UPC - GIA

Project 1: LARGE-SCALE DATA ENGINEERING FOR AI

Advanced Databases

Autors:
Elena Alegret & Sergi Tomàs & Júlia Orteu

April 2024

Contents

Abstract

This project delineates the implementation of best practices in operationalizing data science pipelines, with a specific focus on three critical components: Data Ingestion, Data Engineering Pipeline and Data Analysis Pipelines. The discussion revolves around the organizational structure essential for this process, highlighting key elements such as Data Sources, Data Collectors, and the delineation of separate zones to ensure an optimal flow of data to land on the final usage. Utilizing relational databases sourced from prominent platforms such as *Airbnb*, *Mossos d'Esquadra*, and *TripAdvisor*, the project endeavors to develop a tourist visualization tool tailored for Barcelona. This tool offers multifaceted functionalities, primarily enabling users to explore nearby restaurants based on preferences and providing insights into local crime rates. By demonstrating the seamless integration of data science methodologies into the realm of tourism, this project underscores the practical significance and transformative potential of data-driven approaches in enhancing user experiences and decision-making processes. To access the interactive interface integrated into your web browser, visit the BCN Map4Tourism interface.

Key Words: Data engineering, data pipelines, organizational structure, data ingestion, data analysis, relational databases, tourism visualization, Barcelona

1 Introduction

Barcelona, renowned for its culture, has long been a magnet for tourists from around the globe. However, beneath its well-known facade lies a growing concern: an increase in crime rates that has earned it the reputation of being one of the most dangerous cities in Spain. Despite this concern, tourism in Barcelona continues to grow, raising questions about the potential between these two phenomena.

In response to this believe, our project aims to explore the potential correlation between crime rates and tourism. We intent to construct robust pipelines for data collection, storage, and analysis, with the goal of developing a tool that enables users to visualize the impact and distribution of various crime rates on local businesses and tourist attractions.

Drawing from a diverse array of data sources, including official crime statistics, tourism metrics, and business records, we seek to understand how could affect the tourist experience and its decrease.

0ur hypothesis that tourism industry may inadvertently foster criminal activities, thereby deteriorating the crime rate and potentially jeopardizing the safety and well-being of both residents and visitors alike.

This project makes use of three official data sources, detailed in the section 2, Data Ingestion, to explore the correlation between crime rates and tourism in Barcelona.

2 Data Ingestion

In this section, we present the data sources used and describe the implementation of Data Collectors.

2.1 Data Sources

In this subsection, we dive into the datasets used in this project, providing details of our Crime rates, Airbnb listings, and tourist attractions and reviews in Barcelona.

1. **Airbnb / OpenDataSoft**: OpenDataSoft extracts data from [1], a platform designed to provide information regarding Airbnb's impact on residents' homes. While our project's focus differs, we leverage this source to gather information on apartments listed in Barcelona.
2. **Mossos d'esquadra**: The Catalan government offers an open data platform where various crimes committed in Catalonia, including Barcelona, are published alongside their corresponding police departments. This dataset, accessible through [2], provides valuable insights into the types and frequencies of crimes reported in Barcelona.
3. **TripAdvisor**: TripAdvisor ([3]) is platform for users to submit reviews of restaurants, hotels, and various attractions. Leveraging the TripAdvisor API, we extracted data on the 10 nearest points of interest surrounding a subset of 3000 Airbnb apartments in Barcelona. Additionally, we collected up to 10 reviews (if available) for each point of interest.
4. **GoogleMaps API**: Google Maps API is a platform for accessing geolocation data and related information. By integrating Google Maps API data into our project, we gained spatial context, allowing for a more comprehensive understanding of the geographic landscape under examination.

These datasets collectively allow us to explore the relationship between crime rates, Airbnb listings, and tourist attractions in Barcelona. More details about the datasets' content can be found on the Metadata files provided in the *Data Collectors* directory of the deliverable.

2.2 Data Collectors

In this section, we detail the process of data collection and storage for our project.

All datasets were obtained through API (Application Programming Interface) calls, allowing us to retrieve the necessary information programmatically.

> **Reasoning Behind the Decision:**
> APIs provide structured access to data, ensuring seamless integration across different systems and platforms. They offer real-time or near-real-time data, ensuring the information obtained is up-to-date and accurate. Additionally, they facilitate controlled access to data, allowing organizations to enforce security measures and regulate data usage.

Following, the data was stored in a **.parquet format** in the Data Lake. Each dataset is stored as a separate *.parquet* file in the Data Lake. This approach ensures data integrity and facilitates seamless data access and processing.

