Analysis of niche tourist behaviour based on social network data

Ca' Foscari University of Venice
Department of Environmental Sciences, Informatics and Statistics

Computer Science Master's Thesis
Year 2021-2022

Graduand	Lorenzo Padoan
Supervisor	Claudio Lucchese
Assistant supervisor	Alessandra Raffaetà

Acknowledgments

I want to thank my grandparents and my mother for giving me a roof to sleep under, breads to eat off of and for allowing me access to an advanced education. The Cassoneria group that brought lighthearted breezes in my life starting from high school through master's degree and beyond. An important mention goes to my academic supervisors, Professor Lucchese and Professor Raffaetà, for their helpfulness and compentence in following me. I also want to thank my colleagues in Motion Analytica especially Bruno and Marco for assisting and taking my first steps in this beautiful field.

Abstract

Data Science has become increasingly important in recent years due to the growing volume of generated data. With the help of this discipline, it is possible to make sense of large amounts of data, commonly referred to as "Big Data" which are essential nowadays to drive the decision-making process. Data is produced in all fields, including Tourism and Mobility, where data is used to track the movement of people, identify tourist attractions and better understand customer behavior. This dissertation born through a collaboration with Motion Analytica Srl, an Italian startup that focuses on data analytics in Mobility and Tourism. The set goal is to find niche behaviors of tourists from the raw data from Tripadvisor's platform. The work is organized in four parts. First, an Exploratory Data Analysis (EDA) phase. Second, a refining and cleaning of the data followed by application of TF-IDF (Term Frequency-Inverse Document Frequency) an approach to detect the non-naive behavior of customers - tourists in Italian points of interest. This approach is applied to a large dataset of TripAdvisor reviews, kindly granted by Motion Analytica Srl. Fourth, development of a data visualization of the refined dataset. The last part describes the ETL (Extract, Transform and Load) pipeline implemented for the business. The outcomes of this work include a working ETL pipeline and dashboard based on refined data that allow for the analysis of tourist behavior in Italy and the identification of niche tourist behavior.

Keywords Information Retrieval, Data Mining, Data Analysis, ETL

Glossary

API Application Programming Interface

BI Business Intelligence

DW Data Warehouse

EDA Exploratory Data Analysis

HTML HyperText Markup Language

ML Machine Learning

POI Point Of Interest

TF-IDF Term Frequency-Inverse Document Frequency

TOD Time Origin Destination

TOS Term Of Service

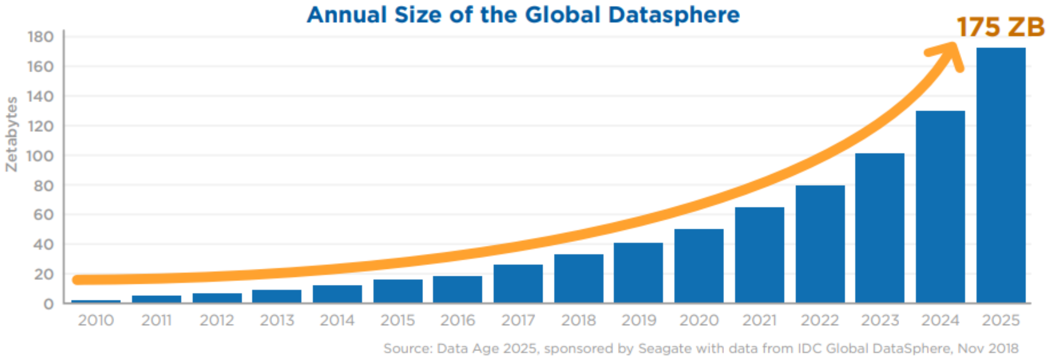
UGC User Generated Content

Contents

Chapter 1

Introduction

The field of Data Science is an interdisciplinary one that incorporates Computer Science, Machine Learning, Statistics, as well as Predictive Analysis and new technologies. This makes it possible for businesses to more effectively comprehend vast amounts of data coming from a variety of sources, generate useful insights, and make decisions that are more data-driven. Both the generation of new information and the application of that knowledge in real-world settings are extremely significant parts of economic activity in the modern world. The only way for raw data to have any value is to be processed into information that can be acted upon, hence transformation is essential. Discovering hidden patterns in massive datasets that could be structured or unstructured is an important part of the data science approach, as well as mining large databases to derive actionable insights is. The International Data Corporation (IDC) is responsible for compiling a report with the title Data Age 2025 [7]. This paper makes the audacious prediction that by the year 2025, global data creation would have increased to 175 zettabytes, from 33 zettabytes in 2018. This is illustrative of the amazing scale of the information that is being generated in the modern day.



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Figure 1.1: Data production trend

As a direct consequence of the remarkable increase in the volume of data created over the course of the past decade and the ongoing process of refining this data for use in a wide variety of applications across many different industries, a whole host of professionals who work in this field have emerged:

- **Date Engineer**: a member of the information technology workforce whose primary responsibility is to prepare data for use in analytical or operational processes. Typically, their responsibilities include the construction of data pipelines, which are used to collect information from a variety of source systems. They arrange the data after integrating, consolidating, and cleaning it in preparation for its use in analytical applications. They want to simplify the process of accessing data while also working to improve the big data ecosystem within their firm.

- **Data Scientist**: a specialist who is accountable for gathering, analyzing, and interpreting very large amounts of data. The function of the data scientist is an evolution of various traditional technical roles, such as that of the mathematician, the statistician, and the computer professional. The employment of sophisticated analytics technologies, such as machine learning and predictive modeling, is a prerequisite for consideration for this position.

- **Machine Learning Engineer**: a highly specialised IT member who concentrates on the investigation, construction, and design of autonomous artificial intelligence (AI) systems in order to automate predictive models, verify data quality and perform model maintenance over time.

Despite the identification of these work figures and the fact that this is a relatively new field, currently workers in this industry are frequently a blend of the figures listed above.

1.1 Social Media Opportunity

The term "social media" refers to a category of different kinds of online communication platforms, the most common examples of which are websites and mobile applications, which give users the ability to generate and share information or to take part in social networking [20].

The rise of social media as an integral part of modern communication has brought about profound shifts in the ways in which individuals engage with one another and the world around them. They have a disproportionately huge influence on our culture as a result of the fact that they enable individuals to communicate with one another and readily disseminate their thoughts and information to a broad group of people. In addition, platforms for social media have emerged as a significant source of news and information, in addition to playing an increasingly important role as a medium for marketing, advertising, and promotion.

The number of people using social media has surged over the last few years, but it has never increased at such a rapid rate as it has since the COVID-19 outbreak. This is because lockdowns and limitations have transformed user behavior and habits toward more digital reliance. In July 2020, there was an increase of 10.4 % in the usage of social media compared to the same period of the previous year [9], and TikTok attracted 12 million individual visitors in the United States in March 2020 [17].

The amazing increase in the use of social media has resulted the further development of a growing collection of both structured and unstructured data from a variety of formats, including photographs, videos, sounds, texts, and geolocations, among other things. The use of social media has evolved to the point that it is now an essential factor in the process of gathering and disseminating information across a variety of fields, including journalism, business, politics, and the sciences. This increase reveals new opportunities for research and discovery of patterns, which may shed light on significant problems, developments, and effects, as well as commercial and societal shifts. The level of difficulty involved in accomplishing tasks such as data discovery, collecting, and preparation for analytical and predictive modeling frequently varies depending on the specific application area, data source and format, methodology, and goals that are being pursued.

Methods from the field of data science, including but not limited to big data analytics, data mining, machine learning, and artificial intelligence, are utilized extensively in this direction.

1.1.1 Data Collection Process

The process of gathering information from websites is known as web scraping.

It requires loading a website and then obtaining the data from the HTML code of the page. This may be done by hand, but in most cases it is accomplished with the assistance of a computer application that can automatically load sites and extract the data. Web scraping is a method that may be used to get data from websites that do not offer an application programming interface (API) or from websites that do not provide access to their data. This practice is almost often mislabeled as unlawful, despite the fact that it is abided by all applicable laws to the letter. Provided that no personally identifiable information is gathered and that the terms of service (TOS) are adhered to. The decision made by the Supreme Court of the United States on April 18, 2022 further validated this reality [6].

The use of this process necessitates non-trivial software engineering expertise; however, this process is frequently useful for the collection of data independently by the company and is thus frequently considered as an alternative to a direct agreement with the company holding the data. This has resulted in the development over the last decade of a whole system of start-ups based on this system and, as a result, an economic development that is not entirely dependent on the medium to large companies, that hold the data. The utilization of scraping in a manner that is compliant with rules can be a tool that enables the formation of new entrepreneurial realities and, as a result, the advancement of technological capabilities [13]. This dissertation describes a work conducted within European borders, that complies with the General Data Protection Regulation, one of the most cutting-edge data protection standards in the world (GDPR).

1.1.2 User-Generated Tourist Data

The past ten years have seen the development of Web 2.0, which has brought with it a broad array of platforms where users actively exchange content with one another. Social networking, picture sharing, wikis, and folksonomy are all examples of social media. The emergence of these platforms has made it possible for travelers to post online accounts of their own personal experiences, which gives tourism and activities related to tourism a major role in many of these services [22]. These platforms contain information that may be used for trip planning, evaluations of destinations and hotels, photos, and recommendations for tourist activities. They also led to a democratization of travel writing and the supply of information since it incorporates the freely stated opinions of travelers who have been to the location in issue or participated in the activity in question.

The proliferation of social media and user-generated content (UGC) gives individuals an unprecedented amount of power to instantly add 'digital traces' when performing tasks such as writing reviews of services provided by airlines, hotels, and restaurants; registering a customer complaint; keeping a travel journal; or uploading photographs and videos to a global big data bank.

TripAdvisor is one of the most widespread platform that vacationers use to discuss and share their experiences. These kinds of social media platforms are growing increasingly popular and are turning into key online travel information sources, often being more extensive and precise than the websites of destination management organizations. The UGC is becoming an increasingly viable option for national tourist organizations, destination management organizations, and other stakeholders as a rich data source.

Businesses can make use of UGC in order to assist them in improving the tourism experience they offer by indicating how to personalize and tailor services and products to different types of visitors. However, policy makers and destination managers can also make use of UGC in order to generate information about tourist behavior, the functioning of the wider destination system, and the perceived image and quality of the various services offered in the destination [11].

Analysis of the vast quantity of data that is generated by UGC platforms, which is always growing, can supply the input that is essential for a more data-driven type of policy making as well as doing business. The transformation of data into information and the application of that information in such a way that it provides knowledge that would not otherwise be accessible without the use of (big) data analysis are the two most important steps [16].

1.2 Tourism in Italy

One of Italy's most important contributors to the country's Gross Domestic Product (GDP) is the tourism industry. According to ISTAT, Italy is the fifth most visited country in terms of international tourism arrivals, with 65 million tourists per year (2019). However, Italy is the second most visited country in terms of nights spent in hotel, coming in behind Spain with 220,7 million foreign visitor nights spent and a total of 436,7 million nights. The Bank of Italy estimates that the tourist industry directly creates more than five percent of the national GDP (13 percent when it is included the GDP that is created indirectly), and it employs more than six percent of the working population [14] [2] [21].

These days, the things that draw tourists to Italy are primarily its culture, food, history, fashion, architecture, art, religious sites and routes, nature beauty, nightlife, undersea sites, and spas. Tourism in both the winter and summer seasons is prevalent in a number of areas inside the Alps and the Apennines, whilst tourism at seaside resorts is prevalent in a number of coastal regions along the Mediterranean Sea.

Milan is the 27th most visited city in the world with a total of 6.8 million visitors, whereas Rome is the third most visited city in Europe and the 12th most visited city in the world with 9.4 million arrivals in 2017. In addition, Venice and Florence are both included in the top one hundred travel destinations in the world. In addition to this, Italy is the nation that is home to the most sites that have been designated as UNESCO World Heritage Sites (58). 53 of them are cultural sites, while the remaining 5 are natural [4].

This industry was particularly vulnerable during the COVID-19 pandemic epidemic, which resulted in a fall of over 60% in tourism income in Italy and a decrease of more than 50% in touristic presences in comparison to 2019. The worldwide public health crisis made it impossible to resume the operations during the first few months of the year 2021 [5]. After the emergency, there was a slow return to normalcy, which is currently experiencing extreme uncertainty due to the present global energy crisis. This situation can affect negatively tourism because it can lead to higher prices for electricity, gas, and other forms of energy. This can make it more expensive for tourists to travel, and it can also make it more difficult for businesses in the tourism industry to operate.

1.3 Context of work & goals

The following work was developed during a curricular internship within the company Motion Analytica [1] an innovative start-up that provides analytical services to the tourism industry and provides insights into the behaviour of tourists to stakeholders both public and private, as well as decision-makers in government, and destination marketing organizations. The work performed is based on a data derived from the social networking platform Tripadvisor. The set goal for this curricular internship is to find a way to observe not naive behaviors of tourists starting from the raw data. To achieve the set goal, the various tasks that comprise most of the data chain work are executed, such as: Data cleaning, the application of data mining techniques and design of data visualizations. What has been obtained from this work is a working ETL system and dashboards that make it possible to identify the niche behaviors of tourists in Italy.

1.4 Related work

Inherent work in this area has been done previously through Motion Analytica's collaboration with Ca' Foscari University of Venice. This previous work was performed on a TripAdvisor data set. The goal of this work was to explore a method for understanding the perceived quality of a POI. The work concluded with an identification of a metric called True Score that was satisfactory for Motion Analytica SRL's business purposes [3].

1.5 Outline

The thesis is arranged according to the following structure. In Chapter 2, It is provided a concise introduction to the particulars pertaining to the data that is available for this study. The cleaning procedure and the application of the data mining algorithm is discussed in Chapter 3. Chapter 4 shows data visualization using refined data. Chapter 5 discusses ETL designs and what tools and technologies were used.

Chapter 2

Data Description & Exploration

Due to the needs of Motion Analytica, information on the behavior of visitors in Italy is of special relevance.

Motion Analytica hired a contractor to build a crawler capable of obtaining all of the sites of interest in the targeted area, as well as reviews and other information posted by users of those locations on the TripAdvisor platform. The length at which the contractor gathers a batch of data and stores it into the company's storage system is determined by the business. This batch may comprise brand-new data or an update to previously collected data. The web scraping service transports the TripAdvisor data stream to the storage system.

In the sections that follow, we will show a range of statistical data and the types of data that are available. As previously noted, data are susceptible to change.

Our observations of the data are based only on a particular moment in time. Using this data image as a reference enables us to operate from a more secure position. Each of our assertions pertaining to this specific data snapshot is readily generalizable and applicable to future and updated data.

The data that were used for this work did not come solely from Tripadvisor. Additionally, data that had been previously stored in the company's system were utilized. This second type of data mostly consists of information on Italian municipalities and provinces that was provided by ISTAT.

2.1 Data Structure

The data that were utilized for this project are kept in a database. More precisely, the database management system that was used was PostgreSQL, which is commonly known as Postgres. Postgres is a free and open-source relational database management system. TripAdvisor and Public are the diagrams that were used for this study. These diagrams will be displayed using only the fields that were actually employed for this work, which will allow the data to be displayed in a manner that is both more concise and easier to understand. They are visualized by means of Entity Relationship Diagram (ERD). ERD is a graphical depiction of relationships between people, objects, places, concepts, or events in an information technology (IT) system. An ERD employs data modeling approaches to assist in the definition of business processes and as the foundation for a relational database.

