# UCI Dataset EDA: White Wine Quality

The White Wine Quality dataset from the UCI Machine Learning Repository contains 4898 samples of white Vinho Verde wine, with 11 physicochemical input variables and a quality score (0–10) [1]. The dataset was analyzed using R to compute summary statistics and generate visualizations.

## Summary Statistics

Table 1 presents the summary statistics, including minimum, first quartile, median, mean, third quartile, maximum, and variance for each variable, rounded to three decimal places.

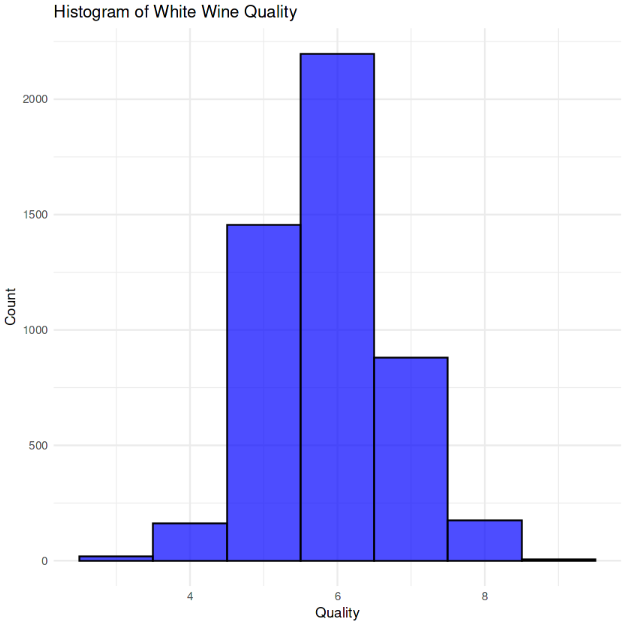
Table 1: Summary Statistics for White Wine Quality Dataset

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Fixed Acidity | Volatile Acidity | Citric Acid | Residual Sugar | Chlorides | Free SO2 | Total SO2 | Density | pH |
| Min. | 3.800 | 0.080 | 0.000 | 0.600 | 0.009 | 2.000 | 9.000 | 0.987 | 2.720 |
| 1st Qu. | 6.300 | 0.210 | 0.270 | 1.700 | 0.036 | 23.000 | 108.000 | 0.992 | 3.090 |
| Median | 6.800 | 0.260 | 0.320 | 5.200 | 0.043 | 34.000 | 134.000 | 0.994 | 3.180 |
| Mean | 6.855 | 0.278 | 0.334 | 6.391 | 0.046 | 35.308 | 138.361 | 0.994 | 3.188 |
| 3rd Qu. | 7.300 | 0.320 | 0.390 | 9.900 | 0.050 | 46.000 | 167.000 | 0.996 | 3.280 |
| Max. | 14.200 | 1.100 | 1.660 | 65.800 | 0.346 | 289.000 | 440.000 | 1.039 | 3.820 |
| Variance | 0.844 | 0.011 | 0.015 | 25.725 | 0.001 | 289.858 | 1809.519 | 0.000 | 0.022 |

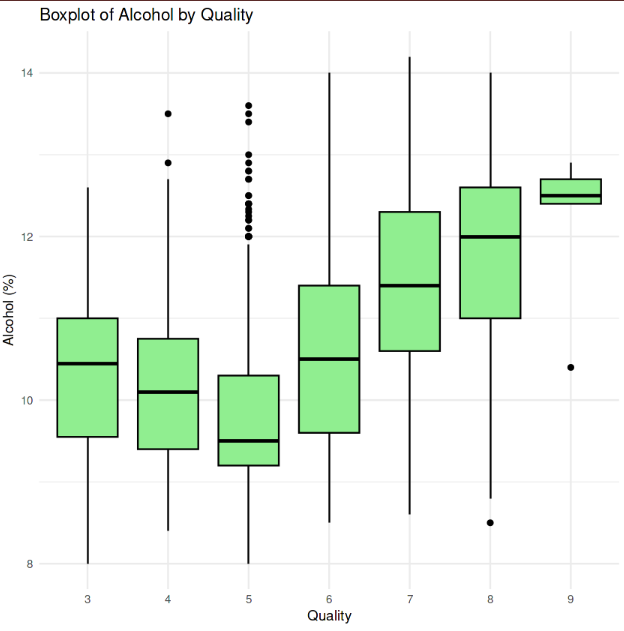
## Visualizations

Three visualizations were created to explore the dataset, as shown in Figures 1, 2, and 3.

