Sketch$^+$ for Visual and Correlation-Based Exploratory Data Analysis: A Case Study with COVID-19 Databases

Mirela T. Cazzolato[1,2], Lucas S. Rodrigues[1], Marcela X. Ribeiro[3],
Marco A. Gutierrez[2], Caetano Traina Jr.[1], Agma J. M. Traina[1]

[1] Institute of Mathematics and Computer Sciences
University of São Paulo (ICMC-USP), São Carlos, Brazil
{mirelac, lucas_rodrigues}@usp.br, {agma, caetano}@icmc.usp.br
[2] The Heart Institute (InCor) – Clinical Hospital of Faculty of Medicine
University of São Paulo (HC-FMUSP), São Paulo, Brazil
marco.gutierrez@incor.usp.br
[3] Computer Sciences Department
Federal University of São Carlos (DC-UFSCar), São Carlos, Brazil
marcela@dc.ufscar.br

Abstract. The amount of data daily generated by different sources grows exponentially and brings new challenges to the information technology experts. The recorded data usually include heterogeneous attribute types, such as the traditional date, numerical, textual, and categorical information, as well as complex ones, such as images, videos, and multidimensional data. Simply posing similarity queries over such records can underestimate the semantics and potential usefulness of particular attributes. In this context, the Exploratory Data Analysis (EDA) technology is well-suited to understand data and perform knowledge extraction and visualization of existing patterns. In this paper, we propose *Sketch*$^+$, a technique and a corresponding supporting tool to compare electronic health records (provided by hospitals) by similarity, supporting correlation-based exploratory analysis over attributes of different types and allowing data preprocessing tasks for visualization and knowledge extraction. *Sketch*$^+$ computes partial and overall data correlation considering distance spaces induced by the attributes. It employs both *ANOVA* and association rules with lift correlations to study relationships between variables, allowing extensive data analysis. Among the tools provided, a pixel-oriented one drives the analysts to observe visual correlations among dates, categorical and numerical attributes. As a running case study, we employed three open databases of COVID-19 cases, showing that specialists can benefit from the inference modules of *Sketch*$^+$ to analyze electronic records. The study highlights how *Sketch*$^+$ can be employed to spot strong correlations among tuples and attributes, with statistically significant results. The exploratory analysis has been shown to be an essential complement for similarity search tasks, identifying and evaluating patterns from heterogeneous attributes.

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous

Keywords: CBIR, correlation, COVID-19, exploratory data analysis, visualization

1. INTRODUCTION

Many real-world applications generate data records that include heterogeneous information. For instance, social media posts contain the date and time of publication, numbers of likes and re-posts, images, and textual information. Shopping transactions from stores and supermarkets collect sets of items, prices, dates, and brands. Health institutions acquire Electronic Health Records (EHRs) from their patients and store personal information of patients and staff [Yadav et al. 2018; Jensen et al. 2012]. The stored information goes from simple data, *e.g.*, dates, diagnosis, and exam results, such as blood counts, to complex data, such as electrocardiograms, X-Ray images, Computed Tomography (CT), and long textual observations. Applications may also include data acquired from different users, laboratories, hospitals, and clinics, requiring the applications to handle data interoperability issues [Gansel et al. 2019; Jensen et al. 2012; Lanucara et al. 2018; Bernier and Thorogood 2020]. The growing amount of available data is also a relevant issue, as well as the transactional characteristics of how the enterprises collect data. Query and analysis tasks must provide timely results, aiding specialists and analysts in data understanding and decision making. In this context, solutions should rely on a Database Management System (DBMS) to organize the available data and perform queries timely, supporting analytical and exploratory tools.

This work aims at exploring similarity search and correlation-based Exploratory Data Analysis (EDA) over records with heterogeneous attribute types acquired from diverse sources. We propose a technique to query, evaluate, present, and help understand the data based on correlations among different and heterogeneous attributes to learn meaningful information from data. Similarity queries have been a relevant topic approached by the Database community for decades now [Farias et al. 2019; Samet 2006]. Attribute types can be scalars, such as dates, numbers, and small strings, or complex, such as images and time series. When posing queries over tuples with such attributes, the query engine must employ a specific representation and comparison measure for each data dimension. Distance functions compare attributes according to their types and application requirements [Deza and Deza 2009; Samet 2006]. For instance, we can compute the difference in physical unities from numbers and dates or compare images according to their color distribution similarity.

Beyond comparing tuples of data, preprocessing and extracting meaningful mining patterns in the database assists analysts in comprehending the available information and making more informed decisions [Müller et al. 2021; Brownlee 2020; Hameed and Naumann 2020; Abedjan et al. 2015]. In EDA, correlation measures can uncover and show relationships among attributes of unlabeled data [Hoshen and Wolf 2018]. Analyzing correlation among variables and their underlying interactions is essential for multi-variable datasets [Huang et al. 2019] and for decades has been the subject of studies [Nouri et al. 2021; Kwon et al. 2021; Abdullah et al. 2020; DSouza et al. 2020; Yang et al. 2019; Kaieski et al. 2016]. Here, we aim at taking advantage of correlation approaches and visual tools to support the exploratory analysis and data understanding.

Several problems occur when performing EDA over datasets composed of heterogeneous types of attributes. For example, functions employed to assess the similarity among pairs of tuples can distort the semantic meaning of the information provided by each attribute, for instance, when considering categorical attributes. Correlation analysis can employ different coefficients and metrics depending on the data types and combinations. Focusing on exploring tuples, the analysis must identify the proper methods to employ considering the data characteristics. Moreover, the correlation-based analysis is guided by metrics such as confidence intervals and frequencies [Han et al. 2011]. Finally, the discovered correlations, patterns, and findings are not always intelligible nor trivial for the domain specialists to understand. Visual tools can improve knowledge readability and understanding, considering distinct data characteristics.

Contributions. In the preliminary version of this work [Cazzolato et al. 2021], we introduced the problem of performing EDA and posing similarity queries over records of heterogeneous attributes. Now, we extend the proposed technique, including additional correlation analysis tools, and provide a further experimental evaluation. We also supplement EDA by adding interestingness measures based on correlation and visualization tools. Unlike existing works, we aim at taking advantage of attributes of different types to both enrich the analysis and better support similarity queries. We explore the impact of attributes in projected distance spaces and provide pixel-oriented visualizations, including a temporal analysis of data.

There are two new main contributions from the proposed *Sketch*$^{+}$ (***S**imilarity and **E**xploratory **T**asks with **C**orrelation-based **H**euristics*) method. First, (i) it allows posing similarity-based queries on tuples, considering heterogeneous attributes and correlation-based distance space weighting. For tuples, we present *Sketch-Corr* to compute the correlation between variables, considering the distance spaces induced by all the attributes. Scatter plots show the multidimensional distance space of tuples and heatmaps to show the global correlations found in the data. Categorical attributes describe different values that have specific semantics to the data domain. Generic distance functions may fail to compare and analyze such data adequately. Thus, the second contribution of this work (ii) focuses on improving the semantics of exploratory analysis obtained by categorical attributes. *Sketch*$^+$ discovers association rules (AR) from different categories and analyzes the corresponding lift correlation scores. Sankey diagrams visually show the discovered rules, with transitions between correlated items. The Analysis of Variance (*ANOVA*) combines categorical and numerical values. Box plots visually show the relationship of categories regarding the numerical variable. Finally, a pixel-oriented visualization using scatter plots assists the analyst in visually identifying the correlation between date, categorical and numerical attributes.

A case study with COVID-19 databases. According to the World Health Organization[1], the Coronavirus disease (COVID-19) is an infectious disease that makes most infected people experience mild to moderate respiratory illness. According to the Brazilian Ministry of Health, COVID-19 has infected more than 22 million people in Brazil, with more than six hundred confirmed deaths in the country[2]. With great effort, diverse health and research institutions have collected, organized, and shared public information from COVID-19 patients, aimed at supporting studies in the understanding and analysis of this pandemic disease [FAPESP 2020; ten-Caten et al. 2021].

Considering the relevance of the current pandemic disease and the amount of up-to-date data available, we performed an experimental analysis over three open datasets related to COVID-19. The experimental results show that *Sketch*$^+$ can find significant patterns for all analysis tools employed. We describe scenarios where *Sketch*$^+$ can combine attributes and have different data insights, relying on correlations and visual tools. We also provide *Sketch-GUI*, a prototype that implements all functionalities and visual tools of *Sketch*$^+$. It is open-source and available for download in a public repository to support future research.

