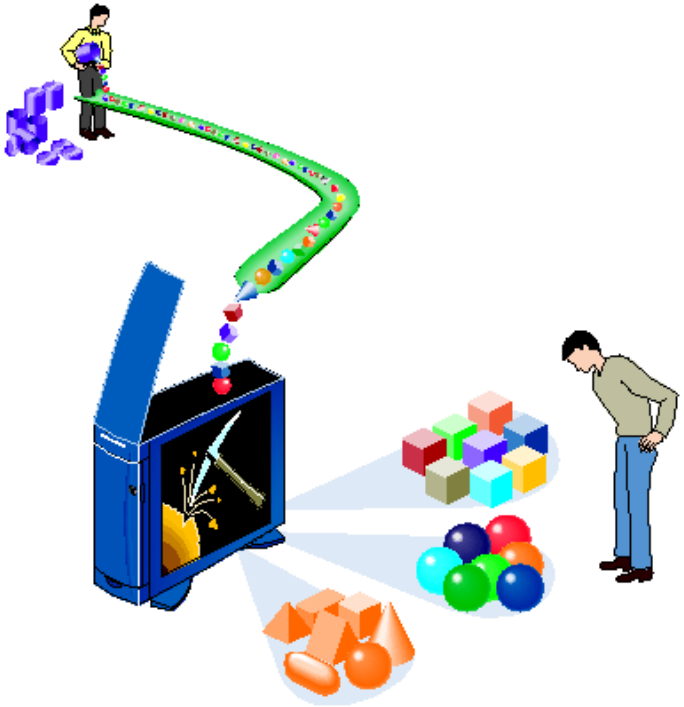
ETM 58D Business Analytics



Instructor: Mustafa Gökçe Baydoğan
Office: M4082

mustafa.baydogan@boun.edu.tr
www.mustafabaydogan.com
blog.mustafabaydogan.com

Survey data

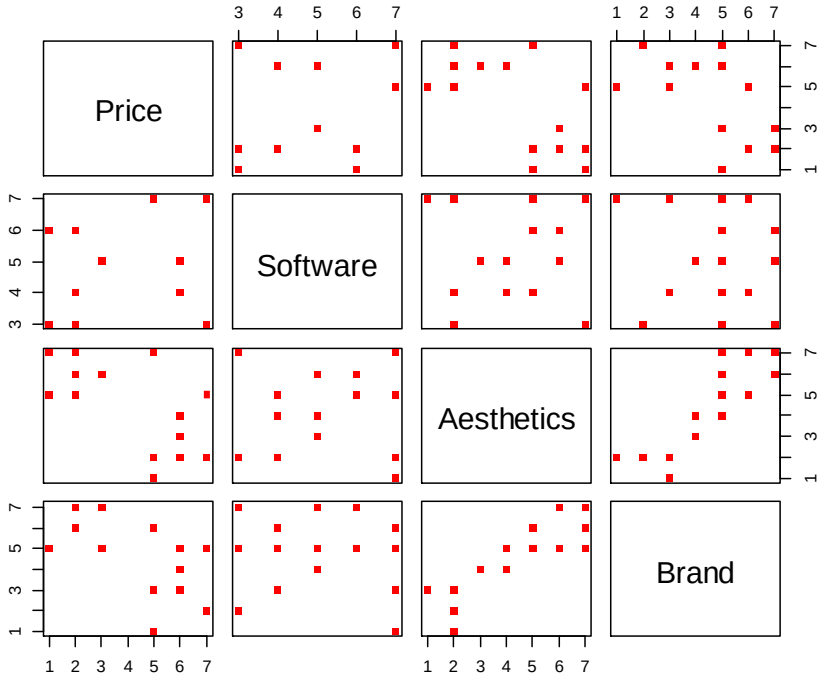
- Simple multivariate analysis of the data with a focus on principal components analysis (PCA)
- Suppose we have survey and asked the participants four 7-scale Likert questions about what they care about when choosing a new computer
  - Price: A new computer is cheap to you (1: strongly disagree – 7: strongly agree)
  - Software: The OS on a new computer allows you to use software you want to use (1: strongly disagree – 7: strongly agree)
  - Aesthetics: The appearance of a new computer is appealing to you (1: strongly disagree – 7: strongly agree)
  - Brand: The brand of the OS on a new computer is appealing to you (1: strongly disagree – 7: strongly agree)

Survey data

Participant	Price	Software	Aesthetics	Brand
P1	6	5	3	4
P2	7	3	2	2
P3	6	4	4	5
P4	5	7	1	3
P5	7	7	5	5
P6	6	4	2	3
P7	5	7	2	1
P8	6	5	4	4
P9	3	5	6	7
P10	1	3	7	5
P11	2	6	6	7
P12	5	7	7	6
P13	2	4	5	6
P14	3	5	6	5
P15	1	6	5	5
P16	2	3	7	7

