HW8

Getong Zhong

2023-04-20

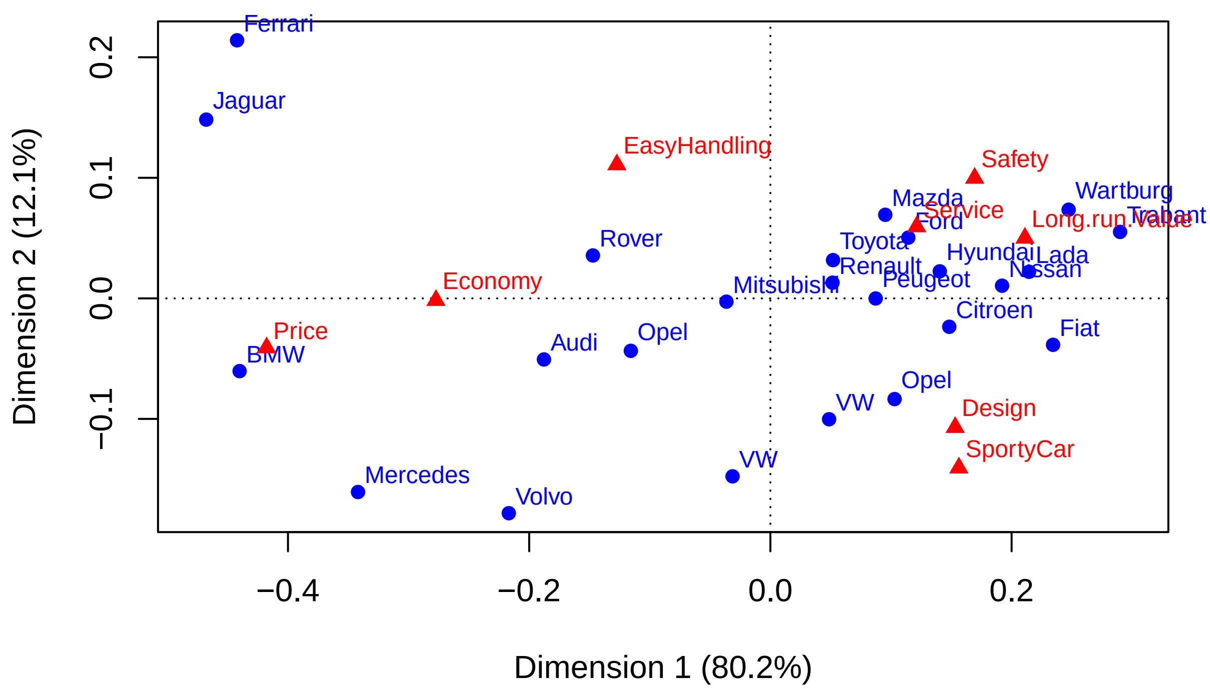
Exercise 1

```
library(ca)
cars <- read.table("C:/Users/tonyg/Desktop/Academic/Grad/HUDM 6122/cars.txt", header = TRUE)
cars
```

```
##          Type   Model Economy Service Long.run.Value Price Design SportyCar
## 1        Audi     100     3.9     2.8            2.2   4.2    3.0       3.1
## 2         BMW 5series     4.8     1.6            1.9   5.0    2.0       2.5
## 3     Citroen      AX     3.0     3.8            3.8   2.7    4.0       4.4
## 4     Ferrari     N/A     5.3     2.9            2.2   5.9    1.7       1.1
## 5        Fiat     Uno     2.1     3.9            4.0   2.6    4.5       4.4
## 6        Ford  Fiesta     2.3     3.1            3.4   2.6    3.2       3.3
## 7     Hyundai     N/A     2.5     3.4            3.2   2.2    3.3       3.3
## 8      Jaguar     N/A     4.6     2.4            1.6   5.5    1.3       1.6
## 9        Lada  Samara     3.2     3.9            4.3   2.0    4.3       4.5
## 10      Mazda     323     2.6     3.3            3.7   2.8    3.7       3.0
## 11   Mercedes     200     4.1     1.7            1.8   4.6    2.4       3.2
## 12 Mitsubishi  Galant     3.2     2.9            3.2   3.5    3.1       3.1
## 13     Nissan   Sunny     2.6     3.3            3.9   2.1    3.5       3.9
## 14       Opel   Corsa     2.2     2.4            3.0   2.6    3.2       4.0
## 15       Opel  Vectra     3.1     2.6            2.3   3.6    2.8       2.9
## 16    Peugeot     306     2.9     3.5            3.6   2.8    3.2       3.8
## 17    Renault      19     2.7     3.3            3.4   3.0    3.1       3.4
## 18      Rover     N/A     3.9     2.8            2.6   4.0    2.6       3.0
## 19     Toyota Corolla     2.5     2.9            3.4   3.0    3.2       3.1
## 20      Volvo     N/A     3.8     2.3            1.9   4.2    3.1       3.6
## 21    Trabant     601     3.6     4.7            5.5   1.5    4.1       5.8
## 22         VW    Golf     2.4     2.1            2.0   2.6    3.2       3.1
## 23         VW  Passat     3.1     2.2            2.1   3.2    3.5       3.5
## 24   Wartburg     1.3     3.7     4.7            5.5   1.7    4.8       5.2
##    Safety EasyHandling
## 1     2.4          2.8
## 2     1.6          2.8
## 3     4.0          2.6
## 4     3.3          4.3
## 5     4.4          2.2
## 6     3.6          2.8
## 7     3.3          2.4
## 8     2.8          3.6
```

```
## 9      4.7          2.9
## 10     3.7          3.1
## 11     1.4          2.4
## 12     2.9          2.6
## 13     3.8          2.4
## 14     2.9          2.4
## 15     2.4          2.4
## 16     3.2          2.6
## 17     3.0          2.7
## 18     3.2          3.0
## 19     3.2          2.8
## 20     1.6          2.4
## 21     5.9          3.1
## 22     3.1          1.6
## 23     2.8          1.8
## 24     5.5          4.0
```

```
rating <- cars[, c("Economy", "Service", "Long.run.Value", "Price", "Design", "SportyCar", "Safety", "Ea
rating <- as.matrix(rating)
rownames(rating) <- cars$Type
ca <- ca(rating)
plot(ca)
```



