VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

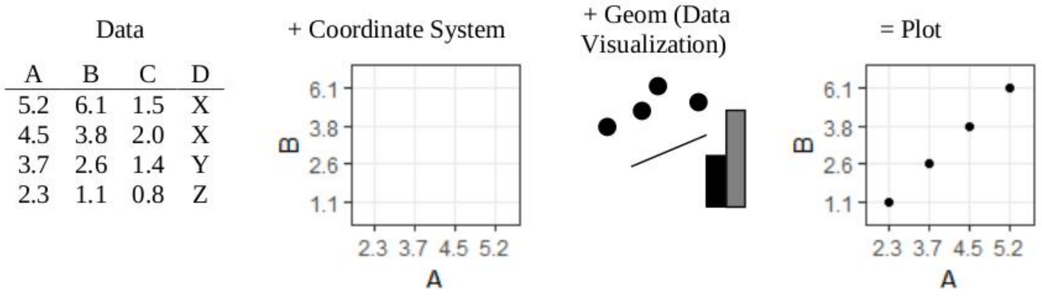
School of Computing Science and Engineering

LAB - 5 Exercises

Course Code	**:**	**CSE3020 - Data Visualization Lab**	**Date**	**:**	**04/01/2018**
Lab Experiment	**:**	**Construction of simple plots using ggplot2 in R**	**Slots**	**:**	**L19+L20**
Instructors	**:**	**Prof. Ramesh Ragala and Dr. Maheswari N**			

1. Introduction: In this lab, we will learn how to do basic exploratory graphs with the ggplot2 package. The “gg” stands for “grammar of graphics” because the package was developed to produce figures that are visually appealing and conform to sound graphical principles.

As with all graphs, the same three components are required: a data set, a coordinate system (x,y- coordinates) and a visual representation of data points (so called geoms in ggplot2). Here is a simple illustration of these three components:



Data

A	B	C	D
5.2	6.1	1.5	X
4.5	3.8	2.0	X
3.7	2.6	1.4	Y
2.3	1.1	0.8	Z

For plotting with the ggplot2 package, data should be in wide format if you have two variables (e.g. to produce a simple scatterplot), but long format is usually best if you have 3 or more variables (e.g. for more complex or multivariate plots).

The coordinate system and geoms must be specified in ggplot2-specific syntax. This ‘grammar’ is usually more efficient than the basic graphic functions of R when creating complicated figures because several key components are developed automatically and elegantly. For example, legends are produced automatically in ggplot2 if the data is in the appropriate format.

Ggplot2 syntax also requires a + symbol after each line of code before another line of code. This signals to the package that more code will follow. Therefore, all of the lines of code to produce a figure have a + symbol except the last line. Forgetting the + symbol is a common source of error.

Bringing these three principles together, the basic syntax therefore looks like this (italics mean user input, bold indicates mandatory commands that specify data, coordinate system and geom):

```
fig1 = ggplot(data=data, mapping = aes(mappings)*) +
     geom_function(stat = stat, position = position ) + coordinate_function()
     + facet_function() + scale_function() + theme_function()

fig1
```

In this lab, you will learn about three types of graphs that can be used to visualize raw data and check for errors, outliers, and the nature of distributions and correlations: histograms and scatter plots. You will work with the iris dataset in this lab. Iris is the genus of a flowering plant and the dataset includes measurement of several features of the flower: Petal length and width and sepal length and width. In addition, there is also a categorical variable of naming three species.

Problem set - I:

Histograms can be used to understand the data frequencies:

1. Produce a colored histogram to understand the data frequencies of Petal Length in iris dataset, where as color is based on the Species.
2. Produce a colored histogram to understand the data frequencies of Sepal Length in iris dataset, where as color is based on the Species.
3. Produce a Colored histogram to understand the data frequencies of Petal width in iris dataset, where as color is based on the Species.
4. Produce a colored histogram to understand the data frequencies of Sepal width in iris dataset, where as color is based on the Species.

Problem set - II:

With scatter plots you can visually check the relationships among variables. Are they linear or curvilinear? Outliers are also easily visible.

1. Produce the colored scatter plot between Sepal Length and Sepal Width, where as color is based on Species.
2. Produce the colored scatter plot between Petal Length and Petal Width, where as color is based on Species.