UNIVERSITÀ DI PISA

Dipartimento di Filologia, Letteratura e Linguistica

Corso di Laurea Magistrale

in Informatica Umanistica

Business Intelligence and Data Driven approaches for the analysis of cultural events

Relatore:
Riccardo Guidotti

Candidato:
Arianna Lisi

ANNO ACCADEMICO 2020/2021

Contents

List of Figures

List of Tables

1. Introduction

Nowadays, we are living in a world made of data. Data is everywhere: it is behind the strategies of a supermarket chain, the decisions of a big company, the study of a scientific research or the organization of an event. Thanks to data science, it is possible to easily manage scientific innovations, algorithms and complex methods to extract knowledge and information that can be useful to a great range of application domains.

Among the various fields to mention, also the cultural one is included. Indeed, the benefits of data analysis, data mining and machine learning studies in the management of cultural organizations are clearly effective. But why and how are they so important?

Let us imagine a very simple situation: the coordination of a cultural event. Events attract people (depending on their scope and resonance) and people bring money to the event's organization, as well as to many other activities that operate in the area, like restaurants, hotels or transports. To figure out an adequate way to measure the economic investment and to guarantee great participation, it is necessary to know the number of people that are going to take part in the event, their movements, behaviors, interests and characteristics. In simple terms, we need knowledge.

Considering that cultural organizations need to estimate their social and economic impact to have a more active role in society, it is therefore evident why it is so relevant to solve these issues. In this kind of situation, data analytics approaches can help us through different strategies and can be extremely useful to extract indicators measuring the impact of cultural events on the area that hosts them, such as the consequences on the economy, the involvement of different segments of the population or the growth of tourism. Furthermore, it is important to study the public and its peculiarities: is it a younger or an older public? What sort of habits, interests or occupation does it have? In this way, it can be quite simple to get an overview of the target that should be considered and to maneuver the organization of events to attract more people and increase profit.

It is important to enhance cultural organizations' awareness about the great power of the data they already have or that they can produce, in order to fully comprehend and

build up different aspects of their world: most organizations make annual reports, project reports, collect financial data, carry out social media and media coverage analyses, and this gives them a great source of knowledge. Moreover, regular reviews of the data collected can produce useful insights for the future events or activities to be carried out. This part of the analysis is mainly carried out with Business Intelligence tools [**11**], which are easier to use and to understand by the majority of the people. But this kind of study can receive a great boost with the exploitation of Data Science technologies [**3**].

Even though a lot of people would not notice the difference between extracting information through Data Mining's algorithms and making simple predictions with Data Warehouse instruments, actually there are many. First of all, using Data Science's techniques can handle much larger volumes of data, which is a very crucial requirement for analyzing the big data we are all surrounded by [**3**]. On the other hand, there are also some advantages in using software programs such as Excel, and one of them is naturally the fact that they are pretty simple tools that can be easily understood by anyone, whilst Data Science methods tend to be more complex to learn. But the improvements that the latter can bring are much more significant: it is not only about making simple statistics on data, indeed Data Science can help to study different kinds of data in much more interesting and thorough ways, such as text processing or speech recognition, convenient for tasks like Sentimental Analysis. One of the most important tasks on known data is the prediction of new data: it is certainly a very useful statistical operation that can be studied with the purpose of anticipating the public's behavior during an event so that it can be organized and structured with foresight and awareness.

Pulling data together and processing them permit to transform raw data into intelligible insights, which has the outcome of, first of all, generating and sharing useful snapshots of what is happening in the activities of the organizations rather quickly [**12**]. Secondly, it enables organizations to be trained in the culture of data measurement. As it is clear, data science can offer to the world of cultural events a new way of thinking, understanding and studying different points of view, through which it will be possible to make people's experiences increasingly enjoyable [**15**].

This is where the aim of this thesis comes in: comparing tools from a Business Intelli-

gence and a Data Mining analysis, in order to enhance the cultural and creative industries. Hence, this study is partially included in the Me-Mind project[1], which is co-financed by the European Commission's "Creative Europe" program, having one major objective, that is to provide cultural and creative industries with different data-driven decision making business models, for communicating and making the impacts more understandable to visitors, to the stakeholders and policy makers. Even though the use cases chosen for this project are the Estonian National Museum as a case of use of permanent collections, and the Pisa Internet Festival for events, here we will consider just the analysis about the Internet Festival.

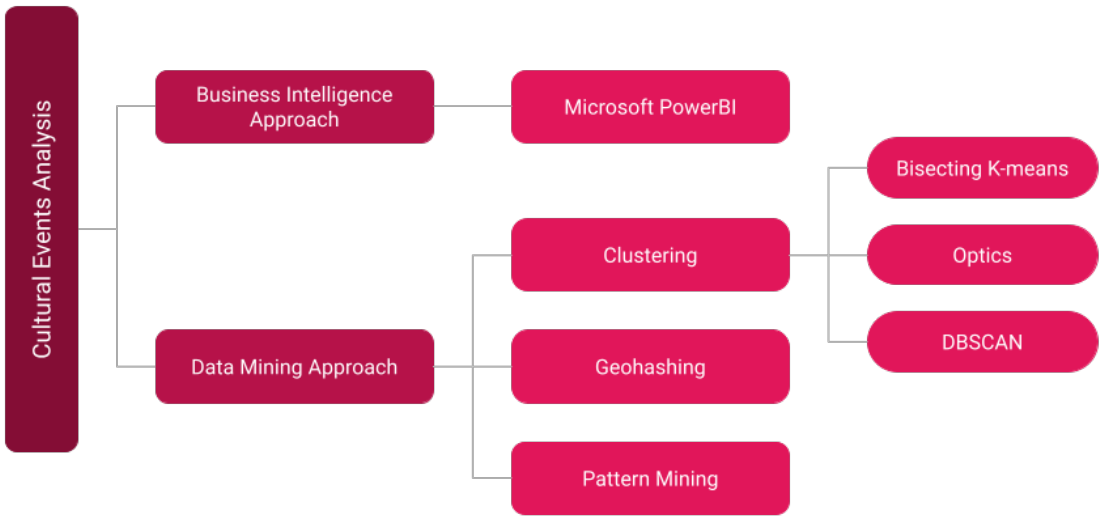

Figure 1.1: Work analysis scheme.

Figure 1.1 shows a graph that represents the analysis scheme that will be described. To analyze cultural events, the Pisa Internet Festival was therefore chosen as a use case, on which two types of studies were carried out: the first is the type that adopts a more business oriented approach, in which the Microsoft PowerBI tool was used, while the second type deals with the analysis with a more analytical and profound approach, using data mining techniques, such as clustering (with three different algorithms: Bisecting K-Means, Optics and DBSCAN), geohashing and pattern mining, in order to study the data

[1]To learn more about this topic, visit the web page at the following link: `https://www.memind.eu/`

in a more sophisticated way.

For a greater comprehension on how to approach this thesis, the work is organized as follows. In chapter 2 a space is dedicated to "setting the stage", that is to briefly illustrate some key concepts or topics for the overall understanding of the study that will be described. In 2.1 will be presented what the Internet Festival is, how it takes place within the city of Pisa and why it can be relevant as an object of analysis. In 2.2 a not too in-depth overview of the main tools and algorithms used to carry out the analyzes will be provided: in the first subsection that is encountered, that is the 2.2.1, PowerBI will be described, while in subsection 2.2.2 the algorithms applied for Data Mining analysis will be illustrated, including DBSCAN, OPTICS and Bisecting K-means for clustering, Geohash for geospatial encoding and PrefixSpan for pattern mining.

In chapter 3 the Business Intelligence Analysis will be outlined, first explaining the type of data used and the way in which they were extracted in section 3.1, then passing through the methodology used for the study in section 3.2, up to the presentation of the results obtained in section 3.3, divided respectively by the results on the internal data of the organization and by the results on the data obtained from the social media accounts.

In chapter 4 the Data Mining Analysis is described, also in this case passing first through the data retrieval phase in section 4.1, then through the methodology adopted in section 4.2, to conclude with the results produced at each step of the work in section 4.3.

Eventually, in chapter 5, the conclusions of the entire study will be summarized and possible future works will be discussed.

2. Setting the stage

Before getting to the heart of the analysis, it is necessary to offer a general overview of some essential notions for fully understanding the project. In section 2.1 a description of what exactly is the Internet Festival will be provided, along with the cultural impact that it has on the city of Pisa and on the visitors. In section 2.2 a definition of the main tools and algorithms used for the entire analysis will be explained, presenting PowerBI, and lastly the techniques used in data mining processes will be illustrated, since they permit to study data in a much more incisive and accurate way.

2.1 Internet Festival

Internet Festival (IF)[2] is a multifaceted event dedicated to the theme of the Internet and the digital revolution, that is set up every year for a few days in the city of Pisa since 2012, generally at the turn of the end of September and the beginning of October. It is a platform that offers a wide range of in-depth, debate and entertainment activities and events that revolve around the theme of the Internet and digital innovation with the aim of stimulating discussions and ideas relating to all aspects of the Network, thanks to the contribution of researchers, experts, administrators, top users, influencers, artists and enthusiasts. In other terms, it is a multidimensional path linked to the city of Pisa, as a place rich in excellence and historically ideal for thought and research in the scientific and humanistic fields, to naturally project itself onto an international dimension [**16**].

It provides interesting activities such as the so called T-Tours (Tutorial Tours), which are an integral part of the Internet Festival, made up of educational and training courses that stimulate curiosity and offer useful tools to navigate the boundless world of the Internet and technology.

Usually, it is organized by thematic tracks. This means that, based on the visitors'

[2]To learn more about this topic, visit the web page at the following link: `https://www.internetfestival.it/`

tastes and interests, every day they will be able to find the event that is right for them, that most intrigues or interests them: from cybersecurity to sport, from robotics to gaming, from exhibitions to shows. Obviously, during the pandemic in 2020, a physical organization of this magnitude throughout the whole city of Pisa could not be realized, so from that moment it has been combined with events and workshops to which everyone can participate online through the typical platforms used for remote meetings.

In this project, we are going to analyze data from the Internet Festival of 2016, hence we are going to focus more on the program and oganization from that year.

The Internet festival in 2016 has been held from the 6th to the 9th of October and it involved:

- 108 events including panels, workshops, keynote speeches, book presentations;
- 127 T-TOUR activities;
- one installation on Ponte di Mezzo (which is the main bridge in Pisa);
- 5 exhibitions;
- 3 cooking shows;
- one hackhathon;
- one running marathon with night tracking of the route;
- one game contest;
- 3 film screenings;
- one DJ-set;
- 3 shows and live musical performances;
- 3 national premieres;
- 16 presentations of books, ebooks and meetings with the authors;
- 212 speakers, including 20 foreigners and 64 T-Tour entertainers.

As it is clear, the impact of this event is really massive. Moreover, the festival program welcomed, presented and launched 34 start-ups involved in the *.itCup*, the business competition of Registry.it[3] (CNR), to encourage opportunities for new business ideas and potential investors and thus to contribute to the technological development of Italy, and in the Bootstrap of Startupitalia. In figure 2.1 it is possible to see an installation in Piazza dei Cavalieri in Pisa.

Figure 2.1: Picture of an installation of the Internet Festival in 2016.

Regarding the locations of the events throughout the city, in figure 2.2 we can observe the points of interest that hosted the installations, presentations and activities related to the Internet festival of that year. As it is evident, the scope of the festival is widespread throughout the city, involving the points of greatest frequentation of citizens and tourists, attracting a great participation.

[3]To learn more about this topic, visit the web page at the following link: `https://www.nic.it/it/progetti/itcup`

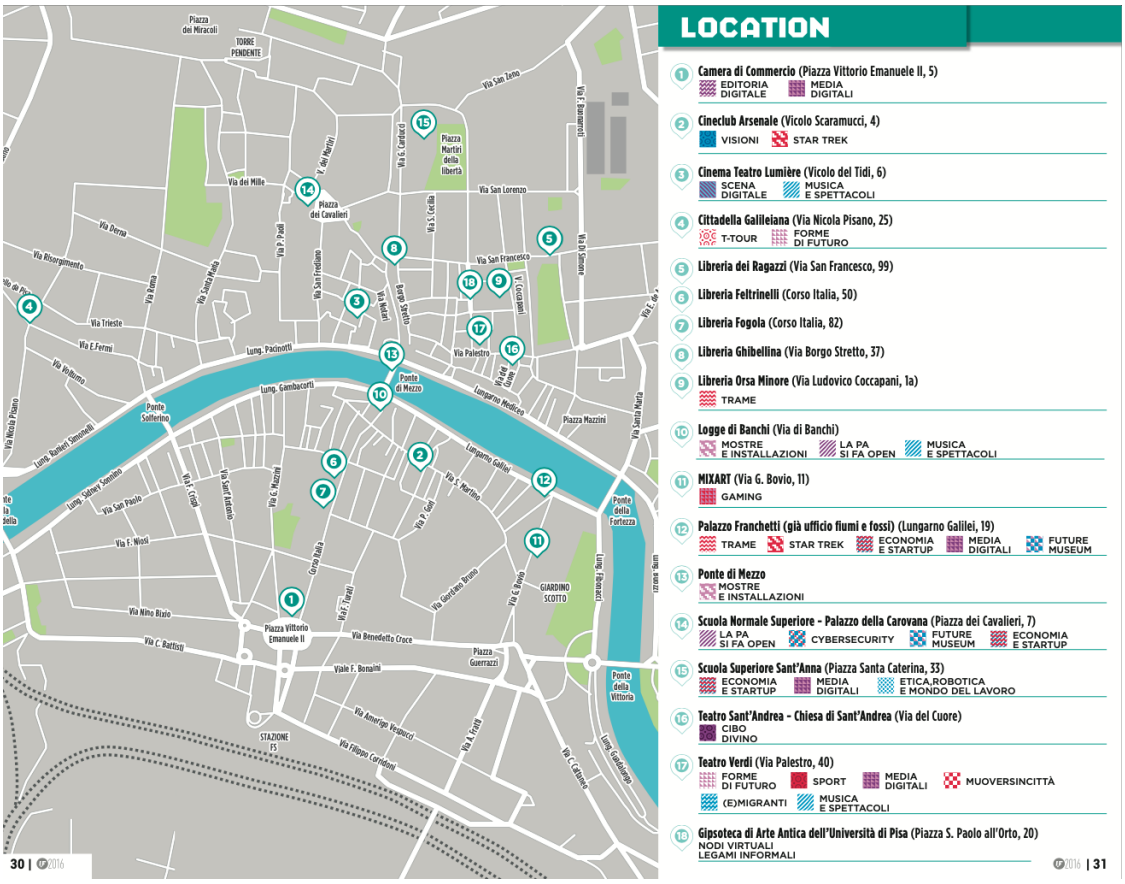

Figure 2.2: Locations of IF 2016's events from the corresponding booklet.

2.2 Tools and algorithms used

2.2.1 PowerBI

Power BI Desktop is the tool used in the Business Intelligence analysis. It is a free downloadable Microsoft software that is indeed included in the "Business Intelligence" category as a business service useful to create reports by most companies. It has a simple and intuitive interface to use, which offers services on the cloud and directly on desktop, with data warehouse functionalities, like preparing data and interactive dashboard.

