POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

A tool to support data exploration and preparation

Tesi di Laurea Magistrale in
Computer Science and Engineering
Ingegneria Informatica

Author: **Lorenzo Vela**

Student ID: 969435
Advisor: Prof. Cinzia Cappiello
Co-advisors: Dott. Ing. Camilla Sancricca
Academic Year: 2021-22

Abstract

Data analysis is becoming increasingly important with the passing of time and represents an important decision factor for strategic and financial choices. Given this key role played by data analysis, its outcome should be more accurate as possible. To this aim, one of the most crucial elements is data preparation, which is the process of preparing the data before analyzing them. Considering that, usually, data preparation takes more or less 80% of the total time of data analysis, the objective of this thesis is to develop a tool to support data preparation and exploration. These phases include different distinct steps: Profiling, Wrangling, Cleaning, and finally duplicate detection and handling. Anyway, the tool implements all the classic data preparation pipeline functionalities. Moreover, the tool is meant to be used by users without any developing experience, so it has been developed to be easy to use and install. The entire design, development, and testing of the tool have been conducted by considering a dataset taken from the database of the Municipality of Milan: it describes personal care sites located within the city. Given that this dataset had quite a poor quality, it has been a very good and strong starting point to assess the tool's potential. Indeed the results obtained with this dataset were very good as the tool solved most of the dataset's issues like the presence of null values, duplicates, and redundancy of information.

Keywords: Data Quality, Data Preparation, Data Exploration, Tool

Abstract in lingua italiana

Dato il ruolo chiave svolto dall'analisi dei dati, il suo risultato dovrebbe essere il più accurato possibile. Per ottenere questo risultato, uno degli elementi più essenziali è il processo di preparazione dei dati prima della loro analisi. Si consideri che di solito la preparazione dei dati richiede più o meno l'80 percento del tempo totale dell'analisi dei dati. L'obiettivo di questa tesi è stato quello di sviluppare uno strumento di preparazione ed esplorazione dei dati. Queste fasi comprendono tre diversi passaggi: Profilazione, Wrangling e Pulizia. Per ognuna di queste fasi sono state implementate funzionalità specifiche. Inoltre, lo strumento è destinato a essere utilizzato da utenti senza alcuna esperienza di sviluppo, quindi è stato sviluppato con l'obiettivo di essere facile da usare e da installare. L'intero sviluppo e la sperimentazione dello strumento si basano su un dataset tratto dal database del Comune di Milano, che rappresenta i siti di assistenza alla persona situati all'interno della città. Dato che questo dataset aveva una qualità piuttosto scarsa, è stato un ottimo e solido punto di partenza per valutare le potenzialità dello strumento. I risultati ottenuti con questo dataset sono stati infatti molto buoni, in quanto lo strumento è in grado di risolvere la maggior parte dei problemi del dataset, come ad esempio la presenza di valori nulli, duplicati e ridondanza di informazioni.

Parole chiave: Qualità dei dati, Preparazione dei dati, Esplorazione dei dati, Strumento

Contents

1 | Introduction

The success of any data analysis project depends on the quality of the input data. Nowadays, data analysis is a crucial step for companies and enterprises: they actually use the outcome of these analyses to take important and strategic business decisions.

Around 80% of the data analysis time is actually spent on data preparation. Data preparation ensures that the data is clean, consistent, and ready for analysis. Unfortunately, real-world data are often affected by errors, inconsistencies, incomplete values, or biases. In order to avoid having erroneous and unusable outputs, data preparation has become a crucial step in the data analysis pipeline for guaranteeing an appropriate quality output [16]. Information extracted by poor quality data may be incorrect, potentially driving organizations towards bad decisions and money loss. In order to avoid this, actions can be put in place before and after the data collection. Data preparation focuses on data that have already been collected, and its process is explained below.

Data preparation is the activity that processes raw data to improve quality and gives several benefits to data analysis applications by achieving more reliable results. Data preparation is not a process that can be carried out blindly; it is necessary to apply it with the knowledge of the effect on the data being prepared and also technical know-how on the actions being carried out. Understanding the preparation functions and their applicability is more important than understanding how they have been executed. Users unfamiliar with the general data preparation pipeline are not supposed to use this tool as they can damage the data instead of enriching them.[15].

Another important characteristic of data preparation is that there is no general structure to be followed. The process is strictly dataset dependent and this means that if a process works very well with a dataset maybe applying it to another dataset collected for another analysis could potentially make things worse.

In this thesis, it is presented a new data preparation tool that helps to ensure the quality of the data used in the analysis. The tool has been entirely developed in Python and it is designed to assist with the profiling, wrangling, and cleaning of the data. By providing a user-friendly interface, the tool enables also non-expert users to perform data preparation and analysis.

The tool provides a wide set of profiling, wrangling, and cleaning functionalities to enhance the quality of the uploaded dataset.
With the profiling section, the user is able to see the dataset's properties from different perspectives with different granularities. Within the wrangling section, the tool provides basic wrangling functionalities like column renaming or column merging. Finally, by using the cleaning functionalities, the user can handle, for example, duplicates or missing values.

The dataset that has driven the design and development of the tool is taken from the database of the Municipality of Milan and some characteristics of the dataset will be also presented within the dataset Section 4.1.
Anyway, an important remark that should be done about the tool is that it is completely dataset independent and can handle both .csv and .xlsx files.
The thesis is structured as follows. In Chapter 2 are described the main concepts related to Data Quality and Data Preparation, given their importance in the Data Analysis pipeline. This chapter also presents some related works. In Chapter 3 all the profiling, explorations, wrangling, and cleaning functionalities offered by the tool are explained. Chapter 4 presents the way in which the tool works and its functionalities. Finally, Chapter 5 concludes the work and presents some future developments.

2 | State of the art

This chapter describes the fundamental concepts on which this entire thesis is based and some related works. Section 2.1 provides a general overview of Data Quality and its importance in the data analysis pipeline.
Section 2.2 introduces the reader to the data preparation process, focusing on data profiling, data wrangling, and data cleaning processes.
Finally, Section 2.3 presents some other tools that focus on data exploration and preparation.

2.1. Data Quality

With the passing of time, data are becoming increasingly important in almost every process regardless of the context. Given this key role in supporting managerial and business decisions, organizations are trying to collect them as much as they can in every possible way.

This "rush to collect" could result in obtaining a large amount of data, but also most probably result in a poor quality collection. Indeed, poor data quality can lead to incorrect conclusions, faulty insights, and wrong decisions, potentially resulting in wasted time, money, and resources.

Data Quality is defined as the ability of a data collection to meet user requirements [4]. This definition is hiding another shade of the Data Quality: it is not possible to achieve a level of data quality at which the data can satisfy all possible use cases, more precisely the absolute data quality [13]. For example, a dataset may have errors, be incomplete, have contradictions, duplication, format issues, or typos, all of which can lower overall quality. However, the quality standards that the dataset must meet, depend strictly on the analysis that should be performed.

Poor data quality can result from a variety of factors. Historical changes that alter the significance of data over time can contribute to this. Neglecting the relevance of data can also lead to the degradation of data quality since it may result in datasets that are unsuitable for their intended purpose. Additionally, privacy rules designed to protect hard-to-collect data can lead to issues of missing values. Moreover, enrichment and integration processes can decrease the quality of the data source. Poor data quality also increases operational costs since time and other resources are spent detecting and correcting errors [9].

In the next section the Data Preparation is presented: a set of actions and assessments with the aim of increasing Data Quality.

2.2. Data Preparation

Data preparation is typically an iterative process of manipulating raw data, which is often unstructured and messy, into a more structured and useful form that is ready for further analysis. The whole preparation process consists of a series of major activities or tasks including data profiling, wrangling, integration, and cleaning that usually take more or less the 80% of work throughout an entire data mining application [1].
Nowadays it is always more and more important to have instruments and tools for understanding the main characteristics of the dataset that is currently being analyzed. Moreover, it is valuable to have interactive tools that do not just show information about the dataset but that also provide functionalities to interact with this information, like imputing in a column that has a lot of null values, or understanding if the dataset has some rows with a lot of null values.
Indeed, the scope of the design and development of this tool is to provide users with an easy-to-use instrument, capable of profiling the data but at the same time able to perform basic wrangling and cleaning operations. The interface has been designed to be as intuitive as possible and best guide users' choices. Note that the data integration is out of the scope of this thesis as the tool accepts only one dataset at a time, so it cannot integrate data from multiple sources.

2.2.1. Data Profiling

Data profiling examines and analyzes a dataset to understand its structure, content, and quality. Data profiling aims to provide insights into the data, identify data quality issues, and prepare the data for further processing, analysis, or modeling.

The profiling is only a preliminary action on the raw data with the aim of producing statistics and metadata like some of the data quality measures. This phase is meant to highlight some of the dataset's characteristics like null values, outliers, and relationships between attributes, which can be used to optimize data processing and analysis.

This process should be constantly performed, not only before processing the data but also to assess all the quality measures required and to check if the wrangling and cleaning process fits the user requirements and effectively increases the dataset's quality. It is important to remark that the dataset is not modified at all during this phase.

2.2.2. Data Wrangling

By definition data wrangling is "the process of profiling and transforming datasets to ensure they are actionable for a set of analysis tasks. One central goal is to make data usable: to put data in a form that can be parsed and manipulated by analysis tools. Another goal is to ensure that data is responsive to the intended analyses: that the data contain the necessary information, at an acceptable level of description and correctness, to support successful modeling and decision-making" [7].

With the data wrangling for example, are carried out actions like converting data types, standardizing the format of the attributes, merging or splitting different columns. These actions are fundamental also when merging two different data sources. In order to perform most of the data wrangling tasks it is mandatory to have some knowledge of the dataset that is currently being analyzed. Indeed, the various steps applied and the transformations are always performed looking towards the final scope of the analysis and of the data.

This is the phase to decide, for example, whether to use or not to use part of the data or how to modify them in order to obtain not just a general "very good dataset" but the "best dataset for the analysis". Data wrangling is a time-consuming and complex process, but it is essential to ensure that the data used for analysis is accurate, relevant, and reliable. By performing data wrangling, analysts can gain a better understanding of the data and generate more accurate and meaningful insights.

2.2.3. Data Cleaning

Data cleaning, also known as data cleansing, is the process of identifying and correcting or removing errors, inconsistencies and inaccuracies in data, and also removing duplicates. It is a critical step in the data preparation process and is essential for accurate analysis and decision-making, as it helps to improve data quality and reduces the risk of incorrect results.

Typical cleaning tasks can be outlier detection, imputing missing values, or removing incorrect functional dependencies. One aspect of data cleaning involves checking individual records within a dataset to ensure they are accurate and consistent with the larger population or dataset. This process can be difficult for humans to do manually, especially for large datasets. Algorithms can help automate the expectation checking process, making it more efficient and accurate. By using algorithms, data cleaning can be done more quickly and with less chance of errors than if done manually by humans.

Especially nowadays, data are collected at an extremely fast pace, and this means that data cleaning should not be seen as a one-time process: this task is continuously ongoing and needs to be carried out regularly to maintain data accuracy and consistency. Therefore, it is crucial to have a well-defined data cleaning process in place that is regularly reviewed and updated.

2.3. Existing data preparation tools

In this section, the existing literature and tools related to data preparation are presented. This chapter aims to identify key features and main gaps in the literature that the tool developed in this thesis aims to fill.

2.3.1. Stanford Data Wrangler

This tool is an interactive system for creating data transformations. Wrangler combines direct manipulation of visualized data with the automatic inference of relevant transforms, enabling analysts to iteratively explore the space of applicable operations and preview their effects. Wrangler significantly reduces specification time and promotes the use of robust transformation methods applied entirely to the chosen column instead of manual editing [11].

The main difference with the tool presented in this thesis is that in the Wrangler the user can interact directly with the dataset by clicking cells and columns and so, the Wrangler can proactively propose actions and possible transformations.

The tool focuses more on data transformation and cleaning with respect to profiling and despite the range of actions in a bit wider with respect to this tool, the lack of profiling implies a huge limitation: the user is not made aware of the dataset's problems.

2.3.2. Trifacta Wrangler

Trifacta is a commercial data preparation and data wrangling tool that helps users to clean, transform, and enrich raw data into a structured format that is suitable for analysis. Trifacta's user interface allows users to visually explore and manipulate data, making it easy to identify and correct errors, merge and split columns, extract and parse data, and perform other data transformations. Trifacta also includes a wide range of built-in functions for data manipulation, as well as support for writing custom functions in Python or JavaScript.

One of the key features of Trifacta is its ability to handle messy, unstructured data from a variety of sources, including CSV files, Excel spreadsheets, JSON, and XML. Trifacta can also work with large datasets, leveraging distributed computing and parallel processing to speed up data processing [2].

Nowadays, every large company uses a data preparation tool, and this is the main reason why a lot of commercial tools have been developed in order to fulfill this need. Trifacta Data Wrangler is one of them, anyway, this category of tools implements very powerful machine-learning functions and also different data preparation pipelines depending on the context of the dataset they're preparing.

In this chapter has been presented the state of the art of the thesis that has covered the main concepts of data quality, data preparation and two related works.

The next chapter will explain from a theoretical point of view all the functionalities implemented in the tool for profiling, exploring, wrangling, and cleaning the data.

3 | Offered Functionalities

This chapter will cover all the functionalities currently implemented within the tool from a theoretical point of view. The general pipeline of data analysis starts with the profiling of the data. This step allows the user to understand better the data his working with and to highlight in an aggregate way statistics and metadata about the dataset. Moreover, some profiling functionalities have the capability to exploit hidden patterns throughout all the attributes. After the profiling has been completed the user is supposed to be aware of some of the problems of the dataset. Is also important to remind that profiling is something that should be constantly performed as it is also the best way to assess if the wrangling and cleaning actions are increasing the quality of the dataset and solving the problem highlighted.

After the profiling, another important step of data analysis is data wrangling. This phase is in charge of unifying the formats of the data and transforming the dataset according to the needs of the user. For example, two wrangling functionalities are column splitting or merging: they are very useful also in the data integration field. In this case, the need would be to have two datasets with the same format and with the columns being as much as possible similar to each other.

Finally is performed the data cleaning that includes various functionalities like the handling of null values. These actions aim to increase the overall quality of the dataset. Data cleaning is the phase meant to fix or remove incorrect, corrupted, incomplete, and duplicate data within a dataset.

3.1. Profiling and Exploration Functionalities

- Single Column Analysis

  Single column analysis provides a summary of the statistical and descriptive properties of the data in that column, such as null values, and unique values but also mean, median, mode as well as the frequency of occurrence of distinct values.

  By analyzing each column one by one, the characteristics and nature of the data could be better understood. Moreover, data quality insights and issues can be easily

detected.
The single column analysis is often used as a first step in data profiling to gain an initial understanding of the dataset before moving on to more complex analyses.

- Correlation table
A correlation table is a powerful tool that provides all the correlation coefficients between all the possible pairs of attributes in a table. It makes easy the identification of patterns between attributes and it works very well also for large datasets. The tool relies on a particular correlation table called the Phi_k correlation table [3]: it is a new and practical correlation coefficient based on several refinements to Pearson's hypothesis test of independence of two variables. The combined features of Phi_K have some advantages with respect to the other existing coefficients: (i) it works consistently well between categorical, ordinal, and interval variables; (ii) it captures non-linear dependency; (iii) it reverts the Pearson correlation coefficient in the case of the bi-variate normal input distribution. Anyway, a weak point of this table is that it is very expensive from a computational point of view.

- Redundancy of Information The tool can also spot possible data redundancies in column names and within the same row for different attributes. Data redundancy occurs when the same piece of data exists in multiple places throughout the dataset, and this could lead to data inconsistency. It is important to spot and handle these redundancies as apart from having misleading and nonlinear information, they also represent a waste of memory, computational resources, and time.

3.2. Wrangling Functionalities

- Column Merging
Column merging is the process of combining or consolidating two or more columns of data into a single column. This is frequently done to simplify data analysis or to prepare data for a specific use case, such as creating a report or visually representing the data. Depending on the data and the desired output, column merging can be accomplished using various techniques such as concatenation, joining, and merging.

- Column Splitting
Column splitting is the process of splitting a single column of data into multiple columns. This is done to organize and simplify the data or prepare it for a specific use case, such as analysis or visualization. Column splitting can be accomplished through various methods, including using delimiters (e.g., commas, tabs, spaces) to separate the data into separate columns or using string manipulation functions

to extract specific parts of the data into new columns. The technique used will be determined by the nature of the data and the desired output.

- Column Renaming
  Column renaming refers to changing the name of a column in a table within a dataset. Renaming a column allows the user to give a more descriptive or meaningful name to a column or to correct a spelling error or other mistake. It can also be helpful when multiple tables to be joined are present, making it easier to understand the relationships between columns in different tables.

