CALIPER: A Toolbox for Exploratory Analysis of Experiment Data

F. Schwager and A. Vijayachandran

Otto-von-Guericke-Universität Magdeburg Fakultät für Informatik (FiN) AG KMD: Wissensmanegement und Wissensentdeckung Gebäude 29, Universitätsplatz 2, 39106 Magdeburg, Germany

Abstract. This study introduces CALIPER, a comprehensive toolbox designed to facilitate exploratory data analysis for experimental data. The paper outlines the tool's capabilities in offering intuitive analysis and visualization functions, enabling users to conduct sophisticated statistical analyses without requiring in-depth programming knowledge. The research delves into the significance of exploratory data analysis in deriving meaningful insights from experimental data, emphasizing its necessity in validating hypotheses and informing further research directions.
The toolbox is built upon Python and Streamlit, providing a seamless user experience through a web-based interface. The tool's functionality includes data upload with schema validation, preprocessing steps for the data, and diverse analytical methods encompassing univariate and bivariate analyses. It effectively handles complex data structures and provides visualizations such as frequency distributions, error rates, box plots, correlation matrices, etc., which are pivotal in understanding the intricacies of the data.
The paper also explores the practical application of the toolbox in an experimental setting across three different sites, illustrating the tool's utility in analyzing variability and patterns in participant responses. The study highlights the potential of the toolbox in streamlining the initial stages of data analysis, thereby saving time and resources, and promoting a more accessible approach to data-driven investigations. The toolbox stands as a testament to the potential of integrating user-friendly software solutions with complex data analysis, broadening the scope of exploratory data analysis to a wider range of users.

Key words. Exploratory Data Analysis, Data Preparation, Data Visualization, Statistics, Electrodermal Activity (EDA) Data, Sociodemographic Data, Univariate Analysis: Distribution Summary, Bivariate Analysis: Correlation Matrix

1. Introduction

Exploratory Data Analysis is the cornerstone of any data-driven endeavor. It lays the groundwork for understanding, interpreting, and extracting meaningful insights from raw data. This crucial initial step equips us to make informed decisions and formulate impactful research questions. Delving into the book Myatt & Johnson (2007), we embark on a journey to demystify the intricacies of Exploratory Data Analysis and equip ourselves with the tools to navigate the vast landscape of information.

1.1. Data Preparation: Laying the Groundwork

Before diving into analysis, it's essential to meticulously prepare the data. This multifaceted process involves:

Understanding Data Sources: Identifying the origin and characteristics of the data helps to assess its reliability and potential biases.

Data Cleaning: Scrutinize the data for inconsistencies, missing values, and outliers. Address these issues through imputation, exclusion, or appropriate transformations.

Variable Exploration: Comprehend the nature of variables – Are they continuous or discrete? What measurement scales do they use? How do they relate and contribute to the analysis?

Frequency Distributions: Visualize how often values occur within each variable using histograms, bar charts, or frequency tables. This unveils patterns and potential skewness in the data.

Data Segmentation: Divide the data into meaningful subgroups based on relevant characteristics. This allows for targeted analysis and identification of group-specific trends.

1.2. Data Visualization: Visualizing Insights

Data visualization is a cornerstone of Exploratory Data Analysis, transforming numerical values into readily understandable formats. Myatt & Johnson (2007) highlights the power of tables and graphs in effectively communicating patterns and relationships:

Tables: Organize and summarize data in a structured manner. Data tables present raw data, while contingency tables unveil relationships between categorical variables, and summary tables condense key statistics.

Graphs: Breathe life into the data through diverse graphical representations. Histograms and frequency polygons illustrate the distribution of continuous variables, while scatterplots depict relationships between two variables. Box plots summarize the distribution of a variable across groups, and multiple graphs can be combined to reveal complex interactions.

1.3. Statistics: Quantifying the Story

Statistics provide the quantitative foundation for Exploratory Data Analysis, enabling to measure, analyze, and draw conclusions from the data. Myatt & Johnson (2007) sheds light on the key statistical concepts:

Descriptive Statistics: Summarize the central tendencies (mean, median, mode) and variability (range, standard deviation) of the data. Understand the shape of the data's distribution (normal, skewed, etc.) using measures like skewness and kurtosis.

Inferential Statistics: Draw conclusions about a population based on a sample. Confidence intervals express the range within which the population parameter is likely to reside. Hypothesis testing helps assess whether observed differences are due to chance or reflect genuine relationships.

Comparative Statistics: Explore relationships between variables. Correlation coefficients like Pearson's R and Spearman's Rho quantify the strength and direction of linear relationships. Techniques like correlation matrices and network analysis can handle more complex relationships involving multiple variables.

Exploratory Data Analysis is an iterative process. As we delve deeper into the data, new questions may arise, prompting us to revisit and refine our initial explorations.

In this project, the goal is to establish a toolbox for the Exploratory Analysis of experimental data collected at three sites, namely Magdeburg (M), Chemnitz (C) and Ilmenau (I).

A. Vijayachandran

2. Experiment

This project aims to create a toolbox for the Exploratory Analysis of the results of an experiment conducted at three sites, namely M, C and I. The experiment given in Rother et al. (2023) is an object classification scenario where the uncertainty of classification

is directly linked to the difficulty of classifying each object. By controlling uncertainty, the experiment builds up a reference dataset and investigates how different sensory inputs can serve as uncertainty indicators for these data. The uncertainty indicators considered here are sensory inputs that can be captured with a smartphone, namely electrodermal activity (EDA) and electrocardial signal (ECG), along with externally measurable indicators, such as task duration.

Rother et al. (2023) explains the experiment in detail as follows:

2.1. Objective

Investigate the ability of participants to estimate metal cylinder diameters and their subjective certainty in those estimations.

2.2. Participants

Participants are students or staff members of universities at sites M, C and I. They are recruited through stratified sampling for gender balance and age homogeneity. International students are included, requiring English for the questionnaire and offering German or English for the introduction session. Participants with self-reported visual or haptic impairments are excluded as these impairments may prevent the proper use of the caliper.

2.3. Procedure

2.3.1. Introductory Session

Instructions on the classification task and caliper usage, in both English and German.

2.3.2. Experiment Session (60-90 minutes)

Participants complete n trials (maximum 150 trials) that are completed in one session. The trials involve:

- Estimating the diameter of a metal cylinder (by heart or with a caliper). Fig. 1. shows an overview of cylinders with different degrees of difficulty, which the participants annotate.
- Classifying the cylinder as "good" (diameter $\leq d + 0.05$ mm) or "scrap" (diameter $> d + 0.05$ mm).
- Rating the certainty of classification using a Likert scale of 3 points (low, medium, high).

