Università degli studi di Milano Bicocca

Master's degree in Data Science

Data Management Project

Analysis of Milan's restaurants using TripAdvisor and Google data to create a recommendation system

Mondini Matteo - 902873 - m.mondini11@campus.unimib.it
Costa Patrick - 858156 - p.costa7@campus.unimib.it

Abstract

Our project is related to restaurant data retrieval, integration, exploratory analysis, sentiment analysis and the development of a recommendation system. The primary goal was to collect information from TripAdvisor and Google related to restaurants in Milan using web scraping techniques and api. The gathered data was then cleaned, integrated, cleaned again and subjected to exploratory analysis to gain valuable insights into the restaurant landscape in Milan.

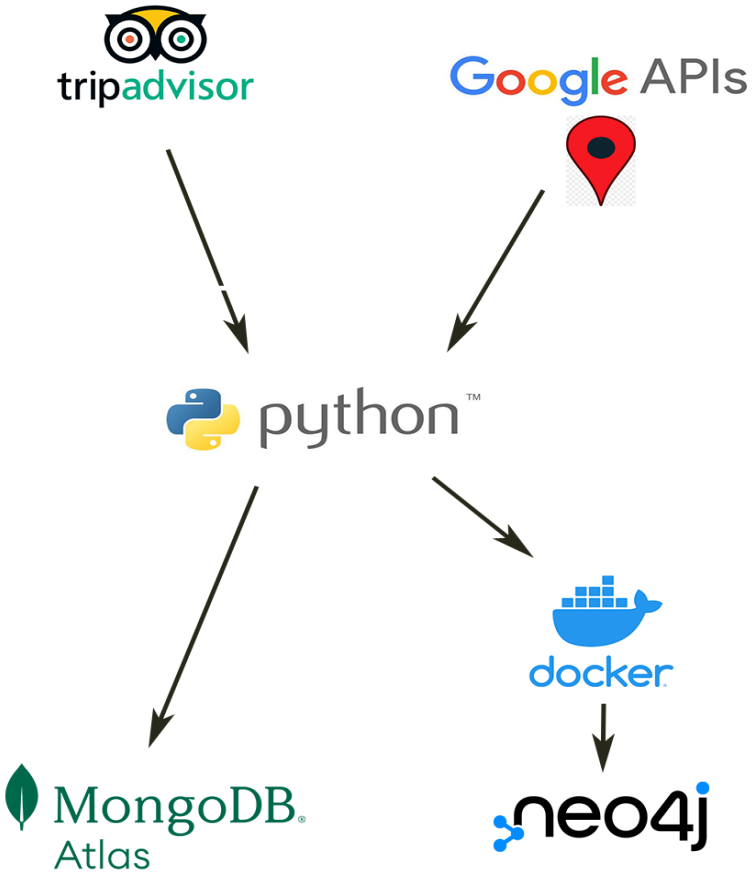
We aim to reply to the following research questions:

1. How effective is sentiment analysis in understanding feelings towards restaurants in Milan based on comments retrieved from restaurants customers?
2. What are the insights gained from the exploratory analysis of restaurant data in Milan, and how do these insights contribute to understanding the restaurant landscape in the city?
3. How used techniques are powerful and effective in order to use the retrieved data to perform reliable analysis?

A sentiment analysis was performed based on comments in order to understand the feelings that customers have for restaurants and for other analysis concerning the best and the worst restaurants in the city.

After the exploratory analysis, the collected data was stored in both MongoDB and Neo4j databases. MongoDB provided a flexible and scalable solution for storing and querying the restaurant information, while Neo4j, allowed us to model the data as a graph and establish relationships between restaurants based on various constraints to develop the basis for a recommendation system that provide restaurant recommendations to users based on the similarity between restaurants .

The entire process, from collecting data through web scraping to integrating it, conducting exploratory and sentiment analysis, storing data in MongoDB, and implementing queries, was showcased. The report demonstrates the potential of using web scraping, data integration, and graph databases for deriving insights and creating valuable applications in the restaurant domain. The findings and techniques presented in this report can serve as a foundation for future research and development in the field of recommendation systems and data-driven decision-making in the restaurant industry.



1. Working pipeline main technologies

Index

Data discovery

Trip advisor scraping made with selenium

Link to GitHub repository:
https://github.com/darthgween/data-management-project/tree/definitive/1_trip_advisor_scraping

To retrieve the data present on the Trip Advisor platform, a Python script was developed using the Selenium library. This script was specifically designed to scrape restaurant information by leveraging the capabilities of Selenium. The main reason for choosing Selenium was its ability to handle the JavaScript-based pagination implemented on Trip Advisor's restaurant pages.

The script implemented an algorithm that simulated user behavior to navigate through the pages and extract the desired data. One of the challenges encountered was the need to accept cookies by clicking the corresponding button. This was essential to ensure the correct flow for the scraping. By emulating user actions, such as accepting cookies and scrolling the page, the script was able to access and retrieve information for all the restaurants listed.

Another minor snag in the scraping process was excluding sponsored restaurants that appeared every 5 restaurants. However, by inspecting the source code, the script successfully overcame this obstacle and only scraped non-sponsored restaurants, ensuring accurate and relevant data retrieval.

Another challenge was due to the asynchronous nature of web pages, where instructions sometimes executed before the page fully loaded. To address this issue, we employed various techniques. Firstly, we utilized time delays (via the "sleep" function) to ensure that instructions

were executed only after the page had loaded completely. Additionally, the script made use of Selenium's "waitUntil" function, which allowed it to wait until the desired elements were visible on the page before proceeding with further actions.

During the execution of the scraping loop, occasional errors occurred when the page was not properly scrolled to facilitate data extraction or when the page failed to load correctly. To handle these situations, the script incorporated exception handling, allowing it to gracefully recover from such errors. Selenium's capabilities were leveraged to virtually scroll the page to the desired location, ensuring that all relevant information was captured.

Sometimes happened as well that Trip Advisor security systems captured our scraping try and they blocked us to interact from a automated page. In order to solve this issue we used fake user agent to around the problem and take the data we needed.

Furthermore, to perform sentiment analysis and retrieve the addresses of the restaurants (which were not available in the restaurant list but only on each individual restaurant's page), the script needed to access each page separately. In order to retrieve the necessary data, a similar script, resembling the one previously described, would need to be executed. By applying similar scraping techniques, including handling JavaScript-based pagination, the script would extract the first 10 reviews from each restaurant's page. We choose to get just the first five reviews for time reasons. Of course our sentiment analysis is weak with only 10 comments, this may be improve in the further developments.

Considering the time-intensive nature of these algorithms, efforts were made to optimize the execution process. Multi-threading was employed to parallelize certain tasks, resulting in modest improvements in overall performance and execution time.
In summary, our Python script, overcame various challenges to successfully retrieve data from the Trip Advisor platform. Through simulated user interactions, handling of JavaScript-based pagination,

and addressing page loading issues, the script efficiently extracted restaurant information while ensuring data accuracy.

Google API

Link to GitHub repositories:
https://github.com/darthgween/data-management-project/tree/definitive/2_google_api

The second source of data collection to build the dataset was Google, one of the broadest and most popular platforms that offers a wide range of services and information. To access the necessary data for our project, we utilized an API provided by Google. The use of an API allowed us to connect to Google's Places system and extract structured and automated information about restaurants in Milan.

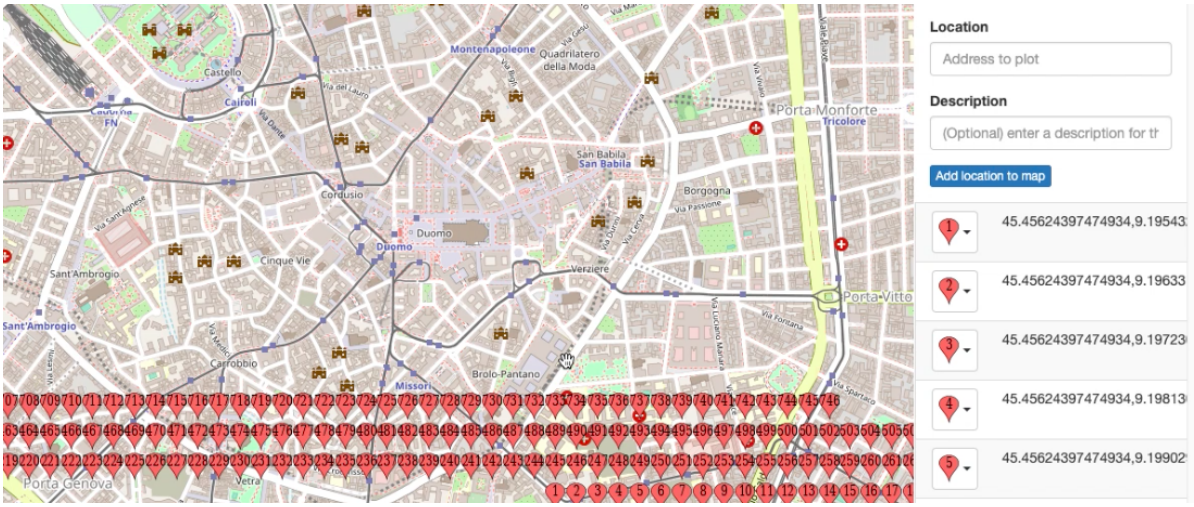
In order to obtain the desired data, we implemented a code that leveraged the functionalities offered by Google Places API. Specifically, we performed a specific text search using the location parameters of Milan and the term "restaurant." This enabled us to acquire a set of significant information for each restaurant, such as its name, address, and other relevant details.

Problems and solutions with the api

However, we encountered some challenges along the way. The biggest obstacle was the fact that Google provides a maximum of 60 restaurant outputs for each API call. Since we needed thousands of restaurants, we developed a clever system to bypass this problem. We created an algorithm capable of generating a network of points across

the entire map of Milan, with each point being 100 meters apart. Iteratively, we made API calls for each point within a 100-meter radius. The results obtained from each request were meticulously processed and aggregated into a comprehensive dataset containing all the restaurant information. This was a computationally expensive operation and pushed the boundaries of legality (Google temporarily blocked our account due to scraping attempts). However, in the end, we achieved the desired outcome: the complete dataset of restaurants in Milan from Google.

The utilization of Google Places API was crucial in integrating our dataset and ensuring its completeness and reliability. Thanks to this data source, we were able to obtain updated and detailed information about the restaurants in Milan, enabling us to conduct a thorough and comprehensive analysis.



