Data management for informed decision-making: analyzing Milan's rental property market with integrated news data

Matteo Altieri, Enrico Mannarino, Christian Persico

June, 2023

Abstract

In recent years, the demand for living in Milan has led to rising rental prices. Navigating the city's real estate market can be confusing, especially for newcomers. To address this, we created a database of apartment rental listings from popular websites like Immobiliare.it and Subito.it. By utilizing web scraping and API protocols, we extracted valuable data on contract prices, listing titles, descriptions, and property characteristics. Also, we obtained address information required, by developing algorithms to parse descriptions and titles. Recognizing the importance of the neighborhood's quality, we enriched the dataset with links to local news, events, and parades. These additions were scraped from MilanToday via the Google News portal. Finally, we merged the datasets into a NoSQL database, offering users easy access to organized and cleaned data for informed decisions on Milan's listings and desired neighborhoods.

Contents

1. Introduction

It has become evident how an ever-increasing demand for living in the city of Milan has led to an increase in the rental prices of houses in the city, fostered by the opportunities for work, study and entertainment that this metropolis offers. The real estate market, made by individuals and agencies, for someone not born and raised in this city can be quite confusing, which is why we felt it was essential to provide an initial database of apartment rental listings, so that anyone can begin to untangle the world of Milan's rentals. To accomplish this we started with two of the major sites: Immobiliare.it and Subito.it. Through web scraping techniques and taking advantage of API protocols we extrapolated useful data within the ads. First, we made two datasets to accomodate all the initial data scraped from the pages, or obtained through the API, about the price of the contract, the title and description of the listing and characteristics of the estate like the squared meters, the floor of the apartment, the number of locals, the number of bathrooms and many other features that make an apartment unique and relevant for a potential tenant. Given that address information are not reported in a particular section of the HTML code, but only in the description entered by the agency or from the land lord, we created ad-hoc algorithms that can extract addresses and locations from the description and title texts, to obtain those information from both sites and have the possibility to standardize both datasets. To allow an enhanced readability, immediacy and machine readability of all this informations, our data went through a meticulous process of data cleaning. We then identified ourselves as the end user of our product, thus thinking about understanding what would actually motivate someone to rent a house in a particular neighborhood, rather than another, not considering the intrinsic features of the property themselves. We thus realized that getting an idea of the quality of the environment in

which the house is immersed is just as important as the characteristics of the property itself, and there is no better way to understand what life is like in a particular neighborhood than to have immediate access to the latest news stories, parades and events, about it. This is why we decided to further enrich the dataset, with links to the latest happenings located near the building, using the same web scraping technique chosen for the two listing sites, from the online newspaper called MilanoToday through the Google News portal. Last, we merged the two datasets into a NoSQL database creating a place where a single user, without in-depth knowledge of data management, can access these cleaned and organized data to make his conclusion about the the listings of Milan and better decide which is the district of the city where he would like to live.

2. Data acquisition and cleaning

For data acquisition, web scraping techniques and extraction through REST API were used. The data are updated as of June 5, 2023, and the necessary computations were performed using the Python programming language. For data cleaning, duplicates and outliers were then eliminated.

2.1. Web scraping of Milan's street directory

There is no official breakdown of streets in specific neighborhoods for the city of Milan. In fact, upon consulting the municipality's website, it has emerged that the separation we are looking for is accomplished through the "Nuclei di Identità Locali" (NIL); however, the list of streets that comprise them is not provided. Both the websites Immobiliare.it and Subito.it group the streets into districts differently, without disclosing the methodology of how these streets are divided. For example, if we wanted to search for an apartment in the Arco della Pace area, on the Subito website, we would have to filter by "Vercelli, Fiera, Sempione," while on the Immobiliare website, we would filter by "Arco della Pace, Arena, Pagano." It is clear, therefore, that by solely relying on the similarity of labels chosen by the platforms themselves, we may end up expanding the search results unnecessarily. To overcome this, it was decided to enrich the two datasets with the relevant area of each listing, obtaining this classification from an external source. In practice, by assigning the neighborhood to each street again, we will avoid any asymmetry issues. The external source consists of the complete street directory of the city, available on the Openalfa website [1], whose information is obtained and regularly updated from freely available data on OpenStreetMap, a collaborative project aimed at creating freely accessible maps of the world. As a result, we now have a list of all the streets in Milan divided by their corresponding neighborhood. To obtain the platform's data, a web scraping process was performed using the BeautifulSoup library in Python. This allowed us to assign the neighborhoods to the keys of a Python Dictionary and the corresponding list of streets to the values, making the data cleaning process more manageable. However, some streets are present in multiple neighborhoods because they form the boundary between them. The method adopted for their removal is automated, so it should be noted that, exceptionally, streets typically associated with certain neighborhoods may be assigned to another. For example, Via Padova, traditionally associated with the Padova neighborhood, has been associated with the Parco Lambro - Cimiano neighborhood.

