ENV 503: Statistics for Bioinformatics
Homework Set #2
Due: September 12, 2018

Instructions:
Use R to complete this assignment.
Assignment is to be submitted via Blackboard.

Use the R dataset **airquality** to answer all questions.

1. Get familiar with the dataset by using ?, str(), and head().

```
?(airquality)
```

New York Air Quality Measurements

Description

Daily air quality measurements in New York, May to September 1973.

Usage

```
airquality
```

Format

A data frame with 154 observations on 6 variables.

`[,1]`	`Ozone`	numeric Ozone (ppb)
`[,2]`	`Solar.R`	numeric Solar R (lang)
`[,3]`	`Wind`	numeric Wind (mph)
`[,4]`	`Temp`	numeric Temperature (degrees F)
`[,5]`	`Month`	numeric Month (1--12)
`[,6]`	`Day`	numeric Day of month (1--31)

Details

Daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973.

- `Ozone`: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
- `Solar.R`: Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park
- `Wind`: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
- `Temp`: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

Source

The data were obtained from the New York State Department of Conservation (ozone data) and the National Weather Service (meteorological data).

References

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983) *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.

Examples

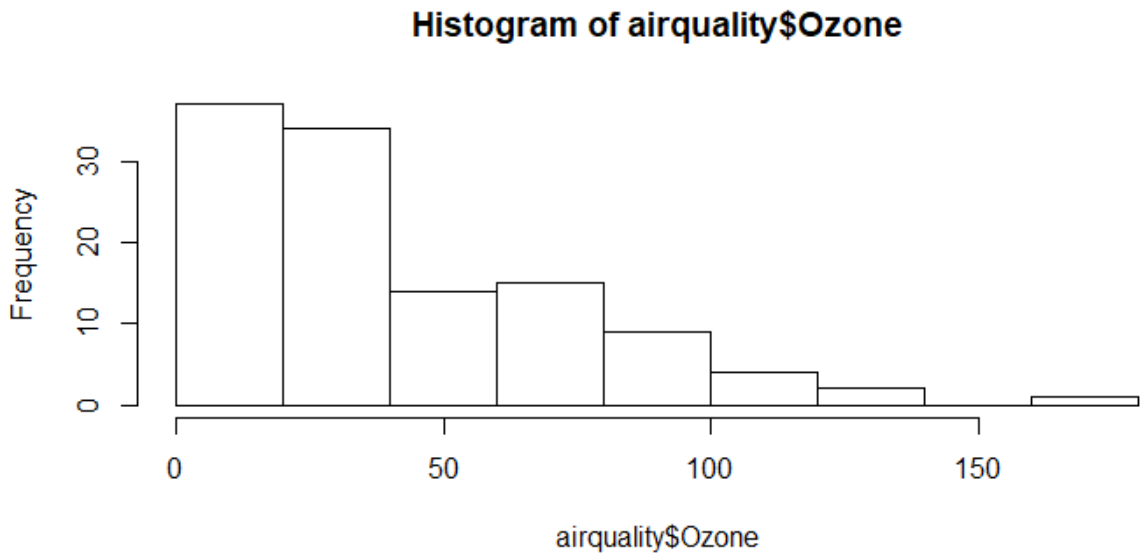
```
require(graphics)
pairs(airquality, panel = panel.smooth, main = "airquality data")
```

```
> head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
```

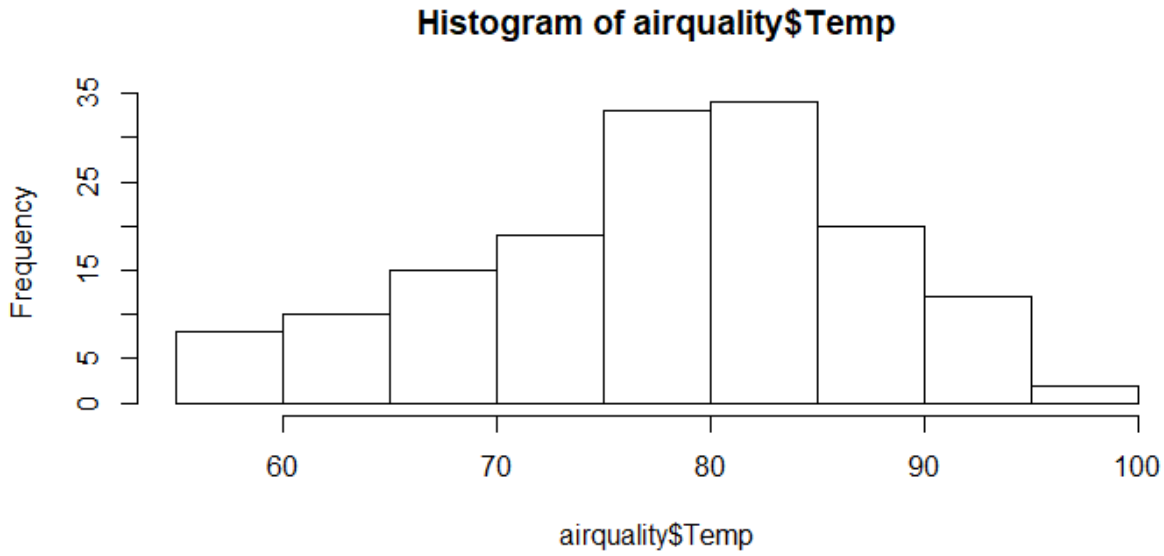
```
> str(airquality)
'data.frame':  153 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month  : int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