2.3.3. Data Collection

The following data are recorded for each trial:

- Trial completion time
- Participant responses for diameter of cylinder, 'good' or 'scrap' classification, certainty of response
- Electrodermal Activity (EDA) signals from non-dominant hand palm
- ECG signal
- Participant sociodemographic information:
  - Age
  - Gender (as stated by the participant)
  - Handedness (because it may affect the use of the caliper)
  - Eyesight (because it may affect classification quality)
  - Program of study
  - Mother tongue language
- Post-Experiment questionnaire:
  - Experiment experience
  - Perceived task difficulty

2.4. Configuration Parameters

- The number of cylinders n (same as the number of trials, upper bounded by 150).
- The number of 'easy' cylinders (the diameter can be easily estimated by heart) vs 'difficult' ones (the diameter is difficult to estimate by heart).
- The number of Likert-scale values that the participants can choose from to describe their uncertainty. There are three values, i.e., low, medium and high.

2.5. Technology Utilized

- Sensor technology EdaMove 4 (movisens GmbH) for measuring electro-dermal activity.
- EcgMove 4 (movisens GmbH) for ECG measurement.
- Data Analyzer software for analysis of both EDA and ECG signals.

A. Vijayachandran

3. Data

The list of data collected from the experiment is given in section 2.3.3. In this current section, we look at the collected data in depth, mainly the Electrodermal Activity (EDA) data and Sociodemographic data.

Electrodermal Activity (EDA), also known as galvanic skin response (GSR), is a measure of the electrical conductance of the skin, which varies with its moisture level (Li et al. (2022)). This activity is primarily influenced by the sympathetic nervous system and is considered a reflection of psychological or physiological arousal.

EDA measurement is based on the principle that the skin's ability to conduct electricity changes with the state of sweat glands. When an individual experiences emotional arousal, stress, or cognitive engagement, the sympathetic nervous system stimulates the sweat glands, increasing skin conductance. Two primary indices measured in EDA are the Skin Conductance Level (SCL) and the Skin Conductance Response (SCR). SCL provides a baseline measure of skin conductance over time, while SCR is concerned with the rapid fluctuations in conductance in response to specific stimuli.

According to Posada-Quintero & Chon (2020), the standard method for measuring EDA involves attaching electrodes to the palmar or plantar surfaces of the skin, where sweat gland density is high. These electrodes are connected to a device that sends a low-voltage current through the skin and measures the resulting conductance. The data obtained can be analyzed to understand the subject's emotional and psychological state. This makes EDA a valuable tool in various fields, including psychology, medicine, and user experience research.

Due to its non-invasive nature and direct correlation with autonomic nervous system activity, EDA is a widely used method in psychophysiological studies. It has applications in areas such as emotion research, stress analysis, lie detection, and even in the evaluation of user responses in human-computer interaction settings.

Lofters et al. (2011) states that sociodemographic data encompasses a range of variables that describe an individual's societal and demographic characteristics. This data is crucial for understanding and analyzing trends within populations. It typically includes age, gender, ethnicity, education level, income, occupation, marital status, and geographic location. The measurement of sociodemographic data is primarily conducted through surveys, censuses, and registration systems. These methods involve structured questionnaires designed to gather comprehensive and accurate information. The accuracy and reliability of sociodemographic data are vital, as it informs policy making, market research, and social science studies. Advanced statistical techniques are often employed to analyze this data, enabling the identification of patterns and correlations within diverse populations. Accurate sociodemographic data measurement is essential for effective decision-making across various sectors including healthcare, education, and urban planning.

The experiment result data are organized into two schema, namely, Answers Schema and EDA Schema. Additionally, we have one more schema named Ground Truth Schema, which stores the actual status of each cylinder used in the experiment, i.e., the ground truth value of whether a cylinder is 'good' or 'scrap'.

The EDA Schema integrates the mean EDA arousal values across all trials for all participants at each site. During the course of the experiment, EDA signals are meticulously recorded as continuous time-series data. To manage this data effectively, the Data Analyzer software is employed. This advanced tool segments the continuous time-series EDA data

Fig. 1. Overview of cylinders with different degrees of difficulty, which the participants are required to annotate. These cylinders have an outer diameter between d and $d + 0.1$ mm, in varied steps of 0.01 **mm (tolerances are Gaussian distributed)** Rother et al. (2023)

into discrete, time-based intervals, typically one-second bins. Subsequently, it computes the average EDA value for each interval. To derive a conclusive metric, the aggregate grand mean of these bin-wise average EDA values is calculated over the full time span of each trial. This final aggregate EDA grand mean is systematically stored within the EDA Schema.

For each trial, the Answers Schema contains trial completion time in seconds, all participant responses for the diameter of all cylinders, 'good' or 'scrap' classification of all cylinders and certainty of classification (low, medium or high). This schema also records the total time in seconds required for completion of all trials by each participant, along with all participant sociodemographic information like age, gender, handedness, eyesight, program of study, mother tongue language, experiment experience and perceived task difficulty.

An overview of the three schema, Answers Schema, EDA Schema and Ground Truth Schema are shown in Fig. 2, 3 and 4.

A. Vijayachandran

4. Proposed Solution: Analysis Tool

For the explorative data analysis of the experiment data, a software tool was developed that is based on the Python programming language. The used app framework is Streamlit, which generates an executable application from Python scripts without the developer needing any prior knowledge of app development (`https://streamlit.io/`). The application consists of three pages: A homepage with an introduction to the topic, a data upload page and a data analysis page, which are described in the following sections.

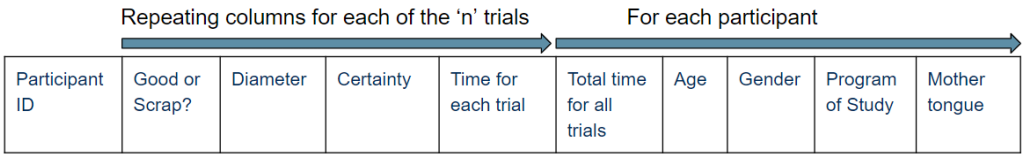

Fig. 2. Overview of Answers Schema

Participant ID/ Response ID	Task/Trial 01 EDA Grand Mean	Task/Trial 02 EDA Grand Mean	.....	Task/Trial 'n' EDA Grand Mean

Fig. 3. Overview of EDA Schema

Task ID/Trial Number	Ground Truth of cylinder

Fig. 4. Overview of Ground Truth Schema

4.1. Data Upload and Preprocessing

Before running any analysis on the collected experiment data, the data has to be uploaded into the database of the analysis tool. The upload includes multiple validation steps which ensure the error-free execution of the subsequent analysis functions:

1. data schema validation
2. naming convention
3. preprocessing of categorical attributes

Each of these steps is explained to the user in the respective sections of the **data upload** page of the app. In the first step, the user is introduced to the data schema of the experiment data that was described in section 3 and shown in the figures 2, 3 and 4. This is done by displaying the schema tables that contain the expected column names and the expected data type for the given column name (attribute name). E.g. the 'age' column of each participant in the survey data may only be of type integer, which is a positive number without decimals. Upon upload the column names and respective data types are automatically validated before adding the data to the database. In case of an error, an error is displayed explaining which column name or data type was unexpected.