2.2. Web scraping of Subito.it

For the Subito website [2], the real estate section was examined, specifically the listings for apartments for rent in the city of Milan. In this case, web scraping was conducted using the BeautifulSoup library to extract relevant information directly from the web page. The obtained and collected information pertains to the first 50 pages of the website, where the relevant listings are divided. Initially, there were a total of 1123 listings and 10 variables, namely:

- *Title*: the title of the listing.
- *Price/month*: the price of the rental.
- *Size (m^2)*: the size of the apartment in square meters.
- *Locals*: the number of rooms in the apartment.
- *Bathrooms*: number of bathrooms in the apartment.
- *Floor*: floor on which the property is located.
- *Description*: a description of the apartment.
- *If_agency*: binary variable indicating whether the listing was posted by a real estate agency ('agency') or by a private individual ('private').
- *Agency_address*: address of the real estate agency managing the listing, if available..
- *URL*: the URL link to the listing.

Since the listing description and the agency address are not present on the main page, they were obtained separately by using the individual URLs of the listings to navigate inside and extract the information. Figure 1 shows an example of a rental listing, in this case belonging to a real estate agency, as can be observed from the banner at the bottom; otherwise, it would not have been present.

Figure 1: Example of a rental listing on the main page of Subito.it

To avoid the presence of fake listings where a user posts an advertisement seeking housing instead of offering it for rent, only those with a thumbnail image were considered. In fact, listings without any image are less appealing and useful to potential users because they lack essential information for decision-making. There are also listings where the floor is not specified as a number but is indicated by values such as 'Rialz.' (which stands for raised), 'Piano T' (ground floor), which have been appropriately recoded as 0, or 'Semint.' (basement), 'Interr.' (sub-basement), which have been recoded as -1. It is now that the custom-built function, described in section 2.4, comes into play. This function allows us to extract the street of the property, if available, from the title or description of the listings with a good degree of accuracy. With the help of this function and the dataset previously obtained from the website related to the street directory of the city of Milan, we are now able to determine the neighborhood where the property is located. As a result, we have two additional useful and important variables: *Address*, which contains the street of the apartment, and *District*, referring to the neighborhood it belongs to. The obtained dataset is then cleaned by removing duplicate observations. This can occur when listings are posted multiple times on the platform to enhance visibility. To identify and address this issue, we first select rows that share common values for variables related to price, size, floor, neighborhood, number of bathrooms and rooms, as well as whether there is an agency involved and its potential address. This results in a total of 191 potentially duplicate observations. Next, to add another level of precision, we compare the descriptions of these potentially duplicate listings using a text correspondence method. To optimize the process, we directly work with the grouped observations that already have the same values for the aforementioned variables. We analyze pairwise combinations and set the threshold at 90% of the words (excluding repetitions) in common. This yields a total of 52 duplicates, which represents 4.6% of the observations. To evaluate the presence of possible outliers, the main descriptive statistical measures were calculated. Unusual values were found for the price and size variables of the apartments. In fact, the minimum recorded size in the dataset is 1 square meter, which is highly inaccurate and likely a result of input error. Therefore, a minimum limit was set, in accordance with Article 96 of the Building Regulations of the Municipality of Milan [3], at 28 square meters. On the other hand, the minimum price is 40 euros per month, while the maximum is 430,000 euros per month. The minimum value suggests the presence of listings referring to short-term rentals, while the maximum value suggests possible apartments for sale, despite being in the wrong section. To address this, a minimum price of 300 euros and a maximum price of 10,000 euros were considered for the available data. The outliers, as defined, amount to 78 observations. Therefore, the final dataset consists of 993 total observations (approximately 11.6% of the initially imported data were removed). Lastly, the file is exported in JSON format for subsequent integration and storage.

