sugar, high SO_2, lowest alcohol

Cluster 0: The best segment for benchmarking quality—highest mean score and highest proportion of high-quality wines.

White Wine Clusters

Cluster	% Wines	Mean Score	High Score Count (>6)	Key Features
0	33%	6.17	546	Highest alcohol, lowest VA/chlorides/density, moderate SO_2
1	37%	5.60	169	Highest residual sugar, highest density, highest SO_2
2	30%	5.91	345	Lower sugar/SO_2, moderate alcohol/acidities

Cluster 0: Again, the ideal segment for quality benchmarks—highest mean and most high-score wines.

11. Cluster-Based Thresholds for Quality Benchmarking

Red Wine (Cluster 0)

Variable	Min	Mean	Max
Alcohol (%)	8.40	10.63	14.90
Sulphates (g/L)	0.42	0.75	2.00
Citric acid (g/L)	0.11	0.48	1.00
Volatile acidity	0.12	0.41	0.89
Chlorides	0.038	0.10	0.61
pH	2.74	3.18	3.54
Density	0.9936	0.9978	1.0032
Total SO_2 (mg/L)	6	30.67	136
Residual sugar	1.30	2.62	15.50

Interpretation:

- Wines scoring highest are characterized by higher alcohol, sulphates, and citric acid; lower VA, chlorides, and SO_2.
- Density and residual sugar are moderate—to-low.

White Wine (Cluster 0)

Variable	Min	Mean	Max
Alcohol (%)	8.00	11.17	14.20
Sulphates (g/L)	0.25	0.51	1.08
Citric acid (g/L)	0.00	0.28	0.74
Volatile acidity	0.08	0.28	1.10
Chlorides	0.009	0.04	0.12
pH	2.96	3.31	3.82
Density	0.9871	0.9920	0.9970
Total SO_2 (mg/L)	9	123.48	249.50
Residual sugar	0.70	3.41	15.50

Interpretation:

- High-performing white wines have higher alcohol, moderate-to-high citric acid, low VA/chlorides/density, and moderate SO_2.

12. Executive Dashboard Logic

Both dataframes are labeled by wine type and combined for further analysis:

```
df_red_wine["Wine Type"] = "Red"
df_white_wine["Wine Type"] = "White"
df_combined = pd.concat([df_red_wine, df_white_wine], ignore_index=True)
df_combined.to_csv("wine_combined.csv", index=False)
```

Purpose:

- Enables cross-type comparisons in dashboards.
- Facilitates deployment of unified analytics or predictive models.