The user is expected to upload the data as CSV files in the following naming convention to be able to distinguish the tables of the different experiment sites:

- answers-site-X.csv
- eda-site-X.csv
- ground-truth-site-X.csv

Where X can be substituted with a single letter abbreviation for the experiment site (e.g. M for Magdeburg)

Once the data is successfully uploaded, the user can view every table to double-check if the data is correct. Additionally, he can use the preprocessing function "get unique values for categorical attributes". In the displayed result, the user can check for unique attribute values, that semantically mean the same. E.g. for the survey question regarding the 'program of study', some participants answered in German and some in English. The resulting answers contained: 'Informatik', 'Computer Science', 'Phd Computer Science' which all have the same meaning. If we would leave the data, each of these values would be treated as a unique value in the analysis and therefore would influence

the distribution analysis. In order to avoid this, the user would have to find a single word for all semantically equal attributes (e.g. use the German version 'Informatik' XOR the English version 'Computer Science'). The deletion of the old data, the replacement of the values and the upload has to be done by the user himself.

The given data is in tabular form, where each row represents a participant of the experiment and each column the different attributes of a participant response. In the preprocessing as well as in the analysis of the data, the following attribute types have to be differentiated by their scales [Fahrmeir et al. (2016)]:

- Nominally scaled: A categorical (or also: nominally scaled) attribute has two or more values (categories) that have no intrinsic order. Example: Program of study
- Ordinally scaled: An ordinally scaled variable is a categorical variable with intrinsic order. However, no differences can be calculated between the values. Example: participant's certainty on answer
- Interval scaled: In contrast to ordinal-scaled variables, the distances between values can be interpreted. Example: Age

F. Schwager

4.2. Data Analysis

4.2.1. Univariate Analysis: Get Distribution Details

Univariate analysis is a fundamental form of statistical analysis where the focus is on a single variable (Cleff (2014)). Its primary purpose is to describe the data and find patterns that exist within it. This type of analysis is often the first step in a statistical examination, offering a detailed look into the characteristics of the variable under consideration.

Key measures in univariate analysis as described by Cleff (2014) include central tendency (mean, median, mode), dispersion (variance, standard deviation, range), and shape (skewness, kurtosis). These measures help in understanding the distribution, central value, and variability of the data.

For visualization of univariate analysis, Cleff (2014) suggests tools like histograms, box plots, and bar charts. A histogram is effective in displaying the frequency distribution of a variable, revealing patterns like bimodality or skewness. Box plots are useful for identifying outliers and understanding the spread of the data. Bar charts are particularly effective for categorical data, showing the frequency or proportion of each category.

Incorporating these measures and visualizations into univariate analysis provides a comprehensive understanding of the single variable, which is essential before moving to more complex multivariate analyses.

In this project, we have defined a function named "Get Distribution Details" to carry out all the univariate analysis. The various measures calculated by the "Get Distribution Details" function, along with their inputs, visualizations and interpretations are described below:

1. Frequency Distribution with Mode: Frequency distribution, in statistical analysis, is a comprehensive representation that showcases how often each distinct value occurs within a dataset. It effectively organizes data by mapping values to their corresponding frequencies, either in a tabular format, known as a frequency table, or graphically, through histograms, bar charts, or pie charts. This distribution provides a visual insight into the data's overall pattern, highlighting concentrations, spread, and outliers.

The mode, a key measure of central tendency within a frequency distribution, identifies the most frequently occurring value(s) in a dataset. It is particularly useful in understanding the commonality or popularity of certain responses or characteristics.

User Input: The user is given the option to choose one or all of the sites as well as the attribute on x-axis of the frequency distribution bar chart. The possible values of attributes on x-axis are:

- Good/Scrap response for all cylinder tasks
- Certainty of response for all cylinder tasks
- Gender
- Handedness

- Program of study
- Mother tongue language

Visualization: The frequency distribution is visualized as a bar chart, with the selected attribute on x-axis and the corresponding count of participants on the y-axis. The mode for the distribution is given on the top right corner of the bar chart, as shown in Fig. 5.

If more than one site is selected, the bars for different sites are grouped as shown in Fig. 6. Note that the data for sites C and I are sample representative values only.

Interpretations: Interpretations derived from the analysis of frequency distribution with mode offer valuable insights into the underlying patterns and tendencies within a dataset.

- Dominant Trends: The mode within a frequency distribution pinpoints the most common value of an attribute. It helps to answer the following questions:
  - What is the gender/program of study/mother tongue of most of the participants?
  - For each cylinder task, what is the good/scrap response of most of the participants?
  - For each cylinder task, what is the certainty of response of most of the participants?
- Comparative Analysis: The values of the x-axis attribute can be compared for the three sites, to unveil similarities and differences among the sites.

2. Error Rate: Error rate is a fundamental metric in statistical analysis, quantifying the frequency at which errors occur within a dataset. It is typically expressed as a percentage, representing the ratio of incorrect values to the total number of cases examined. Error rate serves as a critical indicator of the accuracy and reliability of the data.

User Input: The user is given the option to choose one or all of the sites as well as the sorting criteria based on task ID of cylinder tasks or descending order of error rate.

Visualization: If only one site is selected, the error rate is visualized in the form of a table as well as a bar chart, as seen in Fig. 7 and 8.

If more than one site is selected, a grouped bar chart is shown, as in Fig. 9. Note that the data for sites C and I are sample representative values only.

In the Error Rate table, columns represent participant IDs and rows represent the cylinder task IDs. If the participant's response to a cylinder task matches with the ground truth value of the task, the cell is given the value '0', which means error is absent. If the participant's response to a cylinder task does not match with the ground truth value of the task, the cell is given the value '1', which means error is present.

Error Rate is then calculated by the following equation:
$ErrorRate = \frac{Count of errors present}{Total number of participants}$.

Interpretations: The main interpretation from these visualizations is that "higher the error rate, tougher the cylinder task". In other words, we can get answers to the following questions:

- Which is/are the toughest cylinder task(s)?
- Which is/are the easiest cylinder task(s)?
- Which is/are the cylinder task(s) where the participants made most errors?
- Which is/are the cylinder task(s) where the participants made least errors?

The grouped error rate bar chart for multiple sites helps to compare how the participants of each site responded to each cylinder task.

3. Mean, Median, Percentiles, Interquartile Range (IQR), Range, Standard Deviation, Outliers: In statistical analysis, the central tendency and variability of data are captured by several key metrics:

- **Mean**: The arithmetic average, obtained by dividing the sum of all values by the number of values.
- **Median**: The middle value when data points are arranged in ascending order, effectively splitting the dataset into two halves.
- **Percentiles**: Points in the data below which a certain percentage falls, with the 50th percentile being equivalent to the median.

