Data and quality information

Project report

Ignazio Iacono 11015241

Raffaele Russo 11016624

Project ID : 24 | Dataset N.11

Politecnico di Milano A.A. 2024-2025

Cinzia Cappiello - Camilla Sancricca

1 Setup choices

The pipeline uses fundamental tools such as `*pandas*` for data manipulation, `*numpy*` for numerical operations, and specific libraries like `*ydata-profiling*` for data profiling. In particular for the initial data exploration we use automated profiling tools to identify anomalies, distributions, and relationships within the data, facilitating informed decisions on data cleaning and transformation. For 'preprocessing', we use techniques such as encoding categorical variables (using `*LabelEncoder*`) to handle non-numeric features and preparing features with advanced tools like Random Forests to assess feature importance.

2 Pipeline implementation

2.1 Data profiling

The initial dataset analysis was conducted using the '*ydata-profiling*' library, which generated a comprehensive Profiling Report. This library was chosen for its ability to provide rapid, detailed, and visually intuitive. Analyzing the report we obtained, several critical issues emerged regarding:

1. **Missing Values:**
    - A significant portion of missing data was identified columns: *Insegna*, *Tipo esercizio storico pe*, *Forma commercio*, *Forma commercio prev*, and *Forma Vendita*. These columns exhibited missing values ranging between 20% and 50%, which could impact the quality and reliability of any subsequent analysis. Addressing these gaps required targeted strategies, such as imputation or exclusion, based on their relevance to the dataset.
2. **Inconsistent Formats:**
    - Text variables displayed inconsistencies in formatting, particularly in columns related to location, such as *Ubicazione*. For instance, tuples in the *Ubicazione* column were not standardized, often containing mixed information (e.g., access type and road type) that was already duplicated in other fields. This lack of uniformity made data processing and interpretation more challenging, necessitating a thorough transformation process.
3. **Outliers:**
    - Some columns revealed the presence of outliers, with distributions heavily skewed away from expected values. These anomalies could either represent valid extreme cases or data entry errors, and they required further investigation to determine their origins and impact.
4. **Duplicate Records:**
    - Preliminary analysis suggested the presence of potential duplicate records, though not always exact matches. This called for advanced deduplication techniques to identify and consolidate duplicate entries, ensuring the integrity and accuracy of the dataset.

3 Data quality assessment

The data quality assessment phase was designed to analyze and ensure the completeness, consistency, and accuracy of the dataset. Initially, we evaluated the following characteristics for each column:

- **Uniqueness**: Checking the uniqueness of rows.
- **Distinctness**: Verifying the presence of distinct values.
- **Constancy**: Analyzing the constancy of values.
- **Completeness**: Assessing the presence of missing values.

Subsequently, duplicates were checked, with only one duplicate found.

The phase then focused on the **accuracy** of the columns "Codice Via" and "Municipio":

- For **Codice Via**, the columns "Descrizione Via" and "Codice Via" were taken from the main dataset and merged with the external dataset using the "CODICE_VIA" field. A verification was performed to check if the "Descrizione Via" from the main dataset matched the description retrieved from the join. This yielded an accuracy of 97.09%.
- For **Municipio**, the columns "Codice Via", "Civico", and "ZD" from the main dataset were compared with "CODICE_VIA", "NUMERO", and "MUNICIPIO" from the external dataset. The join on "Codice Via" and "Civico" allowed verifying the consistency of "Municipio", achieving an accuracy of 97.42%.

Finally, **consistency** was assessed through:

- A **group by** on "Codice Via" and "Descrizione Via" to verify if the same street code had different descriptions.
- Checking the "Civico" field for negative numbers.

In both cases, no inconsistencies were detected.

4 Data cleaning

4.1 Data transformation/standardization

During the data transformation and standardization phase, the focus was on ensuring consistency and usability of the dataset by standardizing and transforming the content of all columns, except for Codice_via, which was assumed to be accurate and did not require further processing. The steps taken for each of the remaining columns are outlined below, along with specific enhancements and external reference integration.

1. Identification and Correction of Formatting Errors
   - Formatting errors were addressed using pattern-based corrections.
   - For the column Ubicazione, additional columns were generated using extracted patterns. These derived columns were identified with the _ubicazione suffix, reflecting different aspects of the location data.
2. Validation with Reference Datasets
   - For columns derived from Ubicazione, such as Descrizione_via_ubicazione, matching was performed against existing columns like Descrizione_via to ensure consistency. A reference dataset from the Lombardy region, detailing the road infrastructure of Milan, was used to validate and standardize location details, ensuring high-quality information.