- Edit Values
  Values editing allows the user to make changes to multiple rows or columns in a dataset simultaneously. This is especially useful when working with large datasets or when making the same change to multiple values. Anyway, it is important to exercise caution and carefully review the changes to ensure they are correct and will not result in unintended consequences or errors.

3.3. Cleaning Functionalities

- Column Dropping
  Column dropping is the action of removing one or multiple columns from a dataset. There could be multiple reasons why to drop a column: the high presence of null values, the column always has the same value so it is not descriptive for the dataset, or the column is highly correlated with another column that will be kept. Another possible reason could be that the specific column is useless for the analysis.
  When working with huge datasets, column dropping can be a valuable strategy to minimize content and enhance computation performances. Anyway, it is always important to be cautious when removing a column: the information dropped will never be available for future analysis.

- Imputing missing values
  Often, missing values represent a problem for the outcome of the analysis. A missing value means that the dataset is missing to describe a characteristic of a certain object.
  Input a missing value is trying to guess the real value for that attribute as precisely as possible. There are various imputing techniques in the literature, the most accurate ones are those that often require external knowledge and sources to fill as best the missing attributes. For numeric columns, missing values are often replaced with the mean, the maximum, the minimum, the mode, and the value 0. For the string

columns, the available options are, commonly, the following value, the previous value, a custom value, and the mode.
Anyway, imputing missing values may add bias to the dataset, particularly if they are not missing randomly. Because of this, it is crucial to assess in advance the imputation technique used and its potential side effects on the analysis.

- Duplicate Detection
  Duplicate detection aims to discover multiple representations in a dataset of the same real-world object. It aims to identify multiple representations that refer to the same instance in the real world.
  Someone may say that the duplicate detection would correspond to the comparison of each row with all the other ones, but obviously, this would not be so convenient in terms of efficiency: that's why usually data scientists aim to reduce as much as possible the number of comparisons.
  This process is called space reduction: its goal is to reduce the number of pairwise comparisons and, at the same time, to increase the computation efficiency. By making intelligent guesses about which records have a high probability of representing the same real-world entity, the search space is reduced with the drawback that some duplicates might be missed [6]. The reduction of the comparison space can be obtained with many techniques like blocking, pruning, or sorted neighborhood. Remind that no records have been compared yet at this point since the aim is just to reduce the number of the needed ones.
  Within the tool has been implemented only the blocking technique: the file is partitioned in exclusive blocks and limiting comparisons to records within the same block. Blocking can be implemented by choosing a blocking key and grouping all records with the same values on the block. An example of blocking criteria could be that records that have the same postcode value are inserted into the same block. Finally, the tool will compare only the records within the same block.
  This step will be the actual deduplication process. Also to perform this there are many techniques available in the literature like probabilistic, empirical and knowledge-based. In the tool has been implemented a probabilistic method on a string-based similarity function. For the specific string similarity function has been used the Jaro-Winkler [18] function which is defined as a variant of the Jaro distance. The results of this similarity distance will give an idea to the user of how much 2 rows are similar to each other. The Jaro distance algorithm takes into account the number of matching characters between two strings and the positions of those characters. It gives a score between 0 and 1, where 0 means no similarity and

1 means an exact match.

The "Winkler" is a modification of the Jaro algorithm: it adds a scale factor that gives more weight to the initial characters of the string.

In the next chapter will be presented more in specific how the tool works and how it is structured. Within it will be also explained the dataset that has driven the development of the tool and its characteristics.

4 | Implementation

This chapter presents more in detail the entire tool and all its pages. Moreover, an introduction to the dataset that guided the development of the tool will be also given.

4.1. The Dataset

This section describes the dataset taken into account. A remark that it is important to do is that the tool is developed to work with every dataset since all the functionalities are dataset independent.

4.1.1. General Overview

The dataset is extracted from the database of the Municipality of Milan and represents personal care businesses such as hairdressers, beauticians, and tanning salons. It has 5200 records and 15 attributes that will be described in more detail in the following:

- Codice
  Is the unique code for each record, a primary key for the dataset.
- Ubicazione
  Is the complete address of the shop, with the street address, the street number, and the neighborhood number.
- Area di Competenza
  Is the business field of the shop. The particularity is that it has always the same value.
- DescrizioneVia
  Is the street address of the shop.
- Civico
  Is the street number of the shop.

- CodiceVia
  Is a code that identifies the street of the shop.

- superficie_lavorativa
  Is the surface of the shop.

- tipo_eser_pa
  Is the type of activity of the shop. Some examples from the dataset are hairdressers, beauticians, spas, and many others. This attribute categorizes most the record.

- CAP
  Is the ZIP of the shop.

- ID_NIL
  The NIL (Nuclei di Indentità Locale - Cores of local identity) is an additional geographical indication code adopted by the Municipality of Milan. This is the NIL ID of shop.

- NIL
  This is the NIL name of the shop.

- MUNICIPIO
  Neighborhood name of the shop.

- LONG_WGS84
  Longitudinal coordinates of the shop.

- LAT_WGS84
  Latitudinal coordinates of the shop.

- Location
  Couples of geographical coordinates of the shop: Latitude and Longitude.

This dataset can be queried in order to obtain insights and statistics on the overall presence of these activities in the different areas of the city.
The limitations of the dataset are, at the same time, the elements that best enhance all the wrangling and cleaning features of the tool:

- Null values and Incomplete Rows
  The dataset has 230 rows that have 9 attributes out of 15 equal to null, moreover is present a column with more than 40% of null values

- Column with always the same value
  Within the dataset, there is the column called "Area di Competenza" that has the

same value for all the rows. This column is suggested to be dropped as it is not bringing any additional information to the dataset.

- Duplicates
  The dataset has a huge number of duplicates, there is not a precise number, but from the analysis performed it is possible to state that around 25% of the rows can be dropped because refers to the same real world instance of some other rows.

- Correlation
  Given that the dataset has addresses, areas, ZIP codes, and many other information about the position of the activity, makes sense to suppose that all these geographical attributes are correlated with each other. Indeed, all these correlations are correctly captured by the tool.

- Redundancy
  The columns "DescrizioneVia" and "Civico" are a subset of the information present in the column "Ubicazione". This redundancy is correctly highlighted by the tool.

4.2. The Tool

4.2.1. Streamlit and Streamlit-extras

Streamlit (st) is a free and open-source Python-based library that is meant specifically to rapidly build and share machine learning and data science web apps. A very strong point of this library is to not require any extensive knowledge of web development while still giving a lot of freedom to the developer in building interactive applications.

All the Streamlit applications, including this one, are composed of two main components: user interface and application logic. For the user interface, this app relies for the 90% on Streamlit, while for the rest 10% on plain html code, injected in the page via a Streamlit function. With those injections, it has been possible to customize buttons, layouts, text sizes, and almost everything that Streamlit was not able to customize with its built-in functions. Streamlit anyway, supplies many elements that enable the user to interact with the applications like buttons, select boxes, multi-select boxes, radio buttons, checkboxes, and many others. Moreover, it provides many elements to enrich and enhance the graphic interface like expanders, columns, and different types of messages like warnings and errors. For what concerns the application logic, which is totally transparent to the user, it has been entirely developed in Python and the app shows its results to the user via Streamlit output functions like st.write and st.dataframe.

Anyway, a thing that should be remarked is that the ease to use of Streamlit comes with

huge limitations: Streamlit is not meant to host applications that are too structured or complex. Every time that the user interacts with an element, which could be a button or a slider, or anything else, the entire page is refreshed. To do a concrete example, let us imagine that a long page has been entirely read and at the bottom, there is the button to come back to the homepage: if pressed, the user is not redirected immediately to the homepage, but only after the entire code of the page is executed entirely.
Streamlit-extras could be defined as a spin-off library of Streamlit that starting from Streamlit components provides more powerful elements. In this tool, for example, it has been used for the switching page function that is used everywhere and for some multi-select box components, as they are able to display the entire strings while the built-in one of Streamlit had a predefined limit of displayable characters.

4.2.2. Libraries

- Streamlit Nested Layout: Enhances the Streamlit library by allowing the nesting of elements like columns and expanders
- Streamlit Pandas Profiling
  Integrates the Pandas Profiler report function directly in Streamlit
- Pandas [12]
  Provides powerful data structures and functions for working with structured data.
- Numpy [8]
  Supplies powerful arrays and functions for numerical operations.
- Random [17]
  Provides functions for generating pseudo-random numbers and choices from a range or sequence.
- Os [17]
  Provides a way to interact with the underlying operating system by providing functions for file and directory operations, process management, and environment variables.
- Time [17]
  Supplies functions to work with dates, times, and time intervals.
- Json [17]
  Provides functions to encode and decode JSON data.
- RecordLinkage [5]

Provides tools for linking and deduplicating datasets or rows based on record similarity.

- Jaro [10]
  Provides a function to calculate the Jaro similarity measure between two strings.

- Sklearn [14]
  Provides tools for machine learning, including data pre-processing, model selection, and evaluation.

4.3. Implementation and UI

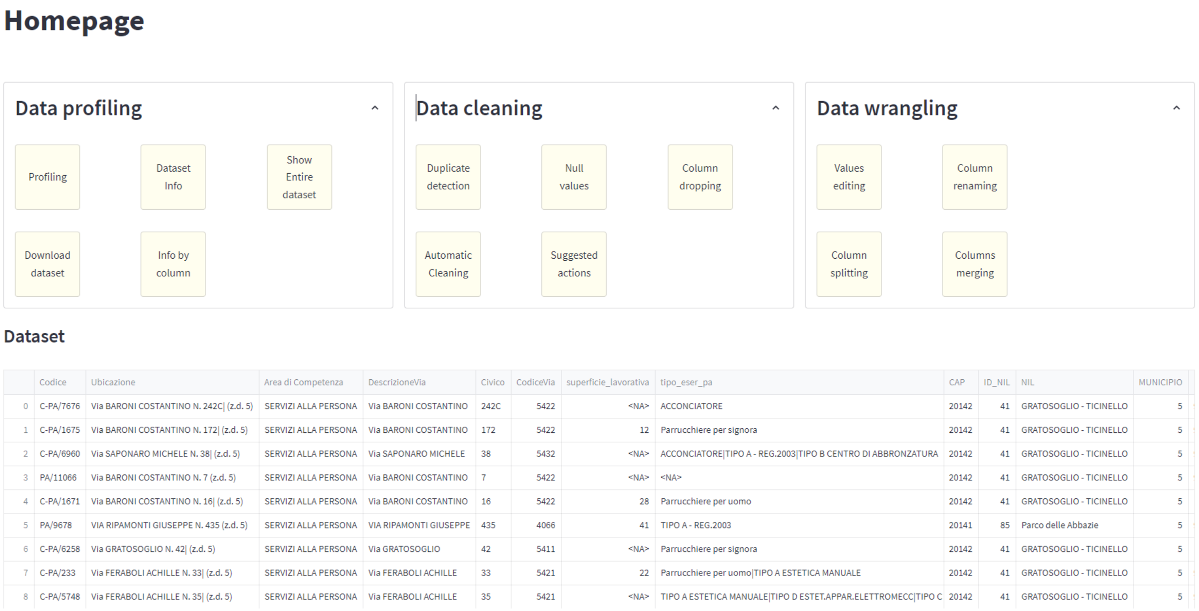
This section explains page by page the UI and its implementation. The Homepage is divided into 3 main parts that are: Profiling, Cleaning, and Wrangling. The first section will cover all the pages that do not have any data-related functionality: Upload page and Homepage.

Upload page

The upload page has the only scope, as the name suggests, to give the possibility to upload the dataset. The tool accepts both .csv and .xlsx files. If the file is accepted it will be immediately profiled and the profiler report will be saved within the tool's folder. The user is then able to continue to the Homepage.

Homepage

The Homepage has been developed with the aim of making all the tool's features immediately available.



Dataset

	Codice	Ubicazione	Area di Competenza	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO
0	C-PA/7676	Via BARONI COSTANTINO N. 242C| (z.d. 5)	SERVIZI ALLA PERSONA	Via BARONI COSTANTINO	242C	5422	<NA>	ACCONCIATORE	20142	41	GRATOSOGLIO - TICINELLO	5
1	C-PA/1675	Via BARONI COSTANTINO N. 172| (z.d. 5)	SERVIZI ALLA PERSONA	Via BARONI COSTANTINO	172	5422	12	Parrucchiere per signora	20142	41	GRATOSOGLIO - TICINELLO	5
2	C-PA/6960	Via SAPONARO MICHELE N. 38| (z.d. 5)	SERVIZI ALLA PERSONA	Via SAPONARO MICHELE	38	5432	<NA>	ACCONCIATORE|TIPO A - REG.2003|TIPO B CENTRO DI ABBRONZATURA	20142	41	GRATOSOGLIO - TICINELLO	5
3	PA/11066	Via BARONI COSTANTINO N. 7 (z.d. 5)	SERVIZI ALLA PERSONA	Via BARONI COSTANTINO	7	5422	<NA>	<NA>	20142	41	GRATOSOGLIO - TICINELLO	5
4	C-PA/1671	Via BARONI COSTANTINO N. 16| (z.d. 5)	SERVIZI ALLA PERSONA	Via BARONI COSTANTINO	16	5422	28	Parrucchiere per uomo	20142	41	GRATOSOGLIO - TICINELLO	5
5	PA/9678	VIA RIPAMONTI GIUSEPPE N. 435 (z.d. 5)	SERVIZI ALLA PERSONA	VIA RIPAMONTI GIUSEPPE	435	4066	41	TIPO A - REG.2003	20141	85	Parco delle Abbazie	5
6	C-PA/6258	Via GRATOSOGLIO N. 42| (z.d. 5)	SERVIZI ALLA PERSONA	Via GRATOSOGLIO	42	5411	<NA>	Parrucchiere per signora	20142	41	GRATOSOGLIO - TICINELLO	5
7	C-PA/233	Via FERABOLI ACHILLE N. 33| (z.d. 5)	SERVIZI ALLA PERSONA	Via FERABOLI ACHILLE	33	5421	22	Parrucchiere per uomo|TIPO A ESTETICA MANUALE	20142	41	GRATOSOGLIO - TICINELLO	5
8	C-PA/5748	Via FERABOLI ACHILLE N. 35| (z.d. 5)	SERVIZI ALLA PERSONA	Via FERABOLI ACHILLE	35	5421	<NA>	TIPO A ESTETICA MANUALE|TIPO D ESTET.APPAR.ELETTROMECC|TIPO C	20142	41	GRATOSOGLIO - TICINELLO	5

Figure 4.1: Screenshot of the Homepage

As visible in Figure 4.1, the entire dataset is also presented to allow the user to perform a preliminary inspection. Moreover, the user has the possibility to change the dataset and upload a new one. The button is placed at the end of the page.

4.3.1. Data Profiling

Profiling

The profiling page has been implemented to visualize entirely the pandas' profile report. Within this report can be found 5 different sections:

- Overview
  In the Overview some dataset's statistics like the number of attributes and the number of rows are present. Moreover, there are also some warnings raised by Pandas itself calculated with standard parameters not based on the dataset's shape or type. Obviously, this default configuration on Pandas does not make these warnings very reliable.

- Variables
  In the Variables section, the attributes are analyzed one at a time: a Single Column Analysis can be performed as described in section 3.1. For each attribute, there are the number and the percentage of distinct values and missing values, also all the

occurrences are counted in order to obtain histograms of the values' frequency. It is also possible to expand the section dedicated to each column and retrieve even more details about the average length of the attribute or the percentages of the value occurrences.

- Interactions
  The interactions table represents the relationship between a couple of attributes strictly based on their values.

- Correlations
  The Correlations table provides an overview of all the correlation coefficients of the dataset. The correlation between two columns is defined the degree to which the values of a column are related to the values of another column. The profiling page displays only the Phik correlation table 3.1. This type of correlation has been chosen since it is the one that was highlighting as best the characteristics and the dependencies of the dataset. The Phik correlation table is defined as a "new and practical correlation coefficient that works consistently between categorical, ordinal and interval variables, captures non-linear dependency and reverts to the Pearson correlation coefficient in case of a bivariate normal input distribution."