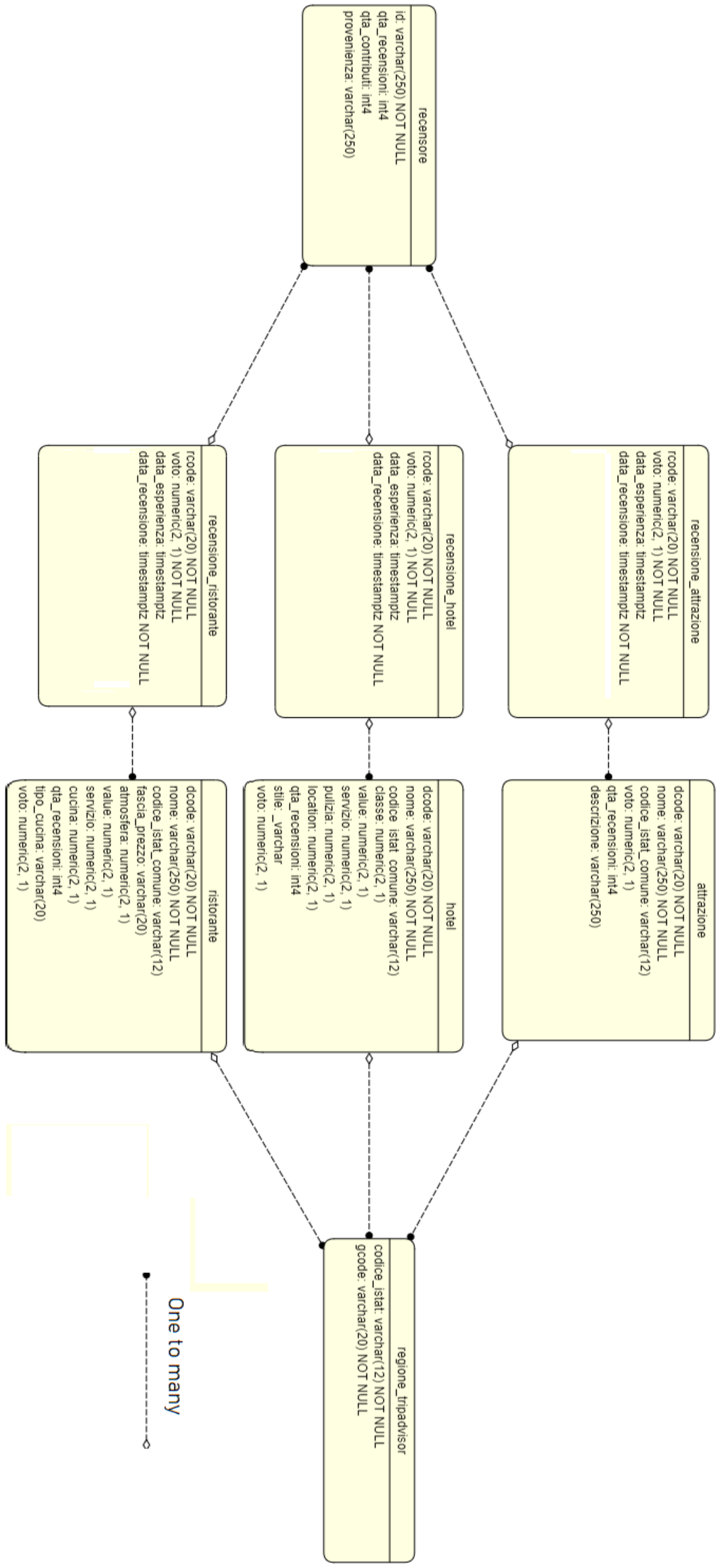
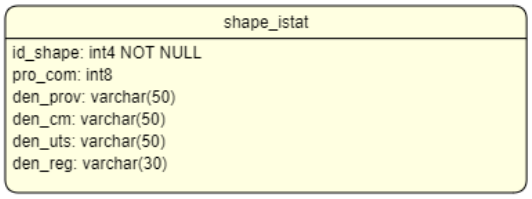

recensore
id: varchar(250) NOT NULL
qta_recensioni: int4
qta_contributi: int4
provenienza: varchar(250)
recensione_ristorante
rcode: varchar(20) NOT NULL
voto: numeric(2, 1) NOT NULL
data_esperienza: timestamptz
data_recensione: timestamptz NOT NULL
recensione_hotel
rcode: varchar(20) NOT NULL
voto: numeric(2, 1) NOT NULL
data_esperienza: timestamptz
data_recensione: timestamptz NOT NULL
recensione_attrazione
rcode: varchar(20) NOT NULL
voto: numeric(2, 1) NOT NULL
data_esperienza: timestamptz
data_recensione: timestamptz NOT NULL
ristorante
dcode: varchar(20) NOT NULL
nome: varchar(250) NOT NULL
codice_istat_comune: varchar(12)
fascia_prezzo: varchar(20)
atmosfera: numeric(2, 1)
value: numeric(2, 1)
servizio: numeric(2, 1)
cucina: numeric(2, 1)
qta_recensioni: int4
tipo_cucina: varchar(20)
voto: numeric(2, 1)
hotel
dcode: varchar(20) NOT NULL
nome: varchar(250) NOT NULL
codice_istat_comune: varchar(12)
classe: numeric(2, 1)
value: numeric(2, 1)
servizio: numeric(2, 1)
pulizia: numeric(2, 1)
location: numeric(2, 1)
qta_recensioni: int4
stile: _varchar
voto: numeric(2, 1)
attrazione
dcode: varchar(20) NOT NULL
nome: varchar(250) NOT NULL
codice_istat_comune: varchar(12)
voto: numeric(2, 1)
qta_recensioni: int4
descrizione: varchar(250)
regione_tripadvisor
codice_istat: varchar(12) NOT NULL
gcode: varchar(20) NOT NULL
One to many

Figure 2.1: Tripdadvisor scheme ER diagram



shape_istat

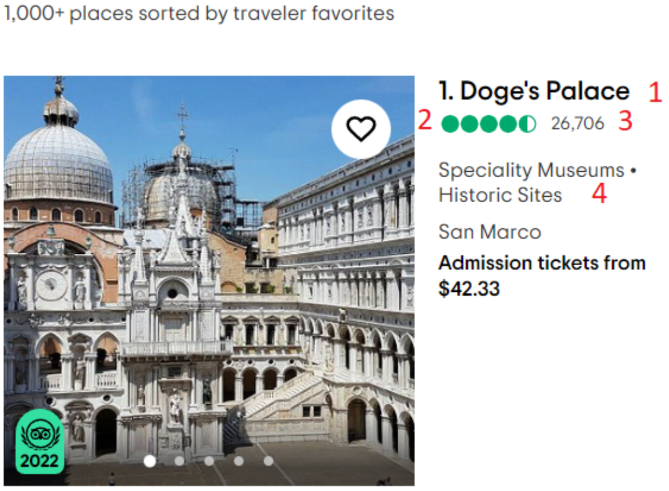
id_shape: int4 NOT NULL
pro_com: int8
den_prov: varchar(50)
den_cm: varchar(50)
den_uts: varchar(50)
den_reg: varchar(30)

Figure 2.2: Public scheme ER diagram

In the next sections some screenshots from the Tripadvisor website are shown. They depict the most relevant fields of the previous tables in order to facilitate the comprehension of their significance.

2.1.1 Attraction

The information about the Attraction POI type that was taken from the TripAdvisor website is the name of the attraction, the average rating given by reviewers, the amount of reviews the attraction has and the category to which the attraction belongs. The mapping of this data with the fields contained in the diagrams described above can be seen in the Figure 2.3.



1,000+ places sorted by traveler favorites

1. Doge's Palace

●●●●◐ 26,706

Speciality Museums • Historic Sites

San Marco

Admission tickets from $42.33

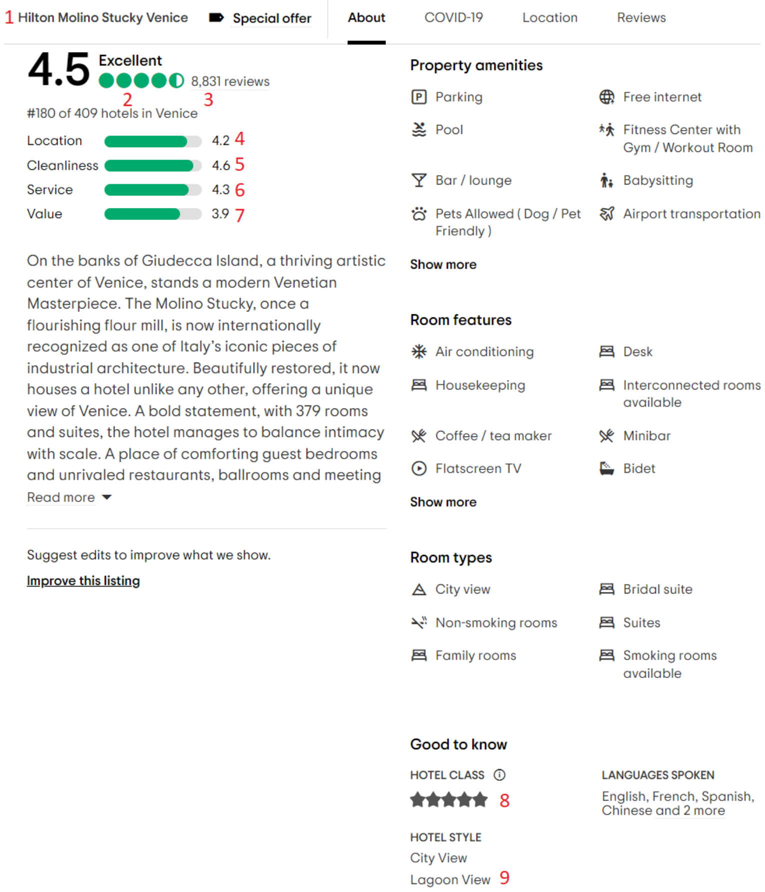
Figure 2.3: Illustration of the attributes of the attraction class

Table [attrazione_info] + [attrazione]

1. nome: varchar(250)
2. voto: numeric(2,1)
3. qta_recensioni: int4
4. descrizione: varchar(250)

2.1.2 Hotel

The fields obtained for the type of Hotel POI are name, average rating, number of reviews, and a slew of attributes specific to the type of hotel POI such as: Average rating of location, average rating of cleanliness, average rating of service, average rating of value, number of stars, and the type of hotel that TripAdvisor identifies. The Figure 2.4 depicts the mapping of these data to the fields in the ERD.



1 Hilton Molino Stucky Venice   Special offer   About   COVID-19   Location   Reviews

4.5 Excellent ●●●●◐ 8,831 reviews
2 3
#180 of 409 hotels in Venice

Location	4.2	4
Cleanliness	4.6	5
Service	4.3	6
Value	3.9	7

On the banks of Giudecca Island, a thriving artistic center of Venice, stands a modern Venetian Masterpiece. The Molino Stucky, once a flourishing flour mill, is now internationally recognized as one of Italy's iconic pieces of industrial architecture. Beautifully restored, it now houses a hotel unlike any other, offering a unique view of Venice. A bold statement, with 379 rooms and suites, the hotel manages to balance intimacy with scale. A place of comforting guest bedrooms and unrivaled restaurants, ballrooms and meeting
Read more

Suggest edits to improve what we show.
Improve this listing

Property amenities
Parking — Free internet
Pool — Fitness Center with Gym / Workout Room
Bar / lounge — Babysitting
Pets Allowed (Dog / Pet Friendly) — Airport transportation
Show more

Room features
Air conditioning — Desk
Housekeeping — Interconnected rooms available
Coffee / tea maker — Minibar
Flatscreen TV — Bidet
Show more

Room types
City view — Bridal suite
Non-smoking rooms — Suites
Family rooms — Smoking rooms available

Good to know
HOTEL CLASS
★★★★★ 8
LANGUAGES SPOKEN
English, French, Spanish, Chinese and 2 more
HOTEL STYLE
City View
Lagoon View 9

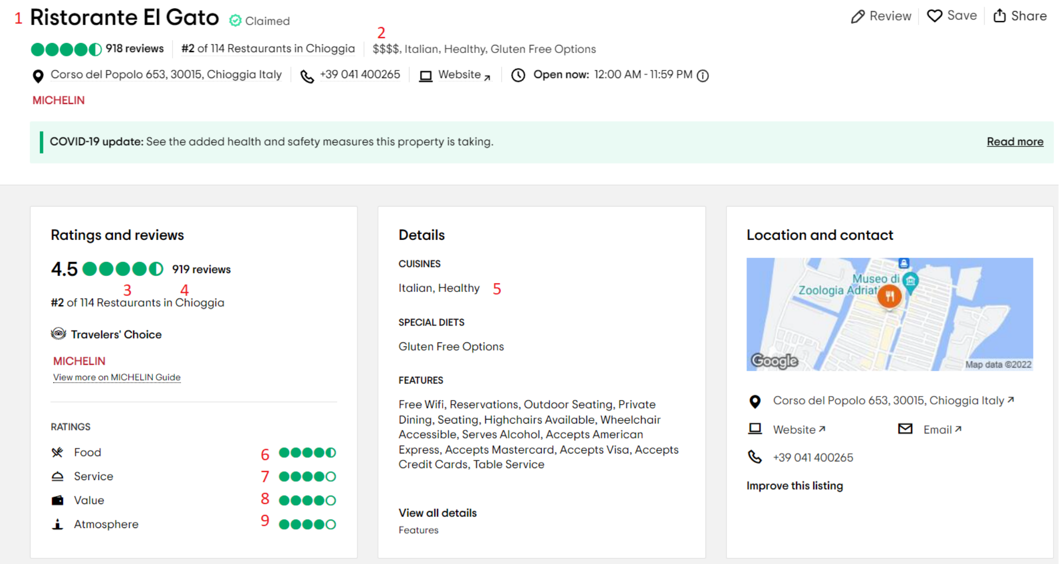
Figure 2.4: Illustration of the attributes of the hotel class

Table [hotel_info] + [hotel]

1. nome: varchar(250)
2. voto: numeric(2,1)
3. qta_recensioni: int4
4. location: numeric(2,1)
5. pulizia: numeric(2,1)
6. servizio: numeric(2,1)
7. value: numeric(2,1)
8. classe: numeric(2,1)
9. stile: varchar(250)

2.1.3 Restaurant

The fields for the type of POI Restaurant are the restaurant name, cost class, average rating, number of reviews, and a number of attributes that are distinctive to that type of POI Restaurant, such as the average food rating, average service rating, average value rating, average atmosphere rating, and type of restaurant cuisine. The Figure 2.5 shows how these data in the TripAdvisor website are linked to the fields in the ERD listed above.



Ristorante El Gato ✓ Claimed

Review | Save | Share

●●●●◐ 918 reviews | #2 of 114 Restaurants in Chioggia | $$$$, Italian, Healthy, Gluten Free Options

Corso del Popolo 653, 30015, Chioggia Italy | +39 041 400265 | Website | Open now: 12:00 AM - 11:59 PM

MICHELIN

COVID-19 update: See the added health and safety measures this property is taking. Read more

Ratings and reviews

4.5 ●●●●◐ 919 reviews

#2 of 114 Restaurants in Chioggia

Travelers' Choice

MICHELIN

View more on MICHELIN Guide

RATINGS

Food	●●●●◐
Service	●●●●○
Value	●●●●○
Atmosphere	●●●●○

Details

CUISINES

Italian, Healthy

SPECIAL DIETS

Gluten Free Options

FEATURES

Free Wifi, Reservations, Outdoor Seating, Private Dining, Seating, Highchairs Available, Wheelchair Accessible, Serves Alcohol, Accepts American Express, Accepts Mastercard, Accepts Visa, Accepts Credit Cards, Table Service

View all details

Features

Location and contact

Corso del Popolo 653, 30015, Chioggia Italy

Website | Email

+39 041 400265

Improve this listing

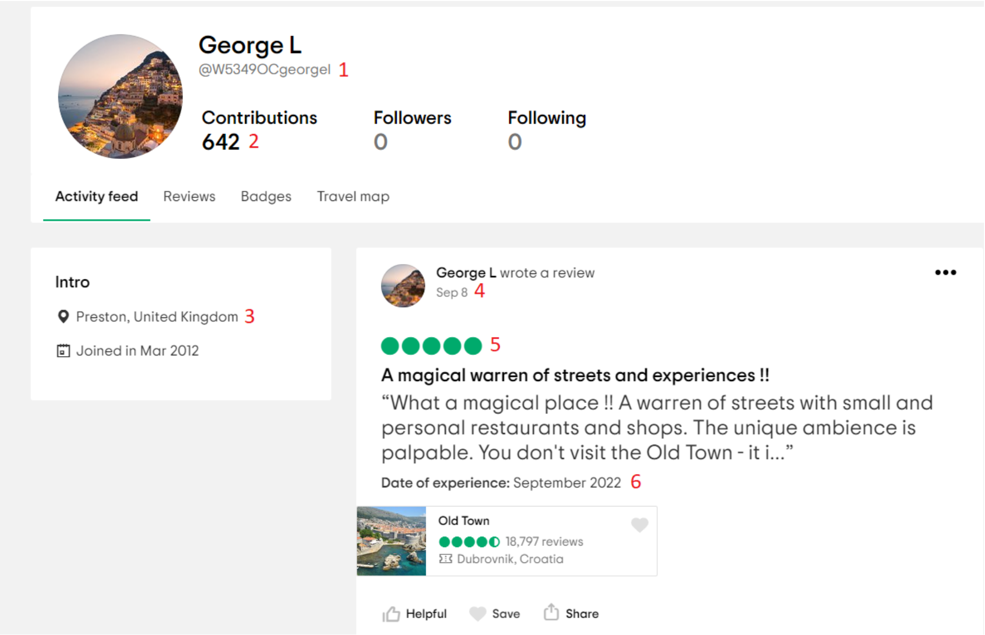
Figure 2.5: Illustration of the attributes of the restaurant class

Table [ristorante_info] + [ristorante]

1. nome: varchar(250)
2. fascia_prezzo: varchar(20)
3. voto: numeric(2,1)
4. qta_recensioni: int4
5. tipo_cucina: varchar(20)
6. cucina: numeric(2,1)
7. servizio: numeric(2,1)
8. atmosfera: numeric(2,1)
9. value: numeric(2,1)

2.1.4 Reviewer

The fields obtained from the TripAdvisor review page are: ID of the reviewer, the total number of reviews he/she made, the date on which he/she made the review, the rating he/she gave to the POI, the date on which he/she made the experience. The mapping of these attributes with the those contained within the ERD are shown in the Figure 2.6



George L
@W5349OCgeorgel 1

Contributions 642 2
Followers 0
Following 0

Intro
Preston, United Kingdom 3
Joined in Mar 2012

George L wrote a review
Sep 8 4

●●●●● 5

A magical warren of streets and experiences !!
"What a magical place !! A warren of streets with small and personal restaurants and shops. The unique ambience is palpable. You don't visit the Old Town - it i..."
Date of experience: September 2022 6

Old Town
●●●●◐ 18,797 reviews
Dubrovnik, Croatia

Figure 2.6: Illustration of the attributes of the reviewer

Table [recensore] + [recensione_attrazione] + [recensione_hotel] + [recensione_ristorante]

1. id: varchar(250)
2. qta_recensioni: int4
3. provenienza: varchar(250)
4. data_recensione: timestampz
5. voto: numeric(2,1)
6. data_esperienza: timestampz

2.2 Data Exploration

Data exploration is the initial step of data analysis, and its purpose is to investigate and display data in order to discover insights from the very beginning of the process or to locate regions or trends that require further investigation.
We are able to process visual information far more quickly and easily than numerical information because humans are visual learners.
Effective data exploration of metadata is facilitated by the use of data visualization tools and elements such as graphs, colors and lines. This makes it possible to identify relationships or inconsistencies in the data. It is possible to have a better understanding of the wider picture and arrive at insights more quickly by making use of dashboards and visualization in general. Tableu, a data visualization tool that offers graphical and pictorial representations of data, is the one that is utilized in order to do this particular work.
We will begin the data investigation of this data sample with a somewhat more general overview, and then move on to somewhat more specific features of the data as we progress.

