BDM600 - Lab 8 - Group 8

Ran Arino; Zubeka Dane Dang; Solmaz Heidar Nassab

2024-04-03

Credentials

All members participated in this lab assignment. The file is the merged version.

Start

```
library(vcd)
```

```
## Loading required package: grid
```

```
library(vcdExtra)
```

```
## Loading required package: gnm
```

```
library(logmult)
```

```
## Warning: package 'logmult' was built under R version 4.3.3
```

```
##
## Attaching package: 'logmult'
```

```
## The following object is masked from 'package:gnm':
##
##     se
```

```
## The following object is masked from 'package:vcd':
##
##     assoc
```

```
library(MASS)
library(ca)
```

```
## Warning: package 'ca' was built under R version 4.3.3
```

6.1

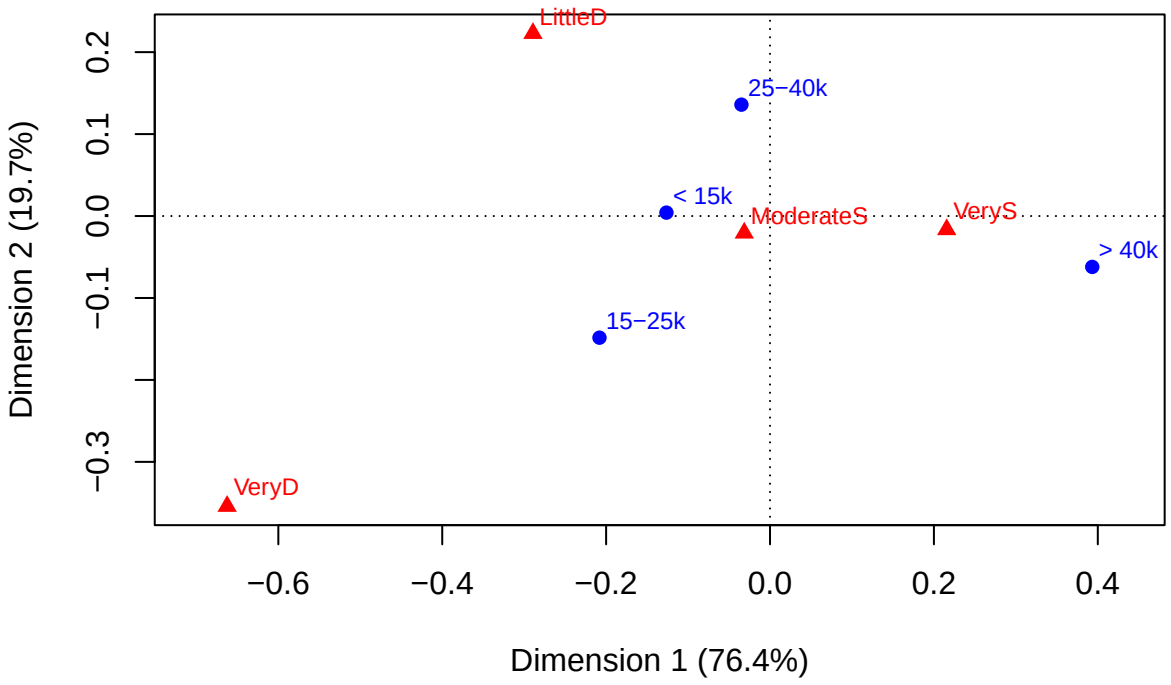
```r
# load data
data(JobSat)
# simple correspondence analysis
JobSat.ca <- ca(JobSat)
summary(JobSat.ca, rows = FALSE, columns = FALSE)
```

(a)

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.047496  76.4  76.4  ******************
##  2      0.012248  19.7  96.1  *****
##  3      0.002397   3.9 100.0  *
##         -------- -----
##  Total: 0.062141 100.0
```

- The inertia accounted for by the one-dimensional solution is 76.4%, which indicates that the first dimension captures a significant portion of the variability in the data.
- The inertia in two-dimensional solution shows an additional 19.7%, which means the the culumative percentage captures 96.1% of variability in the data.

