CptS 575 Data Science
Washington State University

Assignment - 2

Contents

1 Exercise 1

1.1 1.a.

Read the data

```
# reading the data
red_wine_data = read.csv("winequality-red.csv")
```

The wine quality data inputted into R

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
1	7.4	0.700	0.00	1.90	0.076	11	34	0.9978	3.51	0.56	9.4	5
2	7.8	0.880	0.00	2.60	0.098	25	67	0.9968	3.20	0.68	9.8	5
3	7.8	0.760	0.04	2.30	0.092	15	54	0.9970	3.26	0.65	9.8	5
4	11.2	0.280	0.56	1.90	0.075	17	60	0.9980	3.16	0.58	9.8	6
5	7.4	0.700	0.00	1.90	0.076	11	34	0.9978	3.51	0.56	9.4	5
6	7.4	0.660	0.00	1.80	0.075	13	40	0.9978	3.51	0.56	9.4	5
7	7.9	0.600	0.06	1.60	0.069	15	59	0.9964	3.30	0.46	9.4	5
8	7.3	0.650	0.00	1.20	0.065	15	21	0.9946	3.39	0.47	10.0	7
9	7.8	0.580	0.02	2.00	0.073	9	18	0.9968	3.36	0.57	9.5	7
10	7.5	0.500	0.36	6.10	0.071	17	102	0.9978	3.35	0.80	10.5	5

Figure 1: Red Wine Data

1.2 1.b.

Finding median of wine samples and mean alcohol level

```
# finding the median quality and mean alcohol level
median_quality <- median(red_wine_data$quality)
mean_alcohol_level <- mean(red_wine_data$alcohol)
```

values	
mean_alcohol_level	10.4229831144465
median_quality	6L

Figure 2: Median quality and Mean alcohol level

1.3 1.c.

Producing a scatter plot between wine density and volatile acidity

```
# generating scatter plot
plot(
  red_wine_data$density,
  red_wine_data$volatile_acidity,
  xlab = "wine density",
  ylab = "volatile acidity",
  main = "Wine Density vs Volatile Acidity",
)
```

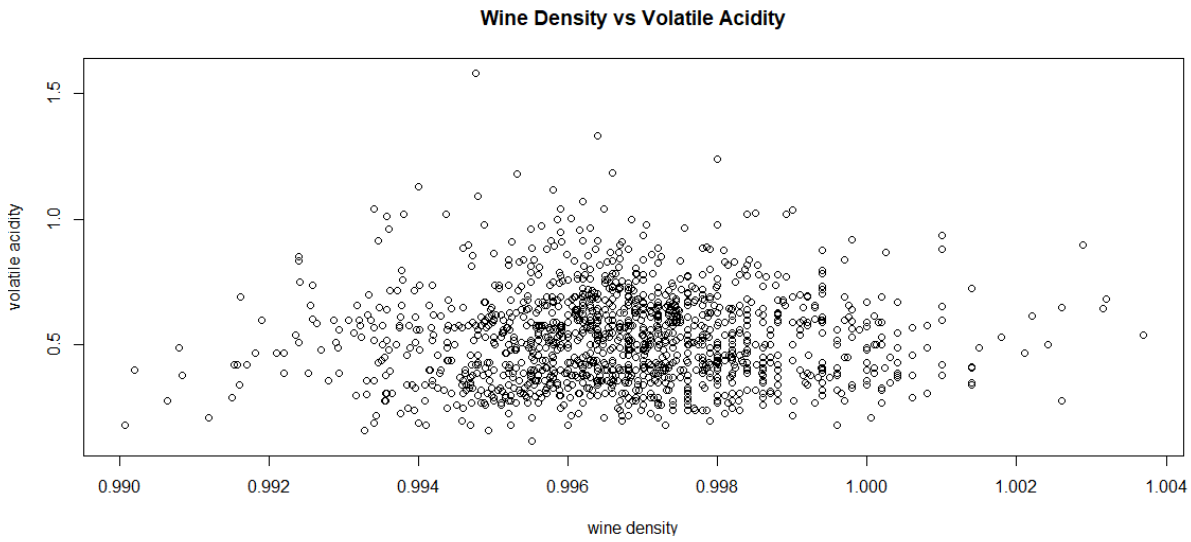

Figure 3: Wine Density vs Voltaile Acidity

From the scatter plot, it is visible that huge concentration of wine density between 0.994 and 0.998. Similarly, the volatile acidity lies between 0.0 and 1.0. However there is no correlation between wine density and volatile acidity since there is no linear pattern observed.