2.2.1 POIs distribution

Within the scope of this section, we will investigate the percentage of POIs that are represented across the dataset.
The pie chart of figure 2.7 makes the disparity among the categories of POIs in the dataset abundantly evident. 61% of the points of interest in the sample are restaurants, followed by 15% hotels and 22% attractions.

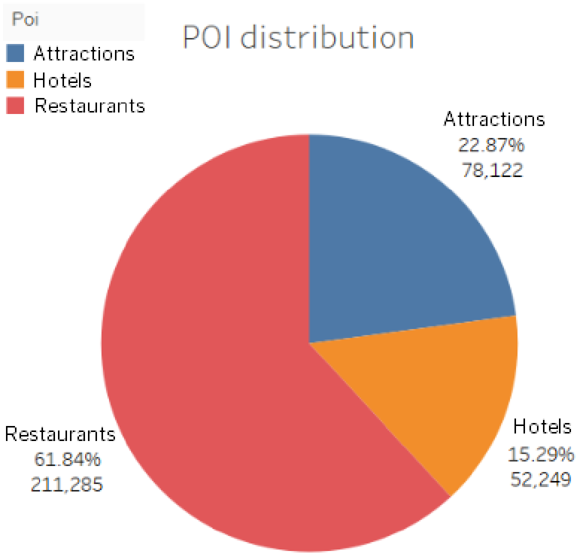


Figure 2.7: POI proportions

Figure 2.8 shows how this dataset distributes POIs across Italy. This distribution of POIs is clearly related to visitor influx in various places [8].

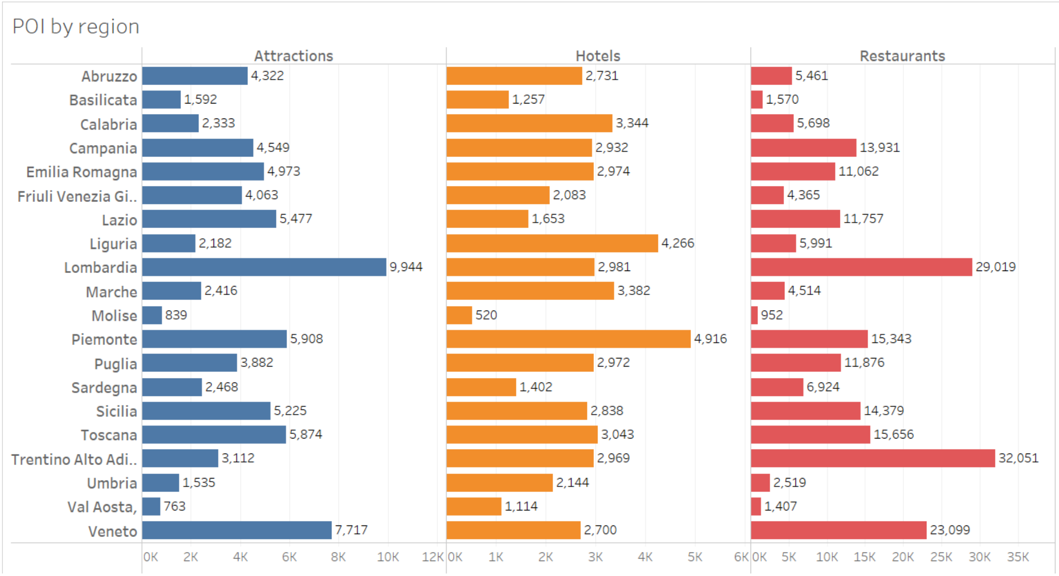


Figure 2.8: POIs distribution by region

2.2.2 POIs reviews distribution

This section examines the percentage of POI reviews in the dataset. Figure 2.9 shows the difference between the dataset's POI categories. 23% of the sample's POIs are hotels, 68% are restaurants. Although there are more attractions in the dataset than hotels as shown in Figure 2.7, this is not the case for the number of reviews. Tripadvisor is recognized for hotel and restaurant reviews, and this dataset reflects that.

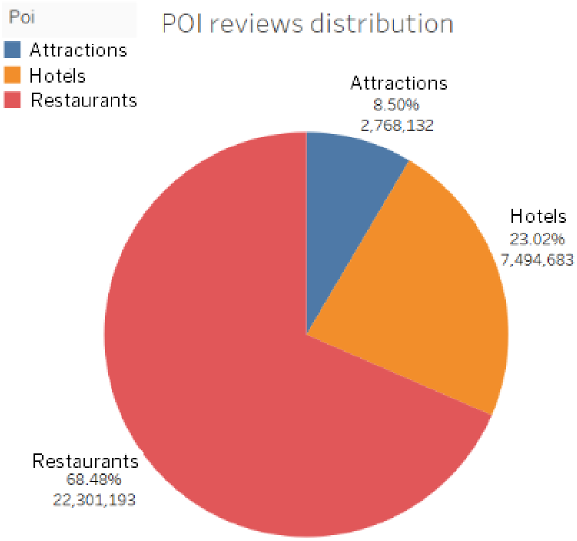


Figure 2.9: POI reviews proportions

Figure 2.10 represents the distribution of reviews in the different POIs. This is in line with the distribution of POIs in Figure 2.9, regarding restaurants and hotels. These numbers are further influenced by the fact that Motion Analytica asked the contractor that provides data to a particular focus and consequent data collection on the Veneto, Trentino Alto Adige and Friuli Venezia Giulia region.

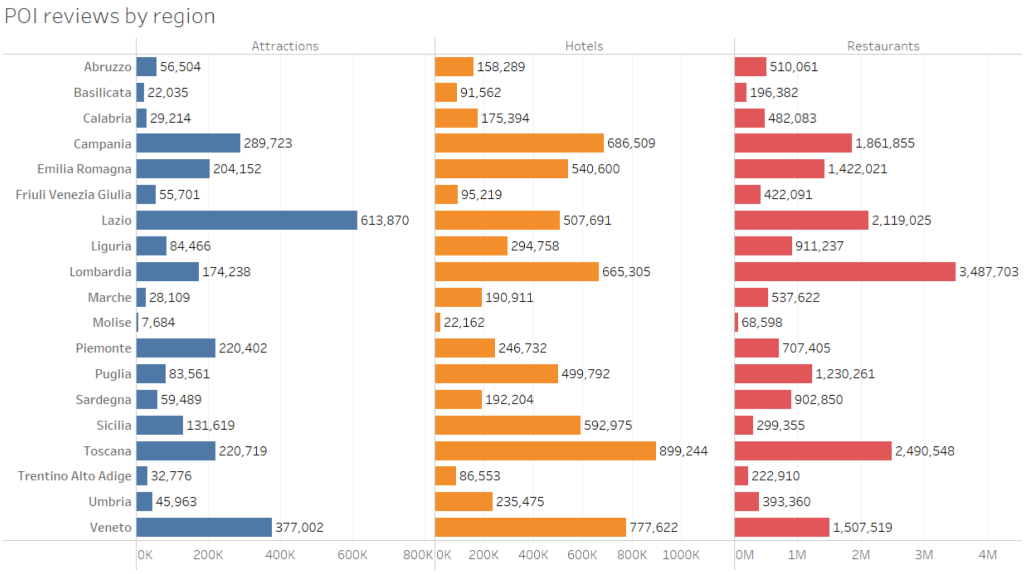


Figure 2.10: POIs distribution by region

Regarding the distribution of votes to the various categories of hotels presented in Figure 2.11, it is possible to observe a Gaussian distribution around the 4.5 level, with the exception of the *Class* field, where the presence of nulls is very pronounced and reaches a level of 65%, while the absence of data in the other fields is approximately 18%.

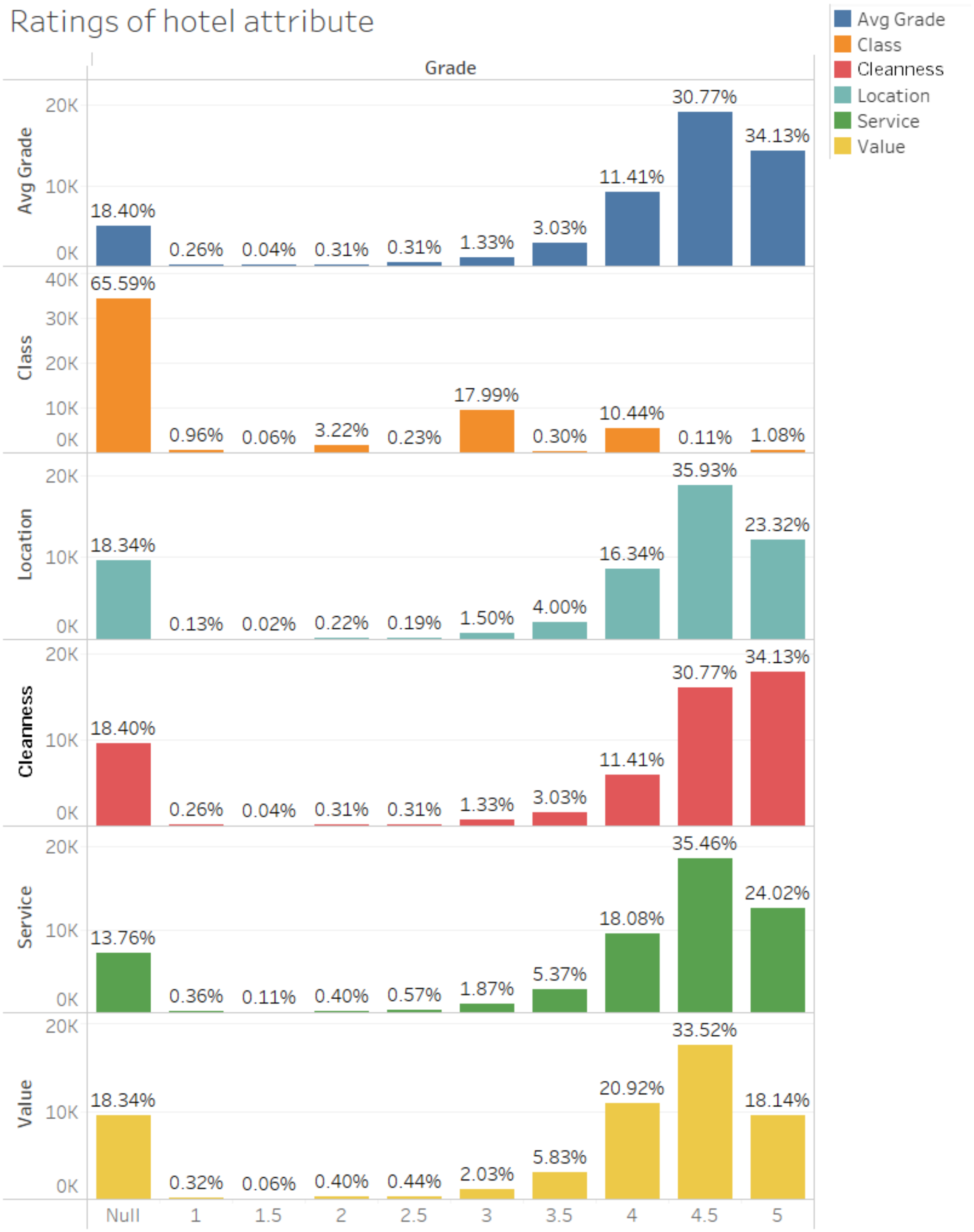


Figure 2.11: POIs distribution by region

Instead, regarding the distribution of ratings to the various restaurant categories depicted in Figure 2.12, it is possible to observe a Gaussian distribution around level 4. Compared to figure 2.11, the lack of data is much more pronounced. Regarding the average grade and *Atmosphere* fields, the value is approximately 67%, whereas for the *Cuisine*, *Service*, and *Value* fields, the number of null values decreases but remains at 39%.

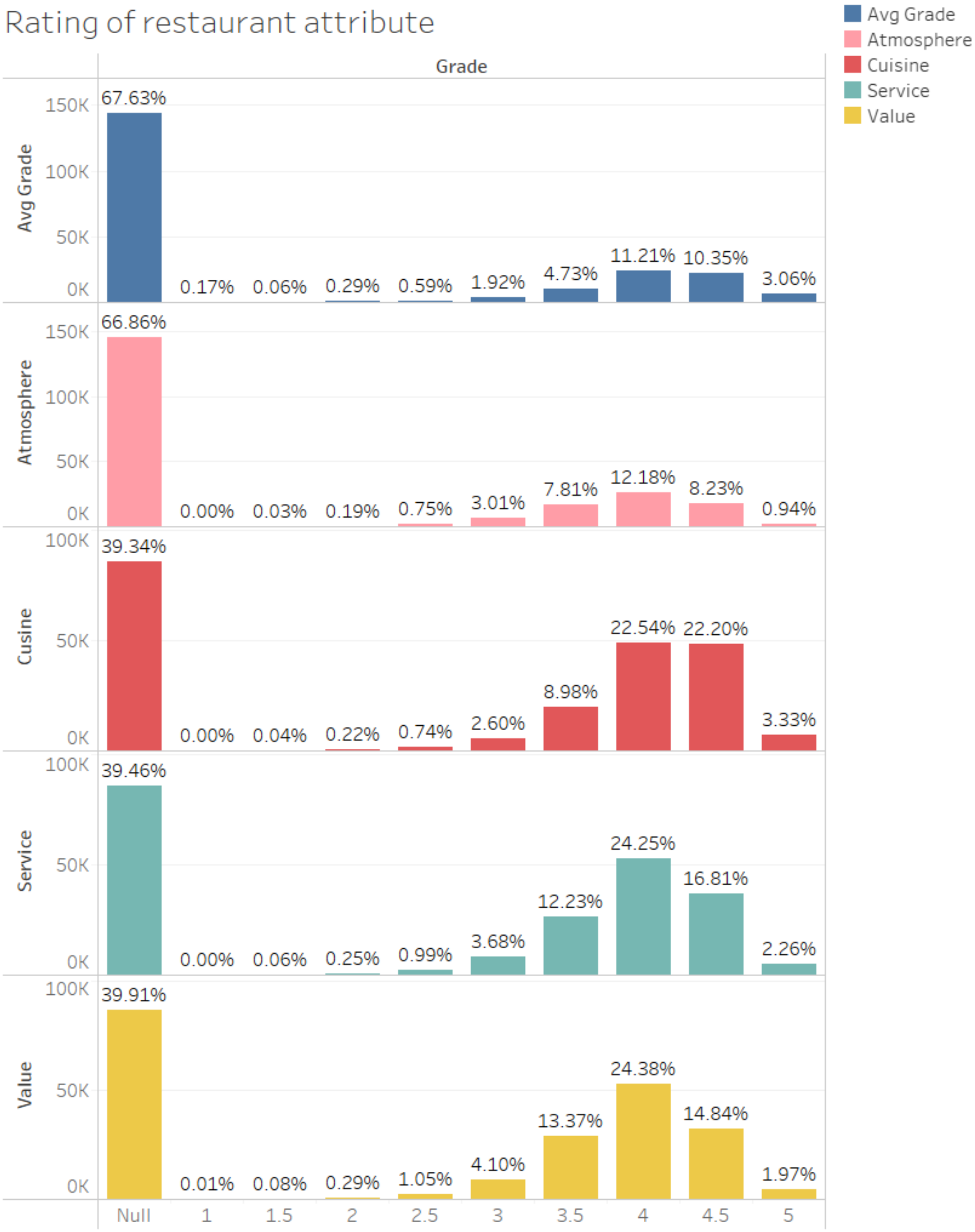


Figure 2.12: POIs distribution by region

2.2.3 Categories by POI type

This section will describe the categories that are present in the dataset regarding attractions, restaurants and hotels.

