Solutions to Homework 1

-
  ```
  rm(list = ls()) # clean up memory
  setwd("path/of/your/working/directory") # change the path accordingly
  cpi <- data <- read.table("CPI.txt", header = T) # first line contains variable names
  str(cpi) # check the structure
  head(cpi) # first few lines
  ```

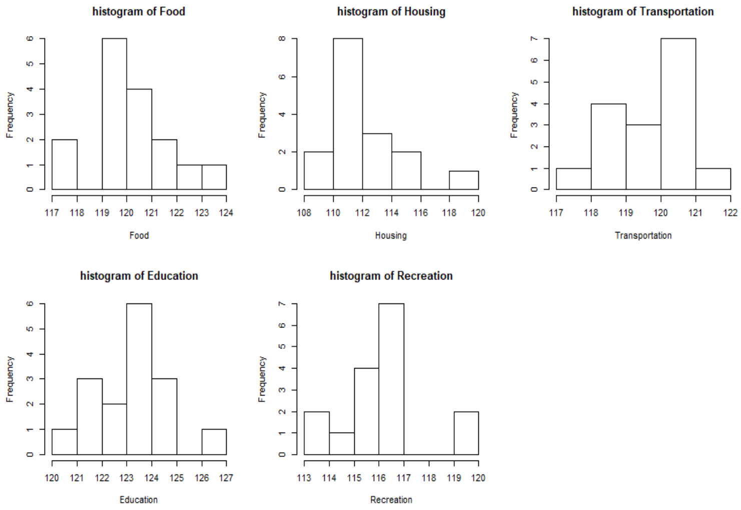
1. Summary of Data by numerical values

   ```
   summary(data)
   apply(data, 2, var)
   ```

- The result of ’summary()’ function is the summary statistics of the data. (Minimum, 1st Quantile, Median, Mean, 3rd Quantile, Maximum)
- By using ’apply()’ function, the variance of each variables can be calculated.
- The variable ’Housing’ has the maximum value of variance which is 5.22667, and the variable ’Transportation’ has the minimum value of variance which is 1.138667.

2. Histogram

   ```
   dev.off()
   par(mfrow=c(2, 3))
   for (i in 1:ncol(data)){
     hist(data[, i], main=paste("histogram␣of", colnames(data)[i]),
     xlab=colnames(data)[i])
   }
   ```

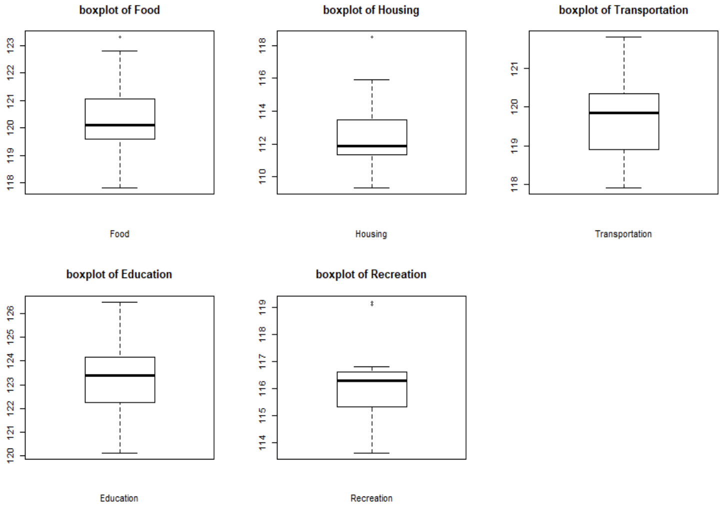


- Using for loop, the histogram of each variables in the data were produced.

- Using 'hist()' function, it is possible to overview the distribution of each variables in the data.
- The distribution of 'Transportation' and 'Education' has little skewness.
- The distribution of 'Food' and 'Housing' is skewed to right.
- The distribution of 'Recreation' is skewed to left.