> **Reasoning Behind the Decision:**
> Parquet is a columnar storage format that is highly optimized for query performance and efficient storage. It offers benefits such as compression, efficient encoding, and support for complex nested data structures.

3 Data Engineering Pipeline

The Data Engineering Pipeline encompasses three key components: Data Formatting, Data Quality and Data Preparation Pipeline.

3.1 Data Formatting Pipeline

In this section, we detail the process of the Data Formatting Pipeline, where data is homogenized according to a canonical data model.

Firstly, a connection with DuckDB, a file-based database, has been established to enable the writing of formatted data. This has been done to ensure that the data is available for further analysis.

Next, a Spark session has been initialized to process the data. It has been configured to use the DuckDB JDBC library, allowing connection to the DuckDB database and writing in the formatted data.

Subsequently, the *.parquet* files containing the raw data from Airbnb, the criminal dataset, and TripAdvisor locations and reviews have been loaded into Spark DataFrames. It is worth mentioning that, due to issues with the Airbnb dataset, preliminary preprocessing has been performed. This includes converting array-type columns to comma-separated strings and removing those that cannot be preprocessed (Spark does not work with array-type columns). This decision has been made to ensure that the data is in a suitable format for further analysis.

Finally, the Spark DataFrames have been written to tables in the same DuckDB database, *barcelona.db*. This allows to homogenized data to an appropriate format. It should be noted that there is one table per dataset.

Reasoning Behind the Decision:
DuckDB is a fast and efficient in-memory database, enabling quick processing of large datasets. By storing Spark tables in DuckDB, query speed is increase and latency is reduced compared to other storage options. Additionally, DuckDB provides a SQL interface, making data access and manipulation easier for future analyses. This offers flexibility and scalability, allowing quick access to stored data and facilitating integration with other systems.

3.2 Data Quality Pipeline

This section delves into the Data Quality Pipeline, highlighting the tasks and techniques employed to ensure data integrity and accuracy.

The three main tasks related to data quality are:

- Identification of Data Quality rules on the datasets: Identify potential data quality issues; errors, inconsistencies, outliers, etc.
- Assessment of the Quality of the Data: Evaluation of the data quality to determine its accuracy, completeness, consistency, and timeliness.
- Application of Data Cleaning processes: Address the issues identified during the data quality assessment; removing duplicate values, correcting formatting errors (column name errors), imputing missing values, standardizing data, etc.

Due to our lack of expert knowledge on the selected fields (criminal and tourism), the preprocessing will be mainly restricted to missing values imputation, the removal of redundant variables and filtering based on our needs.

It is worth highlighting that the decision to correct the formatting errors in the column names is to facilitate the search of relationships between tables in the *Data Explotation Pipeline.* For instance, in order to be able to implement the *Data Analysis Pipeline*, we have seen the need to introduce the *GoogleMaps API*, where we converted the adresses collected into Geographical coordinates using *GoogleMaps API.*

GoogleMaps API was applied on the *Data Quality Pipeline* for reasons associated to the limit of tokens available for API requests.

The details of the quality filtering implementation can be found in the *dataset_ exploration.ipynb* file located on *data_ quality_ pipeline* directory found in the deliverable. Notice how changes are being applied directly on the database instead of being defined as constraints in each table.

> **Reasoning Behind the Decision:**
> As our data ingestion relies on diverse APIs, expanding the database with additional rows implies rerunning the entire system and pipelines. Consequently, imposing constraints on the database may seem less pertinent. Instead, we use the insights gained from dataset exploration to dynamically process our data.

To provide an example of one of the filters applied in the form of a Denial Constraint, we present the following expression:

barcelona_neighborhoods = [Eixample, Sants-Montjuïc, Les Corts, Horta-Guinardó, Sant Martí, Nou Barris, Sarrià-Sant Gervasi, Gràcia, Sant Andreu, Cuitat Vella]

$$\forall t \in \mathrm{R}\neg(\text{t.area\_basica\_policial} \notin \text{barcelona\_neighborhoods})$$

- Ensure only data registered in Barcelona is considered. Since police areas are separated by city district, this has to be assessed using police district rather than the city name.