2.3. Data extraction through Immobiliare.it API

Data extraction from the Immobiliare.it website [4] was performed using APIs obtained by analyzing the HTML code of the web page dedicated to apartment rentals in the city of Milan. Specifically, we have extracted 2000 listings, which correspond to the first 80 pages of the website. Among the provided information, only the most useful and important data were considered. Many of these variables are already shared with Subito.it, such as the listing title (which includes the address of the property), the URL of the listing, the monthly price, the number of bathrooms, the size in square meters, the description, the number of rooms, and the floor on which the apartment is located. Additionally, we have information regarding:

- *If_agency*: owner of the listing, namely real estate agency (*agency*), private individual. (*private*), independent advertiser[1] (*pro*) or

[1]It refers to an advertiser who offers an informational service providing access to a database containing property listings from private owners, which can also be found for free on other portals. They do not charge mediation fees but a single fee to access the service.

construction companies (*constructor*).

- *Features*: accessory components of the apartment such as air conditioning, whirlpool bathtub, and others.
- *Bedrooms_number*: number of bedrooms.
- *Photos*: URLs of images related to the interior spaces of the property.
- *Energy_info*: energy class and fuel type of boilers.
- *Elevator*: elevator's presence.
- *Condition*: condition of the building structure.
- *Agency_url*: link to the web page of the real estate agency managing the listing, if available.

If important details are not specified in the listings, then we will have a null value in their place. The data was then collected into a dataframe using the Pandas library, creating a structure similar to the one obtained previously with Subito.it. In addition, information about the address and neighborhood of the property was also integrated using the algorithm developed and explained in paragraph 2.4. Special cases for variables like *Floor* and others were also recoded, which had invalid strings such as "3-4" to indicate intermediate floors between the third and fourth. Furthermore, in this case, some unusual values for prices and apartment sizes emerged. Therefore, the same criteria used in paragraph 2.2 were adopted to remove duplicate observations and outliers. In fact, 121 duplicates and 119 outliers were identified and removed, resulting in a final total of 1739 records, as rows with null descriptions and floors were also excluded. Finally, the dataset was saved in JSON format in preparation for integration.

2.4. Algorithm for street extraction

Upon an initial analysis of the listings, we noticed that the street where the rental property is located is not specified in any section of the pages on both platforms. This information is only mentioned within the descriptions or titles provided by the ad owners. Therefore, it was deemed necessary to develop an algorithm that could accurately extract the address from the text for each listing where it is present. The chosen approach was to create a Python module containing all the necessary functions to perform this task. The method compares the descriptions and titles with the street names collected from the Milan city street directory, selecting and assigning the one that matches the text. Multiple attempts were made to achieve this goal. Initially, there was a mistaken belief that a suitable regular expression could extract the addresses from titles and descriptions since they were consistently prefixed with terms like "via," "viale," "piazza," etc., although in varying positions within the text. Once it became clear that a more precise approach was required, new strategies were devised. The first insight was to assign a similarity ratio between each street name and the texts, allowing us to identify the correct match based on the highest similarity score. To accomplish this, we opted to use the SequenceMatcher command from the Difflib library in Python, which utilizes an algorithm published by Ratcliff and Obershelp in the late 1980s known as "Gestalt pattern matching." This algorithm finds the longest contiguous matching subsequence in the strings. However, the use of this procedure revealed two problems:

1. Firstly, comparing each individual street with every description made the procedure computationally inefficient. In order to optimize our search, instead of analyzing each complete description, we selected only a limited portion through targeted slicing. To determine the ideal positions for slicing, we conducted a keyword-based search using terms like "via" and "viale" within the text and considered the right context of a length equal to the desired street.
2. The second issue pertains to the accuracy of the information. During the analysis of 50 randomly extracted observations, it was found that 20% of the time, the address extracted from the description was incorrect. This was due to the algorithm making errors in assigning addresses that contained compound names or had strong similarities to others. For example, "Via Pacini" was mistakenly assigned as "Via Pace" due to their close resemblance instead of "Via Giovanni Pacini." Similarly, "Corso Garibaldi" was never correctly assigned because it was listed as "Corso Giuseppe Garibaldi" in the city street directory.