To provide an example of how it can be used, just consider that most of the companies that use it aim to create weekly or monthly reports on some data to show: it is possible to load tables and manipulate them directly through the tool to clean them up or to organize

the data in a different way, in order to make them suitable for the analyzes to be addressed; then it is possible to take advantage of the possibility of creating graphs in a very rapid and intuitive way to represent trends in the data to be included in the report (which can be interactive), together with text to describe the analysis and other components of your choice.

It is undoubtedly one of the simplest tools to use for reporting, which is why it is one of the most used in companies that need to describe data with high frequency.

In the following figure you can see a screenshot of PowerBI Desktop[4], while reporting some data with interactive graphs.

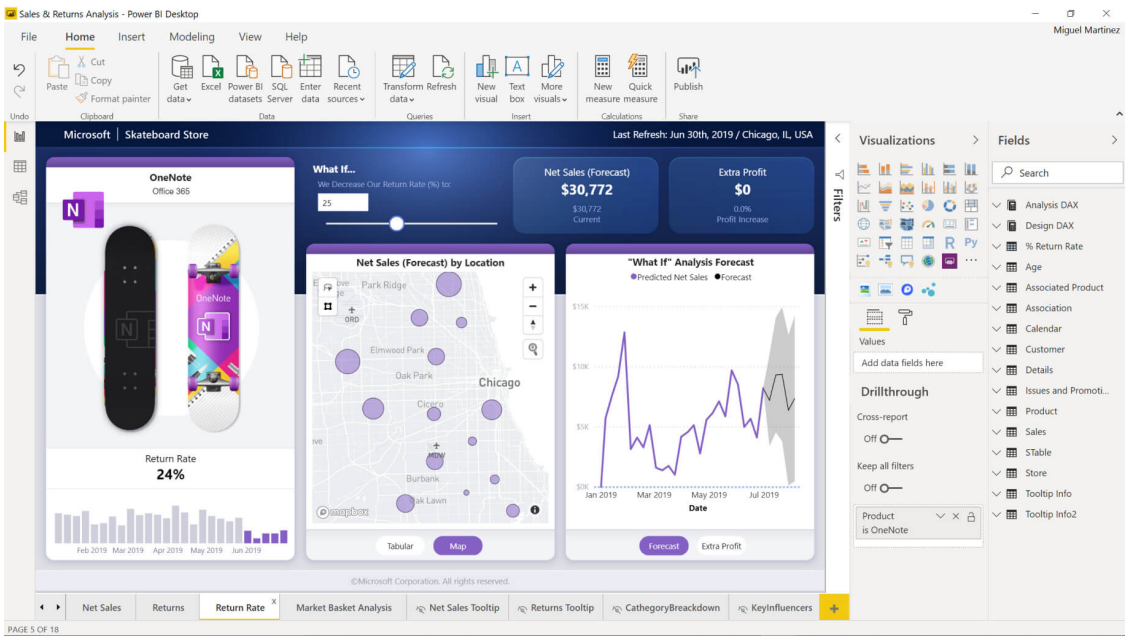

Figure 2.3: Screenshot of PowerBI Desktop.

2.2.2 Data Mining algorithms and techniques

What is Data Mining?

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large

[4]This image was downloaded from the web page at the following link: `https://powerbi.microsoft.com/it-it/desktop/`

> data sets in order to find novel and useful patterns that might otherwise remain unknown. They also provide the capability to predict the outcome of a future observation, such as the amount a customer will spend at an online or a brick-and-mortar store. Not all information discovery tasks are considered to be data mining. Examples include queries, e.g., looking up individual records in a database or finding web pages that contain a particular set of keywords. This is because such tasks can be accomplished through simple interactions with a database management system or an information retrieval system. These systems rely on traditional computer science techniques, which include sophisticated indexing structures and query processing algorithms, for efficiently organizing and retrieving information from large data repositories. Nonetheless, data mining techniques have been used to enhance the performance of such systems by improving the quality of the search results based on their relevance to the input queries.

This is a quotation extracted from *Introduction to Data Mining* [**1**], a very useful manual that explains this topic in a beginner level. Indeed, it describes in a very simple language what exactly Data Mining is and why it is so important nowadays for different kinds of analyses that require a deeper level of detail.

Continuing with the manual reading, at a certain point it indicates that there are different types of Data Mining tasks: *predictive* and *descriptive*. The former are used to predict the value of an attribute based on the values of the other attributes, the latter are used instead for drift patterns that represent "invisible" relationships in the data. The former includes tasks such as classification or regression, while the latter includes techniques such as cluster analysis and association analysis, which are also those used in this project to describe hidden relationships in the data.

Clustering

Cluster analysis groups data objects based on information found only in the data that describes the objects and their relationships. So, it observes the items within a group that are similar to one another and different from the items in other groups. Clustering can

be regarded as a form of classification in that it creates a labeling of objects with class (cluster) labels [**1**].

Also in this case, we distinguish various types of clusterings: *hierarchical* versus *partitional*, *exclusive* versus *overlapping* versus *fuzzy*, and *complete* versus *partial*. A *partitional* clustering is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. Otherwise, if we permit clusters to have subclusters, then we obtain a *hierarchical* clustering, which is a set of nested clusters that are organized as a tree and each node (cluster) in the tree (except for the leaf nodes) is the union of its children (subclusters), and the root of the tree is the cluster containing all the objects. In addition, an *overlapping* or non-exclusive clustering is used to reflect the fact that an object can simultaneously belong to more than one group (class). In a *fuzzy* clustering, every object belongs to every cluster with a membership weight that is between 0 (absolutely does not belong) and 1 (absolutely belongs). A *complete* clustering assigns every objectt to a clusetr, whereas a *partitial* clustering does not, because in some cases there are objects that may not belong to well defined groups, like noise or outliers [**1**].

Among the different clustering techniques and algorithms, in this project we are going to use DBSCAN, OPTICS and Bisecting K-Means.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an algorithm capable of defining clusters following the idea that a point belongs to a cluster if it is close to many points of that cluster. It is density-based, meaning that it produces a partitional clustering locating regions of high density that are separated from one another by regions of low density (see figure 2.4[5]) [**1**].

[5]This image was downloaded from the web page at the following link: `https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py`

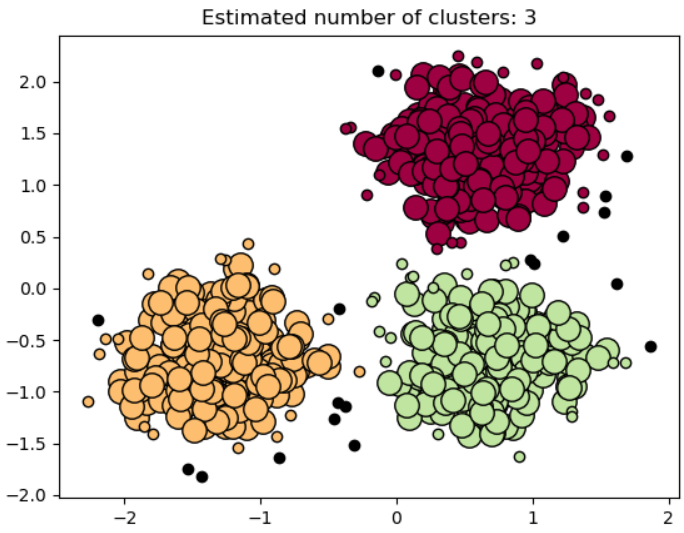

Figure 2.4: Example of DBSCAN clustering algorithm.

In the center-based approach on which DBSCAN is based, density is estimated for a particular point in the data set by counting the number of points within a specified radius (*epsilon*) of that point. It is a simple method to implement, but the density of any point will depend on the specified radius. Any two core points that are close enough, within a distance epsilon of one another, are put in the same cluster. Likewise, any border point that is close enough to a core point is put in the same cluster as the core point. Here is reported the DBSCAN algorithm in a few steps:

1. Label all points as core, border, or noise points.
2. Eliminate noise points.
3. Put an edge between all core points within a distance epsilon of each other.
4. Make each group of connected core points into a separate cluster.
5. Assign each border point to one of the clusters of its associated core points.

OPTICS (Ordering Points To Identify the Clustering Structure) is a clustering algorithm closely related to DBSCAN, since it finds core sample of high density and expands

clusters from them but, unlike DBSCAN, it keeps cluster hierarchy for a variable neighborhood radius. It is better suited for usage on large datasets [**6**].

The last technique we are going to analyze is the **Bisecting K-Means** algorithm, which is a straightforward extension of the basic K-Means algorithm, which uses a simple idea: we first choose K initial centroids, where K is a user-specified parameter, namely, the number of clusters desired. Each point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster. We repeat the assignment and update steps until no point changes clusters, or equivalently, until the centroids remain the same. Bisecting K-Means is slightly different because, to obtain K clusters, we split the set of all points into two clusters, select one of these clusters to split and so on, until K clusters have been produced. So we apply regular K-means with K=2 (that's why the word bisecting) and we keep repeating this bisection step until the desired number of clusters is reached (see figure 2.5) [**1**] [**13**].

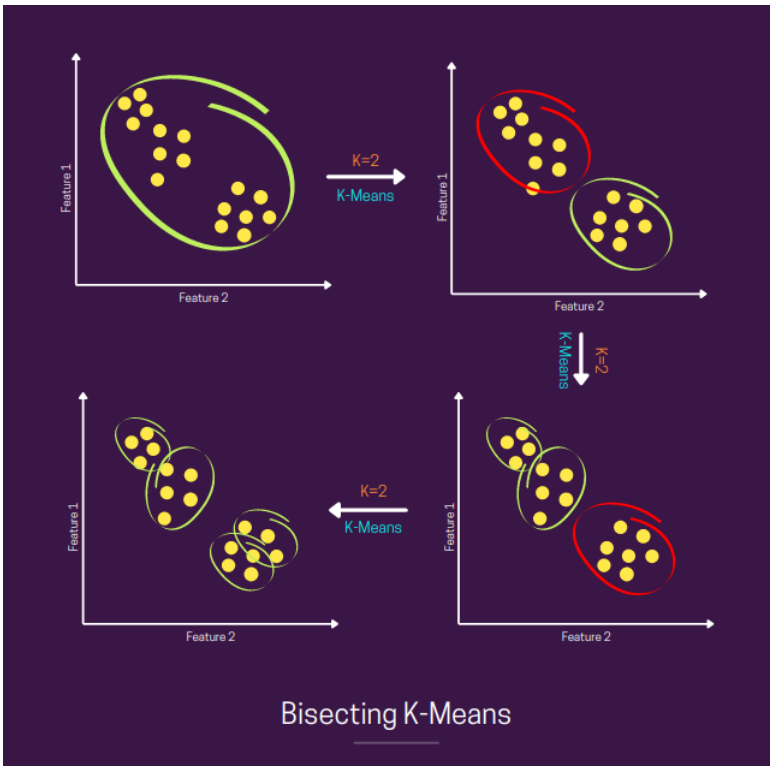

Figure 2.5: Example of how Bisecting K-Means works.

The algorithm can be explained in a few steps:

1. Initialize the list of clusters to contain the cluster consisting of all points (repeat).
2. Remove a cluster from the list of clusters.
3. Perform several “trial” bisections of the chosen cluster.
4. For i=1 to number of trials do
5. Bisect the selected cluster using basic K-Means.
6. End for
7. Select the two clusters from the bisection with the lowest total SSE.
8. Add these two clusters to the list of clusters.
9. Until The list of clusters contains K clusters.

Geohash

Geohash is a latitude/longitude geocode system to convert Geographic coordinate system into a string using a base32 character map. A Geohash string represents a fixed spatial bounding box, thus Geohash divides geographic space into buckets of grid shape [**5**]. Imagine the world is divided into a grid with 32 cells. The first character in a geohash identifies the initial location as one of the 32 cells. This cell will also contain 32 cells, and each one of these will contain 32 cells (and so on repeatedly). Adding characters to the geohash sub-divides a cell, effectively zooming into a more detailed area.

To understand which is the most suitable precision to consider for the subdivision of the geographical area, we can refer to the parameters in table 2.1[6].

[6]The data from this table was extracted from the web page at the following link: `https://en.wikipedia.org/wiki/Geohash`

Geohash length	Lat bits	Lng bits	Lat error	Lng error	Km error
1	2	3	±23	±23	±2500
2	5	5	±2.8	±5.6	±630
3	7	8	±0.70	±0.70	±78
4	10	10	±0.087	±0.18	±20
5	12	13	±0.022	±0.022	±2.4
6	15	15	±0.0027	±0.0055	±0.61
7	17	18	±0.00068	±0.00068	±0.076
8	20	20	±0.000085	±0.00017	±0.019

Table 2.1: Digits and precision in km.

Pattern Mining

In simple terms, Pattern Mining is a data mining technique to discover interesting patterns or relationships hidden in large data. It is also known as association analysis. The associations among the data can be represented in the form of sets of items present in many transactions, which are called frequent itemsets or association rules, that are implication expressions of the form $X \rightarrow Y$, where X and Y are disjoint itemsets.

The strength of an association rule can be measured in terms of its support and confidence. *Support* determines how often a rule is applicable to a given dataset, and it is an important measure because a rule that has a very low support might occur simply by chance. On the other hand, *confidence* determines how frequently items in Y appear in transactions that contain X, so it measures the reliability of the inference made by a rule (for a given rule $X \rightarrow Y$, the higher the confidence, the more likely it is for Y to be present in transactions that contain X).

A common strategy adopted by many association rule mining algorithms is to decompose the problem into two major subtasks: the frequent itemset generation, whose objective is to find all the itemsets that satisfy a minimum support threshold, and the rule generation, used to extract all the high confidence rules from the frequent itemsets found in the previous step.

The frequent itemset generation has a major idea used behind its logic, which is called the *Apriori* principle:

> If an itemset is frequent, then all of its subsets must also be frequent. [**1**]

This strategy of trimming the exponential search space based on the support measure is known as support-based pruning, which is made possible by a key property of the support measure, namely, that the support for an itemset never exceeds the support for its subsets. This property is also known as the *anti-monotone* property of the support measure. You can see an illustration of this principle in figure 2.6[7].

