HW2_jaeyoolkim

Jaiyool Kim

8/29/2019

Problem 1

R is an open source, community built, programming platform. Not only is there a plethra of useful web based resources, there also exist in-R tutorials. To speed our learning, we will use one such tutorial *swirl*. Please install the *swirl* package, install the "R_Programming_E" lesson set, and complete the following lessons: 1-3 and 15. Each lesson takes about 10 min.

From the R command prompt:

```r
install.packages("swirl")
library(swirl)
install_course("R_Programming_E")
swirl()
```

I took $No.1 - 3$, and 15 lectures in the "R_Programming_E" lesson set.

(I also sent the verifications e-mails to your e-mail address : "rsettlag@vt.edu")

Problem 2

Now that we have the R environment setup and have a basic understanding of R, let's add Markdown (choose File, New File, R Markdown, pdf).

Let's go ahead and save the file as is. Save the file to the directory containing the *README.md* file you created and committed to your git repo in Homework 0. The filename should be: HW1_pid, i.e. for me it would be HW1_rsettlag.

You will use this new R Markdown file for the remainder of this homework.

I uploaded my 1st homework in my repository and applied for 'fork' and pull request to your repository.

Part A

In this new Rmarkdown file, please type a paragraph about what you are hoping to get out of this class. Include at least 3 specific desired learning objectives in list format.

My answer to 'Part A'

1. I want to be proficient at the handling (including data cleaning, munging, etc…) of any kinds of data, such as uncensored data, or high-dimensional data.
2. I want to learn a variety of analyzing techniques for various kinds of data, such as Machine Learning.
3. I want to be the specialist of visualization after taking this class in this semester, because visualizing the given data well seems to be one of the most important abilities needed for good statistician !!

Part B

To this, add 3 density functions (Appendix Cassella & Berger) in centered format with equation number, i.e. format this as you would find in a journal.

My answer to ‘Part B’

My chioces were $X \overset{i.i.d}{\sim} N(\mu,\sigma^2)$, $X \overset{i.i.d}{\sim} Pareto(\alpha,\beta)$, and $X \overset{i.i.d}{\sim} Gamma(\alpha,\beta)$.

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0 \cdots (1).$$

$$f(x|\alpha,\beta) = \frac{\beta\alpha^{\beta}}{x^{\beta+1}}, a < x < \infty, \alpha > 0, \beta > 0 \cdots (2).$$

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, 0 \leq x < \infty, \alpha,\beta > 0 \cdots (3).$$

Problem 3

A quote from Donoho (1995): “an article about computational results is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result.” To the document created in Problem 4, add a summary of the steps in performing Reproducible Research in numbered list format as detailed in:
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285.

Next to each item, comment on any challenges you see in performing the step. If you are interested in learning more, a good summary of why this is important can be found in
- https://www.informs.org/ORMS-Today/Public-Articles/October-Volume-38-Number-5/Reproducible-Operations-Research
- https://doi.org/10.1093/biostatistics/kxq028
- http://statweb.stanford.edu/~wavelab/Wavelab__850/wavelab.pdf

Problem 4

Please create and include a basic scatter plot and histogram of an internal R dataset. To get a list of the datasets available use `library(help="datasets")`.

