Homework 1

Zhiyuan Du

8/23/2020

This is the homework 1 of ***Zhiyuan Du***. My PID is ***zhiyuan***.

Homework 1

Zhiyuan Du

08/24/2020

Problem 2

I have a basic understanding of the R environment after checking out the lectures on the canvas. And I am so excited to set off my R journey and cannot wait to use R Markdown to do my homeworks. There are two parts in the Porblem 2.

Problem 2 Part A

There are three **desired learning objectives** that I hope to get out of this class:

- Be proficient about using *R Markdown* through the practice.
- Put the *Reproducible Research Concepts* into the real projects.
- Learn to make good *Data cleaning and munging*.

Problem 2 Part B

1. Normal distribution density function:

$$f(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{2\sigma^2}\right)}; -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0 \tag{1}$$

2. Binomial distribution density function:

$$P(X = x|n,p) = \binom{n}{y} p^y (1-p)^{n-y}, x = 0,1,2,...,n; 0 \leq p \leq 1 \tag{2}$$

3. Poisson distribution density function:

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}; x = 0,1,...; 0 \leq \lambda < \infty \tag{3}$$

Problem 3

The 10 rules revealed in the article vividly express what the reproducible computational research is and how to make it perfect. Here are some basic steps I summarized on executing a reproducible compuational research.

- After getting the raw data, make them stored. In this case, each given figure is easily fetched and the plotting procedure could be slightly modified to get results.
- Do the execution and data manipulation steps on the program instead of manually and archive the exact version of the program.
- Record the intermediate results possibly in Standardized Formats.
- Use a vision control system to store all of the scripts.
- Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected.
- Give statements to all of the underlying results.
- Make public access to all input data, scripts, versions, parameters, and intermediate results.

Problem 4

In this problem, I chose the dataframe *'mpg'* in the package of *'ggplot2'*. In the daraframe, there are 11 variables including *'manufacturer'*, *'model'*, *'trans'*, *'drv'*, *'fl'* and *'class'* as the catergorical variable and *'disp'l*, *'year'*, *'cy'l*, *'cty'*, *'hwy'* as the continuous variable.

For the scattter plot, I choose *'year'* as the x-axis and *'displ'* as the y-axis with model added to show the difference within the model.

```
#####
##### Problem 4: scatter plot #####
#####
ggplot(data= mpg)+
  geom_point(mapping = aes(x = hwy, y = displ , shape = class))
```

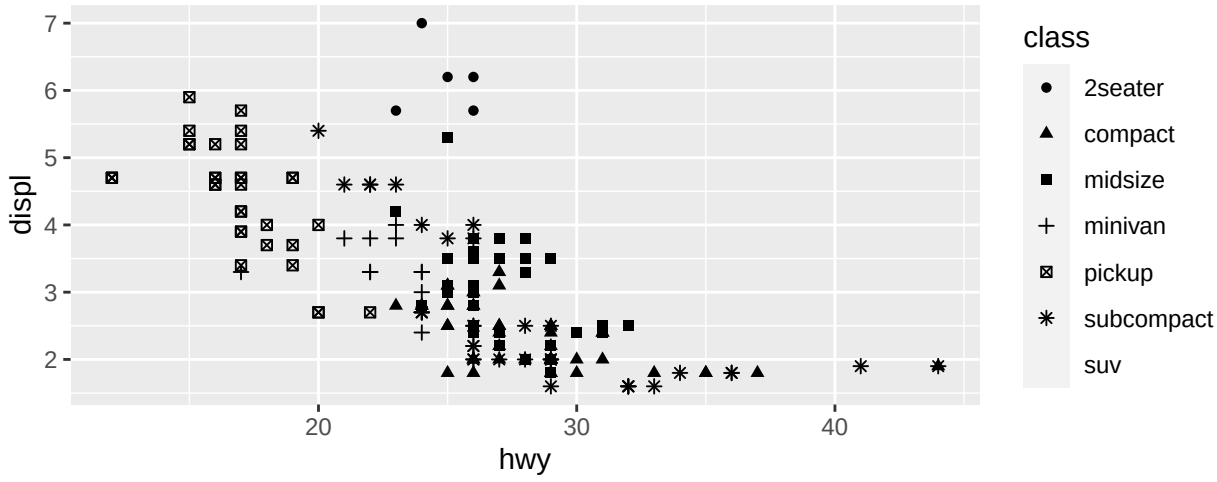

Figure 1: A scatter plot.

