Spring 2017, Seminar 01

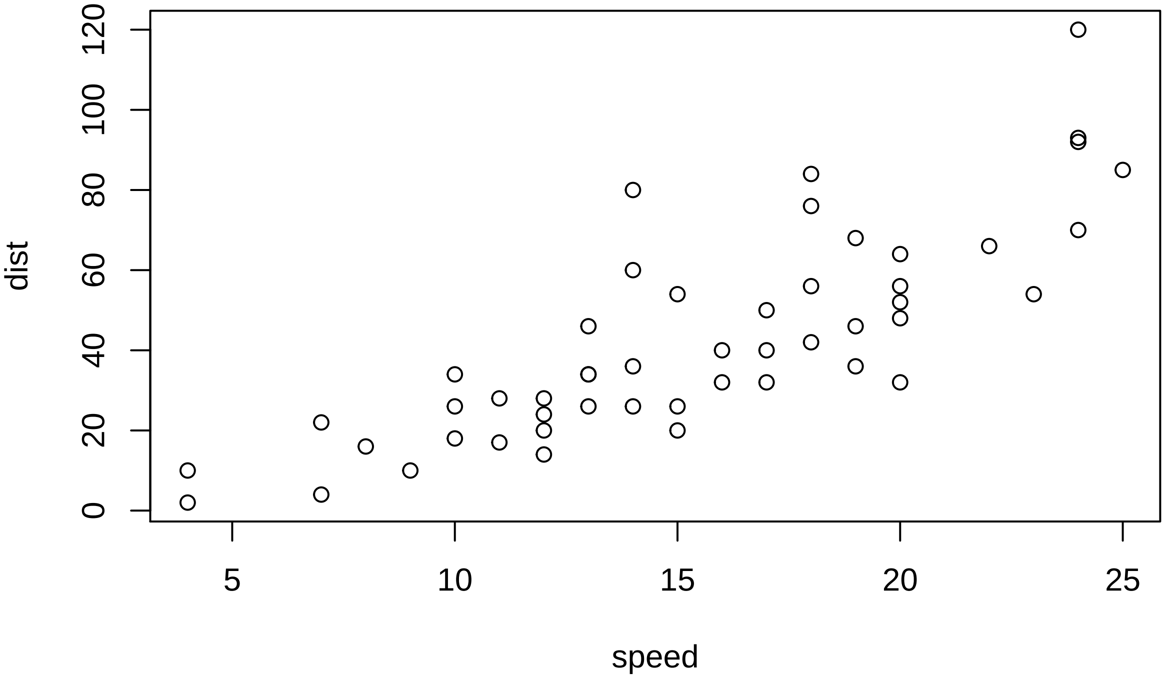
Christopher Prener, Ph.D.

25 Jan 2017

Introduction

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code. Look for the *Run* button, which is a green, right facing triangle shaped icon on the right side of the code chunk.

```r
plot(cars)
```



You'll notice all of the formatting used here with the text. This is *Markdown* formatting, which is a lightweight markup syntax that RStudio increasingly utilizes to produce reproducible research products. You can find out all about Markdown by checking out this tutorial.

Data

Today, we'll use two data sources for our class. The examples we'll discuss come from fivethirtyeight.com. The data was originally used for this article that investigated whether Pulitzer Prizes helped newspapers keep readers.

```r
prize <- read.csv("pulitzer-circulation-data.csv", stringsAsFactors = FALSE)
```

Pay close attention to the structure of the code above. We've assigned the data in the `csv` file to an object named `prize`. We use the `read.csv()` function to do this. You *must* place the filename in quotes

Now, practice writing this code on the other dataset in the seminar's directory: `auto2016.csv`. These data come from the U.S. Department of Energy and have been extensively cleaned by Chris.