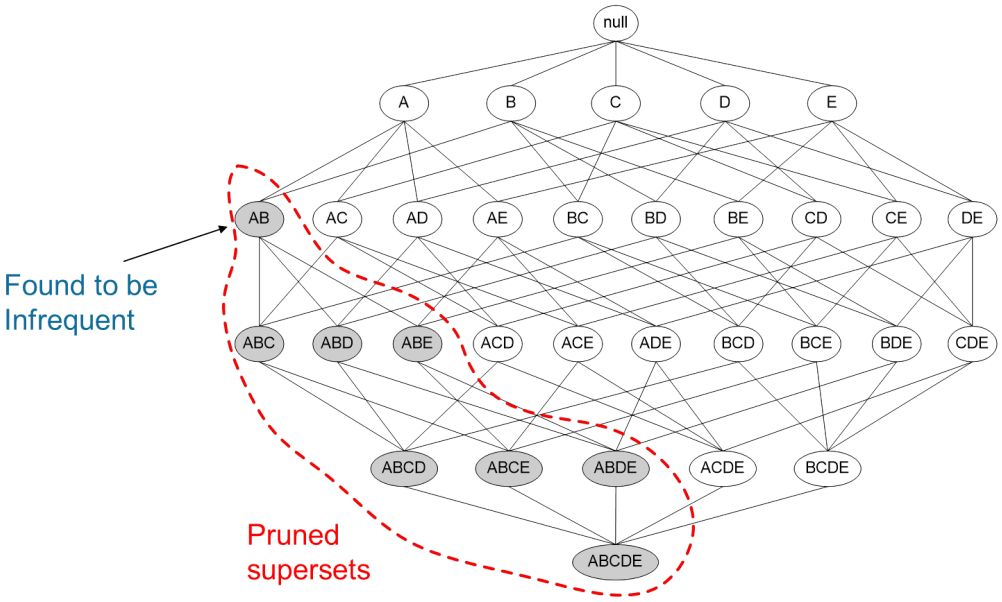

Figure 2.6: Illustration of the Apriori principle.

Considering that the number of frequent itemsets produced from a transaction can be very large, it is useful to identify a small representative set of frequent itemsets from which all other frequent itemsets can be derived: this is why there are the maximal and the closed frequent itemsets. A frequent itemset is *maximal* if none of its immediate supersets are frequent, while it is *closed* if none of its immediate supersets has exactly its same support count.

[7]The image was downloaded from the web page at the following link: `https://chih-ling-hsu.github.io/2017/03/25/apriori`

In this project we are going to conduct an analysis of this kind, using an implementation of the frequent sequential pattern mining algorithm **PrefixSpan**, which finds sequential patterns by prefix-projected pattern growth. Given a set of sequences, where each sequence consists of a list of elements and each element consists of set of items, it finds all the frequent subsequences, which are the subsequences whose occurrence frequency in the set of set of sequences is no less than a minimum support [**4**].

3. Business Intelligence Analysis

Business Intelligence (BI) uses information systems and transaction databases to provide decision-making support and transform data into intelligence within a rational management framework [**2**]. The BI Analysis carried out in this project is somehow a prototype that will be used to obtain meaningful and business-oriented information through the correlation of data, which requires an in-depth knowledge of the context. Indeed, a single KPI (Key Performance Indicator), that is an instrument for measuring and reporting the values of certain operating conditions of a model and that per se does not supply any useful insight, becomes extremely meaningful only when it is put in correlation with other indicators. A correlation can become significant and can also give us an indication of cause-effect relationship, only if it is supported by the stories, i.e. the information that will be provided by the use cases or coming from other external sources.

3.1 Data retrieval

In order to give a framework on what the data retrieval phase actually consists of, it is necessary to distinguish between two types of data: every organization has both internal data that is produced within the organization and additional data that can be sourced from external stakeholders. The first type of data is also called primary data, because it is purposefully collected, concerning a specific challenge of understanding the impact of the activities of an organization. On the other hand, secondary data is extracted to understand external features that can still be definitely meaningful and interpreted for improving different aspects.

In this section we are going to analyze the internal data that Fondazione Sistema Toscana[8] (FST) managed to extract regarding the Internet Festival, with the purpose of

[8]Fondazione Sistema Toscana is a non-profit foundation under private law for the creation and management of the Internet Portal of Tuscany and to promote the regional territory and its identity with integrated digital communication tools: web, multimedia productions, social media.

understanding the progress of certain aspects during the years. It is important to underline that the external data are not going to be analyzed and described in this study, since the Me-Mind project managed to derive them through the production of questionnaires that have been analyzed at a later time to gain also secondary information. Indeed, the availability of external data is not always easy to access, since it is often unstructured and needs to be collected from different sources.

A preliminary collection of quantitative data from internal databases has been done on a sample of 23 indicators covering the time frame 2015-2020 and belonging to economy, knowledge and skills, development of cultural-themed and innovation-themes policies, promotion of cultural development and cultural diversity, impact on the cultural fabric of the city and beyond, gender equality, sensitivity to minorities' issues and communication dimension.

Alongside the analysis of the internal data, Fondazione Sistema Toscana was able to extract quantitative data on user participation insights from Facebook, Instagram, Twitter and YouTube accounts, by simply downloading them in CSV format files[9]. Indeed, companies and organizations need to be aware of the opportunities and challenges of social media data, because these useful platforms can provide cultural organizations with large amounts of data.

Following this procedure, there are two different points of view with which to observe the socio-economic impact of the Internet Festival: one on the internal data of the organization and the other on data obtained from the interactions on the social profiles of the festival. The methodology of the two analyses is described in section 3.2 and the results obtained are illustrated in section 3.3.

[9]The comma-separated values (CSV) is a text file-based file format used for importing and exporting (such as from spreadsheets or databases) of a data table.

3.2 Methodology

Figure 3.1: Workflow followed during the Business Intelligence Analysis.

Figure 3.1 shows the workflow that has been followed for the Business Intelligence analysis on the internal data and on the social media data obtained from Fondazione Sistema Toscana. The first step has already been described in the previous section about data retrieval from the FST's internal databases and the Internet Festival's social media accounts dashboards. The next steps of data cleaning and data structuring in the workflow consisted of cleaning up and reorganizing the tables obtained in order to avoid conflicts due to data type mismatch or disorder. Indeed, for what concerns the internal data from the organization sometimes there were different typologies of data for the same feature of the dataset and, obviously, it had to be corrected and turned into a standard type for all data of that feature. The resulting tables presented a structure in which the remaining indicators were located in the first column and the years indicating which period the data collected refer to were located in the first row.

On the other hand, the CSV tables with the social media data collected directly from the dashboards were of different compositions, so, to solve this problem, the data cleaning and data structuring phases were, also in this case, necessary. In the following points, the datasets and the changes reported on them will be illustrated:

- From the Facebook dashboard, we obtained two datasets dating back to 2019 and 2020, consisting of 2387 columns and 123 rows. There is data about the performance of the account over the years and about the provenance and demographics of the public that interacts with it. Obviously, a great part of the features presented in these tables have been deleted, for redundancy or futility reasons. Moreover, some columns which described the countries reported just the country codes, so, in order

to make them readable by the program, they have been converted into the complete names.

- From the YouTube account it has been possible to download the total number of visualizations daily from 2011 to March 2021.
- Regarding the Twitter account we downloaded a lot of interesting information from the September - December period of 2020.
- Lastly, from the Instagram account we obtained some data which has been useful for having a general overview of the performance of the account.

The fourth point of the work is the analysis of the data, to observe certain characteristics and tendencies that emerge in the comparison of the data during the years, which are eventually shown in the data visualization process through the help of explanatory graphs made up with the PowerBI tool.

3.3 Results

3.3.1 Internal indicators

The main purpose of the analysis is to study the relationships among the indicators and analyze the results in order to seek some correlations related to different sectors. For instance, from the comparison of the number of visitors and the engagement of people in educational activities in figure 3.2, it is possible to notice that the two curves follow a very similar trend, with a pick in 2019, so we may assume that these indicators are correlated.

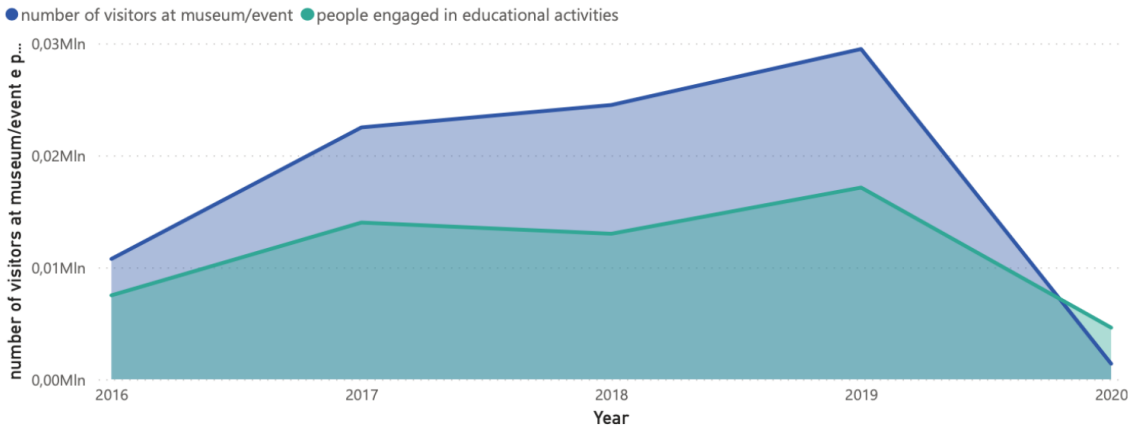

Figure 3.2: FST's Number of visitors at IF and People engaged in educational activities per Year.

Regarding the collateral cultural events and activities, we can see in figure 3.3 that there were just 2 in 2017 and that the highest number is 7 in 2020.

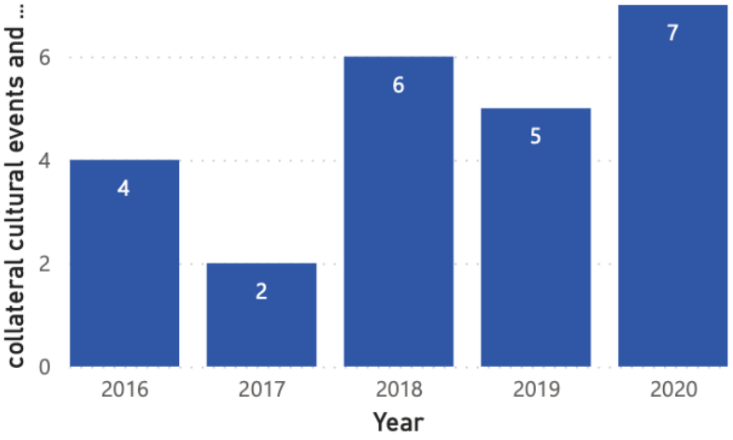

Figure 3.3: FST's Collateral cultural events and activities per Year.

If we look at the coverage of the press, we can see in detail in figure 3.4 the trend of web-papers and blog posts, newspapers and TV services coverage. We can notice that there is a general decreasing tendency, which drops off drastically in 2020, maybe because of Covid-19.

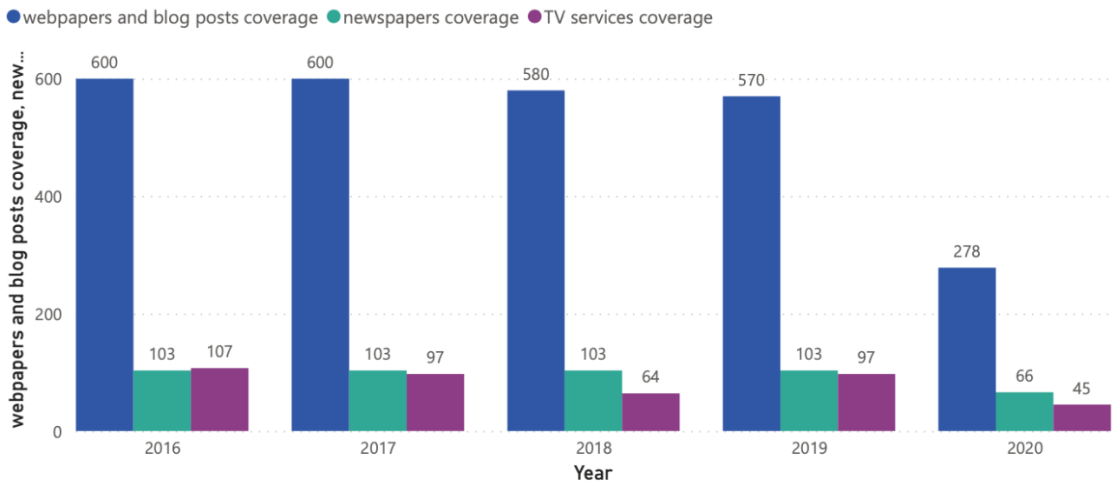

Figure 3.4: FST's Press coverage.

The number of events dedicated to social issues and the number of activities and variety of the offer are two indicators that follow a very similar pattern, going down at the beginning until 2017, then going up rapidly until 2019 and finally decreasing again in 2020 (see figure 3.5).

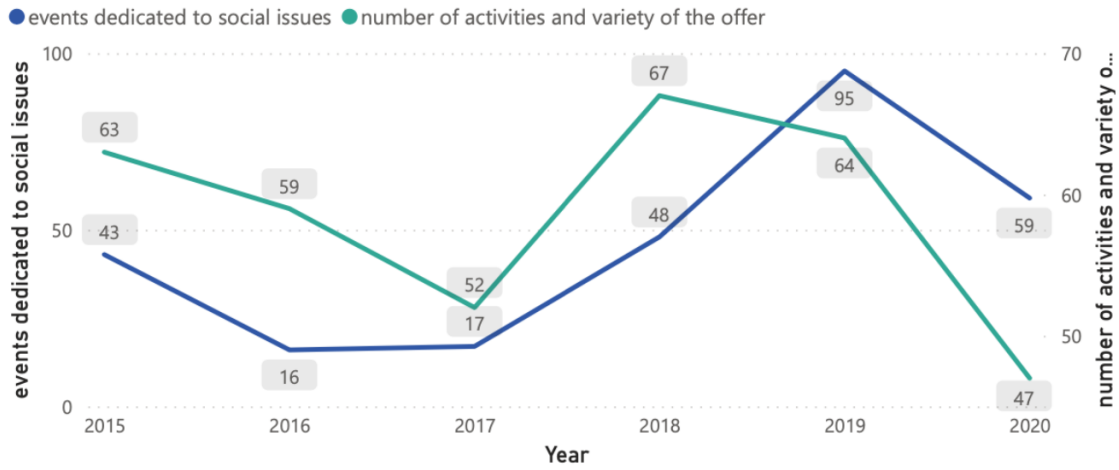

Figure 3.5: FST's Events dedicated to social issues and Number of activities and variety of the offer per Year.

Another interesting indicator is the one dedicated to the training credits acquired through the participation in IF, which are pleasantly increasing during the years: they started from 65 in 2017 and arrived at 530 in 2020 as it is shown in figure 3.6. This trend can be easily

explained by the fact that there are always more agreements with other professionals who participate in the Internet Festival.

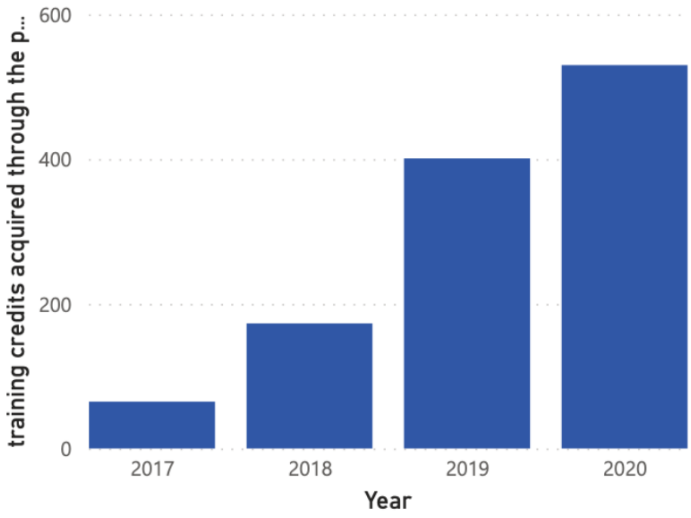

Figure 3.6: FST's Training credits acquired through the participation in IF per Year.