The second problem was effectively solved by adopting a natural language technique known as Bag of Words, which allowed for a comparison between portions of the description and the city streets, regardless of exact spelling match. Each portion of the street was associated with a similarity index using the *token_sort_ratio* command provided by the FuzzyWuzzy library. The algorithm extracts the street with the highest similarity. In

cases where the text mentions multiple streets and they have the same maximum similarity, the street with the most words is selected. This choice was made based on the assumption that an important address would be written in its entirety. Thanks to these new methodologies, we were able to identify and assign the represented streets within the listings more accurately, avoiding incorrect assignments due to compound names or strong similarities. However, the description often includes the agency's address, even though it is already acquired in a separate variable. For this reason, an additional filter was introduced to exclude it from the selection process from the outset. As can be seen from Figure 2, the function is capable of distinguishing the exact street even when the description of the listing includes points of interest in the city with names similar to those of the streets. It is also able to adapt the syntax of the sought-after street to its corresponding representation in the text.

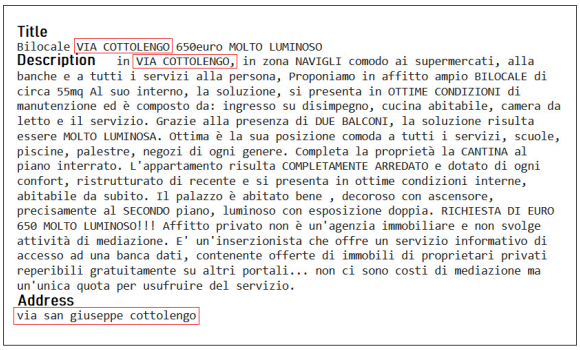

Figure 2: Example of correct extraction of the street from a complex description

3. Data integration and enrichment

We now move on to the phase of data integration and enrichment, whose source proportion is depicted in the Figure. 3.

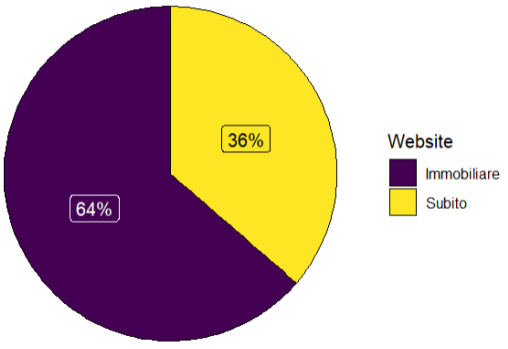

Figure 3: Data Source

We combine the JSON files obtained from the two sources by first working on a dataset that collects the common variables. This allows us to easily identify duplicates between the two sources. By applying the same procedure used earlier in paragraph 2.2 to identify duplicates, we find 35 duplicates (slightly over 1% of the total), which are identical listings found on both Subito and Immobiliare platforms. As a result, we go from a total of 2732 observations to a final value of 2697. The process of removing duplicates was repeated multiple times to accurately determine the number of listings that were actually repeated within each individual website and subsequently between the two sources. We can now consider the *Address* variable of the integrated dataset, which will serve as input for searching news articles from MilanoToday over the past year. To do this, we repeated the information extraction procedure using web scraping from the Google News platform, using the BeautifulSoup library once again. Google News was chosen instead of the news outlet's website because it allows for text-based search of multiple string patterns, making it more efficient in retrieving relevant news articles. During the scraping process, we input the street address of each observation from the integrated dataset. As output, we obtain a list of article titles from the past year that mention the specific street, along with their respective URLs. This provides us with a collection of articles that directly recount and mention events related to that location, with precise correspondence. The newly created variable *News*' consists of an array of dictionaries containing the title and URL of each article related to the street of interest. One concern was to obtain a low number of articles for less well-known streets due to the chosen granularity. However, these streets represent a percentage that is more than acceptable for our purposes, approximately 30%.

