Recall

Recall July 30th, 2019

Data

- Data and its analysis has a rich and wide literature.
- In this course we will try understand Data using Statistical Inference.
- Three kinds of Data:
  - Categorical Data
  - Discrete Numeric Data
  - Continuous Numeric Data
- *On the Theory of Scales of Measurement* By S. S. Stevens Science 07 Jun 1946: Vol. 103, Issue 2684, pp. 677-680, gave a broad classification of data from measurements into 9 categories.

Discrete Numerical Data

- Many data are described in terms of numbers.
- Many variables naturally take on only discrete values.
- Boxplot and Histograms are used to visualise such data.

Discrete Numerical Data: Key features

- Center
- Spread
- Shape

Discrete Numerical Data: Key features

- Center Widely used measure of centre is the `mean` or the average of the data set. Other measures include the `median` and the `mode`
- Spread Understanding variabiity of the given data is very important. If one were to understand `mean` as specifying the center then the range of the data set around it is determined by its variability or spread. It is often measured by the variance(`var`) or standard deviation (`sd`) or the inter-quartile range (`IQR`).
- Shape To understand various distributional aspects of the dataset one needs to understand its "shape". For e.g. if it is symmetric or skewed around its mean. Other aspects include among the data points which are more likely than others.

Scan Function

- Let us use the `Scan()` to get the scores obtained by students in a midterm of a Calculus course.

```
> x = scan("Scores")
> x
 [1] 14 28  0 33 33 42 44 32 34 27 37 26  3 24 32 47 44 14 46 22 33 24 46  8 26
[26] 35 16 20 35 17 31 25 42  2  9 27 31 50 40 19 18 40 32 37 30 25 29  8 41 40
[51] 49 48 23 42 30 17 50 25 25 17 35 43 35  0 36 27 44 42 33 37 49 22 28 28 43
[76] 26 44 38 24 41 42 21  9 14 23 30 22 20 15 21 30 30 18 22 22 40 42
```

Discrete Numerical Data: Summary function

- A key characterisation of the data is given by the five number summary. namely

  `Min.  1st Qu.  Median  Mean  3rd Qu.  Max.`

In the R `summary()` provides the above information.

```
> summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   22.00   30.00   29.28   40.00   50.00
```

Discrete Numerical Data: Trimmed Mean

- `mean` as a measures of center suffers if there is a large tail or many outliers. `median` is resistant to a few very large values or few very small values.
- In R there is a notion of *trimmed mean*.

```
> median(x)

[1] 30

> mean(x, trim=1/10)

[1] 29.8481
```

In the above we trim the dataset by 10% from above and below. If we trim the dataset by 50% the trimmed mean will be the `median`.

Inter-Quartile Range

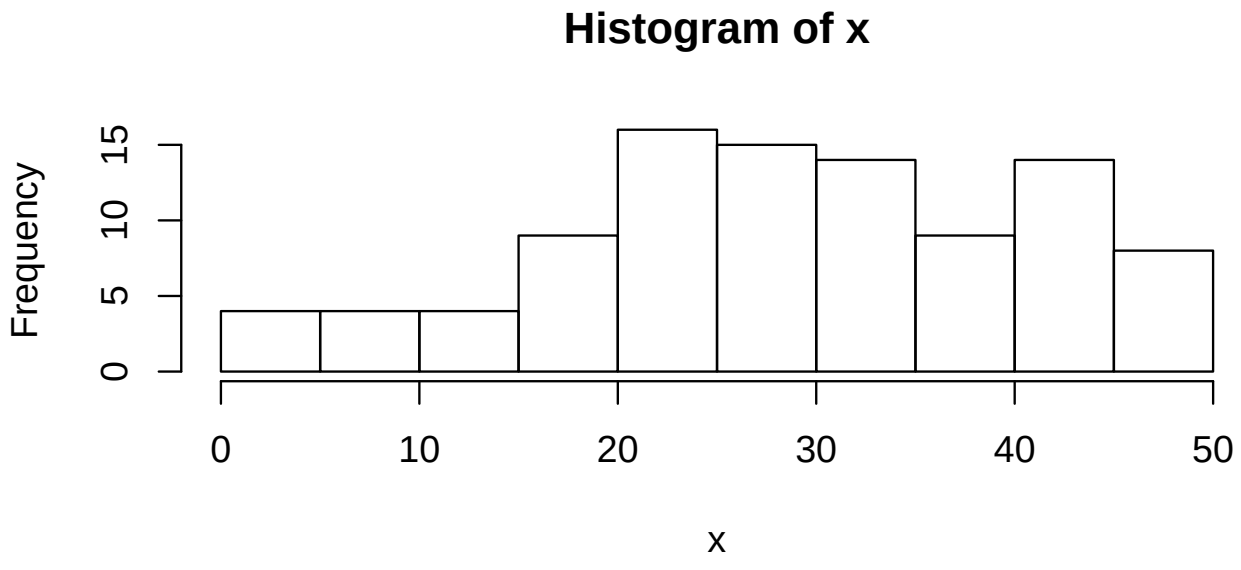
- IQR is a measure that is resistant to outliers.

```
> quantile(x, c(.25,.75))
25% 75%
 22  40
> IQR(x)
[1] 18
```