3. Verification of Resulting Tuples
    - After transformations, tuples were reviewed for consistency to identify and resolve anomalies introduced during the data standardization process.
4. Mapping and Standardization of Specific Columns
    - Forma_commercio: Columns related to the form of commerce were standardized to boolean values (true/false), with a default value of false for missing entries. This ensured uniformity across all rows.
    - Settore_storcio_pe: The Settore_storico_pe column was alphabetically ordered and expanded to maintain consistency and enhance usability in subsequent analyses, in particular for duplicate detection.
    - Ubicazione: a significant portion of the transformation efforts was dedicated to managing and enriching this column, which often contained valuable information, which was extracted and utilized to enhance the dataset. The following steps were applied:
        - Splitting by Content: The Ubicazione column was split into multiple attributes, ensuring maximum retention of meaningful details for future input and analysis stages.
        - Pattern-Based Extraction: Additional parameters, such as ‘Isolato’ and ‘Ingresso’, were extracted using custom patterns, further enriching the dataset with structured location-related information.
        - Cross-checking: If the tuple has a null value in the 'Civico' column, the value extracted from 'Ubicazione' is used and the values for ‘accesso’ and ‘Isolato’ extracted from Ubicazione will be used. On the other hand, if 'Civico' is different from the 'Civico' extracted from 'Ubicazione', the values for 'Isolato' and 'accesso' will not be considered, and specifically, they will be set to null.

4.2 Error detection and correction

4.2.1 Missing Values

Once identified missing columns with missing values, the most appropriate imputation technique for each column was selected, considering both the specific domain of each column and its importance within the dataset. For the "Insegna" column, given its highly subjective and unpredictable nature, it was decided to impute the default value "unspecified." For the "Superficie_somministrazione" and "Settore_storico_pe" columns, after analyzing the distribution of values during the profiling phase, the **mode** was chosen as the imputation technique, as the low number of missing values (less than 2%) allowed for a simple and effective method. For columns with a significant number of missing values, a **Random Forest-based** model was used to predict the missing values in some columns of the dataset. This technique leverages the correlation between the target column (with missing values) and a predictor column to estimate the missing values. For each column, if the target column is categorical, the values are transformed with a **LabelEncoder**, and the prediction model used is the **Random Forest Classifier**. If the target column is numeric, the prediction model uses a **Random Forest Regressor**. **Prediction**: Once the model was trained on the non-missing data, predictions for the missing values in the target column were made. **Data Update**: The predicted values were used to replace the missing values directly in the original dataset. The **Random Forest** model was chosen for imputing missing values for several reasons:

- **Ability to Handle Non-Linear Relationships**: Random Forest is a model that can capture complex and non-linear relationships between the predictor variables and the target variable. This is particularly useful when the relationships between variables are not explicitly clear.

- **Non-Parametric Model**: Since Random Forest does not make assumptions about the distribution of the data (unlike other models), it is a good option for different types of data, both numerical and categorical, without the need for complex data preprocessing.

4.2.2 Outliers

Based on the frequency analysis results for each column, the outlier detection process using categorical data has been thoroughly evaluated, and the following insights can be made:

1. **Tipo_esercizio_storico_pe**:
   - The majority of the categories are well-represented, with the top category "bar caffetteria" comprising over 66% of the data. However, no categories fall below the 1% threshold for rare occurrences, meaning no potential outliers were detected for this column.
2. **Insegna**:
   - The category "unspecified" represents 49.39% of the data, while other categories like "bar caffe" and "bar tabacchi" appear at much lower frequencies. Despite the vast number of distinct categories (2883 unique values), no rare categories (under 1% frequency) were identified, suggesting that even though many unique entries exist, none are significantly underrepresented to be considered outliers.
3. **Tipo_via**:
   - The most common categories are "VIA" (73.75%), followed by "VLE" (12.04%) and "CSO" (5.20%). Again, no rare categories were found under the 1% threshold, meaning no outliers were flagged.
4. **Descrizione_via**:
   - With 1839 unique values, the street descriptions are mostly represented by a few prominent entries such as "MONZA" and "PADOVA". However, no outliers were detected as there were no categories with less than 1% frequency.
5. **Civico**:
   - The most frequent civic numbers are 20 (6.79%) and 10 (6.60%), with other numbers showing lower frequencies. No rare categories under 1% were identified, and thus no outliers were flagged.
6. **Codice_via**:
   - Similar to the civic number column, street codes such as 2274 (1.07%) and 2275 (0.78%) are the most frequent. Again, no rare values below the 1% frequency were detected, indicating no outliers.
7. **ZD**:
   - The ZD column, being numerical, does not contain categorical values, but an IQR-based analysis shows no outliers, as all values fall within the expected range.
8. **Forma_vendita**:
   - Categories like "al banco" and "misto" represent the majority of the data (42.28% and 39.85%, respectively). No rare categories were found, so no outliers are detected here.
9. **Settore_storico_pe**:
   - This column contains a mixture of multi-category values such as "bar caffetteria; bar gastronomici e simili". The most common combinations are still well-represented, and no rare combinations below the 1% frequency threshold were identified.
10. **Superficie_somministrazione**:

- With a broad range of values, the column's distribution suggests that no outliers were detected using the IQR method. The values are within reasonable bounds, with the maximum value at 2336 and a 99th percentile at 414.97.

