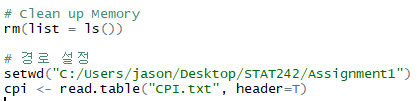
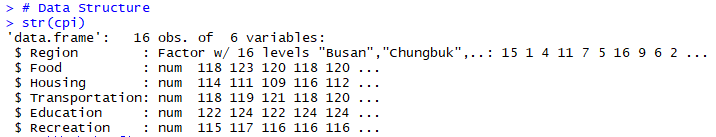
**Statistical Method in Social Science – Assignment 1**

**통계학과 17학번 이재승**

If you want to use R with your own data, you have to set the directory and define your own data.

Let me know about the data structure.

**> str(cpi) # Data Structure**

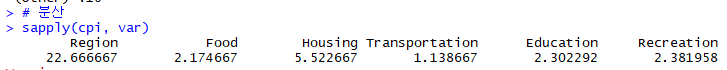
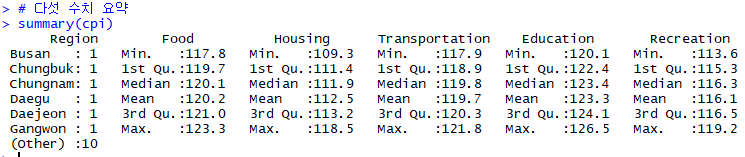


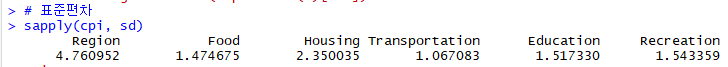
There are 16 objects of 6 variables. 16 objects are ‘Seoul’, ‘Busan’, ‘Daegu’, ‘Incheon’, ‘Gwangju’, ‘Daejeon’, ‘Ulsan’, ‘Gyeonggi’, ‘Gangwon’, ‘Chungbuk’, ‘Chungnam’, ‘Jeonbuk’, ‘Jeonanam’, ‘Gyeongbuk’, ‘Gyeongnam’ and ‘Jeju’. 6 variables are ‘Region’, ‘Food’, ‘Housing’, ‘Transportation’, ‘Education’ and ‘Recreation’. This data can be described by 16 x 6 Matrix.

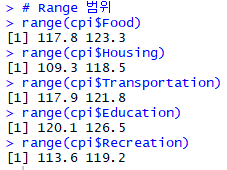
Now, let me know about the 5 number summary(Min, Q1, Q2, Q3, Q4 and Max), Mean, Variance and Standard Deviation.

**> summary(cpi) # 5 number summary**

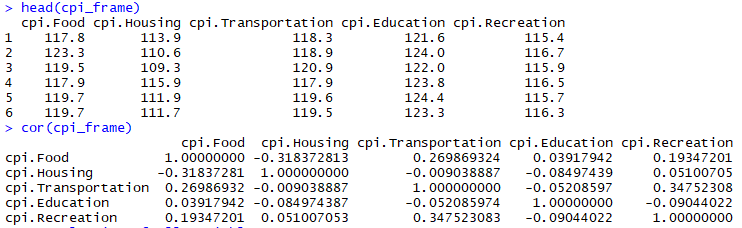
**> sapply(cpi, var); sapply(cpi, sd); # Variance and Standard Deviation**



