POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Data and Information Quality Project

Author(s): **Martina Riva 10756775**
Arimondo Scrivano 10712429

Project ID: **12**

Dataset: **11**

1 | SETUP CHOICES

The setup choices adopted are divided into 3 macro categories:

1. **Inspection/Data Profiling**
2. **DQ Assessment**
3. **Data Cleaning**
   (a) **Data Transformation and Standardization**
   (b) **Error Detection and Correction**
   (c) **Data Deduplication**

The **libraries** involved are (reported once for simplicity but involved throughout the entire process): numpy, pandas, matplotlib.pyplot, recordlinkage(used for data deduplication), seaborn, ydata_profiling and sklearn(for missing values imputation)

1.1. Inspection/ Data Profiling

This section is divided into the following sub-categories:

- **NaN values inspection:** Identifying missing values in the dataset.
- **Unique values inspection:** Analyzing the uniqueness of feature values.
- **Exact matching duplicates:** Detecting duplicate records.
- **Correlation Matrix:** Examining the relationships between features.

Additionally, it includes a general dataset inspection using the Profile Report to provide an overview of the dataset.

1.2. DQ Assessment

This section focuses on evaluating the quality of the raw data. The assessment is based on the following metrics:

- **Completeness**: Measuring the extent to which data is available.
- **Uniqueness**: Ensuring that data entries are not duplicated.
- **Distinctness**: Verifying the diversity of values across features.

- **Constancy**: Checking for consistent patterns within the data.

1.3. Data Cleaning

1.3.1. Data Transformation and Standardization

This section addresses the significant errors found in the dataset (more details will be provided in the following chapter).

This section is divided into:

- **Columns Renaming**
- **Standardization:** focus on 'Settore Storico', 'Tipo via'
- **'Ubicazione' Column Checking and split**
- **New Columns:** "Isolato" and "Accesso"

1.3.2. Error Detection and Correction

This section is divided into:

- **Missing values of:**
  - Settore Storico
  - Tipo esercizio storico
  - Superficie Somministrazione
  - Forma commercio
  - Forma commercio prev
  - Forma vendita
- **Outliers** in Superficie Somministrazione

1.3.3. Data Deduplication

This last section is divided into:

- **Exact Matching on 'Nome Via' and 'Civico' columns**
- **Record Linkage**

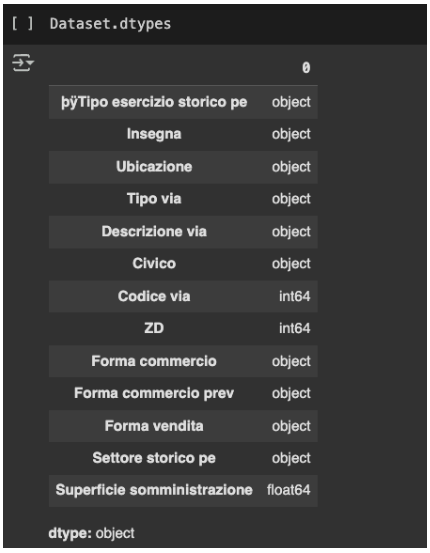
2 | PIPELINE IMPLEMENTATION

For each of the following sections, we provide a description of what we made and why.

2.1. Inspection/Data Profiling

In this section, it's provided an initial description of the data. This is an important step because, before making any changes, it's critical to understand the structure, content, and quality of the dataset.

The dataset is composed initially by 6904 samples and 13 columns:
'þÿTipo esercizio storico pe', 'Insegna', 'Ubicazione', 'Tipo Via', 'Descrizione Via', 'Civico', 'Codice Via', 'ZD', 'Forma Commercio', 'Forma Commercio prev', 'Forma Vendita', 'Settore Storico pe' and 'Superficie somministrazione'.



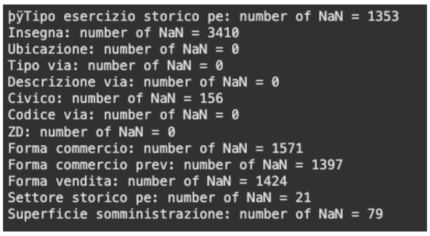
```
[ ] Dataset.dtypes
```

	0
þÿTipo esercizio storico pe	object
Insegna	object
Ubicazione	object
Tipo via	object
Descrizione via	object
Civico	object
Codice via	int64
ZD	int64
Forma commercio	object
Forma commercio prev	object
Forma vendita	object
Settore storico pe	object
Superficie somministrazione	float64

dtype: object

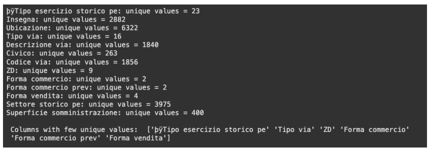
Figure 2.1: datatypes

The first step is the analysis of the null and the unique values.



```
þÿTipo esercizio storico pe: number of NaN = 1353
Insegna: number of NaN = 3410
Ubicazione: number of NaN = 0
Tipo via: number of NaN = 0
Descrizione via: number of NaN = 0
Civico: number of NaN = 156
Codice via: number of NaN = 0
ZD: number of NaN = 0
Forma commercio: number of NaN = 1571
Forma commercio prev: number of NaN = 1397
Forma vendita: number of NaN = 1424
Settore storico pe: number of NaN = 21
Superficie somministrazione: number of NaN = 79
```