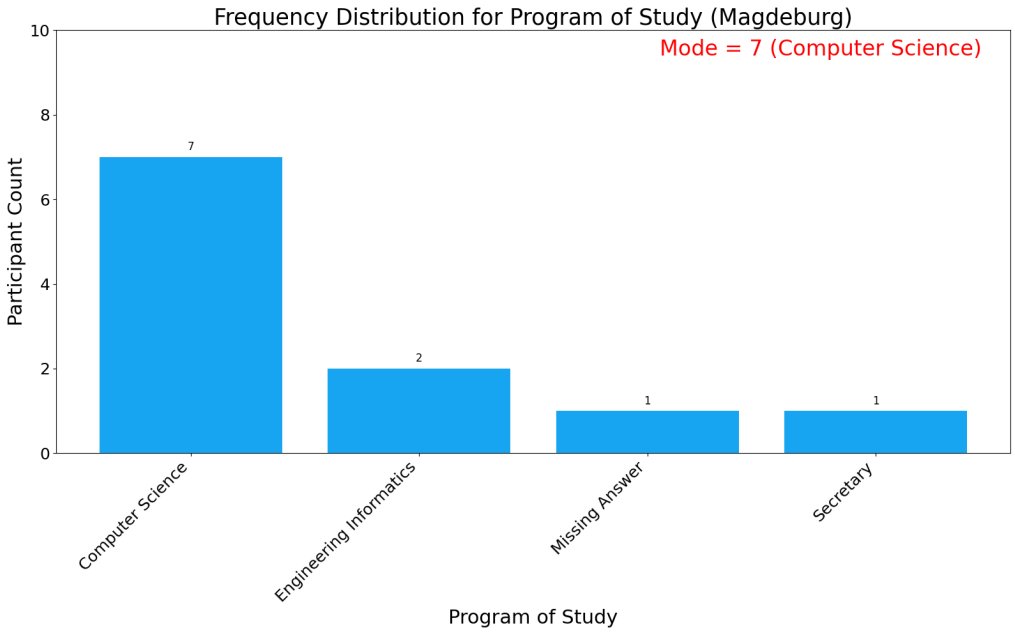

Fig. 5. Frequency Distribution for Program of Study for Magdeburg site

- **Interquartile Range (IQR)**: The range between the 25th and 75th percentiles, providing a measure of the middle spread of data.
- **Range**: The difference between the highest and lowest values, offering a simplistic view of data spread.
- **Standard Deviation**: A measure of dispersion that quantifies the amount of variation or dispersion of a set of values.
- **Outliers**: Data points that deviate markedly from other observations, which can often influence the aforementioned quantities significantly.

User Input: The 'Mean, Median, Percentiles, Interquartile Range (IQR), Range, Standard Deviation, Outliers' are calculated based on the following criteria:

- All Participants, One Task, One or All Attributes
- All Tasks, One Participant, One or All Attributes
- All Participants, All Tasks, One or All Attributes

The user is given the option to choose:

- One or all participants
- One or all cylinder tasks
- One or all attributes

Visualization: Based on the choice of the user, box plot and summary table are created, for each attribute chosen. (Fig. 10 and 11)

If more than one site is selected, the box plots are grouped, as shown in Fig. 12. Note that the data for sites C and I are sample representative values only.

Interpretations: The interpretation of statistical quantities such as mean, median, percentiles, interquartile range (IQR), range, standard deviation, and outliers provides a multifaceted view of a dataset's characteristics:

- **Mean**: Reflects the average value, giving an overall impression of the dataset. A high mean relative to the median may indicate a right-skewed distribution, suggesting the presence of high-value outliers.
- **Median**: Represents the middle value, which is not affected by outliers. A median

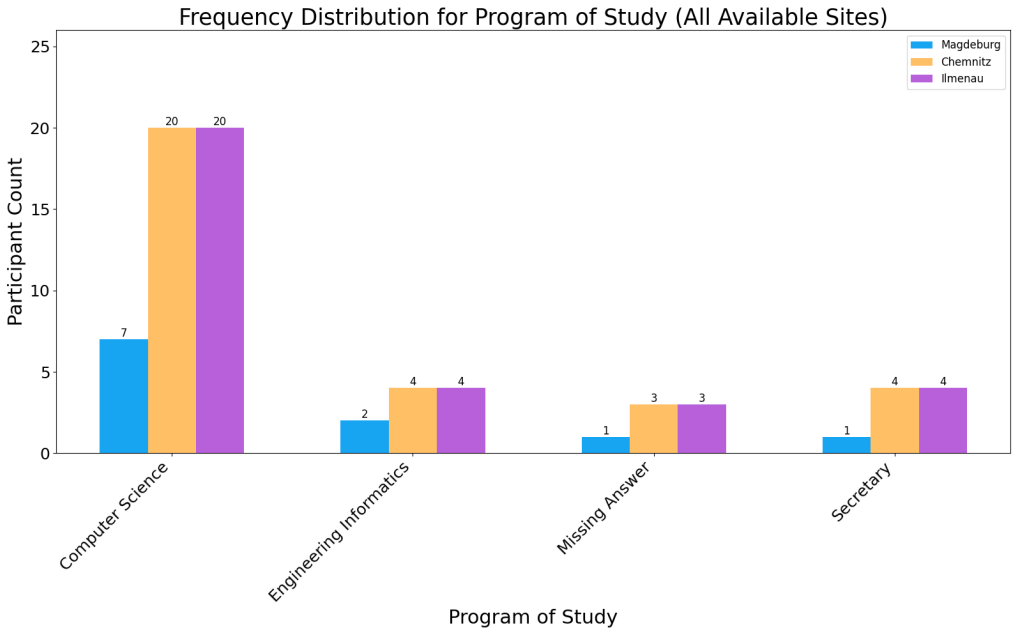

Fig. 6. Frequency Distribution for Program of Study for all available sites

that is significantly different from the mean can signal a skewed distribution.

- **Percentiles**: Offer a deeper understanding of the distribution by indicating the value below which a given percentage of observations fall. For example, the 25th percentile (1st quartile) and 75th percentile (3rd quartile) provide insights into the lower and upper distribution, respectively.
- **Interquartile Range (IQR)**: Measures the range of the middle 50% of the data, providing a sense of the central spread and helping to identify outliers. A larger IQR indicates greater variability around the median.
- **Range**: The difference between the maximum and minimum values, indicating the full spread of data. A wide range may suggest the presence of extreme values.
- **Standard Deviation**: Quantifies the average dispersion of the data points around the mean. A larger standard deviation indicates more spread out data, while a smaller standard deviation indicates that the data points are closer to the mean.
- **Outliers**: They are data points that differ significantly from other observations. They can influence the mean and standard deviation. Identifying outliers is important for understanding potential anomalies or errors in the data.

