Homework #1

Group 4

5/10/2020

First, let's import the required libraries

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:GGally':
##
##     nasa
```

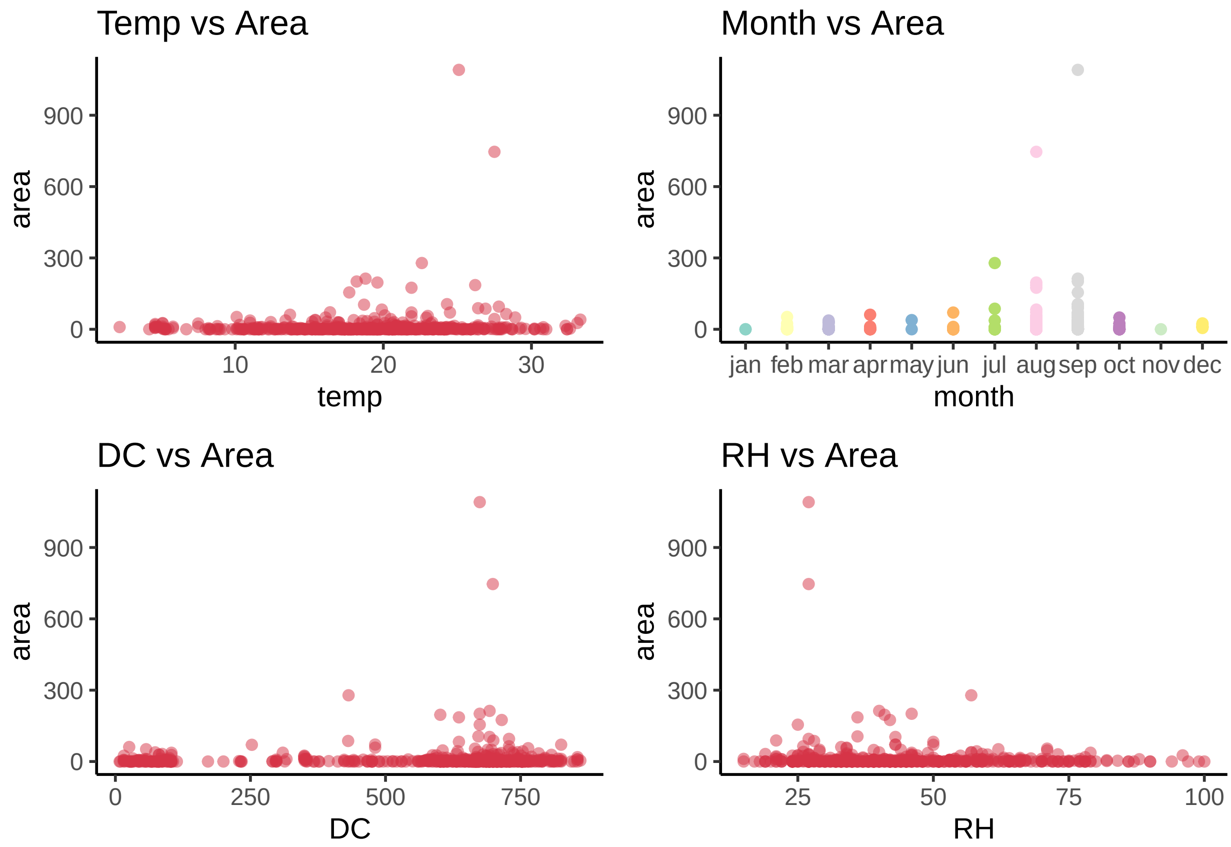
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Problem 1

(a) Plot area vs.temp, area vs. month, area vs. DC, area vs. RH for January through December combined in one graph. Hint: Place area on Y axis and use 2x2 matrix to place the plots adjacent to each other.

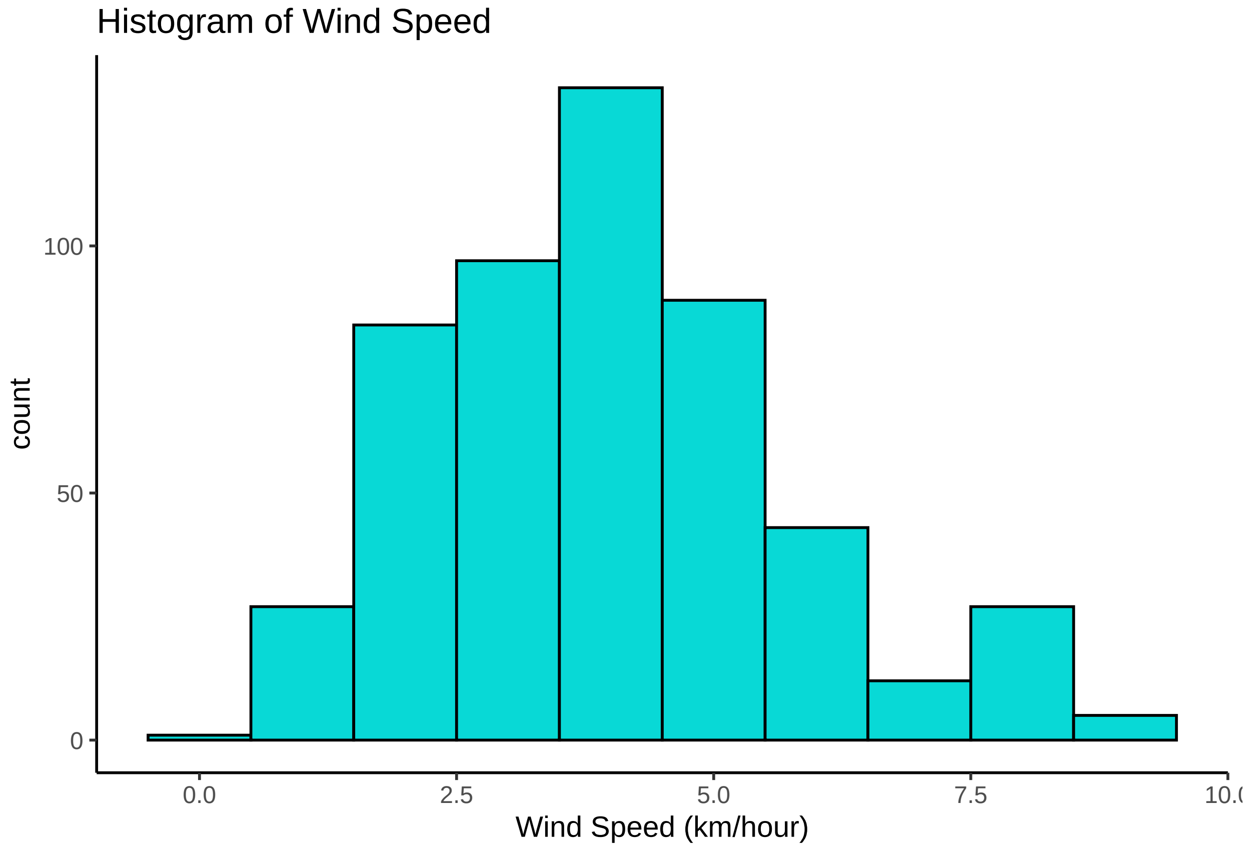
```r
# Import the Dataframe
forestfires <- data.frame(read.csv("./data/forestfires.csv"),
  stringsAsFactors = FALSE
)

