Chapter 6: Visualizing Data using R and ggplot2 package

Overview: So far, we've discussed different ways of collecting data (methods of sampling) and we've seen how to read an external data set into R and how to access particular values from a data set whether it is a matrix or a data frame. We also learned some basic data wrangling techniques. We will now begin exploring how to visualize the data once it is read into R.

Visualization of data is a very important presentation method. It is a quick way to represent the data and to illustrate what the data is telling you. That being said, caution must be taken when choosing how to display the data visually as not all types of plots are appropriate for all types of data.

Motivating Example: Suppose you have several data sets that you want to visualize. These include the final letter grade distribution for a previous Stat 123 class, the annual lynx trappings in Canada, and the number of gears in a variety of manual and automatic cars. What is the best way to display these data sets?

In this chapter, we will cover several types of plots: bar charts, histograms, scatter plots, and line plots. We will not use R's built-in graphics. Instead, we will use the gglot2 package.

Plotting with ggplot2:

The function that we use to create plots is called *ggplot*(). The general form of the code required to create a graph using the *ggplot*() function is:

library (ggplot2)   or   library (tidyverse)

ggplot (data , aes (   )) + type of plot + ...

data frame

specify how you want to present your graphics

Question: What is the aesthetic?

Answer: The aesthetic changes depending on what type of graph you are creating. For a scatterplot, the minimum aesthetic required is the variable you want plotted on the x-axis and the variable you want plotted on the y-axis. You can also include colour in your aesthetic (which will be demonstrated in the examples below).

Question: How do we specify which type of graph we want to create (i.e. what do we type after the + sign?).

Answer: Again, it depends on what type of graph you want:

bar graphs

histogram

scatterplot

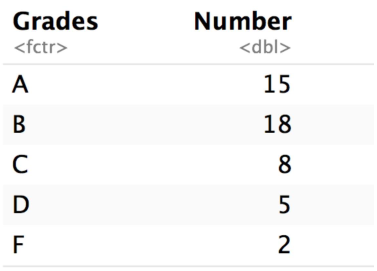
line plot (later)

Bar Graphs: A **Bar graph** works well with categorical variables.

A bar chart can be used:

to display frequencies of categorical variable values.

For example the final letter grade distribution of a former Stat 123 class could be effectively displayed using a bar chart. Suppose you have the following information:



Grades	Number
<fctr>	<dbl>
A	15
B	18
C	8
D	5
F	2

48

Percentage (Relative Frequency)

31.25%

We sometimes refer this as a frequency table.

$$\text{percentage} = \frac{\text{frequency}}{\text{total}}$$

Bar Charts using ggplot2:

Start by creating a data frame with at least these two variables. One which contains the names of the categories of your categorical variable and another which contains the number associated with each category.

```
Grades.distn <- read.csv ("grades.csv")
```

↳ data frame as a frequency table.

Next, use the function ggplot() function using a word geom_bar as the type of plot.

```
ggplot (Grades.distn , aes(x= Grade , y= Count)) + geom_bar (stat = "identity")
```

↑ data

↑ Grade is on the x-axis

↑ bar graph

↑ data are frequencies

We can change the orientation of the bar graph:

```
ggplot (Grades.distn , aes(x = Count , y= Grade)) + geom_bar(stat = "identity")
```

We can add title, labels colour, etc. (see R scripts)

adding + ggtitle( ) and + labs( ) to the above command.

Producing bar graphs from different raw data:

See R Script for some examples.