A. Vijayachandran

4.2.2. Bivariate Analysis: Get Correlation Matrix

In the following, we will look at the relationship (statistical dependence) between two variables with each other, i.e. how to perform a bivariate analysis [Oluleye (2023)].

Baker, Lee (2019) phrased the differentiation between correlation and association in the following: ”When you are looking for a relationship between two numerical variables, such as age and EDA grand mean, then the test you use is called a correlation. If one or both of the variables are categorical, such as the task correctness (True or False), then the test is called an association. When you can phrase

	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	Error Rate
T11	0	1	1	1	0	1	1	1	1	1	1	0.8182
T12	0	1	1	1	0	1	1	1	1	1	1	0.8182
T07	0	1	1	1	0	1	0	1	1	1	1	0.7273
T10	0	1	1	1	0	1	0	1	1	1	1	0.7273
T08	0	0	1	1	0	1	0	1	1	1	1	0.6364
T09	0	1	1	1	0	1	0	1	0	1	1	0.6364
T06	0	0	1	1	0	1	0	1	0	1	1	0.5455
T02	1	1	0	0	1	0	1	0	0	0	0	0.3636
T03	1	1	0	0	1	0	1	0	0	0	0	0.3636
T15	1	1	0	0	1	0	0	0	1	0	0	0.3636
T16	1	1	0	0	1	0	0	0	1	0	0	0.3636
T01	1	0	0	0	1	0	1	0	0	0	0	0.2727
T04	1	0	0	0	1	0	1	0	0	0	0	0.2727
T05	1	0	0	0	1	0	1	0	0	0	0	0.2727
T13	1	1	0	0	1	0	0	0	0	0	0	0.2727
T14	1	0	0	0	1	0	0	0	0	0	0	0.1818

Fig. 7. Error Rate table for Magdeburg, sorted in descending order of error rate

your hypothesis (question or hunch) in the following form, then you are talking about the relationship family of statistical analyses":

- Is the EDA value of the participant related to the correctness of the task?
- Are the age of a participant and the time to solve a task correlated?
- Is the certainty of a participant associated with the task correctness?

User input: Once the user knows, which question he wants answered, he must specify the respective input parameters. The following input must be given to retrieve the data:

- Sites: Should only the data of one or multiple sites be considered for the analysis (Magdeburg, Chemnitz, Ilmenau)?
- Participant: Should all participants be considered or only one (e.g. M3 is the third participant at the Magdeburg site)?
- Tasks: Should all experiment tasks be considered or only one of the sixteen?

There are three different combinations for the selection of the participant and task:

- all participants and one task →How did all participants do over one task?
- one participant and all tasks →How did the participant do for one task compared to another?
- all participants and all tasks →How did all participants do over all tasks?

Results: The displayed correlation matrices show all possible attribute combinations and were created with the Python package Dython. It was written by Shaked Zychlinski in 2018 and uses the following measures to calculate the correlation/association between attributes: [Zychlinski (2018b), Zychlinski (2018a)]:

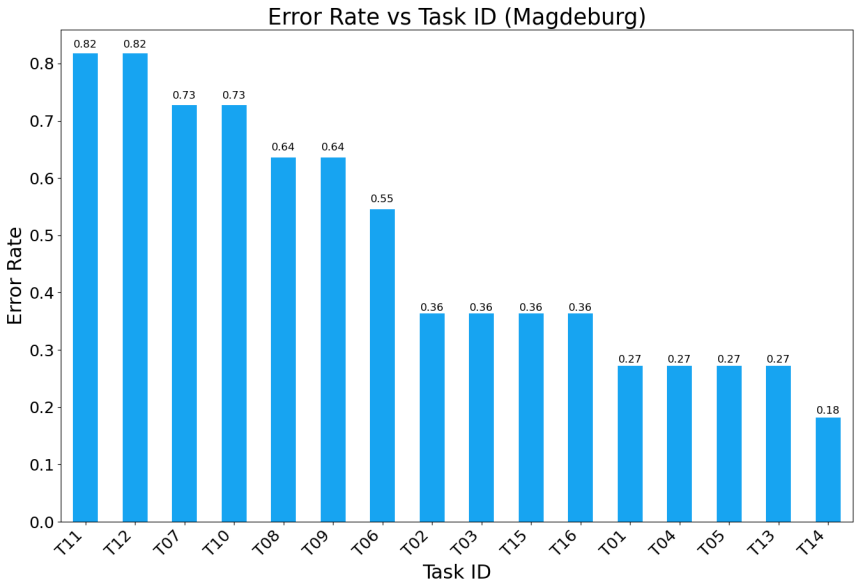

Fig. 8. Error Rate vs Task ID Bar Chart for Magdeburg, sorted in descending order of error rate

Pearson's R is a measure of the linear correlation between two numerical attributes, with an R value ranging from -1 to 1. A positive value indicates a positive linear correlation, a negative value signifies a negative correlation, and an R value close to 0 suggests a weak or no linear correlation between the attributes [Baker, Lee (2019), University of Newcastle (2024)]. The table 1 gives a more detailed overview of the strength of correlation and figure 13 displays different distributions of two numerical attributes and which correlation coefficient is calculated.

The Correlation Ratio assesses the association between a categorical and a numerical attribute. It provides a value between 0 and 1, where 0 implies no association, and 1 signifies a perfect association. Higher values indicate a stronger relationship between the categorical and numerical variables Hazewinkel, M. (2013). It helps to answer the following question "Given a continuous number, how well can you know to which category it belongs to?" Zychlinski (2018b)

Cramer's V is a measure of association between two categorical variables. It produces values between 0 and 1, where 0 indicates no association, and 1 represents a perfect association. The interpretation is that higher values suggest a stronger association between the categorical variables in the analysis Sun et al. (2010).

Interpreation: The resulting three matrices for each user input combination show the different attribute type combinations. One for numerical attributes only (example: figure 15) and the correlation between them, one for categorical attributes only (example: figure 14) and the association between them and one for all attributes that contain correlation and association values (example: figure 16).

