POLITECNICO
MILANO 1863

Data quality Report

Project ID: 43

Silvia Macchi 10730715
Sveva Zanetti 10763926

Code Repository: Link
Dirty dataset: Comune di Milano - Stutture ricettive alberghiere
Clean dataset: Comune di Milano Clean

1 Introduction

The aim of this report is to describe the steps of the data preparation pipeline performed on the dataset "Comune di Milano - Stutture ricettive alberghiere".

The pipeline, shown in Figure 1, is divided into two main parts, the first one is about data profiling and data quality assessment, while the second part concerns the cleaning phase and aims at providing a clean dataset with formatted values and without NaN or duplicated rows.

At the end of the report there is a section that assess the quality of the cleaned dataset.

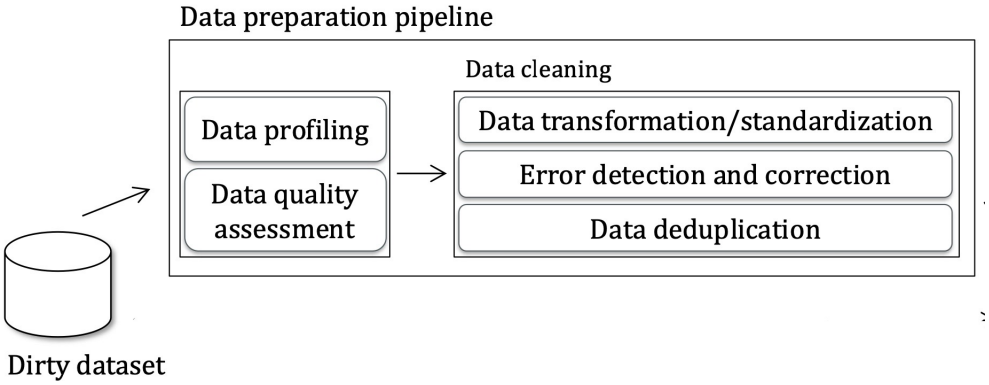

Figure 1: *Data preparation pipeline*

2 Setup choices

The environment used to run the code is python = 3.11, this allows the library "ydata_profiling" to run, and consequently to generate the html report about data profiling. Other library used are pandas, numpy, matplotlib, sklearn, seaborn, recordlinkage and itertools.

3 Data profiling

This section describes how we explored the dataset to find insights about the data that can be later used to clean the dataset.

3.1 Data exploration

The starting point is to import the data in tabular format, as shown in Figure 2.

	Ubicazione	Tipo via	Descrizione via	Civico	Codice via	ZD	Camere	Camere piano	Categoria	Insegna	Piani totali	Piano piano	Posti letto	Posti letto per piano	Tipo attività struture extra
2	CSO BUENOS AIRES N. 18 (z.d. 3)	CSO	BUENOS AIRES	18.0	2129.0	3.0	16.0	16	1	hotel aurora	1.0	1	25.0	25	Albergo
3	CSO BUENOS AIRES N. 26 (z.d. 3)	CSO	BUENOS AIRES	26.0	2129.0	3.0	25.0	NaN	3	hotel buenos aires	NaN	NaN	39.0	NaN	Albergo
4	CSO BUENOS AIRES N. 2 (z.d. 3)	CSO	BUENOS AIRES	2.0	2129.0	3.0	46.0	15;11;8	3	albergo fenice	4.0	1;2;3;4	98.0	24;19;13	Albergo
5	CSO BUENOS AIRES N. 33 (z.d. 3)	CSO	BUENOS AIRES	33.0	2129.0	3.0	65.0	0	4	GALAXY G SRL	NaN	NaN	97.0	0	Albergo
6	CSO BUENOS AIRES N. 3 (z.d. 3)	CSO	BUENOS AIRES	3.0	2129.0	3.0	116.0	4;23;24;24	4	cristoforo colombo	4.0	2;3;4;5	191.0	5;38;40;40	Albergo

Figure 2: *Row 2 to 6 of the original dataset*

Table 1 shows, for every column, the type, the number of NaN values and the number of unique values. As we can see the NaN values are not equally distributed between columns, and columns "Camere piano", "Piano piano" and "Posti letto per piano" have an unusual formatting, since they contain a list of numerical values.

Table 1: *Column analysis*

Column Name	Type	NaN values	Unique values
Ubicazione	object	0	438
Tipo via	object	14	8
Descrizione via	object	14	300
Civico	float64	30	90
Codice via	float64	14	302
ZD	float64	14	9
Camere	float64	1	148
Camere piano	object	106	207
Categoria	object	7	8
Insegna	object	10	437
Piani totali	float64	264	11
Piano piano	object	252	58
Posti letto	float64	1	196
Posti letto per piano	object	106	219
Tipo attività struture extra	object	10	3

3.2 Data quality assessment

The total number of NaN values for each column is used to compute the completeness of the dataset and the obtained value is 87.54%.

