Credit Project for DATA1

Author: Kıvanç Gördü

Expected time consumption of the project: 5h

Absolute time consumption of the project:

Aim

The aim is to apply the learnt concepts (mainly Exp. Data Analysis and Descp. Data Analysis) of DATA1 course content while working the evaluation data of Indonesia. We are responsible to analyse individual features, data as a whole with sufficient illustrations and descriptions of the process and methodology.

Solution

R script reads the complete data. We examine the column names and choose random two features to calculate their sum, mean, quartiles and median. Then we look at the boxplot. At this point, ee may observe a few points based on this method. Those are:

- The most favourable place is Gili_Meno_Beach because its median is the highest, with 6 points. 50% of people voted between 8 and 4 points.
- On the contorary, the least favourable places are Azul_Beach_Club, Prambanan and GiliTrawangan_FreeDive with approximately 2 points.
- There is an evaluation coherance of people in case of Prambanan since the scale is from ~1.5 to 4.
- The highest varience is in GiliTrawangan_FreeDive followed by Mirror_Bali. Meaning that, there is not a significant agreement among people voted. The least coherency among people is in this case. There are high points but also low.
- On the other hand, there is one person that voted "significantly high" / "outlierly" that we can see as 10 and 9 in Bali_Safari and Prambanan respectively.

We put two features to scatter-plot. However, there is not a high distribution and over-crowding. Thus, we cannot much differ their entities.

Subsequently, we are observing the correlation of every dimension with each other.

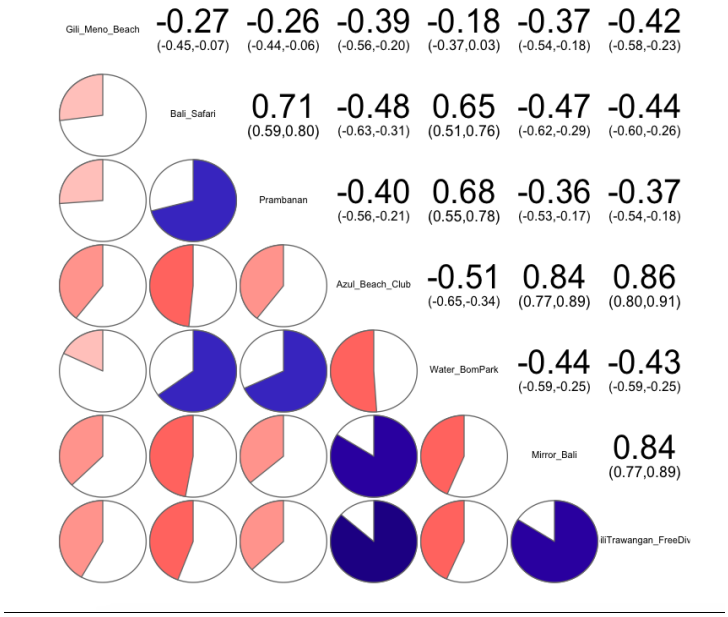


Figure 1.1

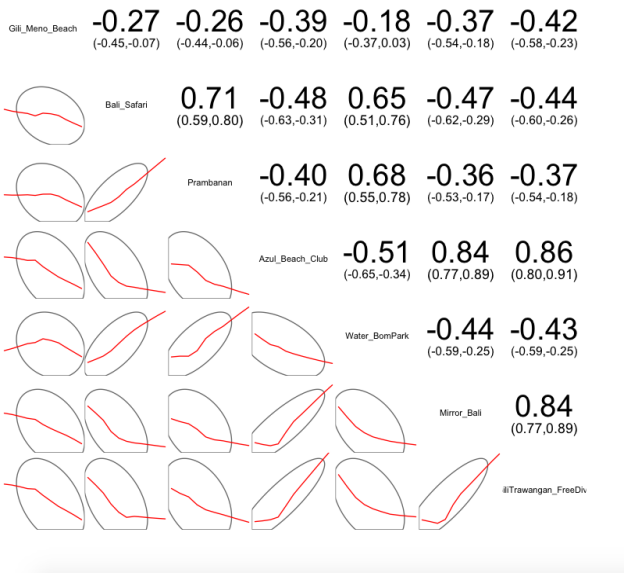


Figure 1.2

*We may notice the high correlations based on the pies. The darker colour / bigger part is occupied the higher correlation is. For ex.: Bali_Safari and Prambanan, Bali_Safari and Water_BomPark, Prambanan and Water_BomPark, Trawangan_FreeDive with Azul_Beach_Club.