Exercise 2

I got the Chi Square test statistics 4354.548, df 147, p-value approximately equal to 0. Since the p-value is extremely small(less than 2.2e-16), therefore R made it approximately 0. I also double checked my anaswer with the R-Based function chisq.test() to make sure I structured my own function correctly.

```
bachelors <- read.table("C:/Users/tonyg/Desktop/Academic/Grad/HUDM 6122/bachelors.txt", header = TRUE)

chi_sq <- function(x) {
  row_totals <- apply(x, 1, sum)
  col_totals <- apply(x, 2, sum)
  grand_total <- sum(row_totals)

  expected <- outer(row_totals, col_totals) / grand_total
  chi_sq <- sum((x - expected)^2 / expected)

  df <- (nrow(x) - 1) * (ncol(x) - 1)

  return(list(chi_sq = chi_sq, df = df, p_value = 1-pchisq(chi_sq, df)))
}

chi_sq(bachelors[,c(-1,-2,-11)])
```

```
## $chi_sq
## [1] 4354.548
##
## $df
## [1] 147
##
## $p_value
## [1] 0
```

```
chisq.test(bachelors[,c(-1,-2,-11)])
```

```
## Warning in chisq.test(bachelors[, c(-1, -2, -11)]): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  bachelors[, c(-1, -2, -11)]
## X-squared = 4354.5, df = 147, p-value < 2.2e-16
```