Data cleaning

Link to GitHub repository:
https://github.com/darthgween/data-management-project/tree/definitive/3_data_cleaning

The data cleaning process begins with loading the restaurant data obtained from TripAdvisor scraping process. The first step in the cleaning process is to remove any count numbers that may be present at the beginning of restaurant names in the TripAdvisor dataset. By

removing the count numbers, we ensure more accuracy for the integration and enrichment process.
Next, the euro symbols in the 'expensive' column of the TripAdvisor dataset are replaced with numerical values to indicate the price level of each restaurant. This transformation allows for easier comparison and analysis of restaurant pricing.
To enhance the readability and understandability of the dataset, column names are modified using a mapping dictionary. This step provides more descriptive and intuitive names to the columns, making the data more interpretable.
In the TripAdvisor dataset, the 'cook_type' column contains certain values that need to be updated. These values are replaced with a '?' to indicate that the information is unknown or not provided. Simultaneously, the corresponding 'expensive' values are adjusted based on the updated 'cook_type' values. This process ensures consistency in representing the cooking type and pricing information.
The cleaning process also involves extracting numeric values from string columns in the TripAdvisor dataset. For example, the 'rating_trip' column contains ratings represented as strings, such as '4.5 out of 5' or '4.0'. Using regular expressions, the numeric rating values are extracted, allowing for easier numerical analysis.
Similarly, the total number of reviews for each restaurant in the 'total_reviews_trip' column is extracted from the corresponding string values. This step ensures that the review count is represented as a numerical value, facilitating quantitative analysis.
After extracting the necessary information and performing the required modifications, unnecessary columns are removed, and duplicate rows in the TripAdvisor dataset were dropped. This step eliminates any redundant data, ensuring a clean and concise dataset for further analysis.
Through these data cleaning operations, the code ensures that the TripAdvisor restaurant data is consistent, standardized, and ready for integration and further analysis. Similar operations are done for the Google dataset

Data integration

Link to GitHub repository:
https://github.com/darthgween/data-management-project/tree/definitive/4_integration

The objective is to merge the datasets based on common attributes such as restaurant names and addresses. The integration process involves data preprocessing, name and address formatting, and comparison of restaurant attributes for matching. This phase was really challenging because the names and addresses in the two dataset in several cases were different even if they represented the same entity. We choose to use a conservative approach to preserve data quality and ensure that just the correct matches were done. This fact has heavily affected the number of matches that were around 60%. Furthermore in order to complete the whole list of restaurants data enrichment procedures were done.
The TripAdvisor dataset contains information about various restaurants, including their names and addresses. The Google dataset provides a collection of restaurant data, including names, addresses, and additional attributes like the most important one for our purposes is geometry that represents the geo coordinates of the places. Both datasets have been preprocessed prior to integration.

Name Formatting

To ensure consistency in the name attributes of both datasets, a name formatting function is applied to each DataFrame. The function removes special characters, spaces, and common keywords such as "restaurant," "sushi," "pizzeria," etc. This formatting step standardizes the restaurant names, enabling better matching during the integration process.

Address Formatting

Similar to name formatting, address formatting is performed to enhance the matching accuracy between the datasets. In the TripAdvisor dataset, the

address information is refined by removing postal codes and additional location details, leaving only the necessary address components. The Google Places dataset uses a similar process, where the trailing ", Milano" part of the address is removed. These steps ensure uniformity in the address attributes, facilitating effective integration.

Integration Process

The integration process involves iteratively comparing each restaurant entry from the TripAdvisor dataset with the corresponding entries in the Google Places dataset. The comparison is based on name similarity and address matching. For efficient processing, a ThreadPoolExecutor is employed to parallelize the comparison task.

Comparison and Matching

Within each iteration, the compare_row function is used to compare the name and address attributes between the current TripAdvisor entry and the entire Google Places dataset. The name attributes are compared using the RapidFuzz library, which calculates a similarity score based on token set matching. A score threshold of 80 is set to consider a match between the names. Additionally, exact matches are considered when one name is a subset of the other or vice versa.
If a match is found based on the name attributes, the address attributes are further compared using the same similarity scoring approach. If the address similarity score exceeds 80 or if one address is a subset of the other, the entries are considered a match. In such cases, the TripAdvisor entry and the corresponding Google Places entry are merged into a result dataset.

Handling Multiple Matches

In situations where multiple potential matches are found in the Google Places dataset, the entry with the highest address similarity score is selected as the final match. If the highest score is 100 and there is a unique TripAdvisor entry with a matching name, it is considered a definitive match. Otherwise, a minimum score threshold of 75 is set to consider a match, taking into account the address similarity.

Results and Output

During the integration process, the matched results are stored in a results DataFrame. Additionally, TripAdvisor entries that could not be matched with any Google Places entry are saved in a not_found_trip DataFrame that will be used further for the enrichment.

Exploratory data analysis

Link to GitHub: https://github.com/darthgween/data-management-project/blob/definitive/explorative_data_analysis.ipynb

The exploratory analysis of the downloaded datasets was conducted using the Python programming language. The code for these operations can be found in the exploratory_data_analysis.ipynb file. The following exploratory analysis was performed after integrating and cleaning the datasets. Statistics were computed using the sklearn and scipy libraries. Graphs were constructed using Plotly to provide interactive visualizations, allowing the end user to interact with the charts and view details about individual restaurants.

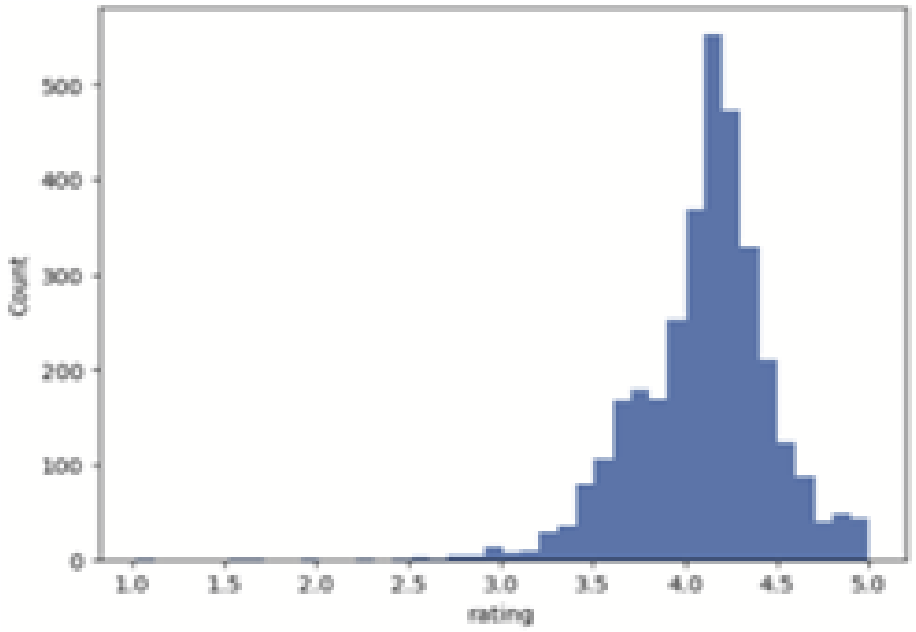
	rating	price_level	total_reviews
count	3341.000000	2996.000000	3600.000000
mean	4.110373	1.927575	662.308333
std	0.366198	0.752969	781.623694
min	1.021739	1.000000	0.000000
25%	3.915714	1.000000	132.000000
50%	4.158038	2.073309	395.000000
75%	4.319588	2.278782	924.250000
max	5.000000	4.000000	10671.000000

1. Rating: The majority of ratings fall within the range of 4 to 5, with a mean rating of approximately 4.11 and a relatively small standard deviation of 0.37. This suggests that the majority of establishments or items in the dataset have high ratings, indicating positive reviews or customer satisfaction.
2. Price Level: The mean price level is approximately 1.93, indicating that most entries have a relatively low price level. The presence of a

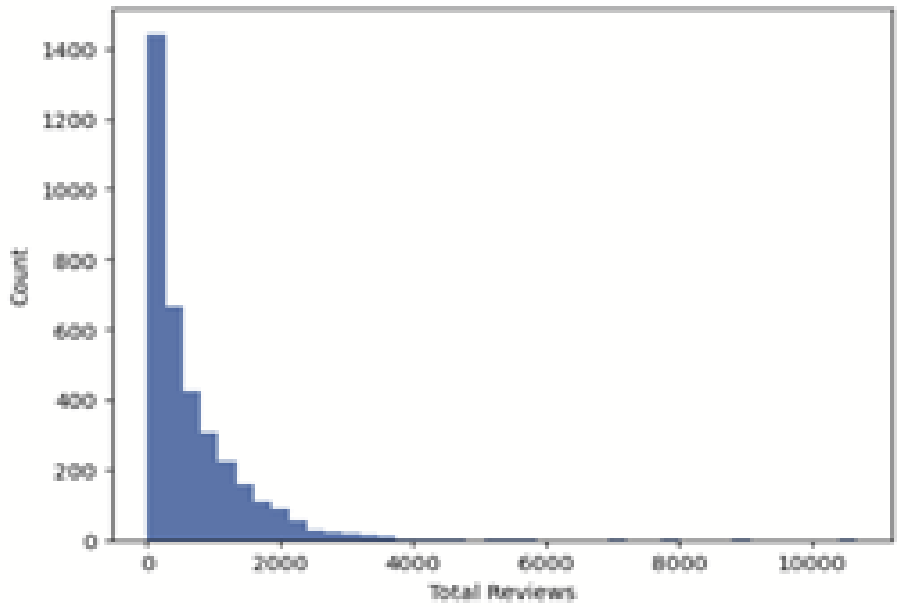
maximum value of 4 suggests that some establishments or items in the dataset may have a higher price range.

3. Total Reviews: The "total_reviews" column has a wide range of values, with a mean of approximately 662.31 and a large standard deviation of 781.62. This indicates that the number of reviews varies greatly across the dataset. The presence of a high maximum value (10671) suggests the existence of some highly popular or widely-reviewed establishments or items.

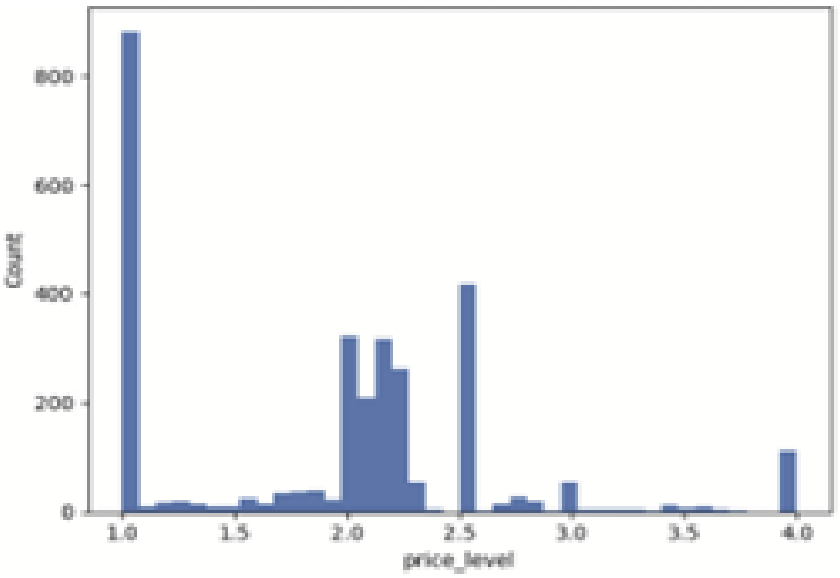
Here some histograms to visualize better the features of the dataset:



Mean: 4.110372966686255
Median: 4.1580381471389645
Skewness: -0.9101084984025848
The distribution is moderately skewed.
Standard deviation: 0.36619787626979666
The data is relatively tightly clustered around the mean.
There are 0 outliers in the rating distribution.



