Analysis of Rental Data for Out-of-Town Students in the Metropolitan City of Milan

Iarocci Luca 894066, Mondella Nicholas 859673,
Prati Davide 845926

Data Management

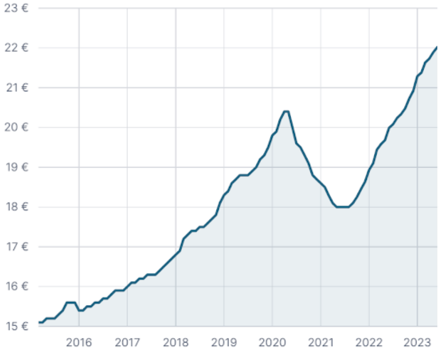
Università degli Studi di Milano-Bicocca
Year 2022/2023

18-07-23

Summary

1. Introduction
2. Sources
3. Database Model
4. Data Extraction
5. Transformation and load approaches
6. Data Cleaning and Integration
7. Data Quality
8. Data Exploration

Due to the recent hike in rent prices, especially in big cities like Milan, off-site students are having a hard time finding an accommodation which is both cheap and near to university. In response to that, students all over the country started organizing protests and demonstrations.



Average rental price on m^2 in metropolitan city of Milan

Introduction - Goal

The goal of this project is to create a tool that may help off-site students find the best housing solution in Milan, taking into consideration the following features:

- characteristics of the accommodation (surface, rooms, location, etc.)
- accommodation monthly rent
- travel time needed for reaching the university daily
- accommodation proximity to the city's attractions and interesting locations

Sources

- **immobiliare.it**: rent listings and related information
- **UrbiStat**: demographic information about municipalities
- **OpenStreetMap**: public transport stations information
- **Google Maps**: travel times computations

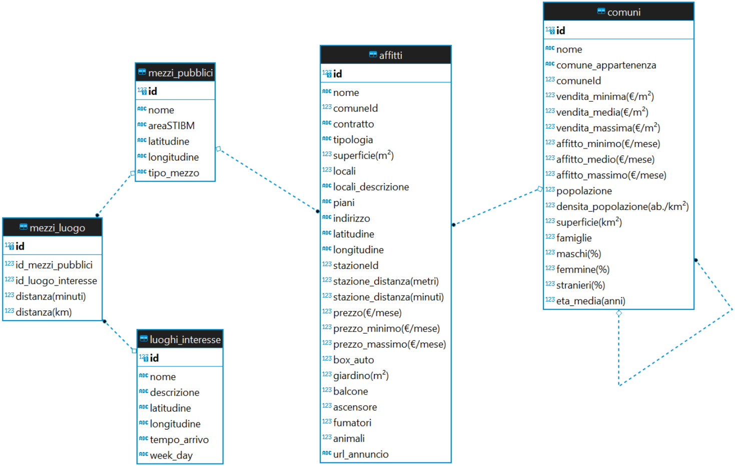
Points of interest were chosen by us and manually listed in a CSV file

Database Model - Data Storage

The proposed solution is a relational model database. This choice was motivated by the following:

- data follows a rigid schema, information is well-structured and mostly standardized (ex. price, surface, rooms, position, etc.)
- relations between different sources information are few and simple
- the data volume and the scope of the project do neither require nor benefit from a distributed approach

Database Model - Schema ER



mezzi_pubblici
id
nome
areaSTIBM
latitudine
longitudine
tipo_mezzo

mezzi_luogo
id
id_mezzi_pubblici
id_luogo_interesse
distanza(minuti)
distanza(km)

luoghi_interesse
id
nome
descrizione
latitudine
longitudine
tempo_arrivo
week_day

affitti
id
nome
comuneId
contratto
tipologia
superficie(m^2)
locali
locali_descrizione
piani
indirizzo
latitudine
longitudine
stazioneId
stazione_distanza(metri)
stazione_distanza(minuti)
prezzo(€/mese)
prezzo_minimo(€/mese)
prezzo_massimo(€/mese)
box_auto
giardino(m^2)
balcone
ascensore
fumatori
animali
url_annuncio

comuni
id
nome
comune_appartenenza
comuneId
vendita_minima(€/m^2)
vendita_media(€/m^2)
vendita_massima(€/m^2)
affitto_minimo(€/mese)
affitto_medio(€/mese)
affitto_massimo(€/mese)
popolazione
densita_popolazione(ab./km^2)
superficie(km^2)
famiglie
maschi(%)
femmine(%)
stranieri(%)
eta_media(anni)

