Today, I am going to cover summarize data graphically using Base graphical procedure. They can produce publication quality graphics in several formats.

Let's look at Histogram. Histogram plots can summarize the distribution of the given continuous variable graphically.

```
library(readxl)
StudentsPerformance <- read_excel("C:/Users/yliu3/OneDrive - Maryville Univer
sity/Online DSCI502 R Programming/DataSets/StudentsPerformance.xlsx")
```

Let's make a histogram of math scores using base R command:

```
hist(StudentsPerformance$MathScore)
```

We can see the graphs shows the distribution of math scores. Most scores concentrate around 65.

We may change several default parameters of this function. For example

```
hist(StudentsPerformance$MathScore, main = "Histogram of Math Test   Scores",
 xlab= " Math Scores", ylab = "Count", xlim= c(0, 100), ylim = c(0,300))
```

- main: the title of the graph
- xlab: horizontal label
- ylab: vertical label
- xlim: the range of x values from the smaller number to a larger number.
- ylim: the range of y values from the smaller number to a larger number.

If we let the length of all bins approach zero, then we can obtain the density plot of the continuous variables. The geometric interpretation of the density plot between two numbers is the area under the curve is the probability between these two numbers.

To generate the density plot, we can use the density() function first, then we plot it.

```
dmath <- density(StudentsPerformance$MathScore)
plot(dmath, main = "Kernel Density Plot of math scores ")
```

If we want to plot histogram and density on the same graph, we can use the following R codes;

```
hist(StudentsPerformance$MathScore,
     main = "Histogram and density plots of Math Test Scores",
     prob = TRUE, # show probabilities instead of frequencies/counts
```

```
     xlab= " Math Scores", ylab = "probability")
lines(density(StudentsPerformance$MathScore), # density plot
 lwd = 2, # line width ratio to the  the default. 2 is twice as wide as the d
efault
 col = "red") #set the color of the density line to be red
```

The codes above perform the following tasks:

- Produces histogram with probability instead of frequency by setting prob to be TRUE
- Add density plot to the histogram by calling line function and specify the line width and color etc.

For continuous variables, we may plot all the points on the dot chart to have a complete picture of the dataset.

```
dotchart(StudentsPerformance$MathScore,
         cex = 0.5,  pch = 19, xlab = "math scores")
```

The two arguments cex show the scaling factor of the dot with respect to the default 1; pch =19 denotes the plotting symbols using solid circles.

For continuous variables, we can summarize the distribution of the data using boxplots. Boxplots summarize five important numbers according to Wiki Box plot and whisker of boxplot

- The bottom of the box shows the 25th percentile, generally denoted by Q_1 (the lower quartile)
- The middle of the box shows the 50th percentile, generally denoted by Q_2 (median)
- The top of the box shows the 75th percentile, generally denoted by Q_3 (the upper quartile)
- The upper whisker approximately shows the “maximum”"
- The Lower whisker approximately shows the “minimum”"

The points below the lower whisker or above the upper whisker are the **outliers**, which differ significantly from other observations.

```
boxplot(StudentsPerformance$MathScore, main = "Box plot of math scores", ylab
="Math scores")
```

We are interested in plotting two continuous variables to find the relationship between them.

For example, we would like to plot math score (y-axis) against reading score (x-axis)

```r
plot(StudentsPerformance$ReadingScore, StudentsPerformance$MathScore,
     main = "Scatter plot of math score against reading score",
     xlab = " Reading score", ylab= "Math scores")
```

Note that the first argument is put on the x-axis, the second argument is put on the y-axis. It is easy to see that students with higher reading scores tend to have higher math scores. We may want to add a trend line between these two scores. We can use the following R codes:

```r
#Scatter plot
plot(StudentsPerformance$ReadingScore, StudentsPerformance$MathScore,
     main = "Scatter plot of math score against reading score",
     xlab = " Reading score", ylab= "Math scores")
#Add trend line
abline(lm(MathScore~ReadingScore, data = StudentsPerformance),
       lwd = 2, col="red")
```

Note that, we need two steps to add the trend line in base R.

- Perform linear regression using the lm() function by specifying the relationship using $R\ formula$ and the data set.
- Draw the line equation using the abline function by setting the line color (col) and line width (lwd).