In order to help a data analyst with the interpretation of the results, we want to give the follwing hints:

- The correlation/association values that are particularly intriguing are those associated with the task's correctness and all other attributes. These values offer insights into the

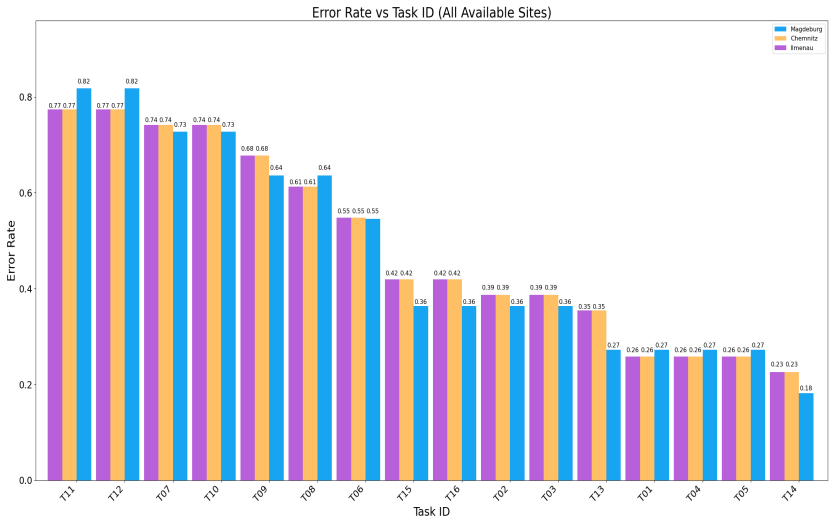

Fig. 9. Error Rate vs Task ID Bar Chart for all available sites, sorted in descending order of error rate

	Time for each Task
Count	16
Mean	40.0975
Standard Deviation	29.1190
Minimum Value	18.5700
25th Percentile	22.6200
50th Percentile or Median	27.3050
75th Percentile	46.6100
Maximum Value	126.2500
Inter-Quartile Range	23.9900
Range of Values	107.6800

Fig. 10. Summary table for Magdeburg, participant ID 11, all tasks and attribute 'Time for each Task'

primary factors (predictors) on which correctness predominantly depends.

– If e.g. in the matrix for the numerical attributes (figure 14) a negative Pearson's R value is present for two numerical attributes (e.g. in the cell of Certainty and EDA Grand Mean), than there is an inverse relationship: when one variable goes up, the other tends to go down, and vice versa (see figure 13).

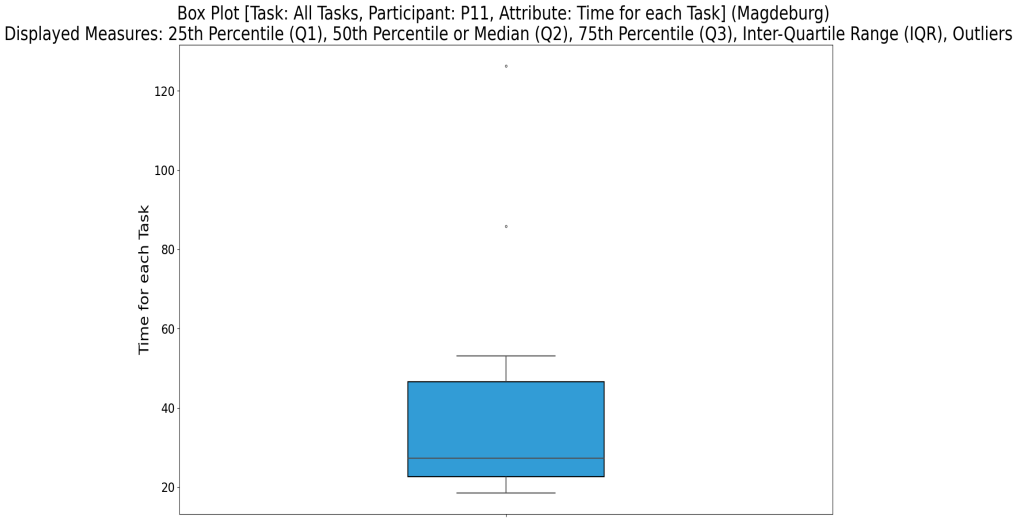

Fig. 11. Box Plot for Magdeburg, participant ID 11, all tasks and attribute 'Time for each Task'

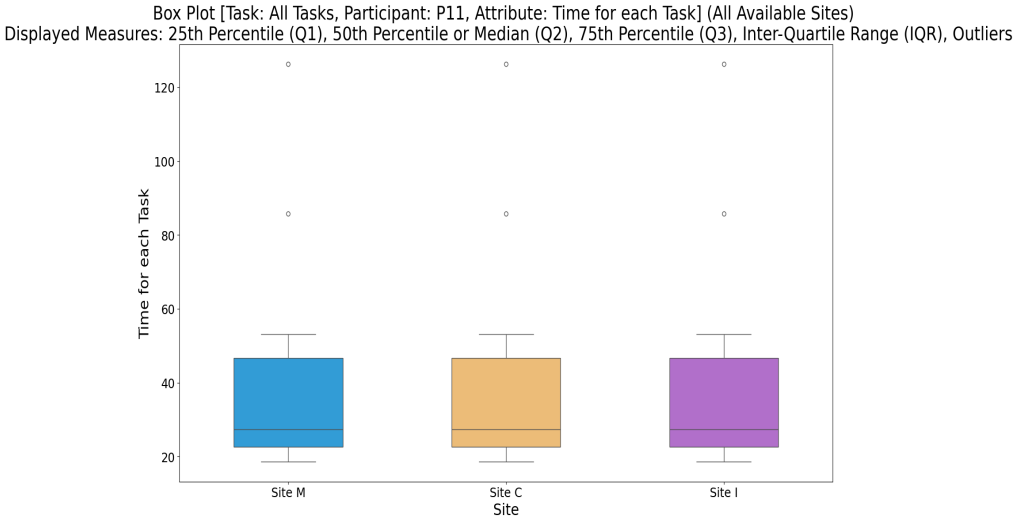

Fig. 12. Box Plots for all available sites, participant ID 11, all tasks and attribute 'Time for each Task'

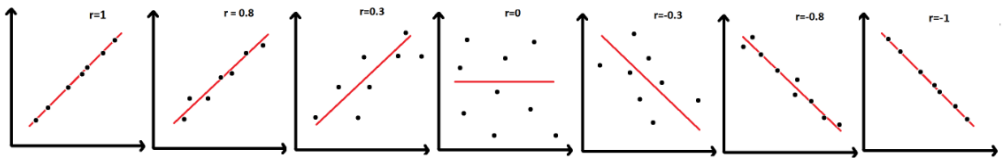

Fig. 13. Strength of correlation How to interpret the Pearson's R correlation coefficient [University of Newcastle (2024)]

Table 1. Value ranges for Pearons's R [University of Newcastle (2024)]

Pearson's R	Range
Perfect negative linear correlation	-1
Strong negative linear correlation	$-0.8 > r > -1$
Moderate negative linear correlation	$-0.4 \geq r > -0.8$
Weak negative linear correlation	$0 \geq r > -0.4$
No correlation	0
Weak positive linear correlation	$0.4 \geq r > 0$
Moderate positive linear correlation	$0.8 \geq r > 0.4$
Strong positive linear correlation	$1 > r \geq 0.8$
Perfect positive linear correlation	1

Table 2. Value ranges for Cramer's V Sun et al. (2010)

Cramer's V	Range
Small effect	$0.07 \geq v > 0.21$
Medium effect	$0.21 \geq v > 0.35$
Large effect	$v \geq 0.35$

– "The balance in the value distribution of variables can influence correlation measures, particularly in the context of skewed or unbalanced distributions. Correlation measures, such as Pearson's correlation coefficient, are sensitive to outliers and extreme values [Kim et al. (2015)]."
– "In statistics you can never prove that there is a relationship between a pair of variables but the strength of the relationship (e.g. see 13) gives an indication of the likelihood of a dependence [Baker, Lee (2019)]."