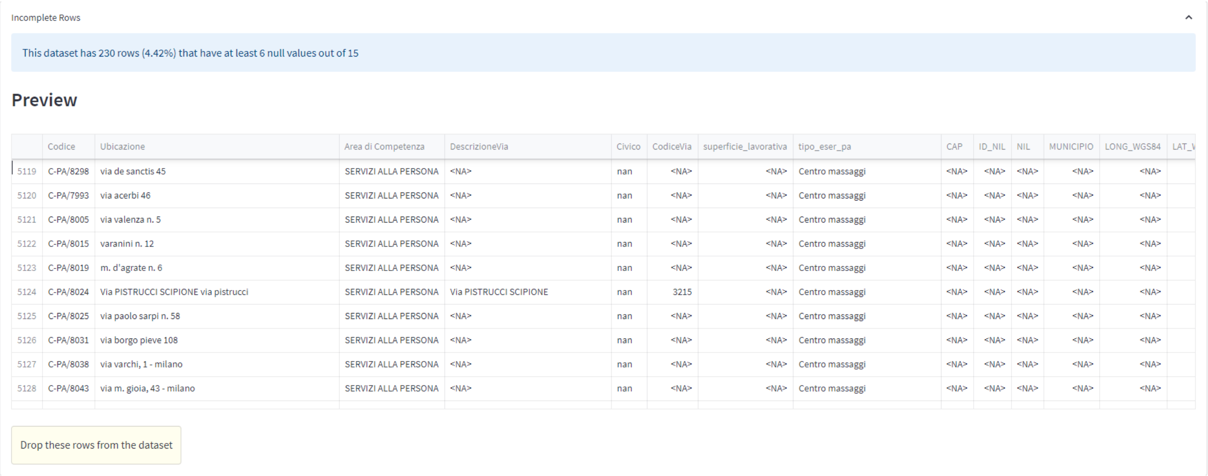
- Missing values
  The Missing values section has 4 different possible visualizations of the missing values: Count, Matrix, Heatmap, and Dendrogram. Anyway, this section does not represent the problem of the null values very well when datasets are particularly large.

All the report is generated by a Streamlit built-in function called "st profile report". This report is powerful but weak at the same time: powerful because is capable of providing some useful elements like histograms and charts that would be hard to obtain without the report, but, at the same time, weak because it is too general and not particularly meant for large datasets.

Dataset Info

With respect to the previous one and like all the other pages the UI generated comes entirely from custom code. This page is divided into different sections and most of the

elements were inserted in an expander in order to provide the cleanest UI possible. After a preview of the dataset, there is the section "Incomplete Rows". A row is considered incomplete by the tool if 40% of the columns are null. Obviously, the 40% is rounded to the closest integer.
The tool then creates a new dataset of support where only the rows that exceed this threshold are inserted. This dataset is then shown, with its length and percentage with respect to the original one, to the users in order to make them aware of these rows.


Incomplete Rows

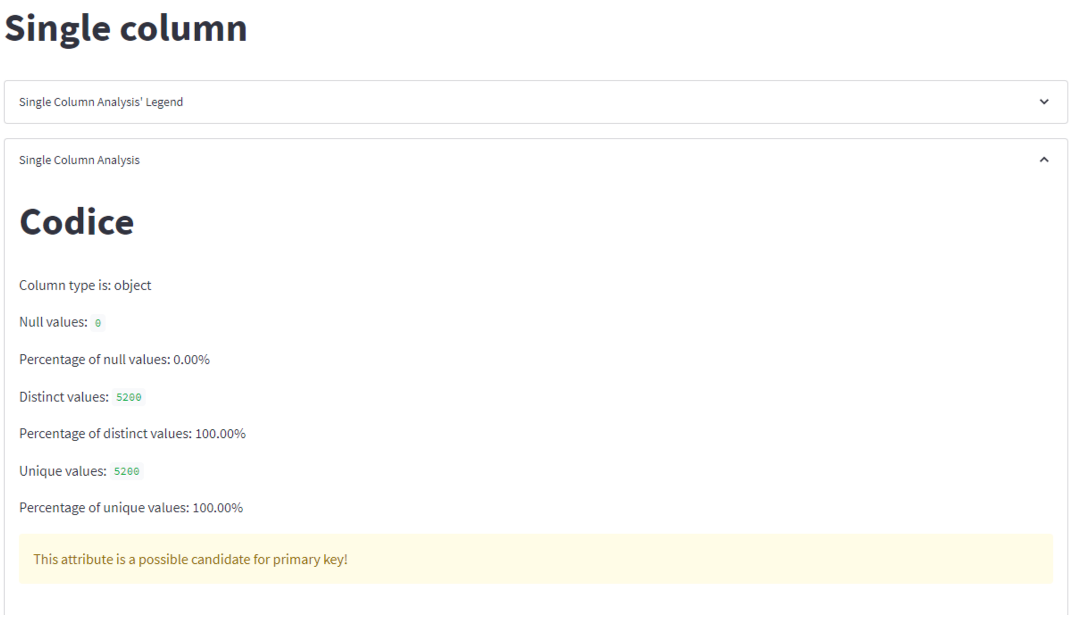
This dataset has 230 rows (4.42%) that have at least 6 null values out of 15

Preview

	Codice	Ubicazione	Area di Competenza	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO	LONG_WGS84	LAT_
5119	C-PA/8298	via de sanctis 45	SERVIZI ALLA PERSONA	<NA>	nan	<NA>	<NA>	Centro massaggi	<NA>	<NA>	<NA>	<NA>	<NA>	
5120	C-PA/7993	via acerbi 46	SERVIZI ALLA PERSONA	<NA>	nan	<NA>	<NA>	Centro massaggi	<NA>	<NA>	<NA>	<NA>	<NA>	
5121	C-PA/8005	via valenza n. 5	SERVIZI ALLA PERSONA	<NA>	nan	<NA>	<NA>	Centro massaggi	<NA>	<NA>	<NA>	<NA>	<NA>	
5122	C-PA/8015	varanini n. 12	SERVIZI ALLA PERSONA	<NA>	nan	<NA>	<NA>	Centro massaggi	<NA>	<NA>	<NA>	<NA>	<NA>	
5123	C-PA/8019	m. d'agrate n. 6	SERVIZI ALLA PERSONA	<NA>	nan	<NA>	<NA>	Centro massaggi	<NA>	<NA>	<NA>	<NA>	<NA>	
5124	C-PA/8024	Via PISTRUCCI SCIPIONE via pistrucci	SERVIZI ALLA PERSONA	Via PISTRUCCI SCIPIONE	nan	3215	<NA>	Centro massaggi	<NA>	<NA>	<NA>	<NA>	<NA>	
5125	C-PA/8025	via paolo sarpi n. 58	SERVIZI ALLA PERSONA	<NA>	nan	<NA>	<NA>	Centro massaggi	<NA>	<NA>	<NA>	<NA>	<NA>	
5126	C-PA/8031	via borgo pieve 108	SERVIZI ALLA PERSONA	<NA>	nan	<NA>	<NA>	Centro massaggi	<NA>	<NA>	<NA>	<NA>	<NA>	
5127	C-PA/8038	via varchi, 1 - milano	SERVIZI ALLA PERSONA	<NA>	nan	<NA>	<NA>	Centro massaggi	<NA>	<NA>	<NA>	<NA>	<NA>	
5128	C-PA/8043	via m. gioia, 43 - milano	SERVIZI ALLA PERSONA	<NA>	nan	<NA>	<NA>	Centro massaggi	<NA>	<NA>	<NA>	<NA>	<NA>	

Figure 4.2: Screenshot of some incomplete rows present in the dataset

The user is then able to drop all these rows from the dataset, in case s/he decides to do so, the new dataset will be automatically profiled again. An important remark to do is that this paradigm is present throughout the entire tool, this is because all the pages of the tool retrieve the data from the json report located in the tool's folder that should be as updated as possible.
Then is present the section "Single column" that focuses on one column at a time. Indeed for every column, the following data are shown: type, null, distinct, and unique values and their percentages with respect to the dataset's length.



Single column

Codice

Column type is: object

Null values: 0

Percentage of null values: 0.00%

Distinct values: 5200

Percentage of distinct values: 100.00%

Unique values: 5200

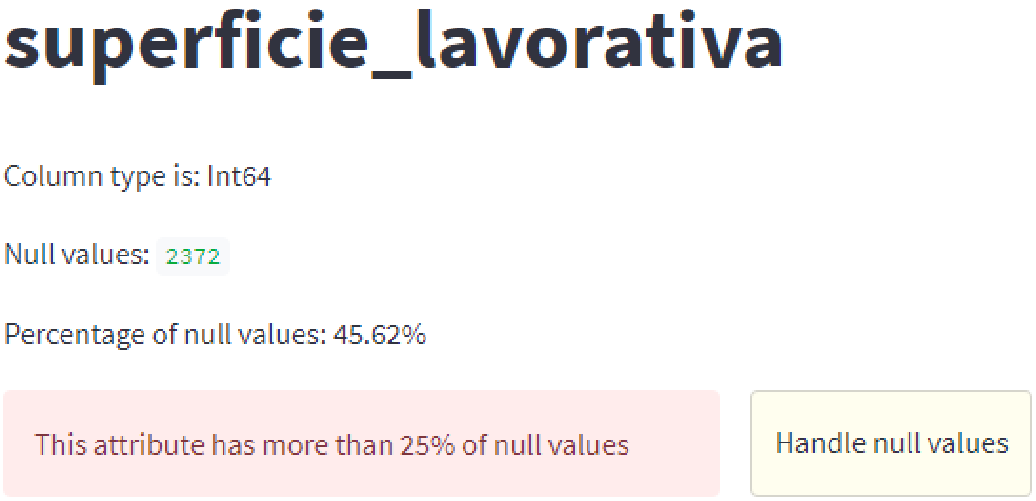
Percentage of unique values: 100.00%

This attribute is a possible candidate for primary key!

Figure 4.3: Screenshot on column 'Codice' in the Single column analysis section

Always within this section warnings and errors are prompted if some percentages are higher or lower than certain thresholds. The following threshold limits have been set:

- Null values
  If the percentage of null values is strictly higher than 25% a warning is displayed and the user is prompted to handle them in the null values page



superficie_lavorativa

Column type is: Int64

Null values: 2372

Percentage of null values: 45.62%

This attribute has more than 25% of null values

Figure 4.4: Example of warning if the number of null values exceeds the threshold

- Distinct values

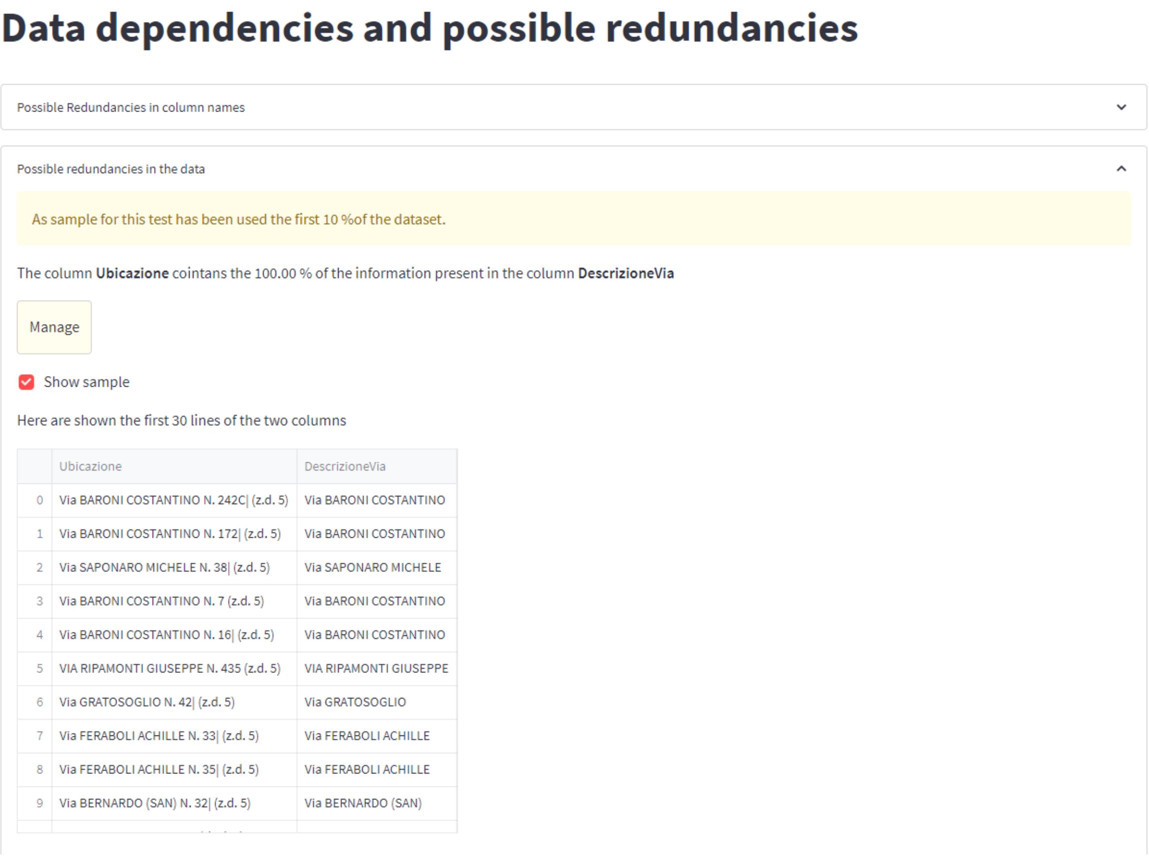
If the percentage of distinct values is strictly less than 4% a warning is displayed and the system advises the users that this attribute can be a possible category and prompts them to manage the categories in the proper page.

- Unique values
  If the percentage of unique values is strictly higher than 95% the user is just informed that the attribute is a possible candidate for the primary key of the dataset.

The next section is the Data dependencies and possible redundancies where the dataset is analyzed in 3 levels: column names, column values, and the correlation parameters between each column taken from the Phik correlation table.

- Redundancies in column names
  The system checks if a column's name is a substring of another one. Then, a couple of columns involved are shown to the users, moreover, if it is more than one, the system will show every pair. Will be then up to them to provide the final answer over the possible redundancy in the data too. Indeed even if the column name is a subset of another one, this does not ensure that also their data are redundant, for this reason, user involvement is required.
  The users are given the possibility to manage this redundancy on a specific page where they'll be able to drop one of the two columns.

- Redundancies in column data
  The first thing that should be remarked is that the pre-check that calculates this redundancy uses only the first 10% of the dataset. This is because the computation of this redundancy is really expensive and applying these checks to the entire dataset would require too much time.
  Anyway with the analyzed dataset, even with the first 10% only, the results were very good because all the dependencies were correctly captured. The threshold of this check has been arbitrarily set to 90%. To explain in other words, if given the first 10% of 2 columns (i.e., column 1 and column 2), if more than 90% of values in column 1 are a subset of the values in column 2, redundancy is detected and shown.



Data dependencies and possible redundancies

Possible Redundancies in column names
Possible redundancies in the data

As sample for this test has been used the first 10 %of the dataset.

The column **Ubicazione** cointans the 100.00 % of the information present in the column **DescrizioneVia**

☑ Show sample

Here are shown the first 30 lines of the two columns

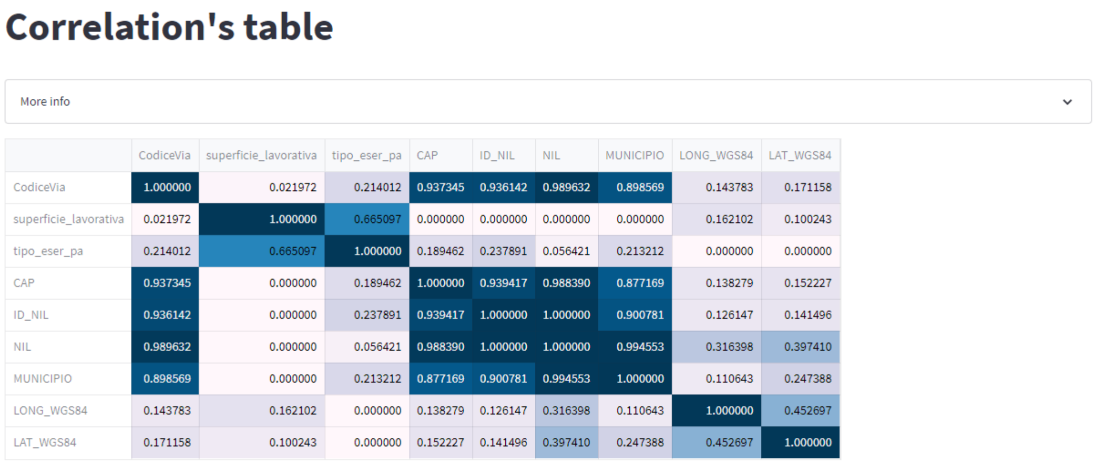
	Ubicazione	DescrizioneVia
0	Via BARONI COSTANTINO N. 242C| (z.d. 5)	Via BARONI COSTANTINO
1	Via BARONI COSTANTINO N. 172| (z.d. 5)	Via BARONI COSTANTINO
2	Via SAPONARO MICHELE N. 38| (z.d. 5)	Via SAPONARO MICHELE
3	Via BARONI COSTANTINO N. 7 (z.d. 5)	Via BARONI COSTANTINO
4	Via BARONI COSTANTINO N. 16| (z.d. 5)	Via BARONI COSTANTINO
5	VIA RIPAMONTI GIUSEPPE N. 435 (z.d. 5)	VIA RIPAMONTI GIUSEPPE
6	Via GRATOSOGLIO N. 42| (z.d. 5)	Via GRATOSOGLIO
7	Via FERABOLI ACHILLE N. 33| (z.d. 5)	Via FERABOLI ACHILLE
8	Via FERABOLI ACHILLE N. 35| (z.d. 5)	Via FERABOLI ACHILLE
9	Via BERNARDO (SAN) N. 32| (z.d. 5)	Via BERNARDO (SAN)

Figure 4.5: Redundancy captured by the system between column 'Ubicazione' and 'DescrizioneVia'

Also in this case, as visible in Figure 4.5, the user can manage this redundancy and has the possibility to drop one of the two columns.