Figure 2.2: Number of Null Values



```
þÿTipo esercizio storico pe: unique values = 23
Insegna: unique values = 2882
Ubicazione: unique values = 6322
Tipo via: unique values = 16
Descrizione via: unique values = 1840
Civico: unique values = 263
Codice via: unique values = 1856
ZD: unique values = 9
Forma commercio: unique values = 2
Forma commercio prev: unique values = 2
Forma vendita: unique values = 4
Settore storico pe: unique values = 3975
Superficie somministrazione: unique values = 400

 Columns with few unique values:  ['þÿTipo esercizio storico pe' 'Tipo via' 'ZD' 'Forma commercio'
 'Forma commercio prev' 'Forma vendita']
```

Figure 2.3: Number of Unique Values

The analysis highlighted some important characteristics:

- The names of some columns were partially corrupted and had format errors (e.g., 'bar caffÿý').
- The "Ubicazione" column is a combination of information from the 'Tipo Via', 'Descrizione', and 'Civico' columns (information inferred from high correlation).
- There were repetitions in the domain of some columns (e.g., 'LGO' and 'LARGO', 'VIA' and 'VIE').
- The "Settore Storico" column is a list of strings in the domain of "Tipo esercizio storico pe".
- A very high presence of NaN values was observed.

Then, we inspected the duplicates and found the presence of one. At the end, using the correlation matrix, we observed that some columns are strongly correlated. This helps us to identify relationships between features and detect potential redundancies or opportunities for feature engineering.

2.2. Data Cleaning

2.2.1. Data Transformation and Standardization

In this section, the goal is to standardize the format of the data and detect and correct typos.

The task involves correcting the names of the values, so that they are coherent with the domains, and renaiming the columns, to enhance readability and make them meaningful. The following changes were made:

- 'þÿTipo esercizio storico pe' -> 'Tipo esercizio storico'

- ’Descrizione via’ -> ’Nome via’
- ’Settore storico pe’ -> ’Settore storico’
- Replace ’LARGO’ with ’LGO’
- Replace ’VIE’ with ’VIA’

Next, the strings in the ’Settore storico’ column were standardized, making them as consistent as possible with the values in the ’Tipo Esercizio Storico’ column while retaining most of the information. The standardization also involved removing duplicate or erroneous strings from the ’Settore Storico’ column, and performing a check on the ’Superficie somministrazione’ column.

Furthermore, the ’Ubicazione’ column was standardized. Some inconsistencies emerged because the information was not aligned with ’Nome Via’ (formerly ’Descrizione Via’). For clarity and code-understandability, this issue was corrected in this section. Based on an exploratory analysis of the real data, the information in the ’Ubicazione’ column was used as the correct one, and was substituted wherever ’Nome Via’ was not aligned.

Two new columns were added: ’Ingresso’ and ’Accesso’, which represent the set of information contained in the ’Ubicazione’ column. Finally, the ’Ubicazione’ column was removed.

2.2.2. Error Detection and Correction

In this phase the task of addressing the NaN is carried out.

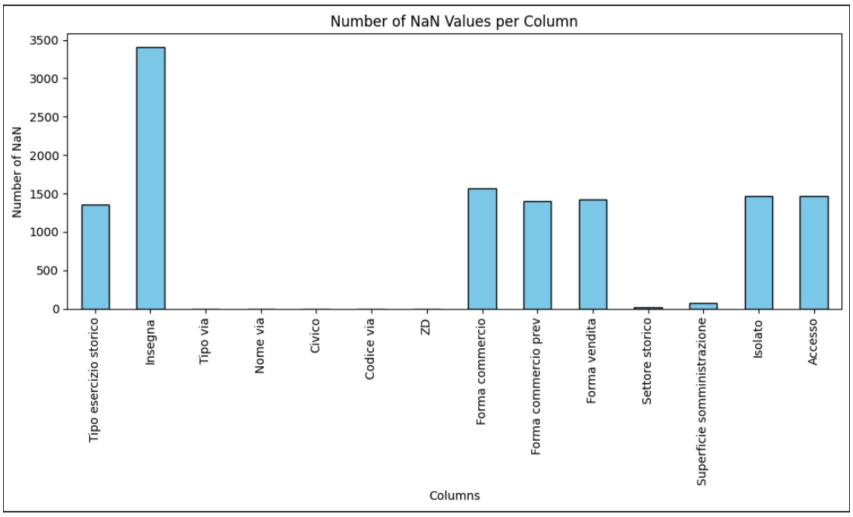


Figure 2.4: Numbers of Null values per Columns

Unfortunately, due to a lack of obtainable information, the columns ’Insegna’, ’Accesso’, and ’Isolato’ could not be completed.