Paper outline. The paper is organized as follows. Section 2 describes the background. Section 3 presents the related work. Section 4 details the proposed approach. Section 5 shows the experimental analysis. Finally, Section 6 gives the discussion and conclusions of this work.

2. BACKGROUND

Similarity Search. Data retrieval in Relational Database Management Systems (RDBMSs) compares pairs of objects based on operators of identity ($=$ and $\neq$) and order ($<, \leq, >$ and $\geq$). Similarity-based comparisons require a function to assess the similarity δ among every pair of objects as a real value in $\mathbb{R}^+$. They can compare both scalars (including numbers, dates, and small strings) and complex data (such as images, videos, time series, long text, etc.), provided a suitable δ is defined over the data domain. It is usual to define δ as a feature extractor followed by a distance function. Thus, to pose similarity queries over a database, a feature extractor f_x must be defined to obtain the feature arrays describing the objects (the Identify function can be used when feature extraction is not required) and a distance function f_d to compare the pairs of features. Given two objects s_1, s_2, a distance function f_d and a feature extractor f_x, it is the composition of the feature extractor and the distance function that assess the pair's similarity $\delta(s_1, s_2) = f_d(f_x(s_1), f_x(s_2))$, and we call δ a "descriptor". Several distance functions are suitable for similarity comparisons, such as those from the Minkowski family for numerical data and Levenshtein (L_{Edit}) for textual data [Samet 2006].

The two basic similarity queries are the Similarity Range and k-Nearest Neighbors (kNN). Let $\mathbb{S}$ be

[1] `https://www.who.int/health-topics/coronavirus`, accessed on January 24, 2022.

[2] `https://covid.saude.gov.br/`, accessed on January 24, 2022.

a data domain where descriptor δ is defined, S be a dataset of complex objects $S \subseteq \mathbb{S}$, $s_q \in \mathbb{S}$ be the query center and $s_i \in S$ be elements in S. A Similarity Range Query retrieves every element $s_i \in S$ where the distance to s_q is less or equal than a similarity radius ξ, *i.e.* $\delta(s_q, s_i) \leqslant \xi$. The k-NN Query retrieves the k objects $s_i \in S$ that are most similar to s_q, measured by a given descriptor δ.

Correlation Heuristics. Correlation coefficients evaluate the existing association between variables in a dataset [DSouza et al. 2020], and enable employing visualization tools such as pixel-oriented scatter plots, boxplots, and heatmaps [Han et al. 2011]. Examples of well-known correlation coefficients employed in the literature are Pearson, Spearman, and *ANOVA*. *ANOVA* (**AN**alysis **O**f **VA**riance) analyzes two or more populations described by a numeric variable and at least one categorical variable [Han et al. 2011]. *ANOVA* test can show significant differences between numerical values and two or more categorical groups. *ANOVA* returns two values, *F-test* and *p-value*. F-test is a correlation score measuring how much the actual means of the groups deviate from the primary assumption, which is that the means of all groups are equal. The higher the F-test score, the larger the difference among the means. As a complement, the *p-values* inform the statistical significance of the score.

Association Rules (AR) look for itemsets that co-occur in a transactional database D. Let I be the itemset of every item in D. AR are implications in the form of $A \Rightarrow B$, where $A \subset I$ and $B \subset I$ are non-empty itemsets, and $A \cap B = \emptyset$. The *support* of a rule is given by the proportion of D that contains $A \cup B$. *Confidence* is the proportion of transactions in D containing A that also contain B:

$$sup(A \Rightarrow B) = P(A \cup B) \qquad conf(A \Rightarrow B) = P(B|A) = sup(A \cup B)/sup(A)$$

The support-confidence evaluation of AR can be supplemented with the lift correlation measure. The occurrence of A is said to be independent of B if $P(A \cup B) = P(A)P(B)$. Otherwise, both itemsets are dependent and correspond to correlated events. The *lift* correlation is evaluated as $lift = conf(A \Rightarrow B)/sup(B)$. It assesses the degree to which the occurrence of one item "*lifts*" the occurrence of the other [Han et al. 2011]. If $lift = 1$, A and B are independent and there is no correlation between them; $lift < 1$ indicates a negative correlation between the itemsets, where the presence of one fosters the absence of the other; and $lift > 1$ indicates a positive correlation between A and B, where the occurrence of one fosters the occurrence of the other. As AR work with transactions, EHRs must be converted before the pattern discovery step. Therefore, every combination of $\{$*categorical_ attribute*, *value*$\}$ is converted to an attribute, which can be present or absent in a tuple.

Data Visualization. Raw data may not explicitly represent semantic domain information. Visual representations can reorganize e represent data characteristics and patterns in a way that amplifies human cognition [Han et al. 2011]. The literature reports a plethora of methods for data visualization, such as in the survey [Wu et al. 2021]. The scatter, bar and line plots, heatmaps, contour, and boxplots are among the most popular information visualization tools. In this work, we employ heatmaps, boxplots, Sankey diagrams, and scatter plots as visualization tools (see Figure 2). We use each tool according to the data type represented by our analysis methods. We detail such representations in Section 4.

3. RELATED WORK

This section presents related work for EDA and visualization tools. Table I summarizes existing methods based on relevant aspects related to our proposal, as follows:

—**EDA Tool**: if the work proposes a practical tool or prototype for EDA

—**Data Cleaning Processing**: whether the solution provides preprocessing tools for EDA tasks

—**Similarity Retrieval**: if the solution performs information retrieval, in particular, similarity queries based on similarity

—**Correlation Analysis**: if the work performs any kind of correlation analysis over the data

—**Visualization and Analysis**: if the work provides visual and analysis tools for EDA

Table I. Overview of recent works for EDA and visualization tasks and the proposed approach.

Work	Year	EDA Tool	Data Cleaning	Similarity Retrieval	Correlation Analysis	Visual Tools
Vis-Health [Kaieski et al. 2016]	2016	✓	✗	✗	✓	✓
SubVIS [Hund et al. 2016]	2016	✓	✓	✗	✗	✓
[Huang et al. 2019]	2019	✗	✗	✗	✓	✓
[DSouza et al. 2020]	2020	✗	✗	✗	✓	✓
[Guo et al. 2020]	2020	✓	✓	✓	✗	✓
VALENCIA [Abdullah et al. 2020]	2020	✓	✗	✗	✗	✓
DPVis [Kwon et al. 2021]	2021	✓	✗	✗	✓	✓
VISUMURE [Nouri et al. 2021]	2021	✓	✓	✗	✓	✓
Sketch$^+$	2022	✓	✓	✓	✓	✓

Vis-Health [Kaieski et al. 2016] is a visual tool analyzing Dengue incidence in Brazil. It combines public health information with climatic factors in the corresponding regions and performs correlation analysis to map strong relationships and significant attributes related to dengue, such as rainfall, temperature, and occurrence counts. The results indicate the same pattern in 6 out of 7 state capitals, relying on maps and pie charts to understand the patterns. In [Hund et al. 2016], the authors proposed SubVIS, a visual analytics tool to analyze high-dimensional data. The tool deletes incomplete records with missing dimensions. Also, SubVIS applies a subspace clustering strategy to analyze high-dimensional data and employs visual techniques over subspaces, such as MDS projection, heatmaps, and dispersion plots. However, the work does not support similarity retrieval or correlation analysis for EDA, focusing on cluster analysis.

In [Huang et al. 2019], the authors introduce a recommendation model based on entropy and decision trees enhanced with correlation analysis. Chord and Sankey diagrams support the discussion of strong relationships between pharmaceutical compositions and their helpfulness in recommending frequent rules. The data integration step performs data dimensionality reduction, which can discard proper knowledge or introduce bias in the model. In [DSouza et al. 2020], the authors propose EDA with visual tools to discover patterns over a COVID-19 dataset from Italy. They analyze variable relationships, crossing cases with statistical indicators from each country region. Visual tools aided recognizing potential tendencies or insights based on the most relevant attributes. The work is an empirical study that does not provide correlation measures nor similarity retrieval mechanisms.