```r
# 2D CA solutiton
plot(JobSat.ca)
```



(b)

- Point "> 40k" shows the longest vector and being close to the x-axis, which shows the most significant impact on the data distribution, its variance captured by the first dimension. - Points "<15K", "15-25K", and "ModerateS" have the same angle of vectors, which means that they have similar pattern of association with other categories, although their impact on variance varies. - Points "VeryD" and "LittleD" close to the y-axis, which shows they are captured by the second dimension. - Points "<15k" and "ModerateS" are close to each other, which means that the nature of their data are similar. - Overall, the first dimension can handle the satisfaction levels (worse in left or better in right side) and the income level (lower&medium in left side vs higher incomes in right side)

6.2

```
data(criminal)
criminal
```

```
##       Age
## Year    15  16  17  18  19
##   1955 141 285 320 441 427
##   1956 144 292 342 441 396
##   1957 196 380 424 462 427
##   1958 212 424 399 442 430
```

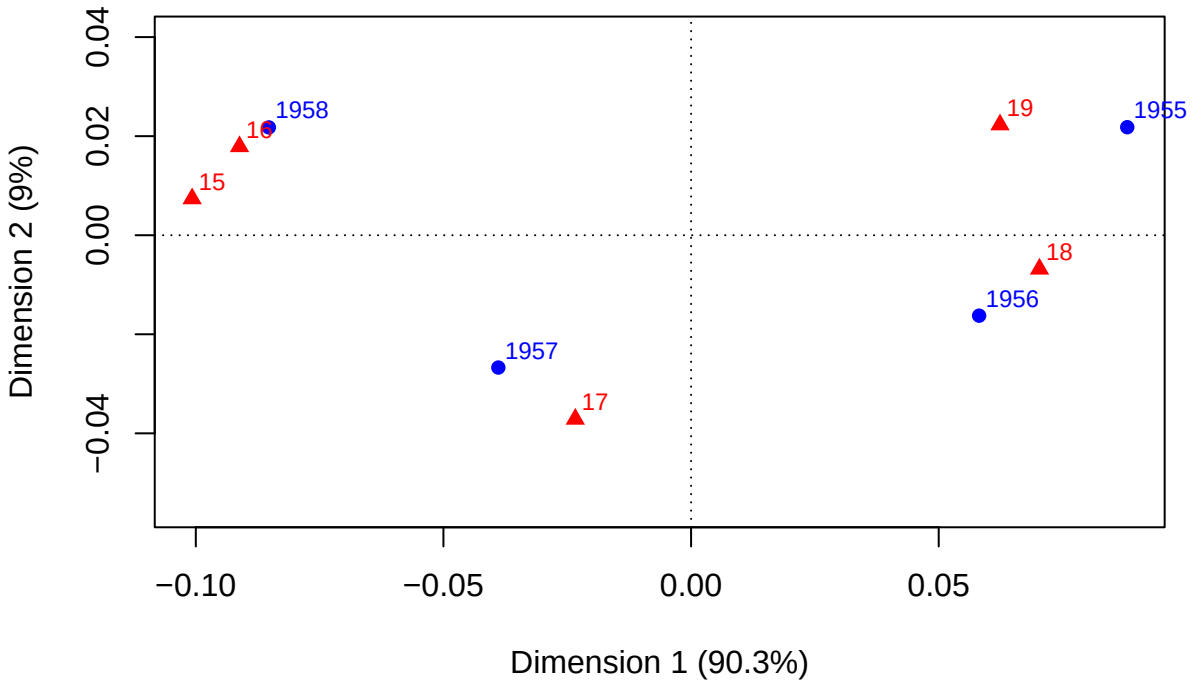
```r
# display ca summary
criminal.ca <- ca(criminal)
summary(criminal.ca, rows = FALSE, columns = FALSE)
```

(a)

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.004939  90.3  90.3  ***********************
##  2      0.000491   9.0  99.3  **
##  3      3.8e-050   0.7 100.0
##         -------- -----
##  Total: 0.005468 100.0
```