# Convert the month column into factors and sort from Jan-Dec
forestfires$month <- factor(forestfires$month,
  levels = c(
    "jan", "feb", "mar",
    "apr", "may", "jun",
    "jul", "aug", "sep",
    "oct", "nov", "dec"
  )
)

# Create 4 scatter plots
p1 <- ggplot(forestfires, aes(temp, area)) +
  geom_point(color = "#d63447", alpha = 0.5) +
  ggtitle("Temp vs Area") +
  theme_classic()
p2 <- ggplot(forestfires, aes(month, area, color = month)) +
  geom_point() +
  scale_color_brewer(palette = "Set3") +
  theme_classic() +
  theme(legend.position = "none") +
  ggtitle("Month vs Area")
p3 <- ggplot(forestfires, aes(DC, area)) +
  geom_point(color = "#d63447", alpha = 0.5) +
  ggtitle("DC vs Area") +
  theme_classic()
p4 <- ggplot(forestfires, aes(RH, area)) +
  geom_point(color = "#d63447", alpha = 0.5) +
  ggtitle("RH vs Area") +
  theme_classic()

# Arrange plots P1-P4 into a 2x2 grid
fig <- ggarrange(p1, p2, p3, p4, ncol = 2, nrow = 2)


plot(fig)
```



```r
rm(list = c("p1", "p2", "p3", "p4", "fig"))
```

(b) Plot the histogram of wind speed (km/h).

```r
# Create the Wind-Historgram
wind_hist <- ggplot(forestfires, aes(wind)) +
  geom_histogram(bins = 10, fill = "#08d9d6", color = "black") +
  theme_classic() +
  ggtitle("Histogram of Wind Speed") +
  labs(x = "Wind Speed (km/hour)")

# Plot
plot(wind_hist)
```



```r
rm(wind_hist)
```

(c) Compute the summery statistics (min, 1Q, mean, median, 3Q, max,) of part b.

```r
# Calculate the Quantiles
quantiles <- quantile(forestfires$wind)