1.4 1.d.

Creating ALevel variable and producing box plot

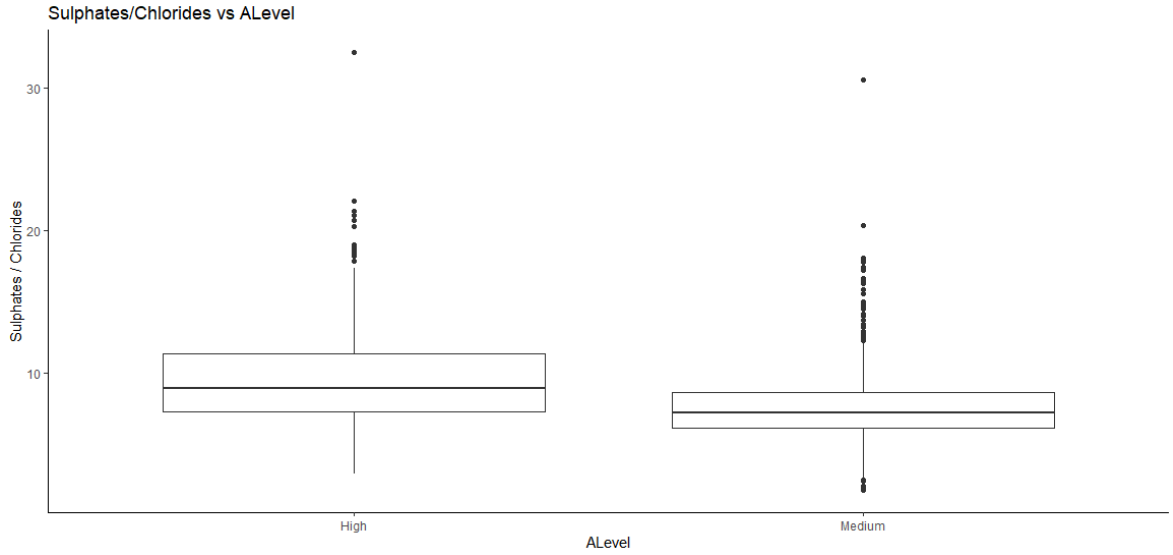

Figure 4: Box plot

```r
# creating ALevel variable
red_wine_data <- red_wine_data %>%
  mutate(ALevel = ifelse(alcohol > 10.5, "High", "Medium"))
# calculating the sulphates/chlorides
red_wine_data <- red_wine_data %>%
  mutate(sulphates_chlorides = sulphates / chlorides)
# generating box plot
box_plot <- ggplot(red_wine_data, aes(x = ALevel, y = sulphates_chlorides)) +
  geom_boxplot() +
  labs(
    title = "Sulphates/Chlorides vs ALevel",
    x = "ALevel",
    y = "Sulphates / Chlorides"
  ) +
  theme_classic()
print(box_plot)
# samples in high category
high_samples <- red_wine_data %>%
  filter(ALevel == "High") %>%
  nrow()
print(high_samples)
```

```
> print(high_samples)
[1] 616
```

Figure 5: Samples in high category

There are 616 samples in the high category

1.5 1.e.

Producing histogram

```r
# Function to create a histogram
create_histogram <- function(data, title, fill_color) {
  ggplot(data, aes(x = citric_acid)) +
    geom_histogram(binwidth = 0.03, fill = fill_color, color = "black", alpha = 0.7)
    +
    labs(
      title = title,
      x = "Citric Acid",
      y = "Number"
    )
}

# Create histograms for High and Medium ALevel
high_Alevel_plot <- create_histogram(filter(red_wine_data, ALevel == "High"), "
    Histogram for High ALevel Wines", "red")
medium_Alevel_plot <- create_histogram(filter(red_wine_data, ALevel == "Medium"), "
    Histogram for Medium ALevel Wines", "blue")

# Display histograms side by side
grid.arrange(high_Alevel_plot, medium_Alevel_plot, ncol = 2)
```