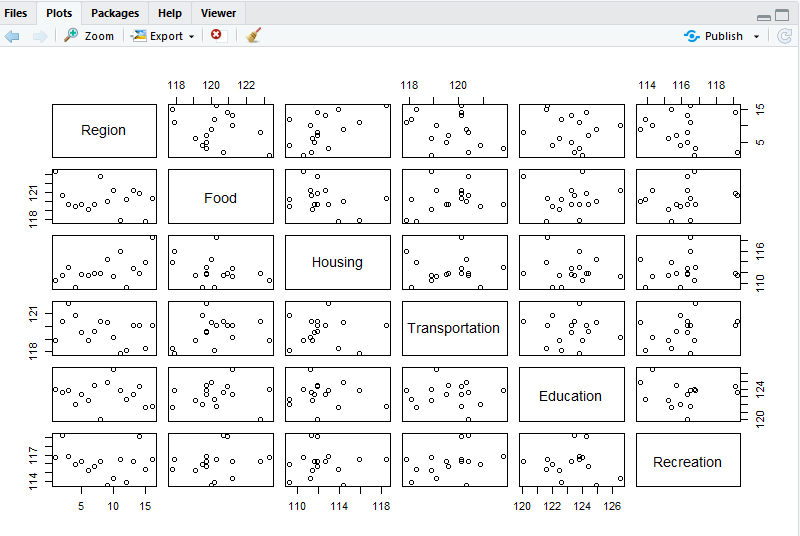


By entering these codes, we can see the Q1, Q2, Q3, Q4, Mean, Variance and Standard Deviation of cpi. The biggest value of cpi is 123.3, which is a data of (Busan, Food). **It means that Busan’s Food price was shaked strongly in 2006.** The biggest variance of cpi is Housing’s 5.522667. **We can translate that in 2006, Housing prices were various in Korea.** On the other hand, the smallest value of cpi is 109.3, which is a data of (Daegu, Housing). **It means that Daegu’s Housing price was shaked weakly in 2006.** Also, the smallest variance of cpi is Transportation’s 1.138667. **It means that in 2006, Transportation prices were not varied in Korea.** How about range? Range is a difference between Maximum and Minimum. We can see range by code ‘**> range(cpi$Food); range(cpi$Housing); range(cpi$Transportation); range(cpi$Education); range(cpi$Recreation);**’. The biggest range is ‘Housing’, which is 9.2. **In advance, I said that Housing’s variance and standard deviation are biggest. In addition, Housing’s Range is the biggest among 5 variables.**

Since I want to know the Correlation between all variables, I made data frame for each variable by entering code such as ‘> food <- data.frame(cpi$Food)’. **After making all variables to data frame, I entered ‘> cor(cpi\_frame)’, which can see the correlation between all variables.**



**Since all abstract values of the correlations are not more than 0.5, there are not strong correlation between all variables. Nevertheless, the biggest abstract value of the correlation is 0.347523083, which is a correlation between Transportation and Recreation.** **Furthermore, it is a positive value.** When I consider this fact, I guess that when we enjoy recreation, travel takes big possession of recreation. And to travel, we usually use public transportation. So, I think that there is some correlation between Transportation and Recreation. Except this correlation, the abstract value of correlation between Food and Housing takes second big value among all correlations. And it is negative value. When we consider this information, we can translate many things.

**You can see all data set by entering code ‘> plot(cpi)’.** Additional information is at the Appendix. You can see numerical and graphical summaries at the appendix. There are Stem and leaf plot, Histogram, Box plot, Plot, and etc.

**Appendix**

# CPI

# Clean up Memory

rm(list = ls())

# 경로 설정

setwd("C:/Users/jason/Desktop/STAT242/Assignment1")

cpi <- read.table("CPI.txt", header=T)

# Data Structure

str(cpi)

'data.frame': 16 obs. of 6 variables:

$ Region : Factor w/ 16 levels "Busan","Chungbuk",..: 15 1 4 11 7 5 16 9 6 2 ...

$ Food : num 118 123 120 118 120 ...

$ Housing : num 114 111 109 116 112 ...

$ Transportation: num 118 119 121 118 120 ...

$ Education : num 122 124 122 124 124 ...

$ Recreation : num 115 117 116 116 116 ...

# 위에 6개

head(cpi)

Region Food Housing Transportation Education Recreation

1 Seoul 117.8 113.9 118.3 121.6 115.4

2 Busan 123.3 110.6 118.9 124.0 116.7

3 Daegu 119.5 109.3 120.9 122.0 115.9

4 Incheon 117.9 115.9 117.9 123.8 116.5

5 Gwangju 119.7 111.9 119.6 124.4 115.7

6 Daejeon 119.7 111.7 119.5 123.3 116.3

# 다섯 수치 요약

summary(cpi)

Region Food Housing Transportation Education Recreation

Busan : 1 Min. :117.8 Min. :109.3 Min. :117.9 Min. :120.1 Min. :113.6

Chungbuk: 1 1st Qu.:119.7 1st Qu.:111.4 1st Qu.:118.9 1st Qu.:122.4 1st Qu.:115.3

Chungnam: 1 Median :120.1 Median :111.9 Median :119.8 Median :123.4 Median :116.3

Daegu : 1 Mean :120.2 Mean :112.5 Mean :119.7 Mean :123.3 Mean :116.1

Daejeon : 1 3rd Qu.:121.0 3rd Qu.:113.2 3rd Qu.:120.3 3rd Qu.:124.1 3rd Qu.:116.5