- Correlation table

  By accessing the json file generated from the pandas profiling report, the system is able to load the Phik correlation table 4.6, which, as already mentioned, works and captures very well all correlations of the dataset between also different types of variables like categorical, ordinal and interval.



Correlation's table

	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO	LONG_WGS84	LAT_WGS84
CodiceVia	1.000000	0.021972	0.214012	0.937345	0.936142	0.989632	0.898569	0.143783	0.171158
superficie_lavorativa	0.021972	1.000000	0.665097	0.000000	0.000000	0.000000	0.000000	0.162102	0.100243
tipo_eser_pa	0.214012	0.665097	1.000000	0.189462	0.237891	0.056421	0.213212	0.000000	0.000000
CAP	0.937345	0.000000	0.189462	1.000000	0.939417	0.988390	0.877169	0.138279	0.152227
ID_NIL	0.936142	0.000000	0.237891	0.939417	1.000000	1.000000	0.900781	0.126147	0.141496
NIL	0.989632	0.000000	0.056421	0.988390	1.000000	1.000000	0.994553	0.316398	0.397410
MUNICIPIO	0.898569	0.000000	0.213212	0.877169	0.900781	0.994553	1.000000	0.110643	0.247388
LONG_WGS84	0.143783	0.162102	0.000000	0.138279	0.126147	0.316398	0.110643	1.000000	0.452697
LAT_WGS84	0.171158	0.100243	0.000000	0.152227	0.141496	0.397410	0.247388	0.452697	1.000000

Figure 4.6: Phik correlation table, enriched with a gradient visualization

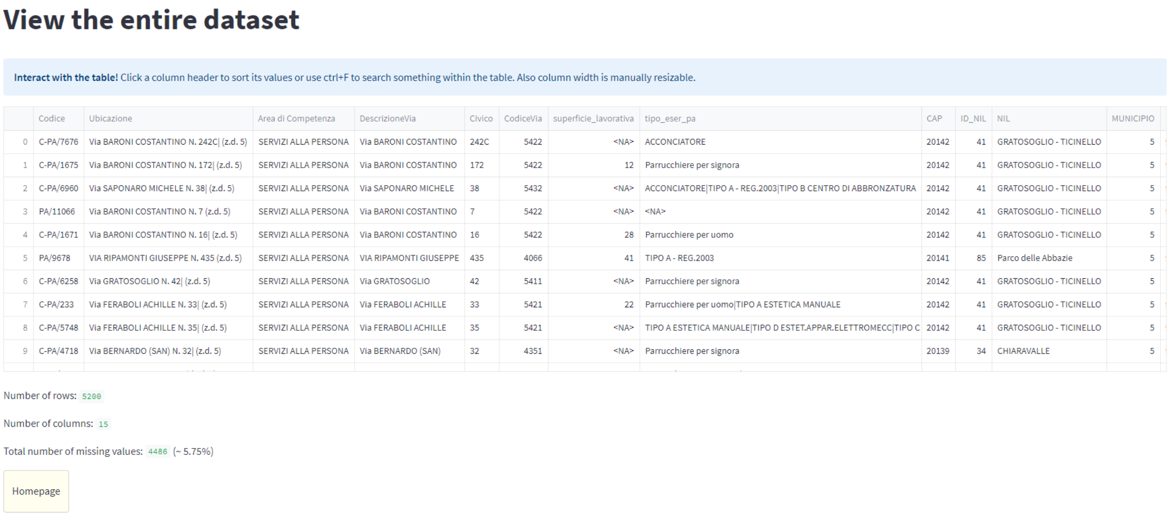
The result is that there are ten correlations across the entire dataset that are higher than 85%. The Phi_k matrix is symmetrical by definition as Corr[a][b] == Corr[b][a], ten is the number already divided by two. An important thing that should be remarked is that unfortunately, the computation of this table is very expensive (it is performed throughout the entire dataset, not just 10% like before). Indeed it takes around 70% of the total computation time of the profiler report.
For every detected correlation is available a 'Manage' button that will give, like in the two previous redundancy types, the possibility to drop one of the two columns involved. Moreover, within this page, the user can check the 2 entire columns and then decide autonomously what to do.

To conclude, the aim of the Dataset Info page was just to show the data to the user from a different perspective. It is moreover true that the user is given a lot of information, it will be then up to him how to exploit the various numbers and percentages, or the correlation and the redundancy present in his dataset in a logical way.

Show Entire Dataset

In this page is presented to the user the entire dataset. The function that is displaying it, the st.dataframe, allows the user to interact with the dataset: it is possible for example to search something within the table, like the crtl+F function, or to order the dataset based on a column by clicking its header. Moreover, some quick statistics of the dataset are shown like the number of rows, of columns, the total number of missing values, and their percentage with respect to the total number of values.
Unfortunately, Streamlit does not provide, at the moment, any functionality to do more,

like clicking directly a cell and retrieving the value and this is a huge limitation for this library.



View the entire dataset

Interact with the table! Click a column header to sort its values or use ctrl+F to search something within the table. Also column width is manually resizable.

	Codice	Ubicazione	Area di Competenza	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO
0	C-PA/7676	Via BARONI COSTANTINO N. 242C| (z.d. 5)	SERVIZI ALLA PERSONA	Via BARONI COSTANTINO	242C	5422	<NA>	ACCONCIATORE	20142	41	GRATOSOGLIO - TICINELLO	5
1	C-PA/1675	Via BARONI COSTANTINO N. 172| (z.d. 5)	SERVIZI ALLA PERSONA	Via BARONI COSTANTINO	172	5422	12	Parrucchiere per signora	20142	41	GRATOSOGLIO - TICINELLO	5
2	C-PA/6960	Via SAPONARO MICHELE N. 38| (z.d. 5)	SERVIZI ALLA PERSONA	Via SAPONARO MICHELE	38	5432	<NA>	ACCONCIATORE|TIPO A - REG.2003|TIPO B CENTRO DI ABBRONZATURA	20142	41	GRATOSOGLIO - TICINELLO	5
3	PA/11066	Via BARONI COSTANTINO N. 7 (z.d. 5)	SERVIZI ALLA PERSONA	Via BARONI COSTANTINO	7	5422	<NA>	<NA>	20142	41	GRATOSOGLIO - TICINELLO	5
4	C-PA/1671	Via BARONI COSTANTINO N. 16| (z.d. 5)	SERVIZI ALLA PERSONA	Via BARONI COSTANTINO	16	5422	28	Parrucchiere per uomo	20142	41	GRATOSOGLIO - TICINELLO	5
5	PA/9678	VIA RIPAMONTI GIUSEPPE N. 435 (z.d. 5)	SERVIZI ALLA PERSONA	VIA RIPAMONTI GIUSEPPE	435	4066	41	TIPO A - REG.2003	20141	85	Parco delle Abbazie	5
6	C-PA/6258	Via GRATOSOGLIO N. 42| (z.d. 5)	SERVIZI ALLA PERSONA	Via GRATOSOGLIO	42	5411	<NA>	Parrucchiere per signora	20142	41	GRATOSOGLIO - TICINELLO	5
7	C-PA/233	Via FERABOLI ACHILLE N. 33| (z.d. 5)	SERVIZI ALLA PERSONA	Via FERABOLI ACHILLE	33	5421	22	Parrucchiere per uomo|TIPO A ESTETICA MANUALE	20142	41	GRATOSOGLIO - TICINELLO	5
8	C-PA/5748	Via FERABOLI ACHILLE N. 35| (z.d. 5)	SERVIZI ALLA PERSONA	Via FERABOLI ACHILLE	35	5421	<NA>	TIPO A ESTETICA MANUALE|TIPO D ESTET.APPAR.ELETTROMECC|TIPO C	20142	41	GRATOSOGLIO - TICINELLO	5
9	C-PA/4718	Via BERNARDO (SAN) N. 32| (z.d. 5)	SERVIZI ALLA PERSONA	Via BERNARDO (SAN)	32	4351	<NA>	Parrucchiere per signora	20139	34	CHIARAVALLE	5

Number of rows: 5200

Number of columns: 15

Total number of missing values: 4486 (~ 5.75%)

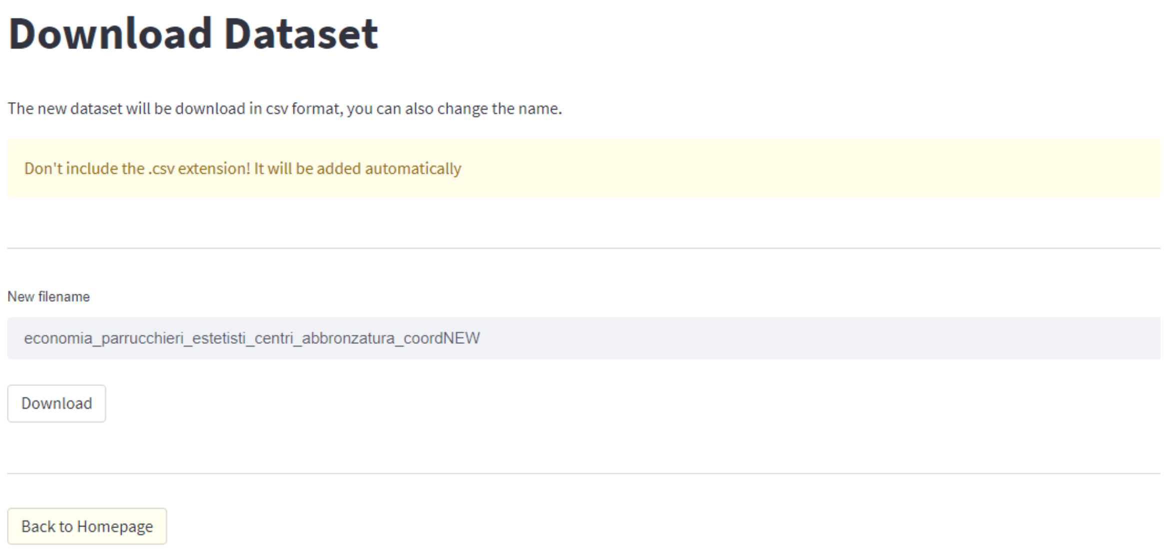
Figure 4.7: UI screen of the Show Entire Dataset page

Download Dataset

The Download Dataset page empowers the users to download their new profiled dataset. Both in the cleaning and in the wrangling section the users are able to transform the dataset and enrich its quality: all these functionalities will be almost useless without giving the users to download the new transformed dataset. The user will just have to select a name and the tool will automatically add the .csv extension just before starting the download.

It will be also possible in the future to upload in the tool this new file and continue the work.

It is important to remark that the tool provides automatically also the .json report from the pandas profiler. This will be already present within the installation folder of the tool and the tool itself will take care of updating it every time the profiling is launched again.



Download Dataset

The new dataset will be download in csv format, you can also change the name.

Don't include the .csv extension! It will be added automatically

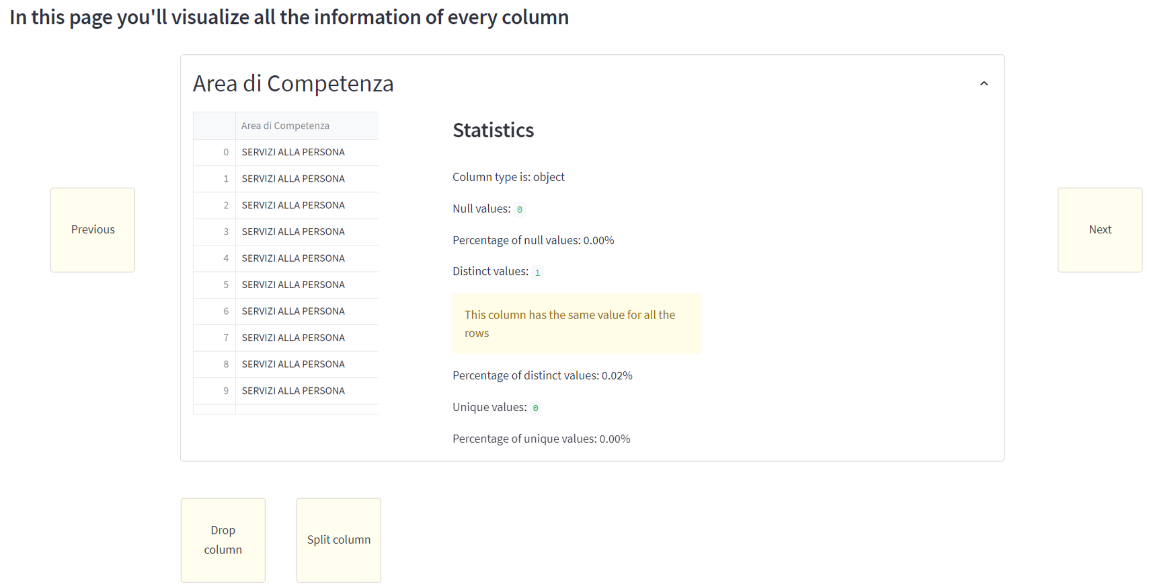
New filename

economia_parrucchieri_estetisti_centri_abbronzatura_coordNEW

Figure 4.8: UI screen of the Download Dataset page

Info By Column

This page was developed with the specific scope of making the single column analysis more readable and clear to the user. The interface is very simple and effective. As it is visible in the Figure below, as visible in the Figure 4.9, there are available some column's statistics like the null, distinct, unique values and their percentages.



In this page you'll visualize all the information of every column

Area di Competenza

	Area di Competenza
0	SERVIZI ALLA PERSONA
1	SERVIZI ALLA PERSONA
2	SERVIZI ALLA PERSONA
3	SERVIZI ALLA PERSONA
4	SERVIZI ALLA PERSONA
5	SERVIZI ALLA PERSONA
6	SERVIZI ALLA PERSONA
7	SERVIZI ALLA PERSONA
8	SERVIZI ALLA PERSONA
9	SERVIZI ALLA PERSONA

Statistics

Column type is: object

Null values: 0

Percentage of null values: 0.00%

Distinct values: 1

This column has the same value for all the rows

Percentage of distinct values: 0.02%

Unique values: 0

Percentage of unique values: 0.00%

Figure 4.9: Column "Area di competenza" displayed in the Info by column page

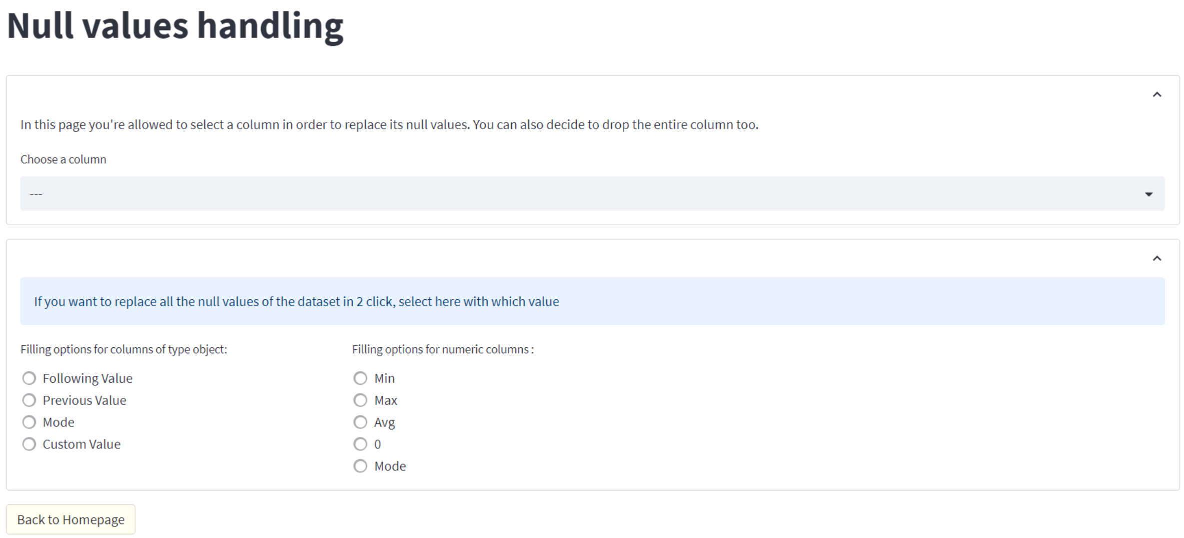
The buttons available at the bottom of the page change dynamically with respect to the column that is currently being shown. There are available 5 different buttons:

- Null values
  If the column has at least 1 null value. This button will redirect the user to the null values handling page 4.3.2 and the column will be automatically set to the one of Info by column.
- Drop column
  If the column, like the case shown in figure 4.9 has only one distinct value, so always the same value for each row, the system will give the possibility to drop the column, as this is not bringing any information to the dataset.
- Manage categories
  If the distinct values of the column are less than 4% out of the length of the dataset, with this button the user will be able to manage the values and apply a better categorization to the column.
- Split column
  If a delimiter, for the moment is only the space, is detected, the tool will make this button available to redirect to the Split column page **??**.
- Correlation with
  In this case, always at runtime, many buttons are created as the number of columns with which the current column is highly correlated. The data are always retrieved from the Phi_k correlation table. Within the button will be added the information with which column is actually correlated to the current one and with which percentage.