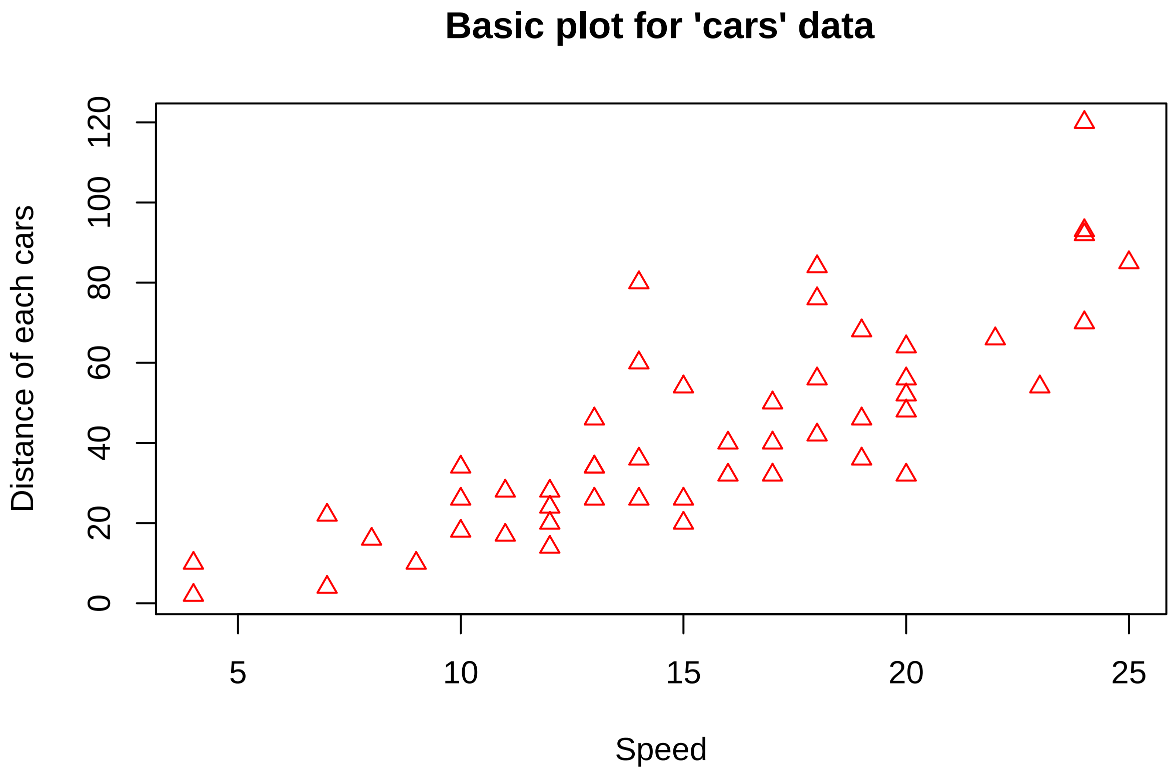
A basic scatter plot and histogram of dataset ‘cars’

```
cars # Firstly, I used the data set named 'cars', one of the internal datasets in 'R'.
```

```
##    speed dist
## 1      4    2
## 2      4   10
## 3      7    4
## 4      7   22
## 5      8   16
## 6      9   10
## 7     10   18
## 8     10   26
## 9     10   34
## 10    11   17
## 11    11   28
## 12    12   14
## 13    12   20
## 14    12   24
## 15    12   28
## 16    13   26
## 17    13   34
## 18    13   34
## 19    13   46
## 20    14   26
## 21    14   36
## 22    14   60
## 23    14   80
## 24    15   20
## 25    15   26
## 26    15   54
## 27    16   32
## 28    16   40
## 29    17   32
## 30    17   40
## 31    17   50
## 32    18   42
## 33    18   56
## 34    18   76
## 35    18   84
## 36    19   36
## 37    19   46
## 38    19   68
## 39    20   32
## 40    20   48
## 41    20   52
## 42    20   56
## 43    20   64
## 44    22   66
## 45    23   54
## 46    24   70
## 47    24   92
## 48    24   93
## 49    24  120
## 50    25   85
```

