HW1_JingBin

Jingbin Xu

2020-09-01

Problem 1

I have finished the Primers labeled as The Basics on Rstudio.cloud.

Problem 2

Part A

People enrolled in this course have a different background. Before I moved to VT, I worked as a Biostatistician and Data manager at LA after I graduated from UC Irvine. I am interested in the following topics:

- 1. How to present high-quality graphs (informative, vivid and easy for all audience)
- 2. Parallel computing (How to break down the assignments to achieve the max efficiency?)
- 3. Handling the data on cloud (I am not sure whether this course would cover the topic, but with the advanced data era, we always expect more than we want.)

Part B

Beta distribution:

$$f(x \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad 0 \leq x \leq 1, \quad \alpha > 0, \quad \beta > 0 \tag{1}$$

Normal distribution:

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

Exponential distribution:

$$f(x \mid \beta) = \frac{1}{\beta} e^{-x/\beta}, \quad 0 \leq x < \infty, \quad \beta > 0 \tag{3}$$

Problem 3

A good start of conducting the data analysis is to have general guidelines/design. The dataset named **mtcars** is very classic and has been written in many trending textbooks. So I am conducting the analysis based on this dataset. Before I perform my analysis, I carefully read the **10 simple rules of reproducible research**. And here are my take-away notes:

Always visualize the data first

I prefer to use a summary function to check each column, and then I will run a scatterplot to identify potential outliers and errors. Based on this preliminary analysis, we could achieve two goals. The first goal is to avoid manual data manipulation steps (for instance, with the analysis, we have a formal and trackable justification for removing outliers). The second goal is to avoid jumping into conclusions too early. Sometimes people go straight to the destination may not be a wise step. As for data analysis, there always lies the trap for careless drivers.

Smart variable name

I enjoy naming my variable in the program with the same pattern and comment on every critical step I made. So when someone reviews my code, they could follow my logic flow and be more efficient.

Be careful with the plot save function in R

There are multiple ways to save the graph in R. However, we do not cover that part. But when I read about **always store raw data behind plots**, it suddenly came to my mind. I did not find a better way to automatically keep the graph on the cloud or at the local device in R. So it is tough to version control the graphs we made (I have tried some automated algorithms but it all does not work very well. So if someone knows, please let me know)

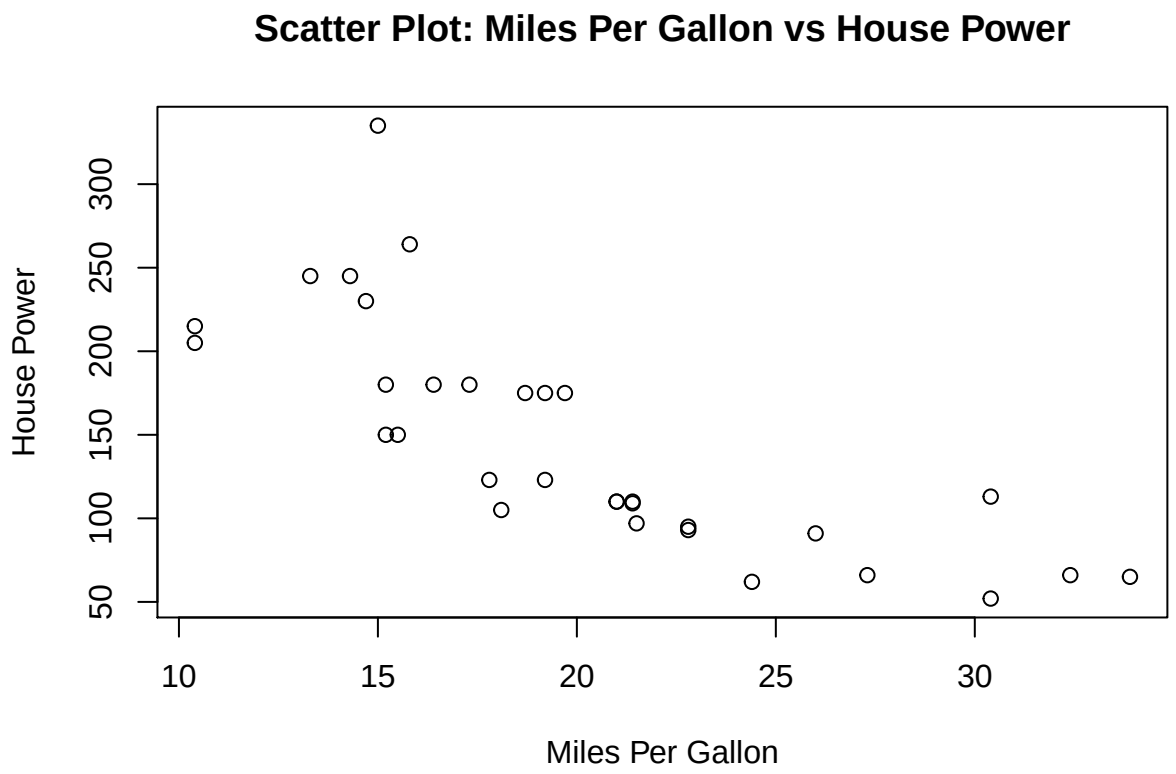
Problem 4

```
# Loading the car dataset
datasets::mtcars
```

