Kritik-Updated-Assignment

The data measure the air quality every day through May to September

```r
dat <- read.table("Student_Dataset_Assignment.txt", header = TRUE, sep = "\t")
data (airquality)
```

Load and quick examine the data

```r
dat <- airquality
str(dat)
```

```
'data.frame':   153 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

```r
summary(dat)
```

```
     Ozone           Solar.R           Wind             Temp      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
 NA's   :37       NA's   :7      
     Month            Day      
 Min.   :5.000   Min.   : 1.0  
 1st Qu.:6.000   1st Qu.: 8.0  
 Median :7.000   Median :16.0  
 Mean   :6.993   Mean   :15.8  
 3rd Qu.:8.000   3rd Qu.:23.0  
 Max.   :9.000   Max.   :31.0  
```

So, there are some NAs in columns Ozone and Solar. R

Check missing values carefully

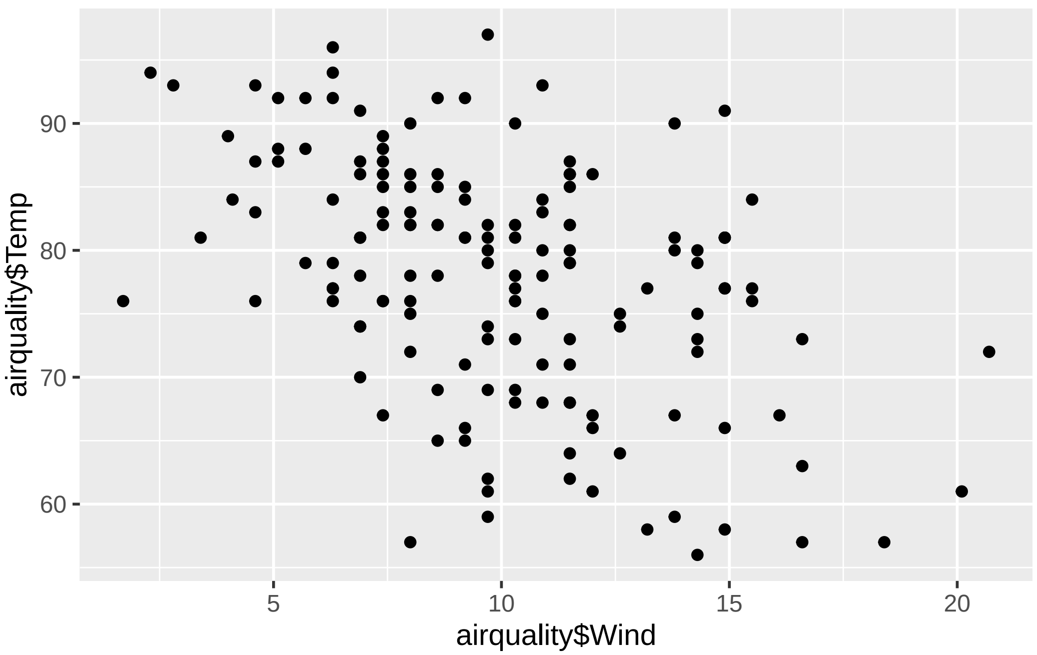
```
colSums(is.na(dat))
```

```
  Ozone Solar.R    Wind    Temp   Month     Day 
     37       7       0       0       0       0 
```

So, there are 37 NAs in column Ozone and 7 NAs in column Solar.R

Examine Variables Wind vs Temp

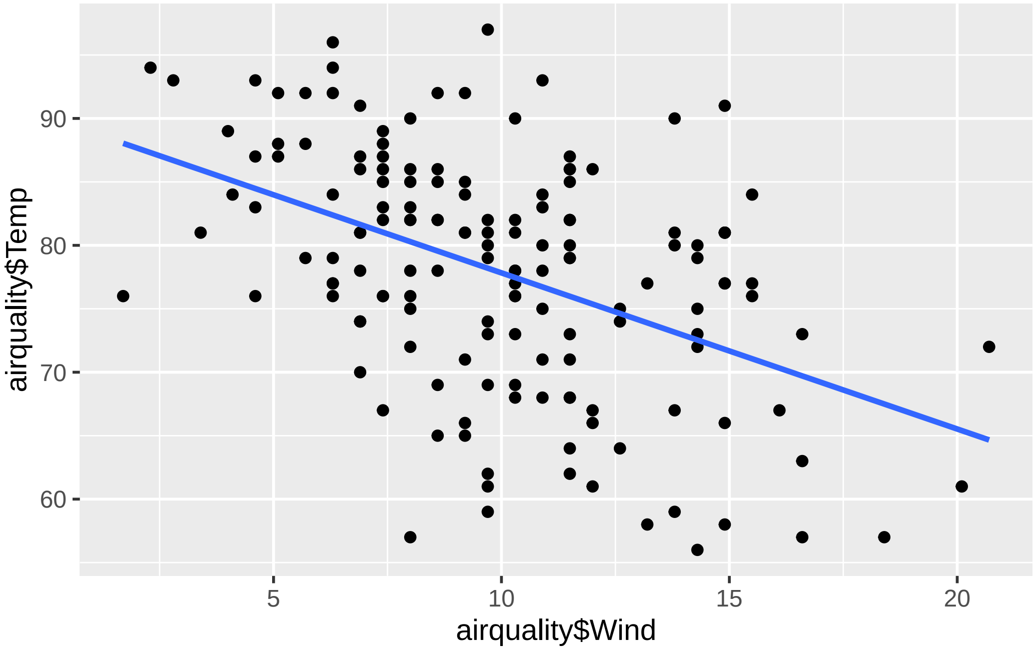
```
library(tidyverse)
library(ggplot2)
#graph data
qplot(airquality$Wind, airquality$Temp)
```