In [Guo et al. 2020], the authors propose an interactive visual system for retrieving and exploiting tendencies over similar and high-dimensional medical records. It provides a preprocessing step based on mean imputation to deal with incompleteness. The system supports similarity retrieval based of data records using with the Dynamic Time Warping (DTW) function. Finally, the system has visual tools for dimensionality reduction, clustering, and similarity queries, assisting in spotting medical alterations, and identifying deceased groups or even differences among similar patients. VALENCIA [Abdullah et al. 2020] is an EDA and visual system for high-dimensional electronic health, clustering data using k-means or hierarchical clustering algorithms. It also provides dimensionality reduction approaches, including PCA, MCA, MFA, and t-SNE. The tool takes advantage of visualization techniques through interactive tools, such as scatter and Sankey plots, heatmaps, and line and bar graphs. However, VALENCIA does not provide data cleaning mechanisms, EDA tasks based on correlation analysis, nor similarity retrieval methods.

DPVis [Kwon et al. 2021] is a web application based on visual tools and EDA tasks employing Hidden Markov Models (HMMs) for disease progression patterns over longitudinal health records. The work supports seven visual tools, such as feature matrix, summarizing patient states and the variable relationships, and pathway waterfall graphs of transitions between patient patterns. However, the application does not provide data cleaning tools or even tools for similarity retrieval. More recently, the authors of [Nouri et al. 2021] proposed VISUMURE, an EDA and visual tool to discover enhanced insights using descriptive statistics into multi-morbidity over EHRs. The cleaning step standardizes categorical attributes and merges categories with a few samples. The EDA employs a dynamic correlation matrix based on logistic regression and decision tree models to estimate the values between the attributes. Visual tools support EDA, such as bar charts for the analytics models and heatmaps for correlation analysis. However, VISUMURE does not support similarity retrieval or equivalent methods.

Table I shows that only a few related studies cover data preprocessing. However, there are many tools focused on data preprocessing providing additional tools, such as data validation, enrichment, and filtering, which encompass many features [Abedjan et al. 2015; Hameed and Naumann 2020]. The preparation steps can be customized according to data requirements, scenarios, and domains.

In this work, we propose a method for correlation-based EDA supported by visual tools and similarity retrieval mechanisms. Our proposal covers several analysis tasks and data scenarios. In particular, we intend to provide multiple tools to exploit and understand distinguished data types with distinct semantics using correlation and visual heuristics, also allowing information retrieval. We present our proposal next.

4. *SKETCH*$^+$: THE PROPOSED METHOD

Sketch$^+$ (***S****imilarity and* ***E****xploratory* ***T****asks with* ***C****orrelation-based* ***H****euristics*) creates a multifunctional environment, providing a holistic approach capable of evaluating the correlation among the attributes that composes a record, regardless of whether they are discrete, continuous values, representing categorical, numerical, date, or complex attributes. The method also provides correlation tools and recurrent patterns to explore individual samples and multiresolution clusters of samples, including similarity-based comparisons.

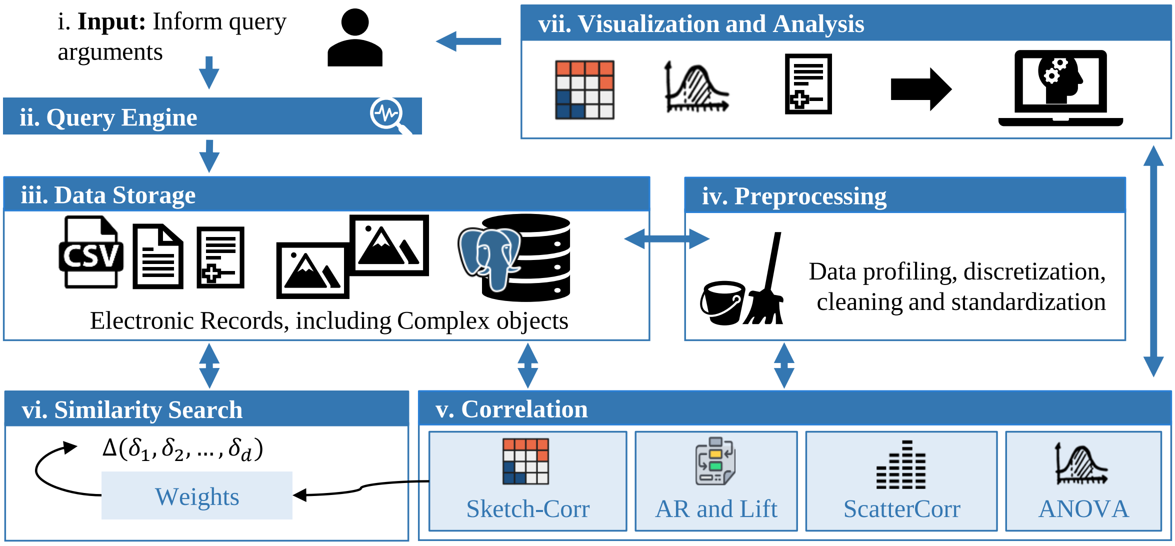

i. Input: Inform query arguments
vii. Visualization and Analysis
ii. Query Engine
iii. Data Storage
CSV
Electronic Records, including Complex objects
iv. Preprocessing
Data profiling, discretization, cleaning and standardization
vi. Similarity Search
$\Delta(\delta_1, \delta_2, \ldots, \delta_d)$
Weights
v. Correlation
Sketch-Corr
AR and Lift
ScatterCorr
ANOVA

Fig. 1. The *Sketch*$^+$ method.

Figure 1 illustrates *Sketch*$^+$ and its seven main steps. In (i), the specialist collects and provides the parameters of the desired analysis to the query engine. By "*sketch*" we imply that the user can choose to have only partial query information for data exploring, such as specific attributes or data filtered by values. We show a few examples further in Section 5.3.1, when we select specific

COVID-19 biomarkers to analyze. They select the relevant attributes and analyze the correlation scores of a subset of attributes according to their types. The query engine (ii) receives the query arguments (*i.e.* selected attributes and types) and verifies the available exploratory analysis tools for the corresponding data types. In each analysis step (v), *Sketch*$^+$ guides the user by showing only the attributes whose types are adequate for the correlation tool being used. *Sketch*$^+$ can be plugged-in into data sources to load files or tables from a RDBMS (iii). A preprocessing (iv) step allows users to perform data profiling, discretization, cleaning, and standardization. The correlation module has four approaches (v) for exploratory data analysis, appropriate for specific combinations of data types. For this, *Sketch*$^+$ implements the *Sketch-Corr* heuristic, which computes the correlation between every pair of attributes (of any type). The method employs such correlations in the similarity search module (vi), which computes the distance of every attribute, allowing the analyst to weigh the variables according to their correlations and observe the impact in the distance space. The analyst can continue to explore the data using AR with lift and *ANOVA* correlations from the similarity results. The *ScatterCorr* tool shows how different values of categorical, numerical, and date attributes correlate with each other, given an initial tuple order. Finally, in (vii) the user can evaluate the results using the available visualization metaphors appropriate for every employed correlation tool and the specific attribute types. Figure 2 shows the available visual tools for (a) *Sketch-Corr*, (b) *ANOVA*, (c) AR with lift, and (d) *ScatterCorr*. We explain each step of *Sketch*$^+$ and the correlation approaches in more detail in the following subsections.

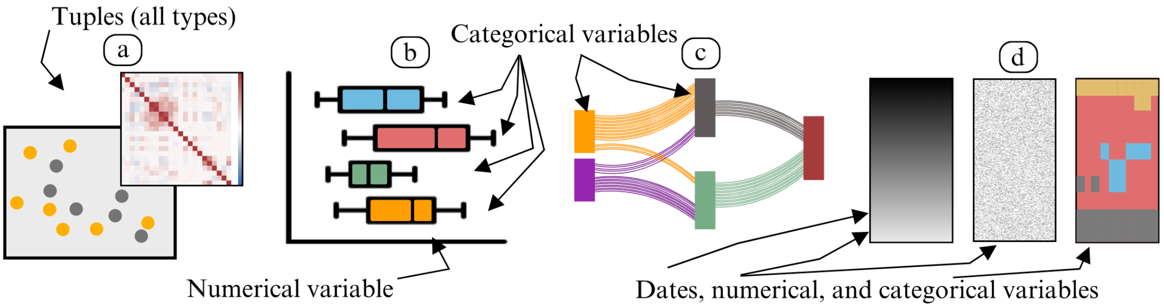


Fig. 2. Visual tools provided by *Sketch*$^+$: (a) heatmaps and scatter plots for *Sketch-Corr*; (b) boxplots for ANOVA; (c) Sankey diagrams for AR and lift; and (d) parallel scatter plots for *ScatterCorr*.