11. **Isolato**:
    - A few block numbers appear to be frequent outliers, such as 476, 466, and 503. These values fall into the upper range of the distribution, suggesting that they may represent special cases or errors in the data.
12. **Ingresso**:
    - The most common category is "ACCESSO ESTERNO", representing 92.92% of the data. Although there are many distinct entries, none of them fall below the 1% frequency threshold, and therefore, no outliers were detected.

Overall, most of the categorical columns show a dominant category or set of categories, with no rare categories falling below the 1% threshold that would indicate outliers. However, some columns like **Isolato** reveal a few potential outliers that might require further investigation or validation. For the categorical data, the applied threshold of 1% successfully filtered out any rare or potentially erroneous categories, ensuring a clean dataset for further analysis or modeling.

4.3 Data deduplication

The process involved two main tasks:

1. handling exact duplicates.
2. identifying similar records using a combination of normalization and record linkage techniques.

To address exact duplicates, particular attention was given to the column `Settore_storico_pe`, where entries often contained identical elements listed in different orders and separated by semicolons (`;`). To standardize these entries, the elements were, in the previous step, sorted in lexicographical order. This approach ensured that records differing only in the order of their elements were treated as duplicates and consolidated effectively.

For identifying similar records, the RecordLinkage library was employed to implement a robust methodology. The first step involved the blocking phase, where the column `Codice_via` was selected as the blocking key. This decision was based on the assumption that each street is uniquely identified by its `Codice_via`. Blocking reduced computational complexity by restricting comparisons to records with the same `Codice_via`. The second step focused on defining and applying comparison rules. Three columns (`Codice_via`, `Civico`, and `Tipo_esercizio_storico_pe`) were required to match exactly because the exact match on the columns `Codice_via`, `Civico`, and `Tipo_esercizio_storico_pe` is sufficient because it was assumed that cases where multiple public establishments exist within a multi-floor building would not be considered. For the column `Settore_storico_pe`, a less stringent approach was adopted using the Jaro-Winkler similarity method with a threshold of 0.9. This allowed slight variations in the list of sectors while still considering records similar. A pair of records was deemed similar if at least three out of the four defined features matched the comparison criteria. Using these rules, the process identified 101 potential matches for further inspection and resolution. The column `Insegna` was excluded from the comparison rules due to extensive imputation, where missing values were replaced with the default value `unspecified`. While this approach achieved 100% completeness for the column, it compromised its relevance as a distinguishing feature for deduplication purposes. This trade-off was necessary to enhance the overall usability of the dataset.

5 Results

The transition from the initial dataset to the final dataset reflects significant enhancements and structural adjustments, improving several data quality dimensions.

1. **Completeness**: Completeness has notably increased across most columns. For example, the `Tipo_esercizio_storico_pe` field achieved full completeness (1.0) in the final dataset, compared to 80.4% in the initial dataset. Similarly, other fields such as `Insegna` also saw completeness rise to 100%, addressing prior gaps.
2. **Uniqueness and Distinctness**: Several columns have experienced changes in uniqueness and distinctness, reflecting the elimination of duplicates and redundancy. For instance, the `Insegna` column's uniqueness and distinctness were maintained at around 42%, indicating limited improvement in reducing overlapping values but ensuring consistency across entries.
3. **Constancy**: Significant improvements in constancy were observed in many fields, such as `Forma_commercio_prev_minuto_boolean`, which reached a constancy of 98.2% in the final dataset, up from a previously unspecified lower value. This reflects a consolidation of values in fields where variability was previously a concern.
4. **Standardization and Schema Alignment**: Adjustments in datatype definitions were implemented effectively. Fields like `Tipo_esercizio_storico_pe`, previously categorized as "object," have now been uniformly defined as "string." Additionally, new boolean fields, such as `Forma_commercio_*_boolean`, have been introduced to simplify categorizations and facilitate analysis.
5. **Data Representation and Consolidation**: The restructuring effort introduced new columns like `Isolato` and streamlined categorical variables into consistent formats. While fields like `Forma_commercio_prev` have been replaced with more specific boolean indicators, aiding in clarity and simplifying downstream processing.
6. **Notable Adjustments in Metrics**:
    - `Superficie_somministrazione`: The uniqueness metric improved marginally, reflecting a better capture of unique data points.
    - `Settore_storico_pe`: Distinctness has been maintained while addressing missing values, further consolidating the quality of categorizations.

In summary, the final dataset demonstrates significant strides in completeness, consistency, and schema alignment, aligning better with data quality principles such as accuracy and usability for analytical purposes.