3. EDA.

3.1 Bảng tổng hợp dữ liệu.

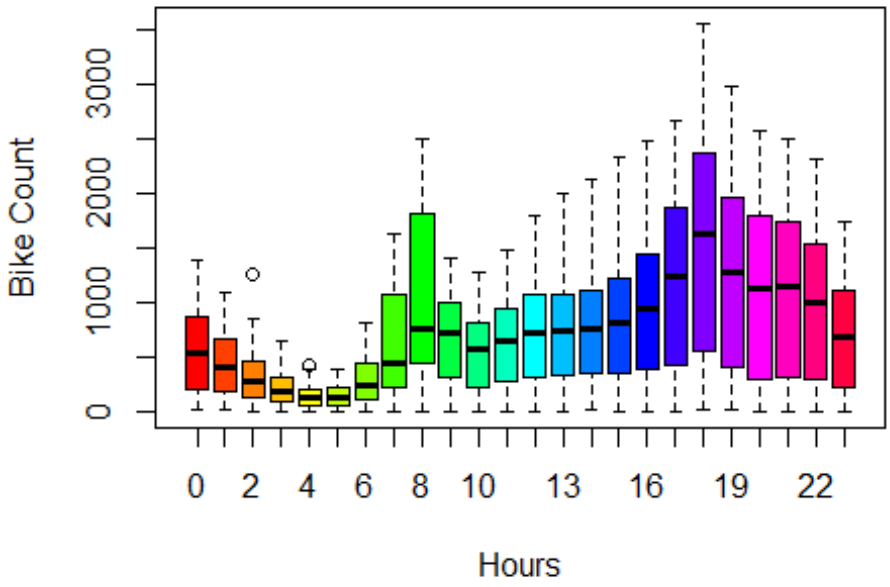
```
summary(data)
```

```
##      date           rented_bike_count      hour       temperature_c
##  Length:8760        Min.   :   0.0    Min.   : 0.00   Min.   :-17.80
##  Class :character   1st Qu.: 191.0    1st Qu.: 5.75   1st Qu.:  3.50
##  Mode  :character   Median : 504.5    Median :11.50   Median : 13.70
##                     Mean   : 704.6    Mean   :11.50   Mean   : 12.88
##                     3rd Qu.:1065.2    3rd Qu.:17.25   3rd Qu.: 22.50
##                     Max.   :3556.0    Max.   :23.00   Max.   : 39.40
##  humidity_percent wind_speed_m_s  visibility_10m dew_point_temperature_c
##  Min.   : 0.00    Min.   :0.000   Min.   :  27   Min.   :-30.600
##  1st Qu.:42.00    1st Qu.:0.900   1st Qu.: 940   1st Qu.: -4.700
##  Median :57.00    Median :1.500   Median :1698   Median :  5.100
##  Mean   :58.23    Mean   :1.725   Mean   :1437   Mean   :  4.074
##  3rd Qu.:74.00    3rd Qu.:2.300   3rd Qu.:2000   3rd Qu.: 14.800
##  Max.   :98.00    Max.   :7.400   Max.   :2000   Max.   : 27.200
##  solar_radiation_mj_m2  rainfall_mm       snowfall_cm        seasons
##  Min.   :0.0000        Min.   : 0.0000   Min.   :0.00000   Length:8760
##  1st Qu.:0.0000        1st Qu.: 0.0000   1st Qu.:0.00000   Class :character
##  Median :0.0100        Median : 0.0000   Median :0.00000   Mode  :character
##  Mean   :0.5691        Mean   : 0.1487   Mean   :0.07507
##  3rd Qu.:0.9300        3rd Qu.: 0.0000   3rd Qu.:0.00000
##  Max.   :3.5200        Max.   :35.0000   Max.   :8.80000
##    holiday          functioning_day
##  Length:8760        Length:8760
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

3.3 Trực quan dữ liệu.

3.3.1 Vẽ biểu đồ boxplot nhằm xem xét phân phối của số lượng xe đạp được thuê (rented_bike_count) vào từng giờ trong ngày (hour)

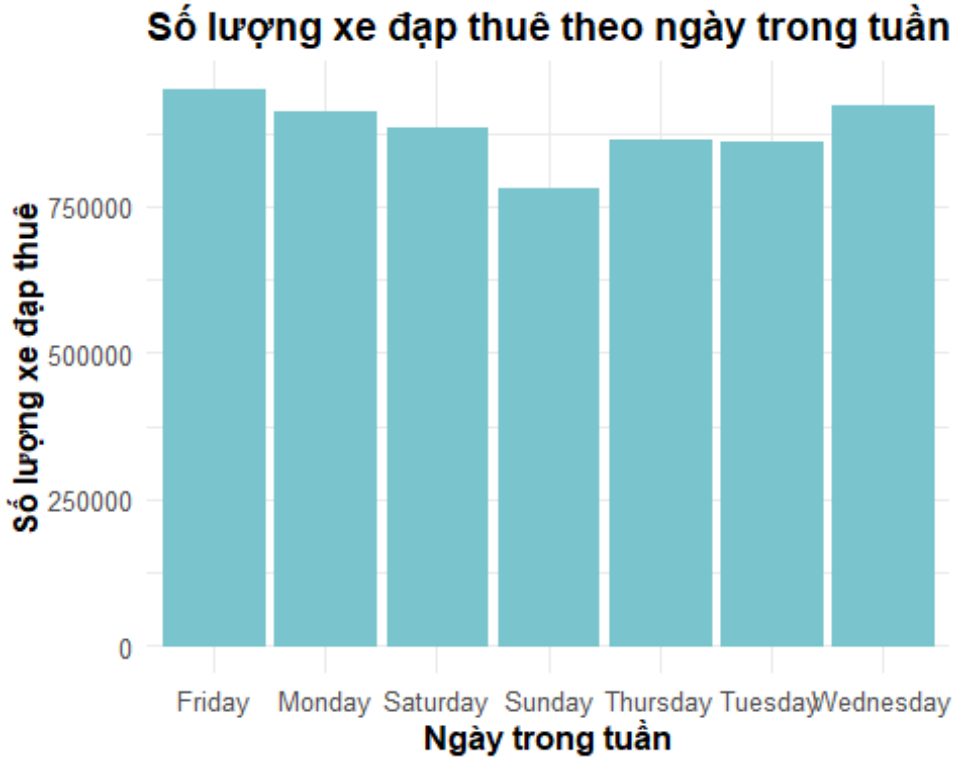
```
# Vẽ biểu đồ boxplot
boxplot(data$rented_bike_count ~ data$hour,
        xlab = "Hours",
        ylab = "Bike Count",
        col = rainbow(length(unique(data$hour))))
```



Biểu đồ ở trên cho thấy số lượng xe đạp được thuê thay đổi theo giờ. Dễ dàng nhận thấy từ biểu đồ rằng nhiều xe đạp được thuê nhất từ 7 giờ sáng đến 9 giờ sáng. Sau đó, số lượng xe thuê duy trì tương đối ổn định cho đến 15:00 chiều và bắt đầu tăng trở lại sau 15:00 chiều.

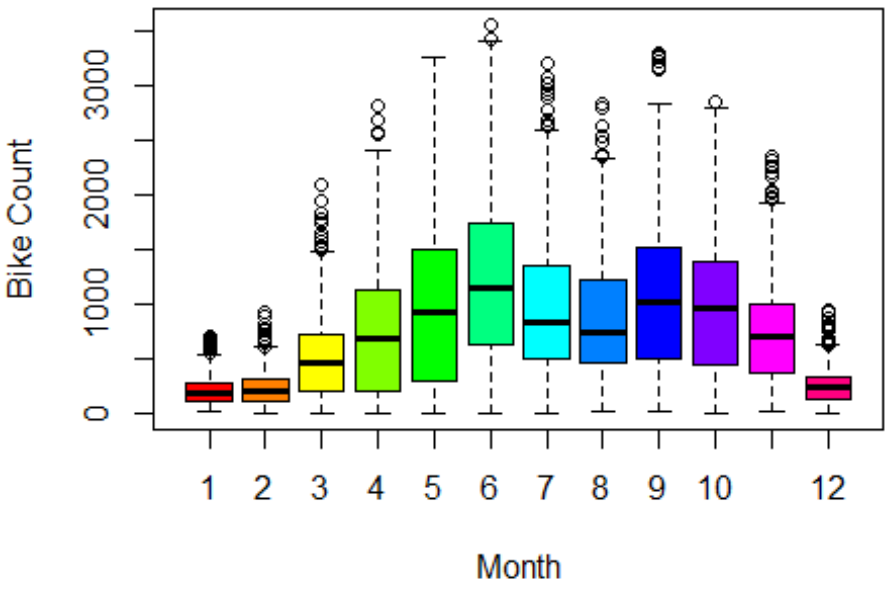
3.3.2 Vẽ biểu đồ thanh cho số lượng xe đạp thuê theo ngày trong tuần.

```
# Biểu đồ thanh cho số lượng xe đạp thuê theo ngày trong tuần
ggplot(data, aes(x = day_of_week, y = rented_bike_count, fill = day_of_week)) +
  geom_bar(stat = "identity", color = "#7AC5CD") +
  labs(
    title = "Số lượng xe đạp thuê theo ngày trong tuần",
    x = "Ngày trong tuần",
    y = "Số lượng xe đạp thuê"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
    axis.title.x = element_text(size = 12, face = "bold"),
    axis.title.y = element_text(size = 12, face = "bold"),
    axis.text.x = element_text(size = 10),
    axis.text.y = element_text(size = 10),
    legend.position = "none"
  ) + scale_fill_brewer(palette = "Set3")
```



3.3.3 Vẽ biểu đồ boxplot cho số lượng xe đạp thuê theo tháng trong năm.

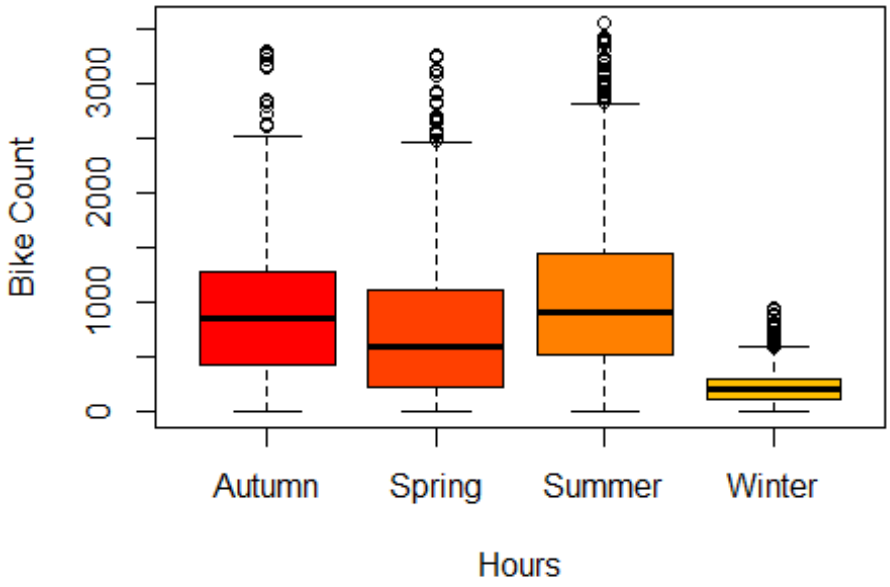
```
# Vẽ biểu đồ boxplot
boxplot(data$rented_bike_count ~ data$month,
        xlab = "Month",
        ylab = "Bike Count",
        col = rainbow(length(unique(data$month))))
```



=> Nhận xét: Qua biểu đồ ta thấy tháng 6 là tháng có số lượng xe đạp thuê cao nhất, trong khi tháng 1 và tháng 12 là các tháng có số lượng xe đạp thuê thấp nhất.

3.3.4 Vẽ biểu đồ boxplot cho số lượng xe đạp thuê theo mùa.