Gangwon : 1 Max. :123.3 Max. :118.5 Max. :121.8 Max. :126.5 Max. :119.2

(Other) :10

# 분산

sapply(cpi, var)

Region Food Housing Transportation Education Recreation

22.666667 2.174667 5.522667 1.138667 2.302292 2.381958

# 표준편차

sapply(cpi, sd)

Region Food Housing Transportation Education Recreation

4.760952 1.474675 2.350035 1.067083 1.517330 1.543359

# Range 범위

range(cpi$Food)

[1] 117.8 123.3

range(cpi$Housing)

[1] 109.3 118.5

range(cpi$Transportation)

[1] 117.9 121.8

range(cpi$Education)

[1] 120.1 126.5

range(cpi$Recreation)

[1] 113.6 119.2

# 줄기 잎 그림

stem(cpi$Food)

The decimal point is at the |

116 | 89

118 | 15777

120 | 0237922

122 | 83

stem(cpi$Housing)

The decimal point is at the |

108 | 33

110 | 63457999

112 | 709

114 | 49

116 |

118 | 5

stem(cpi$Transportation)

The decimal point is at the |

117 | 9

118 | 1399

119 | 156

120 | 1113449

121 | 8

stem(cpi$Education)

The decimal point is at the |

120 | 167

122 | 05633588

124 | 0349

126 | 5

stem(cpi$Recreation)