4.3.2. Data Cleaning

Null Values

This page is meant to replace null values or to drop a column if, for the user, the high number of null values is influencing too much the dataset's quality. It is divided into two sections, as can be seen in figure 4.10



Null values handling

In this page you're allowed to select a column in order to replace its null values. You can also decide to drop the entire column too.

Choose a column

If you want to replace all the null values of the dataset in 2 click, select here with which value

Filling options for columns of type object:

○ Following Value
○ Previous Value
○ Mode
○ Custom Value

Filling options for numeric columns :

○ Min
○ Max
○ Avg
○ 0
○ Mode

Figure 4.10: Null Values homepage

In the upper section is possible to select only one column at a time, instead in the lower section, as shown by the tool, the replacement will take place in every column of the dataset. This function has been implemented to speed up the filling process by letting the user select only 2 filling methods: one for the object columns and the other for the numeric columns. For the object columns, the filling method available are:

- Following Value
- Previous Value
- Mode
- Custom Value

while for the numeric columns, the filling method available are:

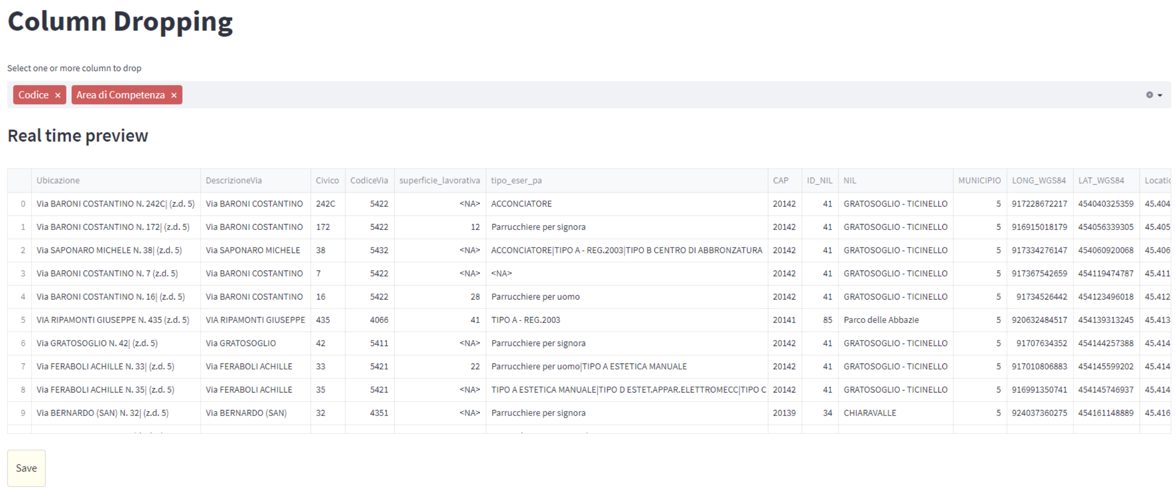
- Minimum value
- Maximum value
- Average value
- 0 value
- Mode value

Obviously, in case the user selects only one column with the appropriate box, he/she will be presented with the same possibilities above, depending on the type of the column chosen.

In the Data Cleaning Section, the user is required to manipulate the data. Anyway, given that some modifications can be performed by mistake, the tool always shows a preview of the new dataset. After the confirmation from the user, the dataset loaded in the tool will be permanently overwritten. Note that the original and loaded file won't be modified. The changes will be limited to the tool's context.

Column Dropping

This page has essentially 2 phases: the first is meant to show the preview of the new dataset while the second is for saving the new dataset. But given the fact this page is reachable both from the Homepage and the Info by Column page, it has slightly different UIs based on the source page. Indeed, if opened by the homepage the UI will be the following:



Column Dropping

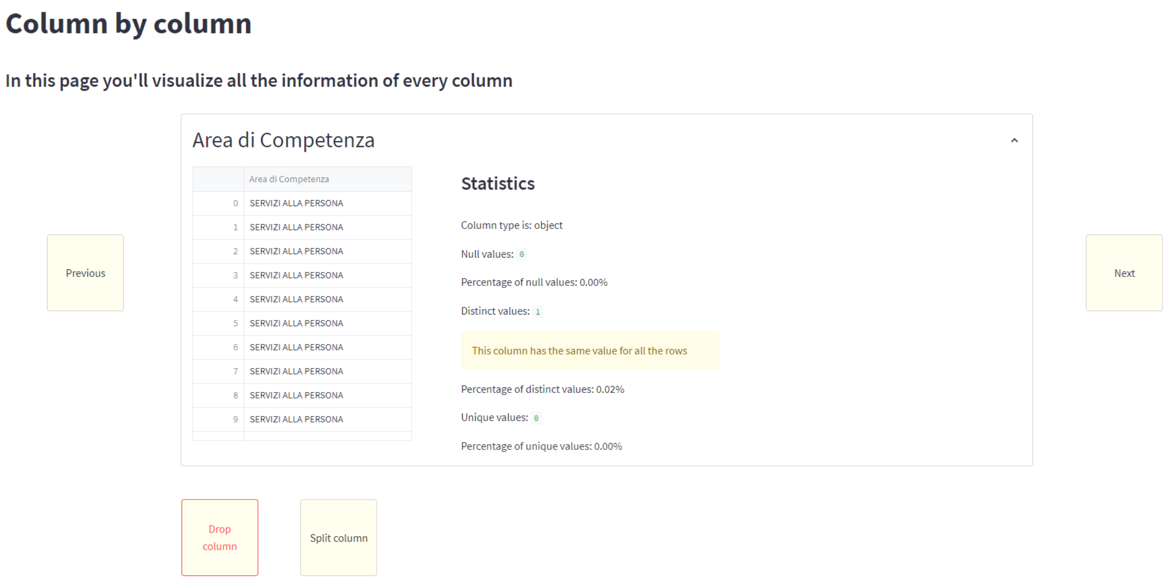
Select one or more column to drop

Real time preview

	Ubicazione	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO	LONG_WGS84	LAT_WGS84	Locatic
0	Via BARONI COSTANTINO N. 242C| (z.d. 5)	Via BARONI COSTANTINO	242C	5422	<NA>	ACCONCIATORE	20142	41	GRATOSOGLIO - TICINELLO	5	917228672217	454040325359	45.404
1	Via BARONI COSTANTINO N. 172| (z.d. 5)	Via BARONI COSTANTINO	172	5422	12	Parrucchiere per signora	20142	41	GRATOSOGLIO - TICINELLO	5	916915018179	454056339305	45.405
2	Via SAPONARO MICHELE N. 38| (z.d. 5)	Via SAPONARO MICHELE	38	5432	<NA>	ACCONCIATORE|TIPO A - REG.2003|TIPO B CENTRO DI ABBRONZATURA	20142	41	GRATOSOGLIO - TICINELLO	5	917334276147	454060920068	45.406
3	Via BARONI COSTANTINO N. 7 (z.d. 5)	Via BARONI COSTANTINO	7	5422	<NA>	<NA>	20142	41	GRATOSOGLIO - TICINELLO	5	917367542659	454119474787	45.411
4	Via BARONI COSTANTINO N. 16| (z.d. 5)	Via BARONI COSTANTINO	16	5422	28	Parrucchiere per uomo	20142	41	GRATOSOGLIO - TICINELLO	5	91734526442	454123496018	45.412
5	VIA RIPAMONTI GIUSEPPE N. 435 (z.d. 5)	VIA RIPAMONTI GIUSEPPE	435	4066	41	TIPO A - REG.2003	20141	85	Parco delle Abbazie	5	920632484517	454139313245	45.413
6	Via GRATOSOGLIO N. 42| (z.d. 5)	Via GRATOSOGLIO	42	5411	<NA>	Parrucchiere per signora	20142	41	GRATOSOGLIO - TICINELLO	5	91707634352	454144257388	45.414
7	Via FERABOLI ACHILLE N. 33| (z.d. 5)	Via FERABOLI ACHILLE	33	5421	22	Parrucchiere per uomo|TIPO A ESTETICA MANUALE	20142	41	GRATOSOGLIO - TICINELLO	5	917010806883	454145599202	45.414
8	Via FERABOLI ACHILLE N. 35| (z.d. 5)	Via FERABOLI ACHILLE	35	5421	<NA>	TIPO A ESTETICA MANUALE|TIPO D ESTET.APPAR.ELETTROMECC|TIPO C	20142	41	GRATOSOGLIO - TICINELLO	5	916991350741	454145746937	45.414
9	Via BERNARDO (SAN) N. 32| (z.d. 5)	Via BERNARDO (SAN)	32	4351	<NA>	Parrucchiere per signora	20139	34	CHIARAVALLE	5	924037360275	454161148889	45.416

Figure 4.11: Column dropping page if reached from the homepage

As it is visible in Figure 4.11, it is possible to select one or more columns to be dropped, and a real-time preview of the new dataset is displayed. Instead, if the page is reached from the Info By Column page, with the button "Drop column" available in "Area di competenza", which has the same values for all the rows, the UI will be different, according to the use case. The first Figure 4.12 is the source page while the second figure 4.13 shows how the page works if reached from the Info By Column page.



Column by column

In this page you'll visualize all the information of every column

Area di Competenza

	Area di Competenza
0	SERVIZI ALLA PERSONA
1	SERVIZI ALLA PERSONA
2	SERVIZI ALLA PERSONA
3	SERVIZI ALLA PERSONA
4	SERVIZI ALLA PERSONA
5	SERVIZI ALLA PERSONA
6	SERVIZI ALLA PERSONA
7	SERVIZI ALLA PERSONA
8	SERVIZI ALLA PERSONA
9	SERVIZI ALLA PERSONA

Statistics

Column type is: object

Null values: 0

Percentage of null values: 0.00%

Distinct values: 1

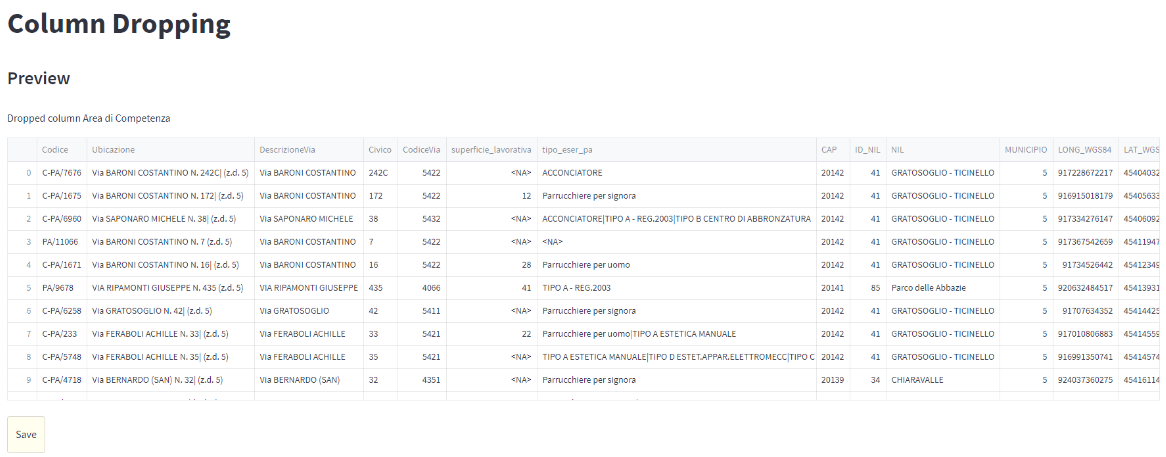
This column has the same value for all the rows

Percentage of distinct values: 0.02%

Unique values: 0

Percentage of unique values: 0.00%

Figure 4.12: One of the possible source page, Info by Column and it is inspected a column with the same value for every row. So, Dropping is available



Column Dropping

Preview

Dropped column Area di Competenza

	Codice	Ubicazione	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO	LONG_WGS84	LAT_WGS
0	C-PA/7676	Via BARONI COSTANTINO N. 242C| (z.d. 5)	Via BARONI COSTANTINO	242C	5422	<NA>	ACCONCIATORE	20142	41	GRATOSOGLIO - TICINELLO	5	917228672217	45404032
1	C-PA/1675	Via BARONI COSTANTINO N. 172| (z.d. 5)	Via BARONI COSTANTINO	172	5422	12	Parrucchiere per signora	20142	41	GRATOSOGLIO - TICINELLO	5	916915018179	45405633
2	C-PA/6960	Via SAPONARO MICHELE N. 38| (z.d. 5)	Via SAPONARO MICHELE	38	5432	<NA>	ACCONCIATORE|TIPO A - REG.2003|TIPO B CENTRO DI ABBRONZATURA	20142	41	GRATOSOGLIO - TICINELLO	5	917334276147	45406092
3	PA/11066	Via BARONI COSTANTINO N. 7 (z.d. 5)	Via BARONI COSTANTINO	7	5422	<NA>	<NA>	20142	41	GRATOSOGLIO - TICINELLO	5	917367542659	45411947
4	C-PA/1671	Via BARONI COSTANTINO N. 16| (z.d. 5)	Via BARONI COSTANTINO	16	5422	28	Parrucchiere per uomo	20142	41	GRATOSOGLIO - TICINELLO	5	91734526442	45412349
5	PA/9678	VIA RIPAMONTI GIUSEPPE N. 435 (z.d. 5)	VIA RIPAMONTI GIUSEPPE	435	4066	41	TIPO A - REG.2003	20141	85	Parco delle Abbazie	5	920632484517	45413931
6	C-PA/6258	Via GRATOSOGLIO N. 42| (z.d. 5)	Via GRATOSOGLIO	42	5411	<NA>	Parrucchiere per signora	20142	41	GRATOSOGLIO - TICINELLO	5	91707634352	45414425
7	C-PA/233	Via FERABOLI ACHILLE N. 33| (z.d. 5)	Via FERABOLI ACHILLE	33	5421	22	Parrucchiere per uomo|TIPO A ESTETICA MANUALE	20142	41	GRATOSOGLIO - TICINELLO	5	917010806883	45414559
8	C-PA/5748	Via FERABOLI ACHILLE N. 35| (z.d. 5)	Via FERABOLI ACHILLE	35	5421	<NA>	TIPO A ESTETICA MANUALE|TIPO D ESTET.APPAR.ELETTROMECC|TIPO C	20142	41	GRATOSOGLIO - TICINELLO	5	916991350741	45414574
9	C-PA/4718	Via BERNARDO (SAN) N. 32| (z.d. 5)	Via BERNARDO (SAN)	32	4351	<NA>	Parrucchiere per signora	20139	34	CHIARAVALLE	5	924037360275	45416114

Figure 4.13: Column dropping page if reached from the Info by Column page, specifically column Area di Competenza

As the column/s to drop is/are chosen, when the Save button is pressed, the new dataset will be profiled and will overwrite the dataset within the tool.

Suggested Actions

The suggested actions page was intentionally developed with the aim of providing the user with just a list of actions proposed by the tool with the aim of increasing the quality of the dataset. By scanning the whole dataset the system will look for different characteristics:

- Completeness of the rows
  If a row has at least 40% of the attributes equal to null it will be shown in the "Incomplete rows" Section. The user has then the possibility to drop all of them.