Data Extraction - 1

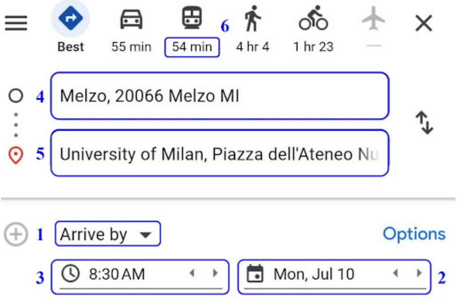
An automated data scraping solution was implemented using the BeautifulSoup4 library for extracting information from immobiliare.it and UrbiStat. Said solution uses the site's structure and manipulates the page URL accordingly to obtain the desired information:

- ➢ immobiliare.it
  - ○ from municipalities summary pages we gathered information about the real estate market
  - ○ from each zone's listing page we gathered information about all rent offers listed
- ➢ UrbiStat
  - ○ from municipalities summary pages we gathered demographic information

Data Extraction - 2

The desired information about public transport station was gathered from OpenStreetMap using OverpassAPI through specific queries.

Regarding travel times, a solution based on the Selenium was implemented. Through means of browser automation Google Maps was set up and queried for all stations and locations pairs returning the desired travel times.



Best | 55 min | 54 min | 4 hr 4 | 1 hr 23

4 Melzo, 20066 Melzo MI

5 University of Milan, Piazza dell'Ateneo Nu

1 Arrive by

3 8:30 AM

Mon, Jul 10 2

Transformation and load approaches

The ETL processes implemented for the different sources was:

- “vertical” record-based approach (immobiliare.it, Urbistat)
  - each listing is extracted, transformed and loaded one at a time
- “horizontal” dataset-based approach (OpenStreetMap)
  - the dataset was extracted, transformed and loaded at once
- “hybrid” batch-based approach (Google maps)
  - travel time and distance from each station to a fixed location was extracted, transformed and loaded one location at a time

Data Cleaning - 1

- immobiliare.it
  - missing numerical data and tags were replaced with zeros
  - unspecified rent prices labeled as “prezzo su richiesta” (rent on demand) were assigned a flag value of zero
  - prices, originally strings, were corrected for italian format (ex 2.000 to 2000) and casted to integers
- Urbistat
  - missing numerical data were replaced with zeros

Data Cleaning - 2

- OpenStreetMaps
  - all excess information was removed by a feature selection process
  - stations with missing geographical coordinates were excluded
- Google Maps
  - travel times, extracted as an “h ora mm min”-type string, were converted to minutes and casted to integers. (ex. 1h 10min to 70)
- URL-based listings deduplication (possible presence in more than one zone’s list)

Data Integration

- Demographic and real estate market data for each zone was joined in COMUNI by zone's name match criteria
- rent listings (AFFITTI) were linked to the respective zone (COMUNI)
- rent listings (AFFITTI) were linked to the nearest station (MEZZI_PUBBLICI)
- MEZZI_LUOGO was populated by querying Google Maps on all the possible combinations of (and linked to) MEZZI_PUBBLICI + LUOGHI_INTERESSE

Data Quality - Completeness

- COMUNI dataset: table completeness = 73%
  Critical attributes are the demographic ones from Urbistat and minimum and maximum sales and rentals data from immobiliare.it

- AFFITTI dataset: table completeness = 98.92%
  Critical attributes are price and minimum and maximum price, from immobiliare.it

- MEZZI_PUBBLICI dataset: table completeness = 98.8%
  The only critical attribute is *areaSTIBM*, from OpenStreetMap

Data Quality - Accuracy / Currency and Timeliness

Accuracy:

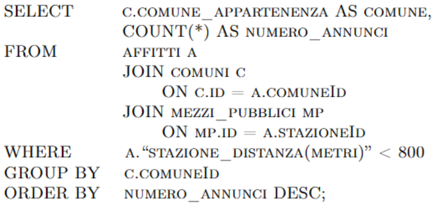
- the accuracy of our datasets has been significantly improved during the data cleaning phase
- tradeoff between accuracy and completeness
- the case of the pair of coordinates (42.76290, 11.11280)

Currency and Timeliness:

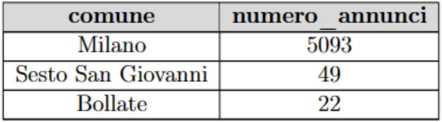
- update of the data from immobiliare.it on a daily basis
- update of the data from OpenStreetMap on a weekly basis
- the case of Repetti metro station and of the whole M4 subway line

Data Exploration - Query 1

Goal: to find, for each municipality, the number of listings, sorted in descending order, which are located within 800 meters of the nearest station