All operations are implemented using Spark and all changes are stored in the Trusted Zone, the tables are the same as the Formatted Zone even though the reliability of the data has been assessed. The final relational database looks like the following:

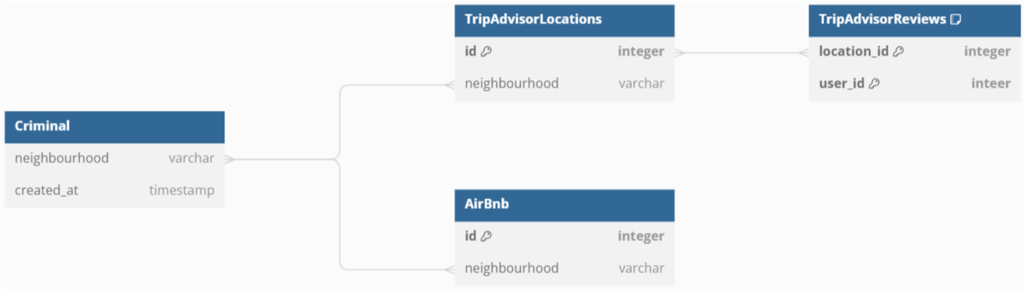


Figure 1: A representation of the variables' role in each dataset.
Note: Not all columns of each table are being represented, only the ones which state as Primary or Foreign Keys.

3.3 Data Preparation Pipeline

In this section, we integrate the datasets located in the Trusted Zone in order to ensure their preparation for the Data Analysis Pipelines.

The pipeline is structured into functions, which apply tasks to consolidate the datasets for use in the Data Analysis Pipeline. It is important to note that the exploration zone only prepares data for analysis. The preprocessing of the dataset is done in the Data Quality Pipeline, see Section 3.2.

> **Reasoning Behind the Decision:**
> Functions make it easier to maintain and update the code. Each function is designed to perform specific tasks which simplifies debugging. Additionally, functions help increasing the scalability of data processing pipelines.

Firstly, the datasets - Airbnb, TripAdvisor, Criminality - are loaded from the Trusted Zone using a combination of JDBC with DuckDB to ensure that the loaded data is the most current.

The relationship between datasets is based on common keys or structures such as geographical locations. The following functions are used to prepare the data:

- `filter_apartments`: Refines the Airbnb dataset based on user preferences configured via the interface, applying filters on:
  - **Review Score**: Filters listings to include only those meeting a specified review score threshold.
  - **Price**: Limits listings to those within a user-defined price range.
  - **Room Type and Amenities Filters**: Enables further refinement by room type and amenities¡.

  The interaction between the Airbnb and TripAdvisor datasets is established, correlating properties near tourist points of interest for combined analyses.

- `criminal_implementation`: Processes crime data to evaluate neighborhood safety by:
  - Aggregating crime incidents by neighborhood and type, identifying areas with higher crime rates.
  - Calculating total and proportional occurrences of each crime type within selected neighborhoods, providing a quantitative assessment of crime.

  Additionally, the average rating per location is calculated by aggregating the ratings provided in the TripAdvisor reviews. This information is then joined with the locations DataFrame. The `df_locations` DataFrame is joined with the average ratings DataFrame using the location_id as the primary key, facilitating the integration of crime and review data for neighborhood evaluation.

The relationship between the Criminal and Airbnb datasets can be when rating the neighborhood where the interesting properties are located.

- `popup_content_review`: Generates interactive popup content for locations on maps based on TripAdvisor reviews by:
  - Filtering and displaying random reviews for specific locations, alongside average ratings and user feedback.
  - Enriching map visualizations with dynamic, user-generated content, enhancing the interactive experience and providing real-time feedback within the application.