```
# Vẽ biểu đồ boxplot
boxplot(data$rented_bike_count ~ data$seasons,
        xlab = "Hours",
        ylab = "Bike Count",
        col = rainbow(length(unique(data$hour))))
```

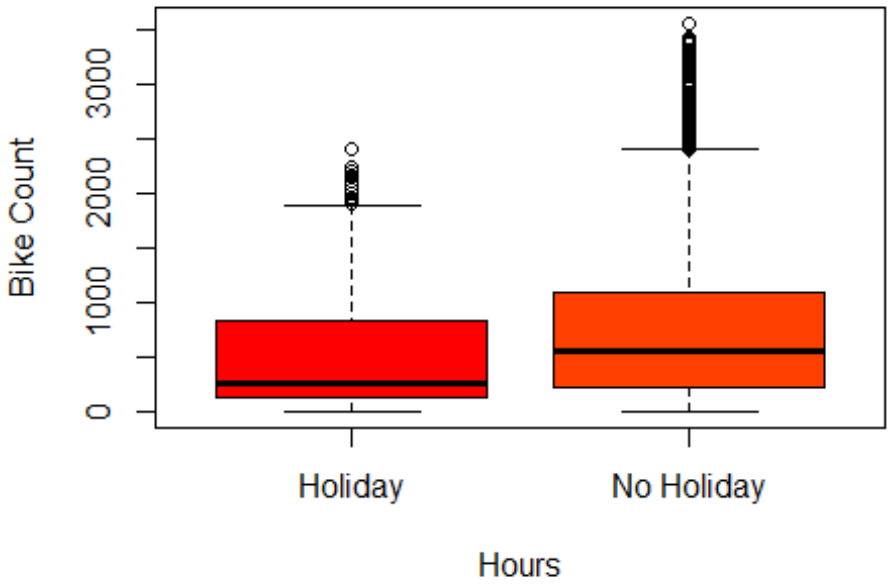


=> **Nhận xét**: Ta có thể thấy rằng trong những thời điểm thời tiết tốt hơn như mùa hè, mọi người thường thích đạp xe hơn và vào mùa đông xe đạp được thuê với giá thấp nhất với số lượng nhu cầu thấp hơn 500 chiếc. Điều này có thể là do thời tiết lạnh cũng như tuyết trong mùa đông. Có một số ngoại lệ với mật độ cao hơn các mùa còn lại khi nhu cầu xe đạp trong Mùa đông là trên 500. Đây có thể là một số lý do cơ bản chẳng hạn như ngày có thời tiết tốt hơn những ngày khác trong mùa đông. + Vào mùa thu, nhu cầu xe đạp cao hơn mùa xuân mặc dù chúng tôi dự kiến nhu cầu vào mùa xuân sẽ cao hơn mùa thu. Điều này có thể là do mùa Xuân có nhiều ngày mưa hơn mùa Thu.

Việc sử dụng xe đạp cao nhất vào mùa thu và mùa xuân, với trung vị và sự biến đổi tương tự nhau. Mùa hè có trung vị thấp hơn nhưng sự biến đổi số lượng xe đạp cao hơn. Mùa đông cho thấy việc sử dụng xe đạp thấp nhất với sự biến đổi ít nhất.

3.3.5 Vẽ biểu đồ boxplot cho số lượng xe đạp thuê theo ngày lễ.

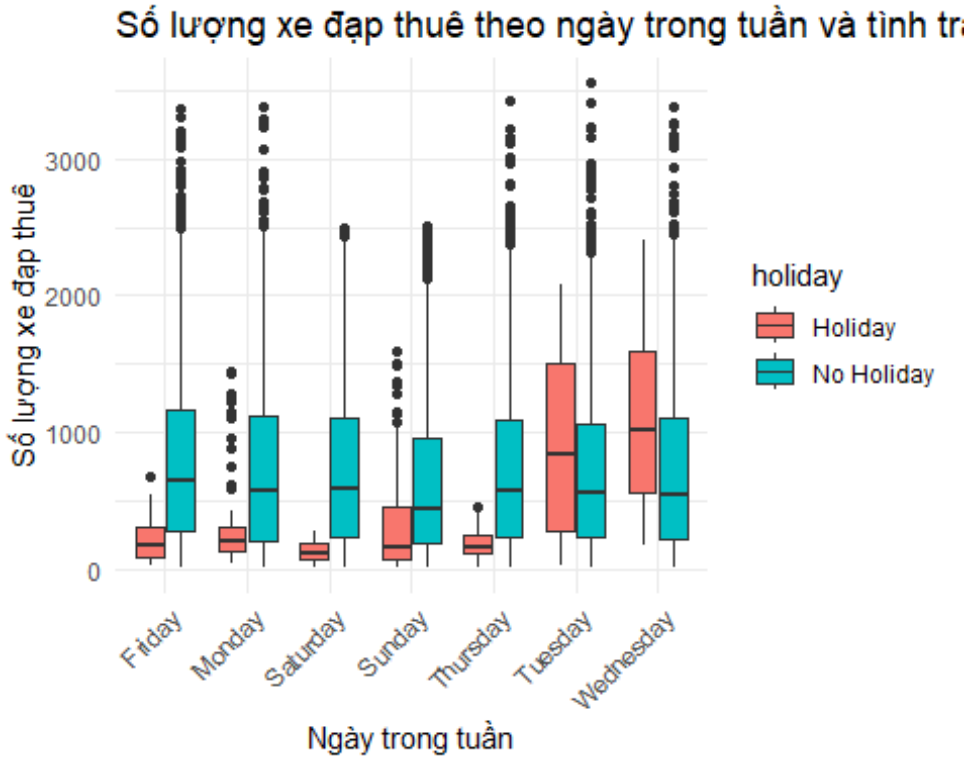
```
# Vẽ biểu đồ boxplot
boxplot(data$rented_bike_count ~ data$holiday,
        xlab = "Hours",
        ylab = "Bike Count",
        col = rainbow(length(unique(data$hour))))
```



=> **Nhận xét**: Ta có thấy rằng nhu cầu về xe đạp cao hơn vào những ngày không phải ngày nghỉ lễ, điều đó có nghĩa là hầu hết người dùng có thể thuê xe đạp vì những lý do khác ngoài mục đích giải trí.

3.3.6 Vẽ biểu đồ boxplot cho số lượng xe đạp thuê theo ngày trong tuần phân loại ngày lễ.

```
ggplot(data, aes(x = day_of_week, y = rented_bike_count, fill = holiday)) +
  geom_boxplot() +
  labs(title = "Số lượng xe đạp thuê theo ngày trong tuần và tình trạng ngày lễ",
       x = "Ngày trong tuần",
       y = "Số lượng xe đạp thuê") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



=> Nhận xét:

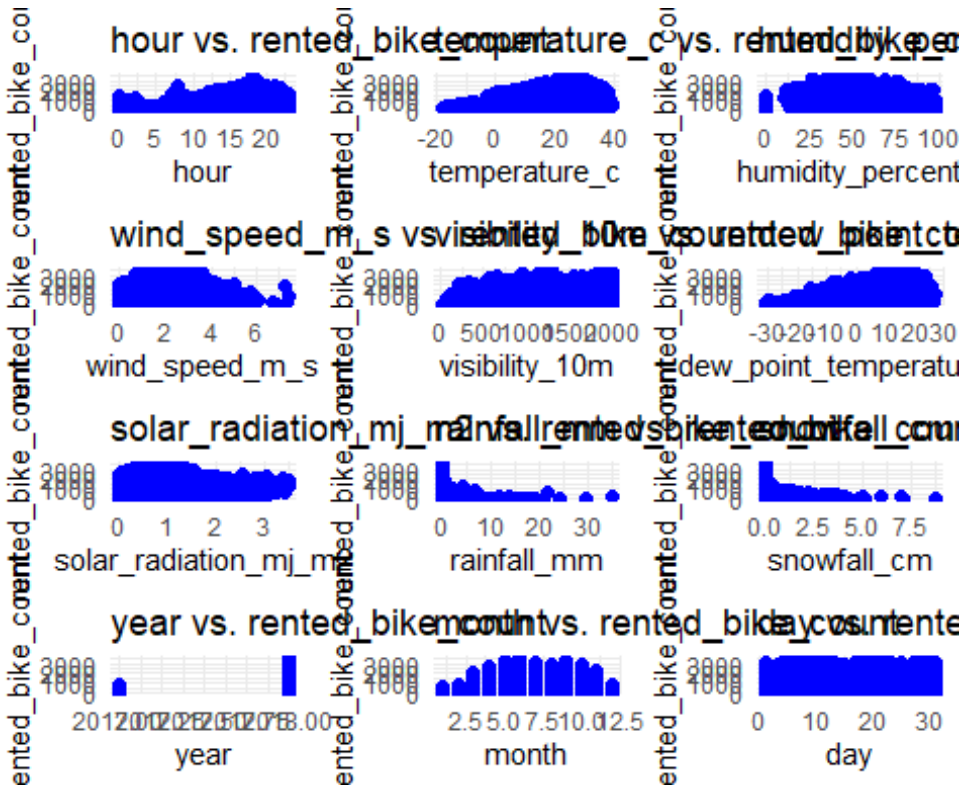
- Biểu đồ cho thấy số lượng xe đạp thuê vào các ngày lễ cuối tuần thường cao hơn so với ngày cuối tuần không phải ngày lễ . Ngày cuối tuần (thứ Bảy và Chủ Nhật) thường có nhiều hoạt động giải trí và du lịch, do đó, số lượng xe đạp thuê vào các ngày này cao và biến động lớn.
- Nếu một ngày lễ rơi vào cuối tuần, có thể tăng cường thêm hoạt động ngoài trời, dẫn đến số lượng xe đạp thuê cao hơn
- Ta cũng có thể thấy thứ Ba và thứ Tư là những ngày làm việc trong tuần. Khi ngày lễ trùng vào những ngày này, nhiều người có thể sử dụng thời gian nghỉ để tham gia các hoạt động ngoài trời, du lịch ngắn ngày hoặc đơn giản là thư giãn => sử dụng xe đạp tương đối lớn.