The decimal point is at the |

112 | 69

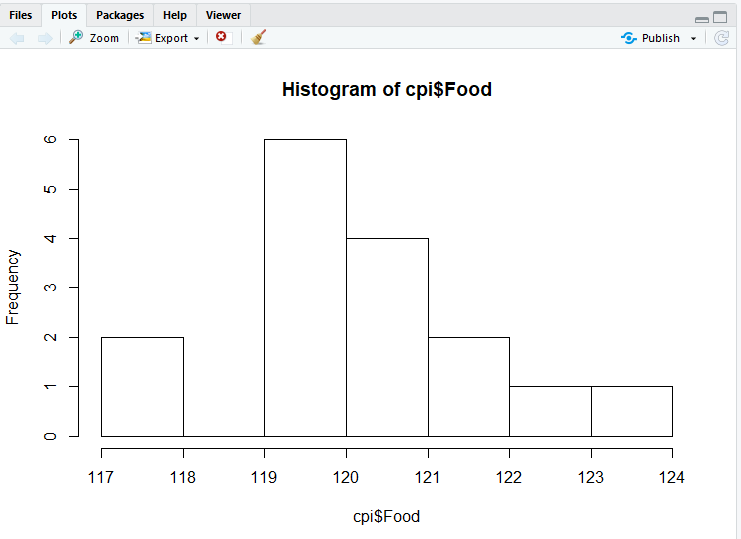
114 | 32479

116 | 3335578

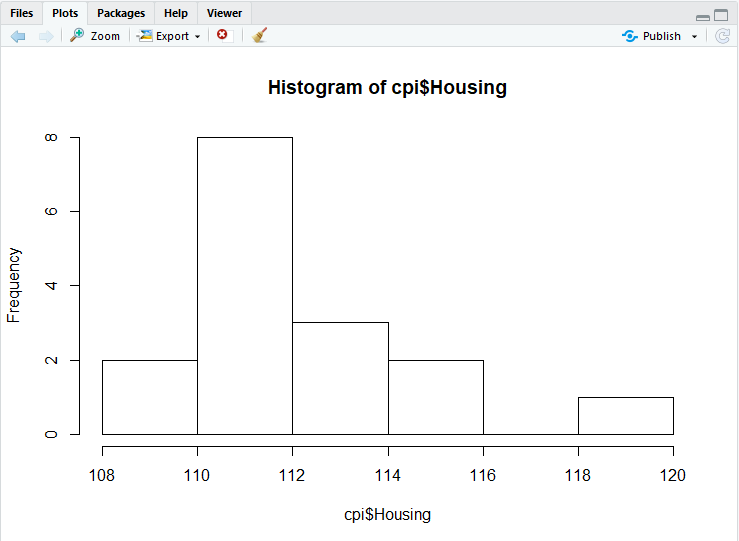
118 | 12

# 히스토그램

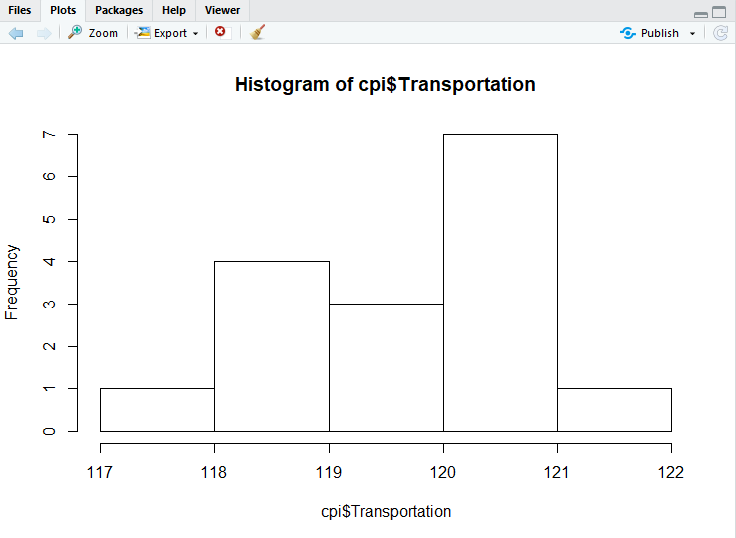
hist(cpi$Food)



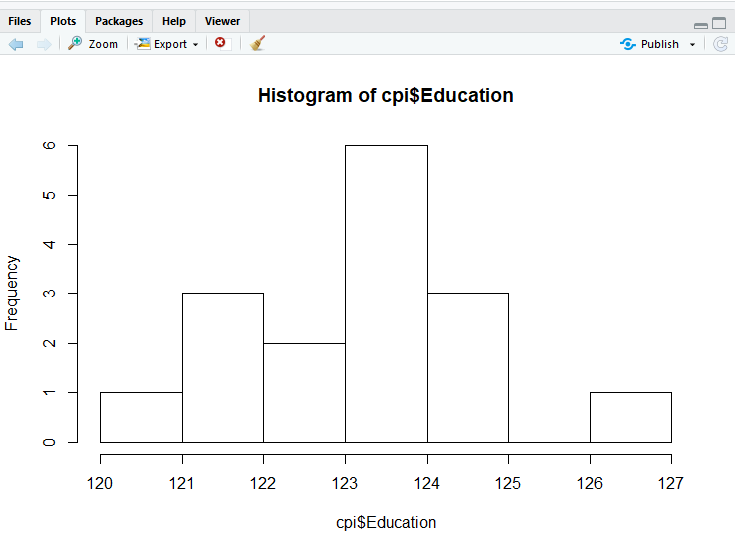
hist(cpi$Housing)



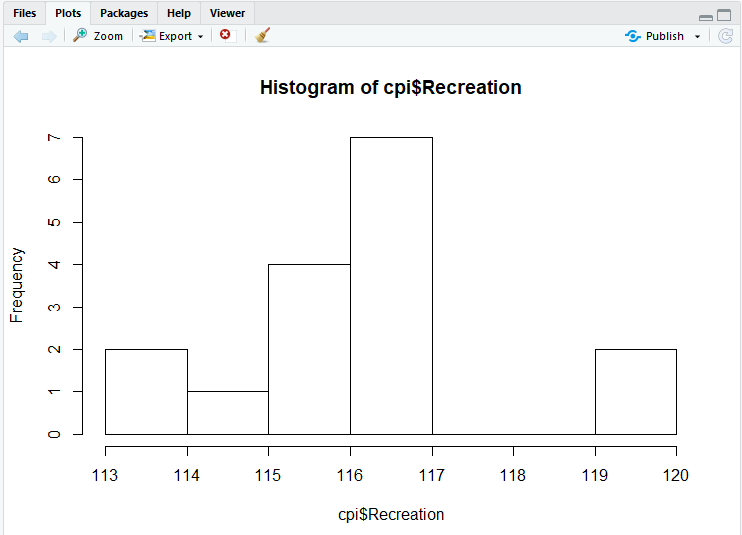
hist(cpi$Transportation)



hist(cpi$Education)



hist(cpi$Recreation)



# 상관계수

head(cpi$Food)