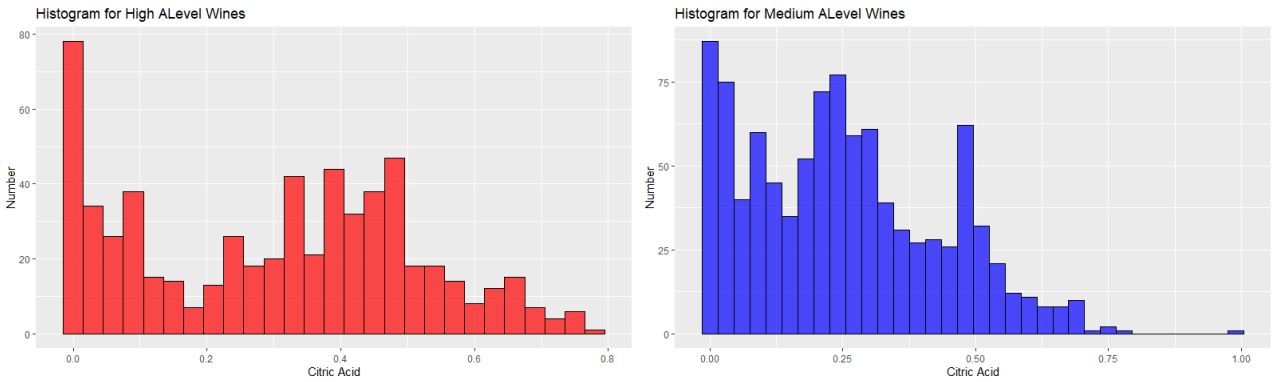

Figure 6: Histogram: Citric Acid Numbers

1.6 1.f.

Exploring the data

```
# Scatterplot: Fixed acidity vs Citric Acid
scat_plot1 <- ggplot(red_wine_data, aes(x = fixed_acidity, y = citric_acid)) +
  geom_point() + geom_smooth(method = "lm") +
  labs(x = "Fixed Acidity",
       y = "Citric Acid",
       title = "Fixed acidity vs Citric Acid")
print(scat_plot1)

# Scatterplot: Total Sulfur Dioxide vs Chlorides
scat_plot2 <- ggplot(red_wine_data, aes(x = total_sulfur_dioxide, y = chlorides)) +
  geom_point() + geom_smooth(method = "lm") +
  labs(x = "Total Sulfur Dioxide",
       y = "Chlorides",
       title = "Total Sulfur Dioxide vs Chlorides")
print(scat_plot2)
```

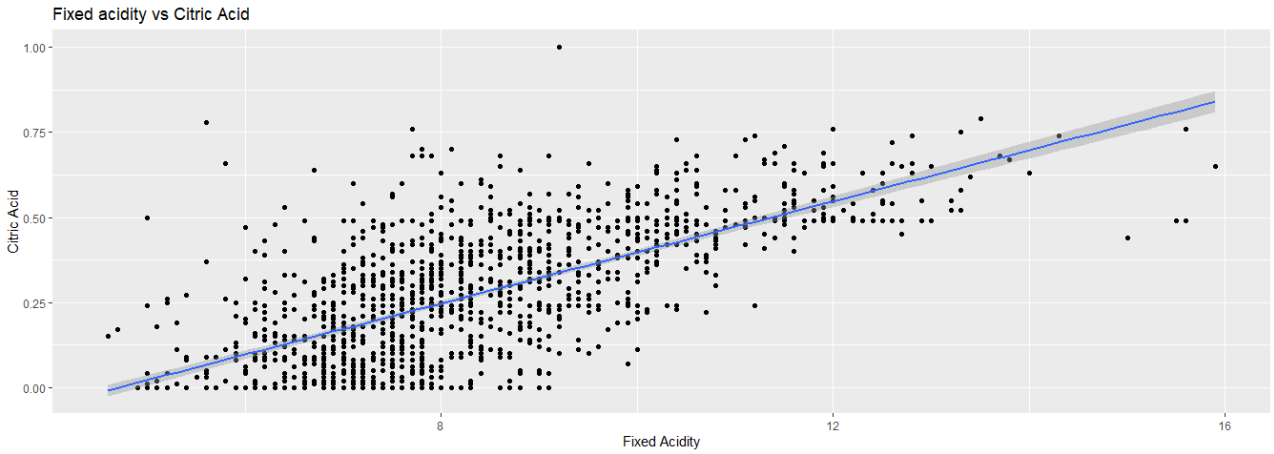

Figure 7: Fixed acidity vs Citric Acid

The scatter plot demonstrates a clear positive correlation between fixed acidity and citric acid in the wine dataset. This indicates that as citric acid levels increase, fixed acidity tends to rise as well. Therefore, citric acid appears to be a key factor contributing to the fixed acidity of the wine.