4.1 Data Storage, Query and Preprocessing

Figure 3(a) shows the steps for storing and preprocessing data, including the support for bulk load in the RDBMS, data schema standardization, filtering, and preprocessing. The user can opt to load the input data by plugging *Sketch*$^+$ into the PostgreSQL RDBMS and loading an existing table, *or* by loading two CSV files, one with the data and one with the respective attribute types. When the user opts to use PostgreSQL, *Sketch*$^+$ allows users to access schemes, tables, and query data promptly, taking advantage of the robustness of a RDBMS. With the RDBMS, *Sketch*$^+$ automatically retrieves the attribute types, which are required to properly set up the (v) correlation tools and (vi) similarity functions to be used. *Sketch*$^+$ also supports complex data, such as images and texts.

The preprocessing step is optional. The user loads the input data through the graphical interface of *Sketch*$^+$. One can see the statistics provided by the *Sketch*$^+$ to choose the attribute to be preprocessed according to the selected approach. *Sketch*$^+$ processes the data, shows the results, and updates initial statistics. Finally, the user can choose to save, or use the modified data for analysis, or discard it. EDA usually requires preprocessing input data before knowledge extraction, thus there are data cleaning methods useful for data standardization and normalization. Among them, there is a data profiling method, allowing users to detect problems using statistics and data quality measures. These measures

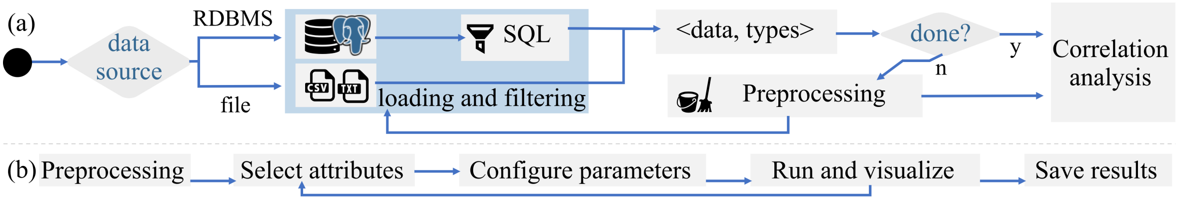


Fig. 3. Pipelines for: (a) Data loading, filtering and preprocessing; (b) Data analysis.

include missingness, uniqueness, and entropy of attributes, which can be helpful in understanding the data. We provide attribute discretization to divide continuous values into a given amount of groups (or bins). Finally, the preprocessing step provides a term standardization method based on a combined strategy of key collision clustering and fuzzy similarity string matching. In practice, this process receives an attribute with heterogeneous or misspelled but similar terms and creates clusters by composing a vocabulary of terms using the cluster centroids. Then, this vocabulary is used in the fuzzy string matching method based on n-gram overlaps and Levenshtein or cosine distances, mapping the most similar terms to replace misspelled ones using unsupervised heuristics. After data preprocessing, the next step is data analysis. As Figure 3(b) shows, the general analysis steps are very straightforward, and we explain the specifics in the next subsections.

4.2 *Sketch-Corr* for Tuples and Similarity Search

The similarity search module aims at providing tools for visualizing the data distribution of tuples and the correlation among attributes of different types. Let D be a dataset with d attributes and n tuples. *Sketch*$^+$ employs predefined distance functions according to every attribute type. Equation 1 gives the *Sketch-Dist*, the weighted *global* distance function Δ to compare a pair of tuples $< t_i, t_j >$, such that $1 \leq i, j \leq n$:

$$\Delta_{i,j} = \frac{1}{w} \sqrt[p]{\sum_{a=1}^{d} |\delta(t_i^a, t_j^a)|^p \times w_a}, \tag{1}$$

where w is the sum of attribute weights, such that $w = \sum_{a=1}^{d} w_a$. The global function Δ works as a Minkowski distance of order p. For every attribute a, such that $1 \leq a \leq d$, the *local* function δ gives the distance between the pair of tuples, given every specific attribute type. The function weights are initialized as $w_a = 1$. Alternatively, the global function can weigh every attribute according to its corresponding relevance in the comparison, represented by w_a. The correlation score gives the relevance value among attributes, which we explain in the sequence.

While correlation coefficients, such as Spearman and Kendall, rely on global references, we aim at determining the correlation locally, that is, in every region of the data space. Also, although such coefficients can work with more than one attribute, the attributes need to be concatenated while preserving the lexicographical order. To overcome these limitations, we propose *Sketch-Corr*, which builds over those monotonic correlation coefficients to handle correlations between the variation of distances among attributes.

Algorithm 1 details *Sketch-Corr*. It receives as input: S (a sample of m tuples from D), with d attributes; the set of distance functions F, that will compare every type of attribute among $Tp = \{numeric, textual, categorical, date, and complex\}$; and the correlation coefficient Φ, such as Pearson and Spearman. In Line 1 *Sketch-Corr* initializes its variables. For every attribute a and every tuple t_i in S (Lines 3 and 4), *Sketch-Corr* computes the array of distances $D_a[i]$ of the tuple t_i to all other tuples in the dataset (Line 5). As a result, D_a has the mean distance variation of the tuples concerning attribute a. For every pair of attributes $< a_r, a_s >$ in S (Line 6), such that $1 \leq r, s \leq d$, the algorithm computes the correlation between the arrays of distance variations D_r and D_s (Line 7). *Sketch-Corr* returns the correlation matrix M of dimensions $d \times d$ (Line 8). Every row q $(1 \leq q \leq d)$ of M has the correlation scores of attribute a_q to all other attributes. The values correspond to the weights w of Δ (see Eq. 1), with the reference attribute a_q.

Algorithm 1: *Sketch-Corr* to compute the correlation between attributes

```
Data: $S$: a sample dataset with $d$ attributes and $m$ tuples
      $F$: a set of distance functions
      $\Phi$: the correlation coefficient
Result: $M$: Matrix of correlations of dimensions $d \times d$
1 begin
2     Initialization
3     foreach attribute $a$ in $S$ do
4         for $i$ from 1 to $m$ do
              /* Mean distance of $t_i$ to all the other tuples */
5             $D_a[i] \leftarrow \frac{1}{m} \sum_{j=1}^{m} \delta_a(t_i, t_j)$   // where $\delta_a \in F$
6     foreach pair of attributes $< a_r, a_s >$ in $S$ do // where $1 \leq r, s \leq d$
          /* Compute the correlation between the attributes */
7         $M[r][s] \leftarrow \Phi\ (D_r,\ D_s)$
8     return $M$
```

Sketch$^+$ employs two visualization tools to evaluate tuples: a heatmap and a scatter plot. The heatmap represents the correlation matrix M with a color pattern, where more saturated colors represent strong correlations (positive or negative) between attributes, and colors close to white represent weak correlations. The scatter plot shows the distribution of tuple distances. *Sketch*$^+$ employs the Manifold Multidimensional Scaling (MDS) method to display the distribution of objects (in this case, tuples) in a two-dimensional space, using the distances between them – see the example in Figure 2). We provide the original and correlation-weighted spaces when posing queries, showing the different results for both options. Accordingly, the user can select an attribute and visualize its impact on the data distribution. For instance, this tool can be employed to check which attribute best adjusts the data in space to best separate classes of data.

Notice that *Sketch-Corr* gives the *overall* correlation among the attributes. Some distance functions can underestimate the semantics of specific attribute values, such as categorical ones, and those relationships with other attributes. The overall data analysis informs the analyst of potential patterns that should be further investigated. However, existing correlation heuristics can take advantage of specific attribute types (and values) to provide meaningful data semantics, as we show next.

4.3 Association Rules (AR) and Lift Correlation for Categorical Attributes

Sketch$^+$ takes advantage of AR to find categories of attributes that frequently co-occur in the database, given the minimum support and confidence values. The lift metric gives the correlation between the antecedent and consequent items (in our case, category values) of discovered AR. Importantly, *Sketch*$^+$ supports users while selecting the relevant attributes and filtering tuples, given a criterion. For many tuples, *Sketch*$^+$ can preprocess the dataset and store the discovered AR for further analyses. In this step, support and confidence give the strength of rules, and the corresponding lift value gives the correlation score of every discovered rule.

