Laboratory work No.1. Dataset Processing and Analysis

1. Select (create) a dataset[1,2,3] to perform this and other laboratory works. Your choice must be approved by the tutor.

   Data set requirements:
   - Numeric (*integer* and *real*) and categorical values must exist.
   - For a dataset, the number of records (rows) m must be at least 500, i.e., $\infty > m \geq 500$ and the number of attributes *n* must be at least 8 (columns) $\infty > n \geq 8$. If there are fewer attributes in the selected dataset, you have to add derivatives (created) (see Figure 1.)

Important. The following tasks must be implemented programmatically using Python (in exceptional cases - *Matlab* if you have no knowledge on *Python*)

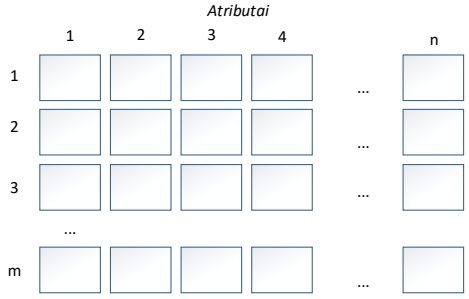

Figure. 1. Dataset graphical view

Perform dataset quality analysis (see Figure 2).

2. For each *countinuous* (numeric) type attribute calculate:
   - total number of values,
   - percentage of missing values,
   - cardinality,
   - minimum (min) and maximum (max) values,
   - 1st and 3rd quartiles,
   - average,
   - median,
   - Standard deviation.
3. For each *category* type attribute calculate:
   - total number of values,
   - percentage of missing values,
   - cardinality. *Cardinality* in mathematics is a property of a set that summarizes the number of members of a finite set concept. Simply put, how many different attribute values are there. For example, the cardinality of the gender attribute equal to 2 - i.e. gender can only have two values.

[1] https://archive.ics.uci.edu/ml/datasets.php
[2] https://vincentarelbundock.github.io/Rdatasets/datasets.html
[3] https://www.kaggle.com/datasets

- Mode. *Mode* is the most frequent sample value.
- The frequency value of the mode
- Percentage value of the mode
- Second mode value (*2nd mode*),
- Frequency value for *2nd mode*,
- Percentage of *2nd mode*.

Compose a table to present each type of variable – continous and categorical

Continuous attribute name	total number of values	percentage of missing values	cardinality	min	max	1st quartile	3rd quartile	average	median	Standard deviation

Categorical attribute name	total number of values	percentage of missing values	Cardina-lity	Mo-de	Frequency value of the mode	Percentage value of the mode	2nd mode value	2nd mode frequency value	Percentage of 2nd mode

Figure 2. Evaluation of quality of continous and categorical attributes of dataset

4. Draw histograms of attributes (recommended number of histogram columns is defined by a formula: $1 + 3.22 \cdot log_e^n$, where *n* is sample size). Provide descriptions of the distribution (e.g., normal, exponential, etc.) and what conclusions can be drawn from it.
5. Identify data quality problems: missing values, cardinality problems, outliers. Provide a plan for resolving these issues, which will be implemented programmatically (e.g. missing categorical attribute values will be included based on an attribute‘s mode estimate, extreme values are eliminated or adjusted).
6. Investigate relationships between attributes using visualization techniques:
   For continuous type attributes:
   - Using a scatter plot type graph, provide *several* (2-3) examples with strong linear attribute dependency (direct or inverse correlation) and *several* examples with non-correlated (weakly correlated) attributes. Comment the results.
   - Provide SPLOM diagram (Scatter Plot Matrix).

   For categorical attributes:
   - Using the bar plot type diagram, give some (2-3) examples of attribute dependency and comment the results.
   - Provide some (2-3) examples of histograms and box plot diagrams depicting relationships between categorical (see 3 fig.for reference) and continuous type variables.
7. Calculate the covariance and correlation values between continuous attributes and graphically represent the correlation matrix. Comment the results.
8. Perform data normalization (boundaries [0; 1] or [-1; 1]).
9. Convert categorical variables to continuous type variables.

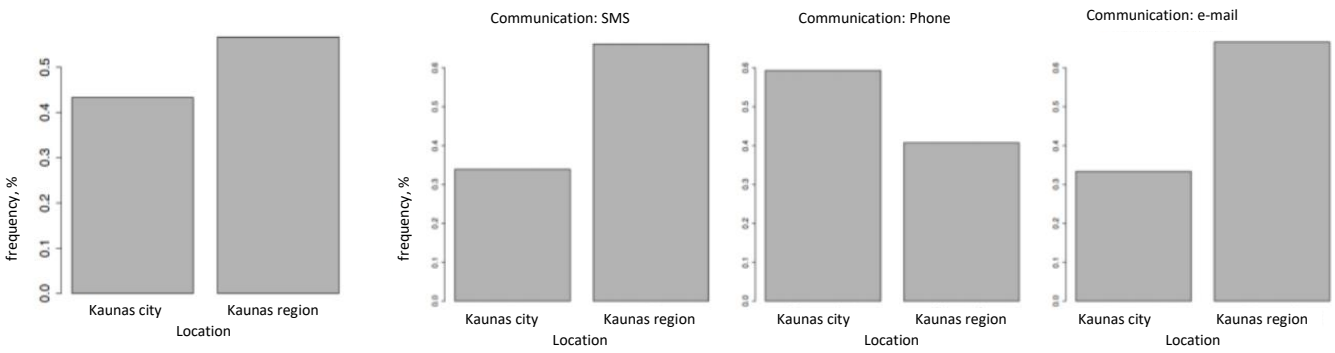

Figure 3. A bar plot type diagram showing: a) A histogram of one categorical type attribute "Location"; b) and the relationship between the two categorical attributes "Location" and "Communication".