The staff of FST remained pretty stable through the years. If we look at the staff analysis per gender in figure 3.7, we can see that there is always a greater number of male employees, which is pretty close to the females' one in the first two years and becomes bigger in 2018 and 2020. Also regarding the gender division in speakers, trainers, performers and curators, we can see that the number of men is much more dominant than the number of women: in 2016 there were 157 more men than women, in 2017 were 169, in 2018 were 163, in 2019 were 208 and in 2020 were 131.

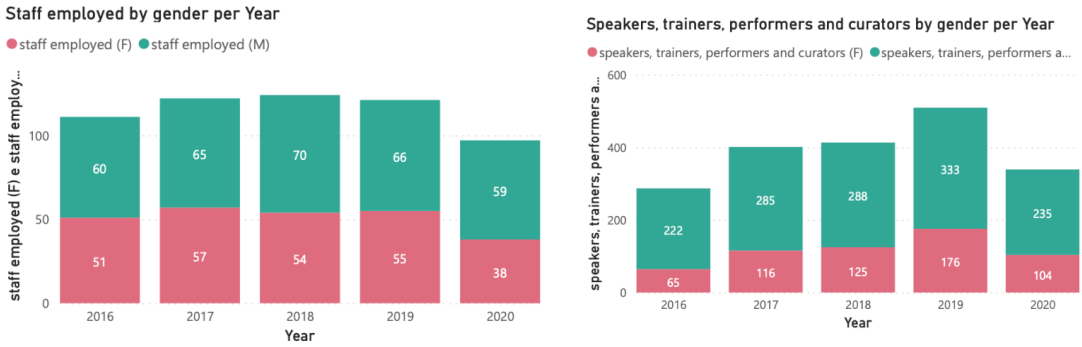

Figure 3.7: FST's staff employed and Speakers, trainers, performers and curators by gender per Year.

Also the people employed permanently and temporarily remained with a similar proportion (see figure 3.8), which is a bit more different in 2018: generally there is always a higher number of people employed temporarily than the ones employed permanently.

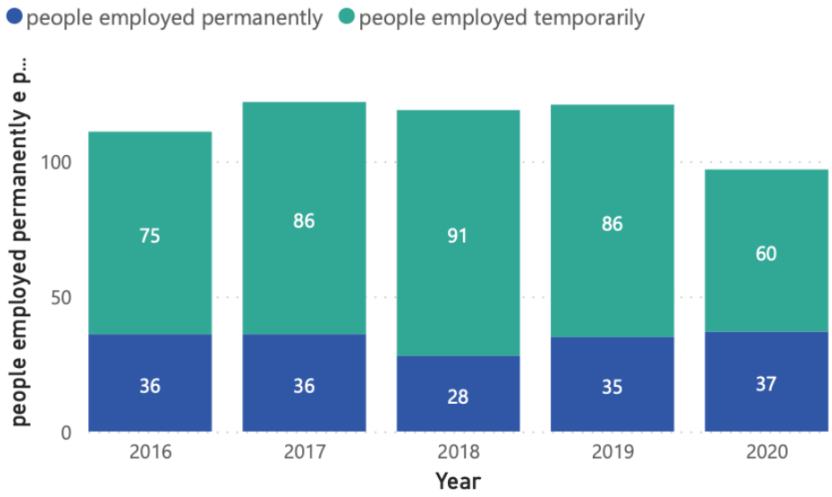

Figure 3.8: FST's People employed permanently and temporarily per Year.

From the budget analysis in figure 3.9 emerges that the major part of the budget comes from the funding by public administration and other public promoters, which is generally around 400K euros, while in 2020 is a bit more than 300K euros. The budget invested at city level is always more than 200K euros and in 2018 reaches more than 300K euros. The budget invested at the regional level was higher than 100K euros in the first two years, then went down to 60K euros in 2018 and 2019 and eventually reached 20K in 2020. The budget invested at the country level is the highest in 2018 exceeding 160K euros. Lastly, the financial contributions provided by private sponsors started decreasing after 2018, reaching a bit more than 20K euros.

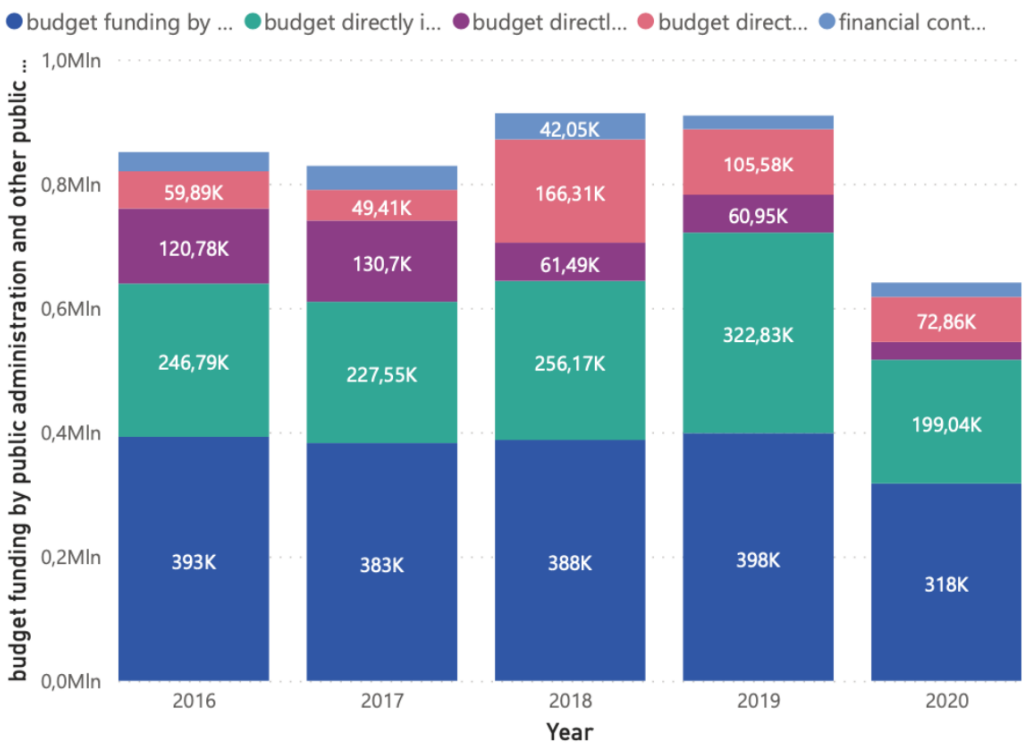

Figure 3.9: FST's Budget comparison: budget funding by public administration and other public promoters, budget directly invested at city, region and country level and financial contributions provided by private sponsors.

Regarding the number of suppliers directly engaged in figure 3.10, there is no big difference among the years, except that in 2020 when it dropped from 85 to 58, while the number of PAs involved per year was slowly improving.

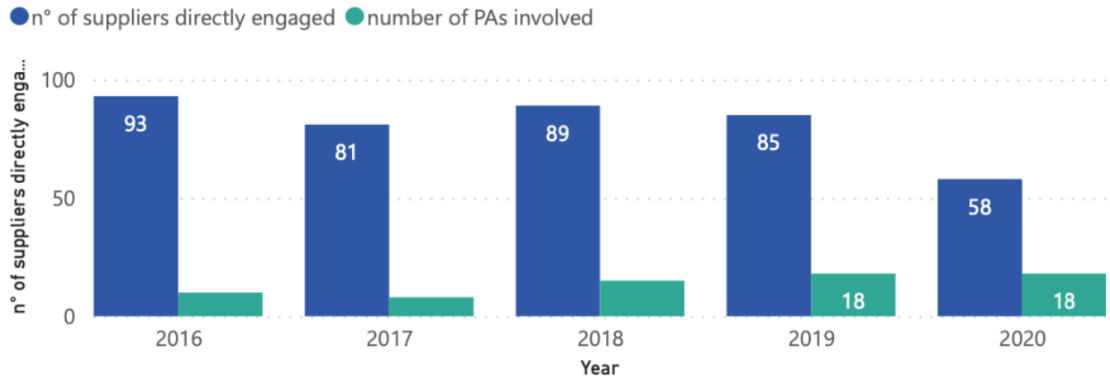

Figure 3.10: FST's Number of suppliers directly engaged and Number of PAs involved per Year.

Concerning the external grants and projects gained in figure 3.11, there is a rapid

descent to 1K units in 2018 and then a sudden increase in 2019 of more than 50K units.

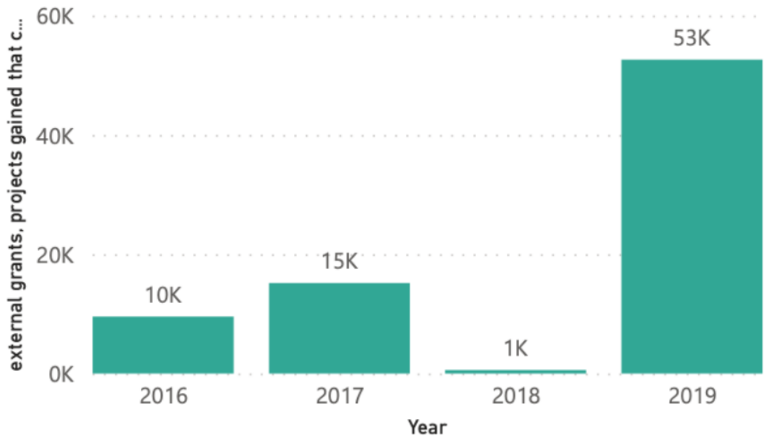

Figure 3.11: FST's External grants, projects gained.

3.3.2 Social media data

In the comparison in figure 3.12 it is shown how the different social profiles attract people: we can see that tendentially the greatest coverage of interactions comes from Facebook and Twitter, while for what concerns Instagram and YouTube, the number of visualizations are not as high as the other two social accounts. In general we can notice that the involvement of the public increases during the period of October, during which usually starts the Internet Festival, that in 2020 has been held from 8 to 11 of October. Moreover, there has been some activity during the month of November (mostly in the Facebook page), and that's explainable by the fact that there have also been online events during that month.

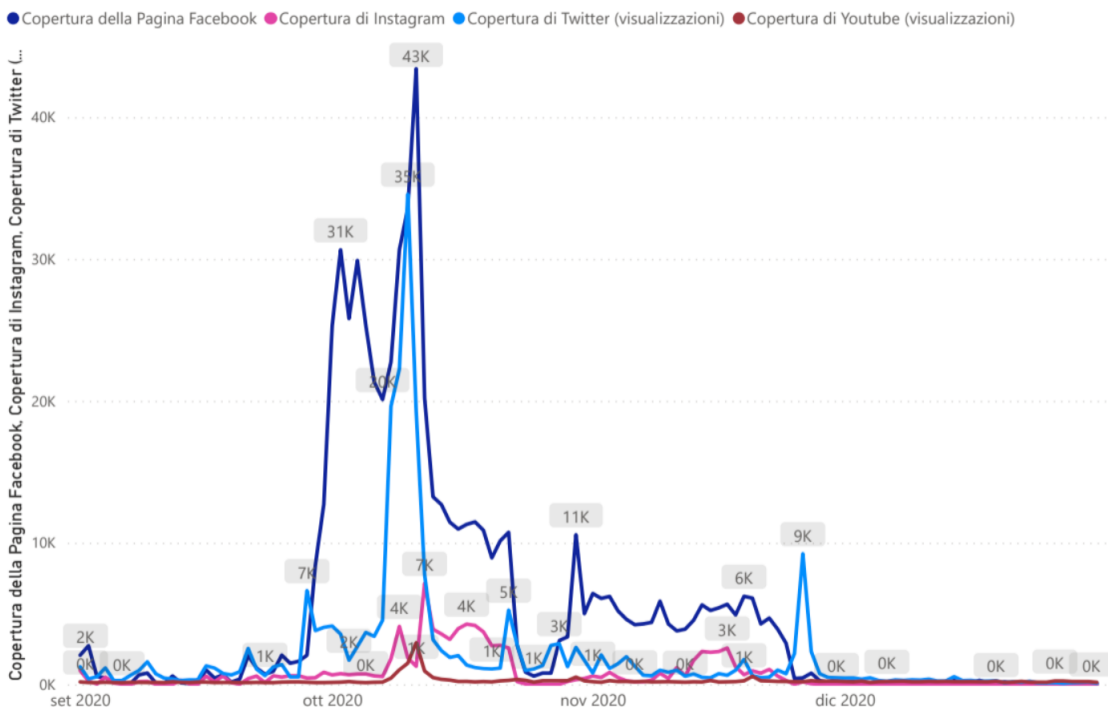

Figure 3.12: Facebook, Instagram, Twitter and YouTube coverage.

Facebook

In this paragraph, we will analyze only data belonging to 2020, excluding 2019, because the trends emerged from the results were very similar, just with a light increment on the overall performance.

From the observation of figure 3.13, it is possible to discover that during the period of IF, the Facebook account is much more used and has a much higher resonance among the audience and that the feedback is mainly positive, with few comments and many likes.

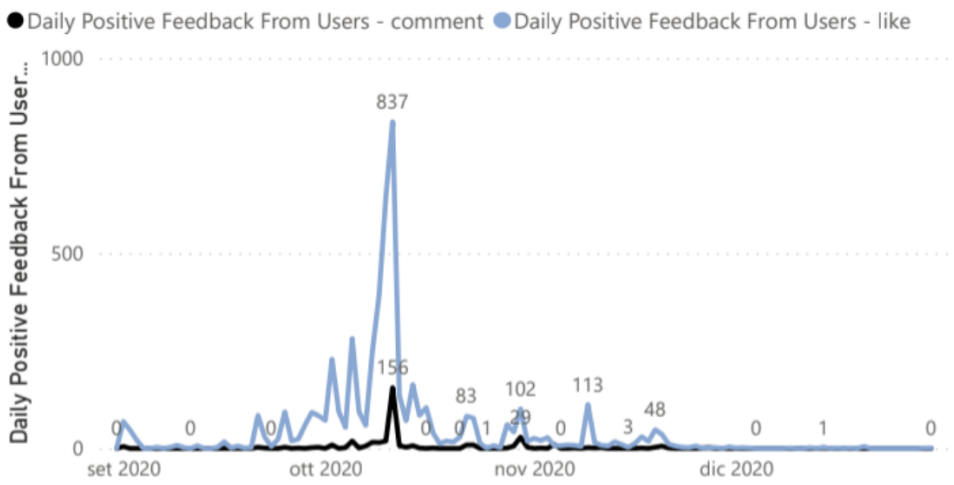

Figure 24. Daily positive feedback from users with likes and comments in 2020

Figure 3.13: Daily positive feedback from users with likes and comments in 2020.

Notice that, in figure 3.14, the daily total reach means the number of people who had any content from the page or about the page enter their screen; this includes posts, check-ins, ads, social information from unique users who interact with the page and more, so obviously the trend is almost stackable on the one describing the impressions, which are the number of times any content from the page or about the page entered a person's screen. The difference is that the impressions are a total count, so not only unique users are considered.