The analyst informs the set of categorical attributes to be used as a reference. Numerical variables can be used after discretizing the value into intervals in the preprocessing step (Figure 1(iv)). *Sketch*$^+$ concatenates the attribute name with every possible value, and each combination becomes a distinct item in the transaction set. *Sketch*$^+$ searches for all patterns that include the given items and shows the corresponding rules with the support, confidence, and lift correlation. *Sketch*$^+$ employs Sankey diagrams as tools to visualize the AR returned by the algorithm – see Figure 2(c). Every item corresponds to a vertical bar in the diagram, and linkages represent items that co-occur in the discovered AR. The thickness of line linkages corresponds to the confidence of the rule, and the diagram shows only those that comply with the minimum support and confidence.

4.4 *ScatterCorr* for Numerical, Categorical and Date Attributes

ScatterCorr is a pixel-oriented technique for observing the visual correlation among several variables simultaneously. The user selects an attribute to sort the dataset. The tool sorts all data and displays the tuples using dates, numerical, and categorical attributes. Figure 2(d) illustrates *ScatterCorr*: the first bar has the sorted values (depicted by a grayscale scatter plot). The two other attributes are depicted according to predefined color criteria, visually showing how the attributes correlate with each other. For instance, let the three scatter plots correspond to attributes *income*, *height*, and *role* from the employees of a company. The first plot shows the *income* attribute. There is no visual correlation between the incoming salary and the *height* of the person, as we see a random pattern in the second plot. However, we observe that different roles (represented by the gray, red, cyan, and yellow regions) are visually correlated to the *income*. Thus, a reasonable hypothesis is that the role represented by the gray color receives lower salaries than the role represented by the yellow color.

4.5 *ANOVA* for the Combination of Numerical and Categorical Attributes

Finally, *Sketch*$^+$ employs *ANOVA* to analyze pairs of attributes, showing existing differences between categorical and numerical variables. *ANOVA* allows specialists to test and check which variables are important to take into account when analyzing a specific numerical measurement. Figure 2(b) employ boxplots to visually show the differences between different groups of the same variable. While *ANOVA* scores allow the analyst to check the correlation among groups, boxplots highlight how the groups overlap each other, considering the range of values of the selected numerical variable.

4.6 *Sketch-GUI*: A Visual Tool for Information Retrieval

We provide *Sketch-GUI*, an application with a visual interface that implements all correlation heuristics and visualization tools of *Sketch*$^+$. The prototype controls the communication among the modules of *Sketch*$^+$, allowing users to query the tuples by similarity and perform an exploratory analysis over the available data using *Sketch-Corr*, AR-Lift, *ANOVA*, and *ScatterCorr* with corresponding visual tools. *Sketch-GUI* is available in a Git repository (ref. to Section 5), and currently supports numerical, categorical, text, date, and complex attributes. Next, we present an experimental evaluation of *Sketch*$^+$ with real-world datasets.

5. EXPERIMENTS

This section describes the datasets employed in the experiments, the implementation details, the validation of the proposed correlation tools by example, and the performance analysis.

5.1 Dataset Description

We evaluate *Sketch*$^+$ with three public datasets, summarized in Table II. *Ds-FAPESP* (FAPESP COVID-19 DataSharingBR) collects and integrates data related to COVID-19 exams from diverse sources [FAPESP 2020]). *Ds-Vaccine* combines vaccination records of the Brazilian population, which we combined with a dataset of daily notifications of suspicious and confirmed and death cases from Brazilian cities and states [Min. Saúde 2022; Gonçalves et al. 2021]. Finally, *Ds-CTMD* has records of patients that either are healthy, have COVID-19, or have Community-Acquired Pneumonia (CAP) [Afshar et al. 2021]. *Ds-CTMD* also includes a complex attribute (`CT-Scan image`).

Table II. Datasets employed in the experiments.

Dataset	Table	# Tuples	Types and attributes
Ds-FAPESP [FAPESP 2020]	Patients	862,571	**ID**: id_patient (PK), id_hospital; **NUMERIC**: aa_birth, cd_zipcode; **CATEGORY**: de_sex, cd_city, cd_state, cd_country.
	Exams	54,763,675	**ID**: id_exam (PK), id_patient, id_attendance, id_hospital; **DATE**: dt_collect; **CATEGORY**: de_source, de_exam, de_analyte, cd_unity, de_reference_value; **TEXT**: de_result.
	Outcomes	307,928	**ID**: id_patient (PK), id_attendance (PK), id_clinic, id_hospital; **DATE**: dt_attendance, dt_outcomes; **CATEGORY**: de_attendanceType, de_outcomes.
Ds-Vaccine [Min. Saúde 2022] [Gonçalves et al. 2021]	Vaccination	196,959,275	**ID**: id_document (PK), id_patient; **DATE**: birth, vaccination; **INTEGER**: age, cityCode, countryCode, codeRace; **TEXT**: sex, descriptionRace, zipCode, city, state, country, nationality, cnesLocal, corporateNameLocal, fantasyNameLocal, groupCode, groupName, categoryCode, categoryName, producerName, producerRef, vacBatch, vacCode, vacName, sourceData.
	Occupation	1,249,167	**ID**: id (PK); **DATE**: notification; **TEXTUAL**: cnes, de_source, State, City; **INTEGER**: suspectCaseCLI, suspectUTI, confirmedCLI, confirmedUTI, suspectDeath, suspectRelease, confirmedDeath, confirmedRealase.

Ds-CTMD
[Afshar et al. 2021]
	Patients	305	**ID**: id_patient (PK); **INTEGER**: age;
			REAL: weight; **TEXT**: gender, clinical_symptoms, surgery, follow-up, pcr, diagnosis;
			CATEGORY: radiologist_1, radiologist_2, radiologist_3;
			COMPLEX: CT-Scan image.

Data collected in January 18, 2022

5.2 Implementation Details

Sketch-GUI was implemented in Python, using well-known open libraries such as *Matplotlib*, *Plotly*, *Seaborn*, *Pandas*, *ScikitLearn*, *Mlxtend*, and *OpenClean*. The user interface is implemented using Python *Tkinter*, and the database employed was PostgreSQL 13.3. Scripts for data preprocessing and loading in the PostgreSQL DBMS are available in a Git repository[3], together with the image features extracted from *Ds-CTMD*, discovered patterns, and a demonstration video of how to use the *Sketch-GUI* prototype. *Sketch-Corr* uses the Levenshtein (LEdit) distance function for categorical and textual attributes and Euclidean for the remaining ones.

5.3 Evaluation of Correlation Tools

Here we present examples of analysis over the real-world datasets using the correlation tools of *Sketch*$^+$.

5.3.1 *ANOVA with Boxplot visualizations.* In this analysis, we aimed at identifying biomarkers that presented significant changes in exams taken from female and male patients, with and without COVID-19. We started this task by selecting biomarkers (analytes) related to COVID-19 reported in [ten-Caten et al. 2021]. To select this sample, we executed the SQL Statements 1 and 2 over the original database *Ds-FAPESP*.

We focus on analyzing three filtered biomarkers: Lymphocytes, Ferritin, and Fibrinogen. Figure 4 presents the corresponding *ANOVA* results. Overall, the higher F-Score values and significant *p*values indicate a strong correlation between the categorical variable ("COVID-19" and "not COVID-19") and the numerical variable (in this case, the result of the selected analyte). The highest differences between the groups with and without COVID-19 appear within male patients, as shown in the charts related to analytes Lymphocytes and Fibrinogen. We also observe that while the Lymphocytes count decreases among patients with COVID-19, Fibrinogen presented higher results for this same group compared to patients who tested negative for COVID-19.