Attractions

Figure 2.13 depicts the distribution of attraction types in the dataset. In the top six positions, a variety of attraction types, that can be embedded in the artistic and cultural type of tourism, stand out. This is consistent with the fact that Italy has one of the world's largest artistic cultural heritages [4].
After these cultural attractions, beaches follows and this is consistent with the fact that Italy has a significant preference for beach and lake tourism.

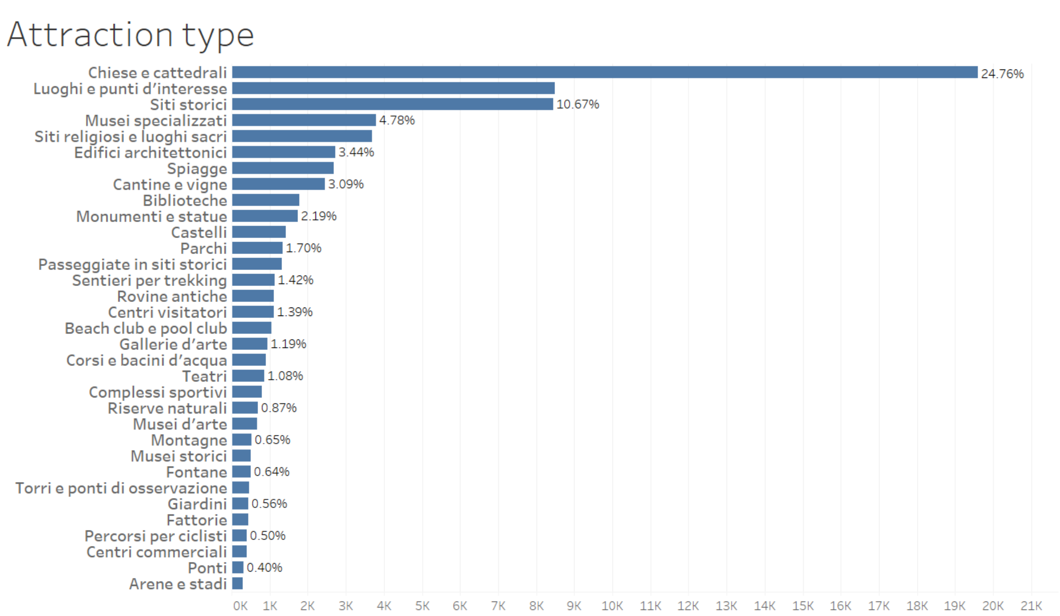


Figure 2.13: Attraction type distribution

Hotels

Figure 2.14 immediately put in evidence that there is a significant lack of data for this field, equal to 28% of the total, in terms of the distribution of the different types of hotels. It is interesting to note that the majority of the hotels in this data sample that have a label, have the *Family* tag attached to them. Another intriguing aspect is that the hotels classified as being suitable for business come in at 3.8%. On the other hand, hotels classified for being appropriate for luxury make up only 0.2% of the whole sample.

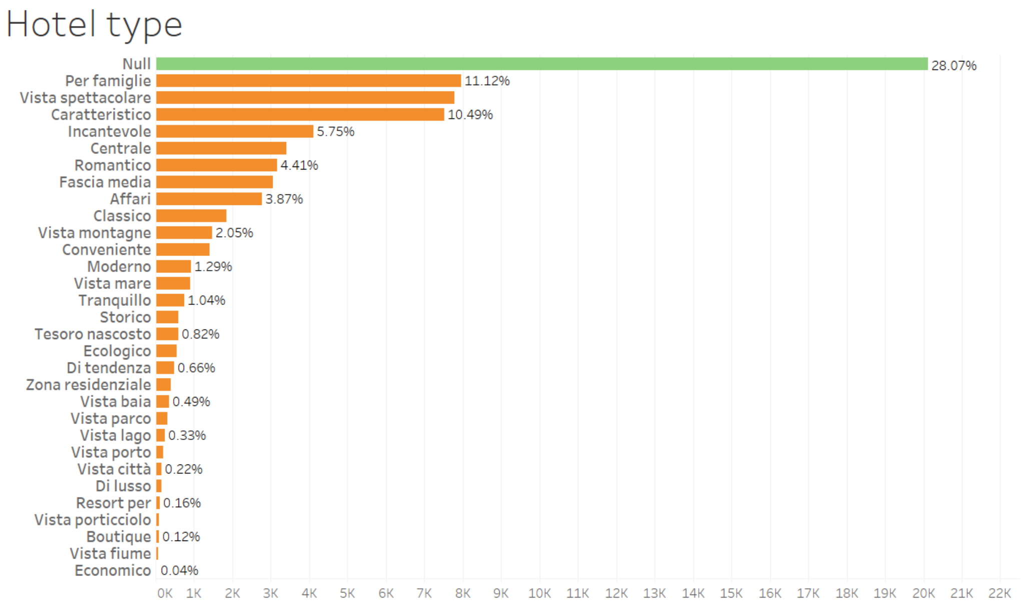


Figure 2.14: Hotel type distribution

Restaurants

In Figure 2.15, an overwhelming majority of restaurants have tags that can be traced back to the Italian culinary tradition. This is in keeping with the actual distribution of the type of restaurants in Italy. As is common knowledge, Italian cuisine is one of the most well-known and well-liked cuisines in the entire world, and as a result, it is widely represented in its country of origin.
An additional noteworthy thing is the fact that even for this type of POI there is a relevant lack of data, amounting to 9%

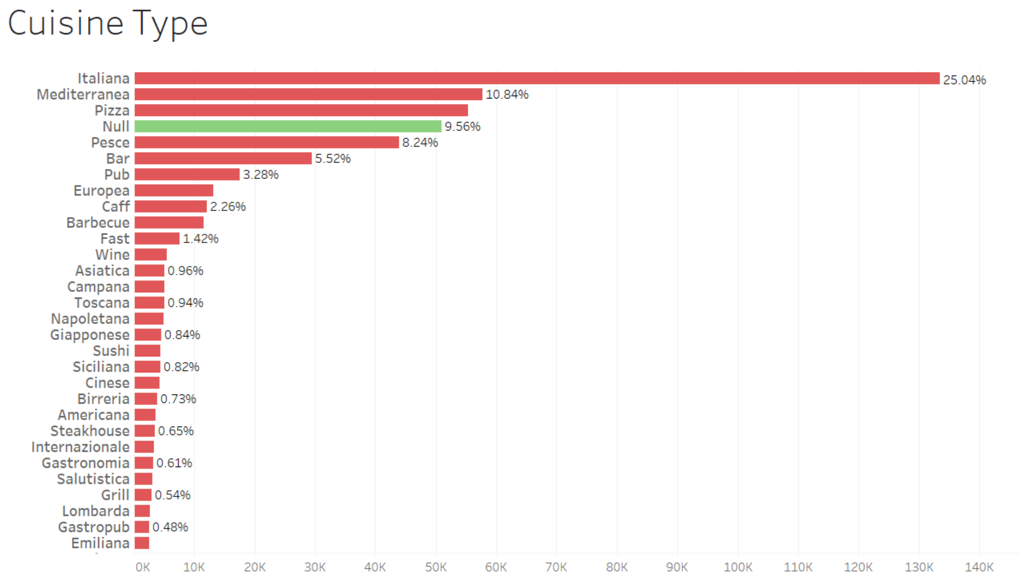


Figure 2.15: Restaurant cuisine distribution

2.2.4 Reviewer analysis

This section examines the behavior of reviewers in terms of the number of reviews they leave, the frequency of their visits, and the interval between reviews.
Concerning Figure 2.16, there is a tendency with peaks for the visitors and reviews portion; these peaks come in August, a time when many people are on vacation. Regarding the reviews - visits section, it can be seen that the number of experiences is significantly higher than the number of reviews left in the peaks, and that this trend reverses in the month following the peak, indicating that many reviews for experiences/visits that occurred in August are disposed of in September. The chart also shows the expected impact of COVID-19, with a clear decline in the number of visits and reviews beginning in 2020.

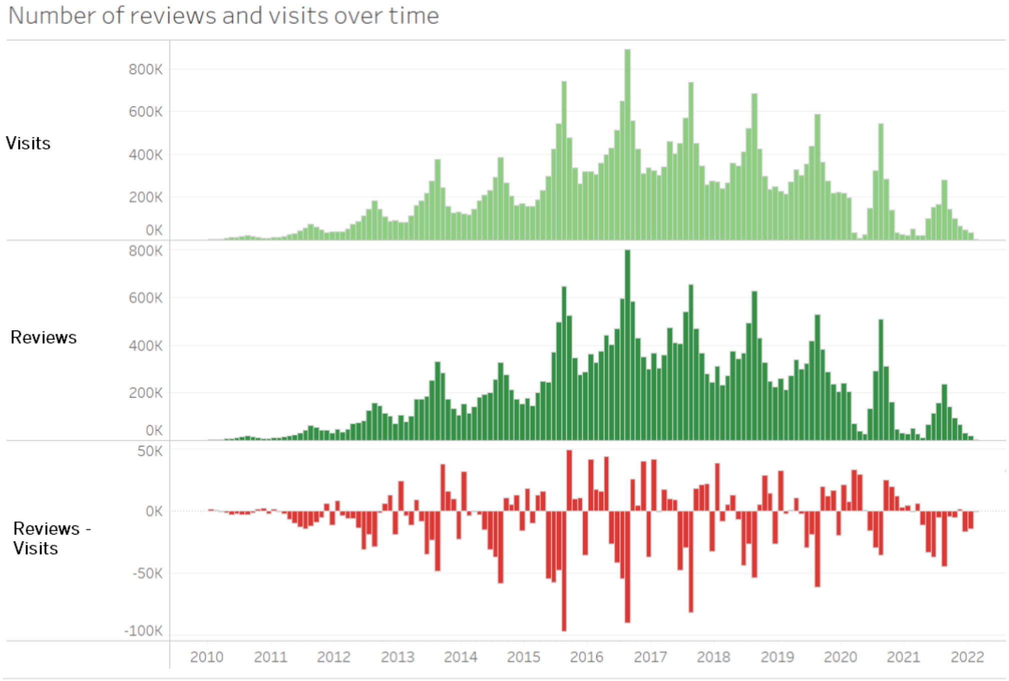


Figure 2.16: Reviews and visits distribution

Figure 2.17 focuses on evaluations left in Veneto, Trentino Alto Adige and Friuli Venezia Giulia. These three regions were taken into account not only due to the fact that they are of particular importance for Motion Analytica but also due to the fact that they have adequate representation within the data set.
More specifically, the amount of time that passed between a client's visit and their review was analyzed, and the results showed that the majority of reviews were posted within the first 30 days after the event. These graphs reflect a behaviour similar to the power law.

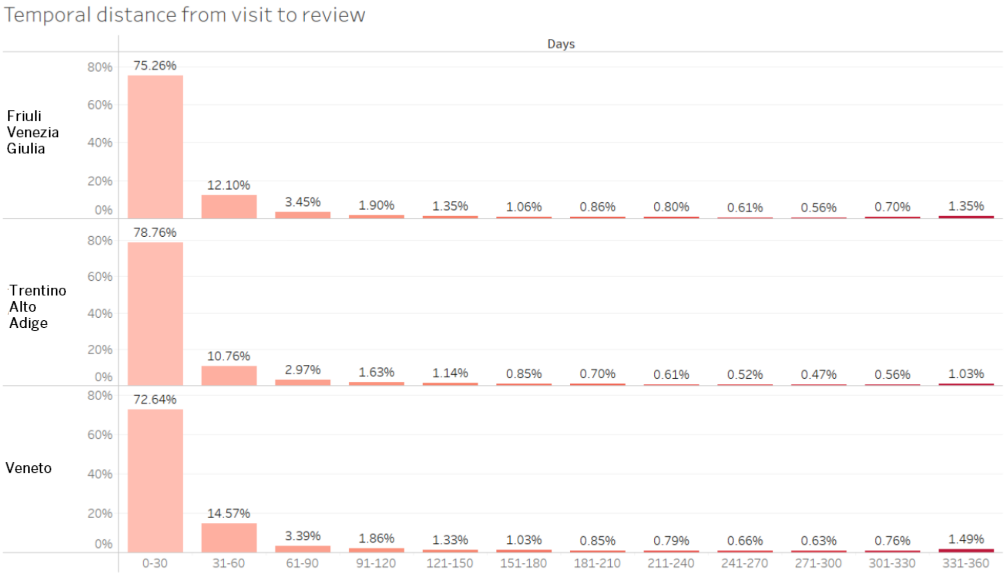


Figure 2.17: Temporal distance visit to review in Triveneto

Figure 2.18 examines the path length of reviewers, which refers to the number of reviews that reviewers leave; for visualization purposes, the bins contain 10 different path lengths; for example, the first bin contains the number of reviewers who left between 1 and 10 reviews. Once more, power law behavior is evident here.

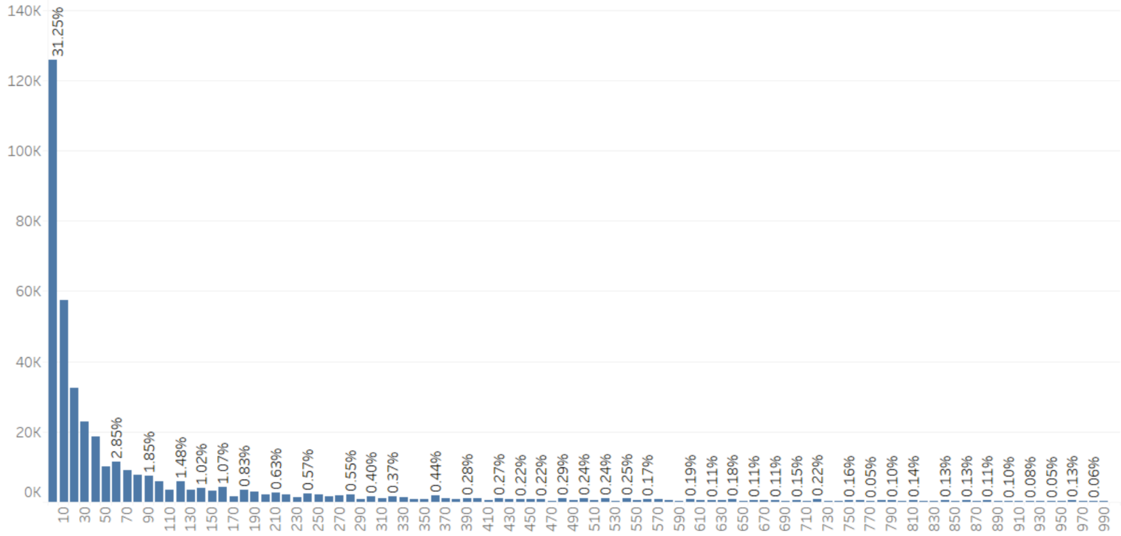


Figure 2.18: Reviewer path length distribution

The last thing that has been explored is the relationship between path length and time span, where path length refers to the number of reviews a Tripadvisor user provides and time span refers to the number of days between the user's first and last review. This is displayed in Figure 2.19. In the left tail, it is possible to observe a phenomenon partially described in Figure 2.18, namely that there are a large number of reviewers in the dataset who have left few reviews. However, in the right tail, a very interesting phenomenon is revealed: if the number of days between the first and last review is greater than 1000, there is a greater variability in the revierwer path length.

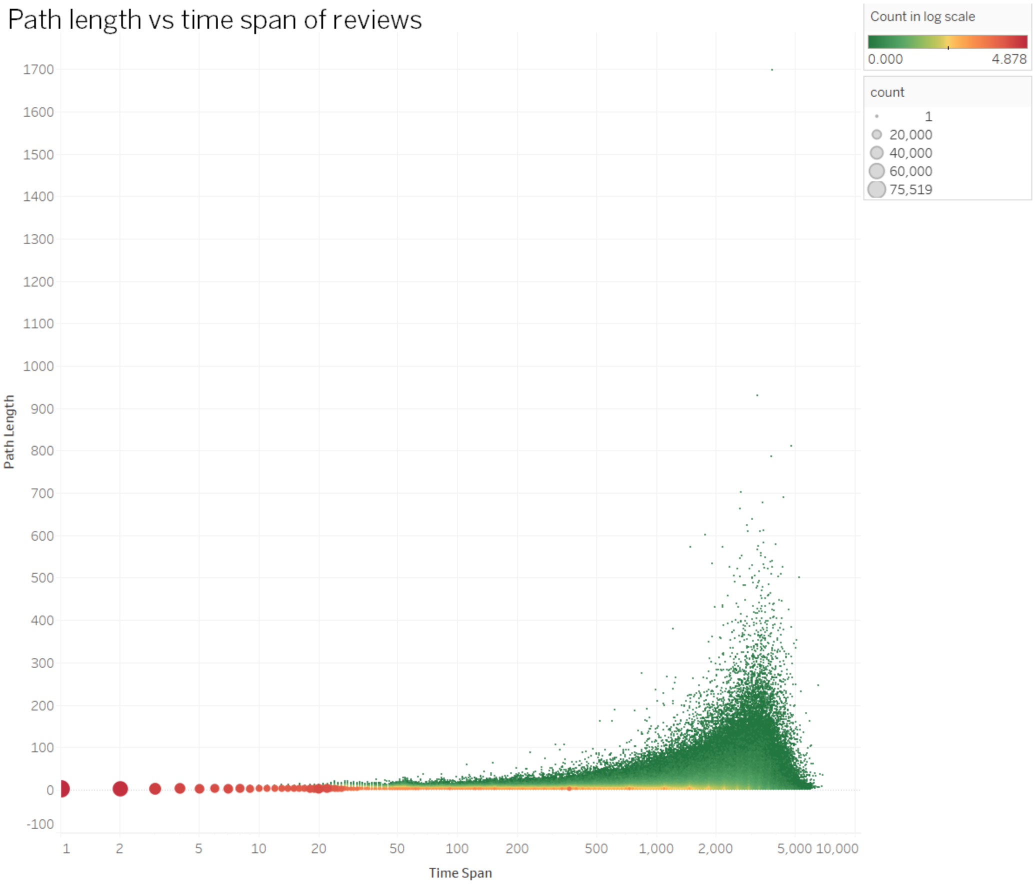


Figure 2.19: Distribution of path length and time span combined

The conclusion drawn from EDA is that the status of the data is good. One thing to mention is that there is a significant lack of data regarding POI ratings. For this reason, the development of this work is focused on what type of POIs the reviewers visit and not what opinion they have of them.