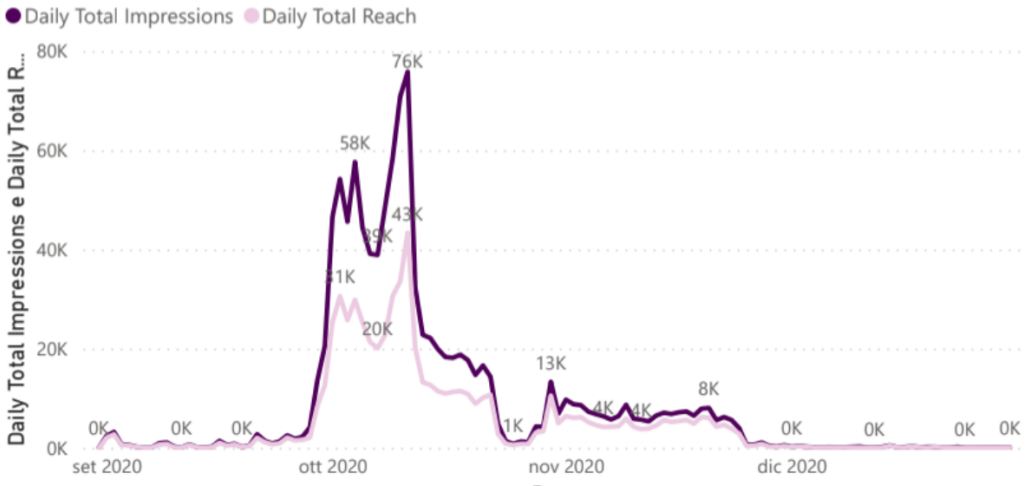

Figure 3.14: Daily total impressions and total reach in 2020.

The map shown below in figure 3.15 represents the percentage of lifetime likes on av-

erage per country. It is possible to see that, as expected, the highest number of likes comes from Italy (with 18.325,08 likes), followed in the top positions by the United Kingdom (with 198,18 likes), Germany (with 131,76 likes), Spain (with 93,85 likes) and the United States (with 85,84 likes).

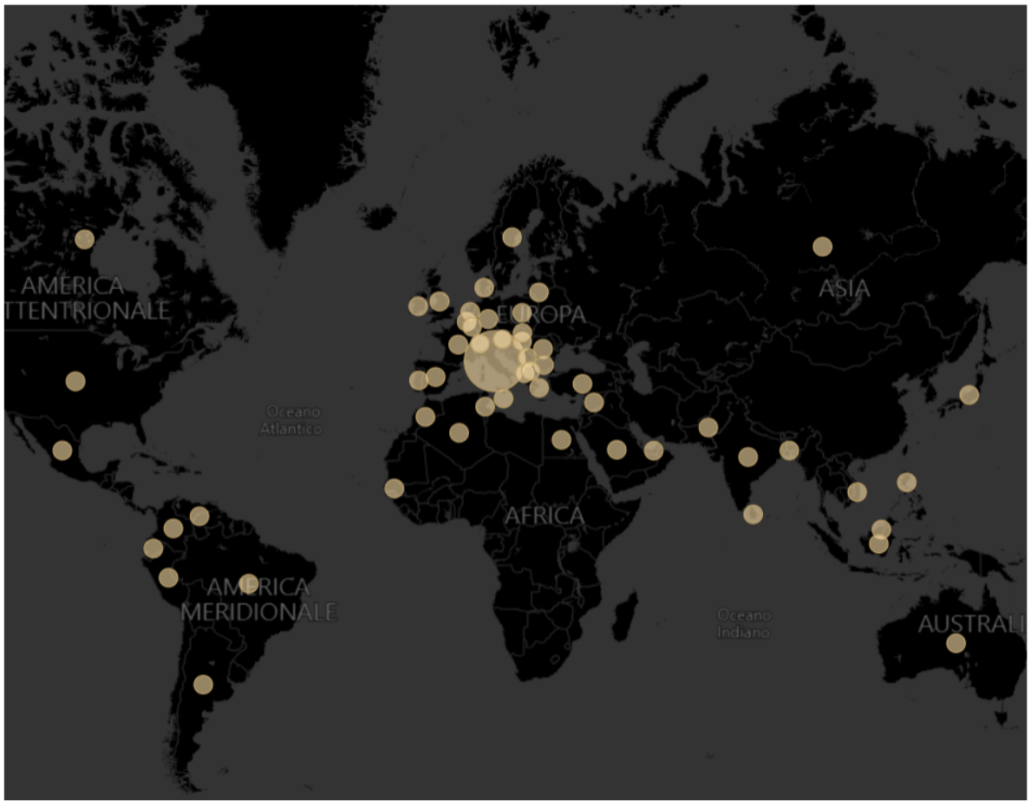

Figure 3.15: Percentage of lifetime likes on average per country in 2020.

In figure 3.16 it is shown the percentage of lifetime likes on average per city. Of course, the city from which the highest number of likes comes is Pisa, with 2.095,57 likes on average, which is strangely followed by other important cities of Italy, like Rome and Milan, that are further from Pisa than, for instance, cities like Livorno, Lucca or Florence (that is at the fourth position).

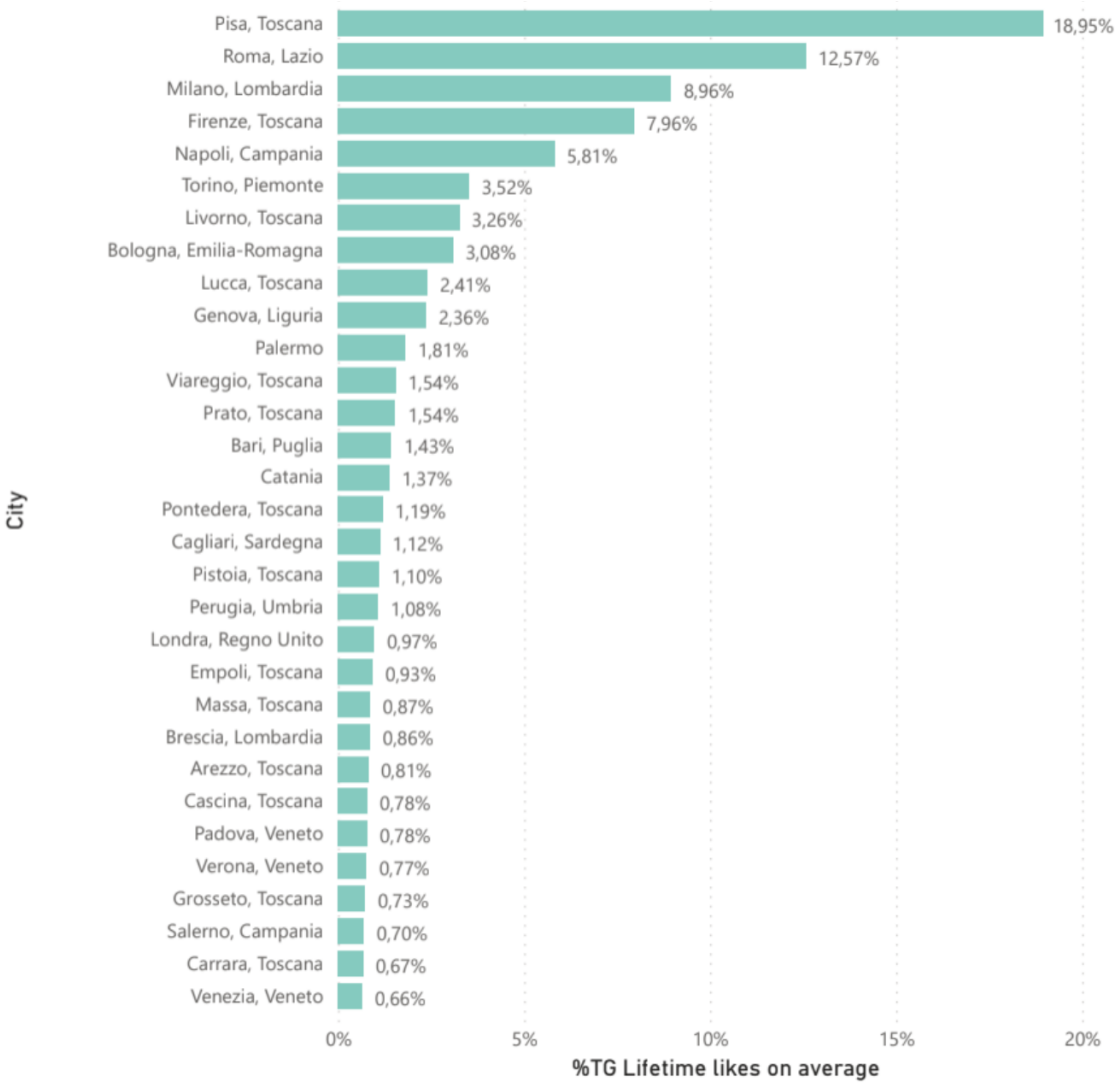

Figure 3.16: Percentage of lifetime likes on average per city in 2020.

In figure 3.17 it is reported the percentage of lifetime likes on average per gender and age. From this analysis we can see that there is a much higher involvement in the Facebook page of the IF in persons of age 25-34, both males and females (with a higher maximum number for the females).

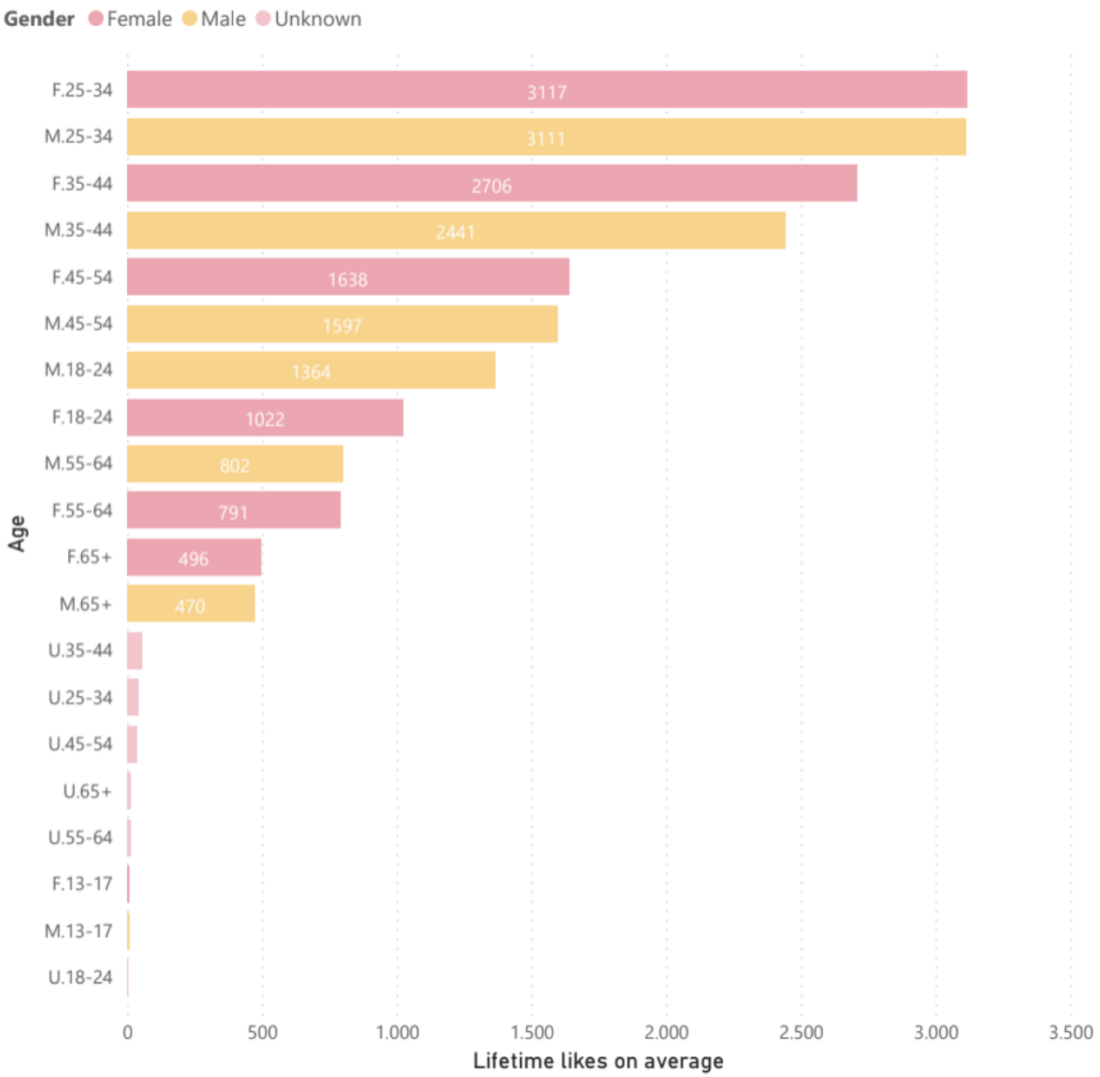

Figure 3.17: Lifetime likes on average per gender and age in 2020.

Twitter

Here an analysis of some of the main interesting metrics (which are the number of interactions, likes, clicks on the profile, retweets and tweets published) registered during the months from September to December 2020 will be explained, in order to see the trend of the Twitter account's popularity (in figure 3.18).

During the first month, the metrics shown are very low, but there is a visible improvement of interactions (that arrive to 131 on the 28th of September) during the last week of September, probably because the beginning of the IF is getting closer.

During October, all the metrics analyzed have a great pick from the 8th to the 11th of October, which is also the period in which the IF is held in the city, so the involvement of the public is much greater than in the other weeks.

The number of tweets published tends to decrease during the month of November, as well as all the other metrics. In the last week of the month instead there are other improvements in the interactions, maybe in relation to other online events that have been shown in that period.

Lastly, in December data is very poor. There are just a few interactions sometimes, but there is not a specific trend, nor some activities from the account, because the IF has ended and there are no more contents published.

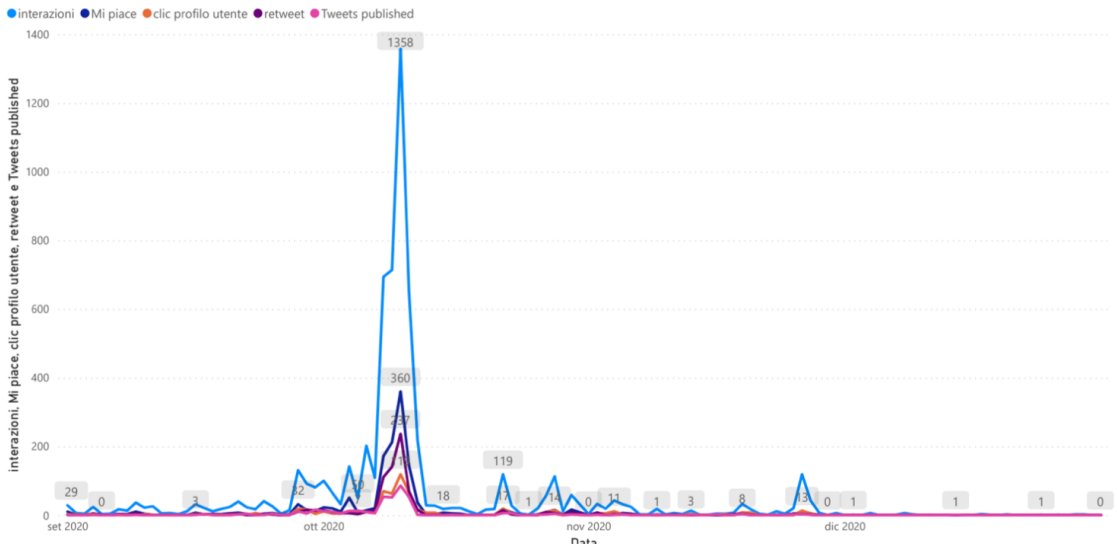

Figure 3.18: Tweets published, clicks on the profile, interactions, retweets and likes from September to December 2020.