```r
cor(airquality$Wind, airquality$Temp)
```

```
[1] -0.4579879
```

```r
qplot(airquality$Wind, airquality$Temp) + geom_smooth(method = "lm", se = FALSE)
```



```r
model <- lm(Temp ~ Wind, data = airquality)
summary(model)
```

```

Call:
lm(formula = Temp ~ Wind, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.291  -5.723   1.709   6.016  19.199 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  90.1349     2.0522  43.921  < 2e-16 ***
Wind         -1.2305     0.1944  -6.331 2.64e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.442 on 151 degrees of freedom
Multiple R-squared:  0.2098,	Adjusted R-squared:  0.2045 
F-statistic: 40.08 on 1 and 151 DF,  p-value: 2.642e-09
```

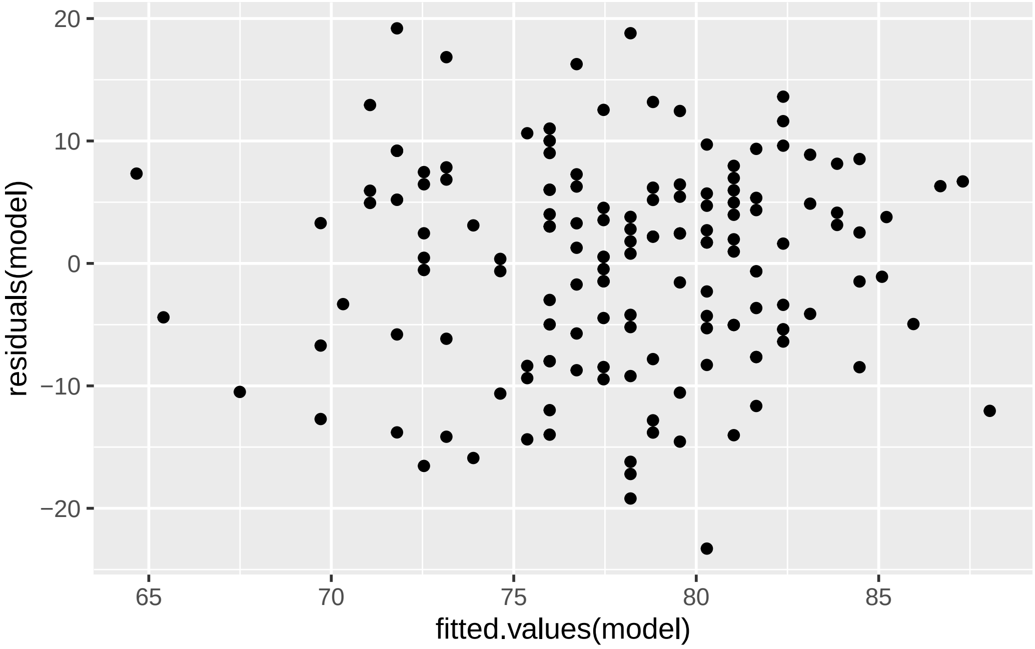
```
coef(model)
```