The other dimension that can be analyzed is the consistency. To do so we create some rules, reported below:

1. Camere = $\sum$ Camere piano
2. Piani totali = $\sum$ Piano piano
3. Posti letto = $\sum$ Posti letto per piano
4. Camere $\leq$ Posti letto

We obtained the following results:

1. Consistency Check 1 (Camere = $\sum$ Camere piano): 31.30%
2. Consistency Check 2 (Piani totali = $\sum$ Piano piano): 94.64%
3. Consistency Check 3 (Posti letto = $\sum$ Posti letto per piano): 26.38%
4. Consistency Check 4 (Camere $\leq$ Posti letto): 100.00%

The results are obtained by considering only the rows in which the values of the two analyzed columns are not NaN. For example, in the analysis of 'Posti letto' and 'Posti letto per piano', only 26.38% of rows without NaN values for these two columns are consistent.

Table 2 summaries some of the main aggregated information about the dataset obtained using the ydata_profiling library. As we can see before the cleaning process there were no duplicated rows.

Table 2: *Report dirty dataset*

Dataset statistics	
Number of variables	15
Number of observations	451
Missing cells	843
Missing cells (%)	12.5%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	53.0 KiB
Average record size in memory	120.3 B

4 Data cleaning

This section outlines the modifications made to the dataset to obtain a cleaned version of it.

4.1 Formatting and normalization

Table 1 shows the original columns in the dataset. We started by assigning more meaningful names to some of the columns because we thought some of the original names were a bit misleading:

- Descrizione via → Nome via
- ZD → Municipio
- Camere → Camere tot
- Camere piano → Camere per piano
- Categoria → Numero stelle
- Insegna → Nome
- Piano piano → Elenco piani
- Posti letto → Posti letto tot
- Tipo attività struture extra → Tipologia

Since "Nome", "Ubicazione", "Tipo via", "Nome via", "Tipologia" contain names and categorical words, we made them uppercase to uniform their values. As a result, the unique values of 'Tipologia' changed to just two categories, instead of three.

['Albergo' 'Residence' nan 'albergo'] → ['ALBERGO' 'RESIDENCE' nan]

The column "Numero stelle" contains these unique values:

['4' '1' '3' '5' nan 'l' '2' 'I' '5 STELLE LUSSO']

Since 'I' and 'l' are not valid values, we set them to NaN and we changed '5 STELLE LUSSO' to six in order to assign the "Int64" type to the column.

"Posti letto per piano" and "Camere per piano" have a lot of NaN values and, as shown in Section 3.1, their values are inconsistent with "Camere tot" and "Posti letto tot". For these reasons and because their formatting is not standard, as shown in Figure 3, we decided to drop the two columns.

	Ubicazione	Tipo via	Nome via	Civico	Codice via	Municipio	Camere tot	Camere per piano	Numero stelle	Nome	Piani totali	Elenco piani	Posti letto tot	Posti letto per piano	Tipologia
4	CSO BUENOS AIRES N. 2 (z.d. 3)	CSO	BUENOS AIRES	2.0	2129.0	3.0	46.0	15;11;8	3	albergo fenice	4.0	1;2;3;4	98.0	24;19;13	Albergo

Figure 3: *Dataset highlighting "Posti letto per piano" and "Camere per piano" columns*

"Elenco piani", instead, is consistent with "Piani totali", as explained in Section 3.1, and can be used to fill some of the NaN in "Piani totali". Then it can be dropped, since its data are not well structured, as shown in Figure 4. This process filled 31 of the NaN values in 'Piani totali'.

	Ubicazione	Tipo via	Nome via	Civico	Codice via	Municipio	Camere tot	Numero stelle	Nome	Piani totali	Elenco piani	Posti letto tot	Tipologia
2	CSO BUENOS AIRES N. 18 (Z.D. 3)	CSO	BUENOS AIRES	18.0	2129.0	3.0	16.0	1	HOTEL AURORA	1.0	1	25.0	ALBERGO
4	CSO BUENOS AIRES N. 2 (Z.D. 3)	CSO	BUENOS AIRES	2.0	2129.0	3.0	46.0	3	ALBERGO FENICE	4.0	1;2;3;4	98.0	ALBERGO
6	CSO BUENOS AIRES N. 3 (Z.D. 3)	CSO	BUENOS AIRES	3.0	2129.0	3.0	116.0	4	CRISTOFORO COLOMBO	4.0	2;3;4;5	191.0	ALBERGO
12	CSO EUROPA N. 9 (Z.D. 1)	CSO	EUROPA	9.0	300.0	1.0	89.0	4	HOTEL GALILEO	7.0	1;2;3;4;5;6;7	159.0	ALBERGO

Figure 4: *Dataset highlighting "Elenco piani" column*

"Ubicazione" doesn't contain NaN values, but most of the information about the address are present in other more structured columns, so we decided to use the information in "Ubicazione" to fill some of the missing values of columns "Tipo Via", "Nome via", "Civico", "Municipio".