# Print
cat("Minimum Wind Speed is :", quantiles[[1]], "\n")
```

```
## Minimum Wind Speed is : 0.4
```

```r
cat("1st Quantile of Wind Speed is :", quantiles[[2]], "\n")
```

```
## 1st Quantile of Wind Speed is : 2.7
```

```r
cat("Mean Wind Speed is :", mean(forestfires$wind), "\n")
```

```
## Mean Wind Speed is : 4.017602
```

```
cat("Median Wind Speed is :", quantiles[[3]], "\n")
```

```
## Median Wind Speed is : 4
```

```
cat("3rd Quartile of Wind Speed is :", quantiles[[4]], "\n")
```

```
## 3rd Quartile of Wind Speed is : 4.9
```

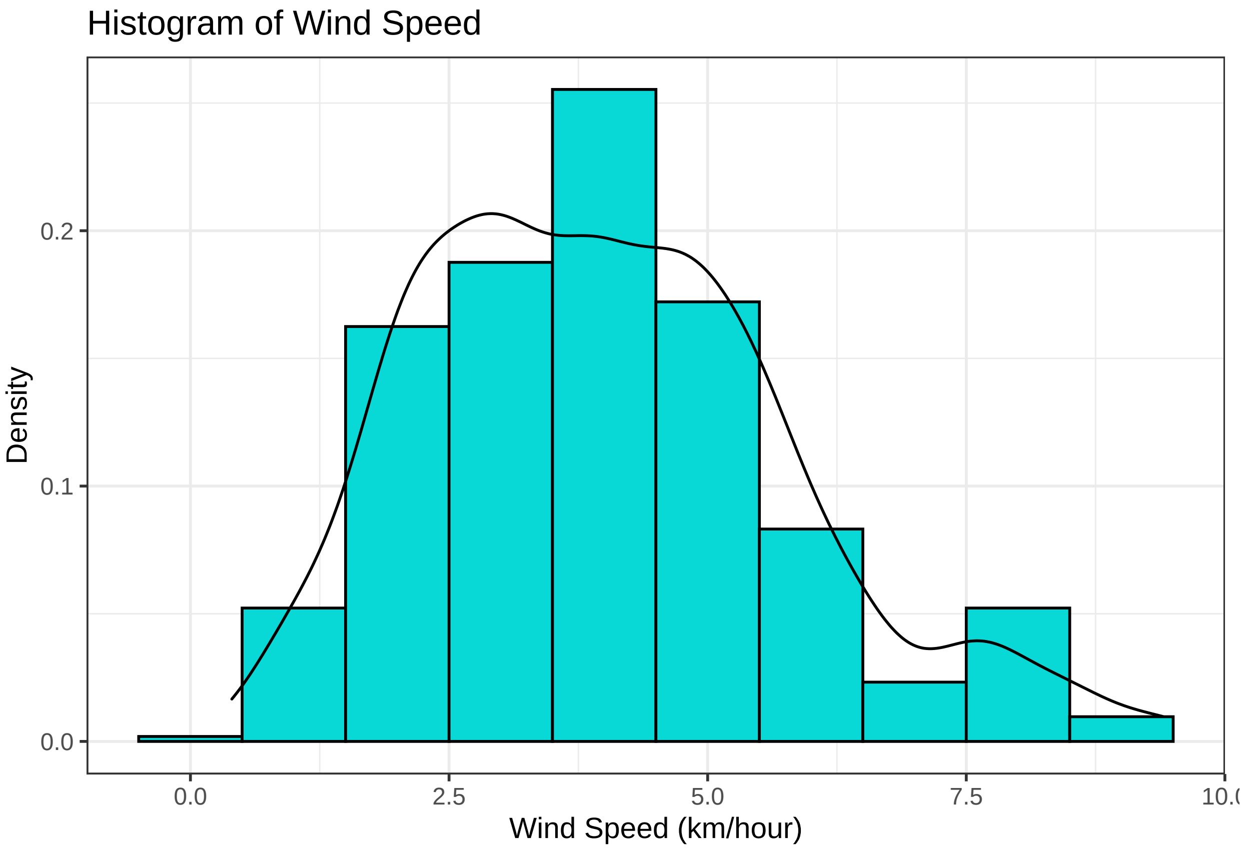
```
cat("Maximum Wind Speed is :", quantiles[[5]], "\n")
```

```
## Maximum Wind Speed is : 9.4
```

```
rm(quantiles)
```

(d) Add a density line to the histogram in part b.

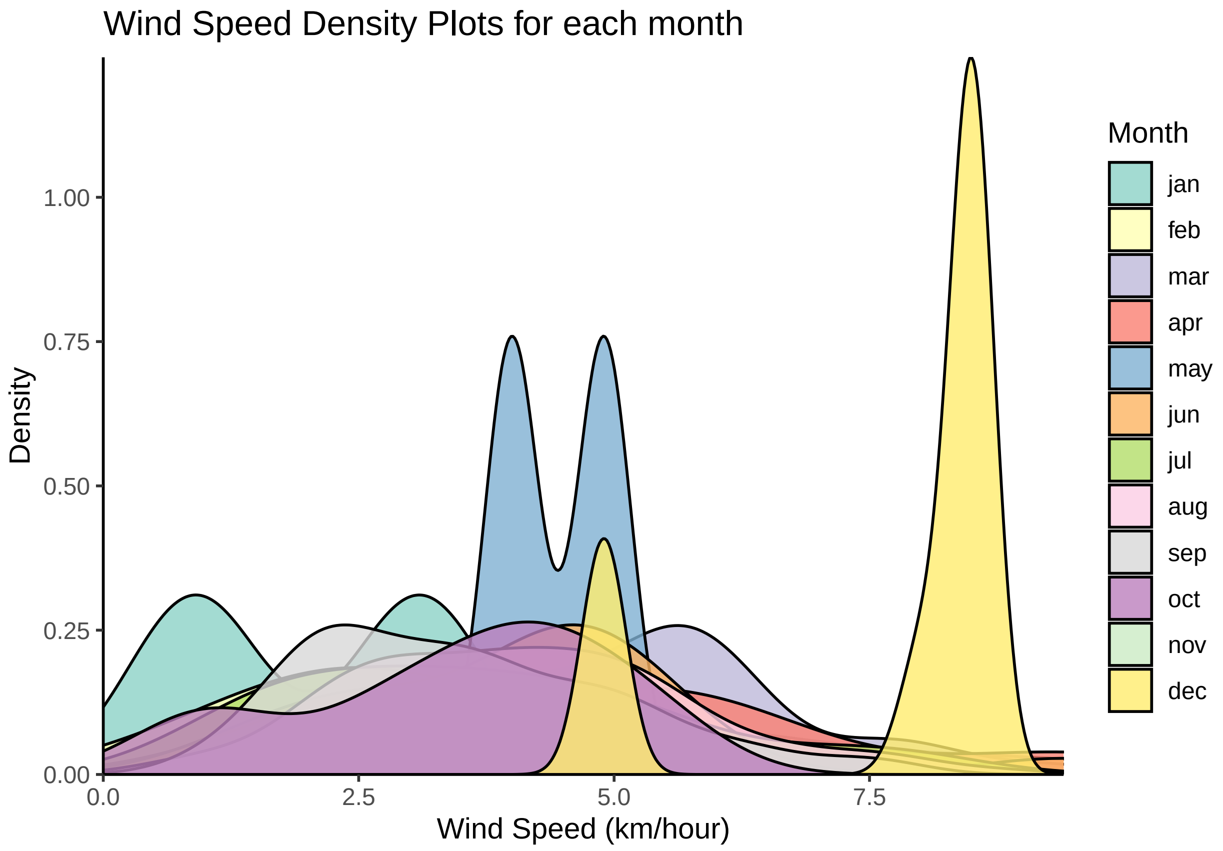
```
# Create the Histogram and Density Plot
wind_hist_plus_density <- ggplot(forestfires, aes(x = wind, y = ..density..)) +
  geom_histogram(bins = 10, colour = "black", fill = "#08d9d6") +
  geom_density(aes(y = ..density..), color = "black") +
  ylab("Density") +
  xlab("Wind Speed (km/hour)") +
  ggtitle("Histogram of Wind Speed") +
  theme_bw()