```
>
```

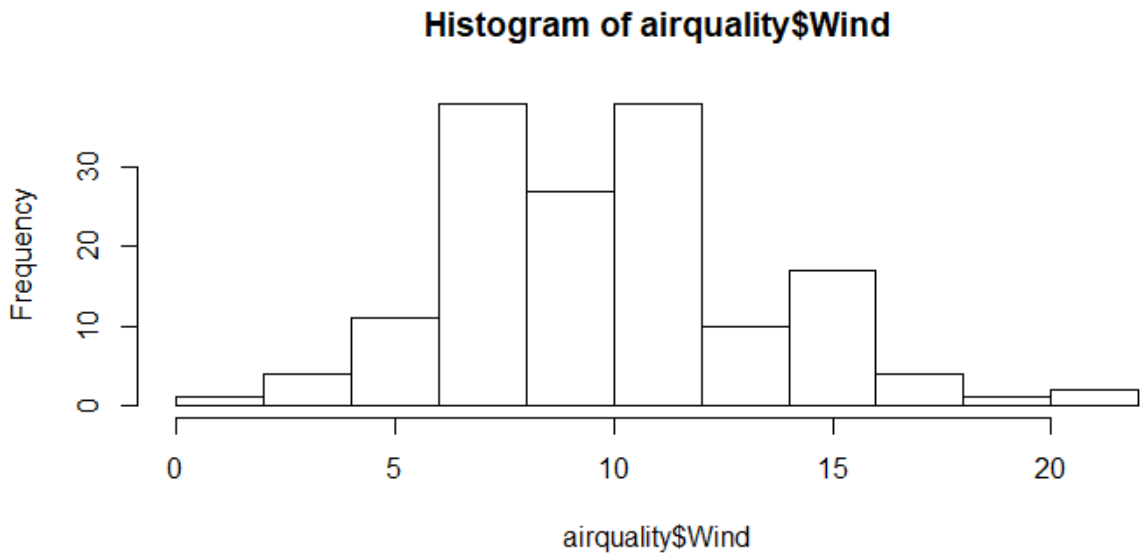
2. Plot a histogram for ozone, temperature, wind speed, and solar radiation. Describe each distribution.



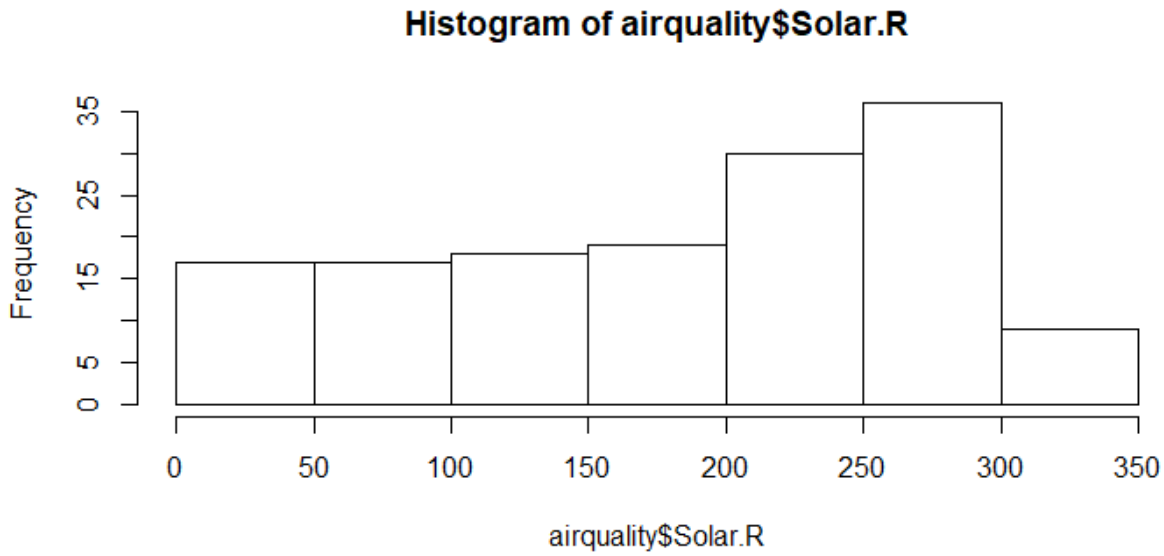
This distribution has skewed right since its longer tail is to the right of the mode. Also, It has outliers.



This is a bimodal distribution because it has two modes or value of high frequency.



This is a symmetric distribution because it's left, and right side are looking almost similar.



This is a bimodal distribution because it has two mode with high frequency than others.

3. Generate summary statistics for each variable using summary(). Which variable has the most missing values?

Ans.