Mean: 662.3083333333333
Median: 395.0
Skewness: 2.980879227234029
The distribution is highly skewed.
Standard deviation: 781.6236941544724
The data is relatively spread out from the mean.
There are 181 outliers in this distribution.

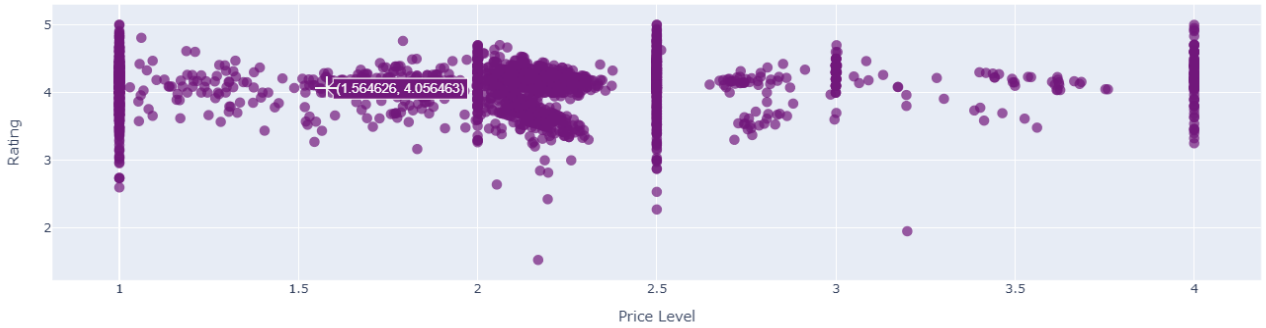


Mean: 1.9275752477204797
Median: 2.073309021568499
Skewness: 0.5095699155678228
The distribution is moderately skewed.

Standard deviation: 0.7529694684791002
There are 0 outliers in the rating distribution.

Corrplots between the variables

In this phase, we will examine the correlations and visualize a correlation plot (corrplot) to explore the relationships between the attributes in the dataset. Our focus begins with the first pair of attributes: Rating and Price level. Surprisingly, there doesn't appear to be a strong relationship between these two attributes, contrary to what one might expect.



The correlation coefficient between "price_level" and "rating" is approximately 0.0466, indicating a weak correlation between these variables. This means that changes in the price level are not strongly associated with changes in the rating. Other factors or variables may have a more significant impact on the ratings of the establishments or items in the dataset.

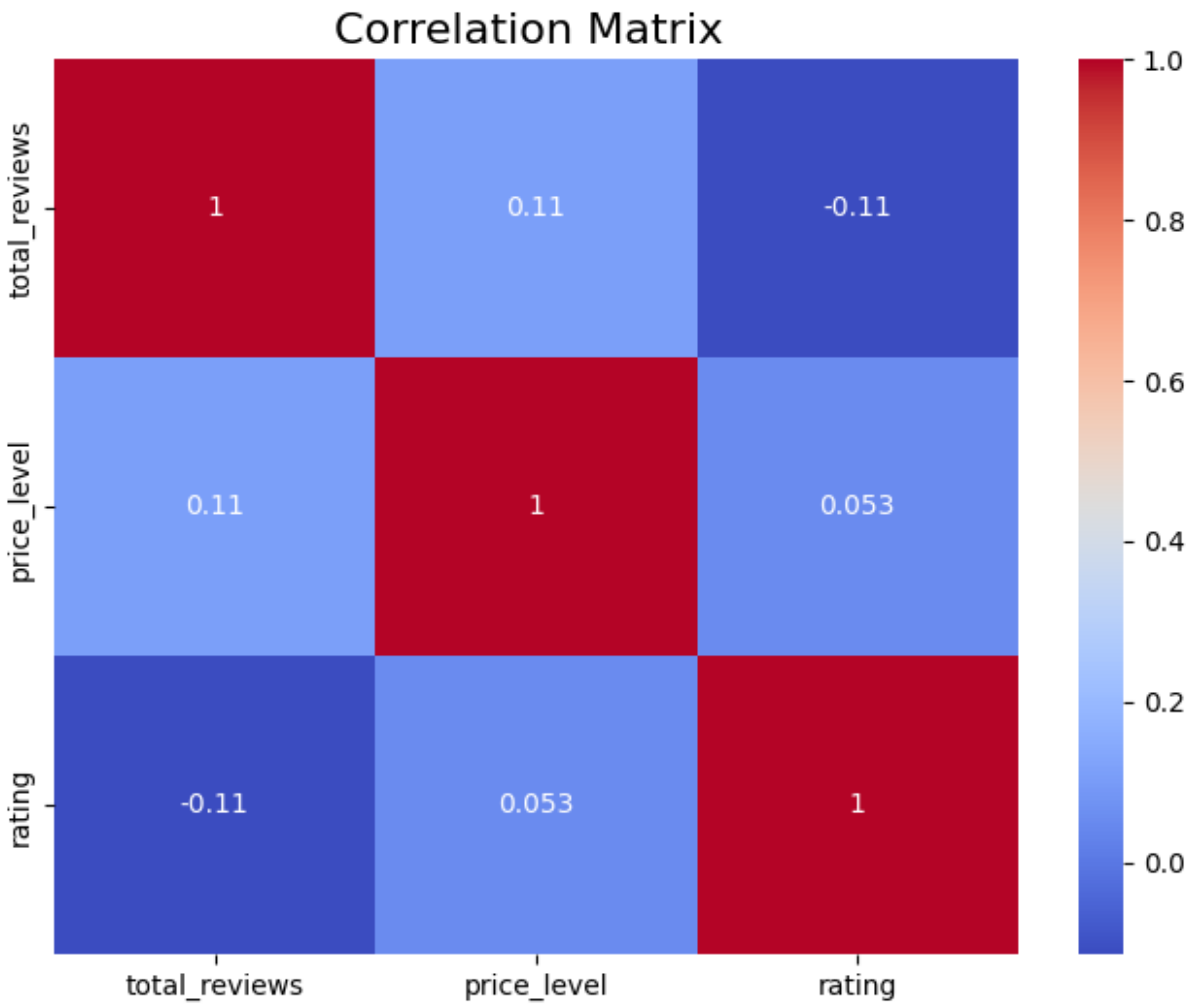
Furthermore, a correlation plot (corrplot) is displayed between the other two variables: rating and total reviews.

In the context of data analysis, an interesting relationship emerged between the rating and the total number of reviews. It was observed that as the number of reviews increases, the rating tends to concentrate within a narrower range between 3.50 and 4.50. This phenomenon suggests that restaurants with a higher number of reviews tend to have more similar ratings, creating a convergence area within the rating range.

Additionally, it was noticed that restaurants with a limited number of reviews exhibit extremely high or extremely low rating values. This indicates that the popularity of a restaurant may play a significant role in determining the rating. It is possible that restaurants with few reviews are subject to greater variability in ratings, as a limited number of opinions can have a more significant impact on their overall scores.

From a statistical perspective, a Pearson correlation coefficient was computed between the rating and the total number of reviews. A weak correlation was observed, indicated by a value close to zero. This suggests that the number of reviews alone does not fully explain the variations in the rating, and other factors may have a more significant influence on restaurant ratings.

Next, we can visualize the correlation matrix to gain further insights into the relationships between the variables.

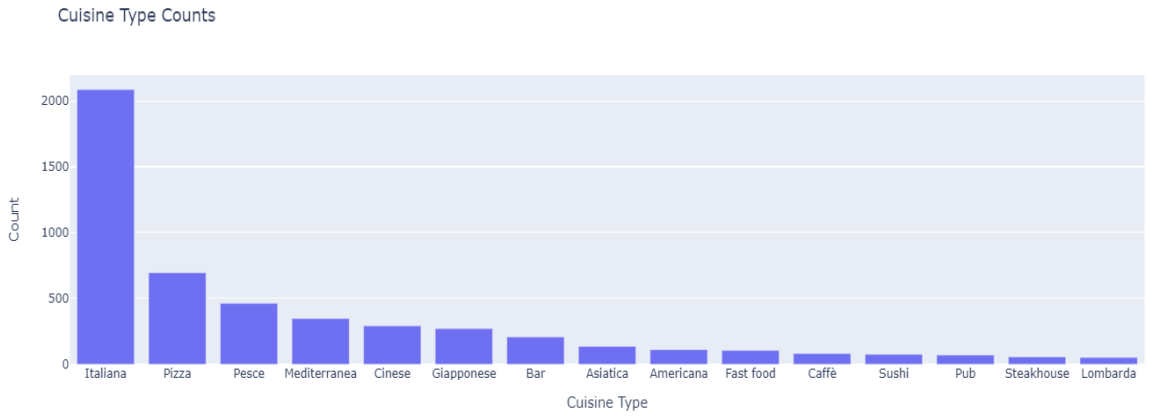


Histograms

Subsequently, an interactive histogram was created to visualize the frequency of different restaurant types in the dataset. It is important to note that many restaurants offer more than one type of cuisine, such as sushi and Asian cuisine. Therefore, the cuisine types were split to calculate their individual frequencies.

From the histogram, it is evident that the most popular cuisines within the dataset are Italian, pizza, seafood, and Mediterranean. These four categories exhibit significantly higher frequencies compared to other cuisines present in the dataset. Conversely, the less represented cuisines, each comprising fewer than 300 restaurants, encompass a variety of other culinary styles.

The interactive histogram provides a clear visual representation of the distribution of cuisine type frequencies in the dataset, highlighting the most common and least represented cuisines.



An analysis of ratings was conducted, considering the different cuisine types present in the dataset. To obtain this evaluation, code was used to extract the unique cuisine types from the DataFrame and calculate the average rating corresponding to each cuisine type. Subsequently, an interactive histogram was created, enabling a clear visualization and comparison of the average ratings for the most common cuisine types. This graphical representation provides an indication of the average ratings for each culinary category, facilitating the analysis of the relationships between cuisine types and ratings. Through the histogram, significant insights can be gained regarding the popularity and average evaluation of the various cuisine types within the dataset.

Data enrichment

Link to GitHub repository:
https://github.com/darthgween/data-management-project/tree/definitive/5_data_enrichment

Data enrichment plays a crucial role in integrating restaurant information from TripAdvisor and Google datasets. In this project, we focused on concatenating the unmatched records from TripAdvisor with the matched records from Google, thereby creating a comprehensive dataset for analysis. To enhance the dataset further, we utilized a Python script leveraging the geopy library. This script accepted the restaurant address as input in text format and returned the corresponding latitude and longitude coordinates. By incorporating this geolocation data into the dataset, we enabled geospatial analysis and enriched the information available for each restaurant. Additionally, we implemented a web scraping algorithm to extract the top 10 reviews for each restaurant listed on TripAdvisor. This additional enrichment step provided valuable user-generated content, which can be used for sentiment analysis and gaining insights into customer experiences. The combination of these enrichment techniques resulted in a more robust and informative dataset, facilitating advanced analysis and a deeper understanding of the restaurant landscape.

