Data Analysis

Task 1: EDA

Tomas Raila

Vilnius University
Institute of Computer Science

February 6, 2024

Dataset selection

Task objective: perform exploratory data analysis on a selected dataset.

Dataset requirements:

- Tabular data in CSV (or similar) format
- At least 500 objects
- At least 5 continuous variables (interval/ratio scales)
- At least 3 categorical variables (nominal/ordinal scales)
- Some domain knowledge
  - At least understand what each variable means
  - More knowledge = better!
- Selection should be unique and confirmed by Tomas.

Tools: Python + libraries (numpy, pandas, matplotlib, scikit-learn), Jupyter notebook

Checklist (1)

The analysis should cover the following:

- General overview of the dataset
  - Number of objects, descriptions of variables
  - Identify variable types (nominal/ordinal/interval/ratio?)
  - Sample rows
- Missing values
  - Detect and describe
  - Identify possible reasons/meaning
  - Suggest and apply actions: ignore, delete, replace, etc...
- Outliers
  - Detect and describe
  - Identify possible reasons/meaning
  - Suggest and apply actions: ignore, delete, replace, etc...

Checklist (2)

- Feature engineering
  - Any redundant/uninformative variables which could be removed from the dataset?
  - Try to derive at least 2 different additional variables which you think might be meaningful and informative
    - ...or provide solid arguments why it is not worth doing.
    - Good example: given mass (m) and volume (V) of a physical object you could compute density ($\rho = m/V$), or apply some nonlinear function (logarithm, exponent, etc), or something else.
    - Bad example: simple linear transformation $y = ax + b$, where a and b - constants.
- Univariate analysis (individual variables)
  - Basic statistics: mean, variance, etc...
  - Visualizations: bar charts, histograms, box plots, etc...
- Bivariate analysis (relationships between variables)
  - Correlation & covariance
  - Visualizations: scatter plots, line plots, 2D histograms, etc...

Checklist (3)

- Result interpretation
  - 5 most important conclusions stated in domain terms
  - Connection between the numbers and charts and your own knowledge
  - Any unexpected, surprising or otherwise interesting facts about the data? Can you prove/disprove any of your initial assumptions based on the results?

Other info

- Maximum grade for the task: **1.0**
- Deadline for task submission to Gitlab: **2024-02-20 16:00**
- Task submission process:
    - Jupyter notebook + dataset (.csv file) uploaded to Gitlab, according to instructions presented in lecture 1.
    - Task presented and explained during exercise class (Tuesdays/Fridays, 18:00 - 19:30)