Data Driven Decision Assignment #1
Data Analytics Assignment #1

April 26, 2018

1 Questions

Consider the data set `dataset-assignment.csv` in the Dropbox shared folder.

1. Import the dataset into a `pandas` `DataFrame`. Then, for each variable:

   - Check that the dataset has been correctly imported in the `DataFrame` looking for `NaN` values.
   - Compute descriptive statistics using `pandas` methods.
   - Draw the frequency histogram and the boxplot using `pandas` methods.
   - Calculate the correlation matrix and represent it with `matplotlib` (see Lecture 8)
   - Represent the scatterplots of the three most correlated pairs of variables

2. Compute *sample mean m* and *sample standard deviation s* for all variables.

3. Write a `python` function that computes *confidence interval* for the population mean under the assumption that all are normally distributed.

4. Calculate confidence interval with $\alpha = 0.05$ and $\alpha = 0.01$ for all variables and check your computation with other tools.

5. Choose the variable that exhibits the *largest variation coefficient*, $\frac{s}{m}$. For such a variable, compute the *empirical confidence interval* for the population mean and the population median using *resampling with bootstrap* (with $\alpha = 0.01$)

6. For the same variable, test the *null hypothesis* that the population mean is equal to the sample median ($\alpha = 0.01$) using resampling with bootstrap.

7. Perform *Principal Component Analysis* on the individuals with scaling. Produce the biplot and the scatterplot of units on the factorial plane. Comment on the % of variance explained, on collinearity between variables, on the main variables that influence the interpretation of the factorial plane.

8. Cluster the individuals using *K-means* for $k = 2, \ldots, 20$. Choose the ideal number of clusters according to a method of your choice and provide motivations for your choice. Exhibit and comment the silhouette plot of the chosen clustering. Provide the plot of the clustered points on the factorial plane.
9. Provide a general comment on the findings of your analysis.

2 Rules

- **Due date: Monday, May 15**
- The assignment must be sent by e-mail to both `giovanni.felici@gmail.com` and `fabrizio.rossi@univaq.it`
- People attending exclusively Data Analytics class must answer Questions 1–4. The confidence interval definition can be found on prof. Felici slides in the shared folder.
- The answers to Questions 1–4 must be provided on a `jupyter` notebook.
- You can use any tool to answer Questions 5–9. Your scripts must contain a clear explanation of the tools and the functions used.
- The answers to questions 5–9 can be provided on a `jupyter` notebook as well as on MS Word, MS Excel or PDF file.
  In the latter case, do not exceed 6 (six) A4 pages length, including tables, plots and charts