Data quality

Link to GitHub repository:
https://github.com/darthgween/data-management-project/blob/definitive/8_data_quality/quality_analysis.ipynb

A very important phase after the integration step is to evaluate the quality of the data obtained following these criteria: unicity, completeness and accuracy

Completeness

The completeness dimension is used to assess whether the data is sufficiently available to make decisions and perform inferences. As a measure of completeness, the percentage of observations without any missing values was evaluated for each dataset.

Column Name	Null Count	Non-Null Count
id	0	6971
name	0	6971
cook_type	936	6035
address	9	6962
latitude	920	6051
longitude	920	6051
rating	770	6201
price_level	1811	5160
total_reviews	0	6971
reviews	2	6969

Since no records have null IDs, we have decided to retain the entirety of the dataset to provide the broadest possible information on all the restaurants found, even at the expense of completeness for individual records in terms of attributes such as price_level, rating, latitude, longitude, and cook_type.
This decision was made because the main objective of our analysis is to have a more comprehensive dataset, despite the absence of certain attributes, in order to provide a list that is as complete as possible.

Accuracy

Accuracy is a parameter that allows us to evaluate whether the collected data accurately represents the reality of the reference data. In our case, considering the nature of the data sources used in the restaurant industry, which are considered almost institutional, on one hand, TripAdvisor, the leading platform for collecting restaurant data, and on the other hand, Google.
By using a systematic approach, we can confidently state that only the records that have been accurately matched between TripAdvisor and Google represent 100% accurate data. However, the data that could not be found in TripAdvisor and was integrated from other sources cannot be considered similarly accurate.
Therefore, we can calculate accuracy using the formula: matched records / (matched records + unmatched records), resulting in

approximately 60%. It is possible to explore future developments to increase this percentage and improve data quality.
It is important to note that the accuracy of the data depends on the quality of the data sources used and the integration process. In our case, accuracy is influenced by the conservative matching approach between data from TripAdvisor and Google, which was implemented to ensure data quality. This may have reduced the number of matches but increased the precision of the matches made.
It is always good practice to carefully evaluate data quality and identify areas for improvement to ensure the reliability of the analyses and conclusions drawn from the collected data.

Unicity

The primary objective is to ensure the representation of each unique real-world object or event within a singular dataset.

To accomplish this goal, duplicate instances were systematically removed from each dataset. These instances were identified as exact replicas of other instances.

In order to quantitatively measure the level of unicity achieved, we utilized a metric that calculates the complement to 1 of the percentage of duplicates to the total number of instances in the CSV file. This metric effectively represents the proportion of distinct observations in relation to the overall dataset.

The results obtained from this analysis are as follows:

```python
df['combined'] = df['name'] + df['address']

# Find duplicates considering name and address
duplicates = df[df.duplicated(subset='combined', keep=False)]
duplicate_counts = duplicates['combined'].value_counts()


for value, count in duplicate_counts.items():
    print(f'{value} - {count} occurrences')
```

✓ 0.0s Python

```
Circolo Arci BellezzaVia Giovanni Bellezza 16, 20136 Milano Italia - 2 occurrences
Long YuanVia Lorenteggio 43, 20146 Milano Italia - 2 occurrences
Bar Viale CertosaViale Certosa 125, 20151 Milano Italia - 2 occurrences
Hot PotVia Paolo Lomazzo 17, 20154 Milano Italia - 2 occurrences
Burgez Corso Como 2Corso Como 2, 20154 Milano Italia - 2 occurrences
Poke House - FiammaVia Galvano Fiamma, 5, 20129 Milano Italia - 2 occurrences
Il Cappellaio MattoVia, Ripa Di Porta Ticinese, 51, 20143 Milano Mi, 20143 Milano Italia - 2 occurrences
Ham Holy Burger - Milano, viale BlignyViale Bligny 39, 20136 Milano Italia - 2 occurrences
CHICCO BistrotAngolo Piazza Meda  Via Adalberto Catena 2 Angolo piazza meda, 20121 Milano Italia - 2 occurrences
Caffe Ambrosiano Bar tavola freddaVia San Vittore, di fronte Ospedale San Giuseppe, Milano Italia - 2 occurrences
Isola Verde La PizzaVia Mario Morgantini Angolo, 20148 Milano Italia - 2 occurrences
I Due Poeti CaffétteriaPiazza Luigi Vittorio Bertarelli 4, 20122 Milano Italia - 2 occurrences
BarlafusLargo Vulci, 7, 20159 Milano Italia - 2 occurrences
Great American GrillVia Lucio Giunio Columella, 36, 20128 Milano Italia - 2 occurrences
```

Numero di istanze presenti prima	duplicati rimossi	istanze rimanenti	Metrica di unicità
6972	14	6958	99.8%

Sentiment analysis

Comparing sentiment scores: Evaluating restaurant reviews based on ratings

Sentiment analysis was conducted using TextBlob to evaluate restaurant reviews and determine if they align with the assigned ratings for each restaurant. The dataset was divided into two subsets: one containing restaurants with a rating below 3 and the other containing restaurants with a rating above 4.5.

For each review in each subset, a TextBlob object was created, and the sentiment score (polarity score) was calculated. Subsequently, the score was divided by the number of reviews in the specific subset,

allowing for the calculation of the average sentiment for the available reviews for each restaurant.

The results of the sentiment analysis display the average sentiment score for each subset. The average sentiment score for the subset of low-rated restaurants was calculated as 0.028, whereas for the subset of high-rated restaurants, it was calculated as 0.208. These scores suggest that reviews for high-rated restaurants are more positive compared to those for low-rated restaurants.

TextBlob utilizes a machine learning algorithm to classify words as positive, negative, or neutral based on a pre-labeled dataset. The sentiment analysis using TextBlob has highlighted the difference in sentiment scores between the subsets of low-rated and high-rated restaurants, indicating that reviews for high-rated restaurants tend to be more positive than those for low-rated restaurants.

Analysis on comment polarization

Now, sentiment analysis is performed using the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis tool. During the analysis, four sentiment scores were calculated using the VADER sentiment analysis tool: "neg" (negative score), "neu" (neutral score), "pos" (positive score), and "compound" (overall score).

During the analysis, the following sentiment scores were calculated using the VADER sentiment analysis tool:

- "neg": Represents the intensity of negative sentiment in the text, ranging from 0 to 1. A higher value indicates a stronger negative sentiment.
- "neu": Represents the intensity of neutral sentiment, ranging from 0 to 1. It indicates how much of the text is considered neutral.
- "pos": Represents the intensity of positive sentiment, ranging from 0 to 1. A higher value indicates a stronger positive sentiment.
- "compound": Represents the overall sentiment of the text, ranging from -1 to 1. It combines the scores of negative, neutral, and positive sentiment. A value above 0 generally indicates a

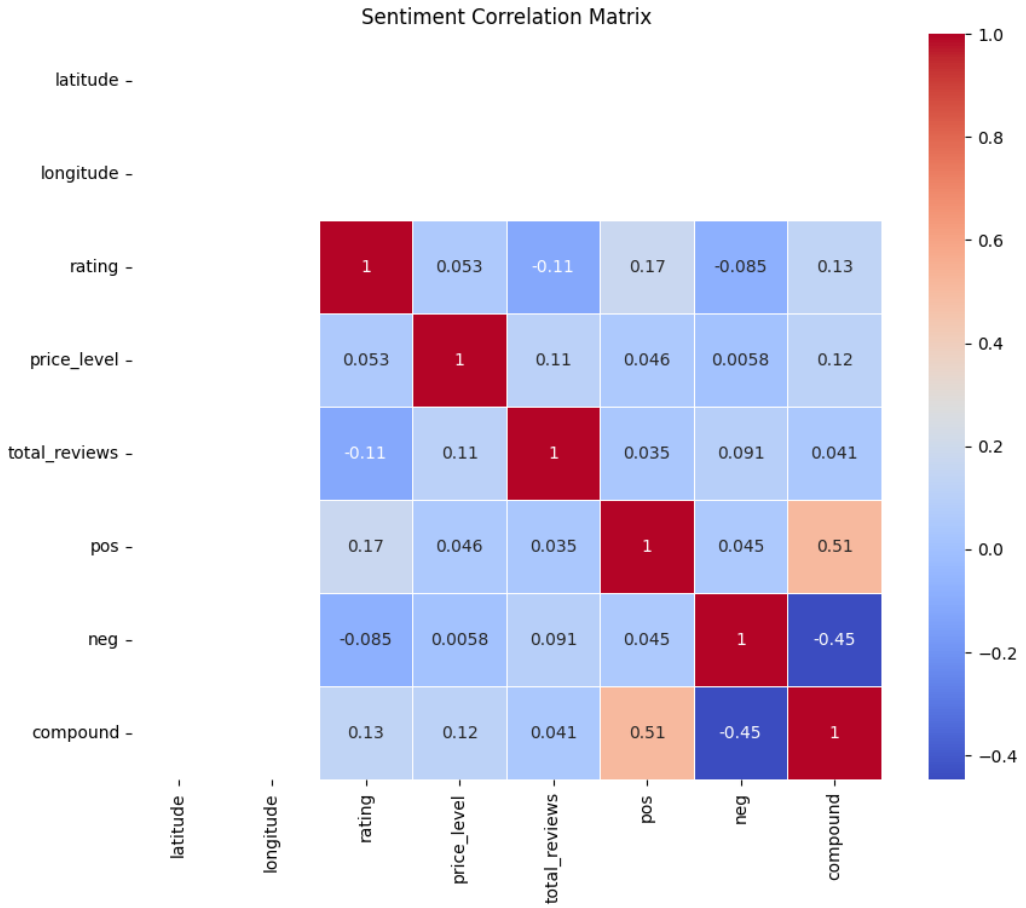
positive sentiment, while a value below 0 indicates a negative sentiment.

These sentiment scores provide insights into the expressed sentiment in the text, enabling a deeper understanding of the reviews in the dataset.

name	cook_type	types	address	latitude	longitude	rating	price_level	total_reviews	reviews	cuisine_type_1	cuisine_type_2	neg	neu	pos	compound
Il Panino	Italiana, Cibo di strada	['restaurant', 'food', 'point_of_interest', 'e...	Via Laghetto 7, 20122 Milano Italia	45.478721	9.185232	4.835329	1.000000	501.0	[{'bubbles': 'bubble_50', 'title': 'Da brevett...	Italiana	Cibo di strada	0.000	0.983	0.017	0.8160
Cantiere Hambirreria	Italiana, Americana	['restaurant', 'food', 'bar', 'point_of_intere...	Corso Garibaldi, 111, 20121 Milano Italia	45.478721	9.185232	4.675467	2.000000	1553.0	[{'bubbles': 'bubble_50', 'title': 'SUPER!', '...	Italiana	Americana	0.000	0.949	0.051	0.9056
Shabby Grill Restaurant	Italiana, Steakhouse	['restaurant', 'point_of_interest', 'food', 'e...	Via Novara, Piazza Carlo Amati, 31, 20147 Mila...	45.478721	9.185232	4.862547	2.500000	801.0	[{'bubbles': 'bubble_50', 'title': 'Ottima car...	Italiana	Steakhouse	0.006	0.980	0.014	0.4885

The result of this analysis was a final dataset with sentiment scores attached to each restaurant. Subsequently, to assess any correlations, a correlation matrix was examined between all variables in the dataset and those representing the sentiment polarization of the reviews.