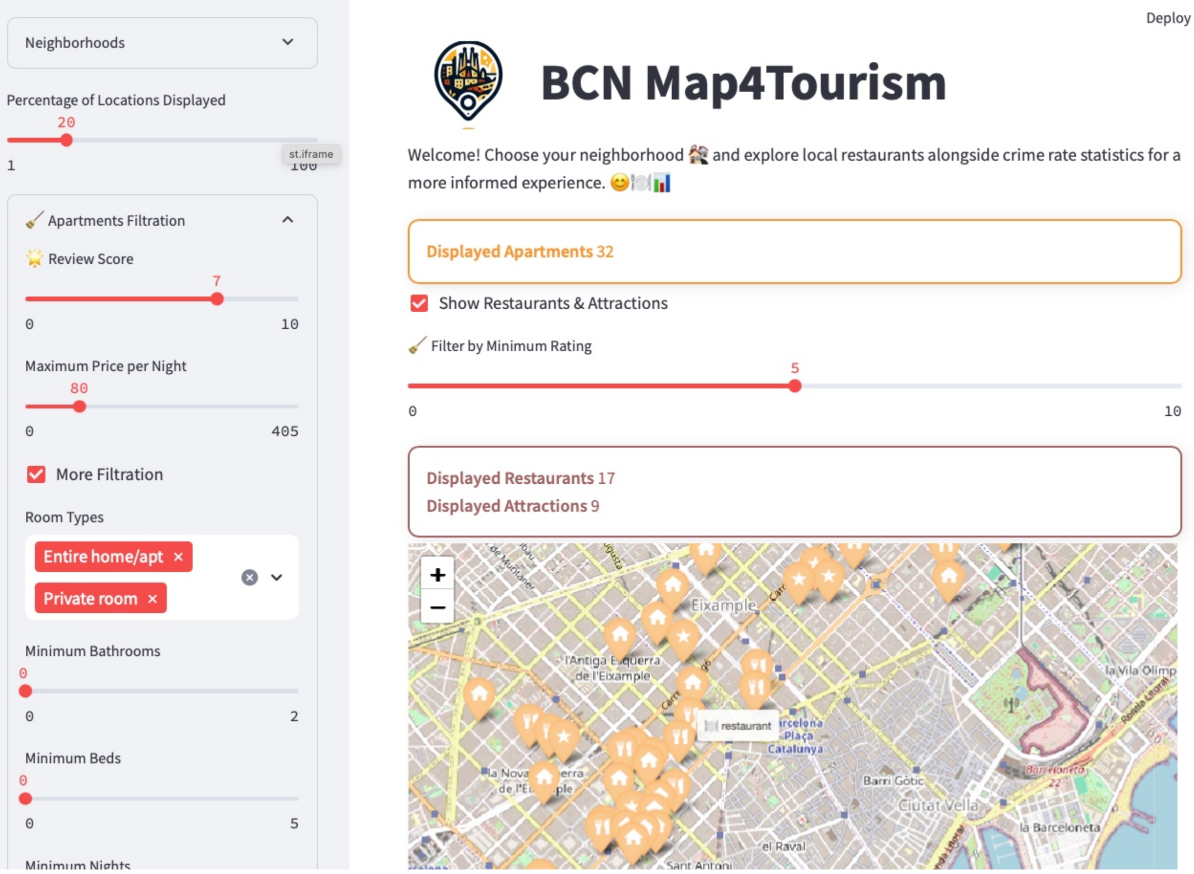
  The relationship between the TripAdvisor and Airbnb datasets is by overlaying reviews and ratings.

4 Data Analysis pipelines

In this section, we present three types of data analysis; where we explain the implementation and the platform designed to offer users a more easy exploration of Barcelona's diverse neighborhoods.

The BCN Map4Tourism interface serves as a tool for tourists and renters in Barcelona, offering a space for neighborhood exploration, restaurant discovery, and crime rate analysis.

Users are provided by a visual interface where they can select specific neighborhoods to explore, aided by a map visualization with markers representing Airbnb listings and nearby restaurants or attractions.



Neighborhoods

Percentage of Locations Displayed: 20 (1–100)

Apartments Filtration

Review Score: 7 (0–10)

Maximum Price per Night: 80 (0–405)

☑ More Filtration

Room Types: Entire home/apt ×, Private room ×

Minimum Bathrooms: 0 (0–2)

Minimum Beds: 0 (0–5)

Minimum Nights

BCN Map4Tourism

Welcome! Choose your neighborhood 🏘 and explore local restaurants alongside crime rate statistics for a more informed experience. 😊🍽📊

