Name:
DNI/Passport:

GRAU D'ENGINYERIA INFORMÀTICA (UPC).
CURS 19-20 Q2 –QUIZ 1
Anàlisi de Dades i Explotació de la Informació (ADEI) .
(Data: 30/03/2019 10:00-11:30 h Room Virtual)

Professor:	Lídia Montero Mercadé
Rules for quiz:	Calculator, statistical tables and customized R script without long comments are allowed. Internet access will be available, emailing and chatting is strictly forbidden. Mobile phones should be switched off.
Duration:	1h 30 min
Marks:	Before 3/4/19 Subject ATENEA WEB site.
Open Office:	Email requests.

Problem 1: All questions account for 1 point

25 personality self-report items taken from the International Personality Item Pool (ipip.ori.org) were included as part of the Synthetic Aperture Personality Assessment (SAPA) web based personality assessment project (SAPA https://sapa-project.org). The data from 2800 subjects are included here. Three additional demographic variables (sex, education, and age) are also included. The first 25 items are organized by five putative factors: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness. The item data were collected using a 6 point response scale: *1 Very Inaccurate 2 Moderately Inaccurate 3 Slightly Inaccurate 4 Slightly Accurate 5 Moderately Accurate 6 Very Accurate*. The items given were sampled from the International Personality Item Pool of Lewis Goldberg using the sampling technique of SAPA. This is a sample data set taken from the much larger SAPA data bank.

Name	Description
A1	I Am indifferent to the feelings of others. (q_146)
A2	Inquire about others' well-being. (q_1162)
A3	Know how to comfort others. (q_1206)
A4	Love children. (q_1364)
A5	Make people feel at ease. (q_1419)
C1	Am exacting in my work. (q_124)
C2	Continue until everything is perfect. (q_530)
C3	Do things according to a plan. (q_619)
C4	Do things in a half-way manner. (q_626)
C5	Waste my time. (q_1949)
E1	Don't talk a lot. (q_712)
E2	Find it difficult to approach others. (q_901)
E3	Know how to captivate people. (q_1205)
E4	Make friends easily. (q_1410)
E5	Take charge. (q_1768)
N1	Get angry easily. (q_952)
N2	Get irritated easily. (q_974)
N3	Have frequent mood swings. (q_1099
N4	Often feel blue. (q_1479)
N5	Panic easily. (q_1505)
O1	Am full of ideas. (q_128)
O2	Avoid difficult reading material.(q_316)
O3	Carry the conversation to a higher level. (q_492)
O4	Spend time reflecting on things. (q_1738)
O5	Will not probe deeply into a subject. (q_1964)
gender	gender Males = 1, Females =2
education	1 = HS, 2 = finished HS, 3 = some college, 4 = college graduate 5 = graduate degree
age	age in years

Source

The items are from the ipip (Goldberg, 1999). The data are from the SAPA project (Revelle, Wilt and Rosenthal, 2010) , collected Spring, 2010 (https://sapa-project.org).

Name:
DNI/Passport:

References

1. Goldberg, L.R. (1999) A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In Mervielde, I. and Deary, I. and De Fruyt, F. and Ostendorf, F. (eds) Personality psychology in Europe. 7. Tilburg University Press. Tilburg, The Netherlands.
2. Revelle, W., Wilt, J., and Rosenthal, A. (2010) Individual Differences in Cognition: New Methods for examining the Personality-Cognition Link In Gruszka, A. and Matthews, G. and Szymura, B. (Eds.) Handbook of Individual Differences in Cognition: Attention, Memory and Executive Control, Springer.
3. Revelle, W, Condon, D.M., Wilt, J., French, J.A., Brown, A., and Elleman, L.G. (2016) Web and phone based data collection using planned missing designs. In Fielding, N.G., Lee, R.M. and Blank, G. (Eds). SAGE Handbook of Online Research Methods (2nd Ed), Sage Publications.


```
> psych::describe(bfi)
          vars    n  mean    sd median trimmed   mad min max range  skew kurtosis   se
A1           1 2784  2.41  1.41      2    2.23  1.48   1   6     5  0.83    -0.31 0.03
A2           2 2773  4.80  1.17      5    4.98  1.48   1   6     5 -1.12     1.05 0.02
A3           3 2774  4.60  1.30      5    4.79  1.48   1   6     5 -1.00     0.44 0.02
A4           4 2781  4.70  1.48      5    4.93  1.48   1   6     5 -1.03     0.04 0.03
A5           5 2784  4.56  1.26      5    4.71  1.48   1   6     5 -0.85     0.16 0.02
C1           6 2779  4.50  1.24      5    4.64  1.48   1   6     5 -0.85     0.30 0.02
C2           7 2776  4.37  1.32      5    4.50  1.48   1   6     5 -0.74    -0.14 0.03
C3           8 2780  4.30  1.29      5    4.42  1.48   1   6     5 -0.69    -0.13 0.02
C4           9 2774  2.55  1.38      2    2.41  1.48   1   6     5  0.60    -0.62 0.03
C5          10 2784  3.30  1.63      3    3.25  1.48   1   6     5  0.07    -1.22 0.03
E1          11 2777  2.97  1.63      3    2.86  1.48   1   6     5  0.37    -1.09 0.03
E2          12 2784  3.14  1.61      3    3.06  1.48   1   6     5  0.22    -1.15 0.03
E3          13 2775  4.00  1.35      4    4.07  1.48   1   6     5 -0.47    -0.47 0.03
E4          14 2791  4.42  1.46      5    4.59  1.48   1   6     5 -0.82    -0.30 0.03
E5          15 2779  4.42  1.33      5    4.56  1.48   1   6     5 -0.78    -0.09 0.03
N1          16 2778  2.93  1.57      3    2.82  1.48   1   6     5  0.37    -1.01 0.03
N2          17 2779  3.51  1.53      4    3.51  1.48   1   6     5 -0.08    -1.05 0.03
N3          18 2789  3.22  1.60      3    3.16  1.48   1   6     5  0.15    -1.18 0.03
N4          19 2764  3.19  1.57      3    3.12  1.48   1   6     5  0.20    -1.09 0.03
N5          20 2771  2.97  1.62      3    2.85  1.48   1   6     5  0.37    -1.06 0.03
O1          21 2778  4.82  1.13      5    4.96  1.48   1   6     5 -0.90     0.43 0.02
O2          22 2800  2.71  1.57      2    2.56  1.48   1   6     5  0.59    -0.81 0.03
O3          23 2772  4.44  1.22      5    4.56  1.48   1   6     5 -0.77     0.30 0.02
O4          24 2786  4.89  1.22      5    5.10  1.48   1   6     5 -1.22     1.08 0.02
O5          25 2780  2.49  1.33      2    2.34  1.48   1   6     5  0.74    -0.24 0.03
gender      26 2800  1.67  0.47      2    1.71  0.00   1   2     1 -0.73    -1.47 0.01
education   27 2577  3.19  1.11      3    3.22  1.48   1   5     4 -0.05    -0.32 0.02
age         28 2800 28.78 11.13     26   27.43 10.38   3  86    83  1.02     0.56 0.21
```

1. Define a binary factor for gender **f.gender** and a polytomic factor for education **f.educ**. Justify with R commands for the procedure and your answer. Calculate **thresholds to identify severe outliers** for the age variable **(age)**.
2. Conduct a suitable data imputation procedure to remove missing data included in dataset for numeric variables. Check imputation consistency for numeric variables.
3. Conduct a suitable data imputation procedure for factors. Summarize imputation results for f.education factor.
4. Can the average of age can be argued to be the same for all education levels (**f.educ)** and gender (**f.gender)?** Which are the groups that show significant greater values than the others? Use graphic, numeric and inferential tools.
5. Let us assume that education **(f.educ)** is the **target** variable. Use a suitable feature selection and profiling tool to discuss global association between target and numerical variables/factors in dataset.
6. Profile **HS education group** according to available data in your dataset.
7. **A Normalized Principal Component Analysis is addressed using as supplementary variables gender, education and age.** How many axes do you have to retain according to Kaiser criteria? What's the inertia explained by retained Kaiser-based principal components?
8. Try to explain the meaning of the axes in the first factorial plane. Which 3 variables have the greatest correlation with each factor in the first factorial plane?
9. **A Non-normalized Principal Component Analysis is addressed using as supplementary variables gender, education and age.** How many axes do you have to retain according to Kaiser criteria? What's the inertia explained by retained Kaiser-based principal components?
10. **A Hierarchical Clustering is addressed**. A non-default criteria for selecting the number of clusters to 3 has to set. Explain the characteristics of cluster number 1.