Histogram

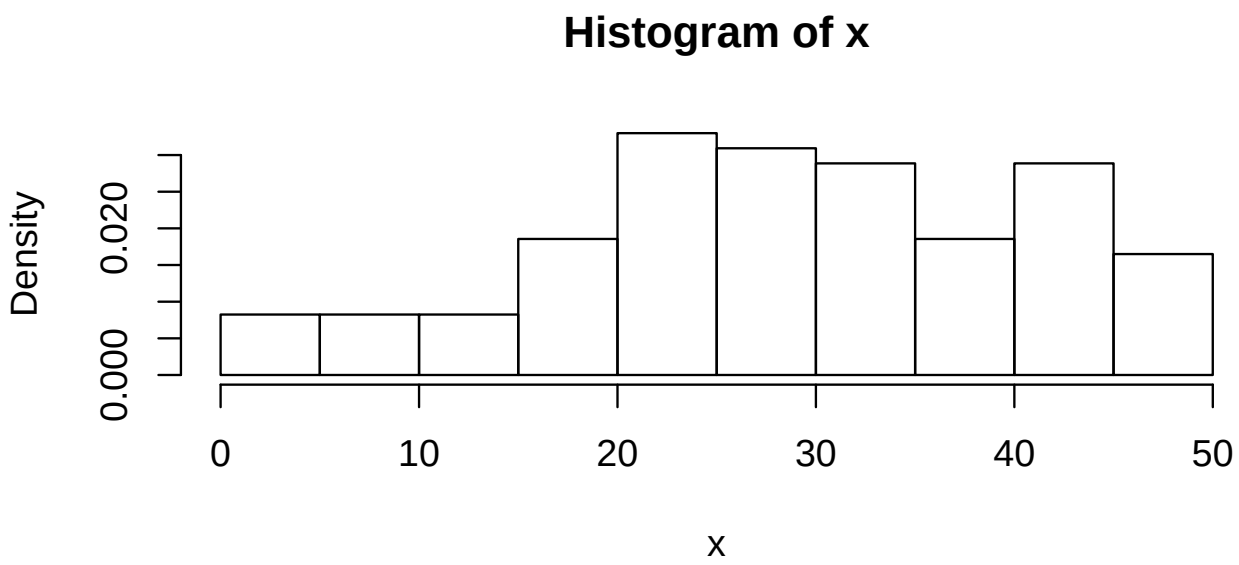
- The Histogram first specifies a sequence of points, called breaks. It counts the number of observation between the breaks, called bins. Place a bar in each bin with base being the length of the bin and height being either the frequency or proportion of observations in the bin.

```
> hist(x)
```