**The confidence ellipse is more circular when two variables are uncorrelated. We also have the correlation vector drawn on this illustration.

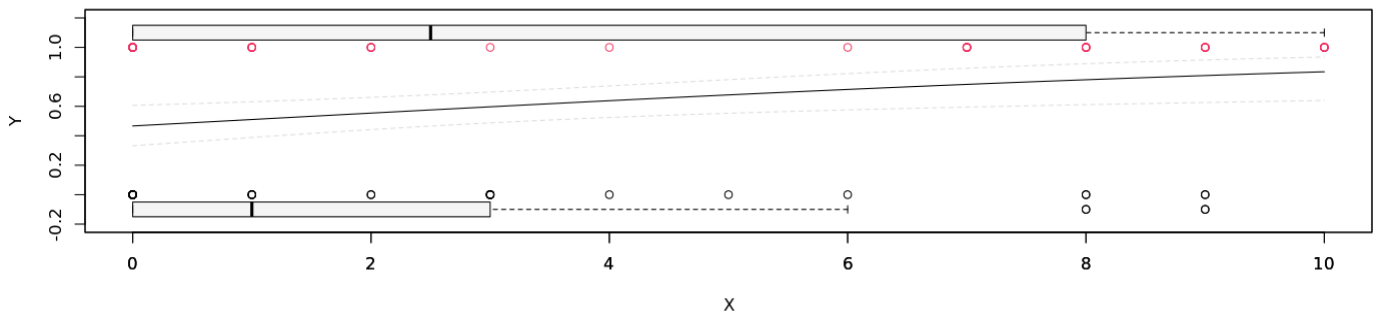
Giving the Logistic Regression, we are in a need of binary data. Therefore, we find the "Gender" column and its comparison with earlier columns to be the most suitable for this method. Here are the results:

```
[✓] Model je s dodaným regresorem významně lepší než bez něj.

Chí-kvadrát = 7.814(1), p-hodnota ≈ 0.005184

Navýšení o 1 jednotku v 'x' změní šanci odpovědi '1' 1.191 krát,
... neboli navýšení o 1 jednotku v 'x' zvýší šanci být odpovědí '1' o 19.1 %.

Logistická regrese dokáže správně přiřadit '0' a '1' v 60 % případů.
```



ExpOddsRatio	Estimate	Std. Error	z value	Pr(>|z|)
0.87	-0.13	0.29	-0.46	0.64
1.19	0.18	0.07	2.63	0.01

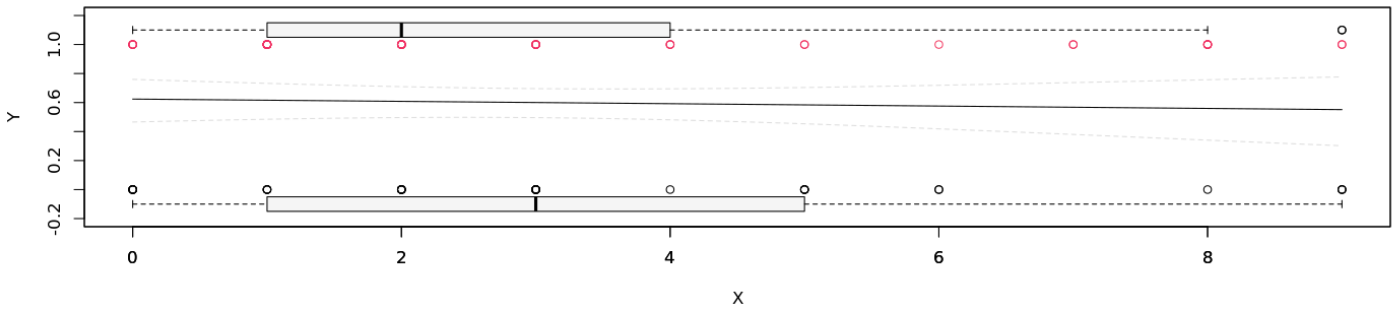
Figure 2.1

```
[!] Model NENÍ s dodaným regresorem významně lepší než bez něj (nemá na určení odpovědí významný vliv).

Chí-kvadrát = 0.168(1), p-hodnota ≈ 0.682135

Navýšení o 1 jednotku v 'x' změní šanci odpovědi '1' 0.967 krát,
... neboli navýšení o 1 jednotku v 'x' sníží šanci být odpovědí '1' o 3.3 %.

Logistická regrese dokáže správně přiřadit '0' a '1' v 60 % případů.
```