plot(wind_hist_plus_density)
```



```
rm(wind_hist_plus_density)
```

(e) Plot the wind speed density function of all months in one plot. Use different colors for different months in the graph to interpret your result clearly. [Hint: use ggplot + geom_density or qplot(geom=density)]

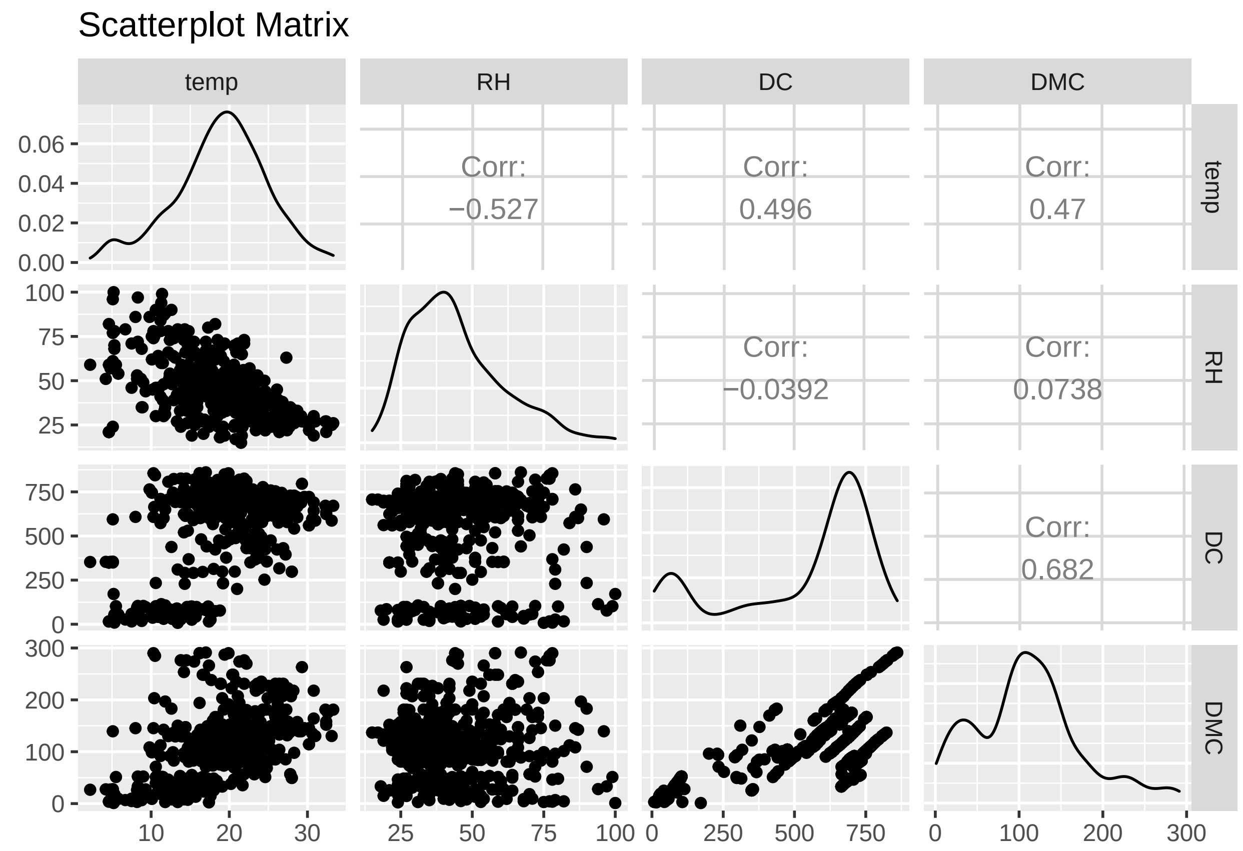
```
# Create month wise density plot
month_density_plot <- ggplot(forestfires, aes(
  x = wind,
  y = ..density..,
  fill = month
)) +
  geom_density(alpha = 0.8) +
  scale_fill_brewer(palette = "Set3") +
  theme_classic() +
  ggtitle("Wind Speed Density Plots for each month") +
  labs(x = "Wind Speed (km/hour)", y = "Density", fill = "Month") +
  scale_x_continuous(expand = c(0, 0), limits = c(0, NA)) +
  scale_y_continuous(expand = c(0, 0), limits = c(0, NA))