Correlation matrix



- Rating and Total Reviews:
  - Negative correlation of approximately -0.11
  - Restaurants with a higher number of reviews tend to have slightly lower ratings
  - This may be because popular restaurants receive more reviews, increasing the likelihood of both positive and negative feedback, which can impact the overall rating.
- Rating and Price Level:
  - Relatively low correlation of approximately 0.053
  - Indicates little to no strong linear relationship between the variables
  - The price level of a restaurant does not significantly impact its rating

- Sentiment scores:
    - Compound sentiment score has a moderate positive correlation (0.51) with positive sentiment score (pos)
    - Higher compound scores are associated with higher positive sentiment scores
    - Compound score has a moderate negative correlation (-0.45) with negative sentiment score (neg)
    - Higher compound scores are associated with lower negative sentiment scores
    - Restaurants with more positive sentiments tend to have higher compound sentiment scores and fewer negative sentiments
- Positive Sentiment Score (pos) and Price Level:
    - Weak positive correlation of approximately 0.046
    - There is a slight association between positive sentiment in reviews and the price level of restaurants
    - Other factors, such as food quality, service, and ambiance, may have a more significant impact on customers' perceptions of the price level.
- Negative Sentiment Score (neg) and Rating:
    - Weak negative correlation of approximately -0.085
    - As the negative sentiment score increases, the rating tends to decrease slightly
    - Other factors, such as positive sentiment, overall experience, or specific aspects of the restaurants, may have a more substantial influence on the rating.

Detailed individual restaurants

We now proceed with the extraction of the best and worst restaurant based on the "compound" indicator, analyzing the reviews and verifying if the result of our analysis is reliable.

To identify the best and worst restaurant, we consider the "compound" sentiment score associated with each restaurant's reviews. The "compound" score represents the overall sentiment expressed in the reviews, combining the positive, negative, and neutral scores.

By sorting the restaurants based on their "compound" score in descending order, we can determine the restaurant with the highest score as the best restaurant, indicating a generally positive sentiment

in its reviews. Conversely, the restaurant with the lowest "compound" score is considered the worst, indicating a generally negative sentiment.

```
Best Restaurant:
place_id                                ChIJAQAAAM3GhkcRpufjyzm5fZE
name                                                    Alla Cadrega
cook_type                                        Italiana, Lombarda
types           ['restaurant', 'food', 'point_of_interest', 'e...
address               Via Vincenzo Viviani 2, 20124 Milano Italia
latitude                                                  45.478721
longitude                                                  9.185232
rating                                                          4.0
price_level                                               2.221863
total_reviews                                               2008.0
reviews         [{'bubbles': 'bubble_10', 'title': 'Pessima es...
pos                                                         0.072
neg                                                           0.0
compound                                                   0.9941
Name: 693, dtype: object

Worst Restaurant:
place_id                                ChIJARR0_qvHhkcR424l1Iato9c
name                                               Spiller Milano Duomo
cook_type                                    Italiana, Birreria
types           ['bar', 'restaurant', 'food', 'point_of_intere...
address                        Via Larga N. 8, 20122 Milano Italia
latitude                                                  45.478721
longitude                                                  9.185232
rating                                                     4.221687
price_level                                                     2.0
total_reviews                                               581.0
reviews         [{'bubbles': 'bubble_50', 'title': 'Scoperta',...
pos                                                         0.008
neg                                                         0.033
compound                                                  -0.9751
```

Reviews of the best restaurants

For the restaurant "Alla Cadrega," the reviews are indeed positive.

{'bubbles': 'bubble_50', 'title': 'Voglia di cose tipiche', 'description': 'Se cercate piatti tipici “ leggermente abbondanti “ senza spendere un esagerazione anche nella scelta del vino , con servizio attento e

premuroso , i piatti di questa cucina ci hanno fatto passare una bella serata in un ambiente pulito e confortevole . Nostra scelta...fiori di zucca ben presentati ottimi , ossobuco con risotto gnam gnam, mondeghili una marcia in più . Tutti soddisfatti . Grazie e felice anno nuovo .Più'}

, {'bubbles': 'bubble_50', 'title': 'Una vera scoperta (A wonderful discovery)', 'description': "Ci siamo finiti in questo posto per caso ma siamo rimasti molto soddisfatti dell'intera esperienza: dall'accoglienza riservata ai tre (bellissimi) piccoli cagnolini dei nostri amici, all'osso buco con risotto alla milanese che è stato il migliore che abbia mai mangiato - era assolutamente delizioso! E...mi sono goduto ogni morso. La carne era molto tenera e gustosa e il risotto aveva una deliziosa nota di limone. Anche il servizio è stato perfetto e molto cordiale. Tutto sommato è stato una cena meraviglioso e ci torneremo sicuramente. \n\nWe happened upon this place by chance, and were very happy with the whole experience: from the welcome we had regarding our friends' three (gorgeous little) dogs, to the osso buco with risotto Milanese which was the best I've ever had - it was absolutely delicious! And I thoroughly enjoyed it. The meat was very tender and tasty and the risotto had a delicious lemon hint to it. The service was perfect and very friendly, too. All in all a wonderful meal, and we'll most definitely go back.Più"},

{'bubbles': 'bubble_40', 'title': 'Pausa pranzo per niente male', 'description': 'Locale molto tradizionale che propone classici piatti milanesi oppure pizza. Proposte tutte promosse e soprattutto in porzioni abbondanti, a prezzi nella media della zona. Servizio davvero davvero veloce. Consigliato!'},

{'bubbles': 'bubble_50', 'title': 'Ottimi piatti tipici milanesi', 'description': "Ambiente caldo ed accogliente, servizio veloce, personale gentile, disponibile e cortese. Ottimo sia l'ossobuco in gremolada di verdure con risotto alla milanese che la cotoletta."},

{'bubbles': 'bubble_40', 'title': 'Buono ma non economico', 'description': 'Locale facilmente raggiungibile e pulito.\nBuoni i piatti proposti, geniìtile il personale.\nI dolci fatti da loro sono molto saporiti.\nLievemente costoso ma probabilmente a causa della zona.'},

{'bubbles': 'bubble_40', 'title': 'Cucina Tradizionale e gentiezza', 'description': "Un locale classico, con molti posto a sedere in cui si respira un'aria di casa, di cura del cliente, un po' agè...che è il suo bello.

Cucina tradizionale milanese fatta bene, eseguita con cura, senza fronzoli, con sistanza. Risotto alla.milanese con ossobuco monumentale, piatto unico...abbondantissimo. Mondeghili classici (senza pan grattato), primi piatti ottimi. Qualche piccola rivisitazione (fiori di zucca ripieni di ricotta) ben eseguita ma la tradizione impera. Ottima anche la pizza, fatta alla napoletana, con il bordo un po' più alto. Bravissimo il capo cameriere che fa di tutto (prende ordini, prepara le pizze, consiglia visita...). Prezzi nella media senza esagerare, soprattutto perché i piatti sono mediamente abbondanti.Più"},

{'bubbles': 'bubble_50', 'title': 'Gran visir de tuc i terun', 'description': 'Ottimo luogo dove evitare i cari ragionieri... ottimo cibo, in particolare le cadreghe sono eccezionali e vi consiglio di provarle a vostro rischio e pericolo'},

{'bubbles': 'bubble_50', 'title': 'MILANO', 'description': 'Ci siamo fermati al ritorno da una vacanza, prima di rientrare a casa. Abbiamo cenato tranquilli e gustato ottimi piatti, buonissimi i fiori du zucca con la ricotta'},

{'bubbles': 'bubble_50', 'title': 'Eccellente', 'description': 'Locale carino e pulito. Piatti squisiti ed abbondanti. Servizio e rapporto qualità/prezzo ottimi. Estrema cortesia. Consigliato.'}]

Reviews of the worst restaurants

For the restaurant "Spiller Milano Duomo" the reviews are indeed negative.

{'bubbles': 'bubble_30', 'title': 'La vita è come una birra: non sai mai quale scegliere!', 'description': "Locale situato a pochi passi dal Duomo di Milano, l'interno è arredato in maniera elegante ed è molto spazioso, ci sono almeno 2 piani, con servizi igienici disponibili in buone condizioni. Il menù propone antipasti, fritti, primi, secondi, dolci e birre. Ho preso 2 birre,...abbastanza normali, i tempi di attesa sono stati nella norma, il personale è amichevole e gentile, il rapporto qualità prezzo non è proprio al top ma comunque mi sono trovato bene. Per il pagamento accettano anche con carta di credito.Più"},

{'bubbles': 'bubble_50', 'title': 'Ottima serata', 'description': 'Siamo stati accolti con gentilezza e fatti accomodare sotto nella sala che viene in

pratica definita "abbandonata da Dio e dagli uomini".\nNoi eravamo attorniati da camerieri solerti che cercavano di accontentarci in ogni modo. \nOttimo il piatto con maiale e patate, ottima la cotoletta...e i fritti, soprattutto le patate dolci fritte. \nBuonissima la zuppa d\'orzo.\nL\'hamburger forse un po\' asciutto e molto gustose le ali piccanti. \nNon siamo riusciti a mangiare i dolci da quanto eravamo sazi. \nUnica pecca le bottigliette piccole di Coca Cola che a me personalmente irritano molto.Più'},

{'bubbles': 'bubble_10', 'title': 'Insalata con “Sorpresa”!!!', 'description': 'Tempi di attesa lunghissimi, nessuno ti considera. Ordiniamo una insalata mediterrane (quindi con tonno). Ma non solo esso ma cava ma trovo un grande pezzo di plastica rigido ed appuntito(pericolosissimo!!!!) lo faccio notare e neanche una scusa, ritornano dopo un bel po’ con una insalata...nuova, ma dopo aver affermato che il tonno probabilmente era terminato.\nNeanche una scusa, solo superficialità molto palese. VergognaPiù'},