Exercise 3

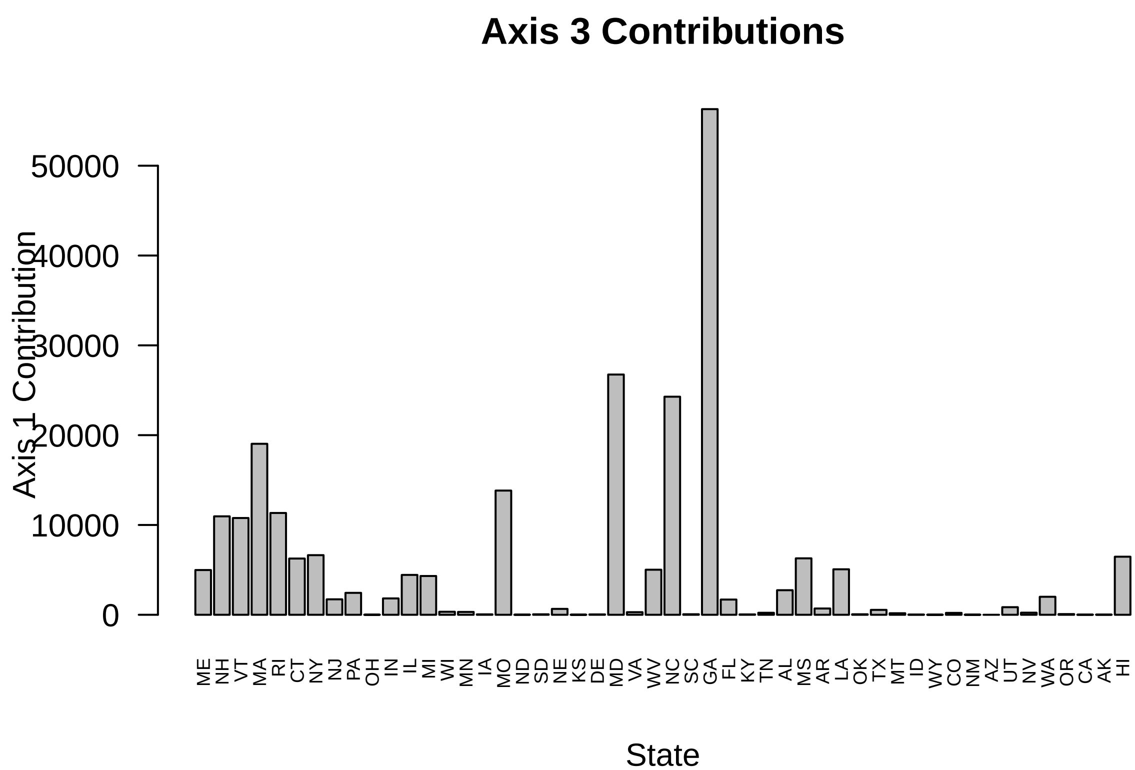
In the third Axis, state GA has the highest absolute contribution. As it has the highest absolute contribution, it means crime pattern in state GA has a strong influence on the structure of the third axis. From the second barchart that colored, we can observe the difference in crime pattern in different region. By comparing the highest absolute contribution, we can identify states that are particularly distinctive in terms of crime patterns along each axis

```r
crime <- read.table("C:/Users/tonyg/Desktop/Academic/Grad/HUDM 6122/UScrime.txt", header = TRUE)
ca <- ca(crime[,4:10])
row_sums <- rowSums(crime[,4:10])
row_masses <- row_sums / sum(row_sums)
squared_coordinates <- ca$rowcoord[, 1:3]^2
ac <- t(t(squared_coordinates) /row_masses) / ca$sv[1:3]^2
absolute_contributions <- data.frame(State = crime$state,
                                     Region = crime$region,
                                     Axis1 = ac[, 1],
                                     Axis2 = ac[, 2],
                                     Axis3 = ac[, 3])
absolute_contributions
```