- The result shows that the accounted for by the one-dimensional solution is more than 90%.
- Hence, almost all portions of the variability are captured in a single dimension.

```r
# plot ca
plot(criminal.ca)
```



(b)

- Similarity is identified among points “15”, “16”, and “1958”; they have the same direction and length of

vector. - The category points for both year and age vary systematically over Dimension 1. - There were more younger men in later years, and more older in earlier years.

6.3

```
# load data
data(AirCrash)
# form a table
AirCrash.tab <- xtabs(~ Phase + Cause,  data = AirCrash)
AirCrash.tab
```

```
##           Cause
## Phase      criminal human error mechanical unknown weather
##   en route       16          63         29      25      24
##   landing         4         114         19      18      55
##   standing        2           0          2       0       0
##   take-off        1          29         24       8       3
##   unknown         0           1          0       1       1
```

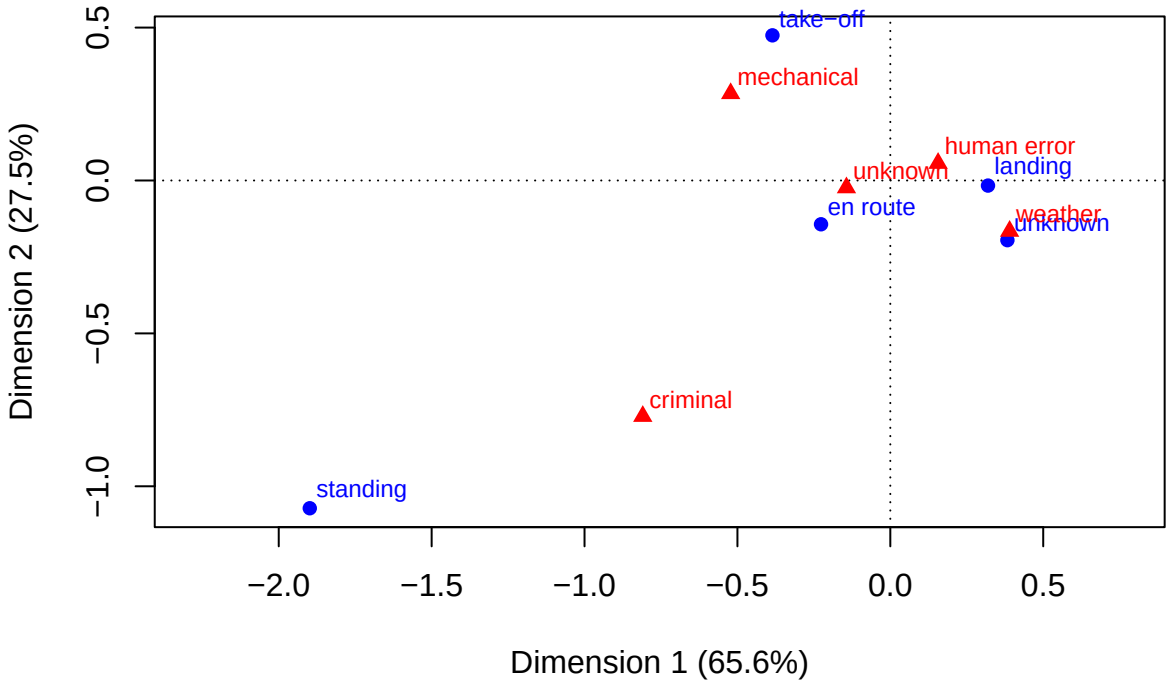
```
# show ca summary
AirCrach.ca <- ca(AirCrash.tab)
summary(AirCrach.ca, rows = FALSE, columns = FALSE)
```