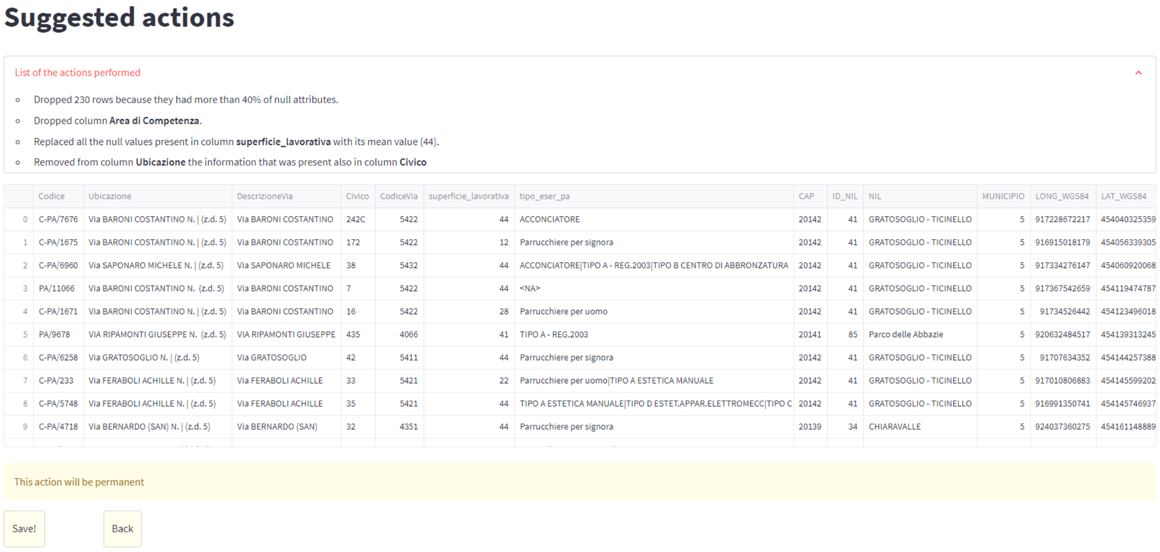
- Correlated columns
  Always by retrieving the data from the Phi_k correlation's matrix, the tool highlights to the user all the correlations that are higher than a threshold of 85%, the actions made available by the tool are to drop one or the other column of the correlation. By doing some computation on the Phi_k matrix, the tool is also able to suggest which column between the 2 is better to drop. The logic behind that is to suggest the column that has the highest sum of all the correlation coefficients, To explain it with a trivial example, column A and column B have a correlation coefficient of 85%. The dataset has in total 3 columns, the correlation between column A and column C is 60% while the correlation between column B and column C is 90%. So between columns A and B, it is better to drop column B, as column C, which will be kept in the dataset, was the "most similar" to column B. So, in the mandatory choice of dropping one column, the tool drops the column that makes the dataset lose the least amount of information possible.

- Useless columns
  If a column has a unique value for every row, the system will suggest to drop it since it is not bringing any additional information to the dataset and is neither characterizing it.

- Null values
  If a column has more than 25% of null values is inserted within this section. The tool's behavior will change, depending on the column's type. In case the column is of Numeric type, it will be given the possibility, apart from dropping the column, to replace all the null values with the mean value. Otherwise, if the column is of type Object, the value proposed for replacing the null values will be the mode value of the column.

- Redundancies in the data
  The system by scanning the dataset will try to detect if, between two column values, some data redundancy is present. If detected, it will be shown a preview of the redundancy and then the user will be able to choose between dropping one of the two columns. In addition, the user has the possibility to remove the information from the column which contains more information.

All the actions chosen are collected at runtime in an array, that stores some keywords and the columns involved. If the user continues, this array will be then parsed and all the actions will be applied to a copy of the dataset. This copied and modified dataset will be then shown to the user. If s/he accepts the changes and decides to save, the new dataset will be profiled again and all the changes will be permanently applied.



Suggested actions

List of the actions performed

- Dropped 230 rows because they had more than 40% of null attributes.
- Dropped column **Area di Competenza**.
- Replaced all the null values present in column **superficie_lavorativa** with its mean value (44).
- Removed from column **Ubicazione** the information that was present also in column **Civico**

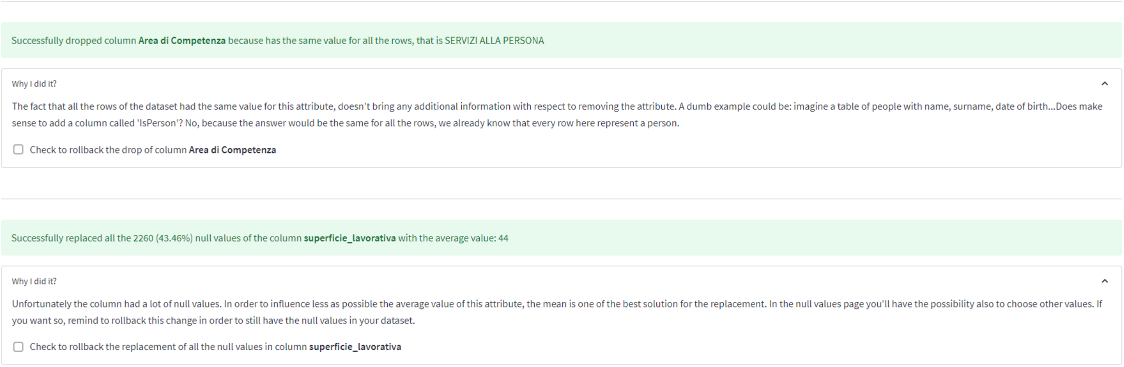
	Codice	Ubicazione	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO	LONG_WGS84	LAT_WGS84
0	C-PA/7676	Via BARONI COSTANTINO N. | (z.d. 5)	Via BARONI COSTANTINO	242C	5422	44	ACCONCIATORE	20142	41	GRATOSOGLIO - TICINELLO	5	917228672217	454040325359
1	C-PA/1675	Via BARONI COSTANTINO N. | (z.d. 5)	Via BARONI COSTANTINO	172	5422	12	Parrucchiere per signora	20142	41	GRATOSOGLIO - TICINELLO	5	916915018179	454056339305
2	C-PA/6960	Via SAPONARO MICHELE N. | (z.d. 5)	Via SAPONARO MICHELE	38	5432	44	ACCONCIATORE|TIPO A - REG.2003|TIPO B CENTRO DI ABBRONZATURA	20142	41	GRATOSOGLIO - TICINELLO	5	917334276147	454060920068
3	PA/11066	Via BARONI COSTANTINO N. (z.d. 5)	Via BARONI COSTANTINO	7	5422	44	<NA>	20142	41	GRATOSOGLIO - TICINELLO	5	917367542659	454119474787
4	C-PA/1671	Via BARONI COSTANTINO N. | (z.d. 5)	Via BARONI COSTANTINO	16	5422	28	Parrucchiere per uomo	20142	41	GRATOSOGLIO - TICINELLO	5	91734526442	454123496018
5	PA/9678	VIA RIPAMONTI GIUSEPPE N. (z.d. 5)	VIA RIPAMONTI GIUSEPPE	435	4066	41	TIPO A - REG.2003	20141	85	Parco delle Abbazie	5	920632484517	454139313245
6	C-PA/6258	Via GRATOSOGLIO N. | (z.d. 5)	Via GRATOSOGLIO	42	5411	44	Parrucchiere per signora	20142	41	GRATOSOGLIO - TICINELLO	5	91707634352	454144257388
7	C-PA/233	Via FERABOLI ACHILLE N. | (z.d. 5)	Via FERABOLI ACHILLE	33	5421	22	Parrucchiere per uomo|TIPO A ESTETICA MANUALE	20142	41	GRATOSOGLIO - TICINELLO	5	917010806883	454145599202
8	C-PA/5748	Via FERABOLI ACHILLE N. | (z.d. 5)	Via FERABOLI ACHILLE	35	5421	44	TIPO A ESTETICA MANUALE|TIPO D ESTET.APPAR.ELETTROMECC|TIPO C	20142	41	GRATOSOGLIO - TICINELLO	5	916991350741	454145746937
9	C-PA/4718	Via BERNARDO (SAN) N. | (z.d. 5)	Via BERNARDO (SAN)	32	4351	44	Parrucchiere per signora	20139	34	CHIARAVALLE	5	924037360275	454161148889

This action will be permanent

Figure 4.14: Preview of the modified dataset. This is the second state of the page.

Automatic Cleaning

The automatic cleaning has the same potential as the suggested actions page from an outcome point of view. The big difference is that this page applies automatically all the suggested actions. It is true that this page works in a very proactive way, but the aim of this development is for users that maybe do not have a deep knowledge of their dataset and want to completely rely on the tool's capacity to clean it. It can happen that the user does not want to perform every action one by one and would like to speed up the entire process. An important remark is that only a copy of the dataset is modified. This will allow the user to see a preview of the new dataset, without modifying the original one. Moreover, a powerful functionality of this page is that it is possible to perform the rollback of every action one by one, just by checking a box. This will update the dataset automatically, without the need of refreshing the page.



Successfully dropped column **Area di Competenza** because has the same value for all the rows, that is SERVIZI ALLA PERSONA

Why I did it?

The fact that all the rows of the dataset had the same value for this attribute, doesn't bring any additional information with respect to removing the attribute. A dumb example could be: imagine a table of people with name, surname, date of birth...Does make sense to add a column called 'IsPerson'? No, because the answer would be the same for all the rows, we already know that every row here represent a person.

☐ Check to rollback the drop of column **Area di Competenza**

Successfully replaced all the 2260 (43.46%) null values of the column **superficie_lavorativa** with the average value: 44

Why I did it?

Unfortunately the column had a lot of null values. In order to influence less as possible the average value of this attribute, the mean is one of the best solution for the replacement. In the null values page you'll have the possibility also to choose other values. If you want so, remind to rollback this change in order to still have the null values in your dataset.

☐ Check to rollback the replacement of all the null values in column **superficie_lavorativa**

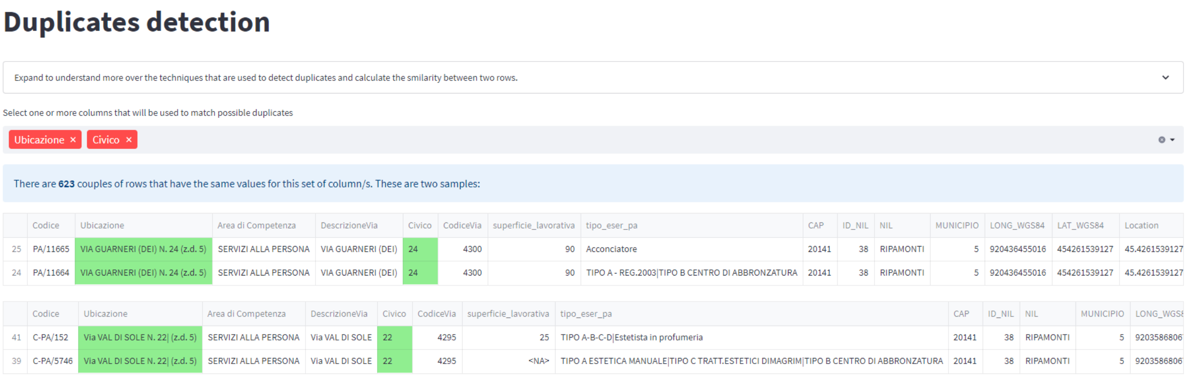
Figure 4.15: Example of how the user can rollback the actions.

After all the actions have been performed the user is able to save the dataset, but in case of need, s/he is also able to save and switch directly to the "Drop Column" page. This function was developed due to the removal of the redundancy of information that is applied automatically throughout all the columns: the tool was not able to recognize autonomously columns that were not adding any information to the dataset. To solve this issue, the Save and go to the Drop button has been implemented .

Duplicate Detection

Before explaining how the page works, it should be remarked that the theoretical process of duplicate detection, which fulfills the question "how does the tool spot two rows that could eventually be similar?", has been explained in Section 3.3. To perform this, the tool uses the blocking technique. To explain briefly what the blocking technique is, a block is composed of two tuples that have a certain set of attributes "blocked" to be equal.

The tool gives freedom on this to the users: it is up to them to choose the best set of blocking attributes. Anyway, the tool partially supports the user in doing this: firstly, it returns the number of couples that respect the blocking constraint and moreover, it shows a preview of the two tuples. This is done in order to make the user more aware as possible of how the blocking works.



Duplicates detection

Expand to understand more over the techniques that are used to detect duplicates and calculate the similarity between two rows.

Select one or more columns that will be used to match possible duplicates

There are **623** couples of rows that have the same values for this set of column/s. These are two samples:

	Codice	Ubicazione	Area di Competenza	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO	LONG_WGS84	LAT_WGS84	Location
25	PA/11665	VIA GUARNERI (DEI) N. 24 (z.d. 5)	SERVIZI ALLA PERSONA	VIA GUARNERI (DEI)	24	4300	90	Acconciatore	20141	38	RIPAMONTI	5	920436455016	454261539127	45.4261539127
24	PA/11664	VIA GUARNERI (DEI) N. 24 (z.d. 5)	SERVIZI ALLA PERSONA	VIA GUARNERI (DEI)	24	4300	90	TIPO A - REG.2003|TIPO B CENTRO DI ABBRONZATURA	20141	38	RIPAMONTI	5	920436455016	454261539127	45.4261539127

	Codice	Ubicazione	Area di Competenza	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO	LONG_WGS8
41	C-PA/152	Via VAL DI SOLE N. 22| (z.d. 5)	SERVIZI ALLA PERSONA	Via VAL DI SOLE	22	4295	25	TIPO A-B-C-D|Estetista in profumeria	20141	38	RIPAMONTI	5	9203586806
39	C-PA/5746	Via VAL DI SOLE N. 22| (z.d. 5)	SERVIZI ALLA PERSONA	Via VAL DI SOLE	22	4295	<NA>	TIPO A ESTETICA MANUALE|TIPO C TRATT.ESTETICI DIMAGRIM|TIPO B CENTRO DI ABBRONZATURA	20141	38	RIPAMONTI	5	9203586806

Figure 4.16: How the blocking on "Ubicazione" and "Civico" works.

In Figure 4.16, there are 623 couples of rows that have the same values for attributes "Ubicazione" and "Civico" and the similarity will be calculated only for these couples.
Here it is important to make the second regression about "how the tool calculates the similarity between two rows". In order to calculate this parameter it is used the string distance Jaro-Winkler.
This string-based distance is calculated between two values of the same attribute, one per row, but not for all the attributes. Some of the attributes are automatically removed by the tool:

- The attributes that are being used for blocking
  These attributes will have for the construction of the Jaro-Winkler similarity the parameter equal to one. The tool is already exploiting this information by investigating only the couples that have these attributes equal and it would make no sense to exploit this information again.

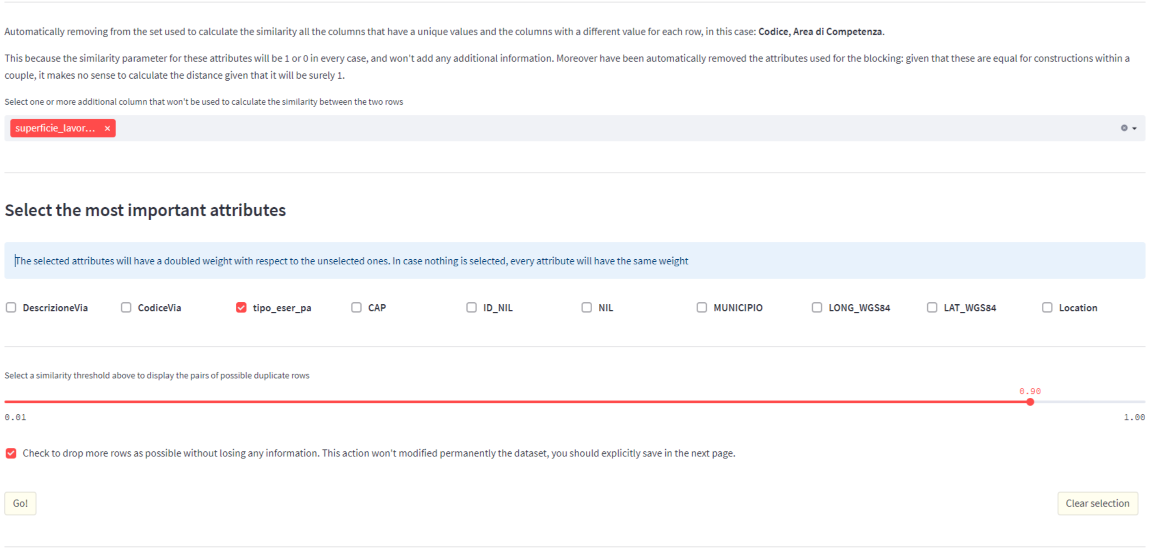
- The attributes that are unique within the dataset
  The uniqueness of an attribute is something that is possible to know in advance. The tool exploits this information by removing the attributes that have a uniqueness equal to 1 (so the value is different for every row) from the set on which the similarity will be calculated as it will always return a zero as result of the string's distance.