3.3.7 Biểu đồ tất cả các biến định lượng so với số lượng xe đạp thuê.

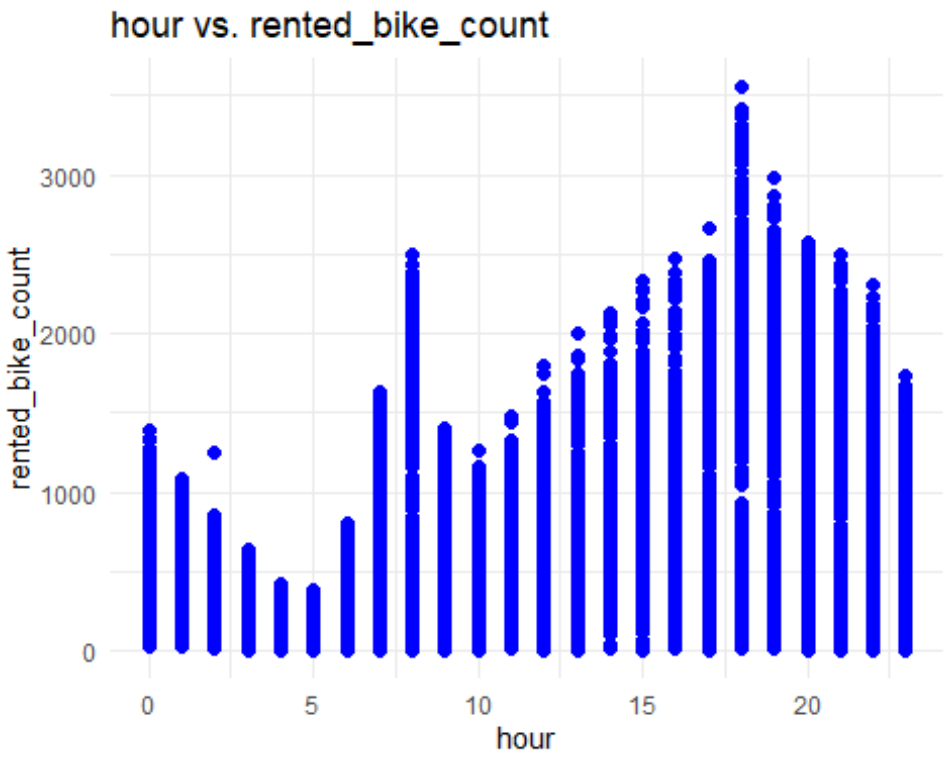
```
multi.scatter <- function(data, target) {
  # Initialize an empty list to store plots
  plots <- list()

  # Loop through each numeric variable
  for (col in names(data)) {
    # Check if current variable is numeric and not the target variable
    if (is.numeric(data[[col]]) && col != target) {
      # Create scatter plot
      scatter_plot <- ggplot(data, aes_string(x = col, y = target)) +
        geom_point(size = 2, color = "blue") +
        ggtitle(paste(col, "vs.", target)) +
        theme_minimal()

      # Add the plot to the list
      plots[[col]] <- scatter_plot
```

```
    }
  }

  # Arrange plots in a grid
  grid.arrange(grobs = plots, ncol = 3)  # Adjust ncol as needed

  # Return the list of plots (optional)
  return(plots)
}

# Call the function with numeric columns and target variable
multi.scatter(data, "rented_bike_count")
```

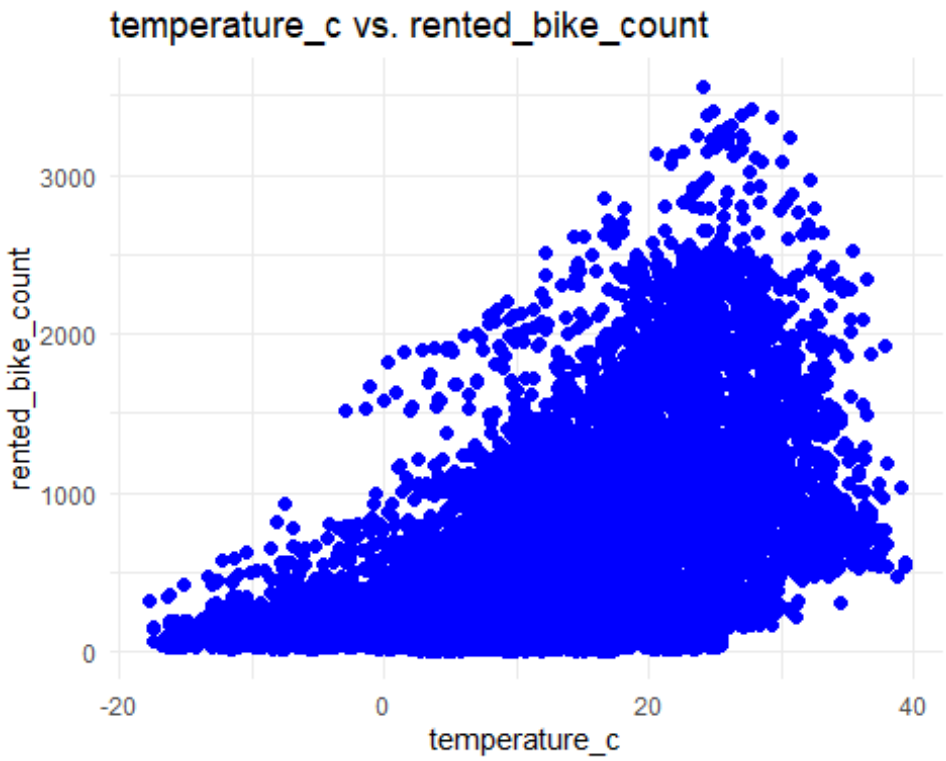
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## ℹ Please use tidy evaluation idioms with `aes()`.
## ℹ See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