(a)

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.123002  65.6  65.6  ****************
##  2      0.051548  27.5  93.1  *******
##  3      0.012340   6.6  99.7  **
##  4      0.000562   0.3 100.0
##         -------- -----
##  Total: 0.187452 100.0
```

- The 65.6% of variance is accounted for the first dimension.
- The 27.5% of variance is accounted for the second dimension; the 93.1% of data can be explained in tatal of the first and second dimensions.

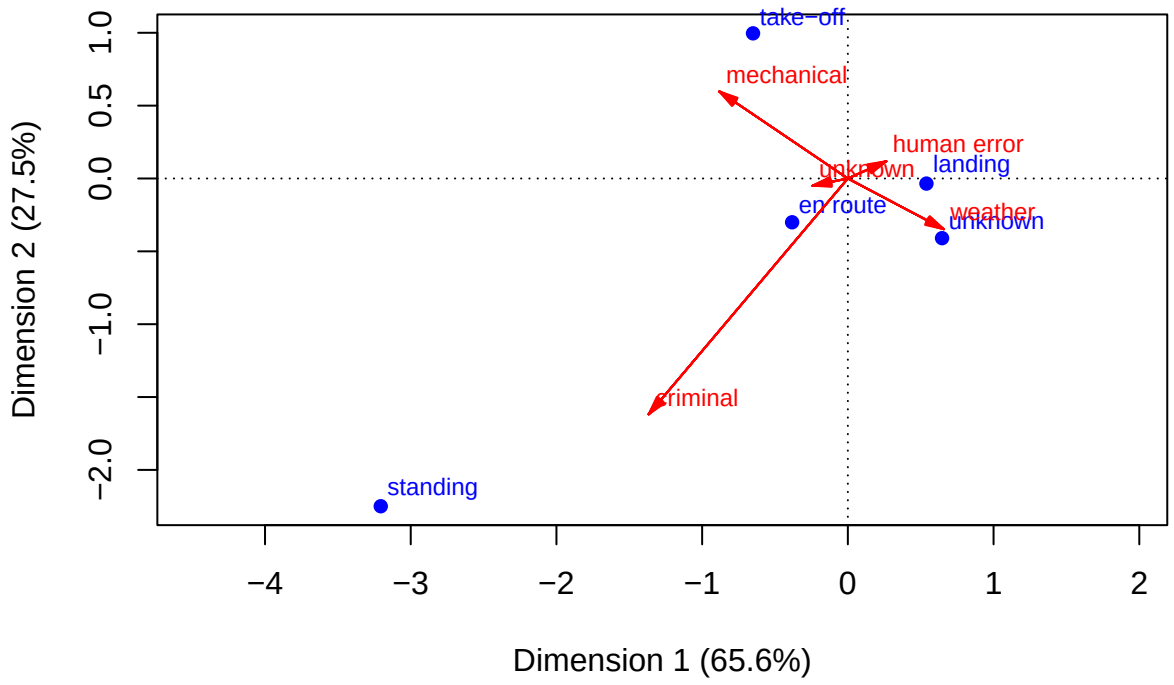
```
# plot ca
plot(AirCrach.ca)
```



(b)

- Overall, "mechanical" and "human-related or weather" causes are more associated with specific flight phases (take-off and landing), respectively. - Both points of "standing" and "criminal" are relatively rare cases, compared to othe categories. - Dimension 1 primarily differentiates incidents occurring during standing and landing phases, with a focus on human error and weather problems. - Dimension 2 primarily differentiates incidents occurring during standing and take-off phases, with a greater emphasis on mechanical problems during take-off.