3. Boxplot

```
opar <- par(no.readonly=TRUE)
par(opar)
for (i in 1:ncol(data)){
  boxplot(data[, i], main=paste("boxplot␣of", colnames(data)[i]),
  xlab=colnames(data)[i])
}
```



- Using 'boxplot()' function, it is possible to overview the distribution of each variables in the data and also detect outliers.
- The x-axis represents the variable and the y-axis represents the value of CPI.
- The variable 'Food' and 'Housing' has one outlier and the variable 'Recreation' has 2 outliers.
- Also the skewness of the data can be checked out in the boxplot.

4. Correlation

4.1. Correlation plots and plots of pairs

```
dev.off()
library(corrplot)
data.cor <- cor(data)
corrplot(data.cor)
pairs(data, col=c("red", "blue"), pch=19, main="plots␣of␣pairs␣between␣items")
```

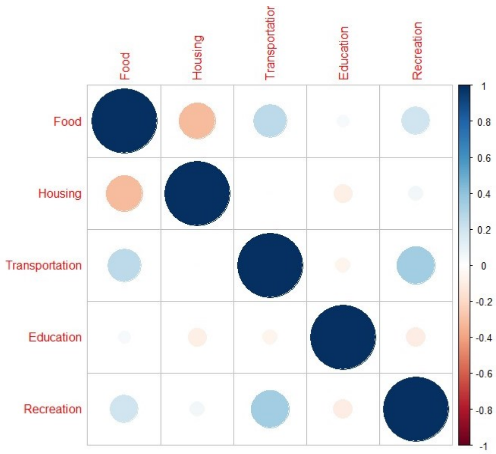


Figure 1: correlation plots

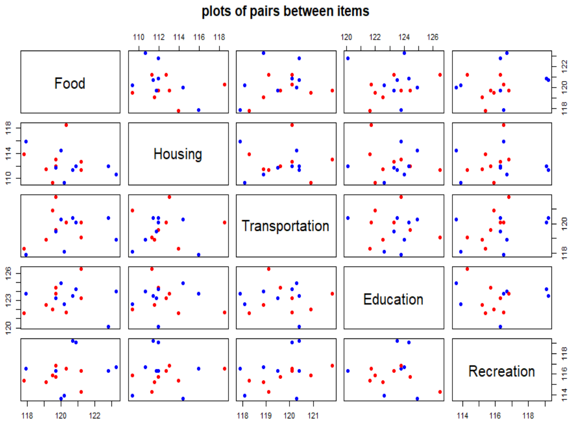


Figure 2: plots of pairs between items

- Refer to Figure 1 and Figure 2.
- The relationship between transportation and food is positive correlation.
- The relationship between transportation and recreation is positive correlation.
- The relationship between housing and food is negative correlation.

4.2. Plots of correlated variables with regions

```
# (transportation-food / transportation - recreation / housing - food)
with(data, plot(Transportation ~ Food, type="n",
             main = "plot␣of␣cpi␣between␣transportation␣and␣food"))
with(data, text(Food, Transportation, rownames(data),
             cex=0.8, col=c(1:nrow(data))))

with(data, plot(Transportation ~ Recreation, type="n",
             main = "plot␣of␣cpi␣between␣transportation␣and␣recreation"))
with(data, text(Recreation, Transportation, rownames(data),
             cex=0.8, col=c(1:nrow(data))))

with(data, plot(Housing ~ Food, type="n",
             main = "plot␣between␣housing␣and␣food"))
with(data, text(Food, Housing, rownames(data),
             cex=0.8, col=c(1:nrow(data))))
```

- Refer to Figure 3
- By making plots with the regions, the relationship between highly correlated variables and region can be confirmed.