Chapter 3

Data Transformation

The data transformation procedure is also known as extract/transform/load (ETL). The extraction phase involves identifying and extracting data from the many data-generating source systems, followed by transferring the data to a central repository. If necessary, the raw data is then cleaned. The data is then converted into a format that can be fed into operational systems, a data warehouse, a data lake, or another repository for business intelligence and analytics applications. The transformation could involve switching data formats, eliminating duplicate data, and improving the source data. In this section, we will discuss the primary processes associated with data preparation and demonstrate the format of the final data that will be used by the scoring alogorithm.

3.1 Origin Cleaning

One of the most important activities that needed to be completed in relation to the preparation of the data was the cleansing of the *Provenienza* field of the reviewer, which is a field that identifies where the reviewer comes from. From now on we will refer to *Provenienza* with *Origin*. For the company's business goals to know the origin of the reviewers is essential. One problem identified during the exploratory analysis is that *Origin* field is poorly constructed, and reviewers frequently misspell their origin and/or use a different language from English upon registration. The solution that has been devised is to create a table containing two fields. The first contains the malformed *Origin* field. The second field contains the corrected and standardized string for the *Origin*. The purpose of this table is that of a loopback. Basically, this table is used for reviewers to have a standardized *Origin* field. Using this constitutes one of the first steps in the ETL process. An example of possible table entries are illustrated in Table 3.1.

Input	**Output**
Mosca, Russia	Russian Federation
Moscov	Russian Federation
Russia	Russian Federation
Москва, Россия	Russian Federation

Table 3.1: Representation of correspondence table

The ETL module that deals with the creation and use of this table is called the Origin Cleaning Module (OCM), this is covered in detail in Chapter 5.

3.2 Time Origin Destination Matrix

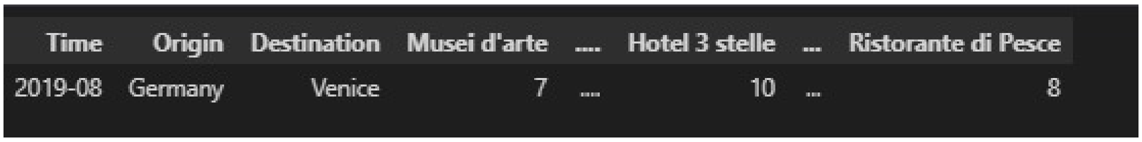
The line of thinking that went into the creation of this particular arrangement of the data will be discussed before moving on to the description of the final refined data.
From the data exploration phase, it is evident that the majority of reviewers have an extremely short review path, and in the cases where this is long, it turns out to be extremely inconsistent. As a consequence, it is difficult to determine the various types of reviewers, the attention is shifted from the reviewers to the locations where the reviews are conducted. The location, which in the context of this study is the province, contains a series of reviews left by tourists. Each review can be seen as a trace left in a particular province. The aggregate of these traces provides a kind of "identity card" for the province. In other words, it indicates what types of POIs tourists visit in a given province.
In addition to that, if the information of time along with the information of nationality is added, it is possible to identify a great deal of different behavior with respect to a given period of time and with respect to a particular nationality, in a given location. This is possible because different people have different ways of behaving in different places and different times.
In light of what has been discussed, the final form of the data that will be the input to the scoring method has been called *Time Origin Destination matrix.* From this point forward, we shall refer to this as the TOD matrix. The name of this matrix suggests that it is constructed with regard to three primary dimensions:

- **Time**: Refers to a particular month, and it is expressed in this format mm-yyyy, In light of the exploratory research, the period under consideration extends from 2012 to the end of 2021, as this span has a high data density.

- **Origin**: Refers to the origin country of the reviewer, expressed with ISO 3166 [19], This is the field that is recovered by the OCM described in subsection 3.1.

- **Destination**: Refers to the Italian province, expressed in Italian language.

Besides these primary dimensions a long list of other dimensions that cover all possible POI types and characteristics is used. Each of these additional dimensions allows for pointing out the number of reviews dealing with a particular POI type in the given Time (month), Origin (Nationality reviewers), and Destination (Italian Province).
A high level tabular representation of the final data is the following:



Time	Origin	Destination	Musei d'arte	....	Hotel 3 stelle	...	Ristorante di Pesce
2019-08	Germany	Venice	7	....	10	...	8

Figure 3.1: Snippet TOD matrix

The following dashboard was built by using Tableau, a data visualization software for organizations working with business information analytics. Moreover it provides a clearer and more graphically appealing representation of the data that was stored in the TOD matrix. An initial prototype dashboard, using data in the form of a TOD matrix, is shown in Figure 3.2. In this dashboard, an in-depth look at the Triveneto is shown:

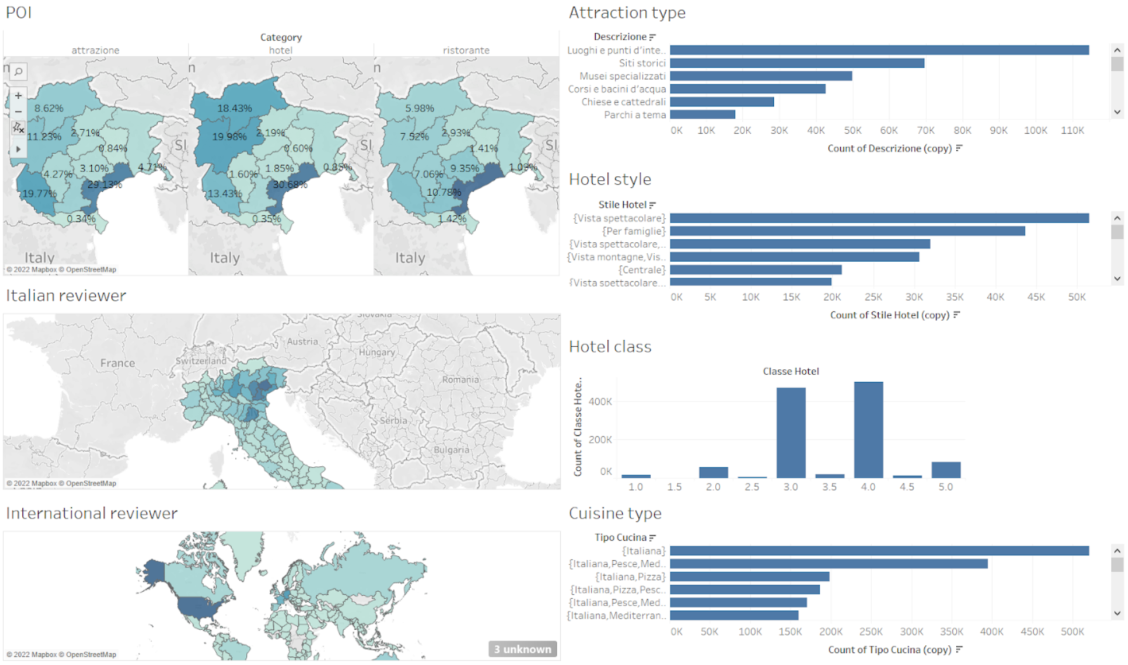


Figure 3.2: Dashboard TOD matrix focused on Triveneto

In the following all the dimensions of the TOD matrix are listed in English but as the TOD snippet shown in Figure 3.1, labels in practice are in Italian for business needs. For clarity of exposition in some cases these dimensions are grouped into classes.
The dimensions related to the Attraction type POI are:

- **Outdoor activity**: Sports camps and courses, Boat tours, Watersports, Boat rentals, Dolphin and whale watching, Gondola rides, Speedboat tours, Kayaking and canoeing Parasailing and paragliding, River rafting and tubing, Diving and snorkeling, Shark diving, Submarine tour, Surfing, windsurfing and kitesurfing, Swimming with dolphins, Water skiing and jet skis, Tour of the ducks, Fishing charters and tours, Stand up paddle, 4 × 4 and off-road tours, Adrenaline and extreme tours, Hot air balloon accommodation, Bike ride, Climbing tour, Eco tours, Hiking and camping, Ride in nature, Running lap, Air tours Descents from the suspended ropes and adventure parks with aerial routes, Horse-drawn carriage tours, Canyoning and abseiling tour, Zoo, Beaches, Ski and snowboard slope, Routes for cyclists, Horseback riding, Hiking trails, Routes for motorcyclists, Off-road routes, Equipment rental, Safari, Ski and snow tours, Other outdoor activities, Beach club and pool club, Golf courses, Horse ride, Paths for jogging and walking, Cross-country ski areas, Scenic roads

- **Gambling**: Casinò, Cynodrome, Hippodrome.

- **Food and beverages**: Wine shops, Farmers' markets, Cooking courses, Cellars and vineyards, Breweries, Distilleries, Other food and drink, Brewery tour, Coffee and tea tour, Food tours, Wine tours and tastings, Distillery tour

- **Concerts and shows**: Theaters and shows, Ballets, Concerts, Opera, Symphonies,Theaters, Dinner theaters, Cabaret, Cirque du Soleil shows, Luau Courses and seminars: Cooking courses, Sports camps and courses, Lectures and seminars, Studies of painting and ceramics

- **Fun and games**: Comedy club, Playgrounds,Studies of painting and ceramics, Game and entertainment centers, Bowling, Sports complexes, Room for escape games, Treasure hunts, Shooting ranges, Other entertainment and games, Wedding chapels, Mini-golf, Rides and activities, Experiences with characters, Cinema

- **Amusement parks**: Water parks, Disney parks and activities, Theme parks

- **Parks and nature**: Aquariums, Playgrounds, Dams, Streams and water, basins, Coral reefs, Islands, Other parks and nature reserves, Canyon, Caverns and caves, Deserts, Forests, Gardens, Geological formations, Hot springs and geysers, Marinas, Mountains, National parks, Natural reserves, Parks, State Parks, Valleys, Volcanoes, Waterfalls.

- **Traveller resources**: Libraries, Lounge at the airport, Visitor Centers, Convention and conference centers.

- **Shopping**: Farmers' markets, Antique shops, Department stores, Outlet, Markets and flea markets, Shopping centers, Specialty and gift shops, Shops at the airport, Tours and fashion shows, Shopping tour.

- **Tours**: Sightseeing tours, Factory visits, Bus tour, Segway tour, Walking tour, Archaeological tours, City tour, Cultural tours, Ghost and vampire tours, Helicopter tour, Historical and heritage tours, Hop on and off tour, Literary, artistic and musical tour, Film and TV tours, Night tour, Private tours, Tour by train, Photo tours, Day trips, Scenic Railways

- **Transports**: Mass Transportation Systems, Taxis and shuttles, Transportation by bus, Railways, Tram, Ferries.

- **Night life**: Wine shops, Tour of bars-clubs-pubs, Comedy club, Blues bars and clubs, Country clubs, Jazz club and bar, Piano bar, Cigar club, Gay bars and clubs, Karaoke, Discos and dance halls, Cafeterias, Bars and clubs

- **Museums**: Art gallery, Sector museums, Art museums, Children's museums, Historical museums, Military museums, Natural history museums, Science museums, Observatories and planetariums

- **Spa and wellness**: Spa, Health / fitness centers and gyms, Onsen spas, Arab baths, Hammam and Turkish baths, Roman thermal baths, Thermal baths, Yoga and pilates.

- **Sites of interest**: Neighborhoods, Scenic roads, Religious sites and sacred places, Educational places, Arenas and stadiums, Farms, Mines, Car race track, Ranch, Fountains, Ghost city, Government buildings, Places and points of interest, Belvedere, Historical sites, Churches and cathedrals, Walks in historical sites, Ancient ruins, Headlights, Military bases and centers, Mysterious sites, Monuments and statues, Other ranches and farms, Universities and schools, Architectural buildings, Observation Decks and Towers, Battlefields, Bridges, Quays and walkways over the sea, Castles, Cemeteries, Panoramic pedestrian paths, Civic Centers, Ships.

The dimensions related to the Hotel type POI are:

- **Type**: Family friendly Spectacular view, Characteristic, Enchanting, Central, Romantic, Mid-range, Business, Classic, Mountain view, Affordable, Modern, Sea view, Quiet, Historic, Hidden Treasure, Ecological, Trendy, Residential area, Bay view, Park view, Lake view, Harbor view, City view, Luxury, Resort, Marina view, Boutique, River view, Cheap, Detail, Lagoon view, Art Deco

- **Stars**: 1 Star, 1.5 Star, 2 Stars, 2.5 Stars, 3 Stars, 3.5 Stars, 4 Star,s 4.5 Stars, 5 Stars

The dimensions related to the Restaurant type POI are:

- **Type**: Italian Mediterranean Pizza, Fish, Bar, Pub, European, Caffe, Barbecue, Fast, Wine, Asian, Bell, Tuscan, Neapolitan, Japanese, Sushi, Sicilian, Chinese, Brewery, Americana, Steakhouse, International, Gastronomy, Healthy, Grill, Lombard, Gastropub, Emilian, Ligurian, Contemporary, Sardinian, Latian, Soups, Roman, Austrian, German, Middle Eastern, Central, Mexican, Turkish, South American, Apulian, Spanish, Indian, Breweries, French, Greek, Calabrian, Brazilian, Thai, Argentinian, Latin, Hawaiian, Romanian, Arab, British, Irish, African, Moroccan, Peruvian, Egyptian,

The data are transformed into their ultimate format, which is the TOD matrix, after passing through the entirety of the ETL process. Data in this shape tallies the total number of reviews given for all POI categories, fixed a specified month (Time), a certain nationality for reviewers (Origin), and a specific region of Italy (Destination). This organization of date as illustrated in Figure 3.2, offers a reasonable overview of the data set. Rearranging the data in the TOD form is not sufficient for the detection of nontrivial information but is necessary for the application of the data mining algorithm. Application of the algorithm will bring up the nontrivial information. In order to meet the goal of this work, which is to extract new knowledge and in particular to identify niche behaviors/trends of particular types of reviewers, a method from the field of Information Retrieval was used, which is explained in the following section.

3.3 Term Frequency–Inverse Document Frequency

In the process of retrieving the significant information from the TOD matrix, the tool known as Term Frequency–Inverse Document Frequency (TF-IDF) from the field of information retrieval was utilized.

TF-IDF, is a statistical metric that determines how pertinent a given word is to a certain document within a collection of documents.

It has a wide variety of applications, the most significant of which is in automated text analysis, and it is highly helpful for scoring words in machine learning algorithms used for natural language processing (NLP).

It functions by growing proportionately with the frequency of a term in a document, but is offset by the number of documents that include the word. Therefore, common terms in every text are ranked poorly, even though they appear often, because they have little relevance and they do not provide a sufficient valuable information. On the other hand, if the same term appears several times in one document but not in others, this is often an indication that the information it contains is particularly pertinent [15].

The TF-IDF score of a word in a document is determined by multiplying the scores of two distinct metrics:

- **Term Frequency (TF)**: There are a few different approaches to computing this frequency, with the most straightforward one being a simple count of the number of times a word appears in a given document. Then, there are methods for adjusting the frequency, either by the whole length of a text or by the frequency of the word that appears the most frequently in a document.

- **Inverse Document Frequency (IDF)**: This indicates how frequent or uncommon a term is throughout the full set of documents. A number approaching zero in this metric indicates that the word is widely used in the document set. To determine this metric, take the total number of documents, divide that number by the number of documents that include a word, and then compute the logarithm of the result.

The TF-IDF score of a word in a given text may be calculated by multiplying these two numbers together. When the score is greater, it indicates that the term in question is more pertinent to the content of the given text.

To put it in more precise mathematical terms:

Definition 1.

$$TF(t, d) = \log(1 + \textit{ freq } (t, d))$$

Definition 2.

$$IDF(t, D) = \log\left(\frac{N}{\text{count}(d \in D : t \in d)}\right)$$

The letter t in the formula above denotes a word; the letter d denotes a specific document; and lastly, the letter D denotes the dataset that contains the set of documents.