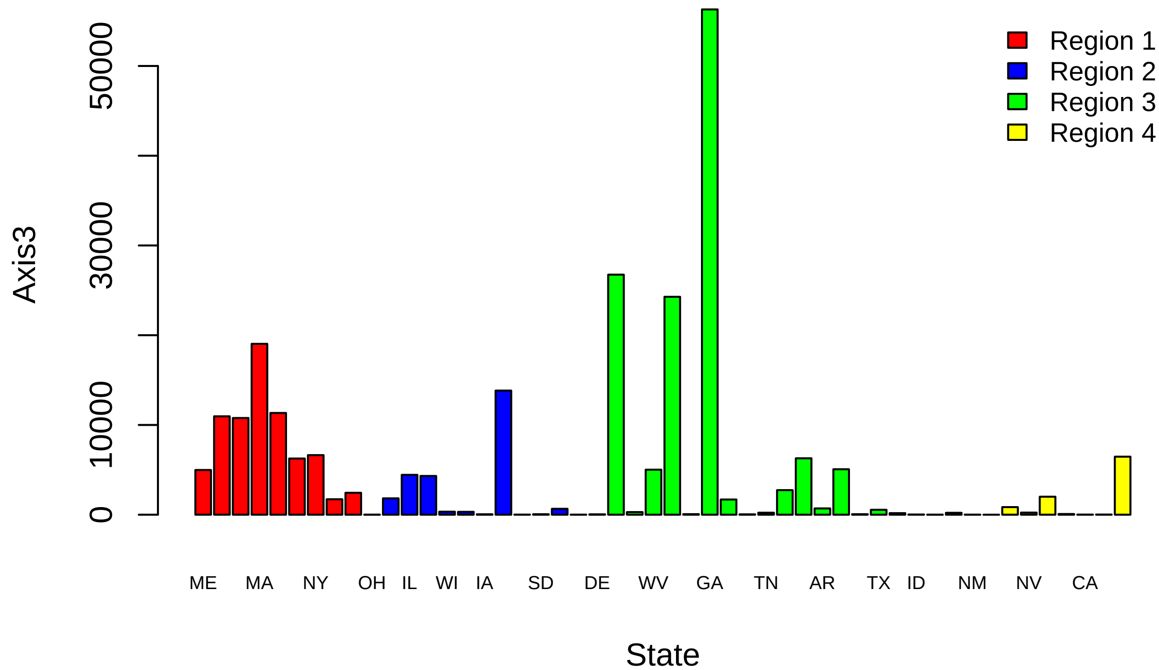
```
##    State Region        Axis1        Axis2        Axis3
## 1     ME      1 7.231765e+02 9.287005e+03  4982.239170
## 2     NH      1 8.933114e+01 9.035354e+02 10961.158569
## 3     VT      1 3.578251e+01 1.868537e+04 10775.279528
## 4     MA      1 4.664414e+03 8.854310e+03 19037.581340
## 5     RI      1 1.146131e+03 5.352771e+03 11337.402698
## 6     CT      1 1.407336e+03 3.015965e-01  6261.895198
## 7     NY      1 1.013825e+04 4.720685e+03  6638.048340
## 8     NJ      1 3.798534e+03 7.446203e+02  1722.753519
## 9     PA      1 4.106360e+05 3.673482e+03  2441.892242
## 10    OH      2 3.032031e+02 2.961278e+03     5.325364
## 11    IN      2 3.634117e+02 7.736323e+02  1823.187919
## 12    IL      2 3.410284e+03 9.205448e+03  4440.031861
## 13    MI      2 3.413806e+02 1.088768e+00  4318.286539
## 14    WI      2 9.150925e+03 5.909839e+02   334.547586
## 15    MN      2 3.134987e+02 1.426822e+03   317.018115
## 16    IA      2 2.555015e+03 9.146651e+02    48.246230
## 17    MO      2 8.155313e+01 3.110538e+02 13831.467398
## 18    ND      2 4.356294e+04 4.408236e+03    11.597208
## 19    SD      2 2.958204e+03 7.364700e+02    52.484513
## 20    NE      2 9.535118e+02 2.920992e+03   655.291290
## 21    KS      2 3.627435e+03 1.898545e+01     8.552350
## 22    DE      3 7.071099e+01 5.726993e+02    41.800261
## 23    MD      3 2.488450e+03 2.946978e+03 26749.796533
## 24    VA      3 8.602186e+02 5.005753e-01   289.724975
## 25    WV      3 8.972166e+01 7.685338e+03  5016.787430
## 26    NC      3 3.272680e+02 9.633671e+03 24280.812899
## 27    SC      3 4.042125e+02 1.800643e+04    60.174761
## 28    GA      3 3.994686e+03 1.095323e+05 56301.601068
## 29    FL      3 3.729267e+01 1.537776e+03  1695.986567
## 30    KY      3 2.913831e+03 3.158382e+01    41.336230
## 31    TN      3 1.549295e+03 6.245470e+03   221.365973
## 32    AL      3 1.460298e+03 3.673513e+03  2737.599894
## 33    MS      3 6.973748e+00 1.102168e+04  6286.351146
## 34    AR      3 1.378152e+01 5.956431e+04   700.818867
## 35    LA      3 4.497747e+03 2.314715e+02  5065.134393
## 36    OK      3 1.157396e+02 1.005841e+03    57.889961
## 37    TX      3 1.863254e+02 9.357597e+02   546.772247
## 38    MT      4 6.891616e+03 1.320398e+02   173.272739
## 39    ID      4 1.425874e+04 1.551300e+02    24.775622
```

```
## 40     WY      4 7.862618e+03 1.508415e+01     6.744867
## 41     CO      4 2.841968e+01 1.718374e+01   210.974295
## 42     NM      4 3.858861e+02 3.163676e+03     7.833747
## 43     AZ      4 5.369683e+02 1.749862e+02     1.145492
## 44     UT      4 9.574373e+03 4.650076e+02   846.674038
## 45     NV      4 1.069363e+03 1.527923e+00   235.050184
## 46     WA      4 4.857503e+02 1.231827e+03  2002.868697
## 47     OR      4 1.746121e+03 2.801611e+02    76.955452
## 48     CA      4 1.204888e+02 3.363801e-02    15.404534
## 49     AK      4 3.040584e+00 2.647302e+03    19.031853
## 50     HI      4 5.156281e+02 1.153835e+02  6465.423753
```

```
barplot(absolute_contributions$Axis3 , names.arg = absolute_contributions$State, xlab = "State", ylab =
        main = "Axis 3 Contributions", las = 2, cex.names = 0.6)
```