{'bubbles': 'bubble_10', 'title': 'Capodanno 2022/23', 'description': 'Cibo abbastanza buono. \nLa tagliata però sapeva di hamburger credo che la griglia non sia stata pulita bene. Ci hanno messo nel piano interrato carino ma isolati con poche persone mentre al piano di sopra si respirava aria di festa , sotto la morte! Allo...scoccare della mezzanotte nessun Brindisi e nessun countdown, TRISTEZZA ASSOLUTA!!! Pensavo che almeno facessero qualcosa !!!!!! Si è stato fatto con gli ospiti del piano superiore dimenticandosi di quelli di sotto (che pagavano il cenone !!!!! E non alla carte) ECC\nPessimo, triste, tristezza infinita delusione all’ennesima potenza\nAlla cassa alla domanda “tutto bene “ ho esternato il disappunto. Mi dicono che se volevano la bottiglia era a disposizione (a pagamento) li in fondo (😱😱😱😱😡😡😡) aggiungendo “la tv giù non c’era ho messo lì il countdown (😡😡) no dico ma serio ????????\nRibadisco che sotto eravamo abbandonati e che il team di camerieri erano tutti li a festeggiare con bicchiere in mano di spumante...per cui sotto c’era il deserto ed eravamo solo in 5 ospiti abbandonati a se stessi \nDELUDERE E TRISTEPiù'},

{'bubbles': 'bubble_50', 'title': 'Molto carino e buona cucina', 'description': 'Locale molto carino, pulito e accogliente. Il personale giovane e gentile. Abbiamo preso dei club sandwich molto buono e piatti molto abbondanti. Ci ritornerò'},

{'bubbles': 'bubble_10', 'title': 'Pessimo e dir poco!', 'description': "Arrivati per pranzare speravamo in tutto, tranne che in quello che

abbiamo ricevuto: una volta che ci hanno indicato dove ci dovevamo sedere il personale è sparito! Ci siamo dovuti alzare dopo un bel po di tempo a richiedere i menù. Stesso tempo di attesa...per la presa delle comande. Panino gourmet al salmone, pane completamente freddo quasi congelato, alcuni ingredienti mancati (oltre a quelli richiesti da noi) e patatine arrivare fredde. Stinco burger consiglio di cambiare il nome in crauti burger in quanto dei 150g di carne,non ne ho visto nemmeno l'ombra era solo ripieno di crauti con relativa assenza di salsa. Unica nota leggermente positiva se non fosse per il fattore di elevata presenza di olio sono i nachos.\nTempi di attesa biblici per niente di buono.\nSCONSIGLIATO!Più"},

{'bubbles': 'bubble_20', 'title': 'Prima e ultima ', 'description': "Scarso/2 stelle giusto per essere clementi, ma pessima esperienza: servizio confuso, scoordinato e inesperto, menù con metà delle proposte non disponibili (a detta del cameriere), qualità della cucina al limite dell'imbarazzante, bretzel decongelato, tagliata di pollo da supermercato e neanche cotto e servito decentemente, idem...i primi ordinati dagli altri del tavolo. Prezzi anche nella media ma nient'altro da salvare. Prima (e penso) ultima volta in uno spiller.Più"},

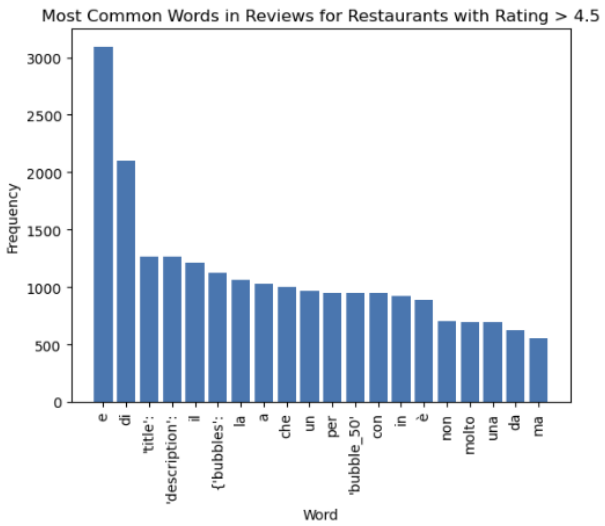
{'bubbles': 'bubble_10', 'title': 'Posto da evitare a pranzo!!', 'description': 'Dopo 50 minuti di attesa per un panino siamo dovuti andare via anche perché dalla cucina non hanno saputo darci nessuna indicazione sui tempi di attesa … bocciato!!'},

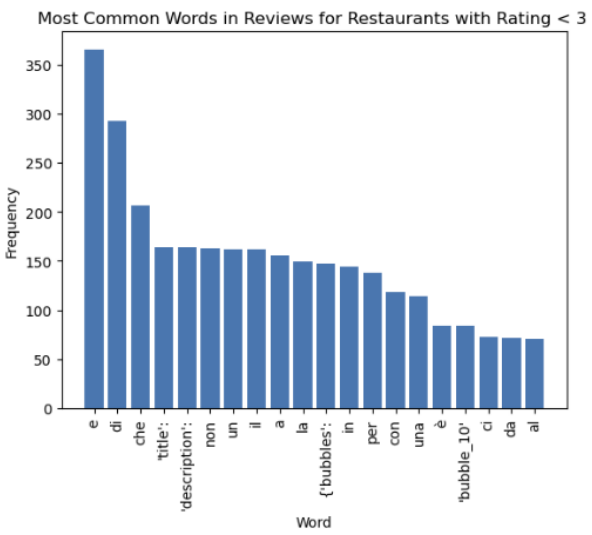
{'bubbles': 'bubble_20', 'title': 'In imbarazzo per loro...', 'description': "Ci rechiamo in questo locale situato vicino al Teatro Gaber (dove avevamo biglietti per uno spettacolo alle 20:45), pensando di riuscire a mangiare e bere qualcosa di veloce in relax. Erano le 19:40. Ordiniamo 1 Bretzel e 1 Pizza.\nIl Bretzel arriva alle 20:25. La...pizza alle 20:35...il tutto dopo che abbiamo più volte sollecitato i camerieri. Cibo di scarsa qualità e ovviamente mangiato di fretta, dopo quasi un'ora di attesa, per non perdere l'inizio dello spettacolo.\nAggiungo che i tavoli vicini erano nella stessa situazione...anzi, una coppia a fianco era arrivata ben prima di noi e ha certamente ricevuto la propria ordinazione dopo che noi ci eravamo già alzati.\nNon so cosa aggiungere. Mi sembra che nel locale manchi un responsabile, un coordinatore...i camerieri sinceramente mi sembrano un pò persi, senza una guida.\nNon ci tornerò.Più"}]

With respect to the work performed, the code calculated sentiment scores using VADER and provided information on the correlation between sentiment scores, rating, total number of reviews and price tier of restaurants. The results indicate that restaurants with a negative compound score actually had negative reviews, while those with a positive compound score had positive reviews. This suggests that the VADER sentiment analyzer has done a good job of assessing the sentiment of restaurant reviews based on their polarities.
It is important to note that sentiment analysis is based on a lexicon-based approach and may not fully capture the full context or nuances of reviews. However, the results obtained so far indicate that sentiment analysis using VADER has produced results consistent with expectations.

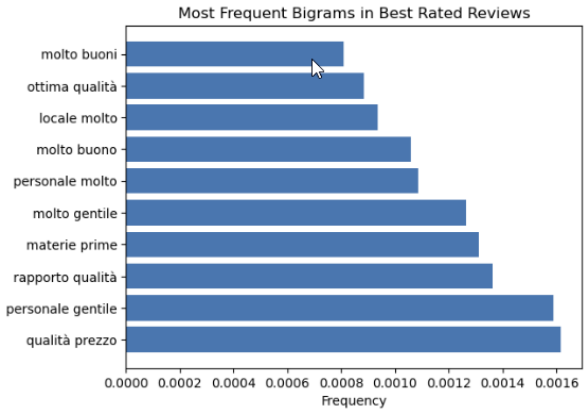
Most frequent words in reviews

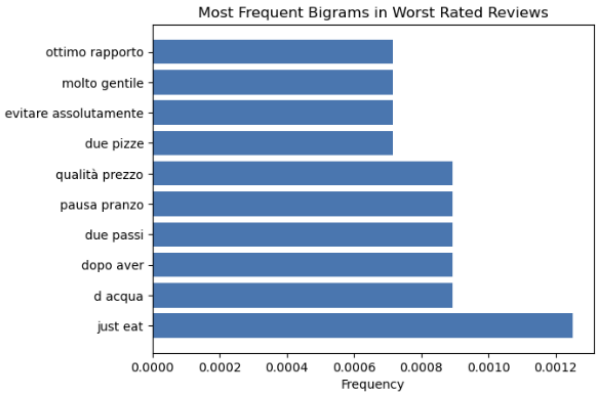
To see the most frequent words in the reviews, the dataset was divided into 2 parts, one part with restaurants with a rating higher than 4.5 and one part with restaurants with a rating lower than 3.





We realize that the most common words are all signs and articles, so a deeper analysis is needed. The process is repeated but looking at the most common bigrams, it is also decided to eliminate the stepwords to avoid outputs similar to the one previously obtained.





In conclusion, the most common words in restaurant reviews with a rating >4.5 are:

molto buoni, ottima qualità, locale molto, molto buono, personale molto, molto gentile, materie prime, rapporto qualità, personale gentile, qualità prezzo.

The most common words in those with a rating < 3:

ottimo rapporto, molto gentile, evitare assolutamente, due pizze , qualità prezzo, pausa pranzo, due passi, dopo aver, d'acqua , just eat

These words indicate the most common trends in restaurant reviews with different ratings. In positive reviews (rating >4.5), people tend to mention the quality of the food, the venue, the staff and the positive value for money. In negative reviews (rating <3), people often mention issues such as bad value for money, disappointing service or against the experience altogether, although there is no shortage of positive reviews even in restaurants with a low rating

Data modeling

We opted to use MongoDB for data modeling and performed several queries to gain insights otherwise not discernible. Specifically, MongoDB was a perfect fit for our needs as it also supports geospatial data. We chose to utilize MongoDB Atlas, the cloud database service provided by MongoDB, which greatly simplifies the management and configuration of MongoDB databases in the cloud. Atlas offers a range

of features, including automatic scalability, data replication management, backup management, security, and more. We were able to easily create and configure our database clusters, manage user access, and monitor database performance and through the Atlas interface. To complete our project, we decided to combine all the available data to create a simple recommendation system that suggests similar restaurants based on selected parameters. We took into account all the available data. To accomplish this, we used Neo4j, a graph-based non-relational database. After loading the restaurants as nodes in the database, we used a query to connect them based on the desired parameters, using arbitrarily chosen predefined thresholds.

Regarding the dataset, we selected a set of attributes that we deemed important for our analyses. Here is a detailed description of each attribute we chose to retain:

1. place_id: This attribute represents a unique identifier for each restaurant in the dataset. It was acquired from the Google dataset, which provides additional information about the restaurant. By using this identifier, we can link the restaurant information to other related data sources.
2. name: This attribute represents the name of the restaurant, which was extracted from the TripAdvisor dataset. The restaurant name is an important element for uniquely identifying it and making it recognizable to users.
3. cook_type: This attribute indicates the types of cuisine offered by the restaurant. There can be one or two cuisine types associated with each restaurant in the dataset. This information can be useful for classifying and filtering restaurants based on users' culinary preferences.
4. address: The address attribute represents the restaurant's address, which was extracted from the TripAdvisor dataset. We preferred using this information because the TripAdvisor dataset provides more complete and clean addresses compared to those in the Google dataset.
5. latitude: This attribute represents the geographical latitude of the restaurant. For most restaurants, we obtained this information from the Google dataset. However, for the remaining restaurants in the "not_found_trip" file, we used the geopy library to obtain the latitude using the address as input.