```
(Intercept)        Wind
  90.134867   -1.230479
```

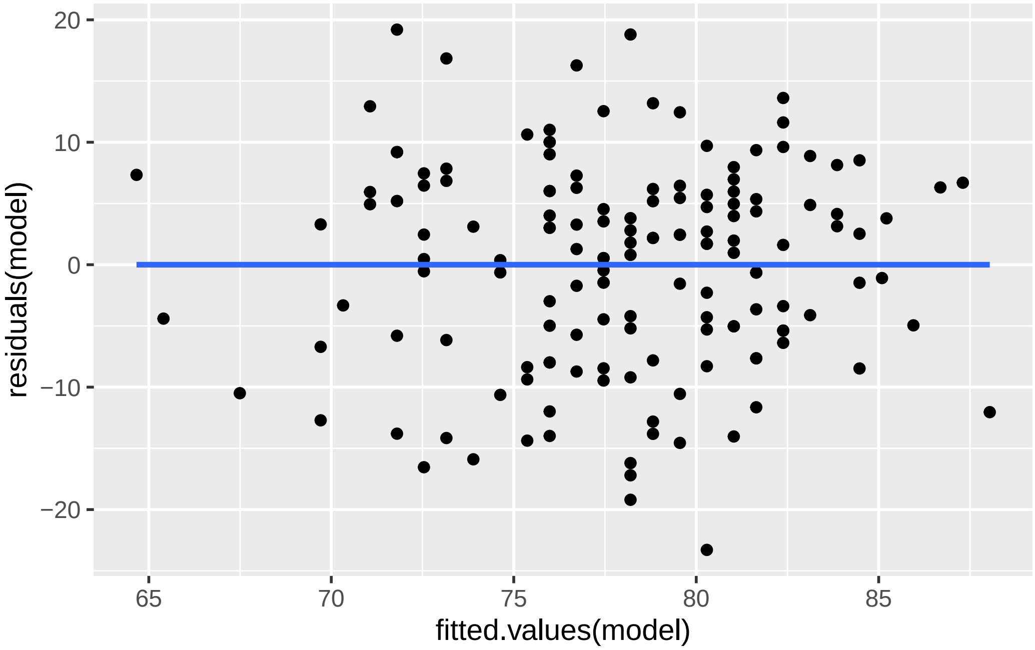
For every unit that wind increases, the temperature will decrease by about 1.23 units.

Residuals vs Fitted Values

```
qplot(fitted.values(model), residuals(model))
```



```
qplot(fitted.values(model), residuals(model))+geom_smooth(method = "lm", se = FALSE)
```

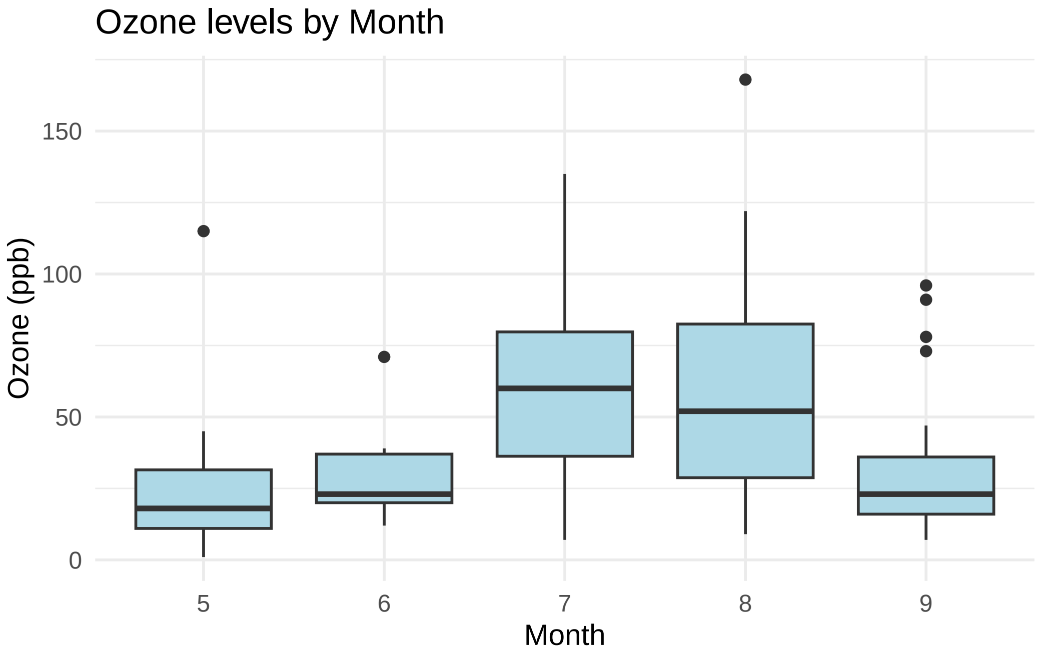


There's no pattern to this data, all the residuals scatter around the 0 line; therefore, Wind vs Temp have a weak linear relationship (since the correlation coefficient is -0.46.

Graph of correlation between Ozone level (y) vs Months (x)