F. Schwager

5. Conclusions

Summary - The objective of this project was to develop a software tool for exploratory data analysis of experimental data. Our primary focus was not to provide an exhaustive interpretation of the analysis results but to enable users to do so. This was achieved by offering users an application where they can upload, preprocess, view and analyze their data. Exploratory data analysis plays a crucial role in uncovering patterns within experiment data, and our tool can help to save valuable time for analysts. While Python allows anyone to create analysis functions, the convenience of a user-friendly interface stands out in comparison to command-line tools or software libraries. To share insights with fellow developers of data analysis tools, we observed during the development process that Streamlit provides a quick solution for creating web apps dedicated to data processing. However, like any framework, it has its limitations, such as the automatic page refresh when an option is selected in a multiple-choice select box.

Future work - Looking towards the future, we propose expanding the tool's functionalities. This could involve implementing an au-

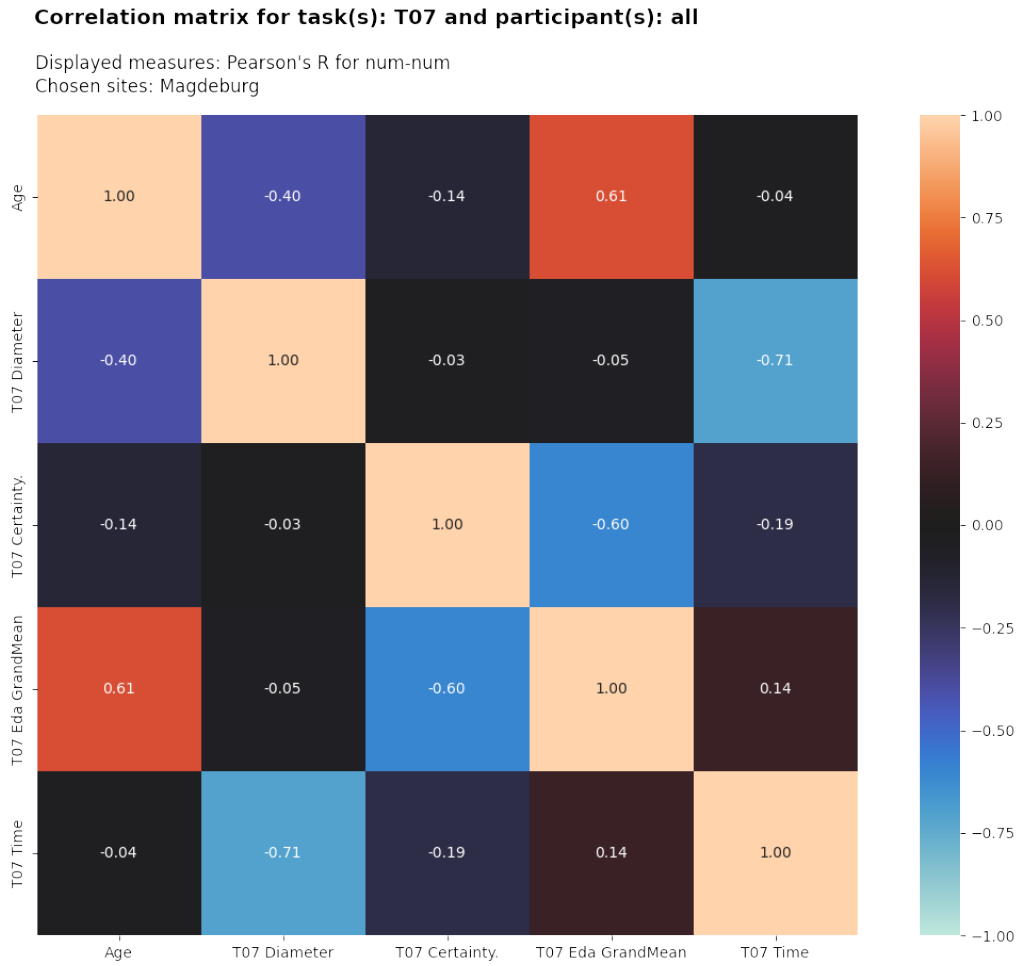

Fig. 14. Numerical attributes only Correlation matrix for displaying the Pearson's R correlation coefficient (created with Dython Zychlinski (2018a))

tomatic check for semantically equal categorical values in the preprocessing phase, introducing a function to calculate the Jaccard-Index for identifying similarities between experiment participants from different sites, and incorporating additional visualizations such as scatterplot matrices with regression lines for numerical attributes.

F. Schwager

Acknowledgements. We want to thank Prof. Myra Spiliopoulou for the invaluable guidance and profound expertise throughout this project, Anne Rother for the helping us with the preprocessing of the EDA data and Giuliana Fiorentino (e-mail: `giuliana.fiorentino@inaf.it`), the author of this freely available LaTeX paper template.

References

Baker, Lee. 2019, Associations and Correlations (Packet Publishing)

Cleff, T. 2014, Exploratory Data Analysis in Business and Economics: An Introduction Using SPSS, Stata, and Excel (Springer), 23

