Data Science
Exploratory Data Analysis
02/2020

UNIVERSITÄT
LIECHTENSTEIN

Exercise 1: Data Preparation

a) Load the wine dataset in the data folder and produce the following textual descriptions:
   - Overview of the data, e.g. what it is about? (Hint: Explore and try to find it on the WWW?)
   - The variables, e.g. how many are there, what are the types (quantitative, nominal, ordinal)

b) Explore: Are there any unusual values in respect to one or more variables?

c) Inspect data quality: Are there missing values? Are there values hinting data quality problems?

d) Clean: Remove observations that have missing values or replace missing value by for example the mean value or the most common element.

Exercise 2: Global Exploration

a) Distribution analysis: Pick an arbitrary “interesting” variable and:
   - Create a histogram to show the distribution of values, and plot the mean and median lines on top of it.
   - What does the histogram tell you, what about the mean and median values, are they a representative measure of central tendency in this particular case?

b) Group analysis: Pick (or create) a second variable to group your first variable with and:
   - Summarize the grouped data by a measure, e.g. mean, median, or standard deviation.
   - Create a bar chart, and write down your observation, e.g. are there differences among groups?
   - Hint: You might create a categorical variable of a continuous variable by defining ranges: “low” for [0,100], “high” for [100,200]

c) Relationship analysis: Pick two continuous variables and plot them against each other on a scatter plot. Is there any visible trend?

Exercise 3: Group Exploration

a) Group comparison: Pick one continuous variable and one categorical variable. Compare the distributions of the continuous variables between each category using one or more of these seaborn functions boxplot, violinplot, stripplot, or swarmplot.

b) Explore all possible relations using Scatter Plot Matrix. What do you see?
   Hint: Use the PairGrid function.

c) Explore the correlations between at least 4 variables by creating a correlation matrix both in table and (scatter) plot form. What do you notice? Can this information be used for classification, ie. predicting a class? Can it be used for regression?