```
library(ggplot2)

ggplot(airquality, aes(x = factor(Month), y = Ozone)) +
  geom_boxplot(na.rm = TRUE, fill = "lightblue") +
  labs(title = "Ozone levels by Month",
       x = "Month",
       y = "Ozone (ppb)") +
  theme_minimal()
```

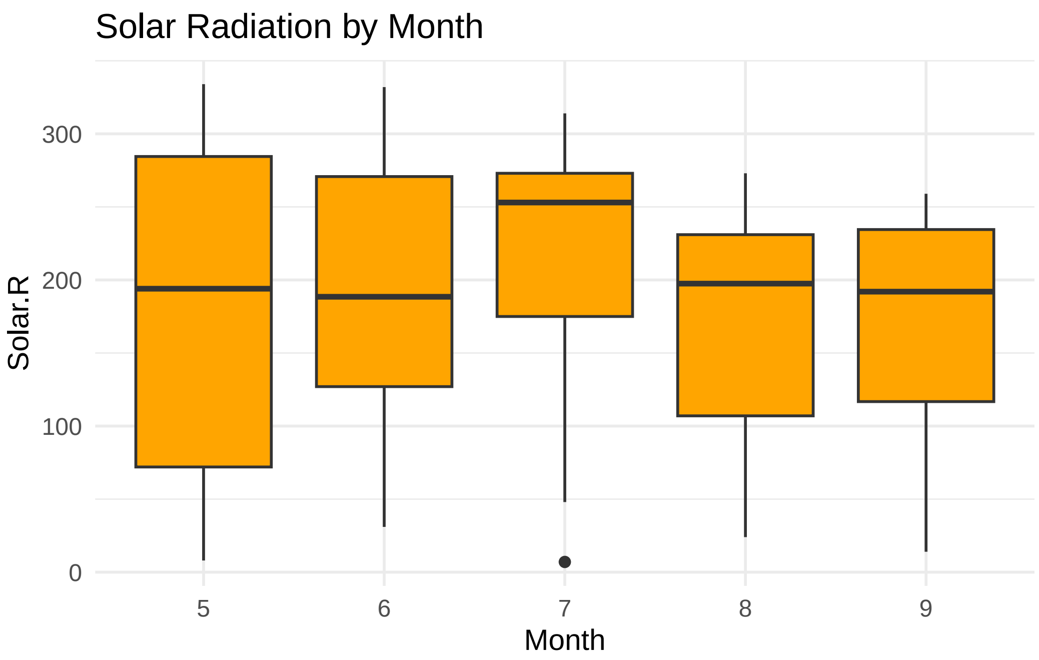


Conclusion: Ozone levels reach their highest values in August, while the lowest values occur in May and July.

Graph of correlation between Solar. R (y) vs Months (x)

```r
library(ggplot2)