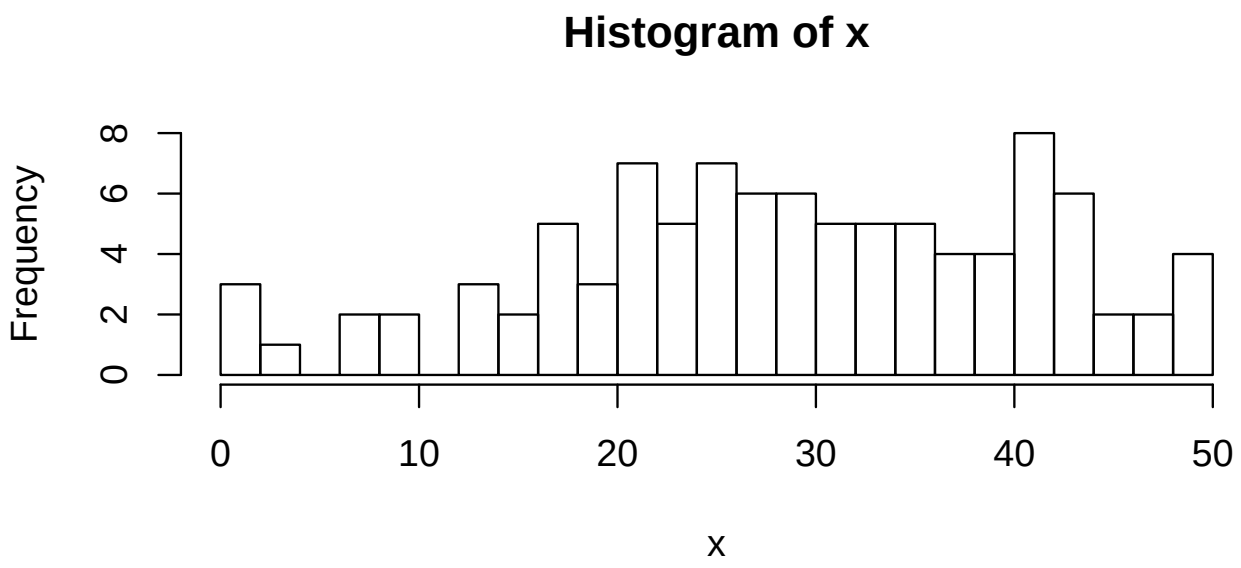
Histogram

```
> hist(x, probability=TRUE)
```