- The attributes that are always the same within the dataset
  The tool removes also the columns that have only one distinct value throughout all the dataset. This means that a value is constantly repeated for every row. In this case, the string's distance will be always equal to one, and this does not add any additional information as it is equal for every couple.

- Every other attribute is explicitly chosen by the user

Apart from the ones that are removed automatically, the user is also free to remove any other attribute from the set meant for the computation of the similarity, using a multi select box.

The system will then perform a weighted average over the remaining attributes to calculate the similarity of every couple of rows. Another configuration that the users can perform is to assign weights to the remaining attributes: by default every attribute has a weight equal to one, anyway if its box is selected it will have a weight equal to two within the weighted average. With the slider, they should select a similarity threshold to display the couples. The couples below the threshold won't be displayed on the next page. Finally, if the users check the box to drop more rows as possible the system will surely drop one row of the couple if the similarity parameter is equal to 1. Moreover, the system will automatically drop one of the rows if the intersection of two rows does not have any conflict except the null values. To explain in other words, row 1(A, B, C) and row 2 (A, None, C) does not have the similarity parameter equal to 1 but the tool will drop row 2 because it is representing the same instance of row 1 in a worse way.



Automatically removing from the set used to calculate the similarity all the columns that have a unique values and the columns with a different value for each row, in this case: **Codice, Area di Competenza**.

This because the similarity parameter for these attributes will be 1 or 0 in every case, and won't add any additional information. Moreover have been automatically removed the attributes used for the blocking: given that these are equal for constructions within a couple, it makes no sense to calculate the distance given that it will be surely 1.

Select one or more additional column that won't be used to calculate the similarity between the two rows

superficie_lavor... ×

Select the most important attributes

The selected attributes will have a doubled weight with respect to the unselected ones. In case nothing is selected, every attribute will have the same weight

□ DescrizioneVia □ CodiceVia ☑ tipo_eser_pa □ CAP □ ID_NIL □ NIL □ MUNICIPIO □ LONG_WGS84 □ LAT_WGS84 □ Location

Select a similarity threshold above to display the pairs of possible duplicate rows

0.90

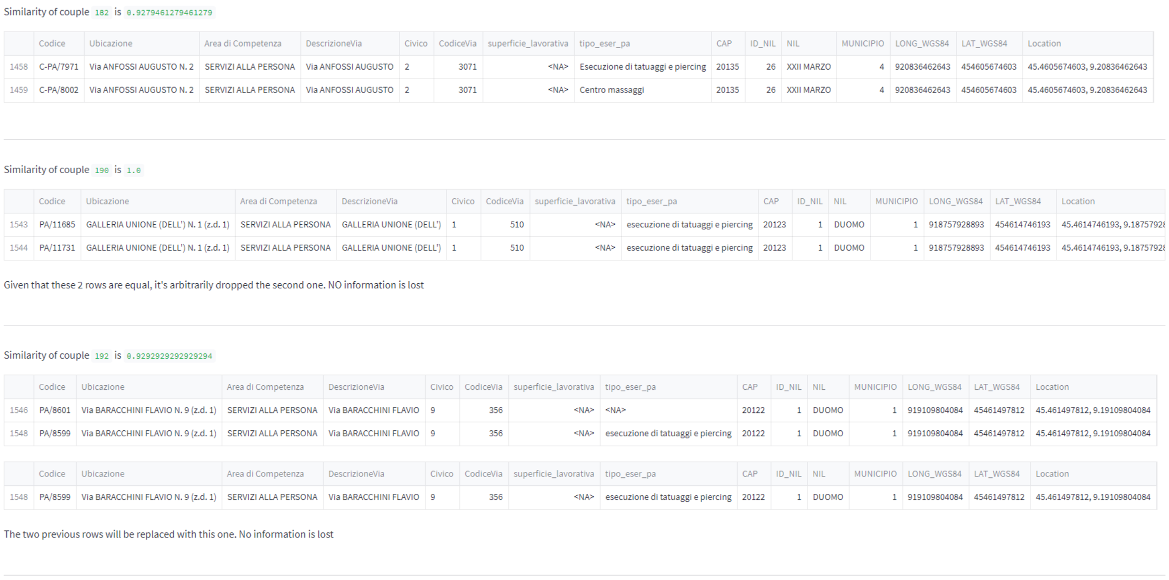
0.01 1.00

☑ Check to drop more rows as possible without losing any information. This action won't modified permanently the dataset, you should explicitly save in the next page.

Figure 4.17: Setting the weights and threshold for the similarity's computation.

An issue that should be highlighted within this page unfortunately given by architectural limits of Streamlit is that the computation takes too much time. This library is not meant for developing heavy pages full of content, and the displaying of a huge amount of couples is expensive from the library's point of view. If the threshold is too low, the system will warn that the page's computation can take up to some minutes.

A concrete use case of the duplicate detection process is visible in the Figure 4.18:



Similarity of couple 182 is 0.9279461279461279

	Codice	Ubicazione	Area di Competenza	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO	LONG_WGS84	LAT_WGS84	Location
1458	C-PA/7971	Via ANFOSSI AUGUSTO N. 2	SERVIZI ALLA PERSONA	Via ANFOSSI AUGUSTO	2	3071	<NA>	Esecuzione di tatuaggi e piercing	20135	26	XXII MARZO	4	920836462643	454605674603	45.4605674603, 9.20836462643
1459	C-PA/8002	Via ANFOSSI AUGUSTO N. 2	SERVIZI ALLA PERSONA	Via ANFOSSI AUGUSTO	2	3071	<NA>	Centro massaggi	20135	26	XXII MARZO	4	920836462643	454605674603	45.4605674603, 9.20836462643

Similarity of couple 190 is 1.0

	Codice	Ubicazione	Area di Competenza	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO	LONG_WGS84	LAT_WGS84	Location
1543	PA/11685	GALLERIA UNIONE (DELL') N. 1 (z.d. 1)	SERVIZI ALLA PERSONA	GALLERIA UNIONE (DELL')	1	510	<NA>	esecuzione di tatuaggi e piercing	20123	1	DUOMO	1	918757928893	454614746193	45.4614746193, 9.1875792
1544	PA/11731	GALLERIA UNIONE (DELL') N. 1 (z.d. 1)	SERVIZI ALLA PERSONA	GALLERIA UNIONE (DELL')	1	510	<NA>	esecuzione di tatuaggi e piercing	20123	1	DUOMO	1	918757928893	454614746193	45.4614746193, 9.1875792

Given that these 2 rows are equal, it's arbitrarily dropped the second one. NO information is lost

Similarity of couple 192 is 0.9292929292929294

	Codice	Ubicazione	Area di Competenza	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO	LONG_WGS84	LAT_WGS84	Location
1546	PA/8601	Via BARACCHINI FLAVIO N. 9 (z.d. 1)	SERVIZI ALLA PERSONA	Via BARACCHINI FLAVIO	9	356	<NA>	<NA>	20122	1	DUOMO	1	919109804084	45461497812	45.461497812, 9.19109804084
1548	PA/8599	Via BARACCHINI FLAVIO N. 9 (z.d. 1)	SERVIZI ALLA PERSONA	Via BARACCHINI FLAVIO	9	356	<NA>	esecuzione di tatuaggi e piercing	20122	1	DUOMO	1	919109804084	45461497812	45.461497812, 9.19109804084

	Codice	Ubicazione	Area di Competenza	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO	LONG_WGS84	LAT_WGS84	Location
1548	PA/8599	Via BARACCHINI FLAVIO N. 9 (z.d. 1)	SERVIZI ALLA PERSONA	Via BARACCHINI FLAVIO	9	356	<NA>	esecuzione di tatuaggi e piercing	20122	1	DUOMO	1	919109804084	45461497812	45.461497812, 9.19109804084

The two previous rows will be replaced with this one. No information is lost

Figure 4.18: The 3 different outcomes of a similarity comparison.

- Couple 182
  Between these two rows, nothing is dropped because there is a conflict on "tipo_eser_pa" and in order to avoid any possible data loss, the system does not do anything.
- Couple 190
  In this case, every attribute belonging to the similarity set is equal throughout the two rows, given as a result of similarity 1. Arbitrarily the second row is dropped.
- Couple 192
  In this case, the two rows do not have a similarity parameter equal to 1. The only different value is the one belonging to the column "tipo_eser_pa": given that in one of the two rows is equal to null the tool drops the first one, as can be fully replaced by the second one.

After the computation is completed, the user is presented with the list of similar couples and the user is able to visualize the preview of the new dataset, remark that the original one hasn't been not overwritten yet: this happens only if the user save it.

4.3.3. Data Wrangling

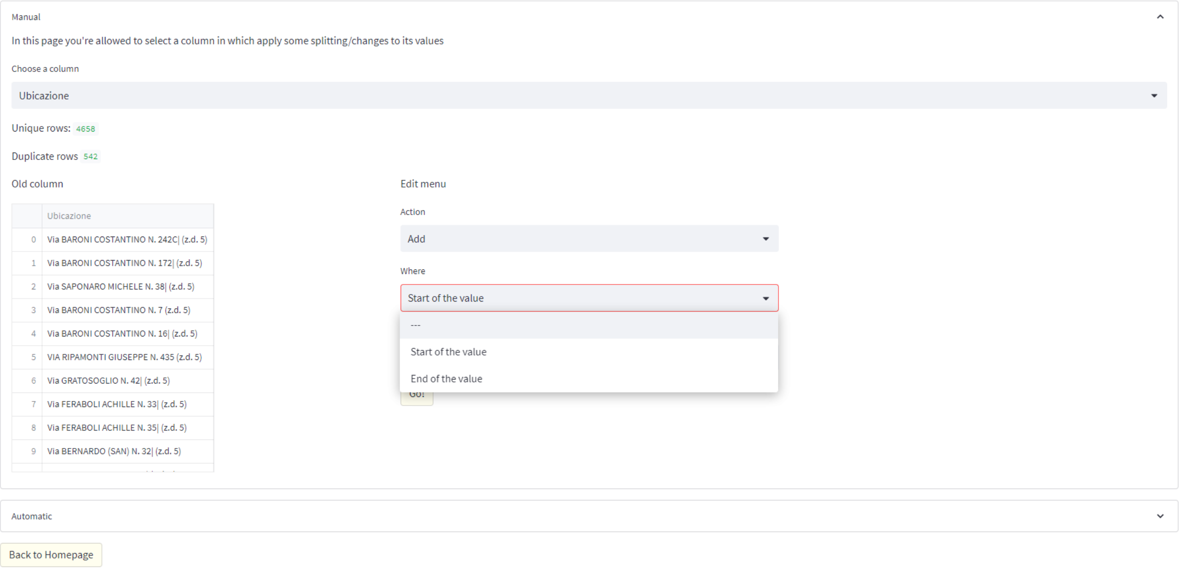
Column Renaming

The page has one selectbox and one input box: with the first, the user selects the column to be renamed, while in the input box, s/he types the name of the new column. A preview of the new dataset is then shown and the user can save and profile it again.

Values Editing

The Values Editing page is meant to edit the content of an entire column at a time. The page is divided in two sections: Manual and Automatic. As the names suggest, the automatic mode proactively proposes some changes to the user, avoiding most of the actions, but with the disadvantage of giving him a bit less flexibility than the manual one.

- Manual
  With the manual mode the user can choose between two actions: Remove or Add. If Remove is selected s/he'll be then asked to type the string to be removed. After clicking the "Go" button the tool will look for that string within the entire column. Every time the string is found it will be removed, if not found the value won't be modified at all. Otherwise, in case the action selected is Add the user can furthermore choose if add the typed string at the start or at the end of the value.



Manual

In this page you're allowed to select a column in which apply some splitting/changes to its values

Choose a column: Ubicazione

Unique rows: 4658

Duplicate rows 542

Old column

	Ubicazione
0	Via BARONI COSTANTINO N. 242C| (z.d. 5)
1	Via BARONI COSTANTINO N. 172| (z.d. 5)
2	Via SAPONARO MICHELE N. 38| (z.d. 5)
3	Via BARONI COSTANTINO N. 7 (z.d. 5)
4	Via BARONI COSTANTINO N. 16| (z.d. 5)
5	VIA RIPAMONTI GIUSEPPE N. 435 (z.d. 5)
6	Via GRATOSOGLIO N. 42| (z.d. 5)
7	Via FERABOLI ACHILLE N. 33| (z.d. 5)
8	Via FERABOLI ACHILLE N. 35| (z.d. 5)
9	Via BERNARDO (SAN) N. 32| (z.d. 5)

Edit menu

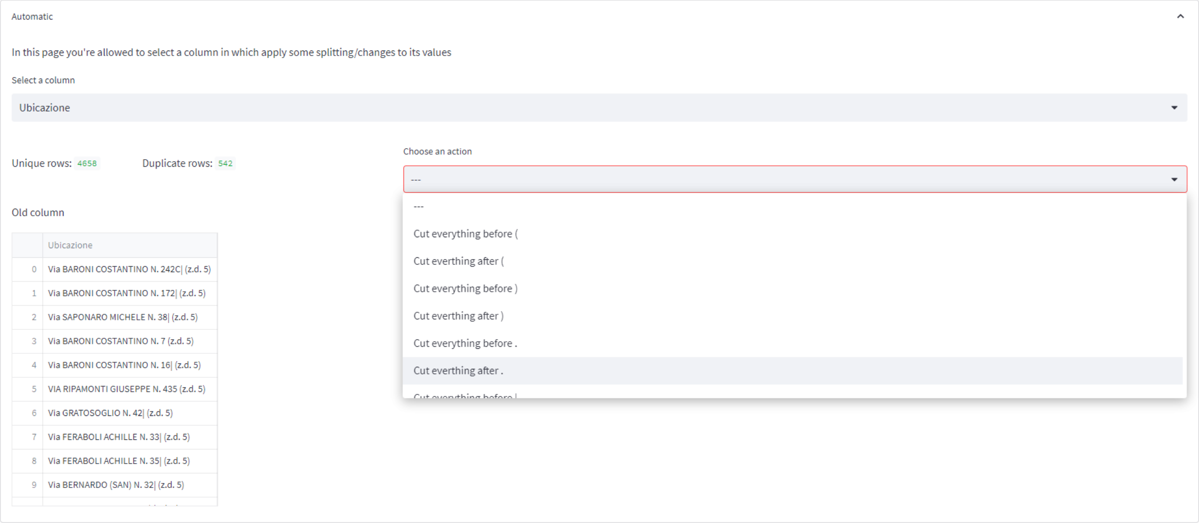
Action: Add

Where: Start of the value

Figure 4.19: Manual section of value filtering.

- Automatic
  In the automatic management of the values, apart from the selection of the column, the tool carries out most of the work. Indeed given the column, the tool at runtime

spots throughout the entire column the most occurring delimiters and prepare the possible actions for the user. The user will then be able to perform only choosing an action present in the list.