plot(month_density_plot)
```



```
rm(month_density_plot)
```

(f) Plot the scatter matrix for temp, RH, DC and DMC. How would you interpret the result in terms of correlation among these data?

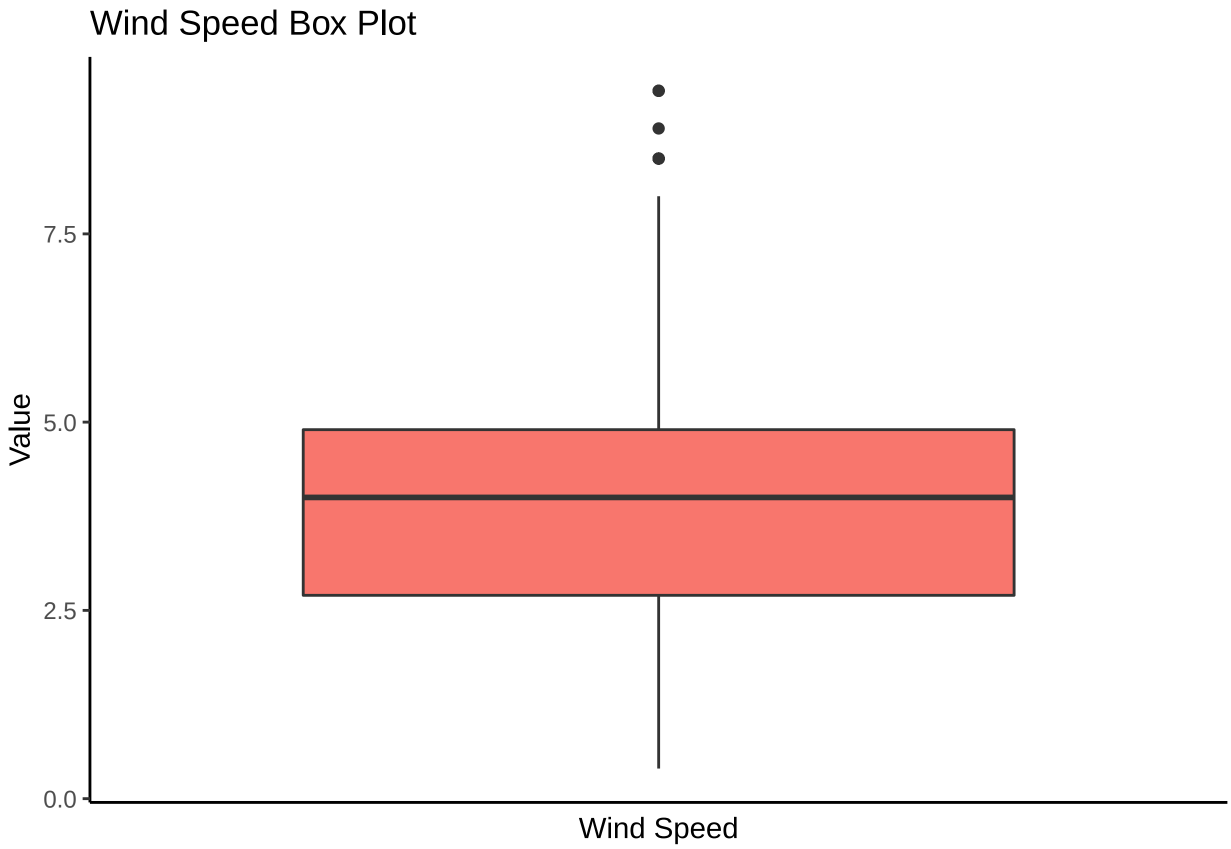
```
# Plot ScatterMatrix
ggpairs(forestfires,
  title = "Scatterplot Matrix",
  columns = c("temp", "RH", "DC", "DMC")
)
```



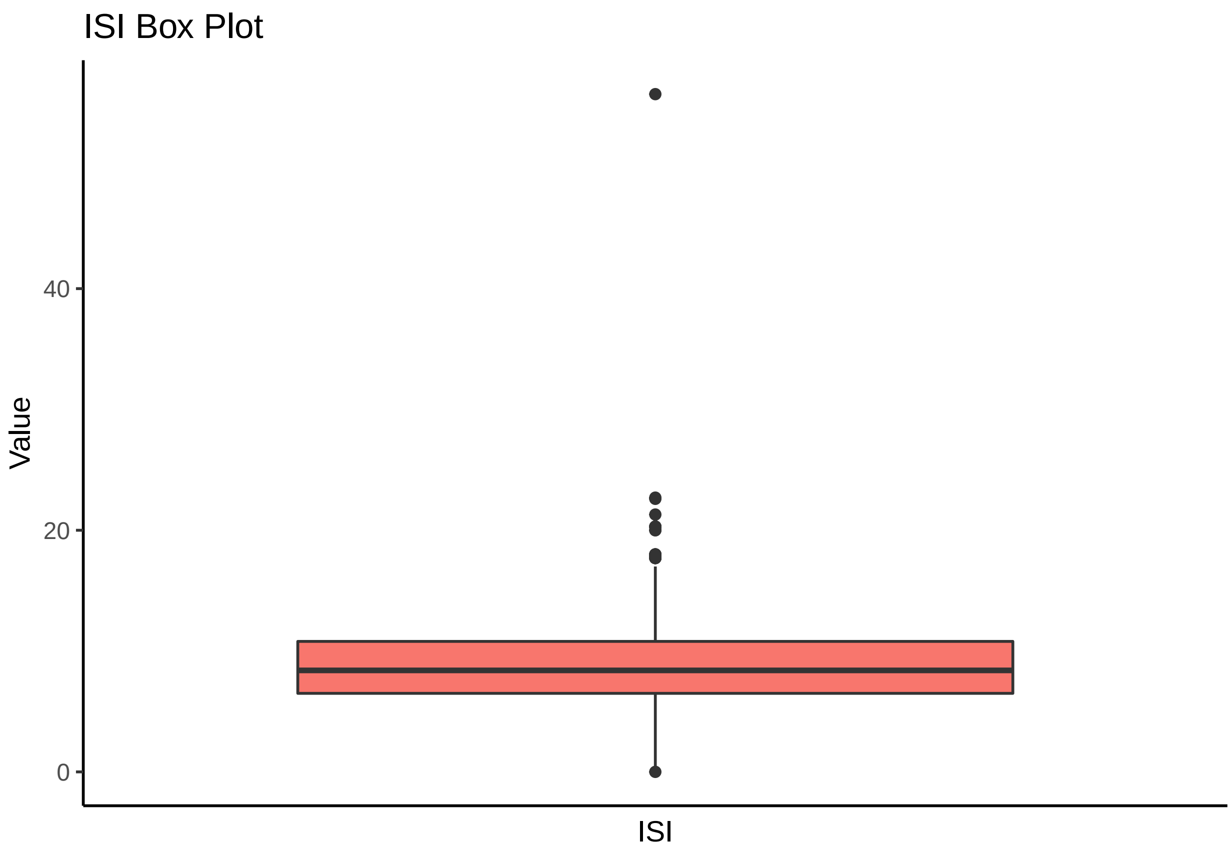
Looking at the Scatter Matrix, we can make the following conclusions - (1) Temp and RH are negatively correlated with corr. coeff of -0.527. (2) Temp and DC are positively correlated with corr. coeff of 0.496. (3) Temp and DC are positively correlated with corr. coeff of 0.47. (4) RH has no correlation with DC and DMS whatsoever. (5) DC and DMC are very strongly correlated with a corr. coeff of 0.682.

(g) Create boxplot for wind, ISI and DC. Are there any anomalies/outliers? Interpret your result.

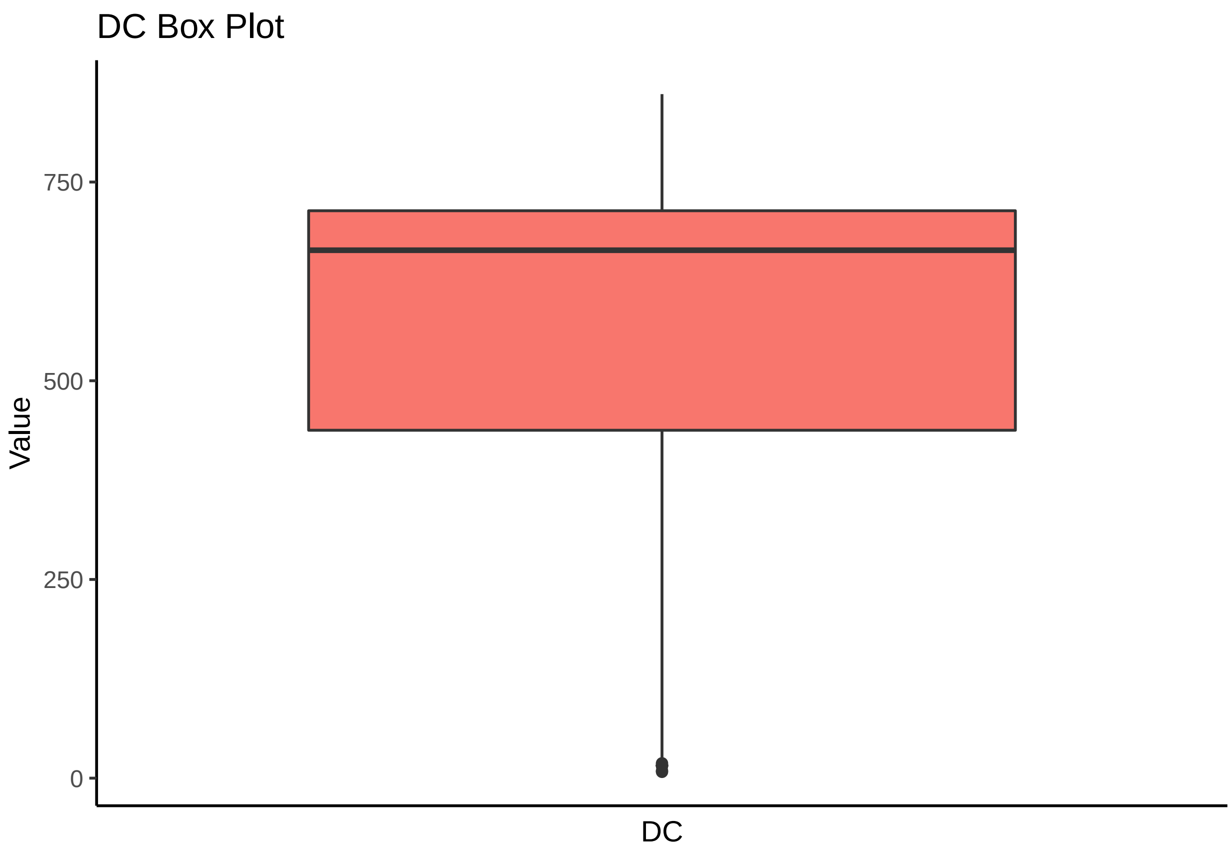
```
# Create a temporary dataframe and reshape it
df <- forestfires %>% select(wind, ISI, DC)
suppressMessages({
  df <- melt(df)
})