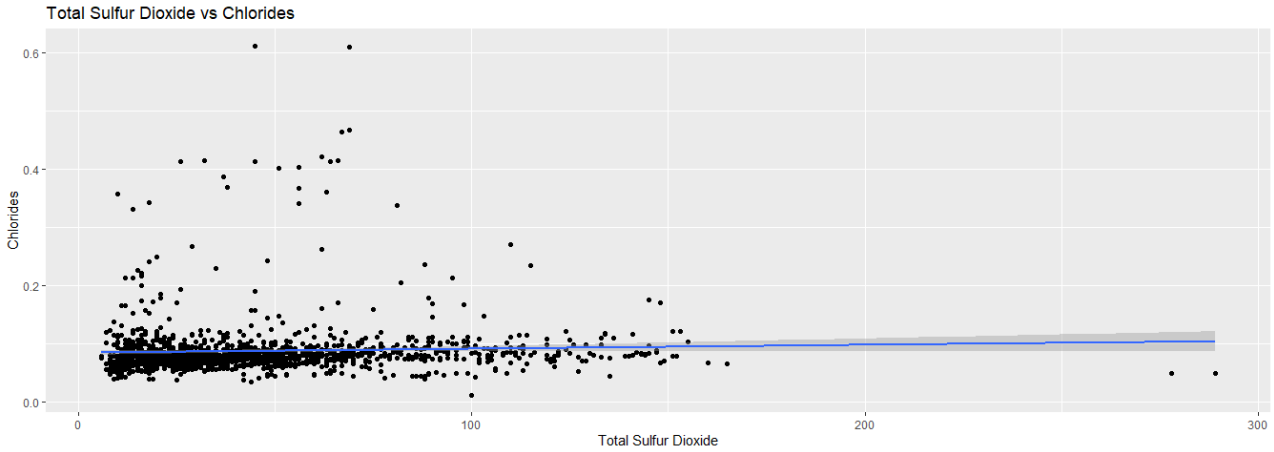

Figure 8: Total Sulfur Dioxide vs Chlorides

The scatter plot shows no clear correlation between total sulfur dioxide and chlorides in the wine dataset. This means that the amount of total sulfur dioxide remains widely distributed, regardless of the chloride levels. As a result, we can conclude that these two components are independent of each other. Wines can have high levels of sulfur dioxide with low chloride levels, or the opposite, without any consistent relationship between the two.

2 Exercise 2

2.1 2.a.

Quantitative Variables

Quantitative variables measure numerical properties that represent some kind of quantity, magnitude, or size. They can take on a wide range of values and are typically used for calculations like sums, means, or statistical analysis.

- **temp** and **atemp**: These represent actual temperatures, which are continuous measurements and can take on a wide range of values.
- **humidity** and **windspeed**: These are also continuous measures, indicating percentages or speeds that vary over a range.
- **count**: This is a numeric value representing the total number of rental bikes on a given day, making it a continuous variable, even though it's discrete in nature (i.e., it only takes whole numbers).

Qualitative Variables

Qualitative variables measure categorical properties or types, representing characteristics, classifications, or categories rather than quantities. They do not have a numerical value that directly signifies a magnitude or size, even if they are represented by numbers.

- **date**: Although it contains numbers (days, months, years), dates represent a point in time rather than a quantity, making this a qualitative variable.
- **season**: The values (1, 2, 3, 4) correspond to seasons, which are categories rather than continuous quantities. The numerical labels are used only for identification.
- **holiday** and **workingday**: These are binary variables, signifying a classification into two groups (yes/no), so they are qualitative.
- **weather**: Similar to season, this variable classifies the weather into categories (clear, cloudy, rain, etc.), and the numerical labels are just identifiers.