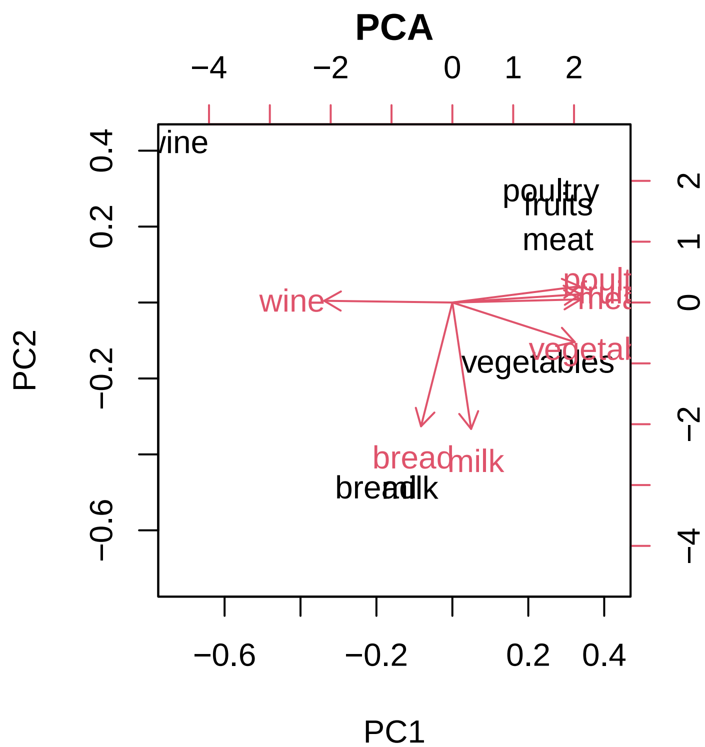

```
colors <- c("red", "blue", "green", "yellow")
colors1 <- colors[absolute_contributions$Region]
barplot(height = absolute_contributions$Axis3, names.arg = absolute_contributions$State, col = colors1,
legend("topright", legend = paste("Region", 1:4), fill = colors, cex = 0.8, bty = "n")
```