"Tipo via", "Nome via" and "Civico" didn't have any inconsistencies with the values extracted from "Ubicazione", while "Municipio" had some inconsistent values. We decided to keep as true the values coming from "Ubicazione". After being used, we dropped "Ubicazione" since all the important data have been added to the other columns.

"Tipo via" contains these unique values:

['ALZ' 'CODVIA' 'CSO' 'GLL' 'LGO' 'PLE' 'PZA' 'VIA' 'VLE']

Since CODVIA is not a valid value, we set it to NaN so that it can be handled in Section 4.2.

Subsequently, we cast to "Int64" all the "float64" columns and we reordered the columns to give them a more logical order, we sorted the rows according to the "Nome" of the hotel.

The formatted dataset is showed in Figure 5 and the main details after this formatting phase are summarized in Table 3.

	Nome	Tipologia	Numero stelle	Tipo via	Nome via	Civico	Codice via	Municipio	Camere tot	Piani totali	Posti letto tot
244	ACCA PALACE	RESIDENCE	4	VIA	NICOTERA GIOVANNI	9	1508	9	41	<NA>	82
412	ALBERGO ACCURSIO	ALBERGO	3	VLE	CERTOSA	88	7174	8	27	<NA>	39
353	ALBERGO DEL SOLE	ALBERGO	1	VIA	SPONTINI GASPARE	6	2141	3	17	2	40
104	ALBERGO FELICE CASATI	ALBERGO	4	VIA	CASATI FELICE	18	2122	3	99	3	145
4	ALBERGO FENICE	ALBERGO	3	CSO	BUENOS AIRES	2	2129	3	46	4	98

Figure 5: *First 5 rows of the formatted dataset*

Table 3: *Column analysis after formatting*

Column Name	Type	NaN values	Unique values
Nome	object	10	435
Tipologia	object	10	2
Numero stelle	Int64	13	6
Tipo via	object	1	8
Nome via	object	1	307
Civico	Int64	0	92
Codice via	Int64	14	302
Municipio	Int64	8	9
Camere tot	Int64	1	148
Piani totali	Int64	233	11
Posti letto tot	Int64	1	196

4.2 Handling missing values

In this section we describe how we removed all the NaN values. Some of them have already been filled in the previous section using the values of "Ubicazione" and "Elenco piani", before dropping these columns.

- "Camere tot" and "Posti letto tot" have just one NaN value, so we decided to start from these columns. The Nan value is in the same row for both columns and also the name of the hotel is NaN for this row, so we decided to drop the entire row.

	Nome	Tipologia	Numero stelle	Tipo via	Nome via	Civico	Codice via	Municipio	Camere tot	Piani totali	Posti letto tot
322	NaN	NaN	<NA>	VIA	SANTA RADEGONDA	14	<NA>	1	<NA>	<NA>	<NA>

Figure 6: *Row with NaN in "Camere tot" and "Posti letto tot"*

- "Nome" has now 9 NaN values remaining, they all have information about the address and the number of rooms and beds, so we decided to keep the rows and assign an "UNKNOWN" value to these rows, since they cannot be predicted.
- "Tipologia" has 9 NaN values, but since its value can be just "HOTEL" or "RESIDENCE" we firstly looked at the "Nome" column to see if the name contained the word "HOTEL" or "RESIDENCE" and then we filled the single NaN value remaining with the mode of the column, that is "HOTEL".
- "Numero stelle" has 12 NaN values, but the data we have are not so useful to infer the missing values, so we used -1 to indicate that the value is not known.
- "Municipio" has 8 NaN values, one of them has been recovered from a row with the same "Tipo via" and "Nome via", for the others a standard value 0 has been used.
- "Codice via" has 13 NaN values, 6 of them have been recovered from rows with the same "Tipo via" and "Nome via", for the others a standard value 0000 has been used.
- "Tipo via" and "Nome via" have just one missing value, and since they cannot be inferred we used the "UNKNOWN" value for both columns.