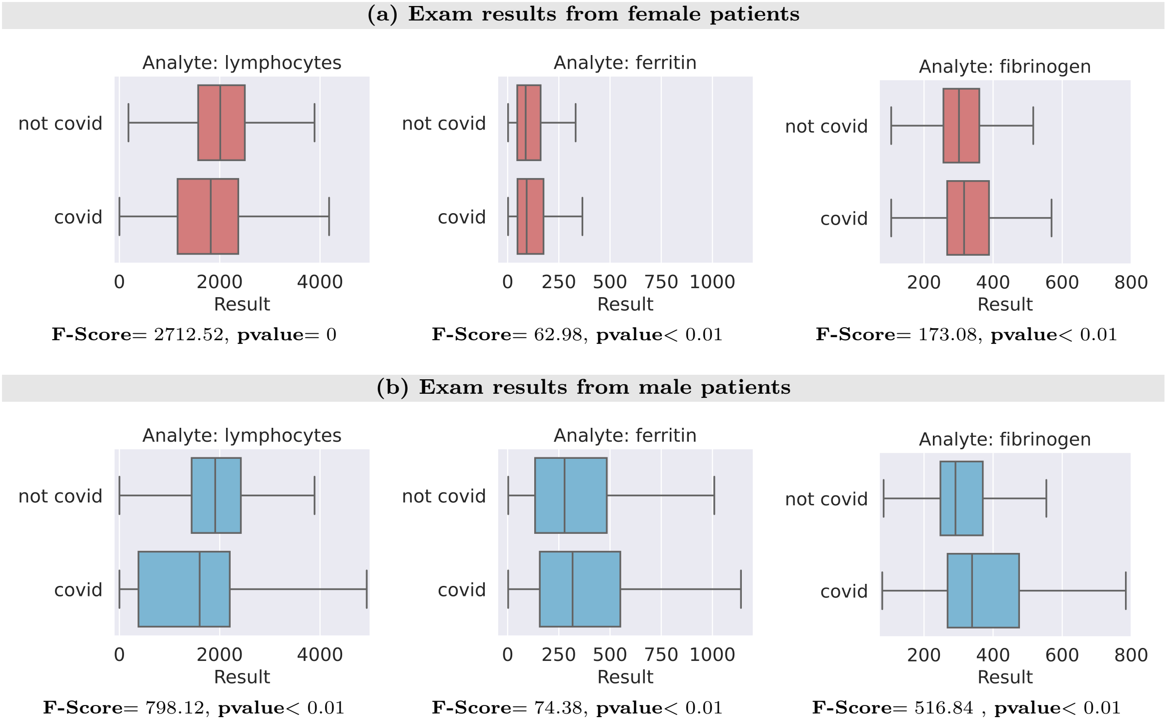
[3] Git repository of *Sketch*$^+$: *https://github.com/mtcazzolato/sketch*.

```
CREATE TABLE patientsCovid AS ( SELECT id_patient,
     ic_sex, aa_birth, de_result FROM exams
  WHERE de_analyte LIKE '%covid 19, antibodies igm%' OR
        de_analyte LIKE '%pcr%'              OR
        de_result  LIKE '%positive'          OR
        de_result  LIKE '%detect'            OR
        de_result  LIKE '%present'           OR
        de_result  LIKE '%reagent'           OR
        de_result  LIKE '%reagent sample'    OR
        de_result  LIKE '%negative'          OR
        de_result  LIKE '%not reagent'       OR
        de_result  LIKE '%not detect'
  GROUP BY id_patient, de_result);
```

Statement 1. Filtering patients with exams related to COVID-19.

```
SELECT * FROM exams ex
JOIN patientsCovid ptC ON ex.id_patient = ptC.id_patient
WHERE ex.de_analyte LIKE '%eosinophils'          OR
  ex.de_analyte     LIKE '%basophils'            OR
  ex.de_analyte     LIKE '%indirect bilirubin'   OR
  ex.de_analyte     LIKE '%alt'                  OR
  ex.de_analyte     LIKE '%gama-gt'              OR
  ex.de_analyte     LIKE '%c-reative protein'    OR
  ex.de_analyte     LIKE '%erythrocytes'         OR
  ex.de_analyte     LIKE '%ferritin'             OR
  ex.de_analyte     LIKE '%neutrophils'          OR
  ex.de_analyte     LIKE '%lymphocytes'          OR
  ex.de_analyte     LIKE '%fibronogen';
```

Statement 2. Selecting exams over selected biomarkers from the set of selected patients.



F-Score= 2712.52, **pvalue**= 0 · **F-Score**= 62.98, **pvalue**< 0.01 · **F-Score**= 173.08, **pvalue**< 0.01

F-Score= 798.12, **pvalue**< 0.01 · **F-Score**= 74.38, **pvalue**< 0.01 · **F-Score**= 516.84 , **pvalue**< 0.01

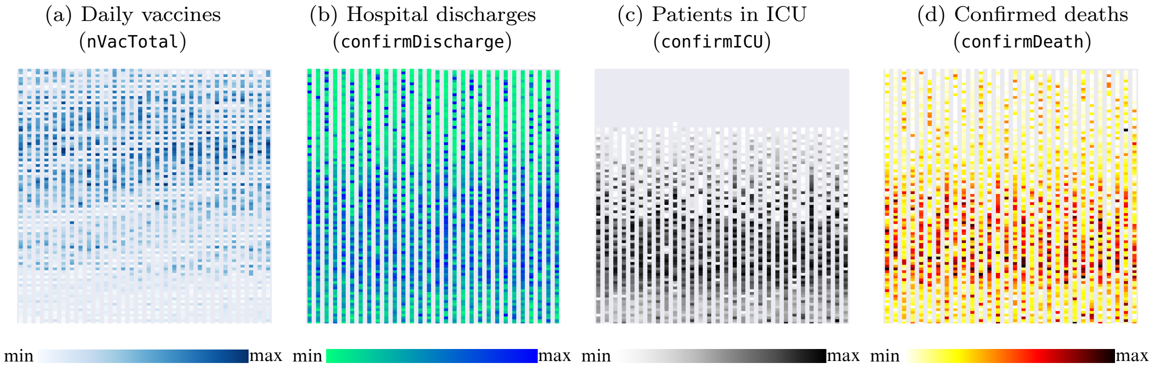
Fig. 4. *ANOVA* results for three analytes related to COVID-19, divided into (a) female and (b) male samples of patients.

5.3.2 *ScatterCorr with pixel-oriented visualizations.* In this experiment, we aim at evaluating different metrics regarding the COVID-19 vaccination of Brazilian residents and the occupation of hospitals in Brazil in the same period. We crossed the information of two datasets to compose *Ds-Vaccine* by executing Statement 3, and summarized the statistics per day.

Figure 5 shows the visual correlation among attributes of the sample data, generated by *ScatterCorr*. We normalized the attributes per state (`stateVac`), and sorted all records by date (`dateVac`) using *Sketch-GUI*. Thus, in this case, the y axis correspond to the date, and the values are plotted linearly over the x axis, bottom-up. The available data goes from 01-01-2021 to 11-23-2021. We observe in (a) that the number of vaccines applied per day increased over time, probably due to the number of available vaccines and the application of the second dose concomitantly with the first dose. Plots (b, c, d) show a visual correlation between the number of hospital discharges, patients in ICU (Intensive Care Unit), and confirmed deaths. With this pixel-oriented visualization, we can observe regions of the plot (mid-low) with the highest concentrations of cases. The top region of plot (c) is empty due to missing data, since *Ds-Vaccine* contains `NULL` values for `confirmICU` after 09-13-2021.

```
SELECT vac.stateVac, vac.dateVAC, vac.nVacMasc, vac.nVacFem, vac.nVacTotal, ocp.suspectCLI, ocp.suspectICU,
        ocp.suspectDeath, ocp.confirmCLI, ocp.confirmICU, ocp.confirmDeath, ocp.confirmDischarge
FROM (SELECT stateVaccination AS stateVac, dateVaccination AS dateVAC,
         COUNT(CASE WHEN patientSex ='M' THEN 1 END) nVacMasc,
         COUNT(CASE WHEN patientSex ='F' THEN 1 END) nVacFem,
         COUNT(CASE WHEN patientSex ='F' OR patientSex ='M' THEN 1 END) nVacTotal
   FROM vaccination
   GROUP BY dateVaccination, stateVaccination) VAC
JOIN (SELECT stateOCP, dateOCP, SUM(suspectCLI), SUM(suspectICU), SUM(suspectDeath), SUM(confirmCLI),
        SUM(confirmICU), SUM(confirmDeath), SUM(confirmDischarge)
   FROM occupation
   GROUP BY stateNotification, dateNotification) OCP
ON vac.stateVac = ocp.stateOCP AND  vac.dateVAC = ocp.dateOCP
```

Statement 3. Crossing vaccination and occupation data and summarizing daily statistics related to COVID-19, such as vaccine indices by sex, number of suspicious and confirmed cases in the clinic or ICU, deaths and medical releases.