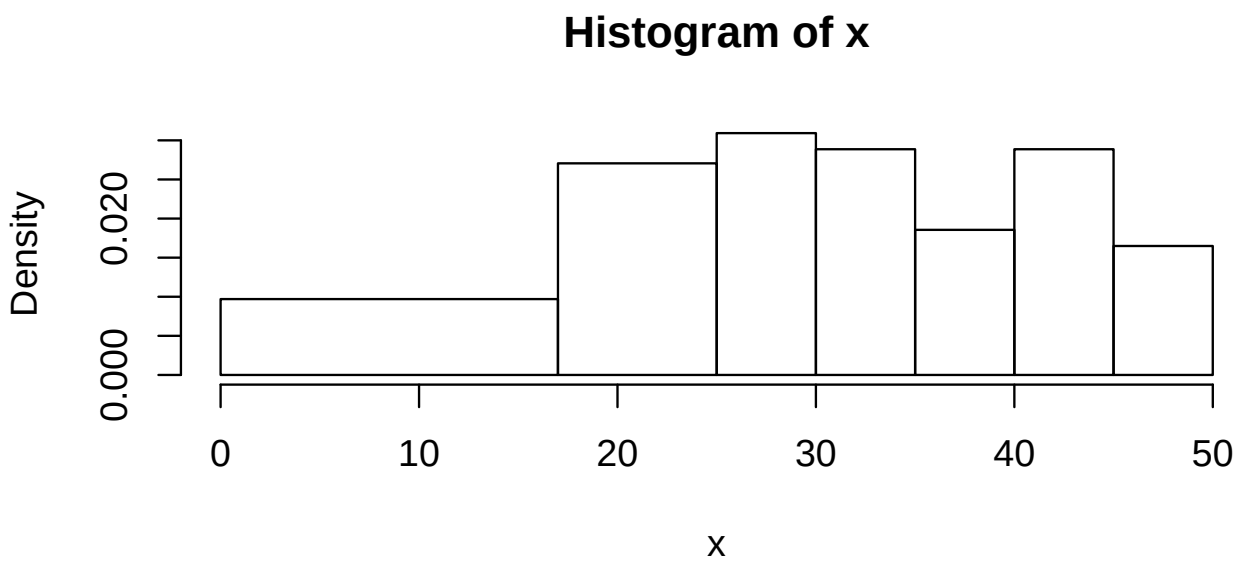
Histogram

```
> hist(x, 20)
```

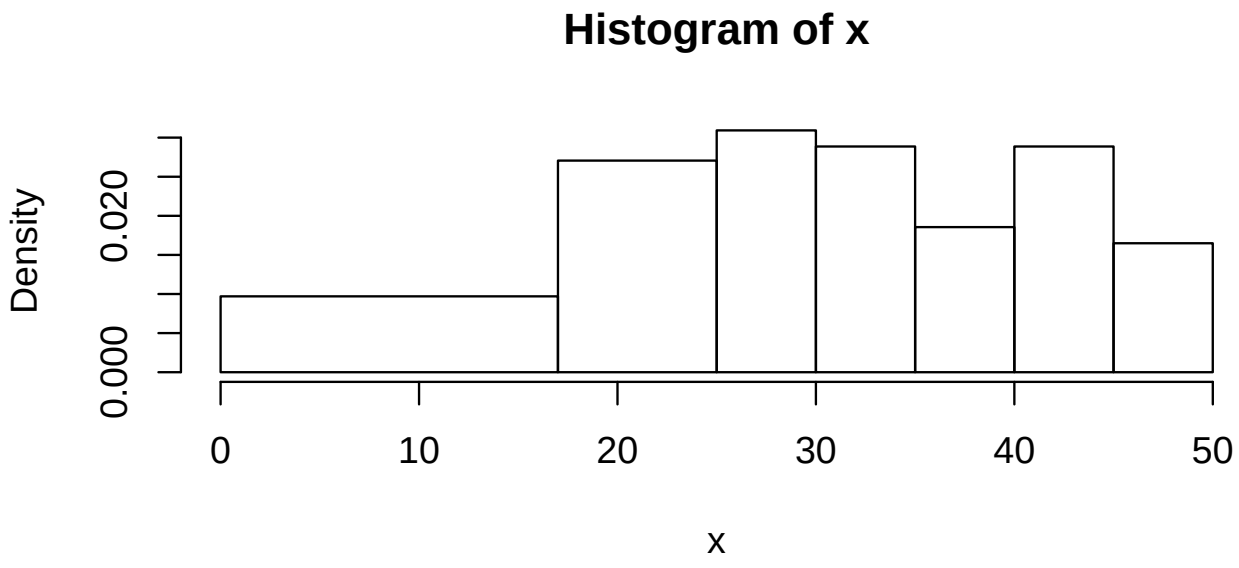


Histogram

```
> hist(x, breaks=c(min(x),17,25, 30,35, 40,45,max(x)))
```



```
> hist(x, breaks=c(min(x),17,25, 30,35, 40,45,max(x)))
```



Datasets in R

R has a lot inbuilt Datasets that one can use. The command :

```
> data()
```

will list currently installed data sets.

Datasets in R

R has a lot inbuilt Datasets that one can use. The command :

```
> data()
```

will list currently installed data sets.

Datasets in R

- R stores many datasets as data frame (often).
- A data frame is a collection of vectors.
- All of these vectors are describing different aspects of the quantity that we are trying to understand.

airquality in R

Let us learn about real data stored as data frame.

```
> ?airquality
```

airquality in R

Let us learn about airquality dataset a bit more.

- we could print the entire data set on the screen

  >airquality

  but this is too much information.

- Let us try the head() function

```
> 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
```

  This provides the first six rows.

airquality in R

Let us learn about airquality dataset a bit more.

- Let us try the tail() function

```
> tail(airquality)
```

	Ozone	Solar.R	Wind	Temp	Month	Day
148	14	20	16.6	63	9	25
149	30	193	6.9	70	9	26
150	NA	145	13.2	77	9	27
151	14	191	14.3	75	9	28
152	18	131	8.0	76	9	29
153	20	223	11.5	68	9	30

This provides the last six rows.

airquality in R

- Below provides the first ten rows.

```
> head(airquality, n = 10)
```

- Data can be called using row and column number

```
> airquality[148,4]
```

```
[1] 63
```

- We can use the variable name for the given column and call it by its position.

```
> airquality$Temp[148]
```

```
[1] 63
```