> head(food)

cpi.Food

1 117.8

2 123.3

3 119.5

4 117.9

5 119.7

6 119.7

# Make a DataFrame

food <- data.frame(cpi$Food)

housing <- data.frame(cpi$Housing)

transportation <- data.frame(cpi$Transportation)

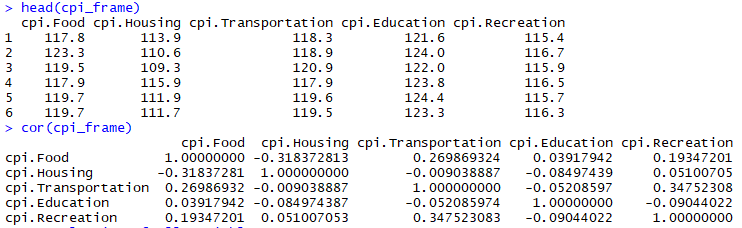
education <- data.frame(cpi$Education)

recreation <- data.frame(cpi$Recreation)

# Data Frame Matrix

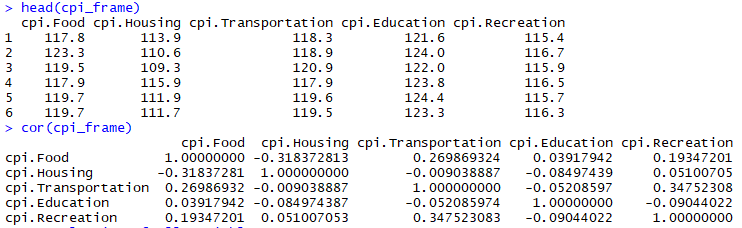
cpi\_frame <- data.frame(food, housing, transportation, education, recreation)

head(cpi\_frame)



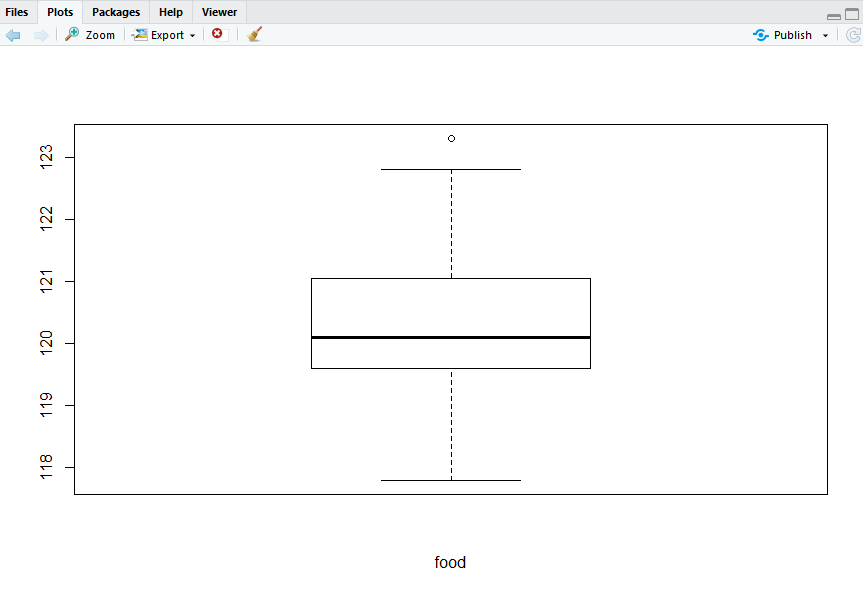
# Corrleation of all Variables

cor(cpi\_frame)

