Chapter 10: Visualizing Data (Good and Bad Plots)

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 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 section we will cover 3 types of plots: a pie chart, a bar graph, and a time series plot.

Pie Charts: A **pie chart** is an effective visualization tool when

for categorical data

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

frecuancy

Grades <fctr>	Number <dbl>	
A	15	$(15/48) \times 100 = 31.25\%$
B	18	$(18/48) \times 100 = 37.5\%$
C	8	$(8/48) \times 100 = 16.67\%$
D	5	$(5/48) \times 100 = 10.42\%$
F	2	$(2/48) \times 100 = 4.17\%$
	total: 48	$total = 100\%$

category

We can manually compute the percentages for each grade and then draw a pie chart with each wedge representing a specific letter grade:

Pie Charts in R: Obviously, we do not want to hand draw a Pie Chart. Luckily, this is something that we can make in R. We will now see one way to do this:

Start by defining 2 vectors. One which contains the names of the categories of your categorical variable and another which contains the number associated with each category.

```
grades = c("A", "B", "C", "D", "F")
grades
#or

grades = c(LETTERS[1:4], "F")

number = c(15,18,8,5,2)
```

Next, use the function pie() to plot the pie chart.

```
pie(number, labels = grades, main = "Simple Pie Chart for Stat 123")
```

This is the most basic way to plot a pie chart in R. If you want more vibrant colours you can put in an additional parameter:

```
pie(number, labels = grades, main = "Simple Pie Chart for Stat 123", col= rainbow(length(grades)))
```

If you also want to display the percentages associated with each category then there is an extra 2 steps:

First, create a vector which calculates the percentages:

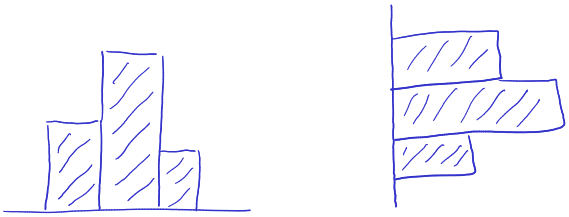
```
percents = round((number/sum(number))*100, digits = 2)
```

Then, use the paste() function in R to concatenate the Grades vector with the Percentages vector:

```
grades2 = paste(grades, percents, "%")

pie(number, labels = grades2, main = "Simple Pie Chart for Stat 123")
```

Bar Graphs: A **Bar graph** also works well with categorical variables but it is more flexible than a pie chart. A bar chart can be used



For example, consider the built-in R data set called mtcars. If you start by looking into the description of the data set you see that column 9 and 10 contain information on whether or not a car is automatic (0) or manual (1) and how many forward gears each car has (3, 4 or 5).

	am <dbl>	gear <dbl>
Datsun 710	1	4
Valiant	0	3
Merc 280	0	4

A bar graph would be a nice way to visualize each categorical variable individually as well as both at the same time.

Creating a Bar Graph in R: First, we need to create a count of how many individuals are in each category. To do this, we use the table() function in R.

```
gearCourts = table (mtcars$gear)
```

Then we use the R function called barplot() to create our plot:

```
barplot (gearCounts, main = " Gear Distribution" , xbar = " Number of Gears")
```

data — title — x-axis label

Practice Question: How many cars in the mtcars data set are automatic?

(A) 13 (B) 19 (C) 32

```
table (mtcars$am)
```

If we wanted to create a bar graph displaying the number of automatic vs manual cars then we could do so using the code:

```
barplot (autvsman, main = "BarGraph of Aut. vs Man cars")
```

Grouped Bar Graphs: When you want to combine categorical variables then you can do so with grouped bar graphs.

For example, suppose we want to compare how many manual vs automatic cars there are with 3 gears, 4 gears and 5 gears. We can do this using a grouped bar chart. To create one of these in R:

We need to have a count table in terms of both variables:

```
groupCounts = table (mtcars$am, mtcars$gear)

rownames (groupcounts) = c("Automatic", "Manual")
```

Then when we use the barplot() function, we need to specify colours for each bar and provide a legend explaining the colours:

```
barplot (groupCounts, main = "Distribution of Greaks and Transmission, xlab ="Number of Gears",
        col = c("darkblue", "red"), legend = rownames (groupCounts), beside = True)
```

put the bars nex to each other

The final type of plot that we will look at in this chapter is called a line graph.

Line Graphs: A **line graph** is often used for numerical variables collected over time
e.g. time series analysis

For example, consider the built in lynx data set. This contains the annual lynx trappings in Canada from 1821 to 1934.

If you use the class() function in R, you will see that this data set is considered to be time series (ts) data.

class (lynx) → R will return "ts" meaning it is a time series

Line Graphs in R: Time series data is very easy to plot in R. You simply use the plot() function and make sure to add titles for your x and y axes. For example, the following code creates a line graph for the lynx time series in R:

```
plot (lynx, main = "Line Graph for Lynx data", xlab ="Year", ylab = "Number of Trappings")
```

Question: What can we notice from plotting a line graph?

Answer: We can notice something called **trend** and **seasonal variation**.

Definition:

- A **trend** in a time series is a long-term or downward movement over time

- A **seasonal variation** in a time series is a pattern which regularly repeats itself over time

Definition: The term distribution is used commonly in statistics. A **distribution** of a variable tells us what values it takes and how often it takes these values

Thus, when we plot data. We are visualizing the distribution of the variable we are plotting.