`airquality` in R

- Provides an entire row

  ```
  > airquality[148,]
  ```

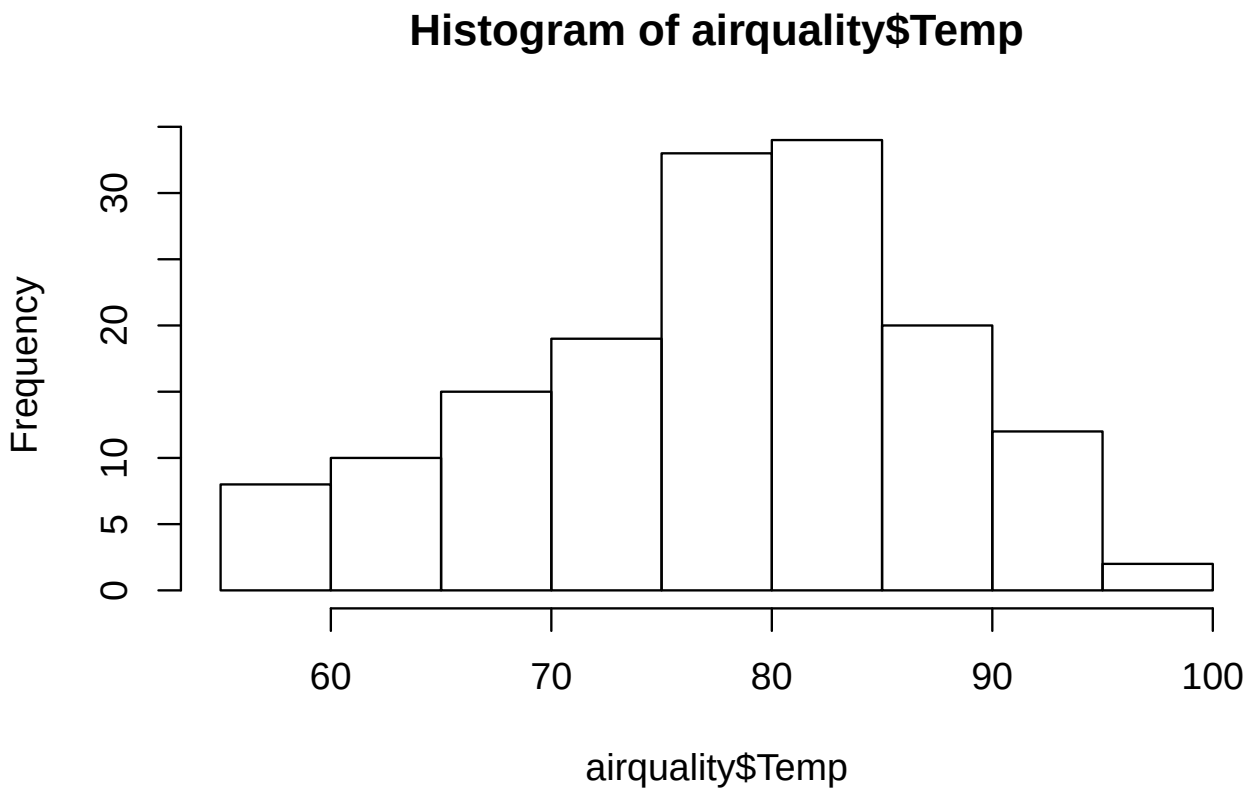
- Provides `Ozone` `Temp` columns

  ```
  > airquality[,c(1,4)]
  ```

  using `c()` function we can form any vector and that will enable display of the respective columns. We did not specify the row, so all rows will be displayed.