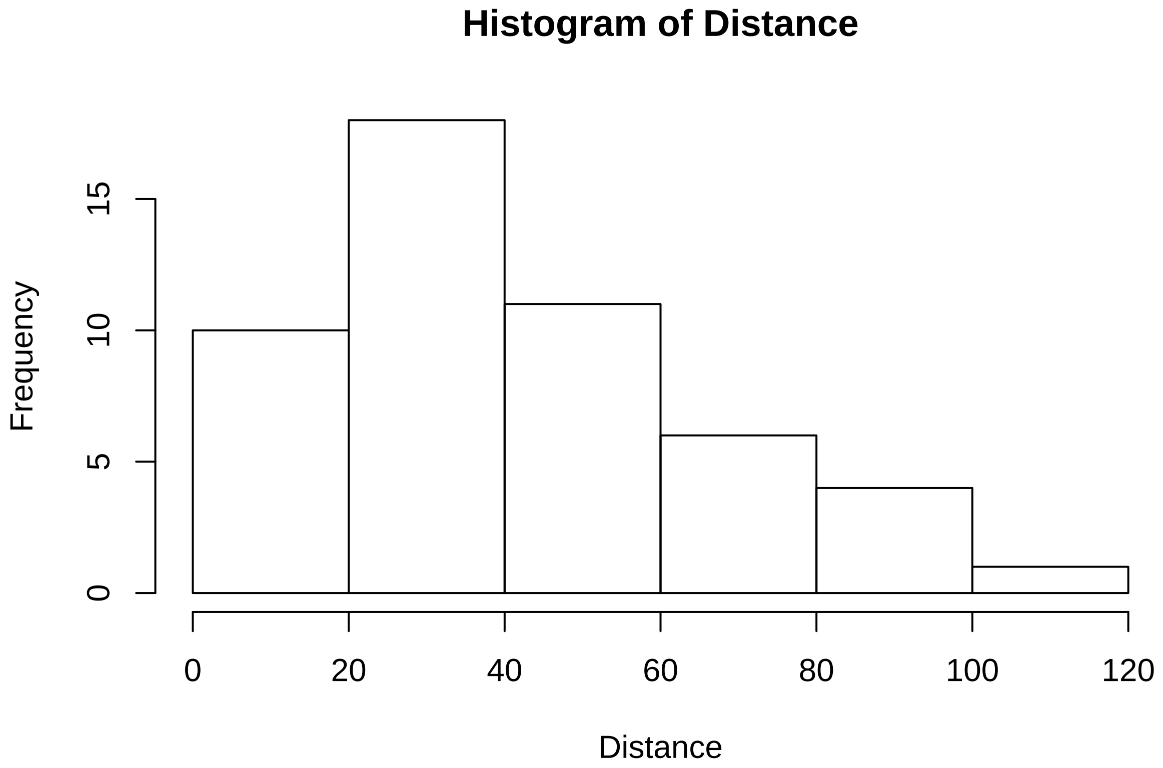
```
## This data set is composed of two column vectors, 'speed', 'Distance'.
## Total number of cars is 50, and there are speeds and Distances of cars corresponding to
## each car by row and column.

plot(x=cars$speed,y=cars$dist,xlab="Speed",ylab="Distance of each cars",pch=2,col=2,
     main="Basic plot for 'cars' data")
```



```
## What I've firstly done is the basic plotting of this dataset via using the basic
## plotting function named 'plot' with 'X'axis : speed, and 'Y'axis : Distance of each car.

hist(cars$dist,main = "Histogram of Distance",xlab="Distance")
```



```
## Plus, I attached the histogram of each car's distance.
```

Brief explanations about the above work related to 10 Rules in Donoho’s paper.

I am going to explain the above results in terms of 10 Rules as detailed in:
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285.

Firstly, I used the data set named ‘cars’, one of the internal datasets in ‘R’.

This data set is composed of **two column vectors, ’Speed’, ’Distance’**.

What I’ve done is just plotting of this dataset via using the basic plot function named **’plot’** with xaxis : speed, and yaxis : Distance of each cars.

Plus, I added the histogram of each car’s distance using **’hist’**.

These are based on **Rule 1: For Every Result, Keep Track of How It Was Produced**,

Rule 7 :Always Store Raw Data behind Plots,

Rule 9: Connect Textual Statements to Underlying Results,

Rule 10: Provide Public Access to Scripts, Runs, and Results.