The final formula is:

Definition 3.

$$TF - IDF(t, d, D) = TF(t, d)IDF(t, D)$$

3.4 TF-IDF adapted for TOD matrix

As discussed in the preceding section, TF-IDF was initially developed for the purpose of document research; however, it is also widely used for keyword extraction from documents. At first glance, this method appears completely unrelated to our data. The TOD matrix stores only numeric values, whereas the traditional formulation of TF-IDF is concerned with words, document retrieval, and similar topics.

This section will describe the repurposing of the TF-IDF concept for data in TOD matrix format. To best expose the idea's readjustment, parts of the formula will be taken in seperate manners, and for each of them the readjustment implemented for the TOD matrix use case is explained.

Starting with TF:

$$TF(t,d) = \log(1 + \text{ freq } (t,d))$$

The scope of TF is a single document d, in the TOD matrix context what is referred to as a "document" is a row belonging to the TOD matrix.

Whereas the term t assumes the role of one of the POI features belonging to the TOD matrix.

Finally, the function freq (t,d) in this context returns the value of row d at column t. This final component represents the number of reviews that took place for a specific category of POI during a specific month of the year (**T**ime), with a specific nationality (**O**rigin), and a specific province (**D**estination).

Proceding with IDF:

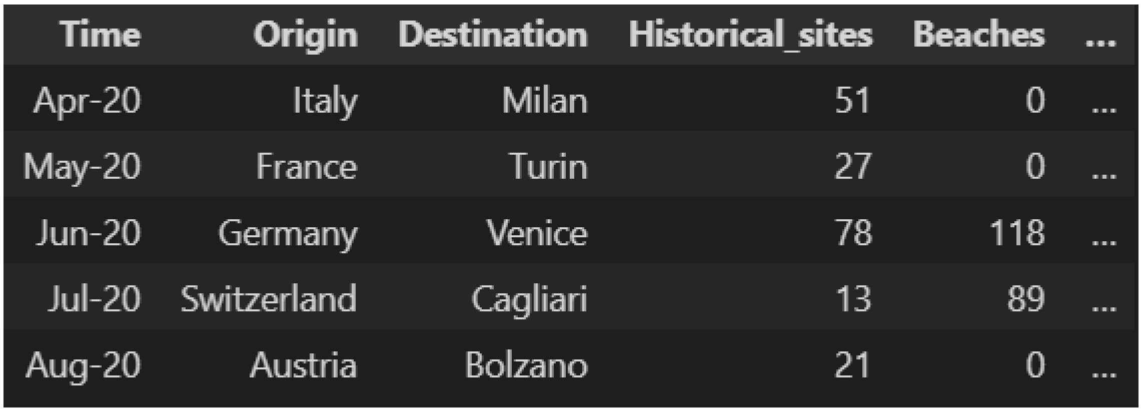
$$IDF(t,D) = \log\left(\frac{N}{\text{count}(d \in D : t \in d)}\right)$$

As in the classical formulation the scope for calculating this value extends to all the set of documents, making a parallelism with respect to our context, D refers to all the rows of the TOD matrix. Thus the number N is equal to the number of rows the TOD matrix possesses. Slightly more complicated is the expression of the denominator: It is the amount of rows in the TOD matrix, such that the number of reviews for a given type-feature of POI is greater than 0. Mathematically:

$$\text{count}(d \in D : t \in d) = \sum_{i=1}^{N} [\![\text{ freq } (t, d_i) > 0]\!]$$

3.4.1 An example of TF-IDF

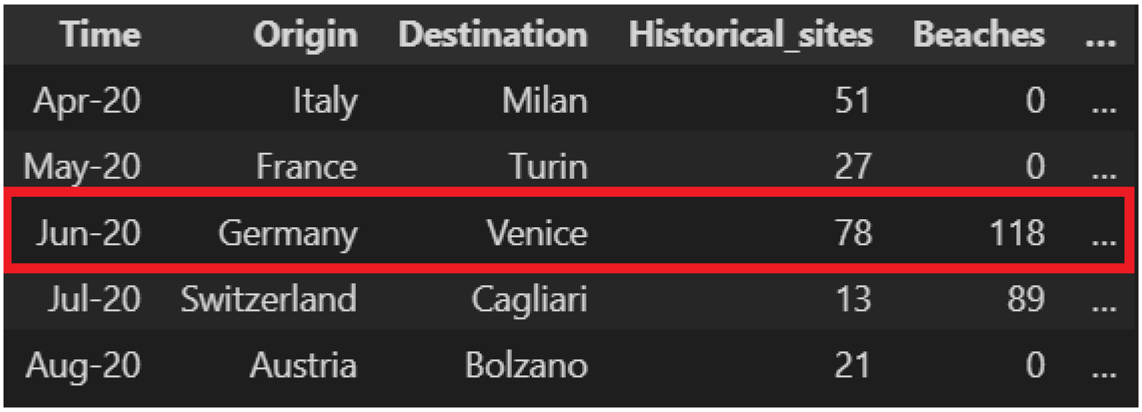
In order to clarify doubts the concept of TF-IDF adapted for TOD matrix, a concrete computation example is provided below.
Consider the TOD matrix shown in Figure 5.2. It contains 5 elements. Thus N in the IDF formula is 5.



Time	Origin	Destination	Historical_sites	Beaches	...
Apr-20	Italy	Milan	51	0	...
May-20	France	Turin	27	0	...
Jun-20	Germany	Venice	78	118	...
Jul-20	Switzerland	Cagliari	13	89	...
Aug-20	Austria	Bolzano	21	0	...

Figure 3.3: TOD matrix TF-IDF computation step 1

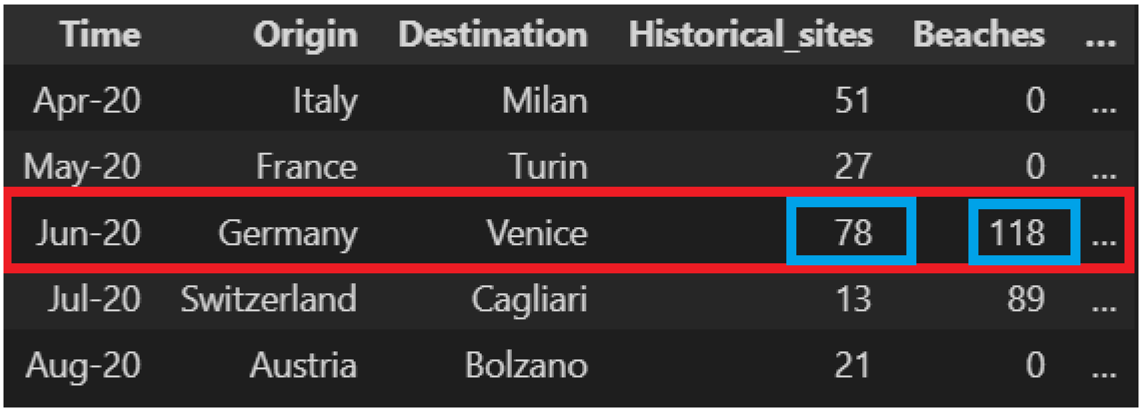
TF-IDF is calculated for the attributes that are part of the POI attraction categories, in this case example *Historical_ Sites* and *Beaches*, for the row having Time = Jun-2020, Origin = Germany, Destination = Venice. It is worth noticing that the combination of Time, Origin and Destination identifies in a unique way a row.



Time	Origin	Destination	Historical_sites	Beaches	...
Apr-20	Italy	Milan	51	0	...
May-20	France	Turin	27	0	...
Jun-20	Germany	Venice	78	118	...
Jul-20	Switzerland	Cagliari	13	89	...
Aug-20	Austria	Bolzano	21	0	...

Figure 3.4: TOD matrix TF-IDF computation step 2

The detection of the elements that are used for TF are the simplest.
Let $d =<$ Jun-2020,Germany,Venice $>$, t_1 = Historical sites and t_2 = Beaches, we have freq $(t_1, d) = 78$ freq $(t_2, d) = 118$ corresponding to the two values in the corresponding columns.

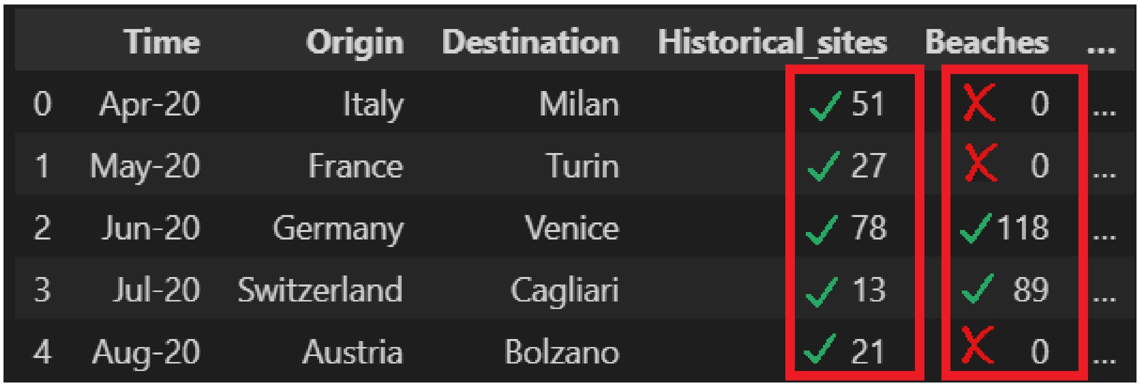


Time	Origin	Destination	Historical_sites	Beaches	...
Apr-20	Italy	Milan	51	0	...
May-20	France	Turin	27	0	...
Jun-20	Germany	Venice	78	118	...
Jul-20	Switzerland	Cagliari	13	89	...
Aug-20	Austria	Bolzano	21	0	...

Figure 3.5: TOD matrix TF-IDF computation step 3

At this point, we have all the necessary components for the TF calculation. Recall that the components necessary to calculate the IDF are N, which, as stated at the beginning of the example, $N = 5$, i.e. the total number of documents belonging to the set.

Regarding count$(d \in D : t \in d)$, for count$(d \in D : t_1 \in d) = 5$, because all rows corresponding to the POI feature Historical_site are greater than 0, instead for Beaches count$(d \in D : t_2 \in d) = 2$.



	Time	Origin	Destination	Historical_sites	Beaches	...
0	Apr-20	Italy	Milan	✓ 51	✗ 0	...
1	May-20	France	Turin	✓ 27	✗ 0	...
2	Jun-20	Germany	Venice	✓ 78	✓118	...
3	Jul-20	Switzerland	Cagliari	✓ 13	✓ 89	...
4	Aug-20	Austria	Bolzano	✓ 21	✗ 0	...

Figure 3.6: TOD matrix TF-IDF computation step 4

At this point, we have all the information necessary to complete the TF-IDF calculation:

$$\begin{aligned}
TF(t_1, d) &= \log(1 + \text{ freq } (t_1, d)) = \log(1 + 78) = 1.89 \\
TF(t_2, d) &= \log(1 + \text{ freq } (t_2, d)) = \log(1 + 118) = 2.07 \\
IDF(t_1, D) &= \log\left(\frac{N}{\text{count}(d \in D : t_1 \in d)}\right) \\
&= \log\left(\frac{5}{\sum_{i=1}^{N} [\![\text{ freq } (t_1, d_i) > 0]\!]}\right) \\
&= \log\left(\frac{5}{5}\right) = 0 \\
IDF(t_2, D) &= \log\left(\frac{N}{\text{count}(d \in D : t_2 \in d)}\right) \\
&= \log\left(\frac{5}{\sum_{i=1}^{N} [\![\text{ freq } (t_2, d_i) > 0]\!]}\right) \\
&= \log\left(\frac{5}{2}\right) = 1.39 \\
TF - IDF(t_1, d, D) &= TF(t_1, d) \cdot IDF(t_1, D) = 1.89 \cdot 0 = 0 \\
TF - IDF(t_2, d, D) &= TF(t_2, d) \cdot IDF(t_2, D) = 2.07 \cdot 1.39 = 2.87
\end{aligned}$$

This example clarifies the idea behind the TF-IDF concept. The TF-IDF for *Historical sites* is 0 because this category appears in all rows and it is likely a piece of information already known. It is important to underline that this is a very rare and extreme case. For *Beaches*, the story is different: In this case, the TF-IDF value is different from 0 and in particular is slightly higher than that provided by TF. This indicates that beaches are a type of POI that is not common within the dataset and may contain nontrivial information.

The use of TF-IDF brings out types of POIs that are not frequent in the dataset but are still reviewed. This forms the key to the identification of tourists' niche behaviors.

Chapter 4

Data Visualization & Analysis

Data into a form that is easier to grasp and showing trends and outliers is known as data visualization. This process is helpful in telling stories. A strong visualization conveys a narrative while minimizing distracting noise in the underlying data and emphasizing important details.

However, it is not as simple as just adding some decorations to a graph in order to make it appear better. A successful data visualization requires striking a precise balance between aesthetics and practicality. It is possible that the most basic graph may be so uninteresting that no one would notice it, or that the most striking depiction could completely fail at delivering the point you were trying to make. Either way, the results could be quite telling. The data and the images need to complement one another, and there is a certain level of skill required to combine excellent analysis with compelling narrative.

The creation of data visualizations is an additional really crucial aspect of this work. The provision of a graphical context, such as that offered by maps and graphs, by means of data visualizations, provides us with a deep comprehension of the significance of the information. This makes the data more natural for the human mind to interpret, and as a result, it makes it easier to discover trends, patterns, and irregularities within enormous data sets.

The data provided by the TOD matrix is easy to understand for a technical eye, but it is not simple to comprehend for a stakeholder such as policy makers and destination managers, so the data visualization aspect aims to target this second type of people. Moreover, since this work aims to be a saleable product for the company, the data visualization component gains further importance.

This chapter deals with the design of dashboards and some use cases of these.

For this purpose is used Tableu, a data visualization tool that provides graphical and pictorial representations of data.

4.1 Dashboard Design

This section describes the final dashboard's design in Figure 4.1. For the sake of illustration, the dashboard holding the TOD matrix and score associated with the *Attraction* type is depicted. However, it is important to note that there is a dashboard for each type of POI.

In the upper left corner there is an interactive map of the world with the title Origin. This is a geographic heat map, a style of visualization that depicts the geographic distribution of data through the use of colors and shading. Typically, heat maps are used to depict data where the geographic placement of items is significant. In this instance, it represents the Origin dimension of the TOD matrix, which displays the distribution of reviewers' nationalities around the globe.

In the centre, there is an interactive geographic heat map of Italy, divided by Italian provinces, which displays through varying shades of red the number of reviews with respect to the attribute *Attraction* type.

To conclude in the upper part of the figure there is a treemaps chart, which is a visual approach for representing hierarchical data using nested rectangles to represent a tree diagram's branches. In addition to each rectangle's area being proportional to the quantity of data it represents, there is also a color type component that takes on various shades of blue in order to remark the proportionality of the amount of data for each attribute.

In this case, it was chosen to show the IDF among the top 20 attributes with the largest TF-IDF, in case there are more than 20 attributes. Recall that IDF represents the rarity of reviews for a certain attribute compared to the entire dataset of a given category, in this specific case the type of attractions. This gives the dashboard viewer a rough idea of what attribute might be a candidate to represent a niche behavior of reviewers.

The final plot at the bottom of the dashboard is a line graph, also known as a line plot or line chart, which connects individual data points using lines. A line graph represents quantitative values over a defined range. In this instance, the first one displays the median value of TF-IDF for each attraction type. This value spans a time interval separated into months, beginning in January 2012 and ending in December 2021 if data is available. Each different attribute regarding to this category is identified with a different color. The second shows the median value of TF, following the same way as the TF-IDF line plot display. These two pieces of information together allow the detection of niche behaviors. As is shown in the next section. Lastly on the far right of the dashboard there are the legends, the first two on the top right refer to the Origin and Destination views, while the last one refers to the attributes displayed in the TF-IDF Score line plot. The latter is interactive and provides the ability to select attributes as desired for display.