4. Data storage and queries

The integrated and enriched data consists of a JSON file that encapsulates a list of dictionaries. The database is stored in MongoDB, a document-oriented non-relational DBMS, by importing the data into the MongoDB Compass interface. In essence, a collection is available with a total of 2697 distinct documents, one for each listing, ready for analysis. A NoSQL database type was chosen for its flexibility, as it allows storing data without a fixed schema. This means that fields can be added, modified, or removed from the data without making changes to the database schema. This is suitable for our case, considering the different structure between the two sources, which

would have been impossible to transform into a relational form. Furthermore, using the PyMongo library, which enables interaction and manipulation of MongoDB from Python, we create queries to interrogate the data.

4.1. First query

First, let's analyze the distribution of prices by creating a query that returns the average monthly rental price per neighborhood. However, this information can be distorted as the price primarily depends on the size of the offered apartment. Therefore, we considered the monthly price per square meter. The results are then sorted in descending order, and the top ten "most expensive" neighborhoods are shown in the Figure. 4.

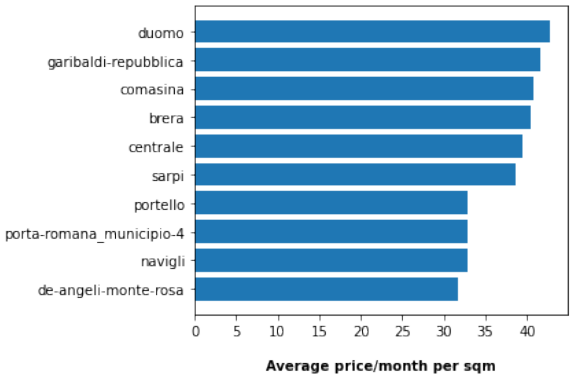

Figure 4: Top 10 districts with the highest average monthly rent per square metre ratio

As expected, the neighborhood with the highest price per square meter is Duomo, which is the city center, with an average of approximately 43 euros per square meter per month. It is followed by other famous neighborhoods such as Garibaldi (41.6 euros per square meter per month), Brera (40.5 euros per square meter per month), Centrale (39.5 euros per square meter per month), and so on. Surprisingly, the neighborhood Comasina is also among the top-ranking neighborhoods (40.7 euros per square meter per month), despite being located on the outskirts. However, upon closer analysis, it appears that this is due to the fact that there are only four listings in that area, with relatively high prices, which significantly inflates the average value.

4.2. Second query

Secondly, we have created a query to display the neighborhoods with the highest number of apartment rental listings. This is to give us an idea of the neighborhoods that contribute the most to the city's real estate market supply. The results of the top ten are shown in the figure. 5:

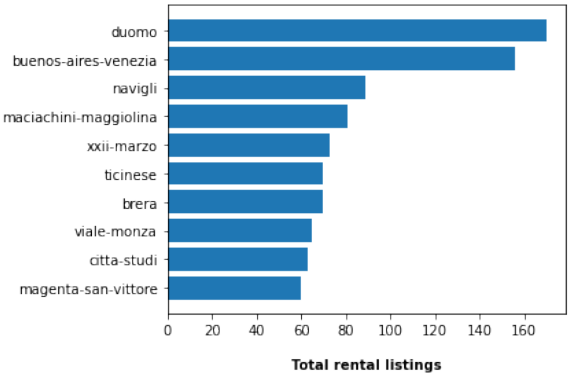

Figure 5: Top 10 districts with the highest number of apartment rental listings

The neighborhood with the highest number of listings is Duomo (170), followed by the Buenos Aires - Venezia area with 156 listings, and the Navigli with 89. The main reason why the Duomo area has the highest supply may be due to its significant size, as shown in the figure. 6.

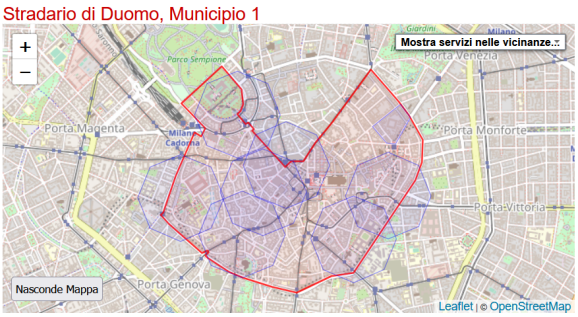

Figure 6: Duomo district

According to the city's street map, indeed, the neighborhood extends along almost the entire inner ring road.