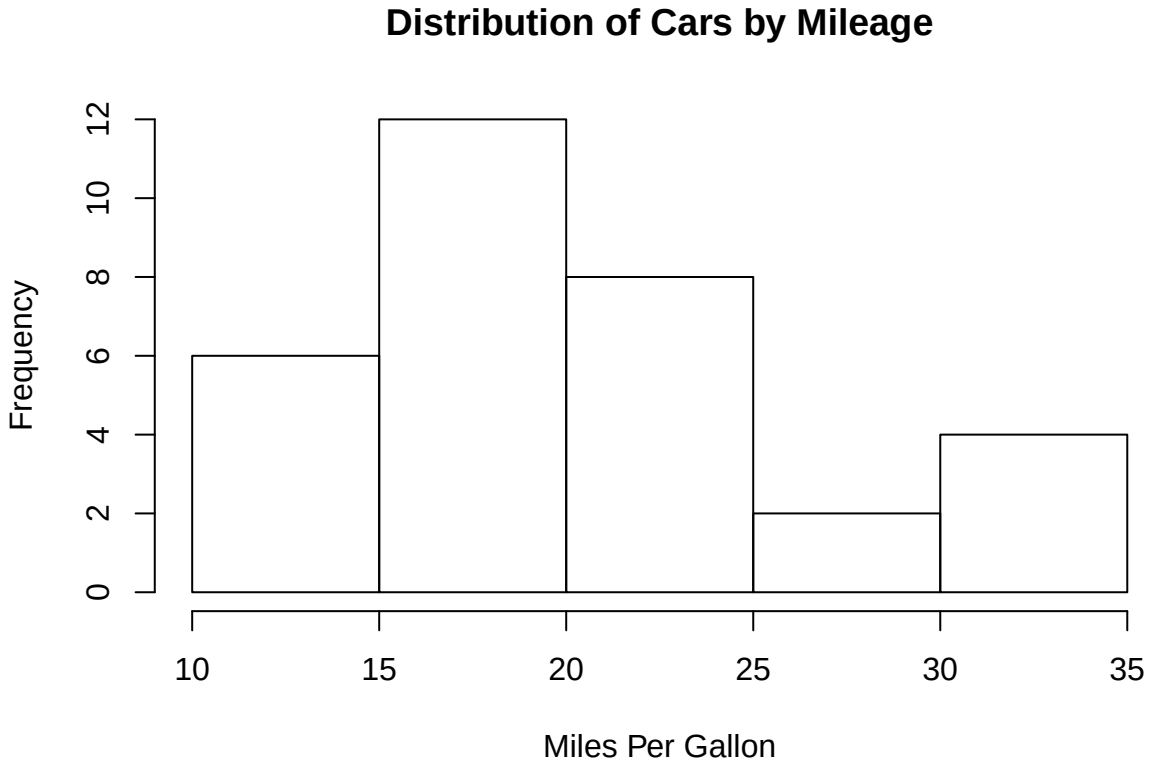
```
##                      mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
```

```
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

```r
# Scatter plot for mpg vs hp
graph_scatter = plot(mtcars$mpg,
     mtcars$hp,
     main = "Scatter Plot: Miles Per Gallon vs House Power",
     xlab = "Miles Per Gallon",
     ylab = "House Power")
```




```r
# Histogram for mpg
graph_hist = hist(mtcars$mpg,
     main = "Distribution of Cars by Mileage",
     xlab = "Miles Per Gallon",
     ylab = "Frequency")
```


Distribution of Cars by Mileage
Frequency
0
2
4
6
8
10
12
10
15
20
25
30
35
Miles Per Gallon