Five Number Summary and Histograms

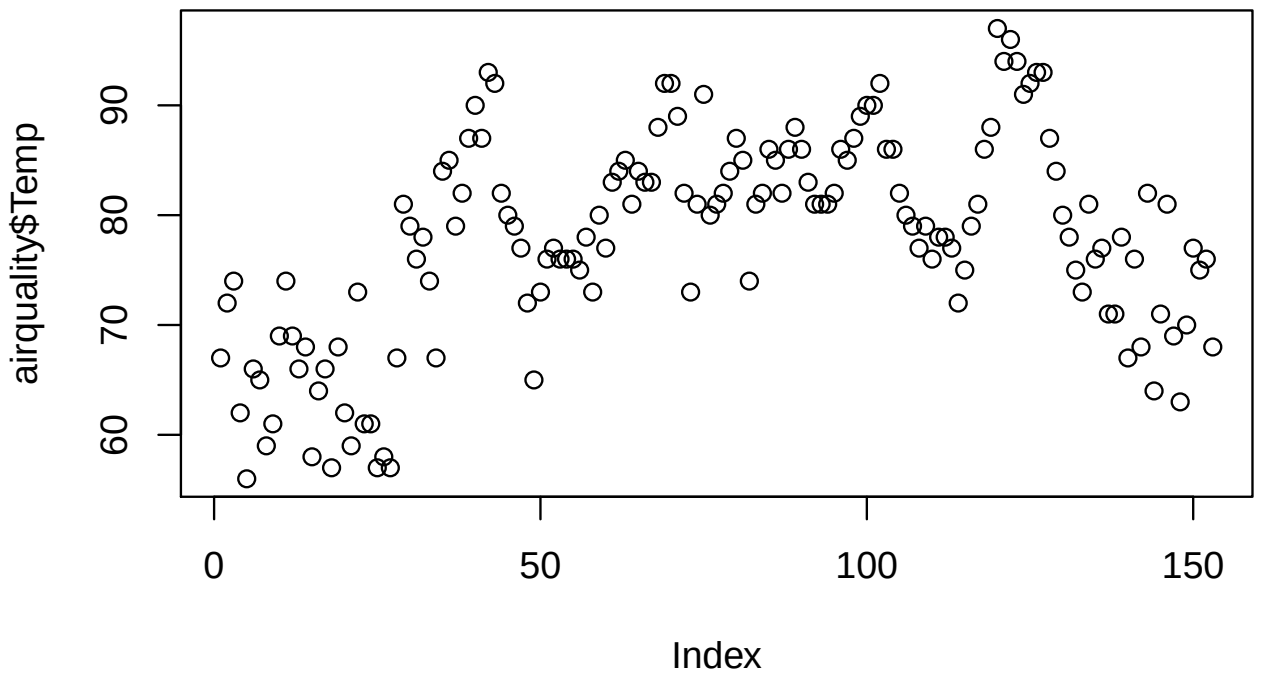
```
> summary(airquality$Temp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  56.00   72.00   79.00   77.88   85.00   97.00
> hist(airquality$Temp)
```



Plot

We can use the `plot` function to just plot.