```
# plot ca; map 'symbiplot'
plot(AirCrach.ca, map = "symbiplot", arrows = c(FALSE, TRUE))
```



(c)

6.4

```
# load data
data(caith)
caith
```

```
##        fair red medium dark black
## blue    326  38    241  110     3
## light   688 116    584  188     4
## medium  343  84    909  412    26
## dark     98  48    403  681    85
```

```
# show ca summary
caith.ca <- ca(caith)
summary(caith.ca, rows=FALSE, columns=FALSE)
```

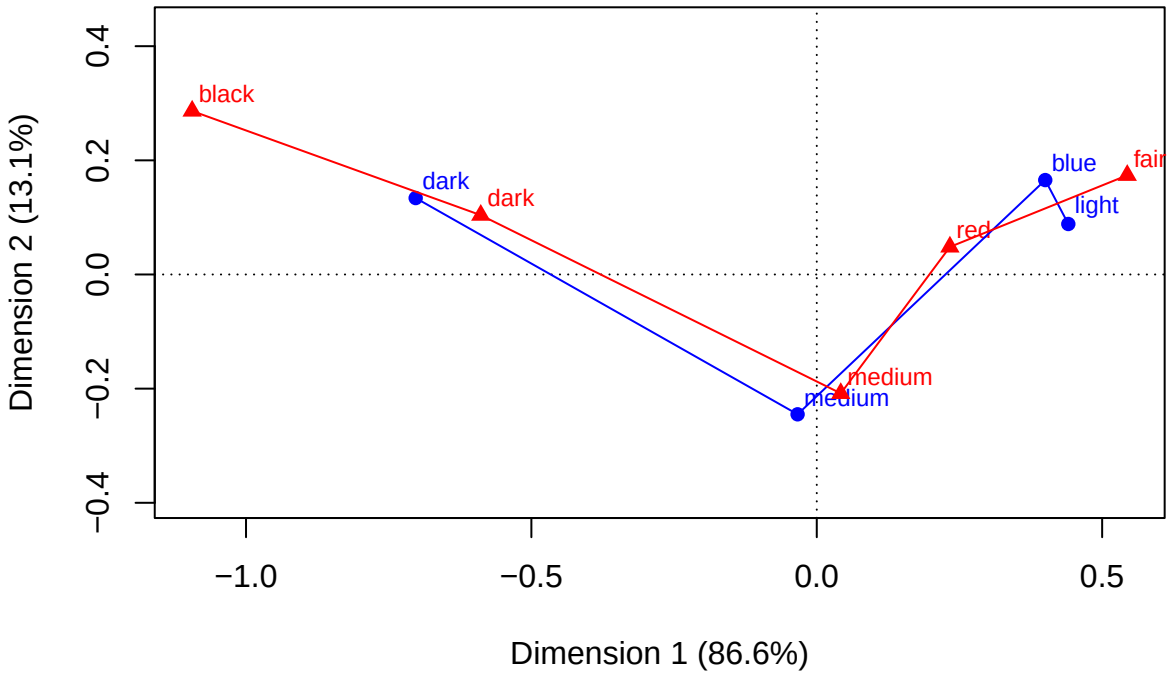
(a)

```
##
## Principal inertias (eigenvalues):
```

```
##
##  dim    value      %   cum%   scree plot
##  1      0.199245  86.6  86.6  **********************
##  2      0.030087  13.1  99.6  ***
##  3      0.000859   0.4 100.0
##         -------- -----
##  Total: 0.230191 100.0
```

- The analysis shows that associations can be explained by a one-dimensional answer, which is accounted by 86.6%. Adding a second dimension solution essentially complete, which achieved 99%.
- Additional third dimension does not add much to the explanation of how the variables are connected.

```
# display ca plot
plot(caith.ca, lines = TRUE)
```



(b)

- We can observe that while Dimension 1 represents the primary light-dark dimension for hair and eye color, accounting for nearly 86.6% of the association between hair and eye color corresponds to dark (left) vs. light (right) on both variables. - Dimension 2 could be called “extremes vs. middle”, but in CA results with largely 1D association.