# Plot 3 boxplots for Wind, ISI, and DC
ggplot(df %>% filter(variable == "wind"), aes(x = variable, y = value, fill = variable)) +
  geom_boxplot() +
  theme_classic() +
  theme(
    legend.position = "none",
    axis.ticks.x = element_blank(),
    axis.text.x = element_blank()
  ) +
  ggtitle("Wind Speed Box Plot") +
  labs(x = "Wind Speed", y = "Value")
```



```r
ggplot(df %>% filter(variable == "ISI"), aes(x = variable, y = value, fill = variable)) +
  geom_boxplot() +
  theme_classic() +
  theme(
    legend.position = "none",
    axis.ticks.x = element_blank(),
    axis.text.x = element_blank()
  ) +
  ggtitle("ISI Box Plot") +
  labs(x = "ISI", y = "Value")
```



```r
ggplot(df %>% filter(variable == "DC"), aes(x = variable, y = value, fill = variable)) +
  geom_boxplot() +
  theme_classic() +
  theme(
    legend.position = "none",
    axis.ticks.x = element_blank(),
    axis.text.x = element_blank()
  ) +
  ggtitle("DC Box Plot") +
  labs(x = "DC", y = "Value")
```

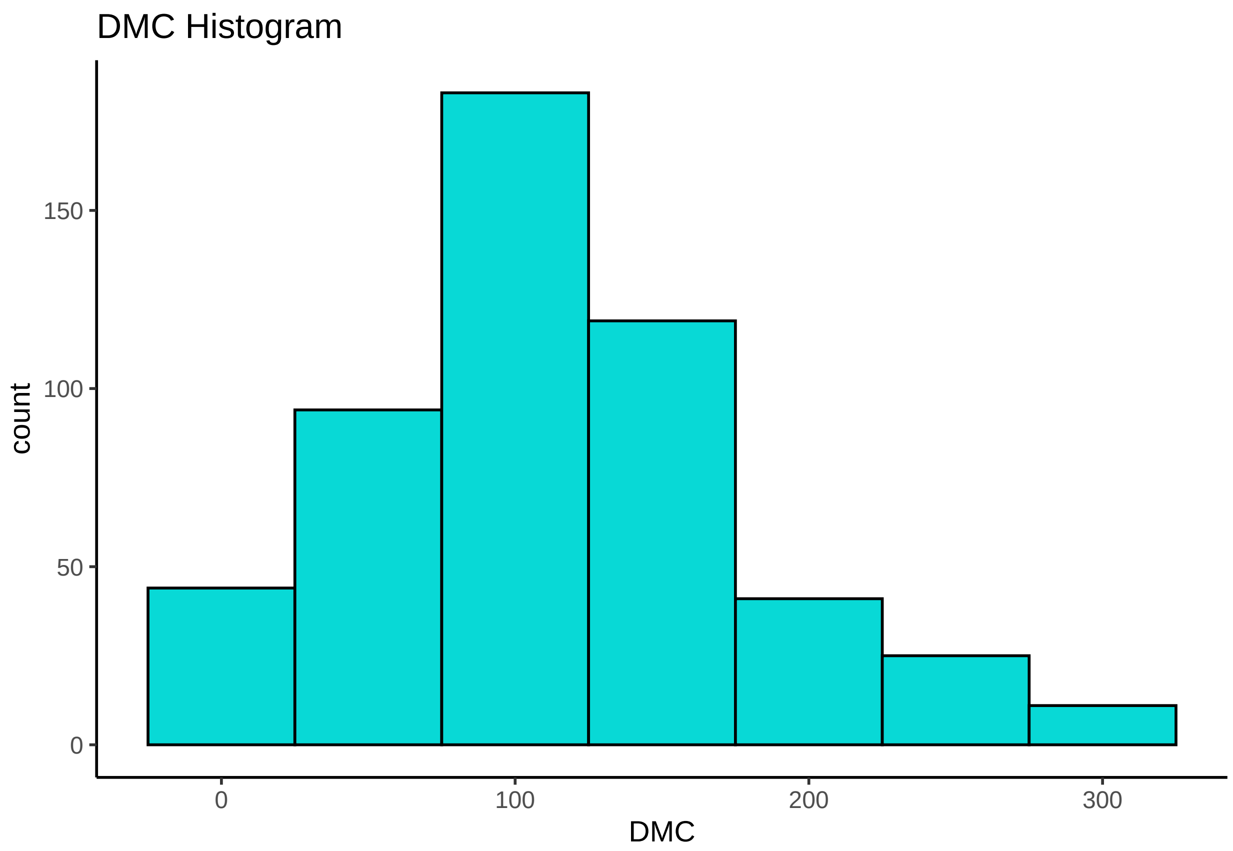