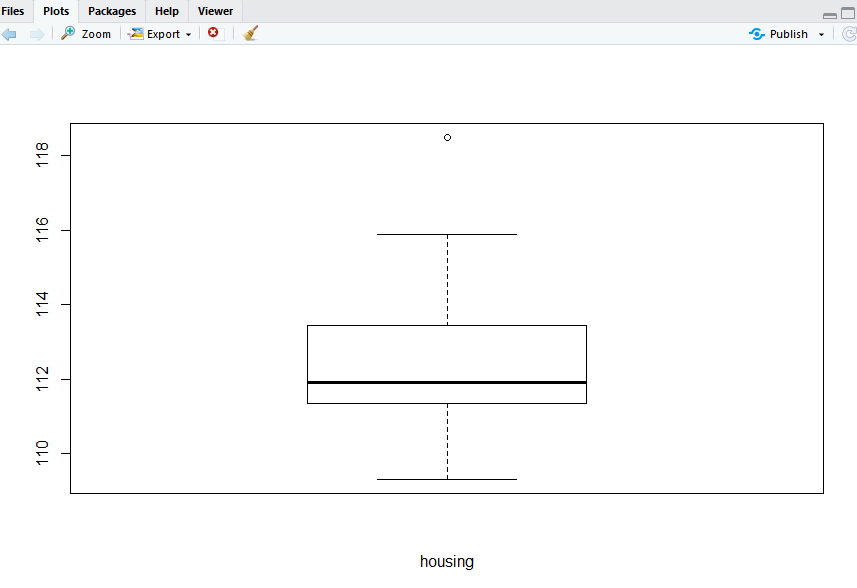


# Boxplot

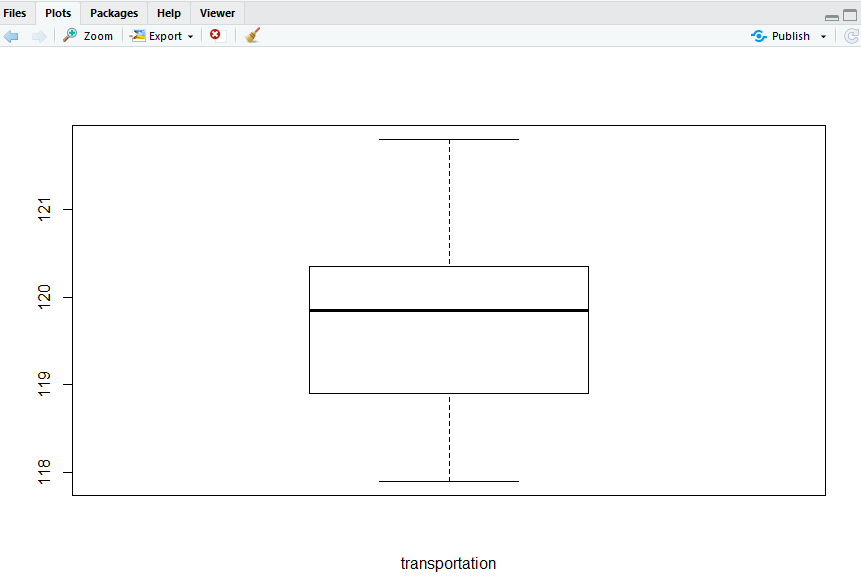
boxplot(food, xlab='food')



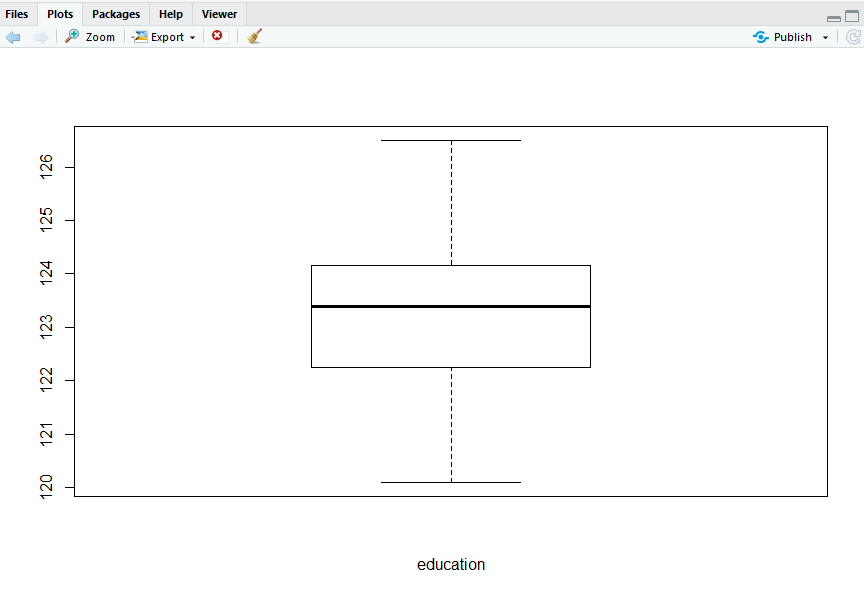
boxplot(housing, xlab = 'housing')