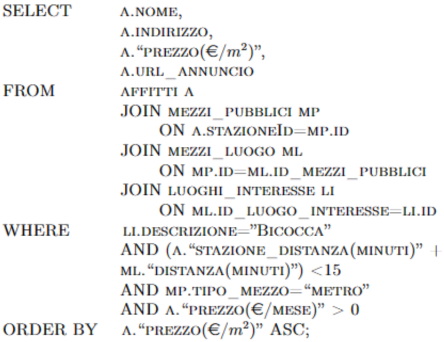


```
SELECT     C.COMUNE_APPARTENENZA AS COMUNE,
           COUNT(*) AS NUMERO_ANNUNCI
FROM       AFFITTI A
           JOIN COMUNI C
                ON C.ID = A.COMUNEID
           JOIN MEZZI_PUBBLICI MP
                ON MP.ID = A.STAZIONEID
WHERE      A."STAZIONE_DISTANZA(METRI)" < 800
GROUP BY   C.COMUNEID
ORDER BY   NUMERO_ANNUNCI DESC;
```

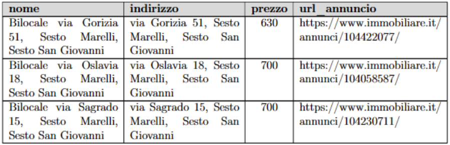


comune	numero_annunci
Milano	5093
Sesto San Giovanni	49
Bollate	22

Goal: to find apartments that are less than a quarter-hour away from Bicocca University by metro



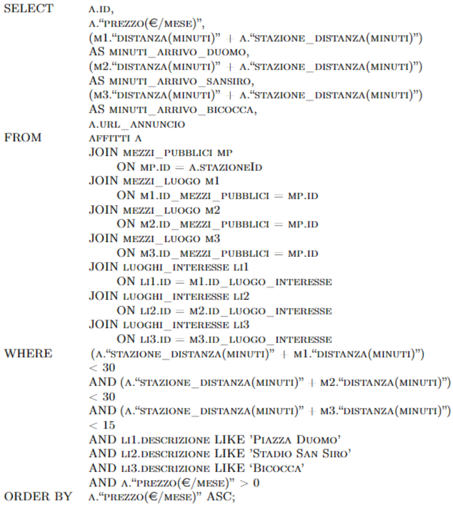
```
SELECT      A.NOME,
            A.INDIRIZZO,
            A.“PREZZO(€/$m^2$)”,
            A.URL_ANNUNCIO
FROM        AFFITTI A
            JOIN MEZZI_PUBBLICI MP
                ON A.STAZIONEID=MP.ID
            JOIN MEZZI_LUOGO ML
                ON MP.ID=ML.ID_MEZZI_PUBBLICI
            JOIN LUOGHI_INTERESSE LI
                ON ML.ID_LUOGO_INTERESSE=LI.ID
WHERE       LI.DESCRIZIONE=”BICOCCA”
            AND (A.“STAZIONE_DISTANZA(MINUTI)” +
            ML.“DISTANZA(MINUTI)”) <15
            AND MP.TIPO_MEZZO=“METRO”
            AND A.“PREZZO(€/MESE)” > 0
ORDER BY    A.“PREZZO(€/$m^2$)” ASC;
```



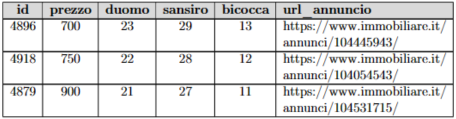
nome	indirizzo	prezzo	url_annuncio
Bilocale via Gorizia 51, Sesto Marelli, Sesto San Giovanni	via Gorizia 51, Sesto Marelli, Sesto San Giovanni	630	https://www.immobiliare.it/annunci/104422077/
Bilocale via Oslavia 18, Sesto Marelli, Sesto San Giovanni	via Oslavia 18, Sesto Marelli, Sesto San Giovanni	700	https://www.immobiliare.it/annunci/104058587/
Bilocale via Sagrado 15, Sesto Marelli, Sesto San Giovanni	via Sagrado 15, Sesto Marelli, Sesto San Giovanni	700	https://www.immobiliare.it/annunci/104230711/

Data Exploration - Query 3

Goal: to find apartments that are less than a quarter-hour away from Bicocca University, and less than half a hour away from both Duomo and San Siro, all three via metro