ExpOddsRatio	Estimate	Std. Error	z value	Pr(>|z|)
1.66	0.51	0.33	1.54	0.12
0.97	-0.03	0.08	-0.41	0.68

Figure 2.2

* Gender and Prombanan would not have correlation. p-value ≈ 0.682135 so p-value is bigger than 5%.

**On the other hand, Gender and Azul_Beach_Club have correlation. p-value ≈ 0.005184 so p-value is way lower than 5%.

Next, we analys PCA. We do not need to scale the data because all the entities are in the same scale. Latent components are ordered based on the importance/proportion of their variance. If we want a 2D graph, the wisest option would be to choose Comp1 and Comp2 with approximately 85% of Cumulative Proportion. Here we can also realise that the first components have a relation with each of the features, thanks to the code `PCA_results$loadings`. Afterwards, we apply MDS. We will be using euclidean becauase there is almost no outlier and we do not care how different is the distribution of the entities. The result shows us that GOF ≈ 80%. We have almost 80 percent of information used for MDS. In other words, %20 of information loss. We see three main clusters. Left two are closer and overlapping with each other. The right one is more separate. To get a sample for each group, let us take one entity of each cluster from each cluster's centre. These are with the indexes 2, 6 and 75.

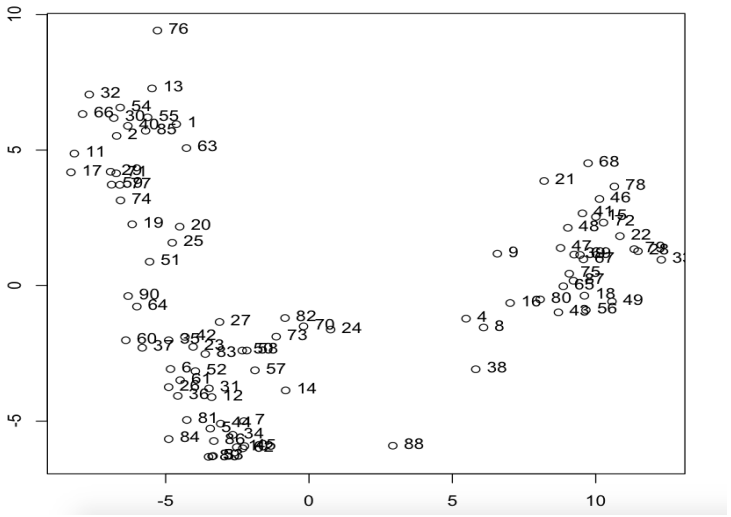


Figure 3

Index 2 has highest points for Azul_Beach_Club and Mirror_Bali. This person may be an individual and likes clubs, night life. Probably a single and enjoying the club life. I would say, this is a preference for a single and young person.

Index 6 has highest point for Gili_Meno_Beach then Water_BomPark. This person would enjoy the place with their partner. Maybe a young couple with middle age.

Index 47 has highest points for Mirror_Bali, Azul_Beach_Club and GiliTrawangan_FreeDive. This would be a senior who likes trying different things and maybe retired and just want tot rest.

Therefore, x might be age of the people in the group and y might be the amount of people in the group.

Conclusions

Demonstrating and applying these methods is utterly benefical for any kind of analysis we make. We analysed the data step by step. Firstly, based on the features, secondly, based on the entities. We realised their statistical indexes. We finished the report with the cluster analysis.