2.2 2.b.

```
# reading the data
bikes_data <- read.csv("bikes.csv")

# Summary statistics for quantitative variables
quantitative_values <- bikes_data %>%
  summarise(across(c(temp, atemp, humidity, windspeed, count),
                   list(
                     range = ~max(.x, na.rm = TRUE) - min(.x, na.rm = TRUE),
                     mean = ~mean(.x, na.rm = TRUE),
                     standarddeviation = ~sd(.x, na.rm = TRUE)
                   ), .names = "{.col}_{.fn}")
```

```
  )
print(quantitative_values)
```

```
> print(quantitative_values)
  temp_range temp_mean temp_standarddeviation atemp_range atemp_mean atemp_standarddeviation humidity_range humidity_mean
1  0.8025366 0.4953848              0.183051   0.7618264   0.474354              0.1629612         0.9725     0.6278941
  humidity_standarddeviation windspeed_range windspeed_mean windspeed_standarddeviation count_range count_mean count_standarddeviation
1                  0.1424291       0.4850713      0.1904862                 0.07749787        8692   4504.349               1937.211
```

Figure 9: Output

- The dataset shows that temperature and windspeed have relatively low variability, while humidity and bike rental counts show much larger variation.
- The high standard deviation for bike rentals suggests that external factors may strongly affect the demand for bike-sharing services (such as weather, holiday status, or workday status).
- Further analysis could investigate the correlation between these factors and the count of rentals, potentially revealing patterns or trends that affect bike-sharing demand.

```
# Calculate the average bike rental count for each season and sort in descending
    order
season_average <- bikes_data %>%
  group_by(season) %>%
  summarise(count_mean = mean(count, na.rm = TRUE)) %>%
  arrange(desc(count_mean))

# Display the average bike rental count for all seasons
print(season_average)

# Display the season with the highest average bike rental count
print(season_average[1, ])
```

```
> print(season_average[1, ])
# A tibble: 1 x 2
  season count_mean
   <int>      <dbl>
1      3      5644.
```

Figure 10: Output

Season 3 (summer) stands out as the peak season for bike-sharing services, as people tend to prefer biking in warm and pleasant weather. This insight could be useful for bike-sharing companies in planning for higher demand during summer by increasing bike availability or implementing special offers to maximize usage.

2.3 2.c.

```r
# boxplot to visualize bike rental counts across different weather conditions
ggplot(bikes_data, aes(x = factor(weather), y = count)) +
  geom_boxplot(fill = "red") +
  labs(
    title = "Bike Rental Counts by Weather Condition",
    x = "Weather Condition",
    y = "Number of Bike Rentals"
  )
```