(a) Daily vaccines (`nVacTotal`)

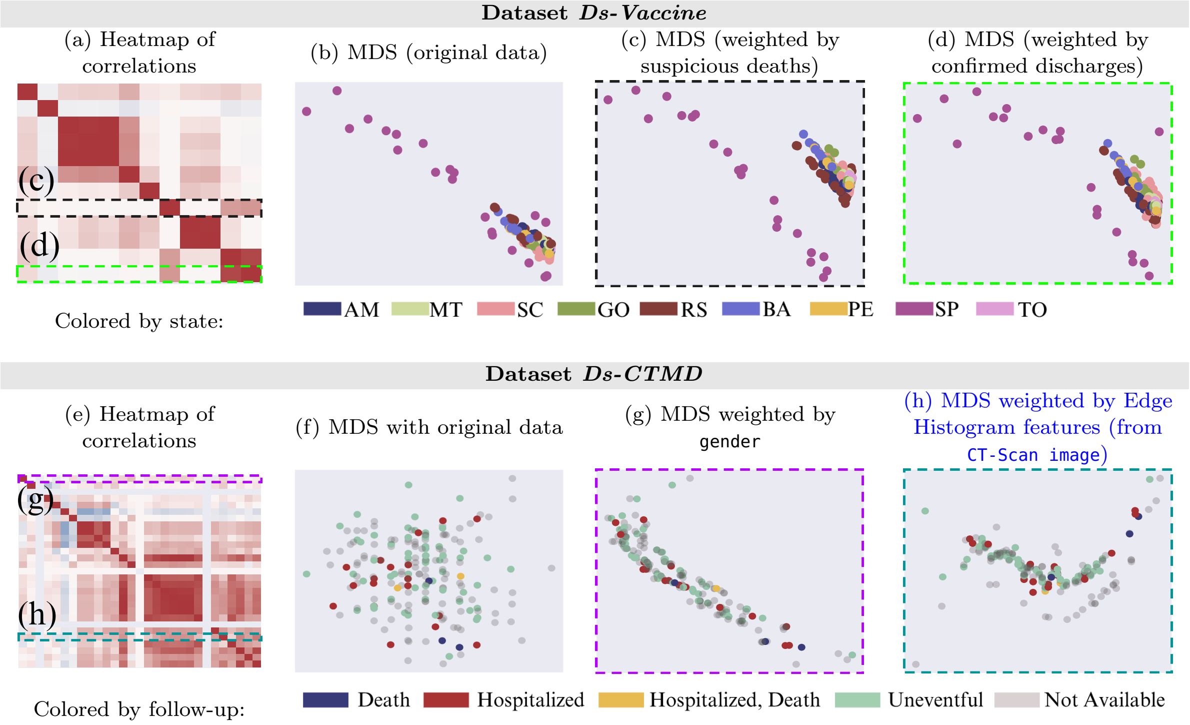
(b) Hospital discharges (`confirmDischarge`)

(c) Patients in ICU (`confirmICU`)

(d) Confirmed deaths (`confirmDeath`)

Fig. 5. Pixel-oriented visual correlation with *ScatterCorr*: the plots show *Ds-Vaccine* sorted by date (y axis) and plotted bottom-up. We observe a pattern suggesting an increasing amount of vaccines applied per day (a), and visual correlations among hospital discharges (b), cases in ICU (c), and confirmed deaths(d).

5.3.3 *Sketch-Corr with heatmaps and scatter plots.* This analysis focused on visualizing the data distributions among tuples from datasets *Ds-Vaccine* and *Ds-CTMD*. Figure 6 shows the visualizations. For *Ds-FAPESP*, we used attribute `stateVac` as color to identify the state of patients from tuples. We selected a sample of 200 records from 01-01-2021 to 02-07-2021. The heatmaps (a and e) show the correlation among the available attributes of every dataset, and the scatter plots in (b and f) show the original distance space generated by MDS. We observe different data distributions when weighting the distance among tuples with attributes related to the number of suspicious deaths (c) and confirmed discharges (d). This tool allows the analyst to observe how each attribute contributes to the distance among records, suggesting how the correlations distort this space. For instance, we observe that by weighting the data space using the number of suspicious deaths, records related to São Paulo state presented a better separability than the other states. For *Ds-CTMD*, we observe how the patient gender (c) impacts the distance among records according to their registered follow-up and how different visual features extracted from images (such as color and texture) separate patients.



Dataset *Ds-Vaccine*

(a) Heatmap of correlations

(b) MDS (original data)

(c) MDS (weighted by suspicious deaths)

(d) MDS (weighted by confirmed discharges)

Colored by state:

Dataset *Ds-CTMD*

(e) Heatmap of correlations

(f) MDS with original data

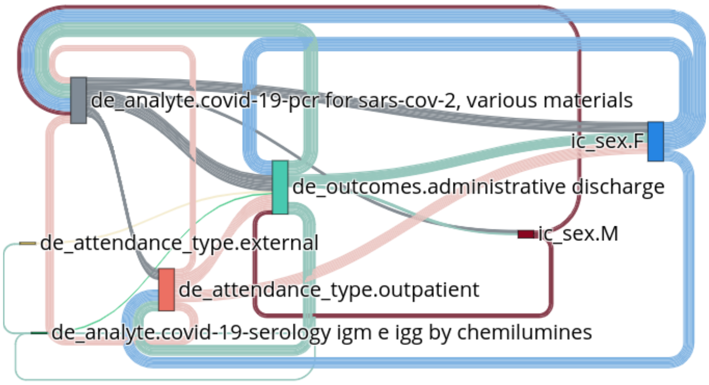
(g) MDS weighted by `gender`

(h) MDS weighted by Edge Histogram features (from `CT-Scan image`)

Colored by follow-up:

Fig. 6. Analyzing the similarity among tuples: (a and e) the heatmaps show how attributes correlate to each other, and the two highlighted ones were used to weight the distance spaces generated by MDS. The original spaces (b and f) generated by MDS show the data distribution, (c-d, g-h) and the weighted spaces show how the data distribution changes according to the correlations of the chosen attribute.

5.3.4 *AR and lift correlation with Sankey diagrams.* For further analyzing the available correlations among categorical attributes, we selected a random sample from *Ds-FAPESP* with patients presenting analytes related to COVID-19, selecting attributes `de_sex`, `de_analyte`, `de_attendanceType`, and `de_outcomes` to provide better AR analysis. Figure 7 shows the discovered AR patterns that present lift correlation different from 1. The diagram items show two analyte values related to COVID-19 found as frequent. We notice reasonable transitions between `de_outcomes` and `de_attendance_type`, which mostly relate to cases of administrative discharge. Also, many patients with administrative discharge were female, as the transitions of `de_outcomes` to `ic_sex` show.



de_analyte.covid-19-pcr for sars-cov-2, various materials
ic_sex.F
de_outcomes.administrative discharge
ic_sex.M
de_attendance_type.external
de_attendance_type.outpatient
de_analyte.covid-19-serology igm e igg by chemilumines

Vertical bar: corresponds to an item, which is composed of the concatenation of an attribute and value; **Transition:** the edges represent the co-occurrence among items with lift different from 1. The diagram shows only the frequent rules discovered.

Fig. 7. Sankey diagram with AR discovered from categorical values of *Ds-FAPESP*.

Table III. Elapsed time (in seconds) of every *Sketch*$^+$ step and tools.

Dataset	SQL Query	Data Cleaning	*Sketch-Corr*	Sim. Search	MDS Projection	AR	ANOVA
Ds-FAPESP	370.20	24.73	26.93	1.59	10.95	27.95	3.41
Ds-Vaccine	499.67	3.44	1.69	0.37	7.67	15.23	0.25
Ds-CTMD	–	3.41	3.63	0.07	8.95	0.79	0.05

5.4 Performance Analysis

We compute the execution time of every step of *Sketch*$^+$ for all datasets, running the tasks ten times and taking the average. Table III shows the results. We selected the experimental data using SQL statements, with an average 370.2 and 499.67 seconds for *Ds-FAPESP* and *Ds-Vaccine*, respectively. We used the original data for *Ds-CTMD*, hence the missing value. In general, the tasks performed over *Ds-FAPESP* presented the highest execution time, due to the data volume. Considering all datasets, the slowest tasks were respectively the SQL query, AR discovery, and the MDS projection. The fastest tasks were respectively *ScatterCorr*, similarity search, and *ANOVA*. We omit *ScatterCorr* from Table III because the plots were generated in very small execution times: 0.06, 0.09, and 0.05 seconds, respectively. All in all, *Sketch*$^+$ is fast when preparing, processing, and analyzing even large amounts of data using the provided methods and algorithms.