Exploring the Dataset

The Structure Function

The variables within `prize` are visible in the environment tab. You can also get a table that summarizes your variables using the structure function - `str()`.

```
str(prize)
```

```
## 'data.frame':    50 obs. of  7 variables:
##  $ Newspaper                                      : chr  "USA Today" "Wall Street Journal" "New York
##  $ Daily.Circulation..2004                        : chr  "2,192,098" "2,101,017" "1,119,027" "983,72
##  $ Daily.Circulation..2013                        : chr  "1,674,306" "2,378,827" "1,865,318" "653,86
##  $ Change.in.Daily.Circulation..2004.2013         : chr  "-24%" "+13%" "+67%" "-34%" ...
##  $ Pulitzer.Prize.Winners.and.Finalists..1990.2003: int  1 30 55 44 52 4 0 23 4 12 ...
##  $ Pulitzer.Prize.Winners.and.Finalists..2004.2014: int  1 20 62 41 48 2 0 15 2 6 ...
##  $ Pulitzer.Prize.Winners.and.Finalists..1990.2014: int  2 50 117 85 100 6 0 38 6 18 ...
```

Note that you can see the variable names, their data types (`chr` for character and `int` for integer, a particular type of numeric variable), and some examples of the data they contain.

Now try using the structure function on the automobile data:

Confirming Data Types

There are ways to automatically confirm the types of variables that you have. For example, we can test whether `Newspaper` is a character or numeric variable:

```
is.numeric(prize$Newspaper)
```

```
## [1] FALSE
```

```
is.character(prize$Newspaper)
```

```
## [1] TRUE
```

Note that the output follows the order of the commands. So line 1 corresponds with command 1, and line 2 corresponds with command 2.

Now, using the automobile data, test whether the variable `brand` is numeric or character:

Looking at Individual Observations

We can use two functions - `head()` and `tail()` - to look at the first 6 and last 6 observations of a dataset. Note that presence of the black right facing triangle icon on the right side of the table. You can use that to scroll right to view more columns.

```
head(prize)
```

```
##             Newspaper Daily.Circulation..2004 Daily.Circulation..2013
## 1           USA Today               2,192,098               1,674,306
## 2 Wall Street Journal               2,101,017               2,378,827
## 3      New York Times               1,119,027               1,865,318
## 4   Los Angeles Times                 983,727                 653,868
## 5     Washington Post                 760,034                 474,767
## 6 New York Daily News                 712,671                 516,165
##   Change.in.Daily.Circulation..2004.2013
## 1                                   -24%
## 2                                   +13%
## 3                                   +67%
## 4                                   -34%
## 5                                   -38%
## 6                                   -28%
##   Pulitzer.Prize.Winners.and.Finalists..1990.2003
## 1                                               1
## 2                                              30
## 3                                              55
## 4                                              44
## 5                                              52
## 6                                               4
##   Pulitzer.Prize.Winners.and.Finalists..2004.2014
## 1                                               1
## 2                                              20
## 3                                              62
## 4                                              41
## 5                                              48
## 6                                               2
##   Pulitzer.Prize.Winners.and.Finalists..1990.2014
## 1                                               2
## 2                                              50
## 3                                             117
## 4                                              85
## 5                                             100
## 6                                               6
```

```
tail(prize)
```

```
##                     Newspaper Daily.Circulation..2004
## 45              Boston Herald                 236,899
## 46              Seattle Times                 233,497
## 47         Charlotte Observer                 231,369
## 48            Daily Oklahoman                 223,403
## 49 Louisville Courier-Journal                 216,934
## 50 Investor's Buisiness Daily                 215,735
##    Daily.Circulation..2013 Change.in.Daily.Circulation..2004.2013
## 45                  95,929                                   -60%
## 46                 229,764                                    -2%
## 47                 137,829                                   -40%
## 48                 124,667                                   -44%
## 49                 131,208                                   -40%
## 50                 157,161                                   -27%
##    Pulitzer.Prize.Winners.and.Finalists..1990.2003
## 45                                               0
## 46                                              11
## 47                                               1
## 48                                               0
## 49                                               0
## 50                                               0
##    Pulitzer.Prize.Winners.and.Finalists..2004.2014
## 45                                               0
## 46                                               5
## 47                                               3
## 48                                               0
## 49                                               3
## 50                                                1
##    Pulitzer.Prize.Winners.and.Finalists..1990.2014
## 45                                                0
## 46                                               16
## 47                                                4
## 48                                                0
## 49                                                3
## 50                                                1
```

Now, use the `head()` and `tail()` functions to explore the automobile data.

Head of the automobile dataset:

Tail of the automobile dataset:

RStudio Output

Everytime you save your RNotebook, an html file will automatically be generated containing all of your output. You can view it from within RStudio, or open it using a web browser.