- "Piani totali" has 232 missing values, we could have eliminated the column, but since it's a numeric value that is correlated with "Camere tot" and "Posti letto tot" and weakly correlated to "Numero stelle" we decided to use a linear regression model to infer the missing values of "Piani totali".

  Figure 7 shows the histogram of the distribution of "Piani totali" before filling the missing values (red) and after (light blue).

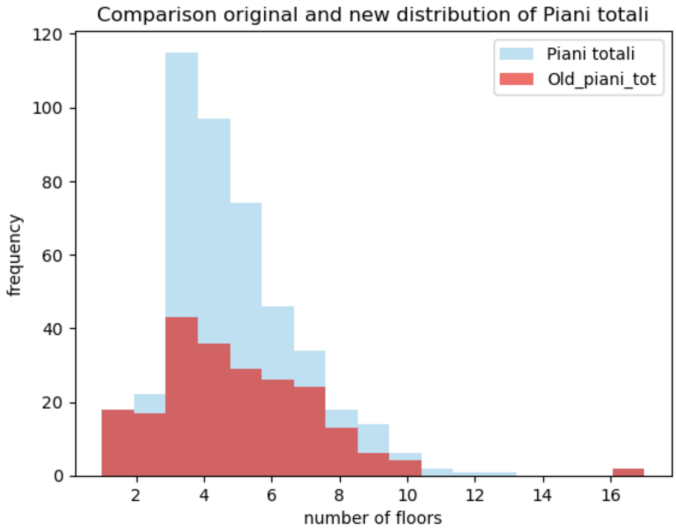

Figure 7: *Histogram of "Piani totali" before and after filling NaN values*

After all these steps, the missing values in the dataset have all been filled and the dataset has now 450 rows and 11 columns.

4.3 Outlier detection

To identify the outlier we produce different boxplots, reported in Figure 8.

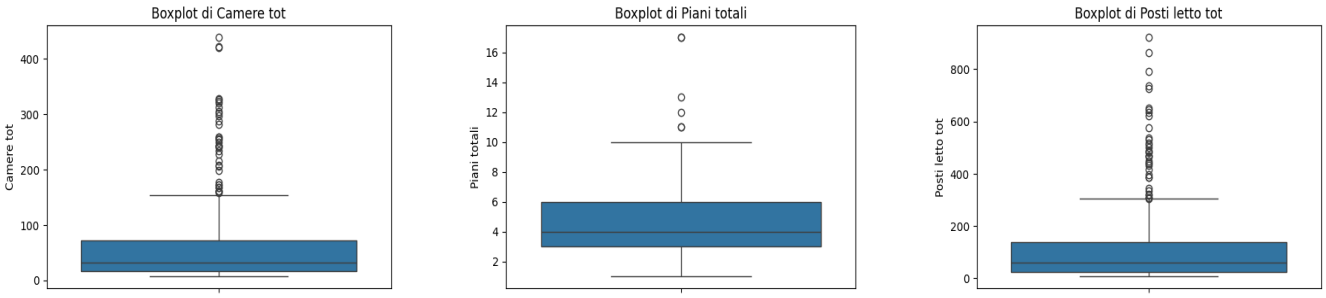

Figure 8: *Boxplot of "Camere tot", "Piani totali", "Posti letto tot"*

Even though the boxplots show that some values are quite far from the distribution of the relative column, we think that since they are referring to hotels in a big city such as Milan all the results are plausible and we decided to keep them. We checked also the distribution of values in "Civico", but the higher one is 371 which is plausible.

The only outliers detected were the ones found during the formatting phase of the categorical columns done in Section 4.1, such as "I" and "L" in "Numero stelle" and "CODVIA" in "Tipo via".

4.4 Data deduplication

In order to define which data is duplicated, we first look for an exact match. Since there isn't one, we need to define some rules to identify possible duplicate rows.

The defined rules are 3, the first two compare attributes based on similarity measure:

- comparison of strings in 'Nome' with the 'jarowinkler' method and a threshold of 90% similarity
- comparison of strings in 'Nome via' with the 'jarowinkler' method and a threshold of 90% similarity

The third one is defined to have an exact match:

- comparison of string in 'Codice via'

We chose these 3 rules because we think that it's very unlikely that two hotel with very similar names are in the same street, it's more likely that they are the same hotel, that maybe underwent some changes and renovation that changed the number of floors, beds and rooms.

With this method we found 9 duplicates, as reported in Figure 9.

As we can see from the reported table with this method we found hotels with names composed by the same words, such as "Green House Residence" and "Green House", as well as hotels with similar names in the same street, such as "Hotel Bernina" and "Hotel Berna".