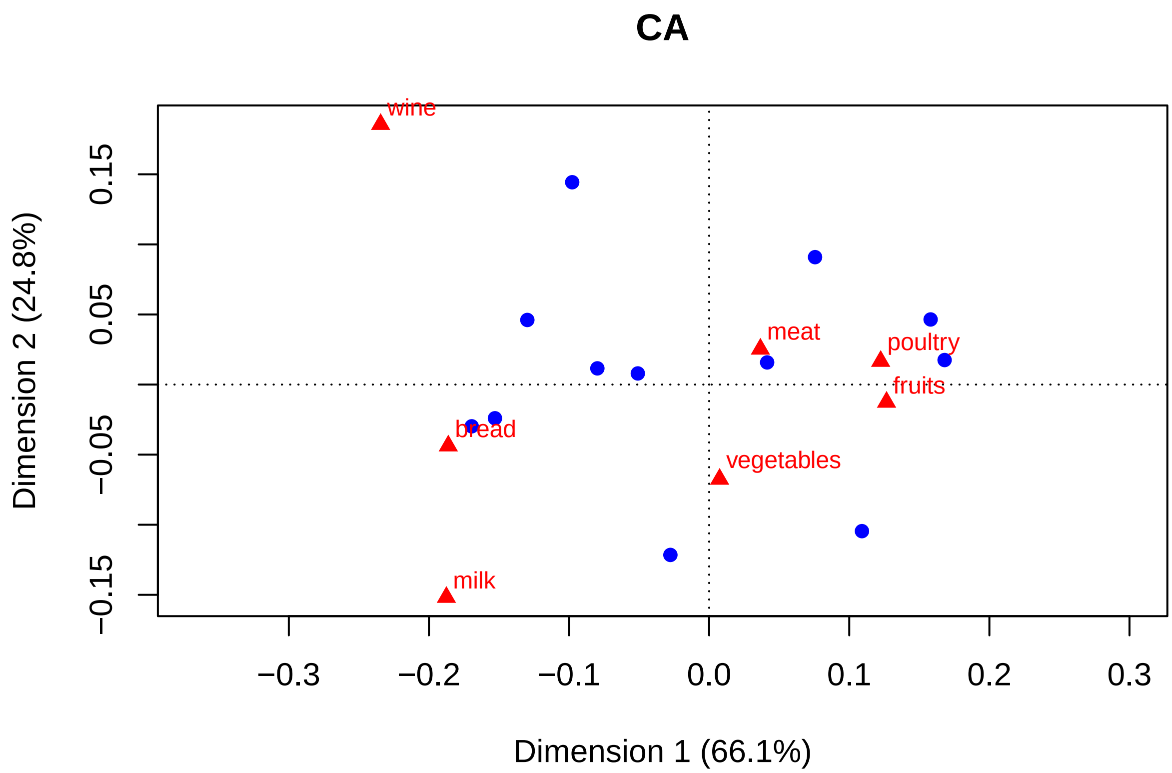
Exercise 4 Since we can treat the columns as categorical variables, we can consider this table as a contingency table. From the two graph, we can see that PCA and CA shows a different pattern but indicate similar relationships between variables and observations. For example, both PCA and CA, we can observe that wine and milk locate on the opposite sides of the origin which indicate they are probably negative associated. In CA, milk & bread and poultry are negative associated, but in PCA , PC1 didn't show the same relationship.

```
food <- read.table("C:/Users/tonyg/Desktop/Academic/Grad/HUDM 6122/food.txt", header = TRUE)

pca <- prcomp(cor(food[, -c(1,2)]), scale = TRUE)
ca <- ca(food[, -c(1,2)])
biplot(pca, main = "PCA")
```




```r
plot(ca, main = "CA")
```


CA
wine
meat
poultry
fruits
bread
vegetables
milk
Dimension 2 (24.8%)
Dimension 1 (66.1%)
0.15
0.05
−0.05
−0.15
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3