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. 2016, Statistik: der

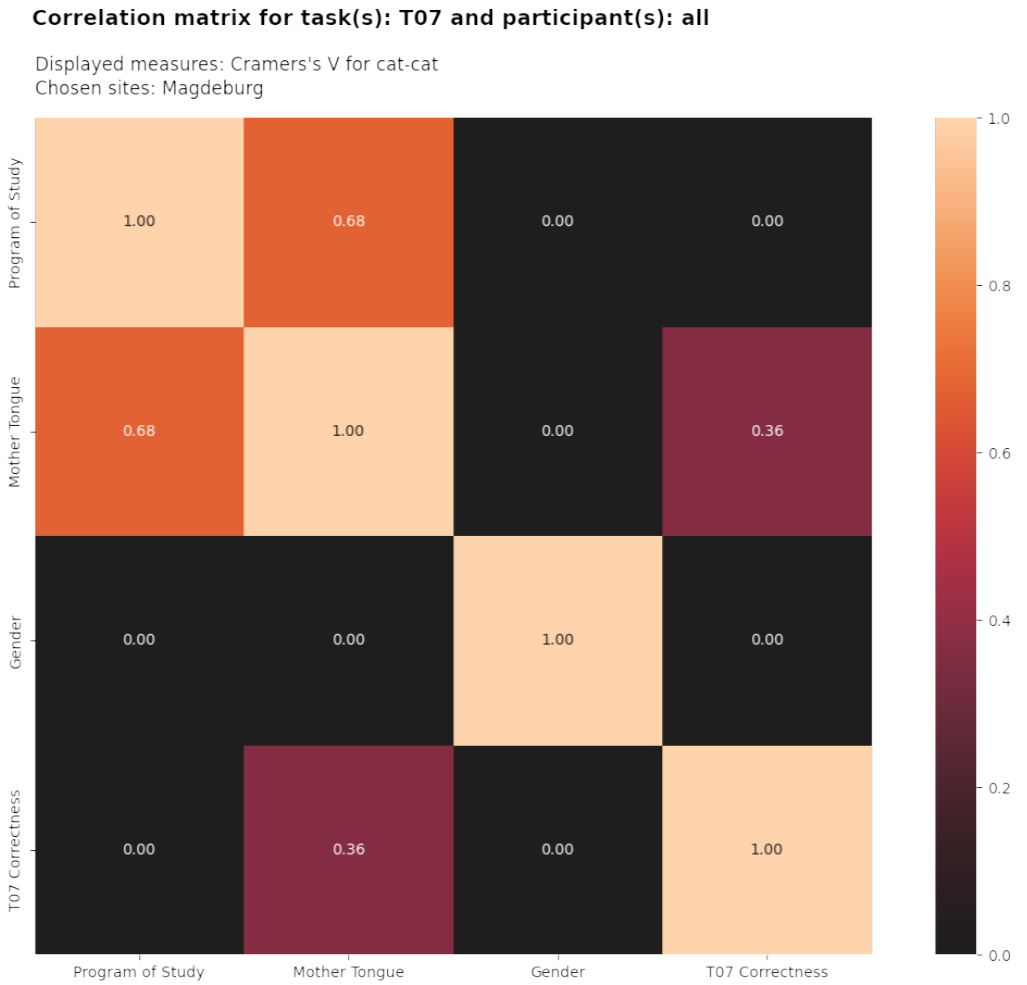

Fig. 15. Categorical attributes only Correlation matrix for displaying the Cramer's V association coefficient (created with Dython Zychlinski (2018a))

Weg zur Datenanalyse (Springer Spektrum)

Hazewinkel, M. 2013, Encyclopaedia of Mathematics (Springer Dordrecht)

Kim, Y., Kim, T. H., & Ergün, T. 2015, The instability of the Pearson correlation coefficient in the presence of coincidental outliers.

Li, S., Sung, B., Lin, Y., & Mitas, O. 2022, Electrodermal Activity Measure: A Methodological Review, Annals of Tourism Research

Lofters, A. K., Shankardass, K., Kirst, M., & Quiñonez, C. 2011, Sociodemographic Data Collection in Healthcare Settings: An Examination of Public Opinions

Myatt, G. J. & Johnson, W. P. 2007, Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining (John Wiley & Sons)

Oluleye, A. 2023, Exploratory Data Analysis with Python Cookbook (Packt Publishing)

Posada-Quintero, H. F. & Chon, K. H. 2020, Innovations in Electrodermal Activity Data Collection and Signal Processing: A Systematic Review

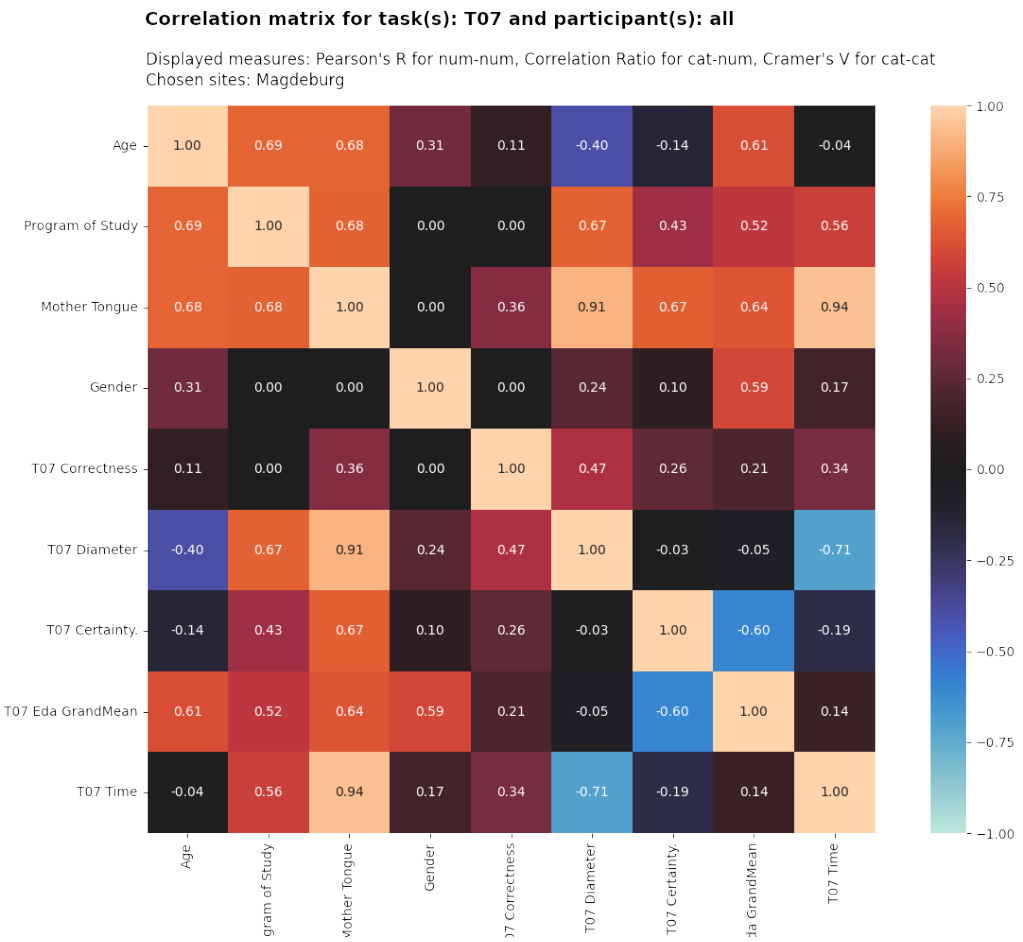

Fig. 16. All attributes Correlation matrix for displaying the Pearson's R correlation coefficient for num-num, Cramer's V association coefficient for cat-cat and the Correlation Ratio for cat-num (created with Dython Zychlinski (2018a))

Rother, A., Notni, G., Hasse, A., et al. 2023, Human Uncertainty in Interaction with a Machine: Establishing a Reference Dataset

Sun, S., Pan, W., & Wang, L. L. 2010, A Comprehensive Review of Effect Size Reporting and Interpreting Practices in Academic Journals in Education and Psychology

University of Newcastle. 2024, Strength of Correlation, https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/strength-of-correlation.html, accessed: 15.01.2024

Zychlinski, S. 2018a, Dython, https://github.com/shakedzy/dython

Zychlinski, S. 2018b, The Search for Categorical Correlation, https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9, accessed: 28.12.2023