Automatic

In this page you're allowed to select a column in which apply some splitting/changes to its values

Select a column

Ubicazione

Unique rows: 4658

Duplicate rows: 542

Choose an action

Old column

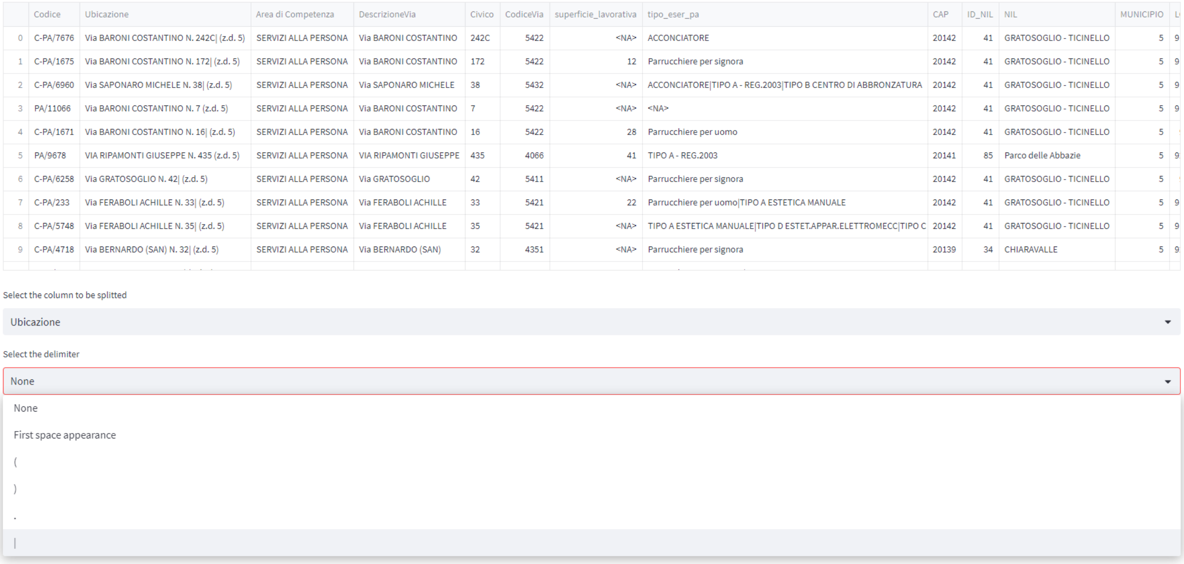
	Ubicazione
0	Via BARONI COSTANTINO N. 242C| (z.d. 5)
1	Via BARONI COSTANTINO N. 172| (z.d. 5)
2	Via SAPONARO MICHELE N. 38| (z.d. 5)
3	Via BARONI COSTANTINO N. 7 (z.d. 5)
4	Via BARONI COSTANTINO N. 16| (z.d. 5)
5	VIA RIPAMONTI GIUSEPPE N. 435 (z.d. 5)
6	Via GRATOSOGLIO N. 42| (z.d. 5)
7	Via FERABOLI ACHILLE N. 33| (z.d. 5)
8	Via FERABOLI ACHILLE N. 35| (z.d. 5)
9	Via BERNARDO (SAN) N. 32| (z.d. 5)

Figure 4.20: Automatic section of value filtering.

After the column has been modified, in both Manual and Automatic cases, a preview of the new column is shown to the user. In case the user accepts the new column and clicks the Save button, the changes will be permanently applied and the dataset will be updated.

Column Splitting

The column splitting page was developed not only with the simple goal of splitting a column but with the goal of proactively guiding the user in finding the most suitable splitting. The user first selects the column to be split with a selectbox. Not all the columns will be available for the splitting because the tool recognizes automatically the numeric-only columns and to them the split is not allowed by design.
The column is then parsed and its delimiters are extracted and attached to a list shown in the next selectbox.



	Codice	Ubicazione	Area di Competenza	DescrizioneVia	Civico	CodiceVia	superficie_lavorativa	tipo_eser_pa	CAP	ID_NIL	NIL	MUNICIPIO	L
0	C-PA/7676	Via BARONI COSTANTINO N. 242C| (z.d. 5)	SERVIZI ALLA PERSONA	Via BARONI COSTANTINO	242C	5422	<NA>	ACCONCIATORE	20142	41	GRATOSOGLIO - TICINELLO	5	9
1	C-PA/1675	Via BARONI COSTANTINO N. 172| (z.d. 5)	SERVIZI ALLA PERSONA	Via BARONI COSTANTINO	172	5422	12	Parrucchiere per signora	20142	41	GRATOSOGLIO - TICINELLO	5	9
2	C-PA/6960	Via SAPONARO MICHELE N. 38| (z.d. 5)	SERVIZI ALLA PERSONA	Via SAPONARO MICHELE	38	5432	<NA>	ACCONCIATORE|TIPO A - REG.2003|TIPO B CENTRO DI ABBRONZATURA	20142	41	GRATOSOGLIO - TICINELLO	5	9
3	PA/11066	Via BARONI COSTANTINO N. 7 (z.d. 5)	SERVIZI ALLA PERSONA	Via BARONI COSTANTINO	7	5422	<NA>	<NA>	20142	41	GRATOSOGLIO - TICINELLO	5	9
4	C-PA/1671	Via BARONI COSTANTINO N. 16| (z.d. 5)	SERVIZI ALLA PERSONA	Via BARONI COSTANTINO	16	5422	28	Parrucchiere per uomo	20142	41	GRATOSOGLIO - TICINELLO	5	
5	PA/9678	VIA RIPAMONTI GIUSEPPE N. 435 (z.d. 5)	SERVIZI ALLA PERSONA	VIA RIPAMONTI GIUSEPPE	435	4066	41	TIPO A - REG.2003	20141	85	Parco delle Abbazie	5	9
6	C-PA/6258	Via GRATOSOGLIO N. 42| (z.d. 5)	SERVIZI ALLA PERSONA	Via GRATOSOGLIO	42	5411	<NA>	Parrucchiere per signora	20142	41	GRATOSOGLIO - TICINELLO	5	
7	C-PA/233	Via FERABOLI ACHILLE N. 33| (z.d. 5)	SERVIZI ALLA PERSONA	Via FERABOLI ACHILLE	33	5421	22	Parrucchiere per uomo|TIPO A ESTETICA MANUALE	20142	41	GRATOSOGLIO - TICINELLO	5	9
8	C-PA/5748	Via FERABOLI ACHILLE N. 35| (z.d. 5)	SERVIZI ALLA PERSONA	Via FERABOLI ACHILLE	35	5421	<NA>	TIPO A ESTETICA MANUALE|TIPO D ESTET.APPAR.ELETTROMECC|TIPO C	20142	41	GRATOSOGLIO - TICINELLO	5	9
9	C-PA/4718	Via BERNARDO (SAN) N. 32| (z.d. 5)	SERVIZI ALLA PERSONA	Via BERNARDO (SAN)	32	4351	<NA>	Parrucchiere per signora	20139	34	CHIARAVALLE	5	9

Select the column to be splitted

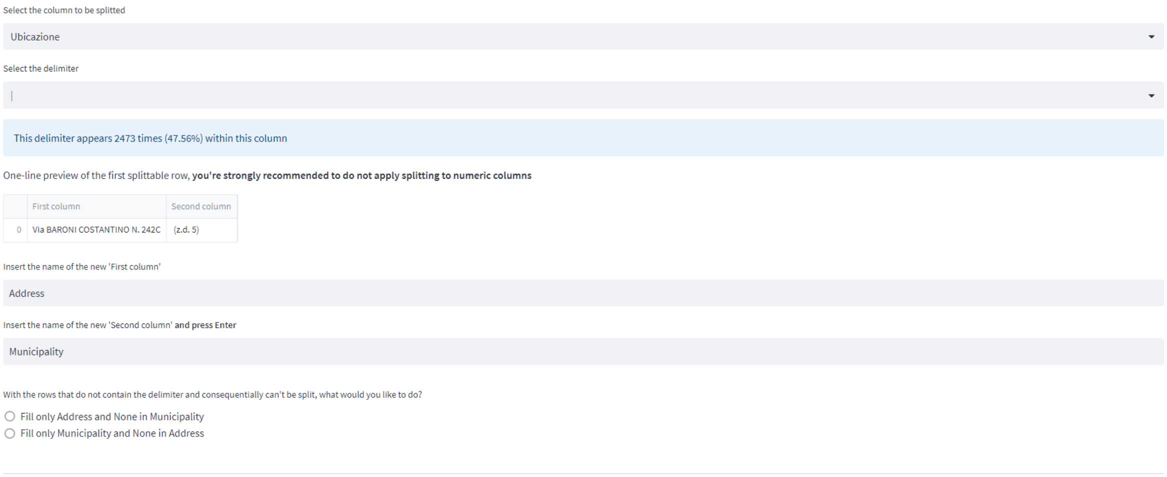
Ubicazione

Select the delimiter

None

Figure 4.21: Delimiters available for the column 'Ubicazione'.

After a delimiter is selected, it is shown to the user the number of times the delimiter is present within the column. This was done specifically to increase the consciousness of the user about the impact of the selected delimiter. This information is relevant because will be up to her/him how to handle all the values where that delimiter is not present. Given that splitting these values is not possible, null values will be imputed in the first new column or in the second new column depending on what the user chooses. The 2 new columns will then replace the old one.



Select the column to be splitted

Ubicazione

Select the delimiter

|

This delimiter appears 2473 times (47.56%) within this column

One-line preview of the first splittable row, **you're strongly recommended to do not apply splitting to numeric columns**

	First column	Second column
0	Via BARONI COSTANTINO N. 242C	(z.d. 5)

Insert the name of the new 'First column'

Address

Insert the name of the new 'Second column' **and press Enter**

Municipality

With the rows that do not contain the delimiter and consequentially can't be split, what would you like to do?

○ Fill only Address and None in Municipality
○ Fill only Municipality and None in Address

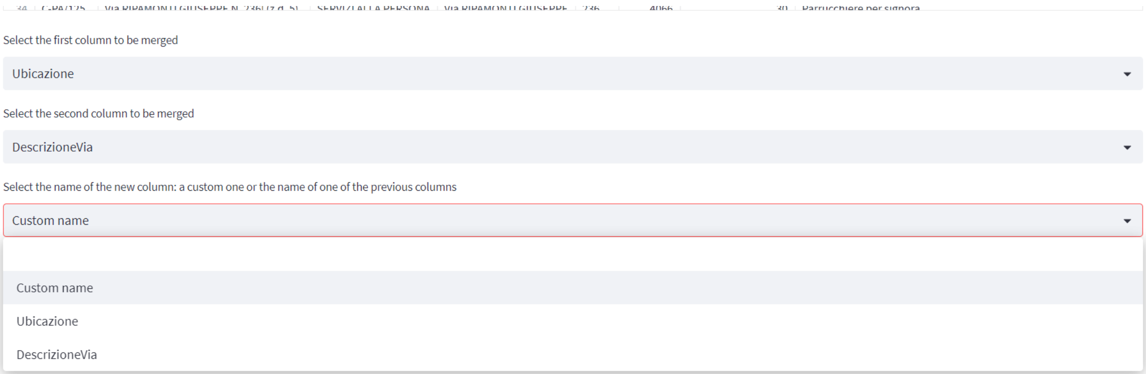
Figure 4.22: Preview of a splitting case and names of the 2 new columns.

After the Go button is clicked, a preview of the possible new dataset is shown. As always,

before modifying the current dataset, the user should explicitly save it with the proper button.
After the saving, the new dataset is profiled again and available for new operations.

Column Merging

The column merging page has the aim of merging two columns. The user should select 2 columns with 2 different selectbox. Obviously, the lists of the available columns are calculated at runtime to prevent the user, for example, to select the same column twice. After the 2 columns have been selected the user should choose a name for the new merged column. S/He can choose between the name of the first column, the name of the second column, and a custom name. If the custom name is selected, an input box will automatically appear.



Select the first column to be merged

Ubicazione

Select the second column to be merged

DescrizioneVia

Select the name of the new column: a custom one or the name of one of the previous columns

Custom name

Figure 4.23: Column merging input and select boxes.

After the new name has been chosen, a preview of the new dataset is shown to the user. In case he/she is comfortable with the changes and decides to save the dataset will be then profiled again and saved permanently.

5 | Conclusion and Future Work

As mentioned in Chapter 2, nowadays companies and enterprises use data analysis as an important decision factor in business and strategy choices. A good outcome from the data analysis pipeline can be reached only if the quality of the data given as input is high and complies with certain standards defined a priori. Because of this, it is important to have instruments and tools to overcome the poor quality issue or at least limit it as much as possible.

This thesis has presented a data preparation and exploration tool developed with the aim of helping the user to profile, wrangle, and clean a dataset. The tool is made also for users without any technical knowledge, as the UI is effective and intuitive. Anyway, an important point is that s/he is required to have at least a basic knowledge of the data preparation tasks that will be applied. A misunderstanding or underrating of the consequences of these actions could even impoverish the dataset that is currently being analyzed. This is because the tool is not aware of the context from which the data has been collected, nor of the final scope that will have the data after the analysis. This additional knowledge is mandatory in order to perform in a reasonable way data exploration and preparation and it can be acquired only by involving a human in the process.

The profiling functionalities of the tool were found to be particularly useful in facilitating data exploration and preparation, providing users with a range of statistical summaries of their data. These functionalities enable users to quickly and easily identify distinct values, missing values, and other issues like the redundancy of information that can impact the quality and reliability of their analysis.
With the wrangling functionalities, the user is able to manipulate the data, for example by splitting a column according to a certain delimiter or editing its values given a pattern: everything in according to the needs and the scope of the analysis.
The implemented cleaning functionalities were discovered to be very effective. The duplicate detection page, for example, allowed to discover that the analyzed dataset had a consistent number of duplicate rows.

The tool developed in this thesis could also be a possible starting point for future implementations and studies in the Data Preparation context. For example, the number of functionalities offered by the tool can be increased or also the already implemented ones can be enhanced. The tool is currently missing all the important functionalities related to data fusion and data integration. Moreover, some context-aware settings could also be implemented that depending on the dataset adjust automatically thresholds or suggested actions by the tool.

An additional element that would enhance by far the power of the tool would be to integrate automation and machine learning technologies as they are becoming increasingly important in data analysis. Moreover, these techniques increase the efficiency of the tool and the precision and reliability of the outcome.

Another important feature that can be improved is the user interface. A user-friendly interface is essential for any software tool, including a data preparation and exploration tool. Simplifying instructions, organizing the interface more effectively, and incorporating user feedback are some options to consider. Conducting user testing can help to identify areas for improvement.

Bibliography

[1] Z. S. Abdallah, L. Du, and G. I. Webb. Data preparation. In C. Sammut and G. I. Webb, editors, *Encyclopedia of Machine Learning and Data Mining*, pages 318–327. Springer, 2017. doi: 10.1007/978-1-4899-7687-1_62. URL `https://doi.org/10.1007/978-1-4899-7687-1_62`.

[2] O. Azeroual. Data wrangling in database systems: Purging of dirty data. *Data*, 5 (2), 2020. ISSN 2306-5729. doi: 10.3390/data5020050. URL `https://www.mdpi.com/2306-5729/5/2/50`.

[3] M. Baak, R. Koopman, H. Snoek, and S. Klous. A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics, 2019.

[4] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3):16:1–16:52, 2009. doi: 10.1145/1541880.1541883. URL `https://doi.org/10.1145/1541880.1541883`.

[5] J. de Bruin, J. Anderson, J. Becker, H. Wong, tylerbinski, T. Waleń, D. Elias, J. Weytjens, T. Knuth, L. Baldassini, M. Antoine, R. Norton, T. Teigman, and V. Deplasse. J535d165/recordlinkage: Release v0.15, Apr. 2022. URL `https://doi.org/10.5281/zenodo.6470616`.

[6] U. Draisbach and F. Naumann. Dude: The duplicate detection toolkit. In *Proceedings of the International Workshop on Quality in Databases (QDB)*, volume 100000, page 10000000, 2010.

[7] P. J. Guo, S. Kandel, J. M. Hellerstein, and J. Heer. Proactive wrangling: mixed-initiative end-user programming of data transformation scripts. In J. S. Pierce, M. Agrawala, and S. R. Klemmer, editors, *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16-19, 2011*, pages 65–74. ACM, 2011. doi: 10.1145/2047196.2047205. URL `https://doi.org/10.1145/2047196.2047205`.

[8] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H.

van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

[9] A. Haug, F. Zachariassen, and D. van Liempd. The costs of poor data quality. *Journal of Industrial Engineering and Management (JIEM)*, 4(2):168–193, 2011. ISSN 2013-0953. doi: 10.3926/jiem.2011.v4n2.p168-193. URL `http://hdl.handle.net/10419/188448`.

[10] M. A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84 (406):414–420, 1989. doi: 10.1080/01621459.1989.10478785. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478785`.

[11] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer. Wrangler: interactive visual specification of data transformation scripts. In D. S. Tan, S. Amershi, B. Begole, W. A. Kellogg, and M. Tungare, editors, *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, Vancouver, BC, Canada, May 7-12, 2011*, pages 3363–3372. ACM, 2011. doi: 10.1145/1978942.1979444. URL `https://doi.org/10.1145/1978942.1979444`.

[12] W. McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[13] A. Nikiforova. Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment. *Balt. J. Mod. Comput.*, 8(3), 2020. doi: 10.22364/bjmc.2020.8.3.02. URL `https://doi.org/10.22364/bjmc.2020.8.3.02`.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] D. Pyle. *Data Preparation for Data Mining.* Morgan Kaufmann, 1999. ISBN 1-55860-529-0.

[16] C. Sancricca and C. Cappiello. Supporting the design of data preparation pipelines. In G. Amato, V. Bartalesi, D. Bianchini, C. Gennaro, and R. Torlone, editors, *Proceedings of the 30th Italian Symposium on Advanced Database Systems, SEBD 2022, Tirrenia (PI), Italy, June 19-22, 2022*, volume 3194 of *CEUR Workshop Proceed-*

ings, pages 149–158. CEUR-WS.org, 2022. URL `https://ceur-ws.org/Vol-3194/paper18.pdf`.

[17] G. Van Rossum. *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020.

[18] W. E. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359, 1990.

List of Figures