The remaining columns, ’Tipo esercizio storico’, ’Forma commercio’, ’Forma commercio prev’, ’Forma vendita’, ’Settore storico’, and ’Superficie somministrazione’, were processed.

Entries with more than 5 null values and those missing the `’Settore storico’` value were removed, as it is difficult to retrieve them. The remaining entries were modified as follows:

- **’Settore Storico’**: If missing, it was set to the value of `’Tipo Esercizio Storico’`, based on the reasoning: "If it was something, for sure it was at some point the current activity." This heuristic ensures consistency and logical alignment between the two columns.
- **’Tipo Esercizio Storico’**: If missing, it was set to the first element (the string before the semicolon `’;’`) in `’Settore Storico’`, provided that the element existed in the dictionary of valid values for `’Tipo Esercizio Storico’`.
- **’Superficie Somministrazione’**: Missing values were addressed using the KNNImputer with `neighbors=10`. This method imputes missing values based on the average of the 10 nearest neighbors in feature space, ensuring consistency with similar data points.
- **’Forma Commercio’** and **’Forma Commercio Prev’**: Missing values in `’Forma Commercio’` were replaced with the more general value (`’somministrazione’`), while missing values in `’Forma Commercio Prev’` were filled using the corresponding value from `’Forma Commercio’`.
- **’Forma Vendita’**: If missing, it was set to `’misto’`, the most general value. This ensures that missing data does not lead to ambiguity or misclassification.

The last step in this section was handle **Outliers in ’Superficie Somministrazione’**. Using the KNN algorithm with `n=3` neighbors, outliers in `’Superficie Somministrazione’` were identified. These outliers were replaced with the mean value of their 3 nearest neighbors, ensuring that extreme values do not distort subsequent analyses.

2.2.3. Data Deduplcation

In this section, duplicate entries were identified and addressed using two methods:

1. **Exact Match**: An entry was considered a duplicate of another if `’Tipo Esercizio Storico’`, `’Insegna’`, `’Tipo Via’`, `’Nome Via’`, and `’Civico’` were identical. These attributes collectively define a unique place. This approach adheres to the most conservative methodology to ensure accuracy.
2. **Record Linkage**: The dataset was sorted by `’Nome Via’` to maximize the likelihood of identifying potential duplicates. While an exact match was required for `’Tipo Esercizio Storico’` and `’Civico’`, a similarity measure (`’jarowinkler’`) was applied to `’Nome Via’`, to consider possible humans mistakes. If these conditions were met, an entry was deleted if the `’Insegna’` value was either identical between the two entries or missing in one of them.

3 | RESULTS

3.1. Results

This section highlights the main outcomes achieved through the data preparation pipeline.

The focus was on ensuring data quality, addressing missing values, standardizing features, and resolving inconsistencies.

The proportion of missing values was significantly reduced. Additionally, using exact matching and record linkage, duplicates were reduced, ensuring a cleaner dataset. The columns such as 'Settore Storico', 'Tipo Via', and 'Nome Via' were standardized, reducing inconsistencies and ensuring uniformity across features.

Outliers in Superficie Somministrazione were identified using the KNN algorithm, and were replaced with the mean of the three nearest neighbors, resulting in a more reliable dataset for analysis.

3.1.1. Final Dataset Overview and Challenges

After processing, the dataset contained 6666 samples and 14 columns. All columns were free of significant errors, duplicates were eliminated, and features were standardized, ensuring high data quality for subsequent analysis.

However some **challenges** remain:

- **Handling Missing Values:** Some columns, such as 'Insegna', 'Accesso', and 'Isolato', lacked sufficient information and could not be completed. These could be removed, but we decide to keep them as additional informations.
- **Inconsistencies:** Aligning 'Ubicazione' with 'Nome Via' required extensive exploration and analysis, highlighting the importance of domain knowledge in data cleaning.

3.1.2. Data Quality Assessment

Evaluating completeness, uniqueness, and constancy ensures that the data meets the minimum quality standards for meaningful analysis. Results:

- **Overall Completeness:**
  - **Before Cleaning:** 89.5%

  - **After Cleaning:** 93.7%

  The overall completeness improved by 4.2%, reflecting a significant reduction in missing data.

- **Key Observations per Column:**
  - **Tipo esercizio storico:** Minimal change in uniqueness (0.33% to 0.31%) and constancy (51.3%). The column appears stable but with limited variety.
  - **Insegna:** Uniqueness increased from 41.7% to 43.2%, due to deletion of duplicates. Constancy remained stable at 5.8%.
  - **Tipo via:** Marginal improvement in uniqueness and constancy, showing better consistency.
  - **Civico:** Uniqueness increased from 3.8% to 6.3%.
  - **Forma commercio:** Constancy improved significantly from 89.3% to 91.8%, indicating stronger standardization in business type entries.
  - **Settore storico:** Uniqueness dropped significantly (from 57.5% to 41.1%) with no change in constancy. This could indicate data aggregation or simplification.
  - **New Columns (Isolato and Accesso):** These columns were added to enhance data granularity. For instance, *Accesso* shows a high constancy of 98.9%.