For each of them we decided to remove the one with smaller amount of total rooms, assuming that establishments with more rooms are linked to the most recent data.

	Nome	Tipologia	Numero stelle	Tipo via	Nome via	Civico	Codice via	Municipio	Camere tot	Piani totali	Posti letto tot
277	AMBROSIANA	ALBERGO	1	VIA	PLINIO CAIO SECONDO	22	2144	3	16	3	22
276	AMBROSIANA	ALBERGO	1	VIA	PLINIO CAIO SECONDO	22	2144	3	14	3	22
422	GREEN HOUSE RESIDENCE	RESIDENCE	4	VLE	FAMAGOSTA	48	5353	6	54	3	97
423	GREEN HOUSE	ALBERGO	3	VLE	FAMAGOSTA	50	5353	6	45	4	92
302	HOTEL ADLER	ALBERGO	3	VIA	RICORDI GIOVANNI	10	2251	3	23	5	43
301	HOTEL ADLER	ALBERGO	-1	VIA	RICORDI GIOVANNI	10	2251	3	23	2	47
377	HOTEL BERNINA	ALBERGO	3	VIA	TORRIANI NAPO	27	2126	2	44	4	75
373	HOTEL BERNA	ALBERGO	4	VIA	TORRIANI NAPO	18	2126	2	122	6	197
106	HOTEL BRIANZA	ALBERGO	3	VIA	CASTALDI PANFILO	16	2119	3	25	2	36
109	HOTEL BAVIERA	ALBERGO	4	VIA	CASTALDI PANFILO	7	2119	2	52	4	92
328	HOTEL CRISTALLO	ALBERGO	3	VIA	SCARLATTI DOMENICO	22	2136	3	104	6	223
330	HOTEL BRISTOL	ALBERGO	4	VIA	SCARLATTI DOMENICO	32	2136	2	68	5	116
7	HOTEL MINERVA	ALBERGO	3	CSO	COLOMBO CRISTOFORO	15	5114	6	39	4	60
8	HOTEL MINERVA	ALBERGO	3	CSO	COLOMBO CRISTOFORO	15	5114	6	44	4	67
47	HOTEL PRINCIPE DI SAVOIA	ALBERGO	5	PZA	DELLA REPUBBLICA	17	1055	2	302	11	725
48	HOTEL PRINCIPE DI SAVOIA	ALBERGO	-1	PZA	DELLA REPUBBLICA	17	1055	2	313	9	623
211	HOTEL VIOLA	ALBERGO	1	VIA	LULLI GIOVANNI	4	2253	3	9	3	14
208	HOTEL LORIS	ALBERGO	1	VIA	LULLI GIOVANNI	11	2253	3	22	3	25

Figure 9: *Duplicated records*

5 Results: data quality assessment on cleaned dataset

Figure 10 shows how the dataset looks after the cleaning process.

	Nome	Tipologia	Numero stelle	Tipo via	Nome via	Civico	Codice via	Municipio	Camere tot	Piani totali	Posti letto tot
0	ACCA PALACE	RESIDENCE	4	VIA	NICOTERA GIOVANNI	9	1508	9	41	5	82
1	ALBERGO ACCURSIO	ALBERGO	3	VLE	CERTOSA	88	7174	8	27	4	39
2	ALBERGO DEL SOLE	ALBERGO	1	VIA	SPONTINI GASPARE	6	2141	3	17	2	40
3	ALBERGO FELICE CASATI	ALBERGO	4	VIA	CASATI FELICE	18	2122	3	99	3	145
4	ALBERGO FENICE	ALBERGO	3	CSO	BUENOS AIRES	2	2129	3	46	4	98

Figure 10: *First 5 rows of cleaned dataset*

Redundant, inconsistent and badly formatted columns have been removed and all the NaN values have been filled, as reported in Table 4.

Table 4: *Column analysis after formatting*

Column Name	Type	NaN values	Unique values
Nome	object	0	431
Tipologia	object	0	2
Numero stelle	Int64	0	7
Tipo via	object	0	9
Nome via	object	0	307
Civico	Int64	0	92
Codice via	Int64	0	303
Municipio	Int64	0	10
Camere tot	Int64	0	147
Piani totali	Int64	0	14
Posti letto tot	Int64	0	194

Table 5 shows some aggregated data provided by the report generated with the library ydata profiling. A total of 10 rows and 4 columns have been removed from the original dataset.

Table 5: *Report cleaned dataset*

Dataset statistics	
Number of variables	11
Number of observations	441
Missing cells	0
Missing cells (%)	0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	38.0 KiB
Average record size in memory	88.3 B