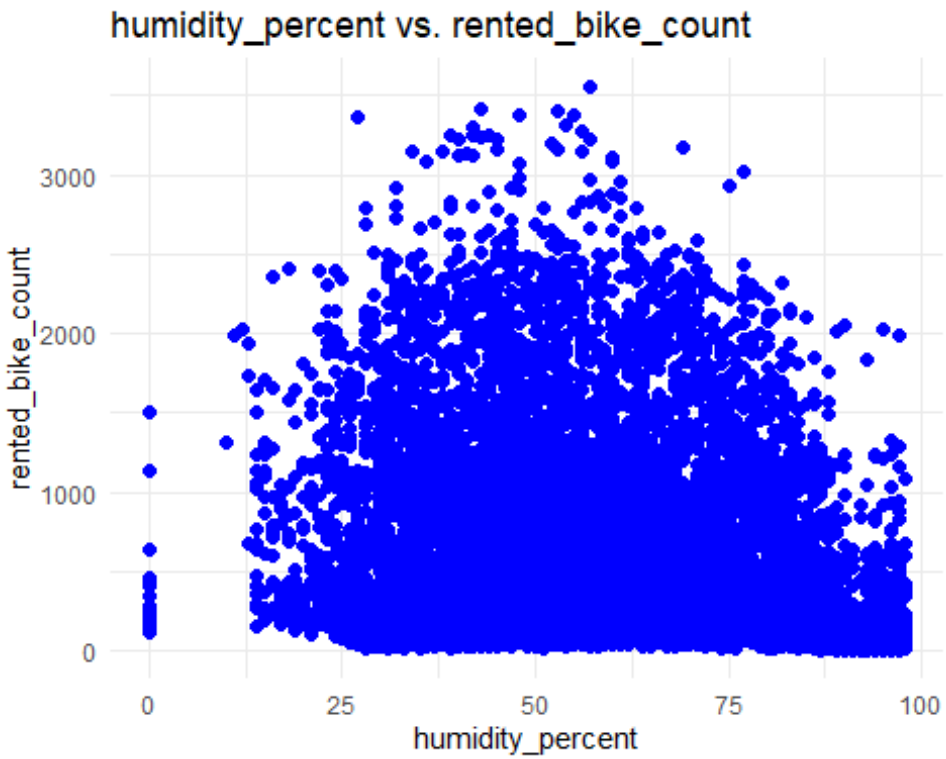

```
## $hour
```



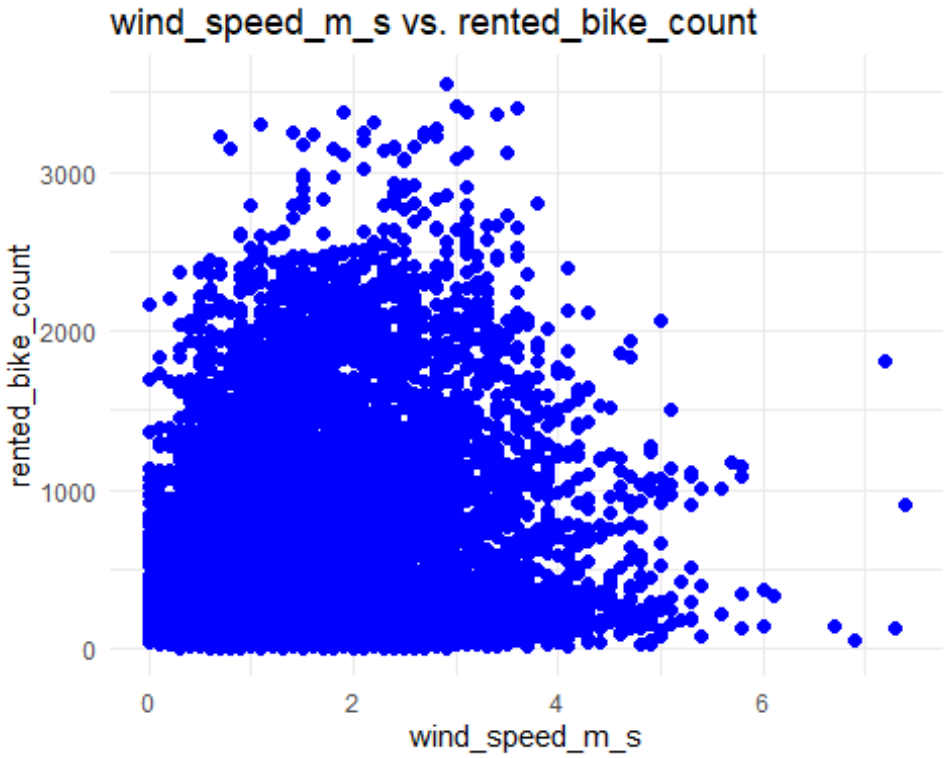

```
##
## $temperature_c
```



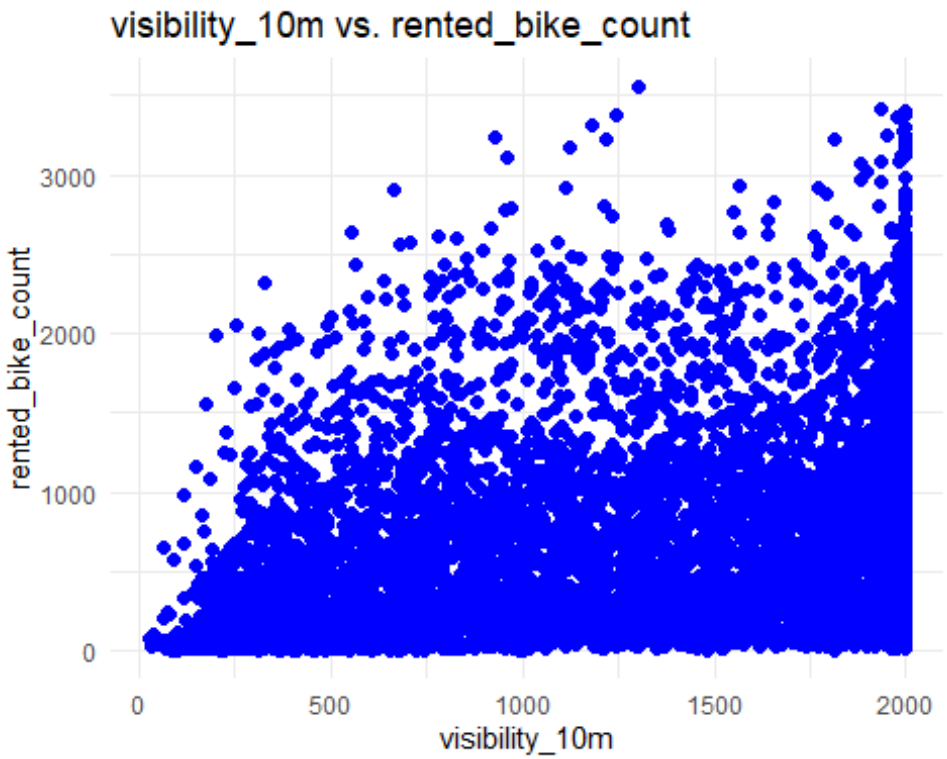

```
##
## $humidity_percent
```



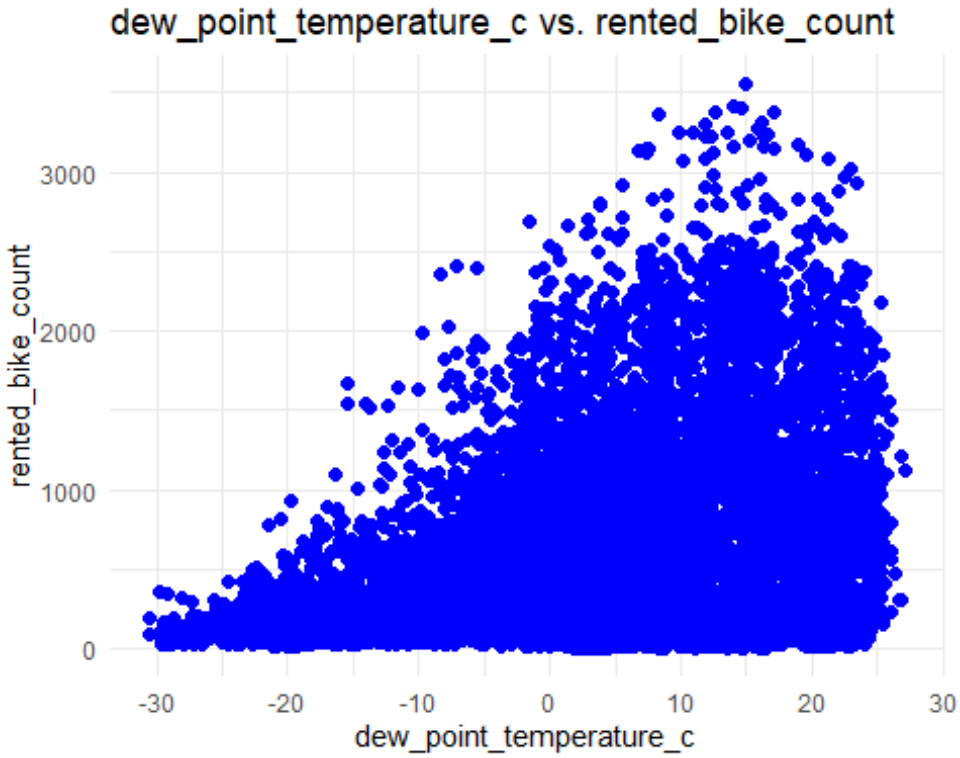

```
##
## $wind_speed_m_s
```



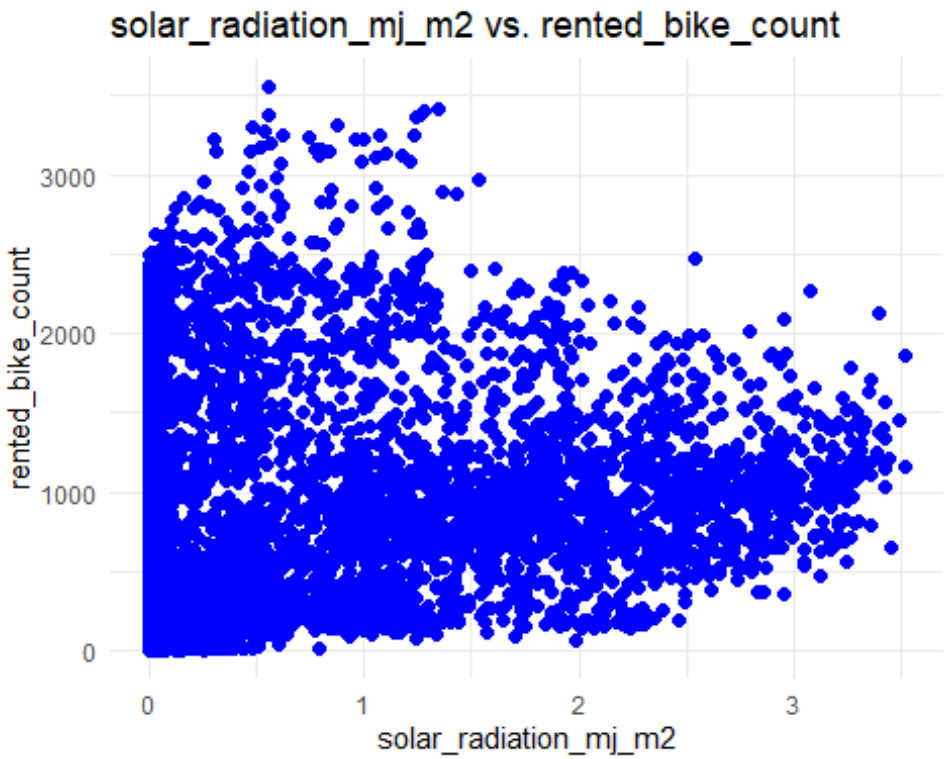

```
##
## $visibility_10m
```



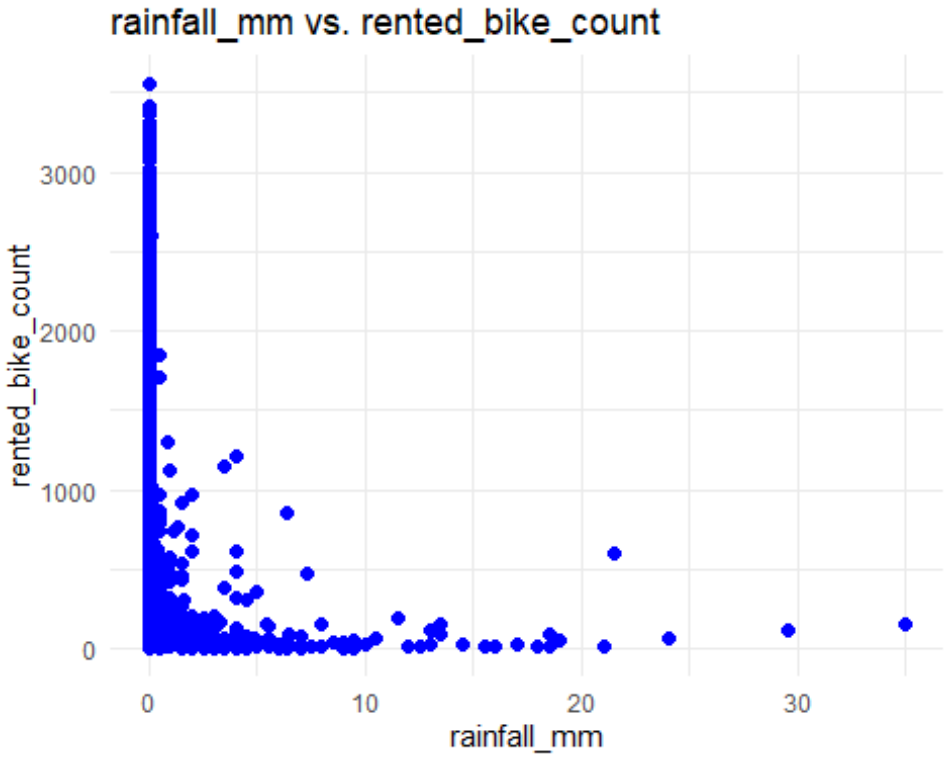

```
##
## $dew_point_temperature_c
```



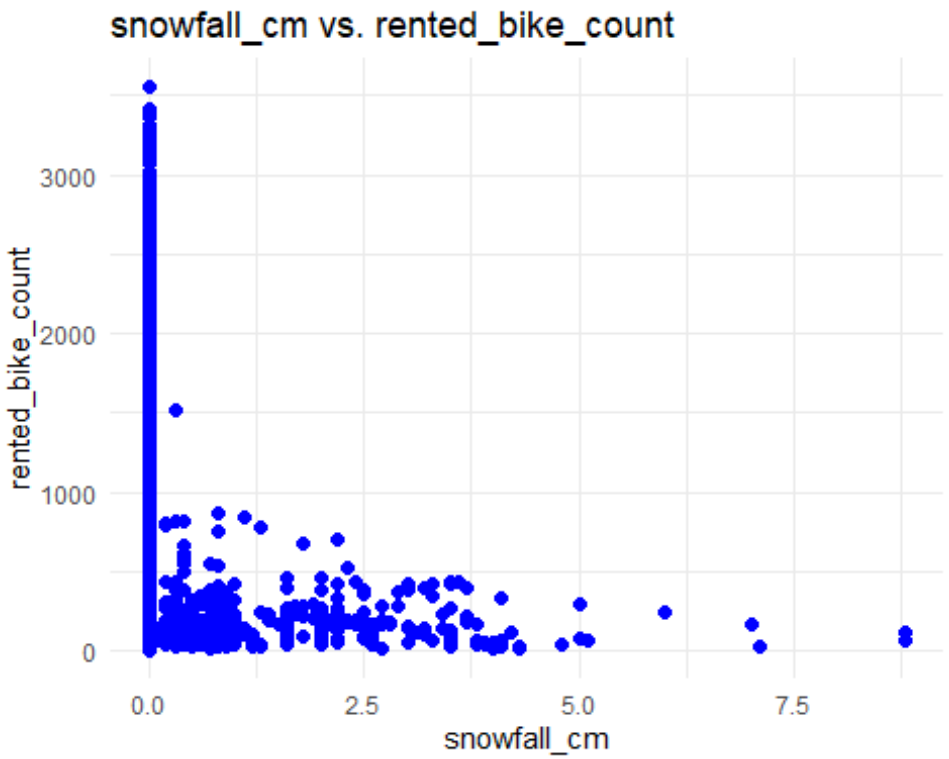

```
##
## $solar_radiation_mj_m2
```



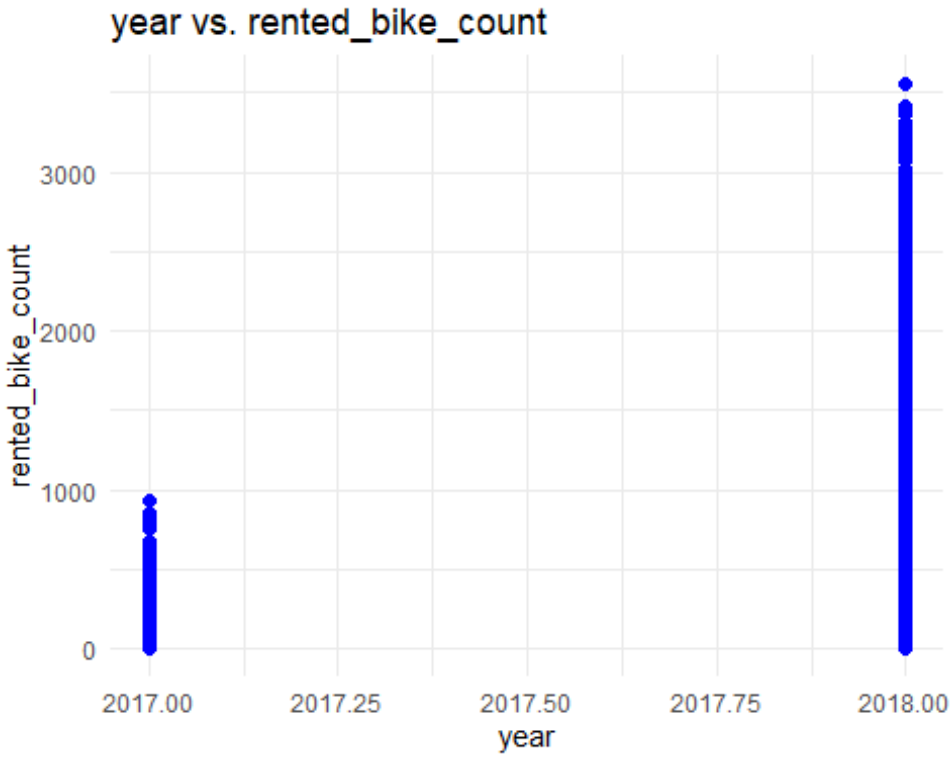

```
##
## $rainfall_mm
```



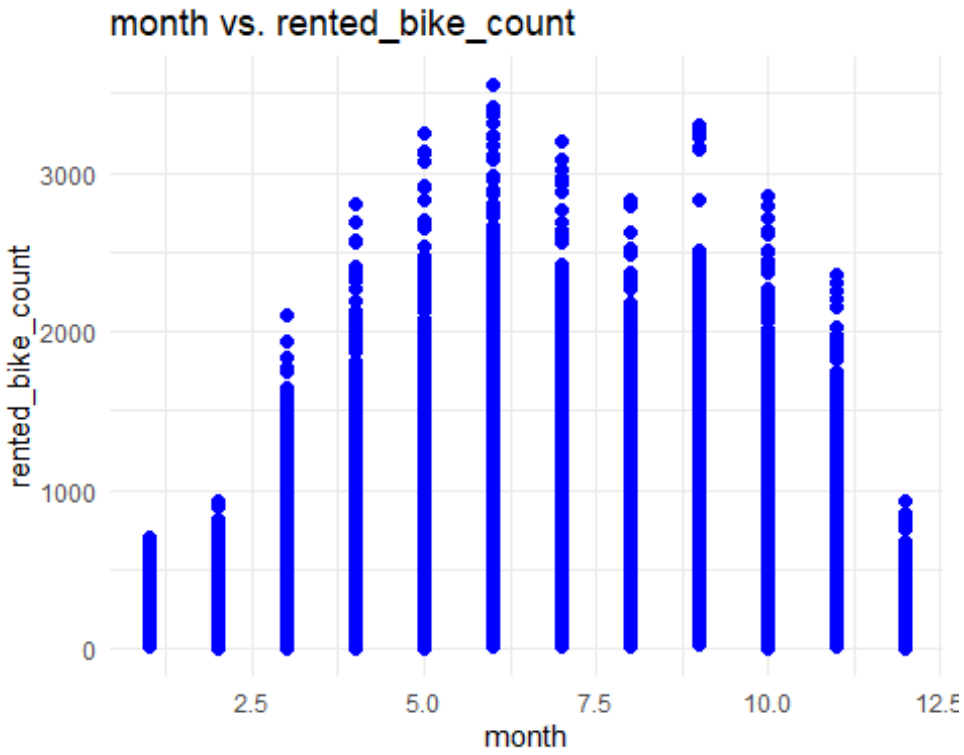

```
##
## $snowfall_cm
```


snowfall_cm vs. rented_bike_count
rented_bike_count
3000
2000
1000
0