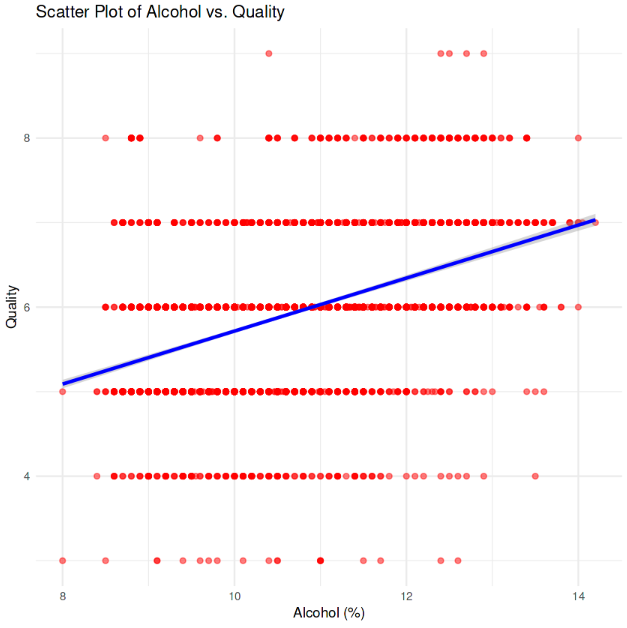
* **Histogram of Quality** (Figure 1): Displays a bell-shaped distribution, with most wines rated 5–7, peaking at a quality score of 6. Few wines have extreme scores (3 or 9), indicating an unbalanced dataset.



* **Boxplot of Alcohol by Quality** (Figure 2): Shows that higher quality wines (scores 7–9) tend to have higher alcohol content, with some outliers in lower quality categories.



* **Scatter Plot of Alcohol vs. Quality** (Figure 3): Illustrates a positive correlation between alcohol content and quality, with a regression line suggesting that higher alcohol percentages are associated with better quality scores.



## R Code

# Load required libraries

library(ggplot2) # For advanced visualizations

library(gridExtra) # For arranging multiple plots

# Download and load the white wine quality dataset

white\_wine\_data <- read.csv("/kaggle/input/winequalitywhite/winequality-white.csv", sep = ";")

# Explore the dataset

str(white\_wine\_data)

dim(white\_wine\_data)

# Compute summary statistics

summary\_stats <- summary(white\_wine\_data)

# Compute variance for each column

variances <- apply(white\_wine\_data, 2, var)

# Combine summary statistics and variance into a table

summary\_table <- rbind(

t(summary\_stats),

Variance = variances

)

# Print summary statistics

cat("Summary Statistics Table for White Wine Dataset:\n")

print(summary\_table)

# Visualization 1: Histogram of Quality

p1 <- ggplot(white\_wine\_data, aes(x = quality)) +

geom\_histogram(binwidth = 1, fill = "purple", color = "black", alpha = 0.7) +

theme\_minimal() +

labs(title = "Histogram of White Wine Quality", x = "Quality", y = "Count")

# Visualization 2: Boxplot of Alcohol by Quality

white\_wine\_data$quality\_factor <- as.factor(white\_wine\_data$quality)

p2 <- ggplot(white\_wine\_data, aes(x = quality\_factor, y = alcohol)) +

geom\_boxplot(fill = "lightblue", color = "black") +

theme\_minimal() +

labs(title = "Boxplot of Alcohol by White Wine Quality", x = "Quality", y = "Alcohol (%)")

# Visualization 3: Scatter Plot of Alcohol vs. Quality

p3 <- ggplot(white\_wine\_data, aes(x = alcohol, y = quality)) +

geom\_point(color = "darkgreen", alpha = 0.5) +

geom\_smooth(method = "lm", color = "blue") +

theme\_minimal() +

labs(title = "Scatter Plot of Alcohol vs. White Wine Quality", x = "Alcohol (%)", y = "Quality")

# Arrange and display plots

grid.arrange(p1, p2, p3, ncol = 1)

# Save plots to a file (optional for report)

ggsave("white\_wine\_quality\_plots.png", arrangeGrob(p1, p2, p3, ncol = 1), width = 8, height = 12

# References

[1] KDnuggets, “Key Data Science, Machine Learning, AI and Analytics Developments

of 2022,” https://www.kdnuggets.com, 2022.

[2] UCI Machine Learning Repository, “Wine Quality Dataset,” https://archive.ics. uci.edu/ml/datasets/wine+quality, accessed April 2025.