```
summary(airquality)
     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
```

Ozone has the most missing values.37

4. Generate side-by-side box plots showing the distribution of each of these variables separately by month. How does each appear to vary by month?

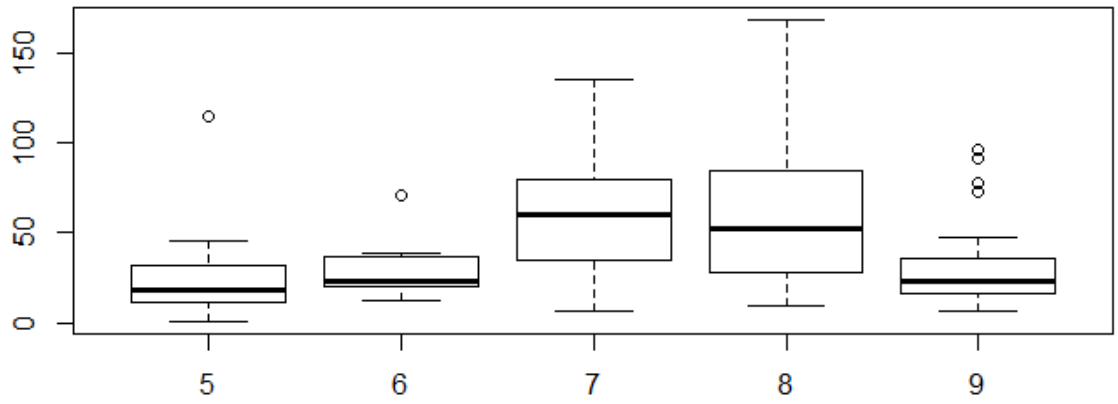


Figure: Side-by-side boxplot for ozone.

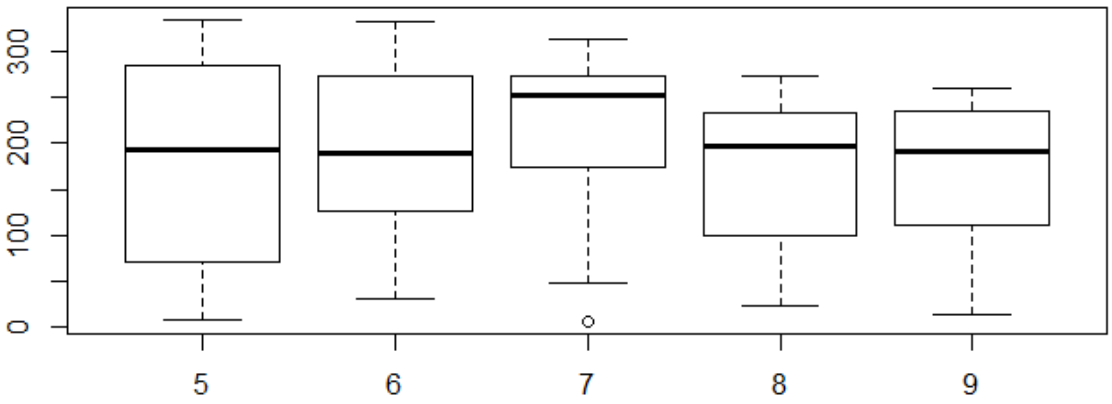


Figure: Side-by-side boxplot for Solar radiation.

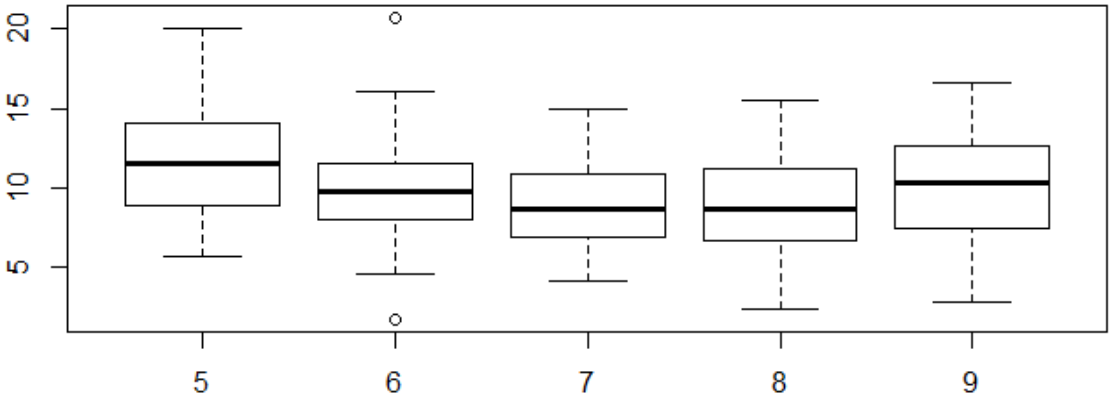


Figure: Side-by-side boxplot for wind.

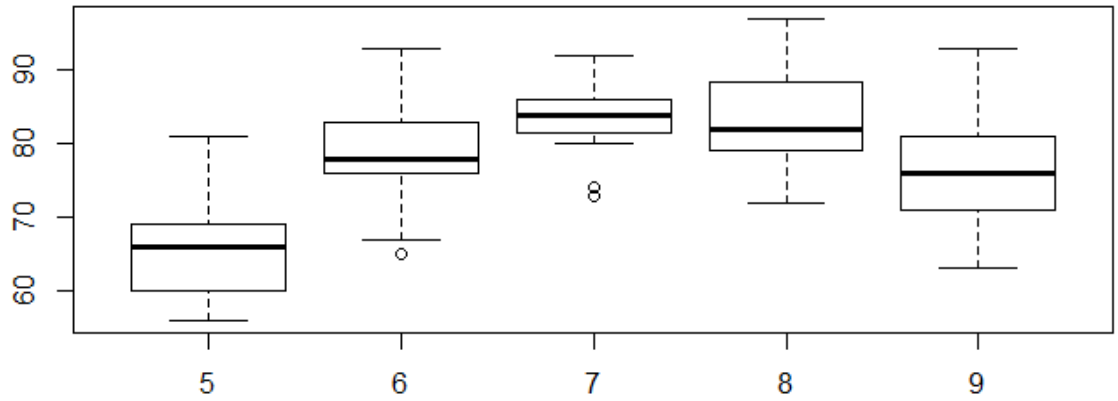


Figure: Side-by-side boxplot for Temperature.

5. Use aggregate() to calculate the mean and standard deviation of each of these variables separately by month. (Hint: the keyword to use for the mean is “mean”, for standard deviation is “sd”)