0.0
2.5
5.0
7.5
snowfall_cm

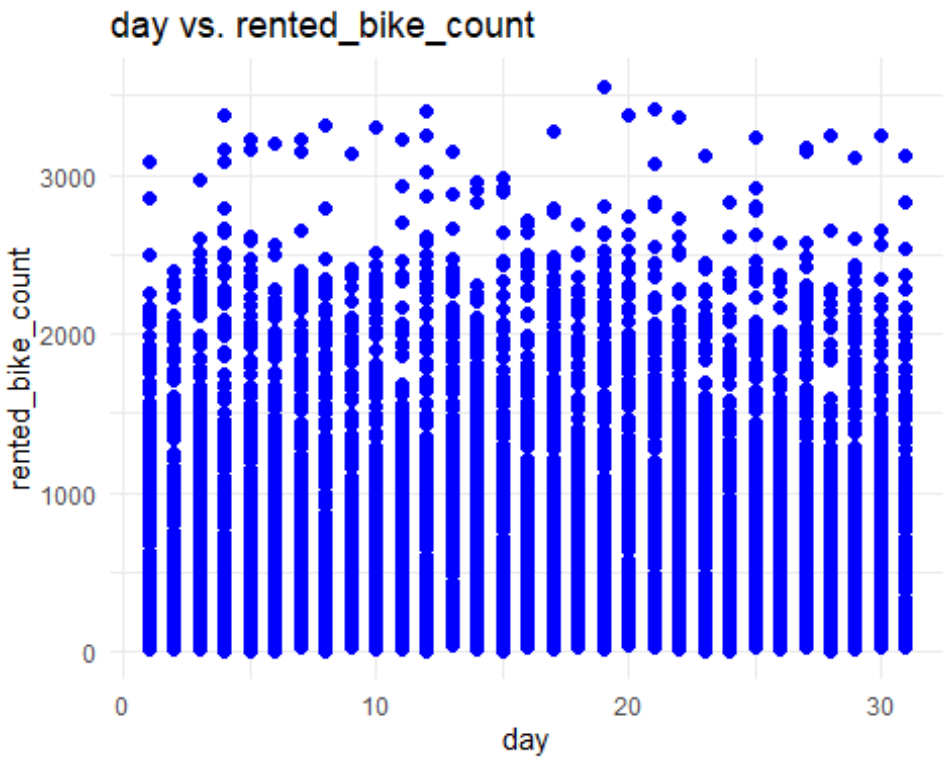
```
##
## $year
```


year vs. rented_bike_count
rented_bike_count
3000
2000
1000
0
2017.00
2017.25
2017.50
2017.75
2018.00
year

```
##
## $month
```



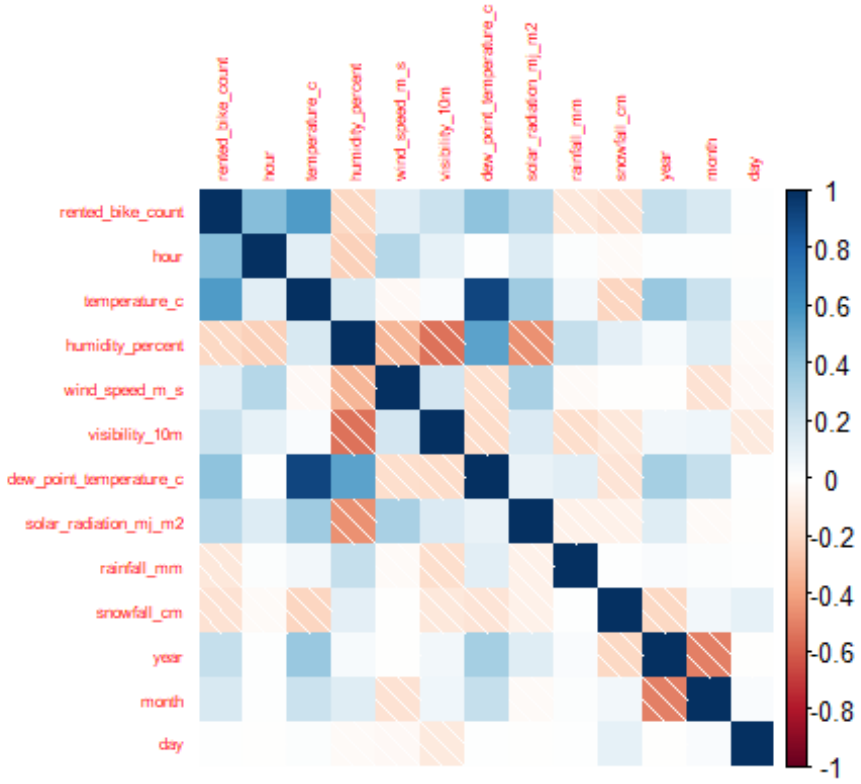

```
##
## $day
```



=> Nhận xét: Hầu hết tất cả các định lượng thể hiện thông tin môi trường bên ngoài điều chỉ là tuyến tính yếu, ngoài ra còn có độc lập tuyến tính.

3.3.8 Correlation Matrix

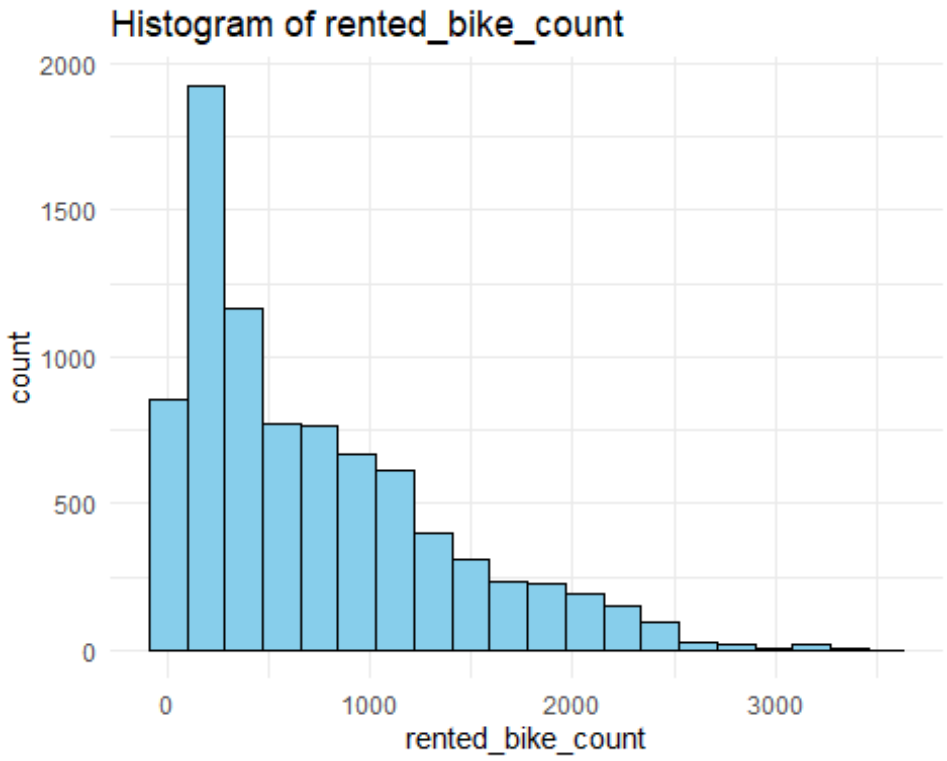
```
numeric_data <- data[, sapply(data, is.numeric)]
correlation_matrix <- cor(numeric_data)

# Load the corrplot package
library(corrplot)

# Plot the correlation matrix
corrplot(correlation_matrix, method = "shade",tl.cex = 0.5)
```