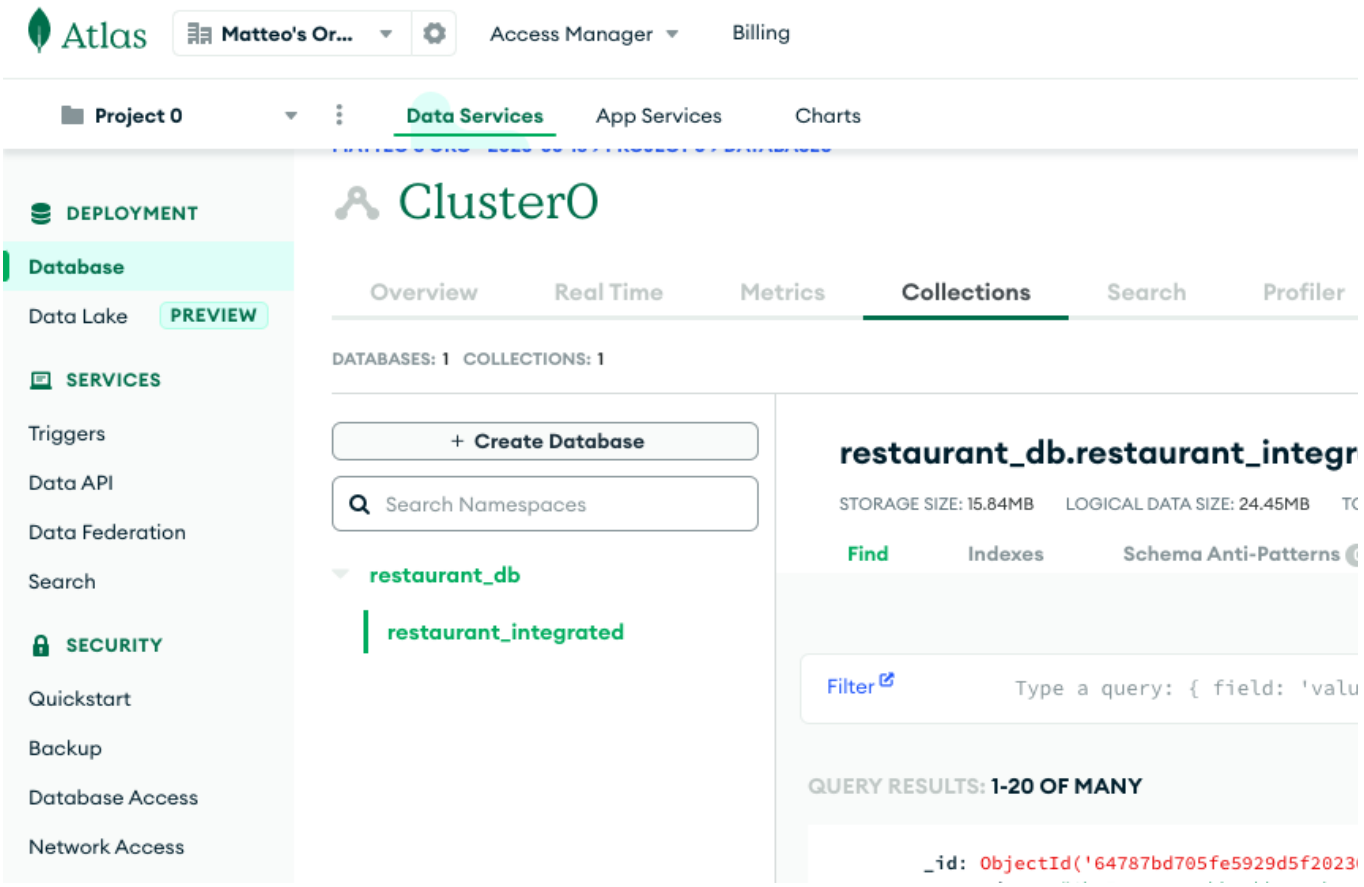
6. longitude: This attribute represents the geographical longitude of the restaurant. Similarly, for most restaurants, we obtained this information from the Google dataset. For the remaining restaurants in the "not_found_trip" file, we used the geopy library to obtain the longitude using the address as input.
7. rating: This attribute represents a weighted average rating of the restaurant, calculated by combining ratings from both the Google and TripAdvisor datasets. The rating provides a general indication of the overall quality of the restaurant based on user feedback.
8. price_level: This attribute represents the price level of the restaurant, also calculated as a weighted average. The price level can provide an indication of the average price range of the dishes offered by the restaurant.
9. total_reviews: This attribute represents the total number of reviews received by the restaurant, summing up reviews from both the Google and TripAdvisor datasets. The number of reviews can provide an indication of the restaurant's popularity and the amount of feedback available.
10. reviews: This attribute contains the top 10 TripAdvisor reviews for each restaurant. These reviews were obtained through web scraping, allowing us to include a representative sample of user opinions about the restaurant. These selected attributes offer a range of relevant information for analyzing and evaluating the restaurants in the dataset, enabling us to gain a more comprehensive and detailed understanding of their characteristics and user experiences.

MongoDB Storage

https://github.com/darthgween/data-management-project/tree/definitive/7_store

MongoDB proved to be a versatile and scalable solution for storing and querying restaurant information. In our research, we harnessed the power of MongoDB to identify and analyze the top 10 pizza restaurants in Milan while considering their proximity to the city center. Our investigation revealed a noticeable trend: restaurants located closer to

the center tend to have significantly higher average price levels compared to those situated in the suburbs.
Moreover, leveraging the available fields such as the restaurant's name, cook type, address, latitude, longitude, rating, price level, and total reviews, we formulated an additional query to delve deeper into the restaurant data. One example of such a query involved exploring the correlation between a restaurant's rating and its price level. By executing this query, we were able to uncover valuable insights and gain a more comprehensive understanding of the restaurant landscape beyond just pizzas in Milan. MongoDB's flexibility and scalability played a crucial role in enabling us to conduct these analyses and derive meaningful conclusions from the data.



Queries

We did some interesting analysis using queries and other processing techniques then we plotted the results on the milan map. The 10 best pizzas in milan:

10 best pizza of milan!!!

	_id	Unnamed: 0	name	cook_type	address	latitude	longitude	rating	price_level	total_reviews	reviews
0	64787bd705fe5929d5f21088	396	Il Pizzaleo	Italiana, Pizza	Corso Cristoforo Colombo 3, 20144 Milano Italia	45.454945	9.172369	5.0	1.0	67.0	[{'bubbles': 'bubble_50', 'title': 'Sfiziosa',...
1	64787bd705fe5929d5f2132c	2427	Il Panzerotto	Italiana, Pizza	Via Padova 147, 20127 Milano Italia	45.499560	9.234306	5.0	1.0	17.0	[{'bubbles': 'bubble_50', 'title': 'ottimi!', ...
2	64787bd705fe5929d5f2148d	3194	Panificio Caffetteria Da Angela	Pizza, Caffè	Via Jacopo Dal Verme 14, 20159 Milano Italia	45.487299	9.188359	5.0	NaN	13.0	[{'bubbles': 'bubble_40', 'title': 'focacce bu...
3	64787bd705fe5929d5f21400	2925	Pronto Pizza	Pizza, Italiana	Via Dei Fiordalisi 3, 20146 Milano Italia	45.450894	9.136821	5.0	NaN	11.0	[{'bubbles': 'bubble_50', 'title': 'Sempre pre...
4	64787bd705fe5929d5f2157b	3649	Yoyogurt - La Paia	Italiana, Pizza	corso XXII Marzo 12 Angolo Via Sciesa, 20135 M...	NaN	NaN	5.0	2.5	7.0	[{'bubbles': 'bubble_50', 'title': 'Pausa pran...
5	64787bd705fe5929d5f21572	3630	New Family	Italiana, Pizza	Via Gallarate 122 Milano, Zona Certosa, 20151 ...	45.501425	9.120325	5.0	NaN	5.0	[{'bubbles': 'bubble_50', 'title': 'FANTASTICO...
6	64787bd705fe5929d5f215f9	3900	Bottega Mascadelli	Pizza	Viale Pasubio 6, 20154 Milano Italia	45.481324	9.185769	5.0	2.5	5.0	[{'bubbles': 'bubble_50', 'title': 'Buona al p...
7	64787bd705fe5929d5f215d3	3827	Play Pizza	Pizza	Via Andrea Solari 41, 20144 Milano Italia	45.453137	9.157223	5.0	1.0	5.0	[{'bubbles': 'bubble_50', 'title': 'La adoro',...
8	64787bd705fe5929d5f21641	4034	L'arte della pizza	Pizza	Via Dalmazia, Milano Italia	45.458166	9.243608	5.0	1.0	5.0	[{'bubbles': 'bubble_50', 'title': 'Una meravi...
9	64787bd705fe5929d5f2165d	4080	Il Forno delle Delizie	Italiana, Pizza	Via Pietro Andrea Saccardo 37 Lambrate, 20134 ...	45.481420	9.243032	5.0	NaN	4.0	[{'bubbles': 'bubble_50', 'title': 'Ottimi pro...

```python
#query the first 10 pizza restaurants that has the highest rating and the highest number of reviews

word = "Pizza"

pattern = f"\\b{word}\\b"
query = {
    'cook_type': {
        '$regex': pattern,
        '$options': 'i'
    }
}

sort_criteria = [
    ('rating', -1),
    ('total_reviews', -1)
]

documents = collection.find(query).sort(sort_criteria).limit(10)
data = [doc for doc in documents]
df = pd.DataFrame(data)

print("10 best pizza of milan!!!")
df
```

The best low cost restaurants in Milan

10 best restaurants with the greatest rating over 4 and lowest price level under 2!!!