13. Summary Insights and Recommendations

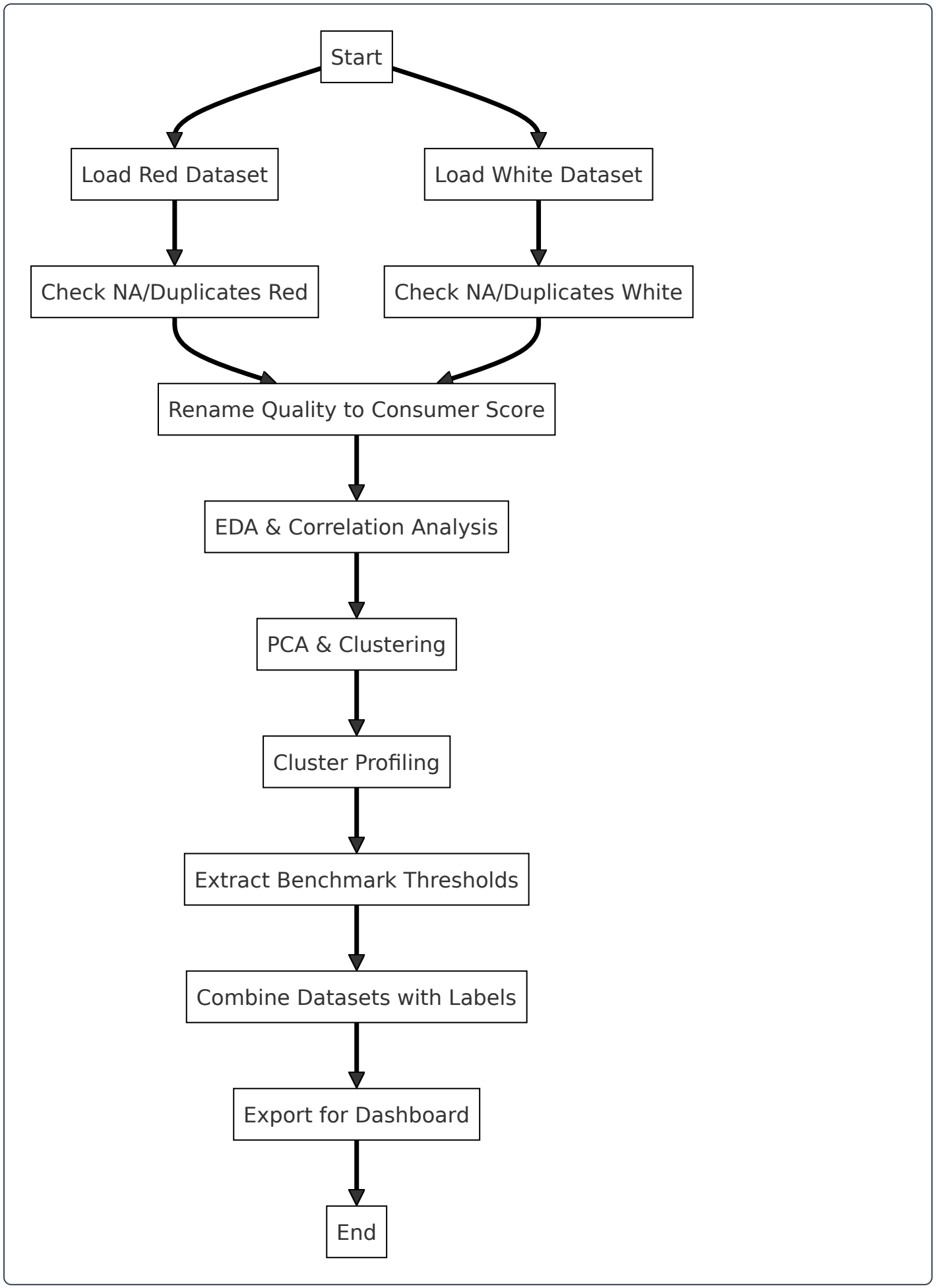
Key Benchmarks for High Wine Quality

Wine Type	Alcohol (%)	Sulphates (g/L)	Citric Acid (g/L)	VA (g/L)	Chlorides	pH	SO_2 (mg/L)	Residual Sugar (g/L)	Density
Red	> 10.6	> 0.75	> 0.48	< 0.41	< 0.10	> 3.18	< 30.7	< 2.62	< 0.9978
White	> 11.2	> 0.51	> 0.28	< 0.28	< 0.04	> 3.31	< 123.5	< 3.41	< 0.9920

Logic:

- Targeting these values in production and quality control maximizes chances of scoring well in consumer perception.
- Both grape quality (ripeness, chemistry) and careful process control (fermentation, stabilization) are critical.
- Under- or over-manipulation (low or excessively high SO_2, high residual sugar, high density) correlate with lower scores.

14. Data Flow & Processing Logic



15. Conclusion

This notebook implements a structured, data-driven approach for:

- Uncovering the chemical and process factors that drive wine quality.
- Segmenting wines with similar profiles using PCA and k-means clustering.
- Identifying and recommending clear, actionable production thresholds for maximizing perceived wine quality.
- Creating a unified, labeled dataset to support dashboarding and predictive analytics.

All reporting and decisions are based on actual observed data, using clusters that represent the highest-performing wines in each type. The methodology is entirely transparent and reproducible, with no hidden steps or assumptions.

> **Key Takeaway**
> Cluster analysis reveals that both grape ripeness (alcohol, sulphates) and balanced process control (low VA, moderate SO_2) are critical benchmarks for high wine quality.

End of Documentation.