4.3. Third query

Finally, we create a query to retrieve, for each neighborhood, the total number of articles written in the last year by MilanoToday that mention the respective streets present in the database. To achieve this, we aggregate by neighborhood and sum the number of articles for each unique street present. This is done to avoid the situation where if a street is present multiple times in the database, the corresponding articles would be further summed. This way, we obtain a good approximation of the actual value, as shown in the figure. 7.

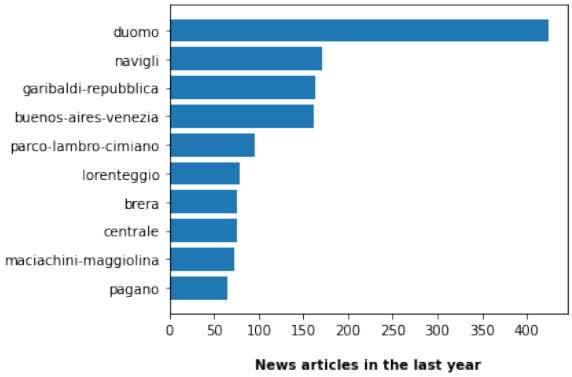

Figure 7: Top 10 districts with the highest number of MilanoToday news articles in the last year mentioning their streets

In this case as well, the Duomo neighborhood is at the top with 423 news articles mentioning the streets within it (present in the database). It is followed by the Navigli neighborhood with 171 articles and Garibaldi - Repubblica with 164. The historical and touristic significance of certain areas attracts the attention of many people and journalists, who write articles to share information, news, or interesting facts about the neighborhood. The Duomo area, in particular, is often the center of important events, celebrations, and manifestations. For example, during Milan Fashion Week, the Duomo square becomes an open-air runway for fashion shows and related events. Additionally, during the Christmas season, a large Christmas tree and a Christmas market are set up in the square. These events and celebrations generate extensive media coverage and, consequently, a higher number of articles about the area.

5. Data quality

Data quality verification is an essential phase as it enables making accurate decisions based on reliable data, potentially reducing the risk of errors and inaccuracies in the decision-making process. The data screening will cover the most critical aspects of the project, such as the completeness, accuracy, and consistency of the information. Additionally, the approaches and strategies employed to enhance the quality of the data will be highlighted.

5.1. Accuracy

Data accuracy indicates how faithfully the data reflects reality or the events it represents. This metric is divided into two parts: syntactic and semantic. The syntactic dimension evaluates the correctness of values within the reference domain, while the semantic dimension assesses the correspondence of values with real-world facts. In our case, the semantic accuracy metric was evaluated to determine if the street assigned by the function, present in the *Address* variable, was consistent with reality. This was done by examining 100 randomly extracted observation descriptions from the database. Out of these, 87 streets were assigned correctly, providing us with an estimate of the overall data accuracy. The errors can be attributed to four possible different scenarios.

1. Error in the street syntax: The street "Via Padova" is incorrectly written as "Viale Padova" and is not recognized.
2. The agency's street mentioned in the advertisement description is not present in the variable *Agency_address*.
3. The text mentions a street indicating a larger area of interest compared to the address also present in the description.
4. In cases where multiple streets are mentioned in the description, all consisting of the same number of words, the extraction among them is random.

5.2. Consistency

Data consistency can be interpreted in two ways: coherence with integrity constraints and uniformity among different representations of the same object. To evaluate data consistency, it is necessary to analyze two elements of the city's street dataset separately: neighborhoods and streets. For both items, a comparison with the official reference from the Municipality's website is required. In the case of neighborhoods, we need to verify if the ones indicated in the dataset correspond to the Local Identification Units (NIL) listed in the official source. Only if there is sufficient correspondence between the two, can we proceed with using such information. Similarly, for streets, they need to be compared with those listed in the reference source. In essence, it is crucial that both neighborhoods and streets in the dataset are coherent and correspond to the neighborhoods and streets reported in the Municipality of Milan's reference point. This is necessary to effectively utilize the association between streets and the neighborhood they belong to. From the analysis, it emerged that 86 out of 88 NILs in the official source match the list of neighborhoods in the street dataset, which counts 96 neighborhoods. The missing ones are "Ronchetto delle Rane" and "Giambellino"; however, they are included within larger neighborhoods. Subsequently, we compared the streets in the street dataset with the official list of Milan's streets, obtained from the Municipality's website. The official number of streets in the city is 4383, while our street list, after a cleaning phase that

eliminated strings with specific patterns such as trails, stations, and intersections, contains 4527 streets. We can confidently state that our street list covers all the streets present in the Municipality of Milan, and the subdivision into neighborhoods reflects the one provided by the official website. The acquired dataset is therefore reliable and consistent with reality.