Instagram

The information extracted from the Instagram dashboard is not as rich as the one from Facebook or Twitter, but there is still some data which describes some interesting features of the public from this social network in order to make some considerations.

Firstly, it is possible to notice in figure 3.19 that there is a great difference between the quantity of followers on Instagram and likes on Facebook. It is clear that the Facebook

page looks much more popular than the Instagram profile and, since in the last few years young people like students tend to use much more Instagram than Facebook, it may be advisable to increase the contents and the attractiveness of the Instagram profile.

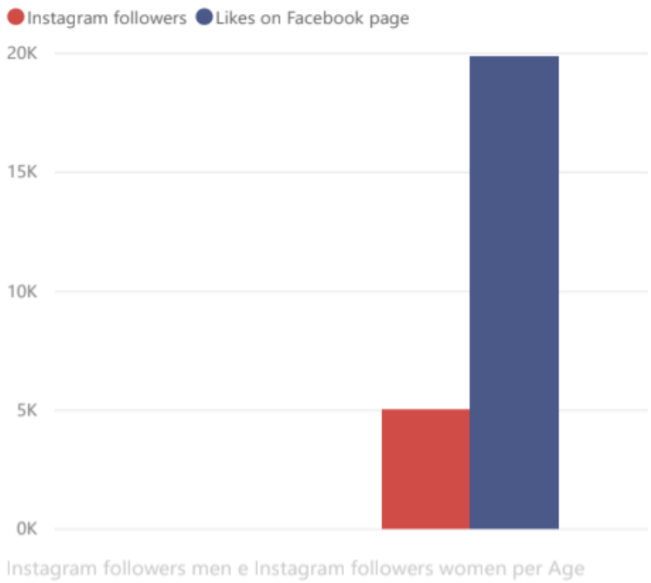

Figure 3.19: Instagram followers VS. likes on Facebook.

In the graph in figure 3.20, the different usage of the Instagram profile by users' gender is reported. Although the gap is not so evident, we can notice that among young users of age 25-34 there are more female followers, while among the older users there is the opposite situation.

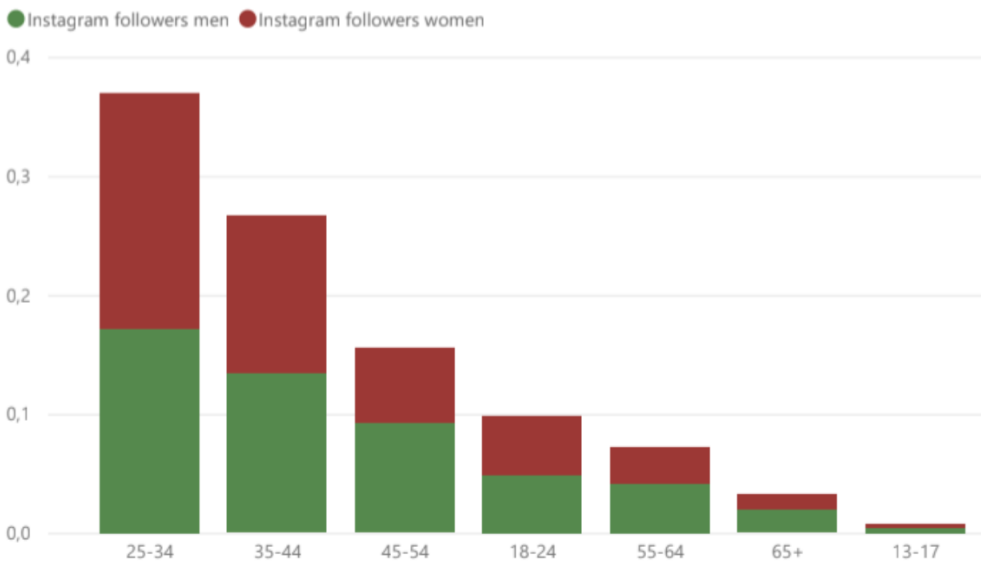

Figure 3.20: Instagram followers per age and gender.

Regarding the Instagram followers per main country in figure 3.21, we only have data about Italy (41,6%), Brazil (1,5%), USA (1,2%), Spain (0,6%) and UK (0,5%). Obviously, the presence of Italian users is dominant over all the other countries.

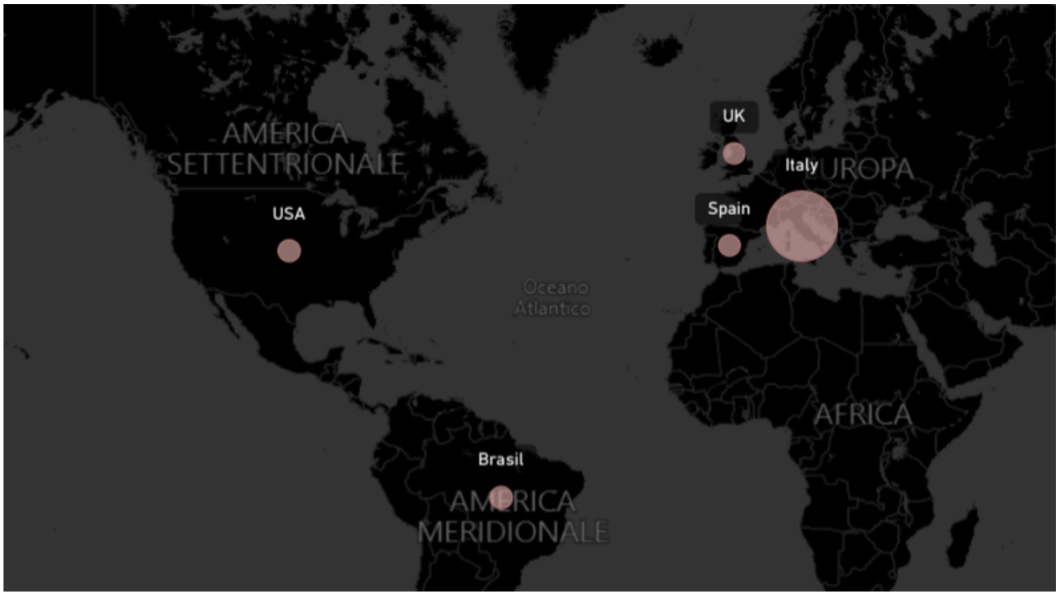

Figure 3.21: Instagram followers per main countries.

Lastly, in figure 3.22 we can look at the cities from which the Instagram followers come: we can notice that, as expected, there is a great number of followers that come from

Pisa (9%). Then Rome follows with 2,30% of users, Florence with 2,20%, Milan with 1,6% and lastly Livorno with 1%.

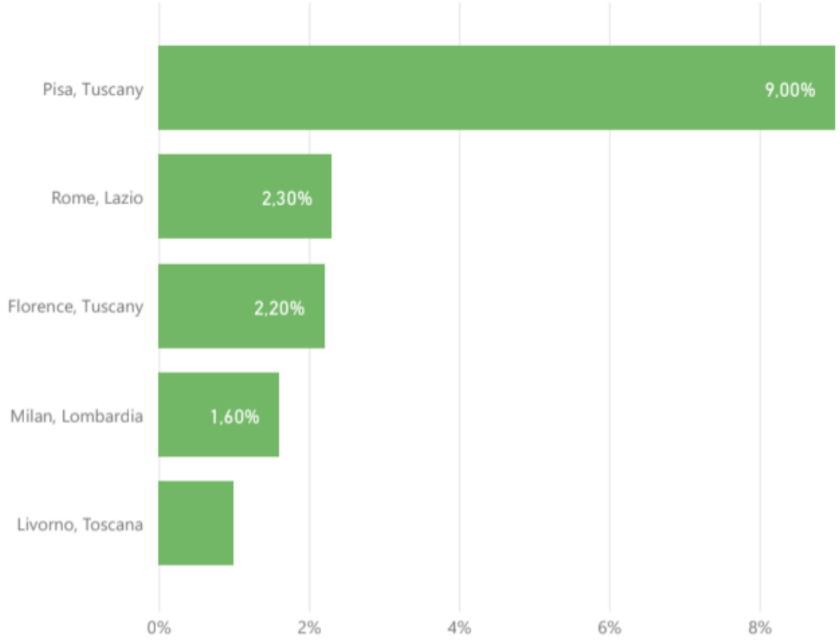

Figure 3.22: Instagram followers per main cities.

YouTube

For what concerns the YouTube channel, it has been possible to collect data about the daily visualizations from 2011 to 2021. As we can see from the first graph in figure 3.23, the number of visualizations have increased a lot from 2015 to 2018, which is the year with the greatest number of visualizations (nearly 118K). In 2020 there is another pick, probably due to the organization of online events because of Covid-19, in which there has been a great usage of YouTube to stream videos and presentations.

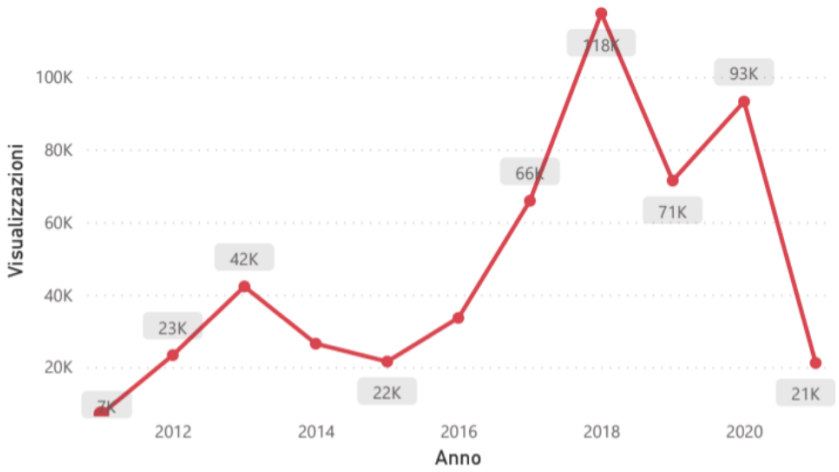

Figure 3.23: YouTube visualizations per year.

In relation to what has been said before, it may be interesting to notice the trend of the year 2020 in figure 3.24: we can see that during the IF (October/November) there is an extremely high pick (which is the highest ever reached from 2011 to 2021 so far) with almost 3000 visualizations on the 11th of October. This is because in those days there were a lot of streaming events shown through the YouTube channel that have been followed by thousands of people.

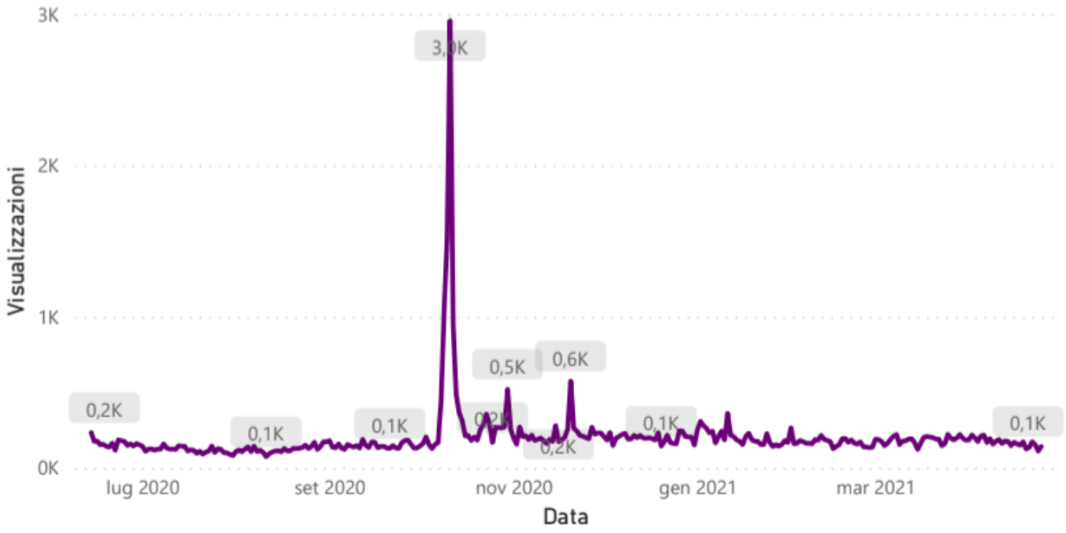

Figure 3.24: YouTube visualizations in 2020.

4. Data Mining Analysis

As it has already been explained in chapter 2 regarding the overview of the project, in this type of analysis a much more invasive approach is used on the extrapolated data, in order to obtain more information that would otherwise be impossible to discover. Through clustering, pattern mining and geohash algorithms it will indeed be possible to study the data to discover hidden trends, behaviors and phenomena that will be helpful to better understand how to organize cultural events based on the results of the past ones.

We decided to study the Tweets produced during the Internet Festival by the people who were in the Pisa area, in order to be able to identify potential groupings (clusters) in some areas of the city and to check if they corresponded to some scheduled events and, consequently, observe the most frequent movements among users to identify preferential routes and patterns that could provide important details for future event organizations.

Also in this case, an overview will be provided on the type of data processed and the ways in which they were extracted (see section 4.1), then the methodology used for this more complex study will be addressed in section 4.2 and finally the results and the various interpretations will be shown in section 4.3.

4.1 Data retrieval

The data analyzed with this method of analysis have been extracted from Twitter APIs, that return Tweets encoded using JavaScript Object Notation (JSON), based on key-value pairs, with named attributes and associated values. These objects all encapsulate core attributes that describe the object: each Tweet has an author, a message, a unique ID, a timestamp of when it was posted, and sometimes geo metadata shared by the user; each User has a Twitter name, an ID, a number of followers, and most often an account bio [**14**].