```r
rm(df)
```

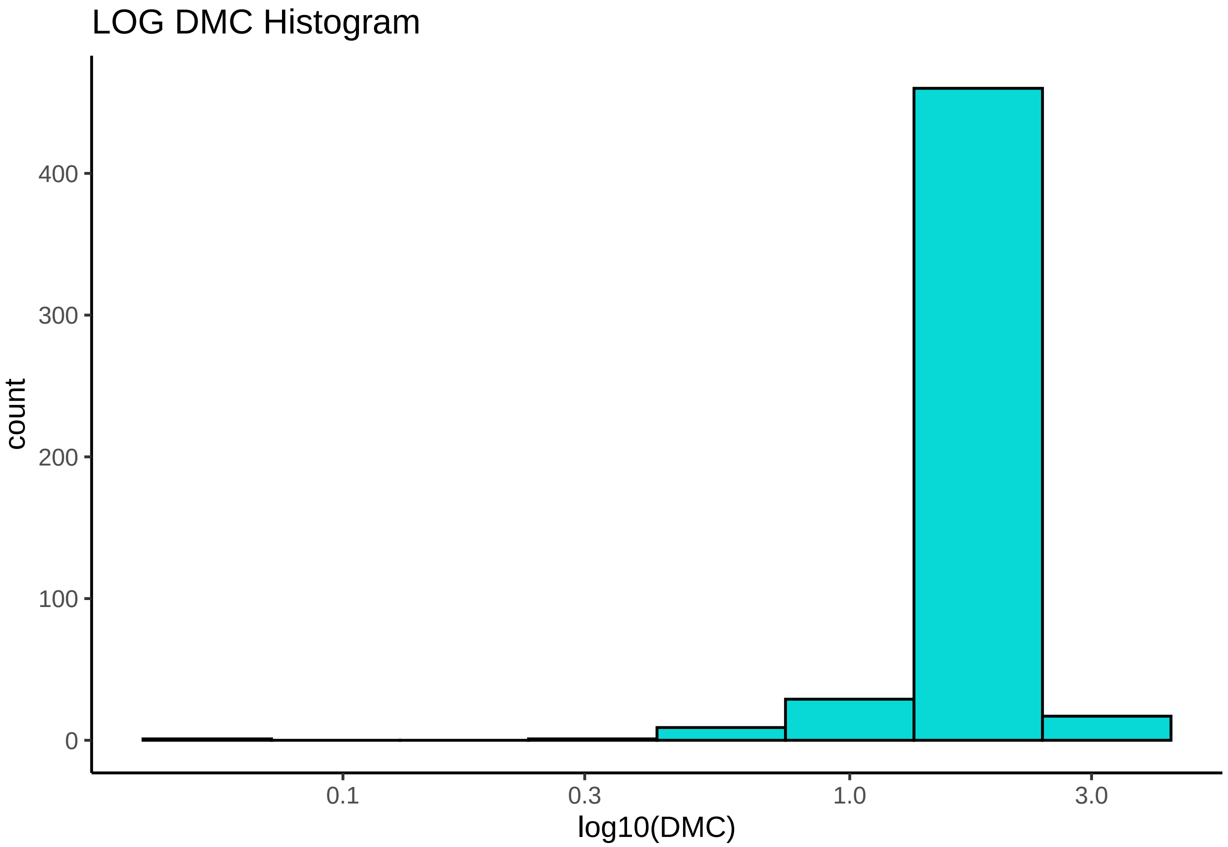
We can make the following observations, (1) Wind has only 3 outliers, the rest of the values are between 0.4 - 7.5 approximately. (2) ISI has quite a few outliers exceeding the value 18. The biggest outlier is of a value 56.1 (3) DC has only 2 outliers which are below the value 15.

(h) Create the histogram of DMC. Create the histogram of log of DMC. Compare the result and explain your answer.

```r
# Plot normal histogram
ggplot(forestfires, aes(DMC)) +
  geom_histogram(binwidth = 50, color = "black", fill = "#08d9d6") +
  theme_classic() +
  ggtitle("DMC Histogram")
```



```
# Plot log scale histogram
ggplot(forestfires, aes(log10(DMC))) +
  geom_histogram(bins = 8, color = "black", fill = "#08d9d6") +
  theme_classic() +
  ggtitle("LOG DMC Histogram") +
  scale_x_log10()
```



We can make the following observations - (1) The normal histogram of DMC tells us that it follows a normal distribution with most values of the range 100-150. It is slightly left skewed.

(2) The Histogram of log(DMC) supports the observations made in (1). Since most of the DMC values are between 100-300, it makes sense that the log(DMC) histogram shows the majority of values at 2 (since log10(X>100) lies in range 1-2).