ggplot(airquality, aes(x = factor(Month), y = Solar.R)) +
  geom_boxplot(na.rm = TRUE, fill = "orange") +
  labs(title = "Solar Radiation by Month",
       x = "Month",
       y = "Solar.R") +
  theme_minimal()
```

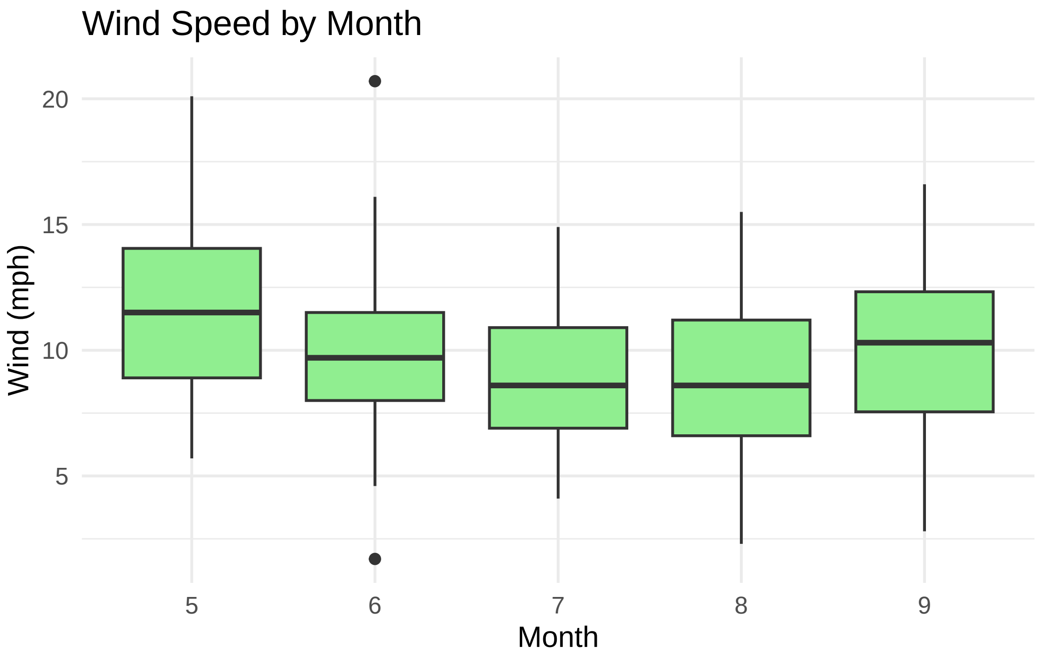


Conclusion: Solar radiation is highest in June and July, and lowest in May and September.

Graph of correlation between Wind (y) vs Months (x)

```r
library(ggplot2)

ggplot(airquality, aes(x = factor(Month), y = Wind)) +
  geom_boxplot(na.rm = TRUE, fill = "lightgreen") +
  labs(title = "Wind Speed by Month",
       x = "Month",
       y = "Wind (mph)") +
  theme_minimal()
```

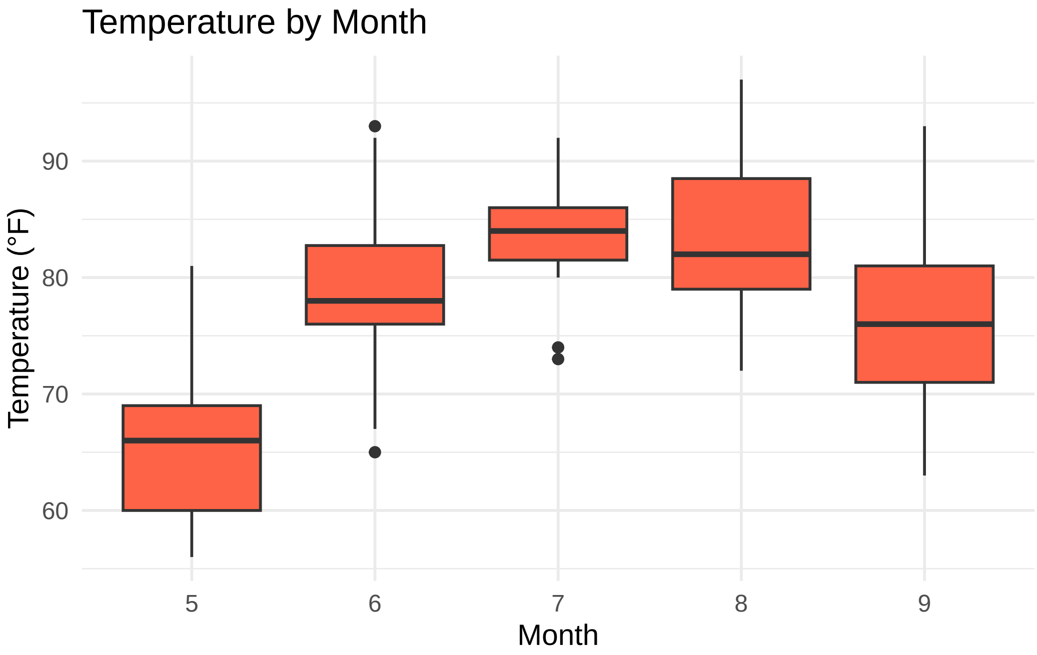


Conclusion: Wind is strongest in May and weakest in August, with June showing moderate levels.

Graph of correlation between Temp (y) vs Months (x)

```
library(ggplot2)

ggplot(airquality, aes(x = factor(Month), y = Temp)) +
  geom_boxplot(na.rm = TRUE, fill = "tomato") +
  labs(title = "Temperature by Month",
       x = "Month",
       y = "Temperature (°F)") +
  theme_minimal()
```



Conclusion: Temperature is lowest in May but peaks in July and August.