From the scatter plot, we can obviously achieve a negative trend between the *'hwy'* and *'displ'*. For the histogram, I choose to use *'model'* as the x-axis and *'hwy'* as the y-axis.

```
#####
##### Problem 4: histogram #####
#####
hist(mpg$hwy, main = "Histogram for highway", xlab="highway")
```

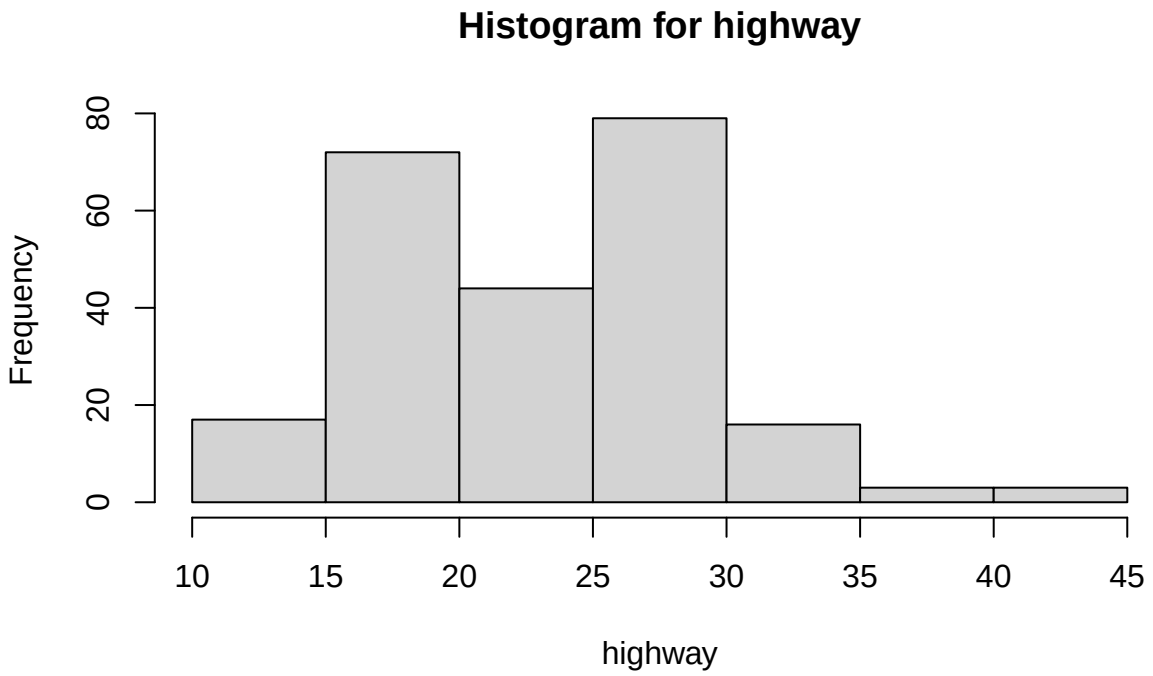

Figure 2: A histogram.

Appendix 1: R code

```
#####
#####
##### Project setup #####
#####
#####
#####
knitr::opts_chunk$set(echo = TRUE, fig.pos = "H", out.extra = "")
library(ggplot2)
#####
#####
#####
##### Problem 4: scatter plot #####
#####
ggplot(data= mpg)+
  geom_point(mapping = aes(x = hwy, y = displ , shape = class))

#####
```

```r
##### Problem 4: histogram #####
#####
hist(mpg$hwy, main = "Histogram for highway", xlab="highway")
```