5.3. Completeness

Completeness refers to the extent to which a real phenomenon is represented within a dataset and is determined by the number of non-null values present for the relevant data. In our scenario, the percentage of missing values was evaluated for the variables:

- Price/month
- Size (m^2)
- Locals
- Floor
- Bathrooms
- Address

These variables were specifically selected because we already know that all the considered observations have a title, description, URL, and so on. Additionally, some variables are not common between the two sources (for example, the agency's address is specified only in some ads from Subito.it, while certain optional features are specified only for those from Immobiliare.it). Therefore, it would not make sense to consider them. Otherwise, we would unfairly penalize a factor that is actually a strength of the non-relational database structure. From Figure 8, we can see that the variable *Bathrooms*, which indicates the number of bathrooms in the apartment, has the highest number of missing values, accounting for approximately 15% of the total. The next highest is the addresses specified in the *Address* variable, with 14.2% missing values. This suggests that the algorithm did not assign any street based on the text and description. This could be because the street is not actually specified in the ad or because the mentioned street differs from the official one (for example, the advertiser specifies "Piazzale Firenze" instead of "Piazza Firenze"). It follows that in this case, the missing values will also correspond to the*District* variable, as *Address* implies *District*.

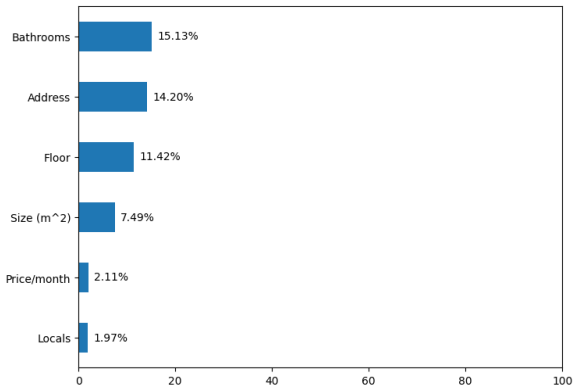

Figure 8: Percentage of missing values for the variables of interest

Finally, considering the monthly rental price, it is only not mentioned in 2% of cases.

6. Conclusions and future developments

The work carried out has successfully achieved the set goal of creating a unified and informative database for collecting rental listings from different sources. The strength lies in considering both the supply from real estate agencies and private individuals, allowing for a broader and more comprehensive view of the apartment rental market in Milan. Additionally, potential users have access to news and recent events related to the location of each individual property. However, this tool is still subject to limitations, such as the availability of up-to-date information, which relies on periodically re-executing the web scraping processes. One solution could be the implementation of an automatic and regular updating process. Furthermore, the implementation of advanced Natural Language Processing methods would allow for the exclusion of listings that mistakenly request a rental property instead of offering one. The same algorithm could also address errors and improve the extraction process of streets from descriptions and titles of the listings, where available, making it more accurate. By addressing these limitations and incorporating automated updates and advanced NLP techniques, the tool would become even more robust and efficient, enhancing the overall user experience and the reliability of the extracted data.

References

[1] https://vie.openalfa.it/milano.

[2] https://www.subito.it/annunci-lombardia/affitto/appartamenti/milano/milano/.

[3] https://www.comune.milano.it/documents/20126/3813098/regolamento+edilizio+-++approvato+con+deliberazione+del+consiglio+comunale+n.+27+del+2+ottobre+2014+e+successive+modificazioni+ed+integrazioni.pdf/02e04741-4c68-35ca-4155-eccffd7fc6e4?t=1558544993206.

[4] https://www.immobiliare.it/affitto-case/milano/?criterio=rilevanza.