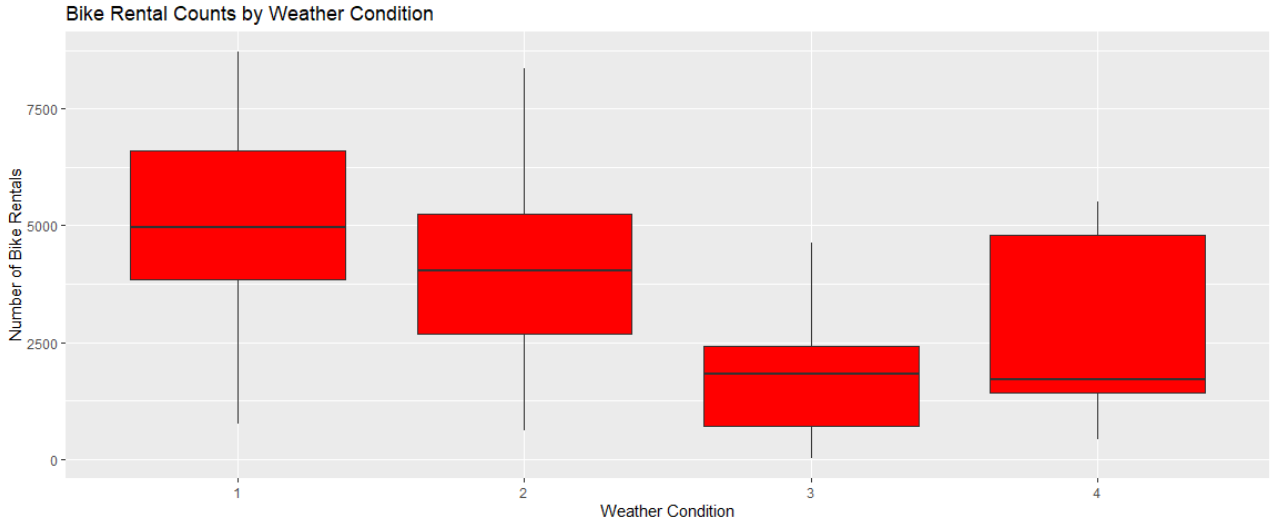

Figure 11: Bike rental counts by weather condition

Weather condition 1 (clear) has the highest median bike rental count, as evidenced by the higher position of the median line (the line inside the box) compared to other conditions. As the weather condition worsens (e.g., heavy rain/snow), the median count of rentals decreases, reflecting the impact of unfavorable weather on bike usage

2.4 2.d.

```r
# Ensure the 'date' column is converted to Date type and extract the month as a new
    column
bikes_data$date <- as.Date(bikes_data$date, format = "%m/%d/%y")
bikes_data$month <- format(bikes_data$date, "%B")  # Extract month in full name

# Create a bar plot showing the total number of rentals for each month
ggplot(bikes_data, aes(x = factor(month, levels = month.name), weight = count)) +
  geom_bar(fill = "blue") +
  labs(title = "Total Bike Rentals by Month",
       x = "Month",
       y = "Total Rentals") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

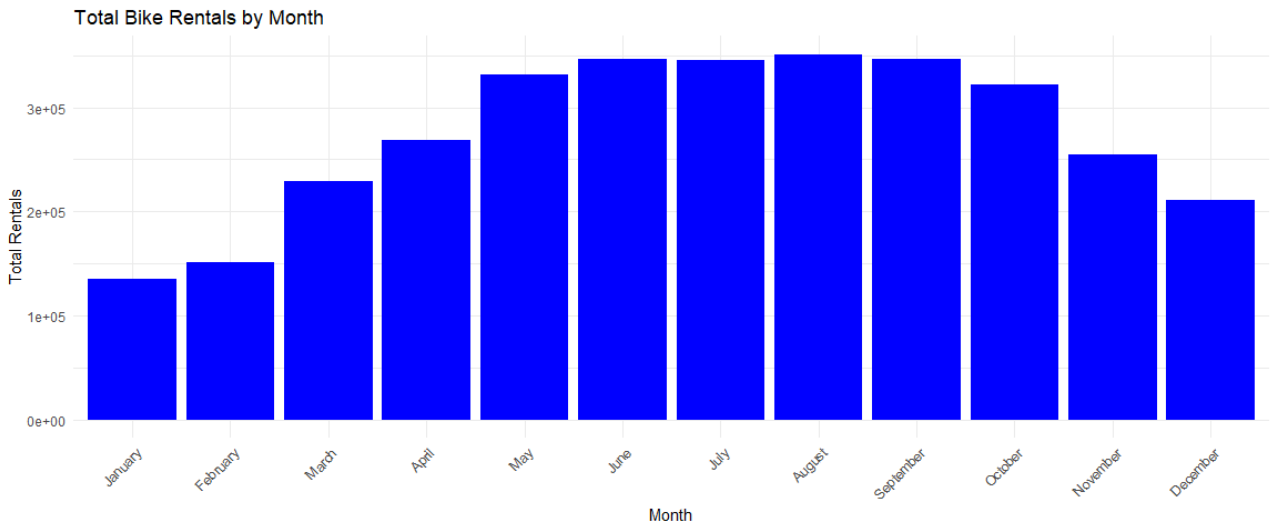

Figure 12: Total Bike Rentals by Month

August has the highest number of bike rentals. The chart reveals a clear seasonal pattern, with rentals peaking during the summer months (June, July, August). This is likely due to favorable weather conditions that encourage outdoor activities such as biking. August stands out as the month with the highest bike rentals, suggesting that late summer is the most active period for biking. This could be due to optimal weather, longer days, and possibly vacation or leisure time.

2.5 2.e.