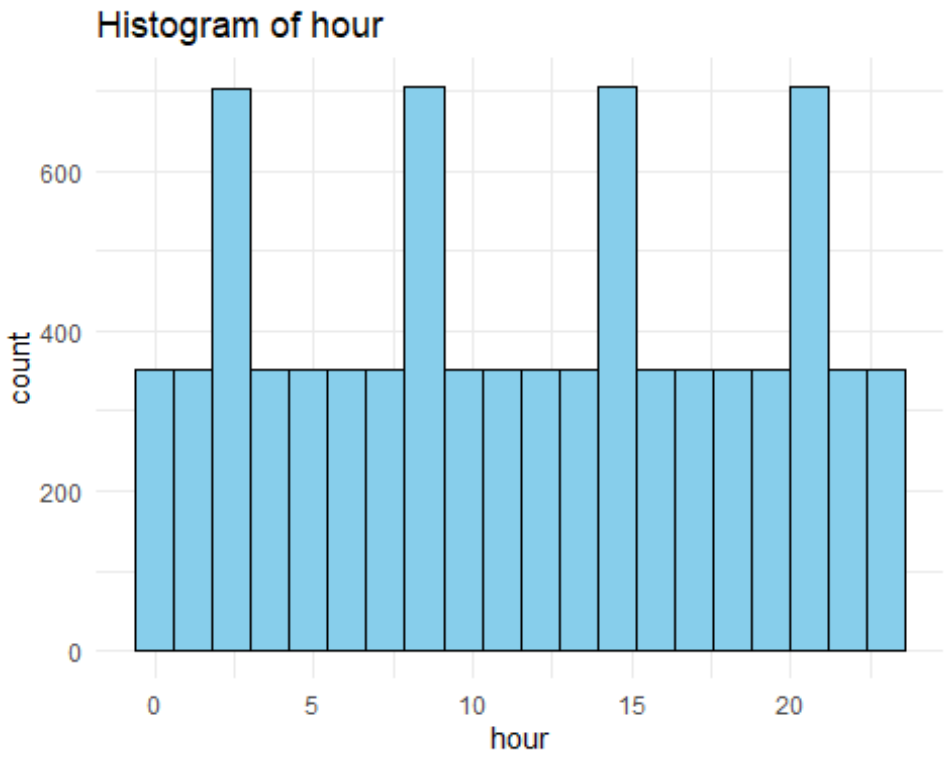

```
# Function to create histograms
multi.hist <- function(data) {
  # Initialize an empty list to store plots
  plots <- list()

  # Loop through each numeric variable
  for (col in names(data)) {
    # Create histogram plot
    hist_plot <- ggplot(data, aes_string(x = col)) +
      geom_histogram(fill = "skyblue", color = "black", bins = 20) +  # Customize fill
and color
      ggtitle(paste("Histogram of", col)) +
      theme_minimal()

    # Add the plot to the list
    plots[[col]] <- hist_plot
  }

  # Arrange plots in a grid
  grid.arrange(grobs = plots, ncol = 3)  # Adjust ncol as needed
```

```r
  # Return the list of plots (optional)
  return(plots)
}

# Call the function with numeric columns of 'bike'
multi.hist(data[, sapply(data, is.numeric)])
```