```
aggregate(Ozone~Month,airquality,mean)
  Month    Ozone
1     5 23.61538
2     6 29.44444
3     7 59.11538
4     8 59.96154
5     9 31.44828
```

```
aggregate(Ozone~Month,airquality,sd)
  Month    Ozone
1     5 22.22445
2     6 18.20790
3     7 31.63584
4     8 39.68121
5     9 24.14182
```

```
aggregate(Solar.R~Month,airquality,mean)
  Month  Solar.R
1     5 181.2963
2     6 190.1667
3     7 216.4839
```

```
4     8 171.8571
5     9 167.4333
```

```
aggregate(Solar.R~Month,airquality,sd)
  Month   Solar.R
1     5 115.07550
2     6  92.88298
3     7  80.56834
4     8  76.83494
5     9  79.11828
```

```
aggregate(Wind~Month,airquality,mean)
  Month      Wind
1     5 11.622581
2     6 10.266667
3     7  8.941935
4     8  8.793548
5     9 10.180000
```

```
aggregate(Wind~Month,airquality,sd)
  Month     Wind
1     5 3.531450
2     6 3.769234
3     7 3.035981
4     8 3.225930
5     9 3.461254
```

```
aggregate(Temp~Month,airquality,mean)
  Month     Temp
1     5 65.54839
2     6 79.10000
3     7 83.90323
4     8 83.96774
5     9 76.90000
```

```
aggregate(Temp~Month,airquality,sd)
  Month     Temp
1     5 6.854870
2     6 6.598589
3     7 4.315513
4     8 6.585256
5     9 8.355671
```