```
# Select relevant quantitative variables
quantitative_values <- bikes_data %>%
  select(temp, atemp, humidity, windspeed, count)

# Create a correlation matrix
correlation_matrix <- cor(quantitative_values, use = "complete.obs")

# Print the correlation matrix
print(correlation_matrix)

# Reshape the correlation matrix for visualization using melt
correlation_data <- melt(correlation_matrix)

# Visualize the correlation matrix using ggplot2
ggplot(correlation_data, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "green", high = "red", mid = "white", midpoint = 0) +
  labs(title = "Correlation Matrix of Quantitative Variables",
       x = "Variables",
       y = "Variables") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
                temp      atemp   humidity  windspeed      count
temp       1.0000000  0.9917016  0.1269629 -0.1579441  0.6274940
atemp      0.9917016  1.0000000  0.1399881 -0.1836430  0.6310657
humidity   0.1269629  0.1399881  1.0000000 -0.2484891 -0.1006586
windspeed -0.1579441 -0.1836430 -0.2484891  1.0000000 -0.2345450
count      0.6274940  0.6310657 -0.1006586 -0.2345450  1.0000000
```

Figure 13: Output

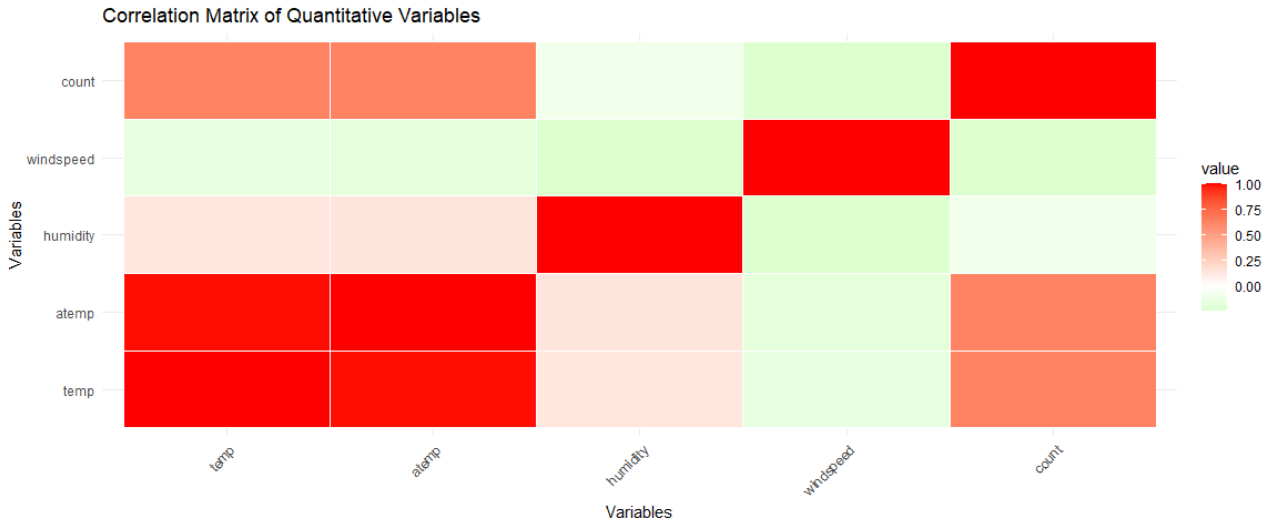

Figure 14: Correlation Matrix of Quantitative Variables

Temperature (both actual and perceived) is the most influential factor on bike rentals, as indicated by its moderate positive correlation with the number of rentals. Windspeed negatively affects bike rentals, which implies unfavorable weather conditions (strong winds) reduce the number of bike users. Humidity has a smaller impact on rentals, indicating that it is not as significant in determining biking activity.

2.6 2.f.

- Temperature (temp) and Feels-like Temperature (atemp) are strong predictors of bike rental counts, with both showing a moderate positive correlation (around 0.63). Warmer temperatures are associated with an increase in bike rentals.
- Since temp and atemp are highly correlated with each other (close to 1), using one of these variables for prediction is recommended to avoid multicollinearity.
- Windspeed has a moderate negative correlation (around -0.23) with bike rentals, suggesting that higher wind speeds tend to reduce the number of rentals, making it a useful predictor in windy conditions.
- Humidity shows a weak negative correlation (-0.10) with bike rental counts. While its influence is less significant compared to temperature and windspeed, it may still provide some predictive value when combined with other weather variables.

- Holiday and working day are not included in the correlation matrix, but based on other analyses, temperature and windspeed appear to be the most important predictors for bike rental counts.
- Overall, temperature and windspeed are the most significant variables to consider for predicting bike rentals, with humidity offering marginal predictive power.