The following JSON object shows an example of a Tweet structure and some of its features:

```
{
  "created_at": "Thu Apr 06 15:24:15 +0000 2017",
  "id_str": "850006245121695744",
  "text": "1\/ Today we\u2019re sharing our vision for the future
  of the Twitter API platform!\nhttps:\/\/t.co\/XweGngmxlP",
  "user": {
    "id": 2244994945,
    "name": "Twitter Dev",
    "screen_name": "TwitterDev",
    "location": "Internet",
    "url": "https:\/\/dev.twitter.com\/",
    "description": "Your official source for Twitter Platform news,
    updates & events. Need technical help?
    Visit https://twittercommunity.com/#TapIntoTwitter"
  },
  "place": {
  },
  "entities": {
    "hashtags": [
    ],
    "urls": [
      {
        "url": "https://t.co/XweGngmxlP",
        "unwound": {
          "url": "https://cards.twitter.com/cards/18ce53wgo4h/3xo1c",
          "title": "Building the Future of the Twitter API Platform"
        }
      }
    ],
    "user_mentions": [
    ]
  }
```

```
}
```

Considering that for our analysis a large amount of information present in each object is superfluous, in section 4.2 we will illustrate the methodology followed for processing the extracted data and for the following analyses.

4.2 Methodology

Figure 4.1: Workflow followed during the Data Mining Analysis.

The flowchart in figure 4.1 shows the steps of the procedure that was adopted for the analysis with data mining techniques. It is immediately clear that, compared to the Business Intelligence analysis methodology, it has a more articulated structure, since it involves more steps.

Also in this case, the first phase involves Data Retrieval, already explained in the previous section, which provided us with a large amount of compressed data in JSON format. The other steps apply the Data Preprocessing, essential for preparing the data in anticipation of the analyzes, the Data Understanding to know fundamental information about the data, the Geohash algorithm application and the Clustering analysis, the creation of an ODMatrix and lastly the Pattern Mining analysis.

Since the data retrieved is usually in a raw status, there is a need for processing it in order to make it suitable for the analysis and because a focus on improving data quality typically improves also the quality of the results: here the **Data Preprocessing** step comes in. In section 4.1 we learned that we have obtained a fair amount of data that presents a lot of useless information, and therefore it is necessary to eliminate it to address our study in a more targeted way: the first step involves skimming the attributes of the collected

Tweet objects, eliminating those that are not of our interest and saving only those that can provide us with information mainly on movements and other important facts, namely:

- *Text*: represents the text of the Tweet, i.e. the actual UTF-8 text of the status update;
- *ID*: represents the unique identifier of the Tweet;
- *Coordinates*: represents the geographic location of this Tweet as reported by the user or client application;
- *User*: contains information on the user, including his identification;
- *Created_at*: UTC time when this Tweet was created;
- *Place*: when present, indicates that the tweet is associated (but not necessarily originating from) a Place and Places are specific, named locations with corresponding geo coordinates.

Subsequently, we move on to a further step of screening the data, to exclude all tweets that are not located within the Pisa area, since we are interested in finding out if there are phenomena of particular importance among the visitors who were in Pisa when the festival was in progress. The area that we selected is a polygon in which the whole city of Pisa is included, as it is roughly shown in figure 4.2, having for coordinates (10.4269000 43.7359000), (10.3693832 43.7353212), (10.3684834 43.6955374), (10.4288578 43.6955193) and (10.4269000 43.7359000).

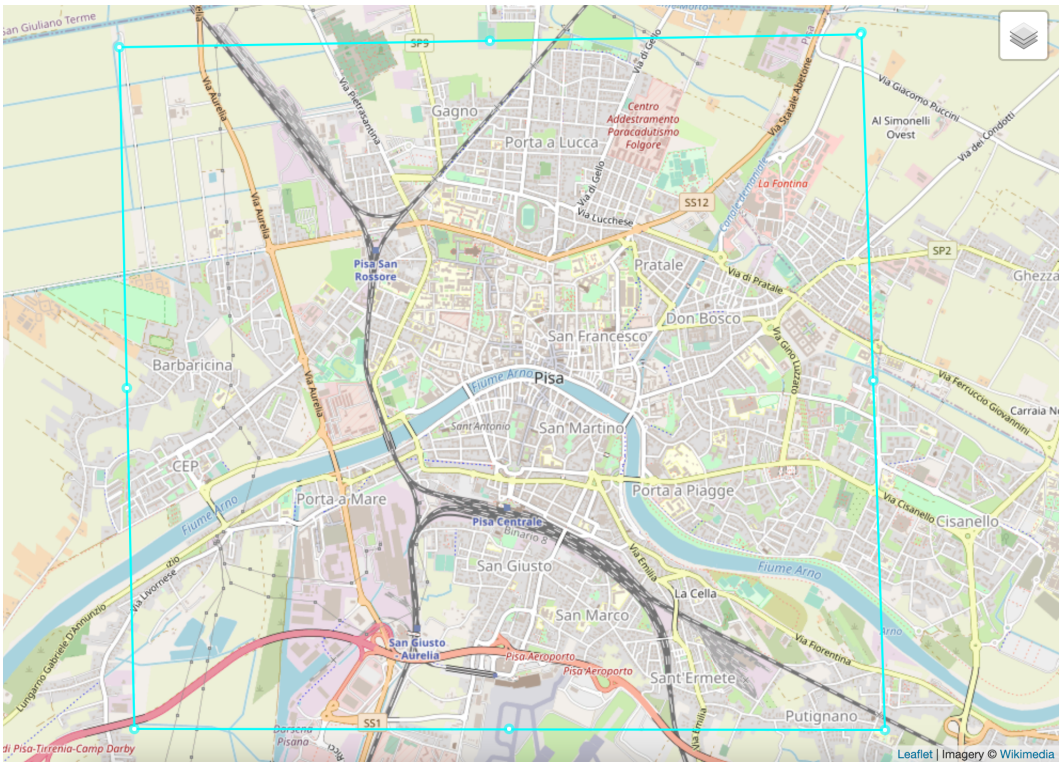

Figure 4.2: Polygon used as area to filter the Tweets.

It must be specified that all the visualizations on the geographical map of Pisa have been made with the use of the **Folium library**[10], which makes it easy to visualize the data that has been manipulated in Python on an interactive map. It allows both data binding to a map for choropleth views and the passing of advanced vector/raster/HTML views as markers on the map.

At this point, it is possible to collect all the skimmed data in a single dataset, in order to manage them more comfortably and to be able to move on to the Data Understanding phase, in which they can be studied better to decide how to deal with the next steps.

Eventually, the created dataset is composed of the features *Screen_name* (i.e. the user's username), *UserID*, *TweetID*, *Coords* (tweet coordinates), *Lat* (latitude only), *Lon* (longitude only), *Created_At* (when it is been published) and *Text* (tweet text). In the following figure it is possible to see a part of the dataset.

[10]To learn more about this topic, visit the web page at the following link: `https://python-visualization.github.io/folium/`

	Screen_name	UserID	TweetID	Coords	Lat	Lon	Created_At	Text
0	madikeeper12	868809325	779072240994234368	[43.72666207, 10.41268069]	43.726662	10.412681	2016-09-22 21:37:51+00:00	Cieli infuocati.\n\n#picoftheday #quotesofthed...
1	madikeeper12	868809325	781615843406819329	[43.72666207, 10.41268069]	43.726662	10.412681	2016-09-29 22:05:13+00:00	Prospettive.. \nunite a casa #ilselfone\n#team...
2	madikeeper12	868809325	781870800156499968	[43.72666207, 10.41268069]	43.726662	10.412681	2016-09-30 14:58:19+00:00	Non occorre essere matti per lavorare qui, ma ...
3	madikeeper12	868809325	780003801260404736	[43.7167, 10.3833]	43.716700	10.383300	2016-09-25 11:19:32+00:00	RunOnSunDay 🏃‍♂️💥\n#run #running #runner #ni...
4	madikeeper12	868809325	779443101123260417	[43.70561, 10.42059]	43.705610	10.420590	2016-09-23 22:11:31+00:00	La vita è come la fotografia sono necessari i ...
...	...	...	...	...	...	...	...	...
656	antoniocassisa	358042635	781879291911016448	[43.7167, 10.3833]	43.716700	10.383300	2016-09-30 15:32:04+00:00	I mì ómini \n#son #figli #boys @ Pisa, Italy h...
657	SefaMermer	293157588	780753755830677504	[43.7167, 10.3833]	43.716700	10.383300	2016-09-27 12:59:35+00:00	#love #tbt #tagforlikes #TFLers #tweegram #pho...
658	SefaMermer	293157588	780756143668953088	[43.7167, 10.3833]	43.716700	10.383300	2016-09-27 13:09:05+00:00	#love #tbt #tagforlikes #TFLers #tweegram #pho...
659	matteluca89	494389053	779638196258811904	[43.71544235, 10.40051616]	43.715442	10.400516	2016-09-24 11:06:45+00:00	Last saturday I went out with my #chinese teac...
660	anabrmotta	98254561	781123690343698432	[43.72263, 10.3948]	43.722630	10.394800	2016-09-28 13:29:35+00:00	Já que é pra tombar, ela tombou (só um pouquin...

Figure 4.3: Dataframe created after cleaning and filtering the JSON data from Twitter APIs.

After making our data readable and ready for any analysis, it is advisable to move on to the **Data Understanding** phase, as it is essential to understand whether the data in our possession is representative or not for the purposes of this study. We chose to observe characteristics such as the frequency of Tweets per user and per day, in order to know how informative the data was.

Regarding the next part of the study, i.e. the application of **Clustering** algorithms for grouping data based on certain characteristics, the methodology that has been chosen to adopt involves the use of the DBSCAN, OPTICS and Bisecting K-Means algorithms. The reasons for these choices can be traced back to the fact that the different implementations of these algorithms could be better adapted to the geo-spatial type of the data that we are treating. After the preparation and filtering phase, the remaining amount of data for conducting an analysis that requires a certain level of detail is usually relatively small, depending on the filters and the initial data dimension. Indeed, choosing a good algorithm to carry out a clustering operation demands several tests, changing and adapting from time to time the parameters to obtain accurate results and this is what happened during the algorithms applications.

With the aim of fully understand the results obtained, it was decided to draw up a table to compare the evaluation metrics of the clustering algorithms for choosing the two most

promising for the analysis. The metrics considered for the evaluation are the following:

- The number of clusters produced;
- The percentage of outliers generated;
- The average and median of the size of the clusters;
- The size of the largest cluster;
- The *Silhouette* coefficient, which is a score used to calculate the goodness of a clustering technique in a value range from -1 to 1.

The other interesting method tried for grouping the visitors is the **Geohash**. As it has already been explained in the section where it was presented, this method involves the definition of a precision parameter for dividing the geographical area in cells: for the size of the city of Pisa, the most suitable value turned out to be 6. Later on, the visualization of the division will be shown on a Folium Map.

Another important factor for the study that can be extrapolated from the data is the movement of the IF's visitors within the area previously selected. In this way it is possible to visualize which are the most popular routes and, consequently, also to understand how to better organize the future events to make them reachable and accessible to the majority of visitors. One way to view the most frequent movements is through a **ODMatrix**, having as columns the clusters of origin of the movements and as rows the destinations. It is implemented according to a simple logic: each time a movement is recorded from one cluster to another, the cell of the matrix describing that movement will have an incremented value of one unit. For its realization it was first necessary to group the data by user, in order to have a list containing all the movements made by each visitor. Subsequently, the lists obtained were ordered temporally, so as to have a temporal continuity in order to accurately establish the paths of the users.

If we had had a greater amount of data, a further step would have been to eliminate "non-movements", i.e. those cases in which users have tweeted several times in the same place, without making a real movement. In our case, however, it would have reached a too minimal level of detail, which would have excluded a large portion of data. On the

other hand, knowing that users have been in the same place for a long time can also be informative, so taking into account the "non-movements" can also reveal other interesting aspects in users' behavior.

Eventually, an empty source-destination matrix was created and, by scrolling the user lists containing their respective movements in pairs, it was updated according to the logic explained above. Then, a Heat Map[11] was used to display the matrix.

Having reached this point of the methodology followed, all that remains is to move on to the **Pattern Mining** analysis, to understand which are the most traveled paths formed by several movements. The application of this analysis was carried out using the PrefixSpan algorithm, trying to extract the patterns with a frequency of at least 3, meaning that it returns itemsets of clusters labels that appear at least 3 times among the users' movements. To make the results visible and understandable, the most relevant movements were shown by drawing arrows on a Folium map, which indicate the most traveled paths by the users.

4.3 Results

It is necessary to state that, although during the execution of the methodology it emerged several times that the data were not completely suitable for the purpose of the analysis (both for quantity and quality), it was preferred to continue in order to provide an analysis prototype that can be applied to other data sets.

In the following sections the results of all the steps of the methodology described above will be shown.

4.3.1 Data Understanding

As a first step, we can see a brief description of some of the main features of the dataframe obtained in the previous phase:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 661 entries, 0 to 660
```

[11]The Heat Map is a graphical representation of data where the individual values contained in a matrix are represented by colors.