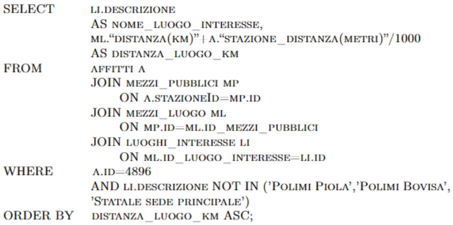
```
SELECT      A.ID,
            A.“PREZZO(€/MESE)”,
            (M1.“DISTANZA(MINUTI)” + A.“STAZIONE_DISTANZA(MINUTI)”)
            AS MINUTI_ARRIVO_DUOMO,
            (M2.“DISTANZA(MINUTI)” + A.“STAZIONE_DISTANZA(MINUTI)”)
            AS MINUTI_ARRIVO_SANSIRO,
            (M3.“DISTANZA(MINUTI)” + A.“STAZIONE_DISTANZA(MINUTI)”)
            AS MINUTI_ARRIVO_BICOCCA,
            A.URL_ANNUNCIO
FROM        AFFITTI A
            JOIN MEZZI_PUBBLICI MP
                ON MP.ID = A.STAZIONEID
            JOIN MEZZI_LUOGO M1
                ON M1.ID_MEZZI_PUBBLICI = MP.ID
            JOIN MEZZI_LUOGO M2
                ON M2.ID_MEZZI_PUBBLICI = MP.ID
            JOIN MEZZI_LUOGO M3
                ON M3.ID_MEZZI_PUBBLICI = MP.ID
            JOIN LUOGHI_INTERESSE LI1
                ON LI1.ID = M1.ID_LUOGO_INTERESSE
            JOIN LUOGHI_INTERESSE LI2
                ON LI2.ID = M2.ID_LUOGO_INTERESSE
            JOIN LUOGHI_INTERESSE LI3
                ON LI3.ID = M3.ID_LUOGO_INTERESSE
WHERE       (A.“STAZIONE_DISTANZA(MINUTI)” + M1.“DISTANZA(MINUTI)”)
            < 30
            AND (A.“STAZIONE_DISTANZA(MINUTI)” + M2.“DISTANZA(MINUTI)”)
            < 30
            AND (A.“STAZIONE_DISTANZA(MINUTI)” + M3.“DISTANZA(MINUTI)”)
            < 15
            AND LI1.DESCRIZIONE LIKE ’PIAZZA DUOMO’
            AND LI2.DESCRIZIONE LIKE ’STADIO SAN SIRO’
            AND LI3.DESCRIZIONE LIKE ‘BICOCCA’
            AND A.“PREZZO(€/MESE)” > 0
ORDER BY    A.“PREZZO(€/MESE)” ASC;
```



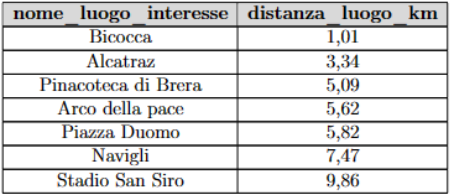
id	prezzo	duomo	sansiro	bicocca	url_annuncio
4896	700	23	29	13	https://www.immobiliare.it/annunci/104445943/
4918	750	22	28	12	https://www.immobiliare.it/annunci/104054543/
4879	900	21	27	11	https://www.immobiliare.it/annunci/104531715/

Data Exploration - Query 4

Goal: once an apartment is chosen from the previous query results (we chose the first one, id=4896), find the distance of that apartment from all the recreational points of interest and from the attended university (Bicocca)



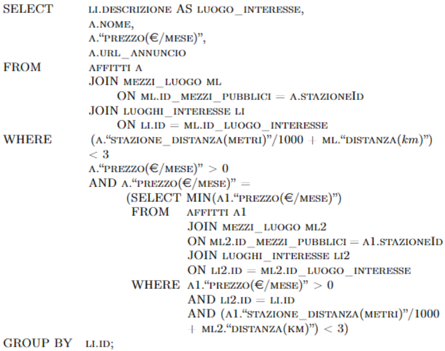
```
SELECT      li.descrizione
            AS nome_luogo_interesse,
            ml.“distanza(km)”+a.“stazione_distanza(metri)”/1000
            AS distanza_luogo_km
FROM        affitti a
            JOIN mezzi_pubblici mp
                ON a.stazioneId=mp.id
            JOIN mezzi_luogo ml
                ON mp.id=ml.id_mezzi_pubblici
            JOIN luoghi_interesse li
                ON ml.id_luogo_interesse=li.id
WHERE       a.id=4896
            AND li.descrizione NOT IN (’Polimi Piola’,’Polimi Bovisa’,
            ’Statale sede principale’)
ORDER BY    distanza_luogo_km ASC;
```



