Multivariate Statistical Techniques
CLPS 2908, Spring 2019
Prof. Bertram Malle

Homework 1:

Exploratory Data Analysis (EDA)

Before starting, be sure you have read the *Homework Guidelines 2019*

The data for this homework are in raw ASCII text format, downloadable from the Homework 1 page on canvas. The data come from a real study, but minor alterations will inspire your statistical intuitions.

In the study, people described themselves in three different ways: (a) they rated themselves on unipolar 9-point Likert scales, (b) three weeks later they freely generated self-descriptive attributes, and (c) also three weeks later, they checked attributes as self-descriptive on an adjective check list. There are 159 cases, with 1 record (= data row) per case. The variables, and their column positions in the data file, are described below.

Column Name Meaning of the variable

```
1-3    SUBNR    Subject identification number
5-6    ADJEXT   Number of extraverted attributes checked (range 0 to 4)
8-9    ADJINT   Number of introverted attributes checked (0 to 4)
11-12  IAMEXT   Number of extraverted attributes generated (0 to 5)
14-15  IAMINT   Number of introverted attributes generated (0 to 5)
18-19  EXTRA    Extraverted unipolar scale (1 to 9)
22-23  RESER    Reserved unipolar scale (1 to 9)
26-27  LIVEL    Lively unipolar scale (1 to 9)
30-31  SHY      Shy unipolar scale (1 to 9)
34-35  TALKA    Talkative unipolar scale (1 to 9)
38-39  INTRO    Introverted unipolar scale (1 to 9)
42-43  OUTGO    Outgoing unipolar scale (1 to 9)
46-47  QUIET    Quiet unipolar scale (1 to 9)
```

Below are the specific things you should do to complete this homework. To develop the appropriate syntax, look at the lecture handouts and consult the SPSS command syntax file.

1. (a) Make SPSS read your data correctly (either directly from the text file that contains the data or via an excel file). Then check the data thoroughly. That is, inspect them visually (in the raw data file), look at the DESCRIPTIVES and the FREQUENCIES (all caps refer to SPSS commands). Be sure to inspect all variables, but send your TA only the DESCRIPTIVES table and two examples of FREQUENCIES output. Briefly comment on any trends or unusual findings. (b) Find the three anomalies that this data file contains. (Missing values don't count as anomalies.) (c) Correct the anomalies and justify your method of correction.

2. (a) EXAMINE the distributions of continuous variables. Do we have reasonably normal distributions? Use the EXAMINE subcommand /PLOT=NPPLOT to this end. Include one example of such a plot and briefly discuss it. (b) Do we have outliers? Inspect all univariate boxplots in EXAMINE as well as bivariate plots (e.g., with the GRAPH /SCATTERPLOTS command). In the boxplots, don't be too concerned about what the EXAMINE procedure calls "outliers." These are values that are 1.5 to 3 IQR (inter-quartile ranges) away from the median, and that is often within the 95th percentile. Serious concern should begin with the "extremes," which are more than 3 IQR away from the median. Include a boxplot from one variable (or variable set) and one bivariate plot. Comment on any extreme values within other variables verbally. (c) To examine the ordinal variables (like ADJEXT or IAMINT), bivariate plots can be difficult to read because the variables have very few values. You can complement your exploration by also asking for CROSSTABS, the command for multivariate crosstabulations. Either way, find out what the distributions of these variables are, on their own and correlated with one another. Select one or two graphs/crosstabulations, interpret them, and comment on the remaining distributions and outliers verbally.
3. Suggest transformations we might want to try for any non-normal distributions in this data set. Proceed as follows: Select two appropriate transformations and apply each to two different variables (for a total of four examples) and compare the results of your transformations within and across variables. Use EXAMINE to inspect any changes in normality, skewness, etc. Do the transformations succeed in improving the distributions?
4. (a) A good alternative to transformation of individual variables is to compute aggregates (= sums or more typically averages across several variables that hang together in some way). With the COMPUTE command, form two comprehensive aggregate variables that make sense). Justify why you are aggregating a particular set of variables. (b) EXAMINE the effects of the two aggregations, especially on normality, skewness, and kurtosis. Include one graph to illustrate your discussion.
5. (a) Edit your output such that you have no more than 11 pages of text, tables, and graphics. Don't forget to always include the SPSS syntax you used! (b) Add as your last page (p. 12) a one-page summary of your exploratory analyses, using the language and format of empirical journal articles.

Enjoy!