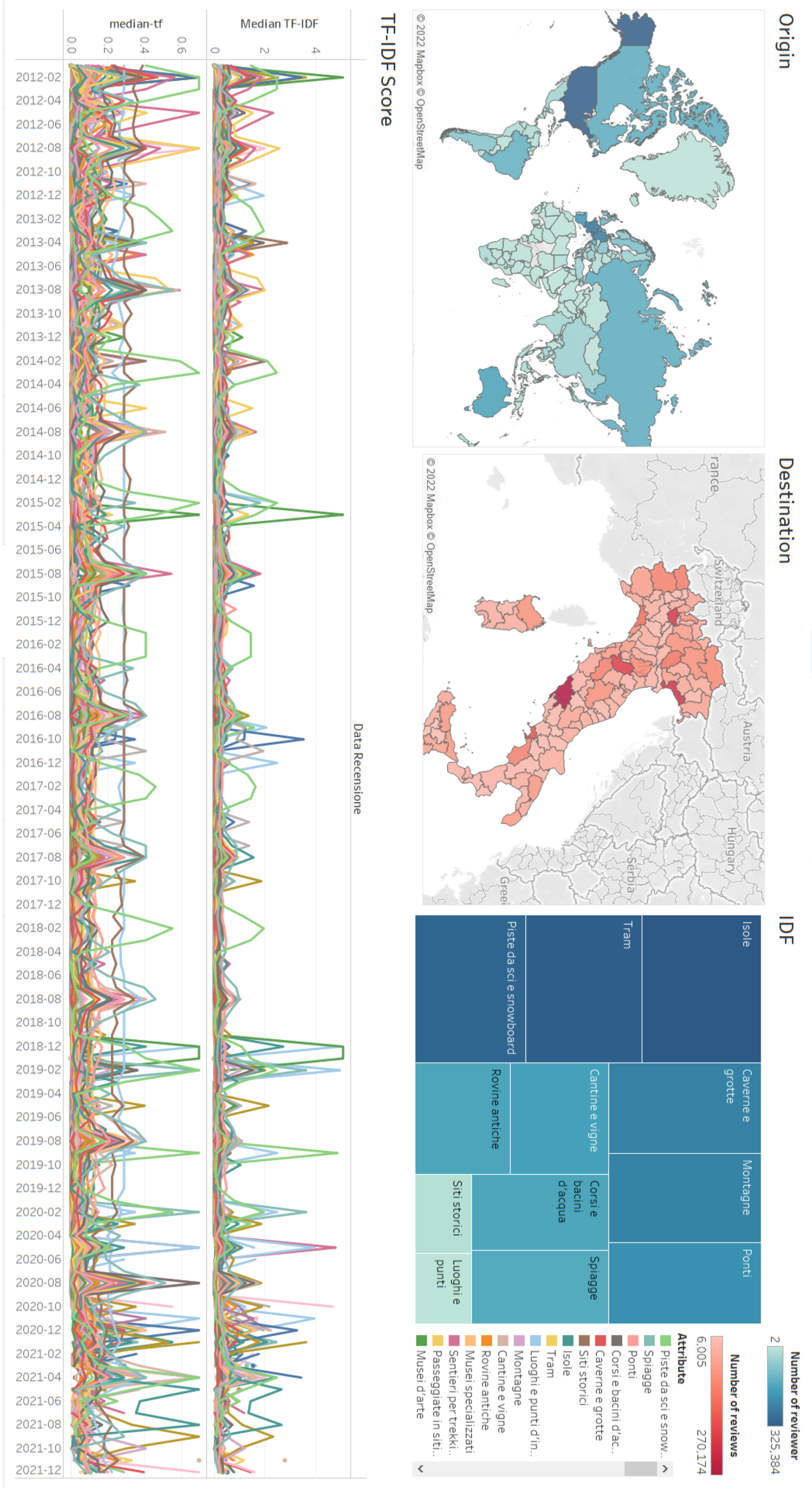


Figure 4.1: Dashboard design

4.2 Extraction of Niche Behavior

In this section, the niche behavior of a type of tourists will be represented. Niche behavior is defined as an attribute that has a fair number of reviews and also is characterized by a relevant IDF value, this combination of values causes this attribute to stand out from the others in the TF-IDF graph and to be somewhat similar to the attribute that is characterized by a large number of reviews. In other words a niche behavior is a type of review that can be charaterizing with respect to a particular place or a particular group of reviewers or the combination of the two.

In Figure 4.2, we compare the behavior of tourists from the United States (Origin) and the province of Rome (Destination). From Figure 4.2, it is possible to determine that the most frequently visited type of attractions are *Historical Sites*, which are characterized by a high TF but a low IDF, indicating that this type of attractions are widely distributed within the data set and are not distinctive to the province of Rome.

The story changes if the focus is moved on the type of attraction named *Ancient Ruins* .The value of TF for this attribute increases from February 2016. This attribute is also characterized by a high IDF. As a result, the value of TF-IDF begins to emerge. The interpretation of this is that the group of tourists who have U.S. nationality showed an interest in a type of attraction that is not common in the data set and is distinctive of Rome province. In other words, niche behavior is identified by a high TF-IDF but in which the IDF component contributes more than the TF.

In Figure 4.2, the red bar in the line plots represents the beginning of niche behavior.

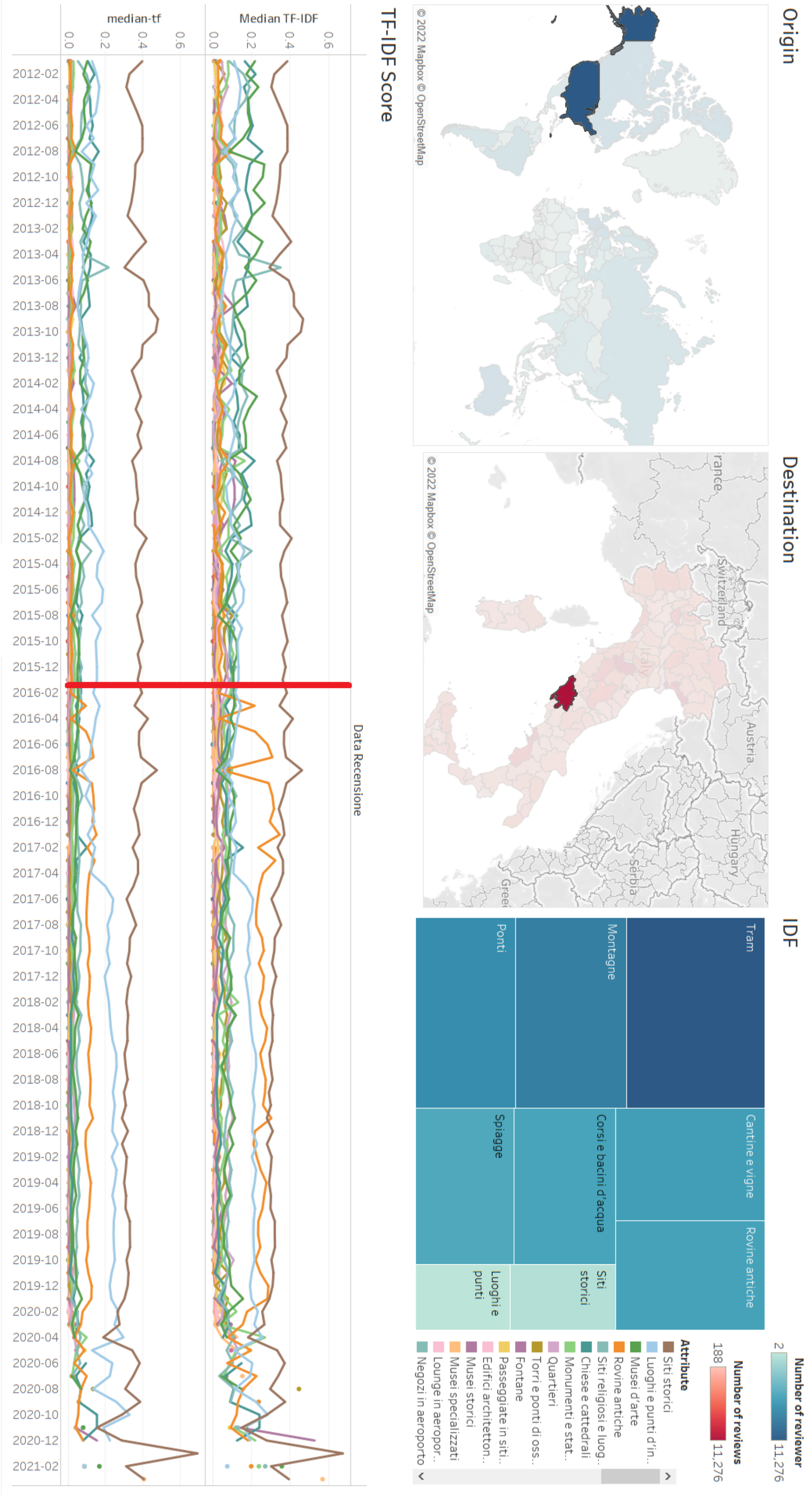


Figure 4.2: Niche behaviour representation

4.3 Dashboard use case

In this part, we will walk through an example of how the dashboard can be used. Regarding the Trentino Alto Adige region, we do research on the activities of visitors from other countries who use the service Tripadvisor. Since our objective is to gain an understanding of the behaviour of people from other countries, we start by excluding Italy as the country of origin for the reviewers and then choose the provinces of Bolzano and Trento, as shown in Figure 4.3.

The TF and TF-IDF graph is unclear and does not provide any meaningful information at this moment, so we proceed to the IDF analysis, which reveals a number of qualities with a high IDF, such as *Trekking Trails* and *Ski Slopes.* All of this makes sense because these types of attractions are not widely spread across the Italian territory, but a user with a minimum of knowledge of Italy already knows that these types of attractions are primarily found in Trentino Alto Adige because it is a predominantly hilly region. The one that arouses particular interest is *Tram*, it has the greatest IDF, while Churches have the lowest.

Figure 4.4 displays the IDF choices of *Tram*, *Hiking Routes*, and *Churches.* Let us now turn our attention to the TF-IDF Score graph. It is possible to observe that the attractions of church kinds are those that obtain the biggest proportion of reviews, but since this attribute is widely distributed in the dataset, it has a low IDF, which negatively impacts the TF-IDF score. Regarding the appeal of *Trams* and *Hiking Trails*, these two types represent a niche behavior of tourist in Trentino Alto Adige.

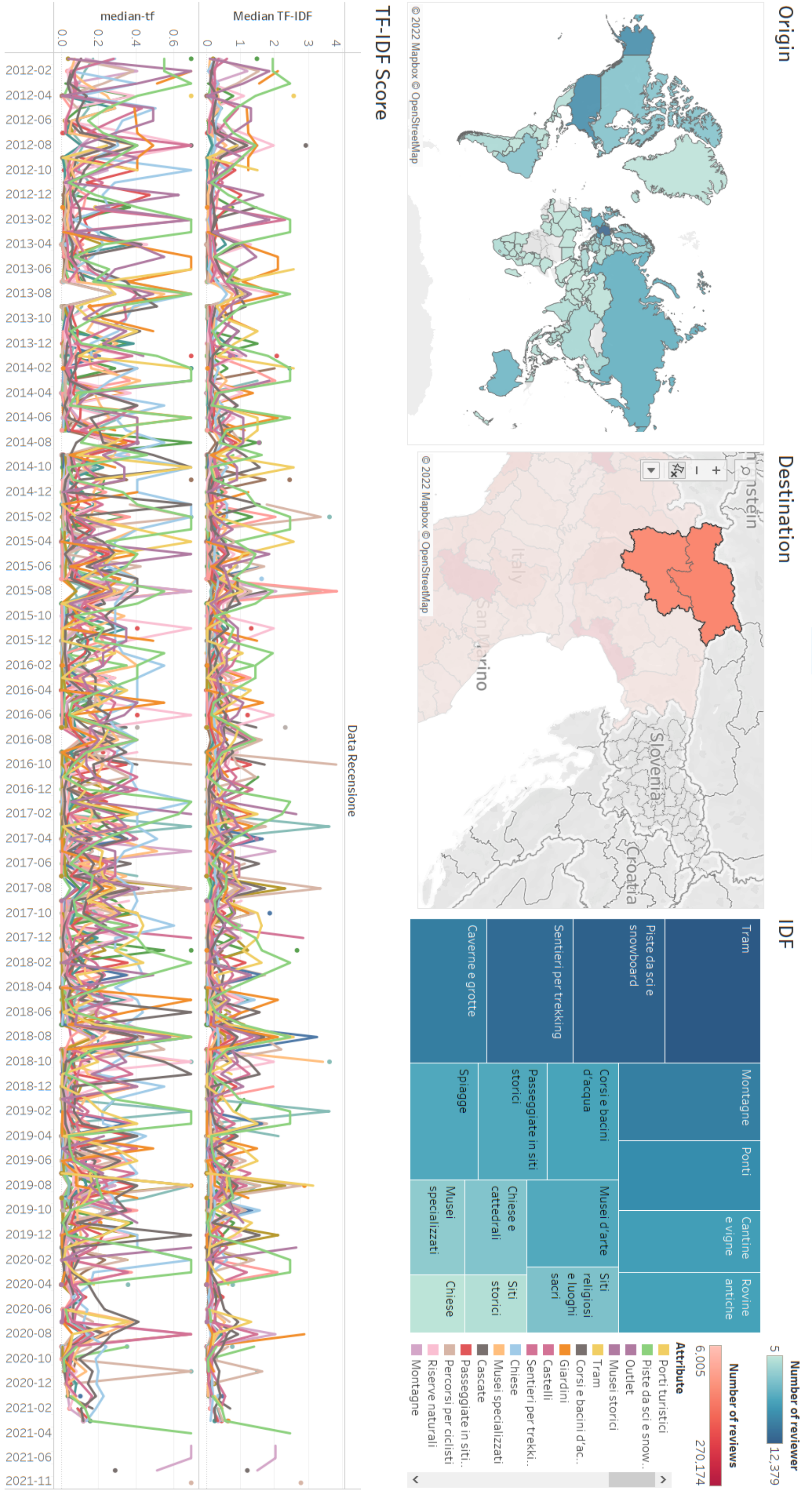


Figure 4.3: Dashboard use case step 1

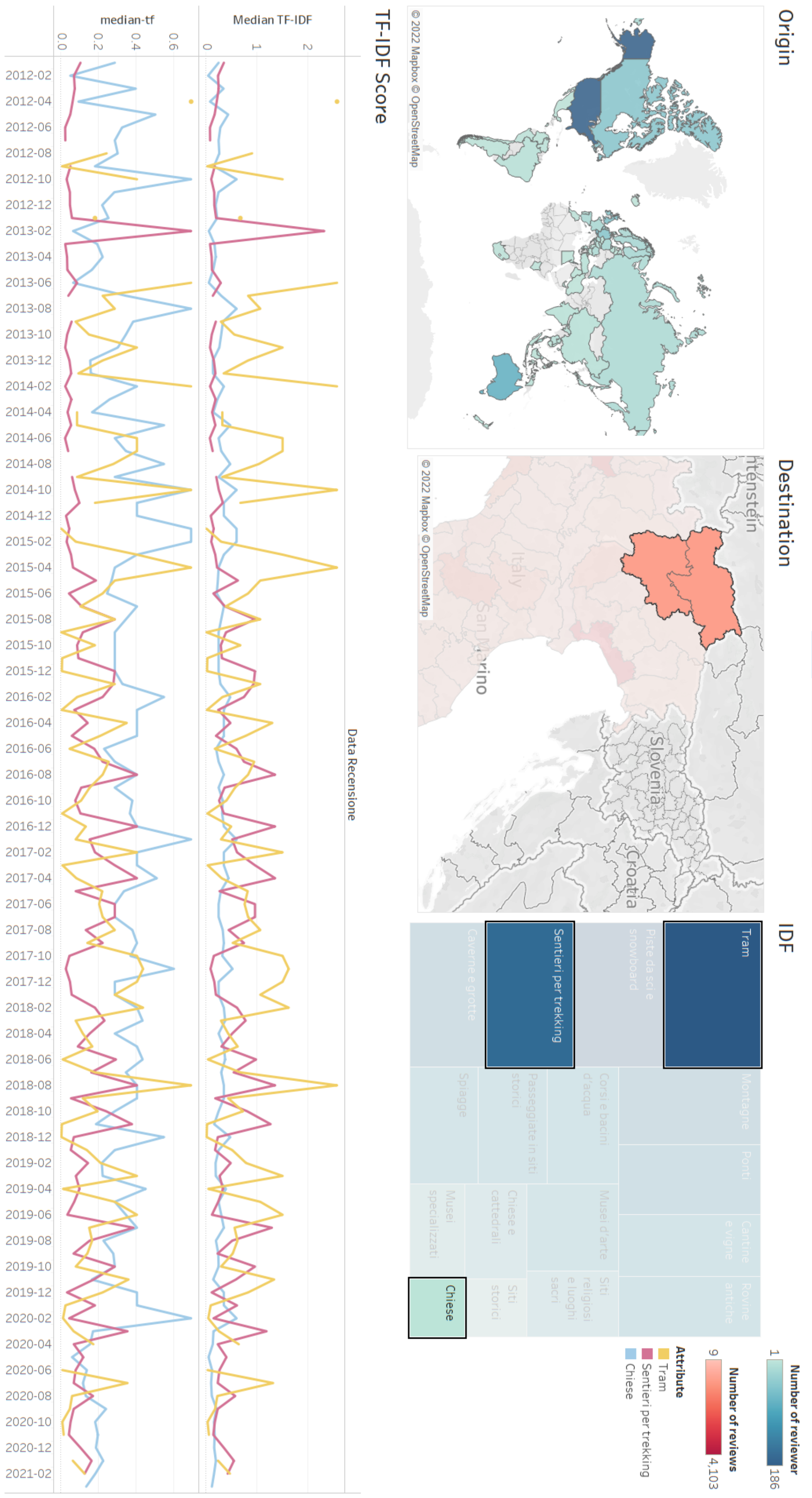


Figure 4.4: Dashboard use case step 2

Chapter 5

ETL Design

The process of designing systems in terms of software engineering contributes its own unique value and significance to the overall system development process. Mentioning it may make it seem as though it is as straightforward as anything else or as though it is simply the design of systems. However, in a broader sense, it implies a methodical and stringent approach to the design of such a system that satisfies all of the practical aspects, including flexibility, efficiency, and security. The process of specifying the architecture, components, and modules of a system in order to ensure that it will fulfill the criteria that have been established is known as systems design. One interpretation of systems design is that it is an application of systems theory to product creation [12].