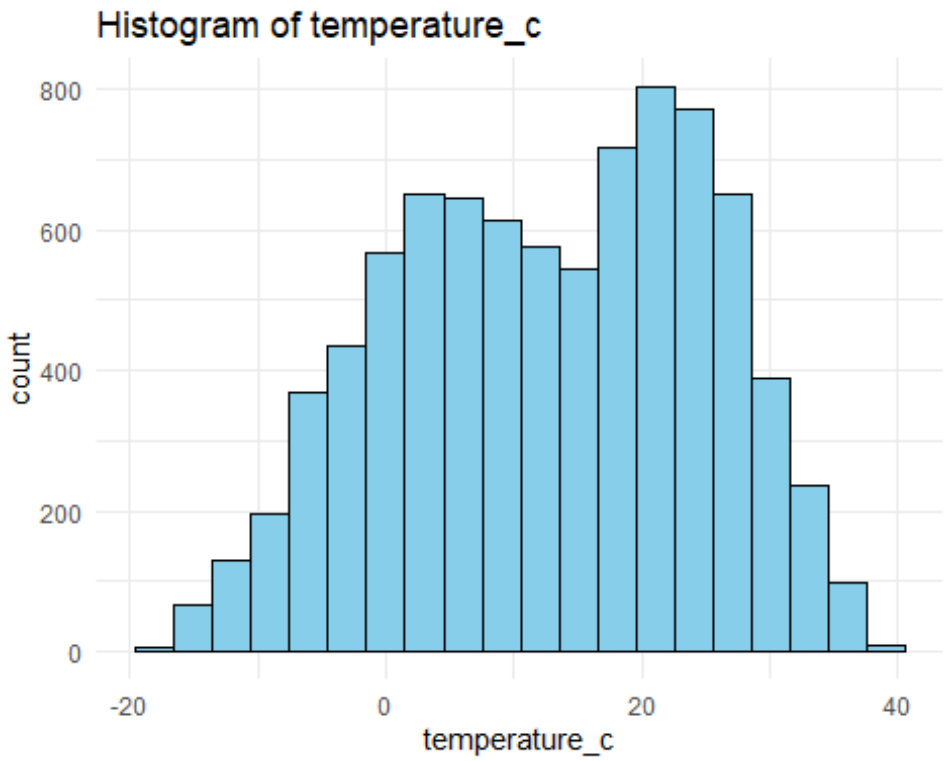
```
## $rented_bike_count
```



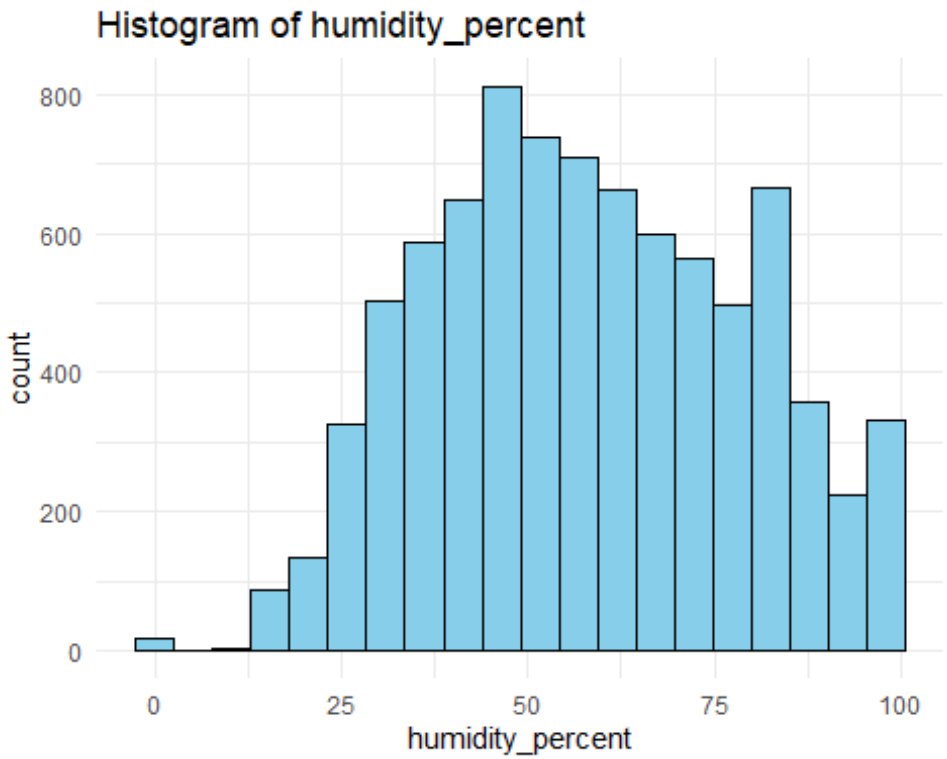

```
##
## $hour
```



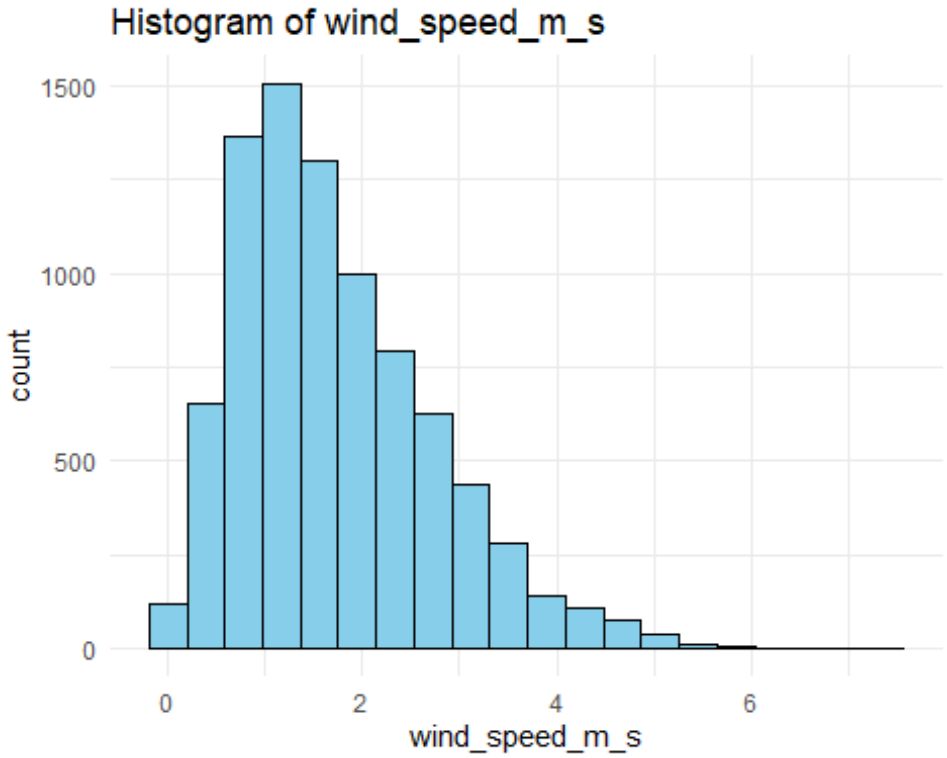

```
##
## $temperature_c
```



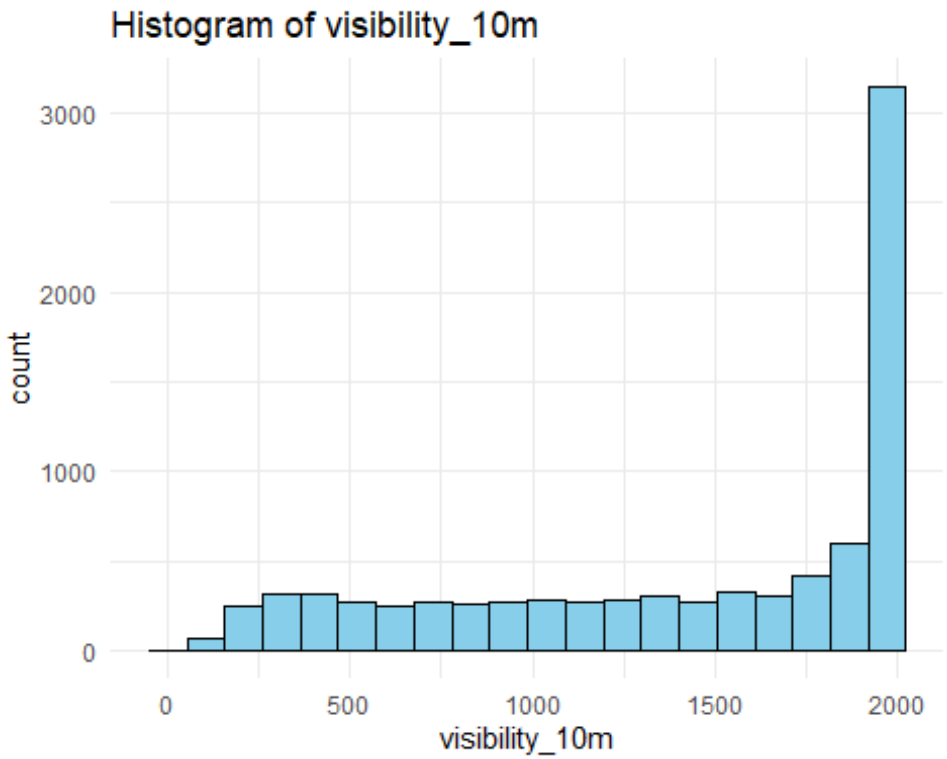

```
##
## $humidity_percent
```



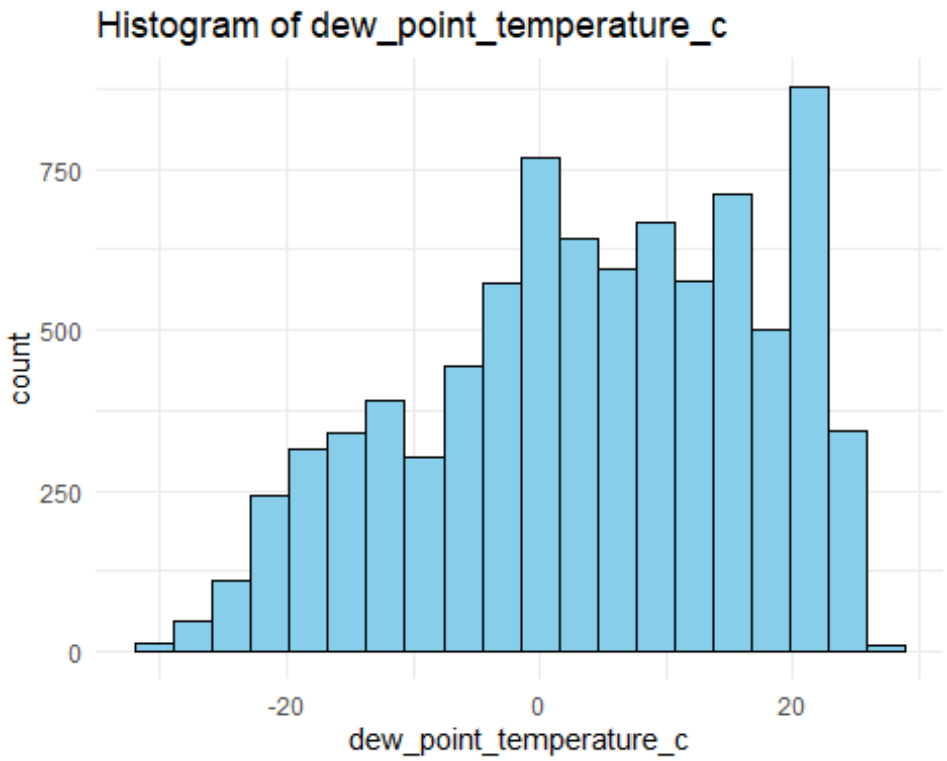

```
##
## $wind_speed_m_s
```



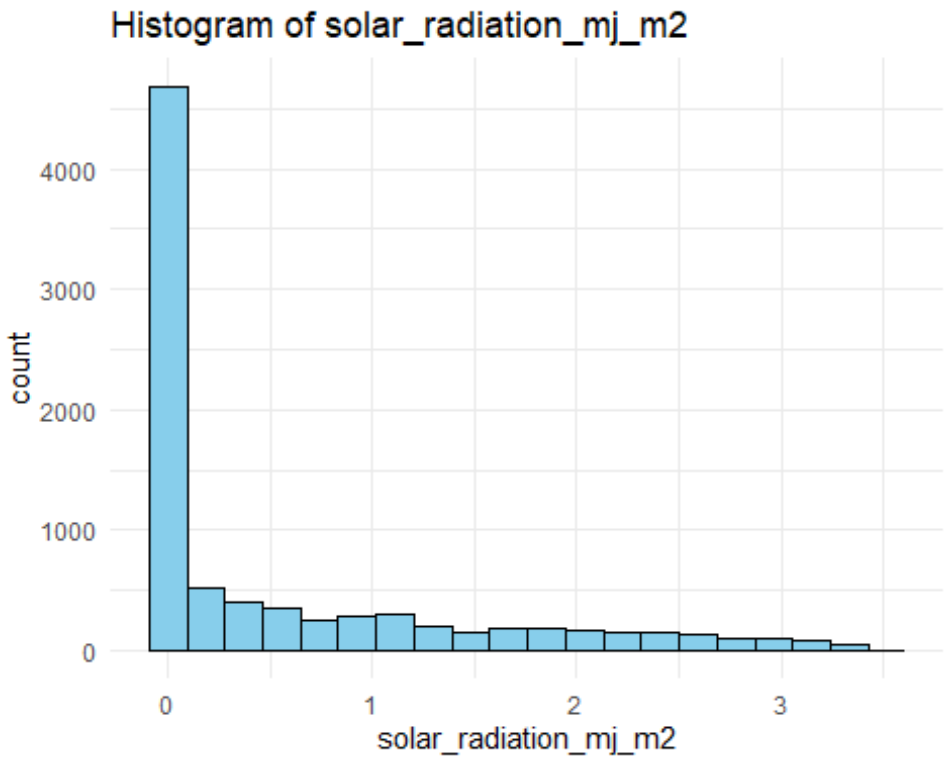

```
##
## $visibility_10m
```



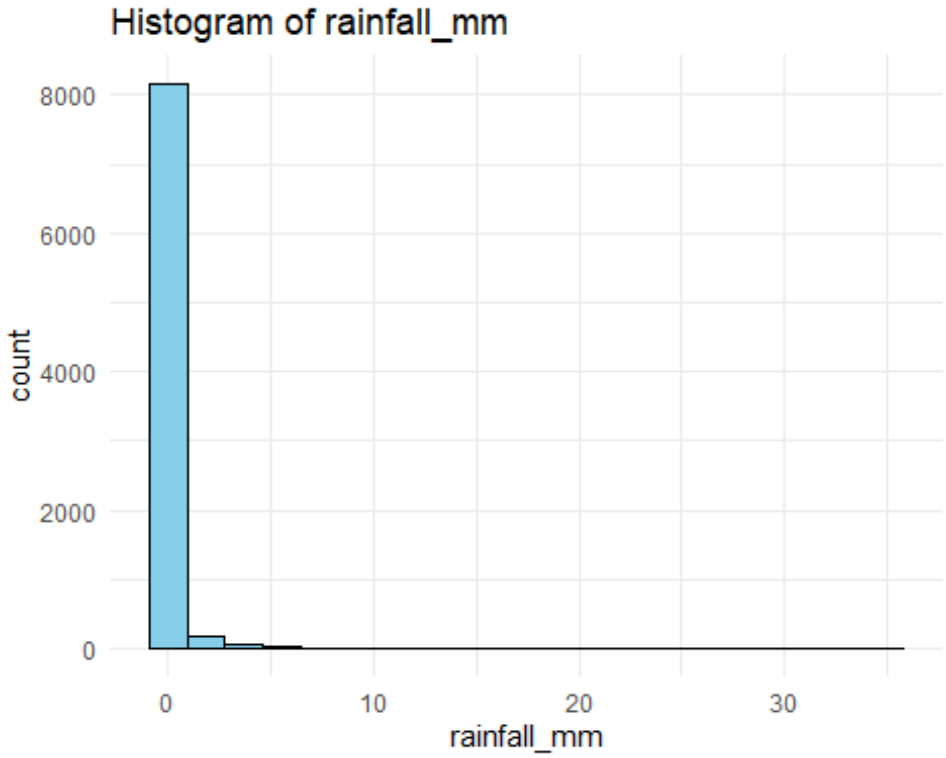

```
##
## $dew_point_temperature_c
```



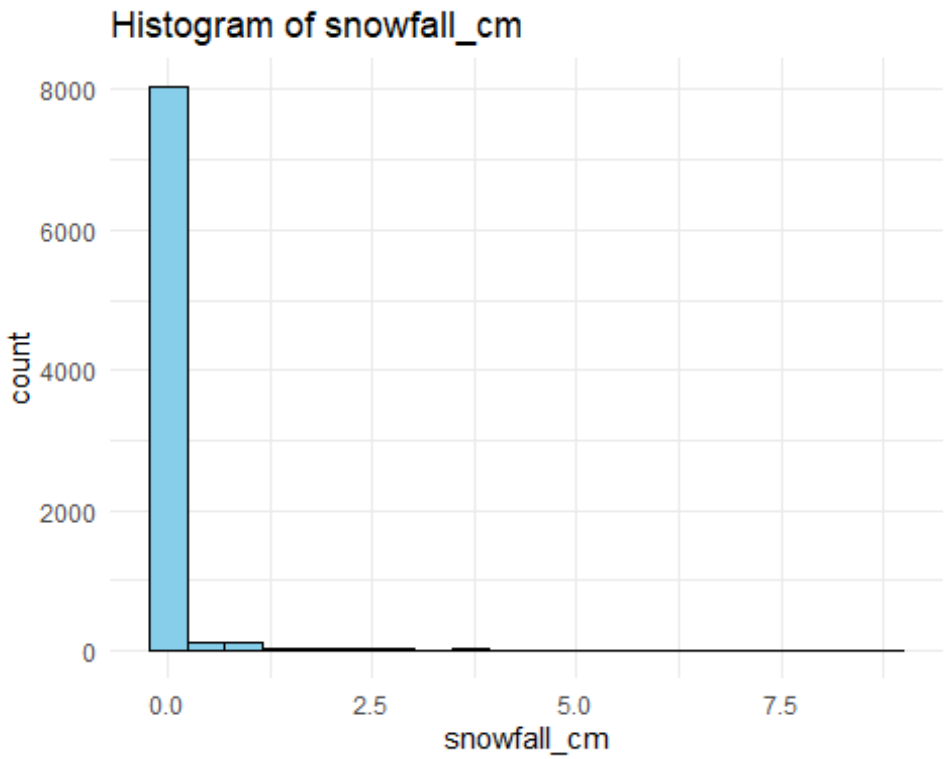

```
##
## $solar_radiation_mj_m2
```



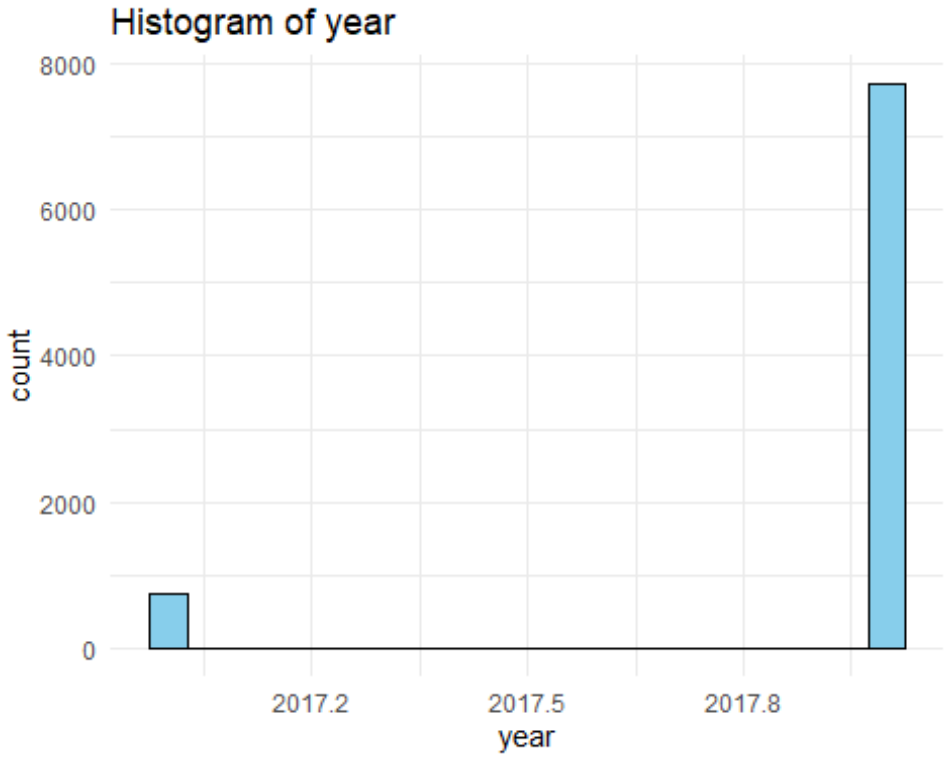

```
##
## $rainfall_mm
```



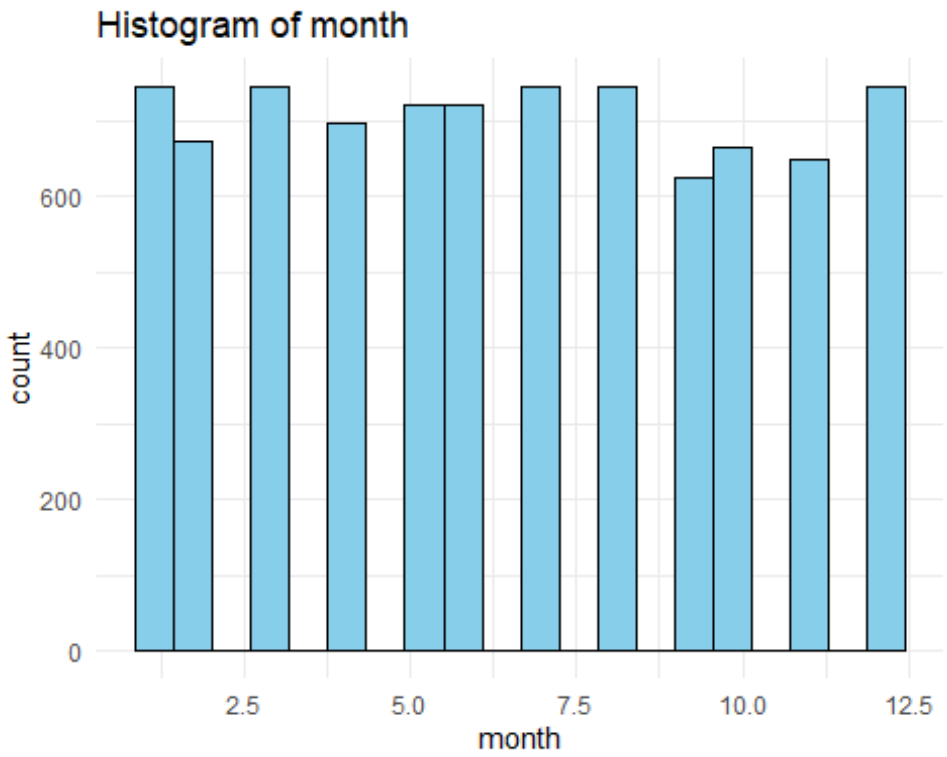

```
##
## $snowfall_cm
```


Histogram of snowfall_cm
count
8000
6000
4000
2000
0
0.0
2.5
5.0
7.5
snowfall_cm

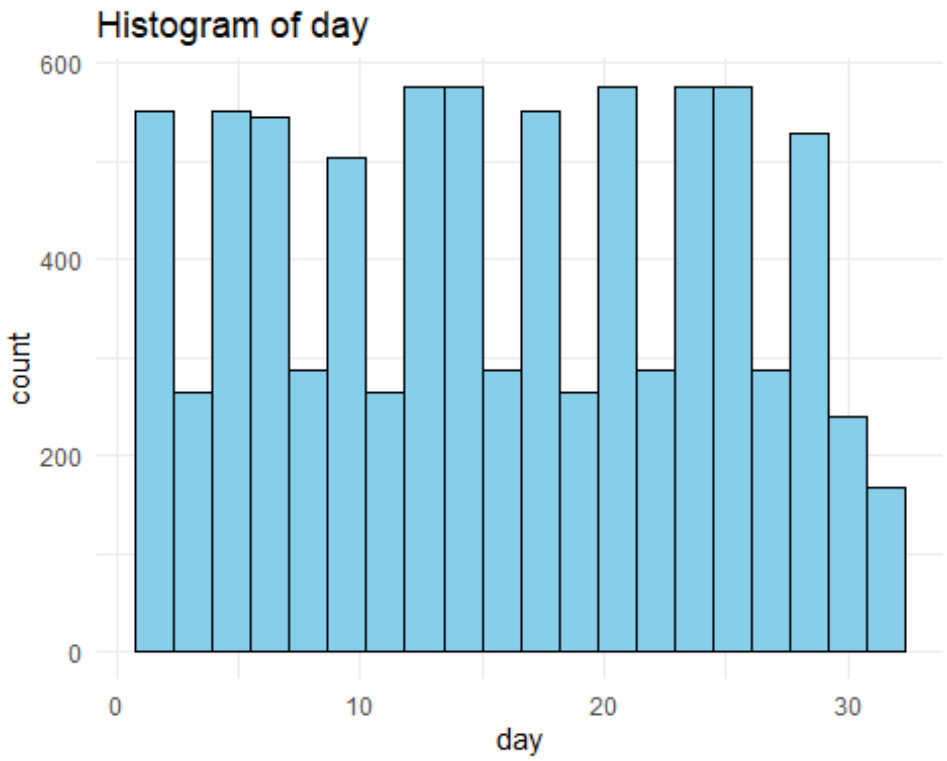
```
##
## $year
```


Histogram of year
count
8000
6000
4000
2000
0
2017.2
2017.5
2017.8
year

```
##
## $month
```




```
##
## $day
```



=> Nhận xét: các thuộc tính không tuân theo phân phối chuẩn.

3.4 Bảng Tổng hợp theo Giờ trong Ngày

```
summary_hour <- data %>%
  group_by(hour) %>%
  summarize(
    total_bike_count = sum(rented_bike_count),
    mean_bike_count = mean(rented_bike_count),
    median_bike_count = median(rented_bike_count),
    max_bike_count = max(rented_bike_count),
    min_bike_count = min(rented_bike_count)
  )
print(summary_hour)
```

```
## # A tibble: 24 × 6
##     hour total_bike_count mean_bike_count median_bike_count max_bike_count
##    <dbl>            <dbl>           <dbl>             <dbl>          <dbl>
##  1     0           197633            561.              532.           1394
##  2     1           155557            442.              412.           1088
##  3     2           110095            313.              277            1254
##  4     3            74216            211.              184.            644
##  5     4            48396            137.              124.            421
##  6     5            50765            144.              136.            383
##  7     6           104961            298.              245             807
##  8     7           221192            627.              443            1629
##  9     8           370731           1050.              767            2495
## 10     9           235784            668.              719            1401
## # ℹ 14 more rows
## # ℹ 1 more variable: min_bike_count <dbl>
```

=> **Nhận xét**: Bảng cho biết tổng số, trung bình, trung vị, giá trị lớn nhất và nhỏ nhất của số lượng xe đạp thuê theo từng giờ trong ngày. Giúp xác định giờ cao điểm và giờ thấp điểm.

3.4.1 Bảng Tổng hợp theo Ngày trong Tuần

```
data$day_of_week <- weekdays(as.Date(data$date, format="%d/%m/%Y"))

summary_day_of_week <- data %>%
  group_by(day_of_week) %>%
  summarize(
    total_bike_count = sum(rented_bike_count),
    mean_bike_count = mean(rented_bike_count),
    median_bike_count = median(rented_bike_count),
    max_bike_count = max(rented_bike_count),
    min_bike_count = min(rented_bike_count)
  )

print(summary_day_of_week)
```

```
## # A tibble: 7 × 6
##   day_of_week total_bike_count mean_bike_count median_bike_count max_bike_count
##   <chr>                 <dbl>           <dbl>             <dbl>          <dbl>
## 1 Friday               950334            776.              600            3365
## 2 Monday               911743            731.              533            3380
## 3 Saturday             885492            728.              560            2497
## 4 Sunday               780194            637.              424            2514
```

```
## 5 Thursday              861999           718.              538           3418
## 6 Tuesday               858596           745.              572.          3556
## 7 Wednesday             923956           770.              570.          3384
## # ℹ 1 more variable: min_bike_count <dbl>
```

=> Kết quả sẽ cho thấy sự khác biệt về nhu cầu thuê xe đạp giữa các ngày trong tuần.