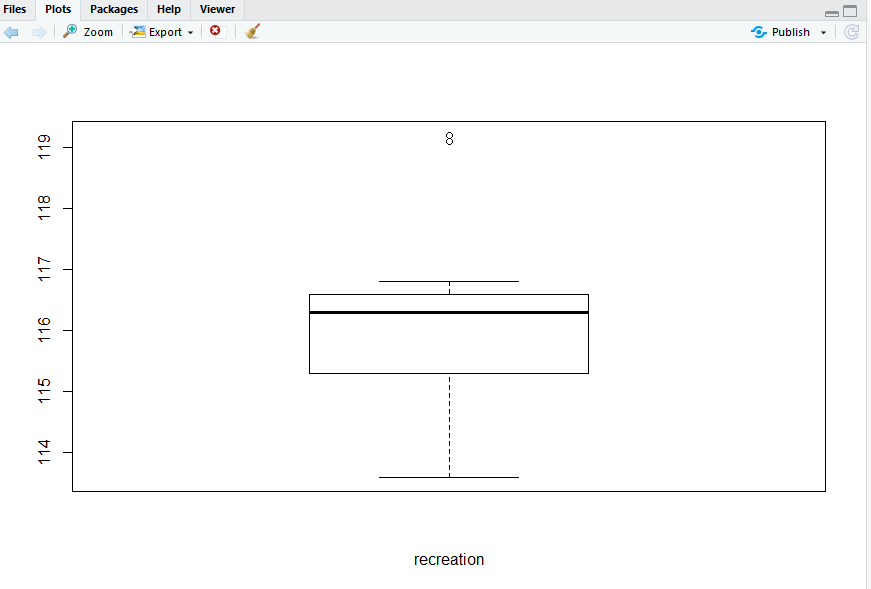
boxplot(transportation, xlab = 'transportation')



boxplot(education, xlab = 'education')

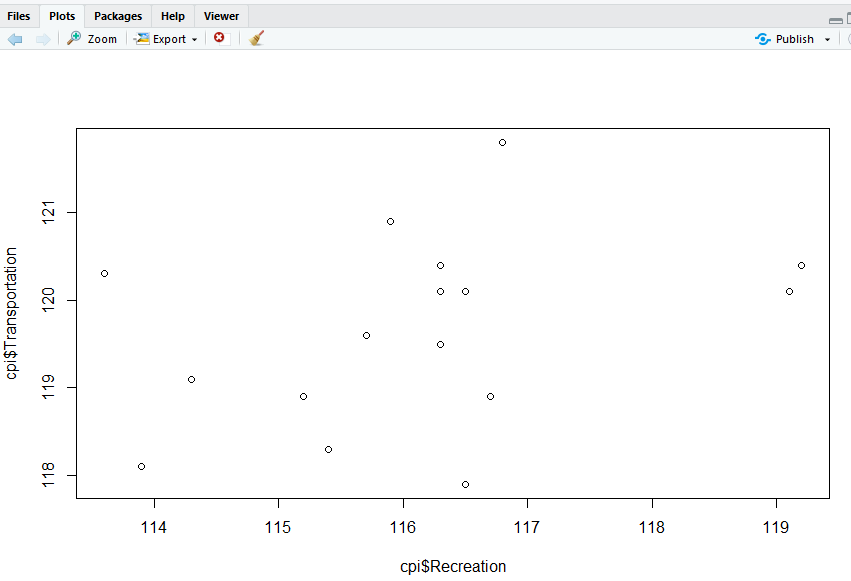


boxplot(recreation, xlab = 'recreation')