In this context, systems are developed in order to fulfill the requirements expressed by the users. The purpose of the whole process is not only to find acceptable answers to the issues that now exist but also to devise solutions that may be used in the event that new problems arise in the future. The entire process of system development starts with the creation of the blueprint and ending with the delivery of the finished product, entails taking into account all of the important aspects, determining which specifications are necessary, and developing a functional system by relying on the strong technical, analytic, and development skills of trained professionals.

This section will provide a description of the design of the prototype ETL system that was created to generate the data that are eventually utilized in the dashboards that were shown in chapter 4.

Building a data warehouse or business intelligence (DW/BI) environment requires a significant amount of time and work, the majority of which is taken up by the extract, transformation, and load (ETL) system. The development of the ETL system is difficult due to the fact that there are many external constraints that put pressure on its design. These external constraints include the business requirements, the realities of the source data, the budget, the processing windows, and the skill sets of the available staff.

When it comes to establishing the architecture of an ETL system, one of the most difficult problems is to first gather all of the system's needs. This entails compiling a list of all the known needs, realities, and constraints that are relevant to the ETL system and gaining an understanding of them. Before beginning work on the ETL system, it is vital to compile the whole list of needs. [10].

5.1 Requirements

In order to provide for the implementation of the system, it is important to address and fix certain types of requirements that will be impacted by the system. The following is a list of the category of needs that were taken into consideration, each of which turned out to be an important component in the design of the system.
From the perspective of an ETL designer, **business needs** are the information needs of DW/BI system users. We use the phrase business needs in a restricted sense here to refer to the information content that business users require to make informed business decisions.
In recent years, **security** awareness has expanded significantly throughout IT, but it remains an afterthought and an undesirable expense for the majority of DW/BI teams. The data warehouse attempts to distribute data broadly to decision makers, whereas security interests believe data should be confined to those with a need to know.
The purpose of **data integration** is to have all systems operate in unison. In a data warehouse, data integration often takes the form of conforming dimensions and facts. Conforming dimensions entails defining common dimensional properties across distinct databases in order to build drill-across reports utilizing these features. Conforming facts involves reaching agreements on common business measurements, such as key performance indicators (KPIs), across distinct databases so that these statistics can be analytically compared by calculating differences and ratios.
Data latency refers to the speed with which source system data must be supplied to business users via the DW/BI system. Clearly, data latency requirements have a substantial impact on ETL design. Intelligent processing methods, parallelization, and robust hardware can accelerate batch-based data flows. However, at some point, if the data latency need is severe enough, the design of the ETL system must transition from batch to microbatch or streaming-oriented. This is not a progressive or evolutionary transition, rather, it is a paradigm shift that requires nearly every stage of the data supply pipeline to be reimplemented. The ETL team and data modelers must collaborate closely with the **BI application** developers to define the precise data handoff needs. Each BI tool has specific sensitivities that should be avoided and specific capabilities that may be utilized provided the physical data is in the proper format [10].
The requirements of the implemented ETL system are as follows:

Business Needs
The company requires that the data in the form of a TOD matrix, to which the TF-IDF-based algorithm has been applied, be updated once a month, once new data from Tripadvisor has been deposited in the corporate database.

Security
The complete data processing chain as well as the final refined data should only be accessible to the components that are a part of the firm. On the other hand, the dashboards that are based on this data should be accessible to the customers of the company and should be shared with them.

Data integration
The enterprise data base and the enterprise data sharing space called sharepoint, which is used among the many members of the organization. It is the primary components with which this ETL system must connect and integrate.

Data latency
After the contractor firm that owns the crawler sends the Tripadvisor data update to the company database, the data must be available within one day at the latest.

BI Delivery Interfaces
The program used to view the data is Tableu and the dashboard used can be seen in chapter 4, so the data must comply with the fields that were used to make the dashboard.

5.2 Tools and Technologies

Technology is in a permanent state of flux, fresh information generates new technical prospects, and there are always opportunities for advancement. This continual innovation generates a multitude of technology options from which a business may select and deploy. In addition to this extensive variety of options, technologies are growing more complicated and advancing at a quicker rate, making choices more challenging.

If technologies are chosen right, they can offer competitive advantages. However this is one of the most difficult decisions to make. The picks must be compatible with the organization's existing technology and systems, as well as its human factors, culture, strategy, and objectives.

In the software industry, developments are made in complex and dynamic systems involving several interrelated elements such as people, processes, methods, products, technologies, tools and techniques. These interrelations create a feedback system where a change or improvement in one area creates effects in others parts of the system directly or indirectly. Additionally, system complexity might stem from its risks and uncertainties, dynamic behavior or changes over time. This section will provide the list of technologies used in the development of the ETL system.

Python
It is a popular computer programming language that is utilized often in the process of developing websites and applications, as well as automating chores and doing data analysis. Python is what is known as a general-purpose programming language, which means that it may be put to use in the creation of a wide range of applications and is not tailored to solve any particular issues. Because of its flexibility and the fact that it is easy to learn, it has quickly become one of the most popular programming languages in use today.

Python has established itself as an indispensable tool in the field of data science. This has made it possible for professionals working in the field to use the language to perform intricate statistical calculations, build machine learning algorithms, create data visualizations, manipulate and analyze data, and complete other data-related tasks.

Spark
It is a data processing framework that can analyze very large data sets fast and distribute data processing jobs over numerous computers, either alone or in conjunction with other distributed computing technologies. These two characteristics are crucial to the fields of big data and machine learning, which need the deployment of tremendous computer capacity to sift through vast data collections. Spark also alleviates some of the programming constraints associated with these activities by providing developers with an easy to use API that abstracts away the majority of the grunt work associated with distributed computing and large data processing.
In this work is used PySpark the Python API for Apache Spark.

PostgreSQL
PostgreSQL, more often known as Postgres, is a relational database management system (RDBMS) that is open-source and free to use. It places an emphasis on flexibility and SQL conformance. It is built to manage a wide variety of workloads, ranging from single computers to data warehouses or Web services with a large number of users logged in at the same time.

Docker
Docker is a platform that is open source and allows developers to build, deploy, run, update, and manage containers. Containers are standardized, executable components that combine application source code with the operating system libraries and dependencies necessary to run that code in any environment. Docker enables developers to do all of these things. Containers make it easier to design and deliver programs that run on distributed systems. As more and more businesses move toward cloud-native development, their use has become increasingly widespread.

Amazon Web Service
It is comprehensive cloud computing platform supplied by Amazon that combines infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) solutions. AWS services may provide an organization with computing power, database storage, and content delivery services, among other capabilities. In detail, the services used are:

- **S3 bucket**: is a public cloud storage resource available AWS Simple Storage Service (S3), an object storage offering. Amazon S3 buckets are similar to file folders, store objects, which consist of data and its descriptive metadata.
- **Amazon Elastic Container Service (ECS)**: It is a highly scalable, high-performance container management solution that supports Docker containers and enables easy application execution.

5.3 System Design

This section will give the ETL pipeline design. The architecture is organized into two modules. The first of which is known as the Origin Cleaning Module (OCM). As its name suggests, this is responsible for cleaning the *Origin* field, which is a component of the reviewers. Additional information regarding this task was presented in section 3.1. The second component is called the Transformation-Evaluation Module (TEM). Its job is to produce the TOD matrix by making use of the *Origin* field once it has been cleaned by the OCM. After that it is also responsible to applying the TF-IDF scoring method, which was covered in chapter 3.

5.3.1 Origin Cleaning Module

In order to accomplish the task set for this module, a Nominatim-provided API is utilized. Nominatim is one of the sources for the search box on OpenStreetMap (OSM). It provides a tool to search OSM data by name and address (geocoding) and produces synthetic addresses for OSM points (reverse geocoding) [18].
This module consists of a component called *sql executer*, which is responsible for executing sql queries to the enterprise database. The object of these queries is the *Id* of the reviewer and its *Origin*. At this level, the *Origin* field may not be formatted correctly. The result of these queries is stored in the *S3 temporary bucket*. At the end of the execution of the *sql executer*, the *checker* component is activated. This component is responsible for filtering in the *S3 temporary bucket* and it basically checks which *Origin* is already present in the correspondence table contained in the *S3 main bucket*. Those which are already in such basket are deleted from the *S3 temporary bucket*. Once the filtering has taken place, only reviewers with unregistered and uncleaned provenances are in the *S3 temporary bucket*.
At that point, the *cleaner* component comes into play. It sends a request with the malformed string to the *Numinatim API*. It can respond in two different ways. The first is a response with a JSON. In this case the API was able to reconstruct the malformed string. Within the JSON there is a field named country, in this case the pair (bad format provenance, country) is created. If the API cannot get any match this returns an empty JSON and in this case the checker module provides to create a pair (bad format provenance, NULL). Once all provenances have been validated by the *S3 temporary bucket*, they are appended to the correspondence table in the *S3 main bucket*, which will be used later in the data preparation process.
Sql executes, *Checker*, *Cleaner*, contained within a Docker container, which runs inside to an ECS instance. The schema for this module is as follows:

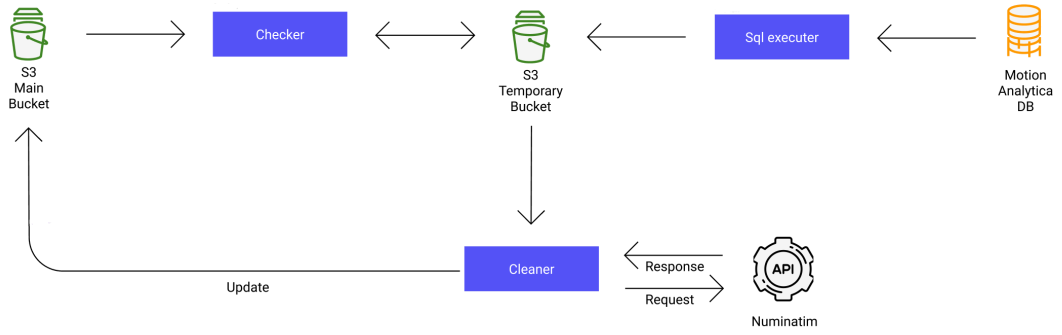


Figure 5.1: Origin Cleaner Module design

5.3.2 Tansformation-Evaluation Module

The task of this module is to create the final form of the data. This module consists of the TOD builder component, this through processing using the PySpark tool the data from the tables shown in Chapter 2, contained within the company's Postgres DB. In addition to this the module is responsible for replacing the Origin field with the correctly formatted one from the lookup table contained within the *S3 main bucket*, which was created by OCM. After this is completed, the data passes within the TF-IDF scorer component. This applies the TF-IDF algorithm, explained in chapter 3, for all POI categories. For each one a column is added with its TF-IDF, TF and IDF. This TOD matrix with the scores is written to the *S3 main bucket*.

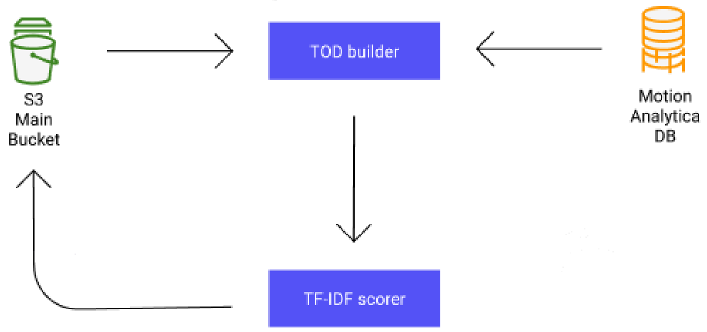


Figure 5.2: Tansformation- Evaluation Modulee design

Chapter 6

Conclusion & Future Work

This dissertation presented a project based on a curricular internship at Motion Analytica Srl, an innovative startup with the goal of proposing an innovative way of analyzing and understanding the movement of people and things in motion. This project was based on social data, specifically from Tripadvisor, the purpose of the work was to employ a process of processing, reconstruction, aggregation and analysis of the raw data in order to obtain a refiend data that contained suggestions with respect to certain niche behaviors of reviewers in Italian points of interest. The work was carried out in different parts.
Chapter 2 dealt with exploratory data analysis (EDA). Data visualization was heavily employed in this part to help this task.
Chapter 3 dealt with the issue of cleaning and refining the data. The refined form of the data is called the Time Origin Destination (TOD) matrix. Also in this chapter, the Term Frequency Inverse Document Frequency (TF-IDF) algorithm for scoring within the TOD matrix was explained.
The design and a use case of the implemented dashboards was presented in Chapter 4. Moreover what is meant by niche tourist behavior was visually presented in this chapter.
In conclusion, in Chapter 5, the technologies and tools used to implement the ETL were discussed in addition to the design of the ETL. One result of this work is a working ETL pipeline with attached data visualization. Another result is that through the use of the TF-IDF algorithm it was possible to identify niche behaviors of tourists. The company Motion Analytica Srl found the results of this work satisfactory.
Regarding the breadth of Motion Analytica Srl's social media data analysis, this study serves as an starting point. In the future, this data will be combined with data from other platforms, such as Airbnb and Google Reviews. In addition, there is the possibility of a significant integration of these data with those from the telco network. This is made possible through a relationship between the company and one of the largest European telephone operators.

Bibliography

[1] Motion analytica website. https://www.motionanalytica.com/.

[2] World Tourism Organization (2015), UNWTO Tourism Highlights, 2015 Edition, UNWTO, Madrid, DOI: https://doi.org/10.18111/9789284416899.

[3] Lorenzo Boccalon. Social media data analysis for tourism. Università Ca' Foscari, 2021.

[4] UNESCO World Heritage Centre. "italy". 2021.

[5] Valerio Della Corte, Claudio Doria, and Giacomo Oddo. The impact of Covid-19 on international tourism flows to Italy: evidence from mobile phone data. Questioni di Economia e Finanza (Occasional Papers) 647, Bank of Italy, Economic Research and International Relations Area, October 2021. URL https://ideas.repec.org/p/bdi/opques/qef_647_21.html.

[6] United States Supreme Courts. Hiq labs v. linkedin. 04 2022.

[7] John Rydning David Reinsel, John Gantz. Data age 2025: The evolution of data to life-critical. 2017.

[8] ISTAT. Movimento turistico in italia nel 2018. 2018.

[9] Simon Kemp. Digital 2020: July global statshot. 07 2020.

[10] Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.* Wiley Publishing, 3rd edition, 2013. ISBN 1118530802.

[11] Weilin Lu and Svetlana Stepchenkova. User-generated content as a research mode in tourism and hospitality applications: Topics, methods, and software. *Journal of Hospitality Marketing & Management*, 24(2):119–154, 2015. doi: 10.1080/19368623.2014.907758. URL https://doi.org/10.1080/19368623.2014.907758.

[12] Robert C. Martin. Clean architecture: A craftsman's guide to software structure and design. USA, 2017. Prentice Hall Press. ISBN 0134494164.

[13] KunalF Mehta, Maya Salvi, Rumil Dand, Vineet Makharia, and Prachi Natu. A comparative study of various approaches to adaptive web scraping. In *ICDSMLA 2019*, pages 1245–1256, 2020.

[14] Bank of Italy. The weight of tourism in italy, the characteristics of the demand and the accommodation capacity. 12 2018.

[15] Bruno Stecanella. Understanding tf-id: A simple introduction. MonkeyLearn, 05 2018.

[16] Egbert Van der Zee and Dario Bertocchi. Finding patterns in urban tourist behaviour: a social network analysis approach based on tripadvisor reviews. volume 20, 12 2018. doi: 10.1007/s40558-018-0128-5.

[17] Debra Aho W. Us consumers are flocking to tiktok, 04 2020. URL "https://www.insiderintelligence.com/content/us-consumers-are-flocking-to-tiktok". [Online; accessed 1-October-2022].

[18] OpenStreetMap Wiki. Nominatim — openstreetmap wiki,, 2022. URL https://wiki.openstreetmap.org/w/index.php?title=Nominatim&oldid=2379097. [Online; accessed 1-October-2022].

[19] Wikipedia contributors. List of iso 3166 country codes — Wikipedia, the free encyclopedia, 2022. URL https://en.wikipedia.org/w/index.php?title=List_of_ISO_3166_country_codes&oldid=1112403451. [Online; accessed 1-October-2022].

[20] Wikipedia contributors. Social — Wikipedia, the free encyclopedia, 2022. URL https://en.wikipedia.org/w/index.php?title=Social&oldid=1103873083. [Online; accessed 1-October-2022].

[21] Wikipedia contributors. Tourism in italy — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Tourism_in_Italy&oldid=1111568479, 2022. [Online; accessed 29-September-2022].

[22] Zheng Xiang and Ulrike Gretzel. Role of social media in online travel information search. *Tourism Management*, 31(2):179–188, 2010. ISSN 0261-5177. doi: https://doi.org/10.1016/j.tourman.2009.02.016. URL https://www.sciencedirect.com/science/article/pii/S0261517709000387.