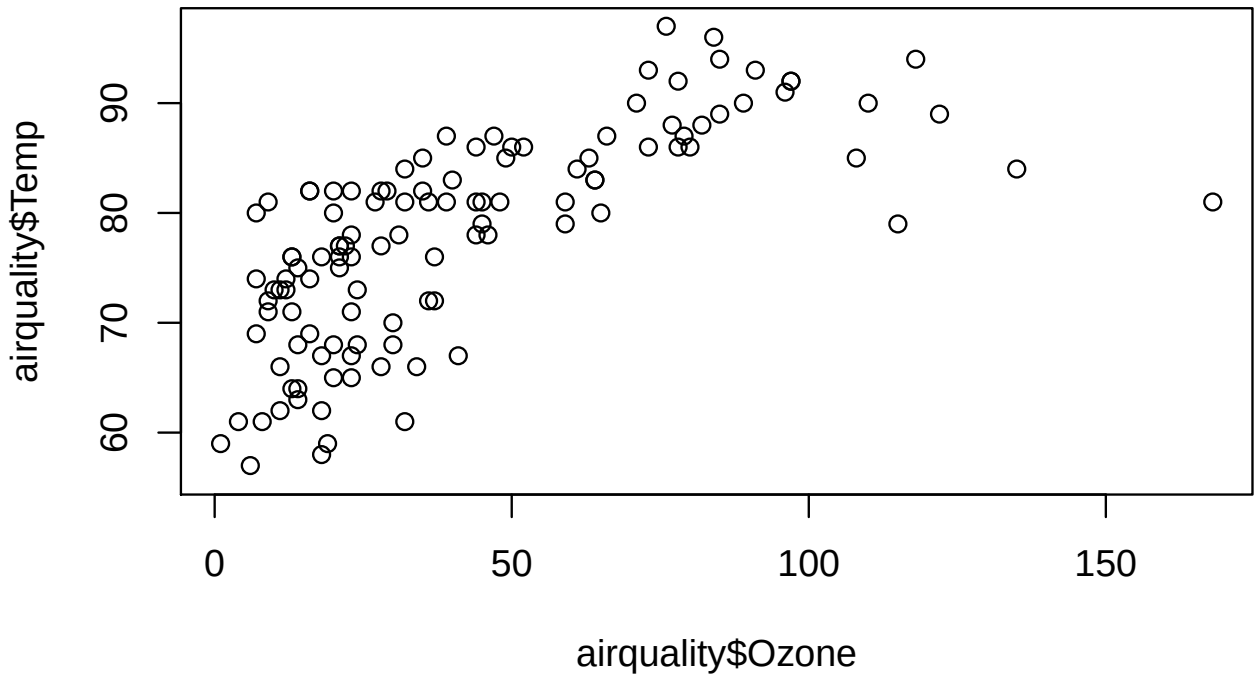
```
> plot(airquality$Temp)
```



Scatter Plot

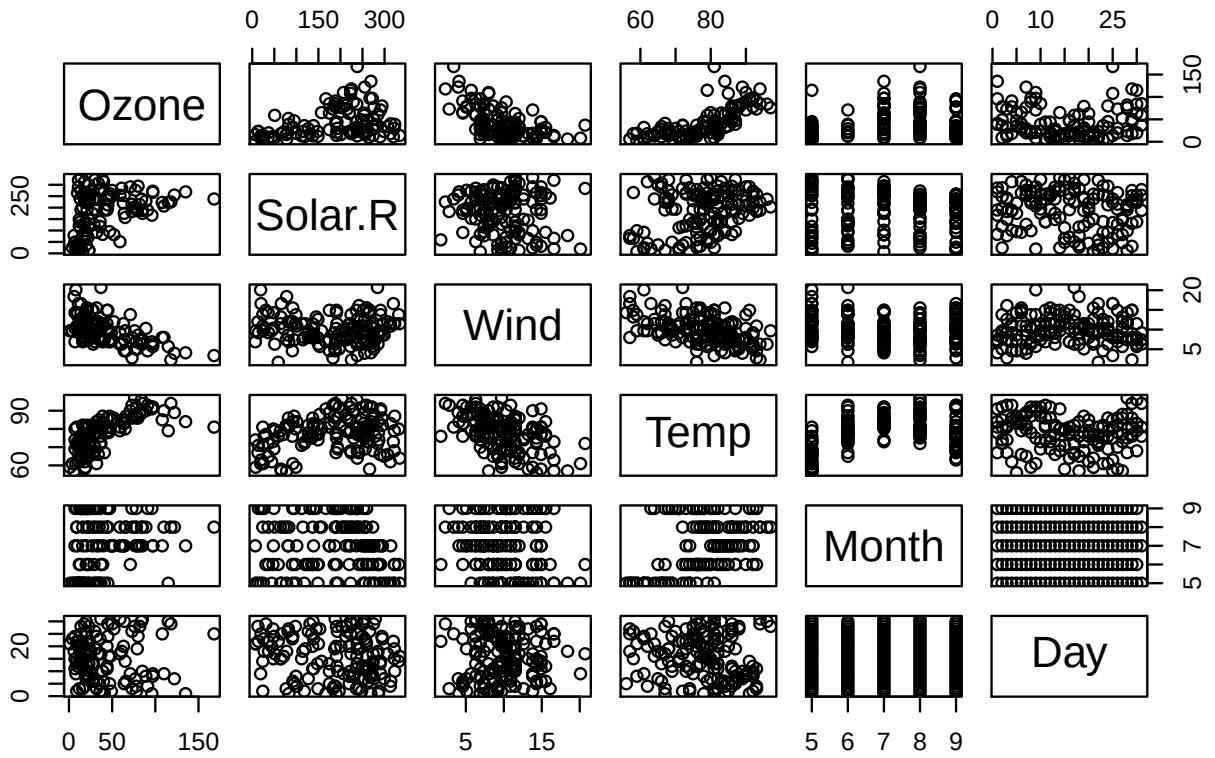
We can use the `plot` function to get a Scatter plot.

```
> plot(airquality$Ozone, airquality$Temp)
```



Plot!

```
> plot(airquality)
```



External Packages in R

R has can be enhanced with a lot of external packages that are available. The package `UsingR` has many datasets loaded in it.

```
> install.packages("UsingR")
```

Once installed then to add to current workspace

```
> require("UsingR")
```