3.4.2 Bảng Tổng hợp theo Mùa

```
summary_season <- data %>%
  group_by(seasons) %>%
  summarize(
    total_bike_count = sum(rented_bike_count),
    mean_bike_count = mean(rented_bike_count),
    median_bike_count = median(rented_bike_count),
    max_bike_count = max(rented_bike_count),
    min_bike_count = min(rented_bike_count)
  )

print(summary_season)
```

```
## # A tibble: 4 × 6
##   seasons total_bike_count mean_bike_count median_bike_count max_bike_count
##   <fct>              <dbl>           <dbl>             <dbl>          <dbl>
## 1 Autumn           1790002            924.              856           3298
## 2 Spring           1611909            746.              599           3251
## 3 Summer           2283234           1034.              906.          3556
## 4 Winter            487169            226.              203            937
## # ℹ 1 more variable: min_bike_count <dbl>
```

=> Kết quả sẽ cho thấy sự khác biệt về nhu cầu thuê xe đạp giữa các mùa. Ví dụ, mùa hè có thể có nhu cầu cao hơn mùa đông.

3.4.3 Bảng Tổng hợp theo Ngày Lễ

```
summary_holiday <- data %>%
  group_by(holiday) %>%
  summarize(
    total_bike_count = sum(rented_bike_count),
    mean_bike_count = mean(rented_bike_count),
    median_bike_count = median(rented_bike_count),
    max_bike_count = max(rented_bike_count),
    min_bike_count = min(rented_bike_count)
  )

print(summary_holiday)
```

```
## # A tibble: 2 × 6
##   holiday    total_bike_count mean_bike_count median_bike_count max_bike_count
##   <fct>                 <dbl>           <dbl>             <dbl>          <dbl>
## 1 Holiday              215895            529.              259           2400
## 2 No Holiday          5956419            739.              561           3556
## # ℹ 1 more variable: min_bike_count <dbl>
```

Kết quả sẽ cho thấy sự khác biệt về nhu cầu thuê xe đạp giữa các ngày lễ và ngày thường.

3.4.4 Bảng Tổng hợp theo Nhiệt độ

```
data$temp_range <- cut(data$temperature_c, breaks = seq(floor(min(data$temperature_c)),
ceiling(max(data$temperature_c)), by = 5))

summary_temp <- data %>%
  group_by(temp_range) %>%
  summarize(
    total_bike_count = sum(rented_bike_count),
    mean_bike_count = mean(rented_bike_count),
    median_bike_count = median(rented_bike_count),
    max_bike_count = max(rented_bike_count),
    min_bike_count = min(rented_bike_count)
  )
data <- data |> dplyr::select(-temp_range)
print(summary_temp)
```

```
## # A tibble: 12 × 6
##    temp_range total_bike_count mean_bike_count median_bike_count max_bike_count
##    <fct>                 <dbl>           <dbl>             <dbl>          <dbl>
##  1 (-18,-13]             10113            119.               94            469
##  2 (-13,-8]              42724            161.              136.           812
##  3 (-8,-3]              123377            195.              168.          1516
##  4 (-3,2]               227180            255.              216           1883
##  5 (2,7]                440565            408.              341           2000
##  6 (7,12]               627171            614.              555           2254
##  7 (12,17]              775206            829.              814           2857
##  8 (17,22]             1245008            967.              931           3130
##  9 (22,27]             1421600           1212.             1117           3556
## 10 (27,32]              937282           1188.             1078           3418
## 11 (32,37]              305711           1076.              821           2965
## 12 <NA>                  16377            780.              665           1929
## # ℹ 1 more variable: min_bike_count <dbl>
```

Kết quả sẽ cho thấy mối quan hệ giữa nhiệt độ và số lượng xe đạp thuê. Điều này có thể giúp dự đoán nhu cầu thuê xe đạp dựa trên dự báo thời tiết.

=> Những bảng tổng hợp và biểu đồ này sẽ cung cấp cái nhìn tổng quan và chi tiết về các yếu tố ảnh hưởng đến nhu cầu thuê xe đạp, giúp đưa ra các quyết định quản lý và phát triển dịch vụ hiệu quả hơn.