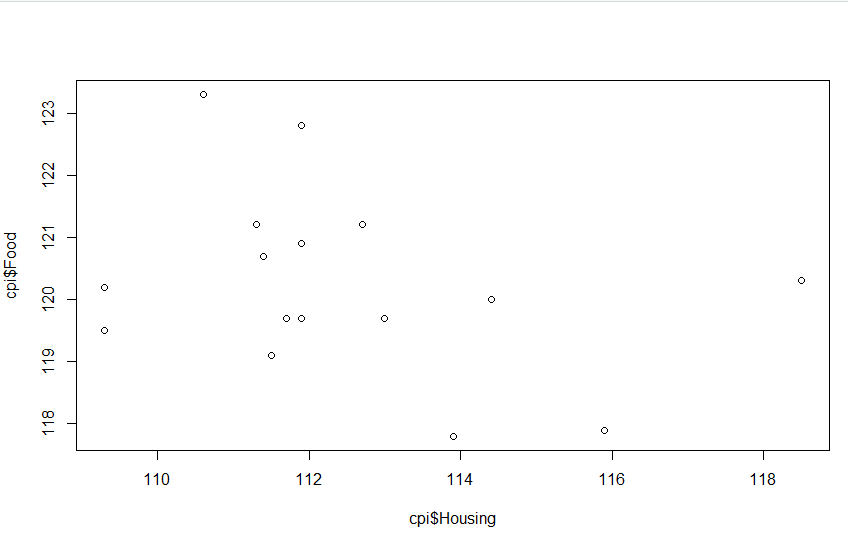


# Plot

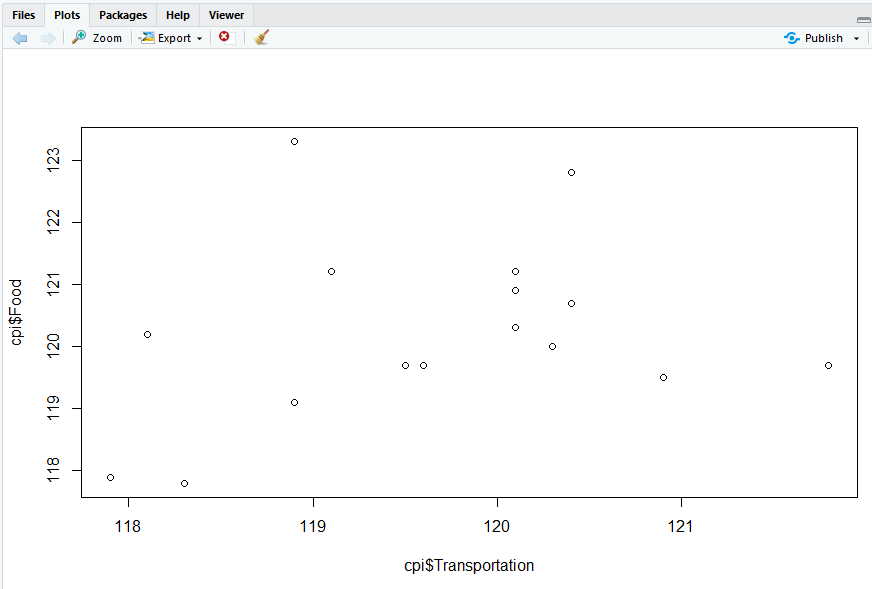
plot(cpi$Recreation, cpi$Transportation)



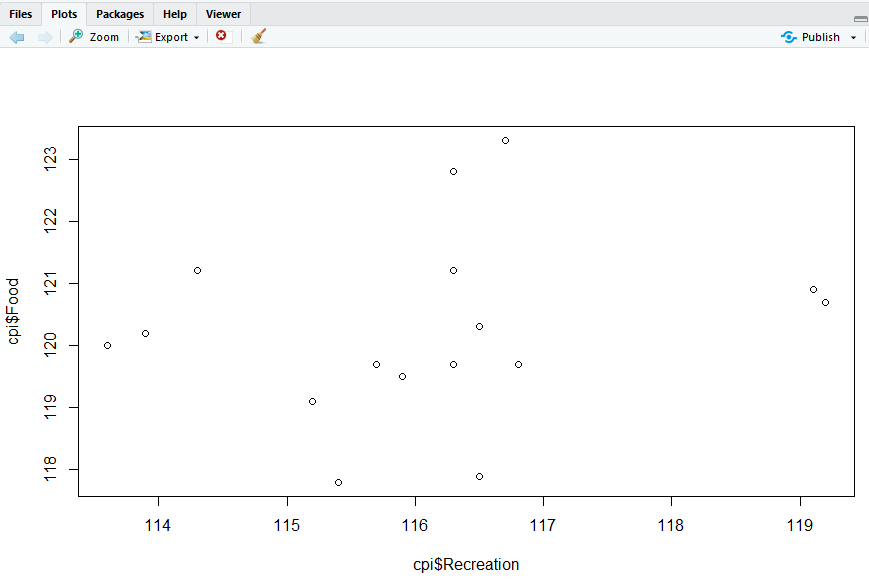
plot(cpi$Housing, cpi$Food)



plot(cpi$Transportation, cpi$Food)



plot(cpi$Recreation, cpi$Food)



plot(cpi)