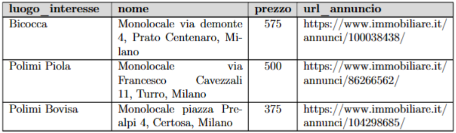
nome_luogo_interesse	distanza_luogo_km
Bicocca	1,01
Alcatraz	3,34
Pinacoteca di Brera	5,09
Arco della pace	5,62
Piazza Duomo	5,82
Navigli	7,47
Stadio San Siro	9,86

Data Exploration - Query 5

Goal: to extract the apartment with the minimum price within a 3-kilometer radius from each point of interest



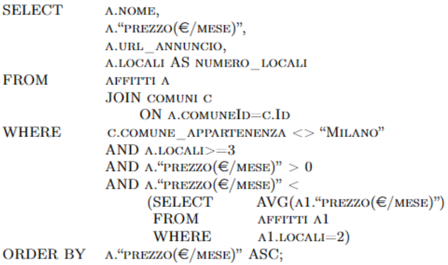
```
SELECT      LI.DESCRIZIONE AS LUOGO_INTERESSE,
            A.NOME,
            A."PREZZO(€/MESE)",
            A.URL_ANNUNCIO
FROM        AFFITTI A
            JOIN MEZZI_LUOGO ML
                ON ML.ID_MEZZI_PUBBLICI = A.STAZIONEID
            JOIN LUOGHI_INTERESSE LI
                ON LI.ID = ML.ID_LUOGO_INTERESSE
WHERE       (A."STAZIONE_DISTANZA(METRI)"/1000 + ML."DISTANZA(km)")
            < 3
            A."PREZZO(€/MESE)" > 0
            AND A."PREZZO(€/MESE)" =
                (SELECT MIN(A1."PREZZO(€/MESE)")
                 FROM    AFFITTI A1
                         JOIN MEZZI_LUOGO ML2
                         ON ML2.ID_MEZZI_PUBBLICI = A1.STAZIONEID
                         JOIN LUOGHI_INTERESSE LI2
                         ON LI2.ID = ML2.ID_LUOGO_INTERESSE
                 WHERE   A1."PREZZO(€/MESE)" > 0
                         AND LI2.ID = LI.ID
                         AND (A1."STAZIONE_DISTANZA(METRI)"/1000
                         + ML2."DISTANZA(KM)") < 3)
GROUP BY    LI.ID;
```



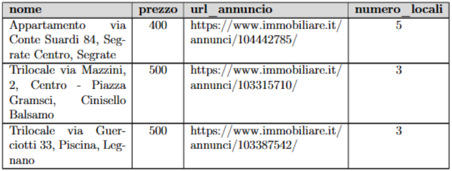
luogo_interesse	nome	prezzo	url_annuncio
Bicocca	Monolocale via demonte 4, Prato Centenaro, Milano	575	https://www.immobiliare.it/annunci/100038438/
Polimi Piola	Monolocale via Francesco Cavezzali 11, Turro, Milano	500	https://www.immobiliare.it/annunci/86266562/
Polimi Bovisa	Monolocale piazza Prealpi 4, Certosa, Milano	375	https://www.immobiliare.it/annunci/104298685/

Data Exploration - Query 6

Goal: to extract the three-bedroom and four-bedroom apartments (or even more) outside of Milan that have a rental price lower than the average rental price of two-bedroom apartments, sorted by price



```
SELECT      A.NOME,
            A."PREZZO(€/MESE)",
            A.URL_ANNUNCIO,
            A.LOCALI AS NUMERO_LOCALI
FROM        AFFITTI A
            JOIN COMUNI C
                ON A.COMUNEID=C.ID
WHERE       C.COMUNE_APPARTENENZA <> "MILANO"
            AND A.LOCALI>=3
            AND A."PREZZO(€/MESE)" > 0
            AND A."PREZZO(€/MESE)" <
                (SELECT    AVG(A1."PREZZO(€/MESE)")
                 FROM      AFFITTI A1
                 WHERE     A1.LOCALI=2)
ORDER BY    A."PREZZO(€/MESE)" ASC;
```



nome	prezzo	url_annuncio	numero_locali
Appartamento via Conte Suardi 84, Segrate Centro, Segrate	400	https://www.immobiliare.it/annunci/104442785/	5
Trilocale via Mazzini, 2, Centro - Piazza Gramsci, Cinisello Balsamo	500	https://www.immobiliare.it/annunci/103315710/	3
Trilocale via Guerciotti 33, Piscina, Legnano	500	https://www.immobiliare.it/annunci/103387542/	3

Thank you for your attention!