5. Dot chart

```
par(mfrow=c(1, 2))
item <- colnames(data)
```

```
region <- rownames(data)
for (i in 1:ncol(data)){
  order.1 <- order(data[, i], decreasing=F)
  dotchart(data[order.1,][, i], labels=region[order.1], xlab=item[i],
           cex=0.8, main=paste("regional␣rank␣of", item[i]))
}
```

- Refer to Figure 4
- Item - (maximum CPI region, minimum CPI region)
- Food - (Busan, Seoul)
- Housing - (Ulsan, Daegu)
- Transportation - (Chungnam, Incheon))
- Education - (Gyeongnam, Gyeongbuk)

5-1. Minimum and maximum value of CPI in each variable and the region

```
min.item <- matrix(colnames(data), nrow=ncol(data), ncol=3)
max.item <- matrix(colnames(data),nrow=ncol(data), ncol=3)
for (i in 1:ncol(data)){
  min.item[i, 3] <- rownames(data)[which.min(data[, i])]
  max.item[i, 3] <- rownames(data)[which.max(data[, i])]
}
min.item[, 2] <- apply(data, 2, min)
max.item[, 2] <- apply(data, 2, max)
min.item
max.item
```

		min		max	
1	Food	117.8	Seoul	123.3	Busan
2	Housing	109.3	Daegu	118.5	Ulsan
3	Transportation	117.9	Incheon	121.8	Chungnam
4	Education	120.1	Gyeongbuk	126.5	Gyeongnam
5	Recreation	113.6	Gyeonggi	119.2	Chungbuk

6. Regional comparison

6.1. Regional comparison by variation

```
data.city=data[1:7,]
data.country=data[8:16,]
apply(data.city,2,mean)
apply(data.country,2,mean)
```

- The mean of ’Housing’ and ’Recreation’ CPIs are larger in city than country.
- The mean of ’Food’, ’Transportation’ and ’Education’ CPIs are larger in country than the city

6.1. Regional comparison by variation

```
apply(data.city,2,var)
apply(data.country,2,var)
```

- The variance of 'Housing' CPI is larger in the city than country.
- The variance of 'Recreation' CPI is larger in the country than city.