	_id	Unnamed: 0	name	cook_type	address	latitude	longitude	rating	price_level	total_reviews	reviews
0	64787bd705fe5929d5f21040	10	Kebabbar Star Zagros	Mediorientale, Turca	Corso 22 Marzo,38, 20135 Milano Italia	45.462221	9.223890	5.0	1.0	416.0	[{'bubbles': 'bubble_50', 'title': 'Una piacev...
1	64787bd705fe5929d5f20c22	ChIJQ8ITS37BhkcRE0NcBAYFR3M	Trinacriami - Via Caminadella	Italiana, Fast food	Via Caminadella 18, 20123 Milano Italia	45.460395	9.176737	5.0	1.0	5.0	[{'bubbles': 'bubble_50', 'title': 'Ottimo mar...
2	64787bd705fe5929d5f210f7	843	Piadà	Italiana, Fast food	Via Andrea Solari 72, 20144 Milano Italia	45.453091	9.155656	5.0	1.0	74.0	[{'bubbles': 'bubble_30', 'title': 'Buona', 'd...
3	64787bd705fe5929d5f21062	223	I Sapori Della Pasta	Italiana, Mediterranea	Via Alessandro Volta 15, 20121 Milano Italia	45.480038	9.183191	5.0	1.0	86.0	[{'bubbles': 'bubble_50', 'title': 'Superlativ...
4	64787bd705fe5929d5f20ac6	ChIJof-w2wTFhkcRl0oqtvTU_TA	Crostone.it	Italiana, Fast food	Via Giuseppe Ripamonti, 190, 20141 Milano Italia	45.432850	9.200985	5.0	1.0	17.0	[{'bubbles': 'bubble_50', 'title': 'Ottima sco...
5	64787bd705fe5929d5f21088	396	Il Pizzaleo	Italiana, Pizza	Corso Cristoforo Colombo 3, 20144 Milano Italia	45.454945	9.172369	5.0	1.0	67.0	[{'bubbles': 'bubble_50', 'title': 'Sfiziosa',...
6	64787bd705fe5929d5f2124d	1879	Peter Bar	Italiana	Via Larga 31, 20122 Milano Italia	45.460684	9.191605	5.0	1.0	19.0	[{'bubbles': 'bubble_50', 'title': 'Gentile', ...
7	64787bd705fe5929d5f21184	1347	Fun Food United Nations	Latino americana, Internazionale	Via Edmondo de Amicis 35, 20123 Milano Italia	45.459451	9.176316	5.0	1.0	39.0	[{'bubbles': 'bubble_40', 'title': 'Eh...decis...
8	64787bd705fe5929d5f210e7	779	Sbunda	Italiana, Mediterranea	Piazzale Antonio Baiamonti 1 angolo via Paolo ...	NaN	NaN	5.0	1.0	55.0	[{'bubbles': 'bubble_50', 'title': 'Ottimo. Da...
9	64787bd705fe5929d5f210d7	720	L'Alter Bar	NaN	Via Vincenzo Monti 15, 20123 Milano Italia	45.467276	9.173755	5.0	1.0	48.0	[{'bubbles': 'bubble_40', 'title': 'Per feste ...

```
#best low cost restaurants


query = {
    'rating': {
        '$gt': 4.0

    },
    'price_level': {
        '$lt': 2
      }

}

sort_criteria = [
    ('rating', -1),
    ('price_level', 1)
]

documents = collection.find(query).sort(sort_criteria).limit(10)

# Convert the documents to a list of dictionaries
data = [doc for doc in documents]

# Create a DataFrame from the data
df = pd.DataFrame(data)

print("10 best restaurants with the greatest rating over 4 and lowest price level under 2!!!")
df
```

We also made a study to check if the distance from the central zone of milan influences the average price, more precisely difference between the restaurants in the circle from the duomo with 6km or diameter; this is the result:

```
the average price close to the center is  2.0616978347930384
the average price far to the center is  1.726750453218178
```

This evidences that closer to the center the prices are pretty higher

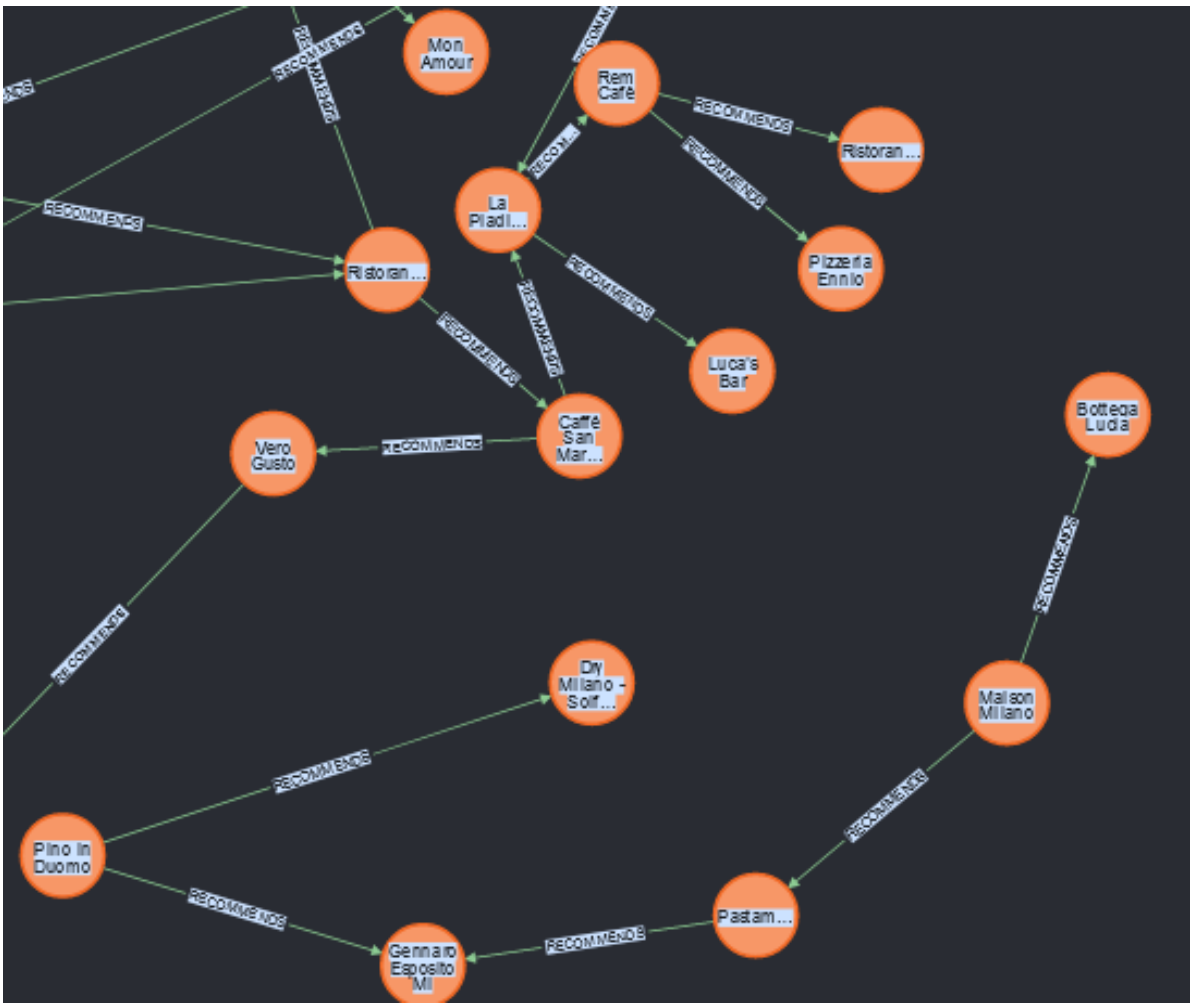
Neo4j storage and recommendation system basis

https://github.com/darthgween/data-management-project/tree/definitive/7_store

Initially, the restaurants were loaded into the database as nodes, creating a connection with the Neo4j database and inserting the data from the DataFrame as nodes representing the restaurants.

Next, a query was run to connect nodes based on certain characteristics. Conditions were applied regarding the difference in rating, reviews, price, type of cuisine and distance (calculated using latitude and longitude coordinates). Only the nodes that met these conditions were connected to each other.

Finally, the recommended restaurants were displayed based on the selected restaurant. These recommended restaurants will have characteristics very similar to those of the starting restaurant, respecting the selection criteria defined in the query.



Take a closer look at the code to gain a detailed understanding of its functionality. The following explanation provides an overview of how the code operates:

1. Matches two nodes labeled "Restaurant" such as r1 and r2.
2. Matches two nodes labeled "Restaurant" such as r1 and r2.
3. Apply several conditions in the WHERE clause:
    - Verify that r1 and r2 are different nodes.
    - Check if both r1 and r2 have the cuisine type "Fast Food", or if both do not have the cuisine type "Fast Food" and if they contain at least one kitchen in common.
    - Compare the price levels (price_level) of r1 and r2 considering a maximum difference of 0.5.
    - Compare the ratings of r1 and r2 considering a maximum difference of 0.5.
    - Compare the total number of reviews (total_reviews) of r1 and r2 with a maximum difference of 30.

    - Verify that the latitude and longitude coordinates of nodes r1 and r2 are valid.
    - Verify that the latitude and longitude coordinates are within acceptable values (-90 to 90 for latitude, -180 to 180 for longitude).
    - Calculate the distance in meters between the coordinates of r1 and r2 and verify that it is less than or equal to 3000.
4. Using the clause WITH, the nodes are selected, along with various calculations and aggregations to gauge the difference between restaurants and get an overall score.
5. The results are sorted by total score in descending order.
6. Using the clause WITH, recommended restaurants are collected and limited to the top three for each node r1.
7. Using the clause FOREACH, "RECOMMENDS" relationships are created between r1 and recommended restaurants.
8. Finally, the restaurants and recommendations are returned as a result of the query.

Future developments

The project's success in collecting and analyzing restaurant data in Milan, as well as developing a recommendation system, sets the stage for exciting future developments in the field.By the way there are some possible future developments that would improve the project:

- Real-Time Data Updates: The project focused on retrieving and analyzing restaurant data at a specific point in time. However, future development could involve implementing real-time data updates, ensuring that the restaurant information remains up to

date. This could be achieved through periodic web scraping, API integration, or integrating with restaurant management systems to obtain real-time data feeds.

- Geographic Expansion: While the project focused on restaurants in Milan, future development could involve expanding the scope to cover other cities or regions. This would require adapting the data collection and analysis processes to incorporate data from multiple locations
- User Interface and Visualization: Enhancing the user interface and visualization capabilities of the recommendation system can improve the overall user experience. Presenting the recommendations in an intuitive and visually appealing manner, providing filtering options, and incorporating interactive features can make the system more user-friendly and engaging.
- A Sophisticated Recommendation System: The current recommendation system is very simple and can be further improved by incorporating advanced machine learning algorithms and deep learning techniques.

SITOGRAPHY

1. Pandas: A flexible and easy-to-use open-source data analysis and manipulation tool. URL: https://pandas.pydata.org/
2. Selenium with Python: A Python library for web automation and testing. URL: https://selenium-python.readthedocs.io/
3. The Neo4j Cypher Manual: A comprehensive guide to using the Cypher query language for Neo4j graph databases. URL: https://neo4j.com/docs/cypher-manual/current/
4. TripAdvisor: A popular website for finding and reviewing restaurants, hotels, and other travel-related information. URL: https://www.tripadvisor.com/

5. Google Cloud Console: Google's web-based interface for managing and accessing various Google Cloud services, including APIs. URL: https://console.cloud.google.com/apis/library?pli=1
6. MongoDB in Python: Official MongoDB documentation for using MongoDB with Python. URL: https://www.mongodb.com/languages/python
7. Cartopy: A Python library for geospatial data processing and mapping. URL: https://pypi.org/project/Cartopy/