

```

# PART 1 - DATA LOADING AND PREPARATION
# coder: Jatesh Joshi
# Project: Wine Quality Analysis
#####
## 1. Load necessary library
library(dplyr) # loads a data package and gives us functions

# now, we can load our datasets
red <- read.csv("winequality-red.csv", sep = ";") # sep here is a
separation operator
white <- read.csv("winequality-white.csv", sep = ";")

# add a new variable for each wine type
red$type <- "red"
white$type <- "white"
# we have to bind or combine our two wine datasets now
wine <- rbind(red, white) # row bind
# we know r treats numbers different from categories and type and quality
are not numerical
wine$type <- as.factor(wine$type)
wine$quality <- as.factor(wine$quality) # now, we are safe when we create
models and plots
colSums(is.na(wine)) # check missing

write.csv(wine, "wine_cleaned.csv", row.names = FALSE)
# this is now our cleaned dataset that we can use for our project.
#####
#####

# now we have to get our basic descriptive statistics
# let us list out all of our continuous variables:
continuous_vars <- c("fixed.acidity", "volatile.acidity", "citric.acid",
                      "residual.sugar", "chlorides",
"free.sulfur.dioxide",
                      "total.sulfur.dioxide", "density", "pH",
                      "sulphates", "alcohol")
# mean
mean_values <- sapply(wine[, continuous_vars], mean)
mean_values
#median now
median_values <- sapply(wine[, continuous_vars], median)
median_values
# Standard Deviation
sd_values <- sapply(wine[, continuous_vars], sd)
sd_values
#####
#####
#meaningful plots
#1. Histogram of alcohol content to show us the distribution and shape.
hist(wine$alcohol,
      main = "Distribution of Alcohol Content",
      xlab = "Alcohol (%)",
      col = "lightblue",

```

```

    border = "black")
# Boxplot of alcohol content by quality level to show us how alcohol
varies with wine wuality
boxplot(alcohol ~ quality, data = wine,
        main = "Alcohol Content by Quality Rating",
        xlab = "Wine Quality",
        ylab = "Alcohol (%)",
        col = "lightgreen")

# sactterplot to see if sweeter wines have different amount of alcohols
plot(wine$residual.sugar, wine$alcohol,
      main = "Alcohol vs Residual Sugar",
      xlab = "Residual Sugar (g/L)",
      ylab = "Alcohol (%)",
      col = "purple",
      pch = 16)
# BarPlot to show which quality rating is more common
quality_counts <- table(wine$quality)
barplot(quality_counts,
        main = "Frequency of Wine Quality Ratings",
        xlab = "Quality Rating",
        ylab = "Count",
        col = "orange")
# correlation heatmap to show stength of relationships between all
contiuos variables
library(corrplot)

# Compute correlation matrix for continuous variables
cor_matrix <- cor(wine[, continuous_vars])

# Plot heatmap
corrplot(cor_matrix, method = "color")

```