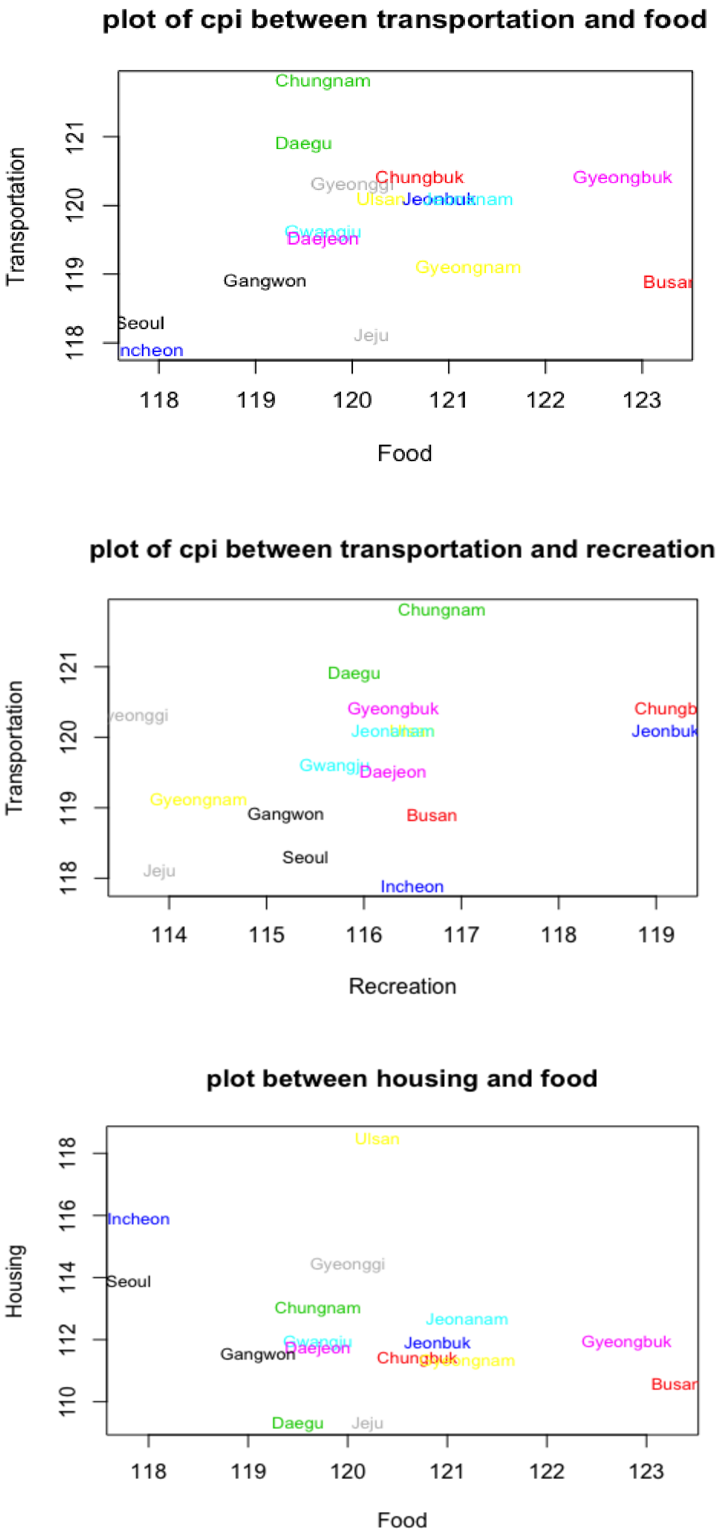
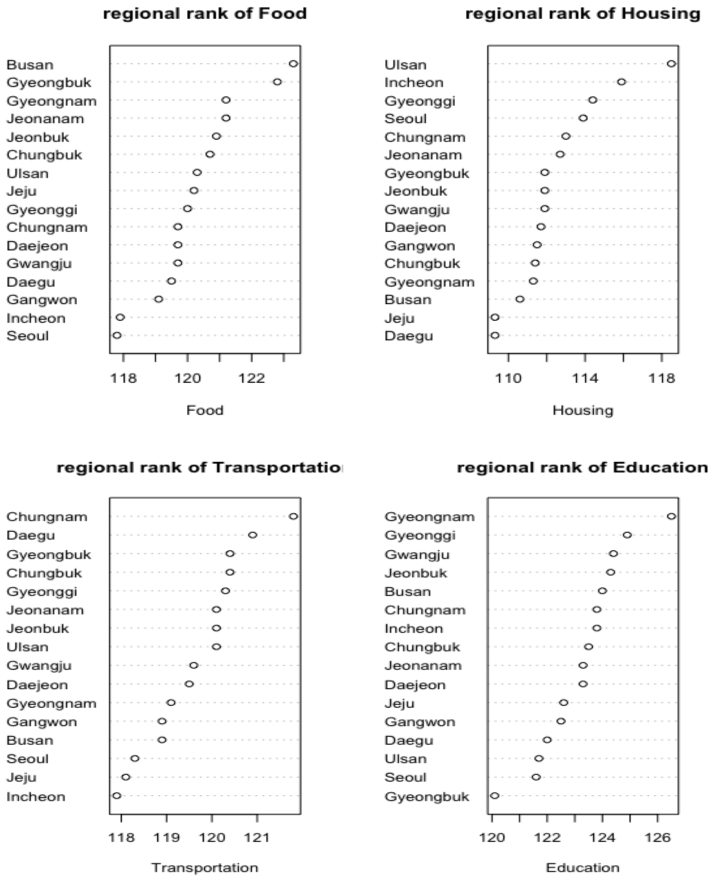


Figure 3: Plots between highly correlated variables with region



regional rank of Food

regional rank of Housing

regional rank of Transportatio

regional rank of Education

Figure 4: Dot plots with region rank