Displayed Apartments 32

☑ Show Restaurants & Attractions

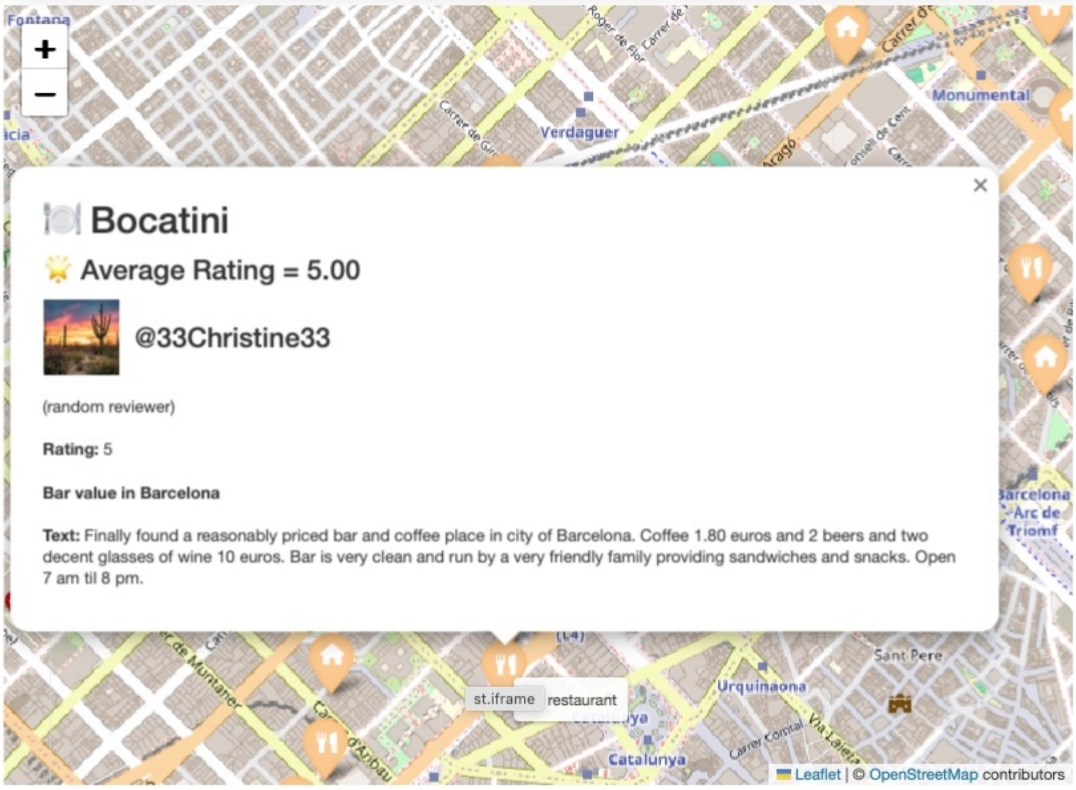
Filter by Minimum Rating: 5 (0–10)

Displayed Restaurants 17

Displayed Attractions 9

Figure 2: Interface.

Through interactive sliders and checkboxes, users can define their preferences, filtering Airbnb listings by review scores, price range, and property features like the number of bathrooms and beds. Additionally, they have the option to display TripAdvisor-rated restaurants and attractions on the map.



🍽️ Bocatini

🌟 Average Rating = 5.00

@33Christine33

(random reviewer)

Rating: 5

Bar value in Barcelona

Text: Finally found a reasonably priced bar and coffee place in city of Barcelona. Coffee 1.80 euros and 2 beers and two decent glasses of wine 10 euros. Bar is very clean and run by a very friendly family providing sandwiches and snacks. Open 7 am til 8 pm.

Figure 3: Review selected randomly from the dataset.

Furthermore, the interface provides insights into neighborhood safety through an analysis of crime rates. By visualizing the top crime types in each selected neighborhood and identifying the highest risk area based on crime ratio, users can prioritize safety in their accommodation.

The BCN Map4Tourism interface helps users to focus their exploration on their preferences by neighborhood, nearby amenities and attractions.

All details about the interface implementation can be found in the file *app.py* in the *data_preparation _pipeline* directory of the deliverable.

Reasoning Behind the Decision:
The file is placed in this location to facilitate the importation of functions from the preceding pipeline. However, it represents our various analysis and modeling pipelines.

It is worth mentioning that the inclusion of three data visualizations — TripAdvisor, crime rates, and Airbnb — is observed. Each visualization serves a distinct purpose:

- TripAdvisor aids in discovering nearby amenities.
- Crime rates prioritize safety considerations.
- Airbnb facilitates accommodation search.

They provide a platform to inform decision-making and exploration to tourists.

5 Conclusions

This project has been a great way to understand how more complex data processes work. Designing a full set of pipelines to land on a final framework has been both challenging and insightful. Addressing the collection and processing of data is crucial and defiantly not a trivial task. We also have had the opportunity to work with `pyspark` and `sparksql` to virtually (since all is executed in a single machine) parallelize our code and processes.

We are proud of the final result, BCN Map4Tourism is a very useful tool for minorities to look for they ideal apartment and attractions in Barcelona while being able to develop their awareness around the dangers of the city of Barcelona.

References

[1] Airbnb. *Inside Airbnb*. URL: https://insideairbnb.com/ (visited on 04/22/2024).

[2] Mossos d'esquadra. *Criminal Dataset of Barcelona*. URL: https://analisi.transparenciacatalunya.cat/Seguretat/Fets-delictius-i-infraccions-administratives-de-l-/y48r-ae59/about_data (visited on 04/22/2024).

[3] Tripadvisor. *Tripadvisor Dataset*. URL: https://www.tripadvisor.com/ (visited on 04/22/2024).