6. DISCUSSION AND CONCLUSIONS

In this work, we presented the *Sketch*$^+$ for correlation-based exploratory data analysis supported by visual tools. The method combines categorical, numerical, date, and complex (such as text and images) attributes in the analysis. *Sketch*$^+$ also allows the analyst to pose similarity queries with heterogeneous types. Algorithm *Sketch-Corr* computes the correlation among attributes based on individual distance spaces. *Sketch*$^+$ evaluates the overall correlation among heterogeneous attributes with *Sketch-Corr*, categorical ones with AR and lift, and the relationship between categorical and numerical attributes with *ANOVA* correlation and pixel-oriented parallel visualizations, which also deals with date values. Also, we provide *Sketch-GUI* as the prototype that implements *Sketch*$^+$ and provides a preprocessing module to assist the analyst in discretizing and adjusting categorical terms in the working data. *Sketch-GUI* is openly available for research purposes.

The experimental analysis performed over three real-world databases shows the application of correlation and visual tools of *Sketch*$^+$ in analyzing heterogeneous data. *Sketch*$^+$ runs over an RDBMS, enabling fast and robust data operations. A performance analysis over the provided tools has shown that *Sketch*$^+$ can run different tasks timely. The correlation-based analysis proposed in this work aimed at objectively improving the interestingness of pattern recognition during EDA. As future work, we aim to assess the subjective quality of our tool with domain specialists, assisted by *Sketch-GUI*. We also intend to include missing data treatment into the preprocessing module, reducing the impact of incompleteness over similarity queries and EDA [Rodrigues et al. 2020].

Acknowledgment

This research was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, by the São Paulo Research Foundation (FAPESP, grants No. 2016/17078-0, 2020/07200-9, 2020/10902-5, 2020/11258-2, 2021/11403-5), and by the National Council for Scientific and Technological Development (CNPq).

REFERENCES

Abdullah, S. S. et al. Visual analytics for dimension reduction and cluster analysis of high dimensional electronic health records. *Informatics* 7 (2): 17, 2020. DOI: 10.3390/informatics7020017.

Abedjan, Z., Golab, L., and Naumann, F. Profiling relational data: a survey. *The VLDB Journal* 24 (4): 557–581, 2015. DOI: 10.1007/s00778-015-0389-y.

Afshar, P., Heidarian, S., et al. COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning. *Scientific Data* 8 (1): 121, 2021. DOI: 10.1038/s41597-021-00900-3.

Bernier, A. and Thorogood, A. Sharing bioinformatic data for machine learning: Maximizing interoperability through license selection. In *Bioinformatics*. SCITEPRESS, Valletta, Malta, pp. 226–232, 2020. DOI: 10.5220/0009179502260232.

Brownlee, J. *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery, San Juan, PR, USA, 2020.

Cazzolato, M., Rodrigues, L., Ribeiro, M., Gutierrez, M., Traina-Jr., C., and Traina, A. J. M. Similarity search and correlation-based exploratory analysis in ehrs: A case study with covid-19 databases. In *SBBD Conference*. SBC, Porto Alegre, RS, Brasil, pp. 25–36, 2021. DOI: 10.5753/sbbd.2021.17863.

Deza, M. M. and Deza, E. Encyclopedia of distances. In *Encyclopedia of distances*. Springer, Berlin, Heidelberg, pp. 1–583, 2009. DOI: 10.1007/978-3-642-00234-2.

DSouza, J. et al. Using exploratory data analysis for generating inferences on the correlation of COVID-19 cases. In *ICCCNT Conference*. IEEE, Kharagpur, India, pp. 1–6, 2020. DOI: 10.1109/ICCCNT49239.2020.9225621.

FAPESP. FAPESP COVID-19 Data Sharing/BR, 2020. https://repositoriodatasharingfapesp.uspdigital.usp.br.

Farias, J. d., Barioni, M. C., and Rezende, H. Explorando o uso de árvores b+ na indexação de dados por similaridade. In *SBBD Conference*. SBC, Porto Alegre, RS, Brasil, pp. 163–168, 2019. DOI: 10.5753/sbbd.2019.8817.

Gansel, X., Mary, M., and van Belkum, A. Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review. *EJCMID* 38 (6): 1023–1034, 2019. DOI: 10.1007/s10096-019-03501-6.

Gonçalves, M. V. F. et al. Datasets Cured and Enriched with Provenance from the National Vaccination Campaign Against COVID-19, 2021. DOI: 10.5281/zenodo.5193920.

Guo, R. et al. Comparative visual analytics for assessing medical records with sequence embedding. *Visual Informatics* 4 (2): 72–85, 2020. DOI: 10.1016/j.visinf.2020.04.001.

Hameed, M. and Naumann, F. Data preparation: A survey of commercial tools. *SIGMOD Rec.* 49 (3): 18–29, dec, 2020. DOI: 10.1145/3444831.3444835.

Han, J., Kamber, M., and Pei, J. *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann, USA, 2011. ISBN: 978-0123814791.

Hoshen, Y. and Wolf, L. Unsupervised correlation analysis. In *CVPR Conference*. Computer Vision Foundation / IEEE Computer Society, Salt Lake City, UT, USA, pp. 3319–3328, 2018. DOI: 10.1109/CVPR.2018.00350.

Huang, H., Zhang, R., and Lu, X. A recommendation model for medical data visualization based on information entropy and decision tree optimized by two correlation coefficients. In *ICICM Conference*. ACM, Prague, Czech Republic, pp. 52–56, 2019. DOI: 10.1145/3357419.3357436.

Hund, M., Böhm, D., Sturm, W., Sedlmair, M., et al. Visual analytics for concept exploration in subspaces of patient groups. *Brain Informatics* 3 (4): 233–247, 2016. DOI: 10.1007/s40708-016-0043-5.

Jensen, P. B., Jensen, L. J., and Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* 13 (6): 395–405, 2012. DOI: 10.1038/nrg3208.

Kaieski, N., de Oliveira, L. P. L., and Villamil, M. B. Vis-health: Exploratory analysis and visualization of dengue cases in brazil. In *HICSS Conference*. IEEE, Koloa, HI, USA, pp. 3063–3072, 2016. DOI: 10.1109/HICSS.2016.385.

Kwon, B. C., Anand, V., et al. Dpvis: Visual analytics with hidden markov models for disease progression pathways. *IEEE Trans. Vis. Comput. Graph.* 27 (9): 3685–3700, 2021. DOI: 10.1109/TVCG.2020.2985689.

Lanucara, S. et al. Harmonization and interoperable sharing of multi-temporal geospatial data of rural landscapes. In *Int. Symp. on New Metropolitan Perspectives*. Springer, Italy, pp. 51–59, 2018. DOI: 10.1007/978-3-319-92099-3_7.

Min. Saúde. Campanha nacional de vacinação contra COVID-19, 2022. https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao.

Müller, H., Castelo, S., Qazi, M., Freire, J., et al. Openclean - data cleaning for Python, 2021. https://github.com/VIDA-NYU/openclean.

Nouri, M., Lizotte, D. J., Sedig, K., and Abdullah, S. S. VISEMURE: A visual analytics system for making sense of multimorbidity using electronic medical record data. *Data* 6 (8): 85, 2021. DOI: 10.3390/data6080085.

Rodrigues, L. S., Cazzolato, M. T., Traina, A. J. M., and Traina-Jr., C. Taking advantage of highly-correlated attributes in similarity queries with missing values. In *SISAP Conference*. LNCS, vol. 12440. Springer, Copenhagen, Denmark, pp. 168–176, 2020. DOI: 10.1007/978-3-030-60936-8_13.

Samet, H. *Foundations of multidimensional and metric data structures*. M. K. series in data management systems. Academic Press, USA, 2006. ISBN: 978-0-12-369446-1.

ten-Caten, F. et al. In-depth analysis of laboratory parameters reveals the interplay between sex, age, and systemic inflammation in individuals with covid-19. *IJID* vol. 105, pp. 579–587, Apr, 2021. DOI: 10.1016/j.ijid.2021.03.016.

Wu, A. et al. Survey on artificial intelligence approaches for visualization data. *CoRR* vol. abs/2102.01330, pp. 1–20, 2021.

Yadav, P., Steinbach, M., Kumar, V., and Simon, G. Mining electronic health records (EHRs): A survey. *ACM Computing Surveys* 50 (6): 85:1–85:40, Jan., 2018. DOI: 10.1145/3127881.

Yang, F. et al. Correlation judgment and visualization features: A comparative study. *IEEE TVCG Journal* 25 (3): 1474–1488, 2019. DOI:10.1109/TVCG.2018.2810918.