Survey data (Analysis in R)

```
Price <- c(6,7,6,5,7,6,5,6,3,1,2,5,2,3,1,2)
Software <- c(5,3,4,7,7,4,7,5,5,3,6,7,4,5,6,3)
Aesthetics <- c(3,2,4,1,5,2,2,4,6,7,6,7,5,6,5,7)
Brand <- c(4,2,5,3,5,3,1,4,7,5,7,6,6,5,5,7)
data <- data.frame(Price, Software, Aesthetics, Brand)
plot(data,col=2,pch=".",cex=7)
```



Provides information about the correlation between ratings but no information about the subjects

Survey data (Analysis in R)

- Check correlations and summary statistics

```
summary(data)
cor(data)
```

```
> summary(data)
     Price          Software       Aesthetics       Brand
 Min.   :1.000   Min.   :3.000   Min.   :1.00   Min.   :1.000
 1st Qu.:2.000   1st Qu.:4.000   1st Qu.:2.75   1st Qu.:3.750
 Median :5.000   Median :5.000   Median :5.00   Median :5.000
 Mean   :4.188   Mean   :5.062   Mean   :4.50   Mean   :4.688
 3rd Qu.:6.000   3rd Qu.:6.250   3rd Qu.:6.00   3rd Qu.:6.000
 Max.   :7.000   Max.   :7.000   Max.   :7.00   Max.   :7.000
```

```
> cor(data)
                Price   Software Aesthetics      Brand
Price       1.0000000  0.1856123 -0.6320222 -0.5802668
Software    0.1856123  1.0000000 -0.1462152 -0.1185864
Aesthetics -0.6320222 -0.1462152  1.0000000  0.8528544
Brand      -0.5802668 -0.1185864  0.8528544  1.0000000
```

Survey data (Analysis in R)

- Apply PCA for visualization purposes

```
pca <- princomp(data, cor=T)
summary(pca, loadings=T)
```

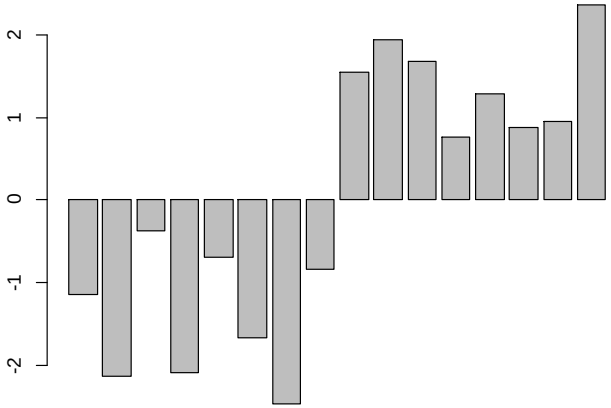
```
> summary(pca, loadings=T)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     1.5589391 0.9804092 0.6816673 0.37925777
Proportion of Variance 0.6075727 0.2403006 0.1161676 0.03595911
Cumulative Proportion  0.6075727 0.8478733 0.9640409 1.00000000

Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4
Price      -0.523         0.848
Software   -0.177  0.977 -0.120
Aesthetics  0.597  0.134  0.295 -0.734
Brand       0.583  0.167  0.423  0.674
```

*Comp.1 = -0.523 * Price - 0.177 * Software + 0.597 * Aesthetics + 0.583 * Brand*

Survey data (Analysis in R)

- What is hidden in the components?

```
#first component
barplot(pca$scores[,1])
```



Recall first component!

*Comp.1 = -0.523 * Price - 0.177 * Software + 0.597 * Aesthetics + 0.583 * Brand*

this new variable indicates whether a user cares about Price and Software or Aesthetics and Brand for the computer. These variables are so called latent variables. We can interpret this as “Feature/Fashion index” or something. However there is no definite answer for this part of PCA. It all depends on the data.

Survey data (Analysis in R)

- Suppose we also obtain the information about the operating system being used from the participant.

Participant	Price	Software	Aesthetics	Brand	OS
P1	6	5	3	4	0
P2	7	3	2	2	0
P3	6	4	4	5	0
P4	5	7	1	3	0
P5	7	7	5	5	1
P6	6	4	2	3	0
P7	5	7	2	1	0
P8	6	5	4	4	0
P9	3	5	6	7	1
P10	1	3	7	5	1
P11	2	6	6	7	0
P12	5	7	7	6	1
P13	2	4	5	6	1
P14	3	5	6	5	1
P15	1	6	5	5	1
P16	2	3	7	7	1

Survey data (Analysis in R)

- Let's plot two components based on the operating system

```
#OS
OS <- c(0,0,0,0,1,0,0,0,1,1,0,1,1,1,1,1)
plot(pca$scores[,1],pca$scores[,2],col=OS+1,pch=".",cex=7)
```