```
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Screen_name  661 non-null    object
 1   UserID       661 non-null    int64
 2   TweetID      661 non-null    int64
 3   Coords       661 non-null    object
 4   Lat          661 non-null    float64
 5   Lon          661 non-null    float64
 6   Created_At   661 non-null    datetime64[ns, UTC]
 7   Text         661 non-null    object
dtypes: datetime64[ns, UTC](1), float64(2), int64(2), object(3)
```

It is noted that this is a relatively small dataframe, which contains the 8 columns, which we selected in the preprocessing phase to describe the most useful information for the study, and 661 records. We quickly discover that in the initial skimming phase a large amount of data was taken away and, for a more accurate and reliable analysis, it would be preferable to use larger datasets.

Having examined the characteristics of the dataframe, it is possible to move on to a deeper knowledge of the data we are examining: we want to observe the distribution of the frequency of Tweets by user.

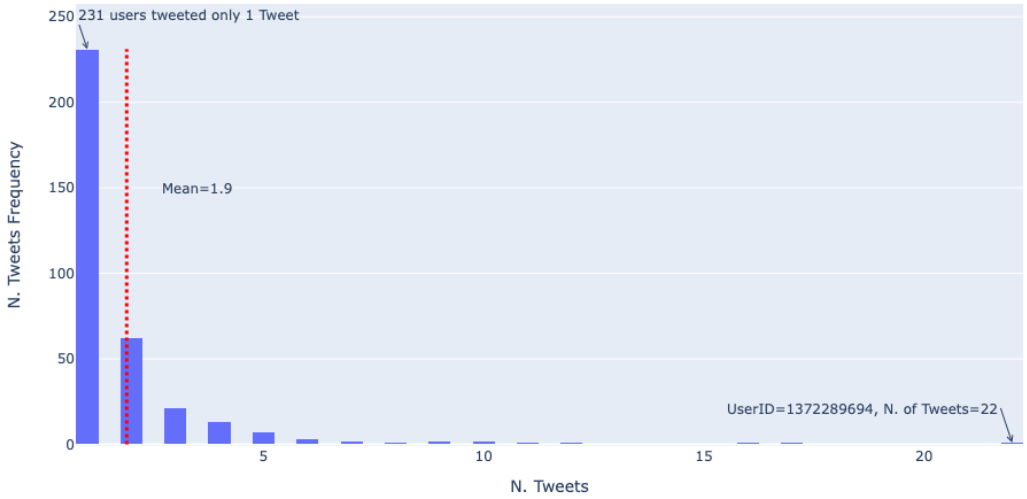

Figure 4.4: Number of Tweets Frequency Histogram.

The histogram in figure 4.6 shows the zipfian distribution[12] of the frequency of the number of Tweets posted by users: it is possible to observe that among the 661 records registered in the dataframe:

- **231** is the number of users who have tweeted only one Tweet;
- **22** is the highest number of Tweets that have been posted by a single user;
- **1.9** is the average number of Tweets published per user, i.e. about 2 Tweets per user.

It can be noted right away that the analyzes that will be carried out will not provide very generalized phenomena and, if so, it will be the result of a poor quality and quantity of data.

At this point, we can study the amount of Tweets posted per day during the festival, to see how informative the data can be if divided over time.

[12]Zipf's Law is a statistical distribution in certain datasets in which the frequencies of some items are inversely proportional to their ranks. It was originally formulated in terms of quantitative linguistics.

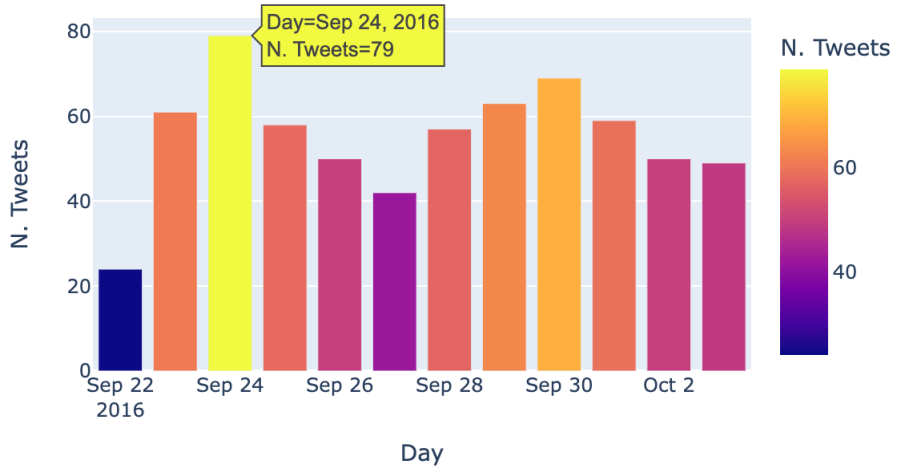

Figure 4.5: Number of Tweets per day.

We notice that the day the highest number of Tweets occurred, i.e. 79, was the 24th of September and the day with the second highest number was the 30th of the same month.

Another interesting factor can be observed in the users with the highest number of Tweets per day: in this case, the user who tweeted the most was with 22 Tweets on the 30th September (see figure 4.6).

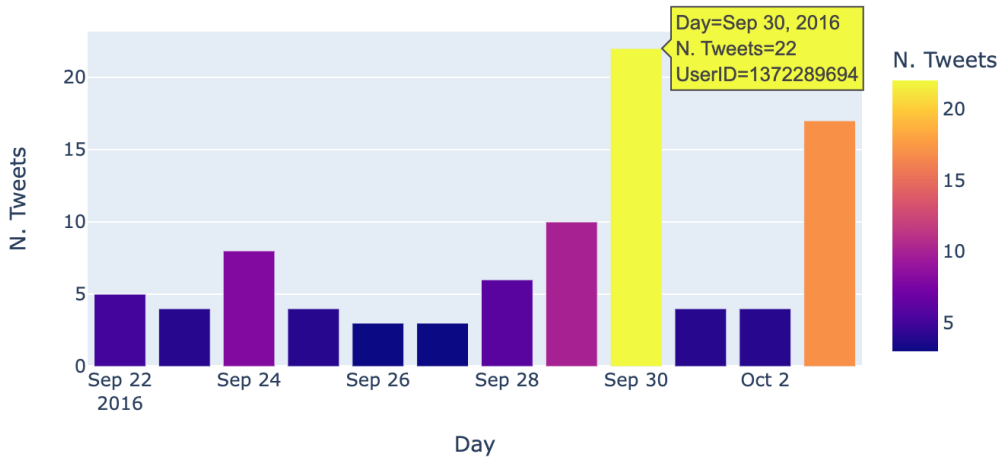

Figure 4.6: Users with the highest number of Tweets per day.

Now that we have studied some interesting details about the data we have, we can proceed with the analysis. As a first view of the quality of the data, we cannot consider

ourselves fully satisfied, so we should expect that subsequent analyzes will not be perfectly representative of our aims.

4.3.2 Geohash

To use a non-data-driven method of zoning the city, geohash encoding with a percision of 6 was also used, as it was the most appropriate value for a balanced division, which corresponds to approximately ± 0.61 kilometers per cell into which to divide the city's area. Then, starting from the database also used for the clustering phase, which contains the latitudes and longitudes of each Tweet, the corresponding geohash code for that geographical location was generated and added as an additional feature in the database. In order to view the grid, the map with the folium library was also used in this case, for which it was necessary to add an extra step to create a dictionary containing several parameters, shown in the following example:

```
'id': '0',
'type': 'Feature',
'properties': {'geohash': 'spz2ub',
               'location': '[43.72666207, 10.41268069]',
               'value': 1},
 'geometry': {'type': 'Polygon',
               'coordinates': [[[10.404052734375, 43.7255859375],
               [10.4150390625, 43.7255859375],
               [10.4150390625, 43.7310791015625],
               [10.404052734375, 43.7310791015625],
               [10.404052734375, 43.7255859375]]]}
```

As you can see, there are characteristics such as the *ID* and the *type* of the record and then the most useful ones for the realization of the grid, that is the *properties*, which contain the geohash code (in this case '*spz2ub*'), the *location* in coordinates and a *value* to determine how many times that location occurs in the database (in this case 1) and finally the *geometry*, which incorporates the characteristics of the data type, that is a polygon,

and the coordinates that describe it. All this additional information was transposed into a database (see figure 4.8).

	location	value	geohash	geometry
0	[43.72666207, 10.41268069]	1	spz2ub	POLYGON ((10.40405 43.72559, 10.41504 43.72559...
1	[43.72666207, 10.41268069]	5	spz2ub	POLYGON ((10.40405 43.72559, 10.41504 43.72559...
2	[43.72666207, 10.41268069]	5	spz2ub	POLYGON ((10.40405 43.72559, 10.41504 43.72559...
3	[43.7167, 10.3833]	5	spz2sq	POLYGON ((10.38208 43.71460, 10.39307 43.71460...
4	[43.70561, 10.42059]	5	spz2th	POLYGON ((10.41504 43.70361, 10.42603 43.70361...

Figure 4.7: Dataframe created for the map with the geohash codes subdivision.

The final result of this process is shown in figure 4.8, where you can clearly see the grid generated by the geohash codes.

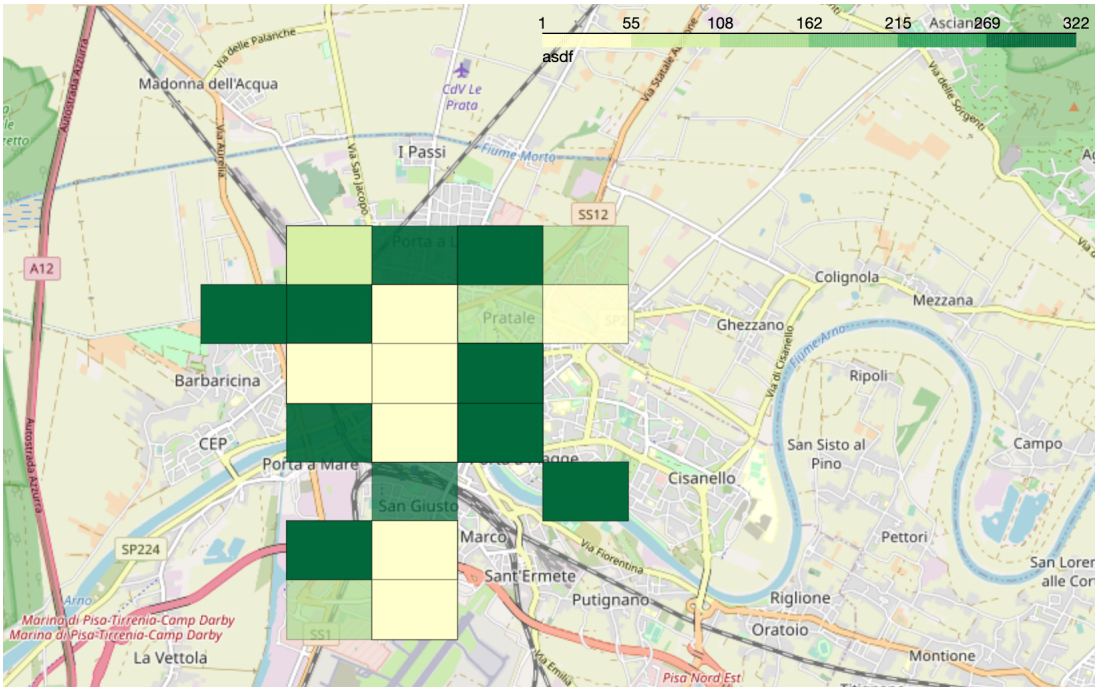

Figure 4.8: Map with the geohash subdivision.

Although the use of Geohash encoding can be an excellent alternative to split in cells geographical areas, in reality we will observe that the application of a clustering algorithm represents the most appropriate solution, both because the geographical scale in relation to the geohash precision value is too large to be able to identify all the groups in detail,

both because it does not carry out a targeted analysis, but rather a random one.

4.3.3 Clustering

In order to start the clustering analysis we need to standardize data. Since it is necessary for the data to have the same scale to avoid bias in the outcome, data standardization is used in machine learning to make model training less sensitive to the scale of features, which in our case allows the algorithm to converge to better weights and leads to a more accurate analysis.

Firstly, we create a smaller dataset from the original one, that contains only the features that we need to analyze, i.e. latitude and longitude. Afterwards, it is possible to proceed with the standardization applying **StandardScaler**[13], which produces a new dataframe composed by values scaled to unit variance.

DBSCAN

The fist attempt of the clustering analysis is made with the DBSCAN algorithm presented in section 2.2.2. With the aim of finding an appropriate set of parameters, we use the *Knee Method* with nearest neighbor distances: to determine the best epsilon value, we calculate the average distance between each point and its closest/nearest neighbors, then we outline a k-distance plot and choose the epsilon value at the "elbow" of the graph. On the y-axis, we plot the average distances and the x-axis all the data points in your dataset.

From the plot in figure 4.9, it seems that the best epsilon value may be around 0.2.

[13] StandardScaler is a scikit-learn method to preprocess data for machine learning and it standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation.

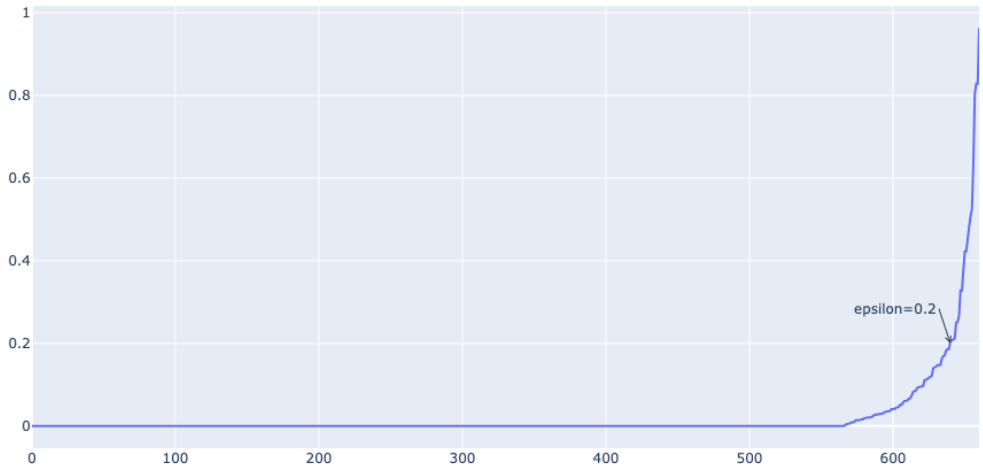

Figure 4.9: Knee method with Nearest Neighbors to find the best epsilon.

According to this estimation, the parameters set for the first analysis are *epsilon*=0.2, *min_samples*=5 and the default distance metric=euclidean. In table 4.1 the metrics' values obtained after evaluating the results are shown.

DBSCAN 1 metrics	
Parameters	eps=0.2, min_samples=5
N. Clusters	15
Biggest cluster dimension	308
Silhouette	0.6883002735680692
Mean cluster dimension	47
Median cluster dimension	13.0
% Outliers	10.287443267776096

Table 4.1: Metrics from evaluation of DBSCAN 1.

As a first attempt, it is by no means adequate to the level of capillarity necessary to find groups of visitors to the events of the IF. Therefore, it is possible to see that the clusters on the Folium map are too confused and mixed to be able to identify clear and precise divisions (see figure 4.10).

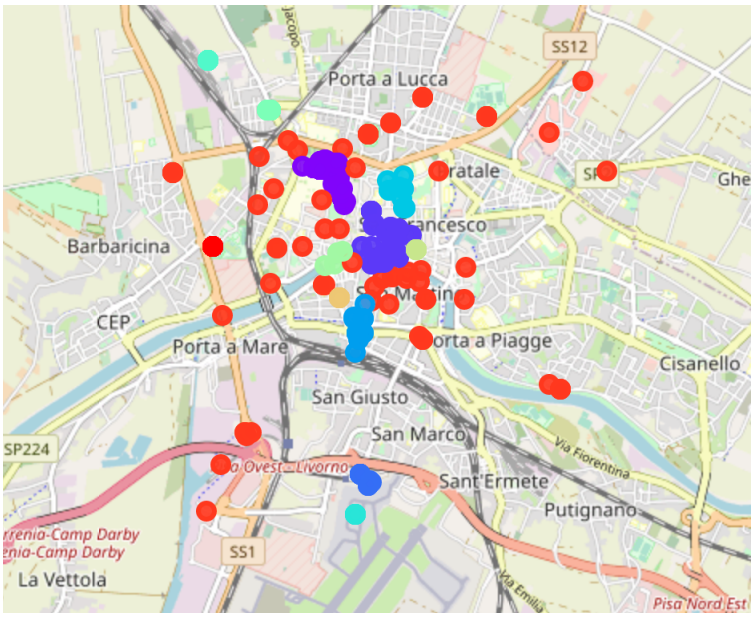

Figure 4.10: Application of DBSCAN with epsilon=0.2 and min_samples=5.

It is indeed necessary to try other parameters: let us decrease *min_samples* to 2 and see the consequent evaluation and visualization.

DBSCAN 2 metrics	
Parameters	eps=0.2, min_samples=2
N. Clusters	29
Biggest cluster dimension	308
Silhouette	0.7378839468500435
Mean cluster dimension	23
Median cluster dimension	4.0
% Outliers	3.177004538577912

Table 4.2: Metrics from evaluation of DBSCAN 2.

The resulting metrics in 4.2 from this latest application have greatly improved, as it is clear also from the map in figure 4.11. Indeed, we see that the number of clusters has increased from 15 to 29, with a much higher level of detail, just as the percentage of outliers has decreased from 10.28 to 3.17 and the silhouette score from 0.68 to 0.73. On the other hand, for what concerns the biggest cluster size, there may still be improvements.

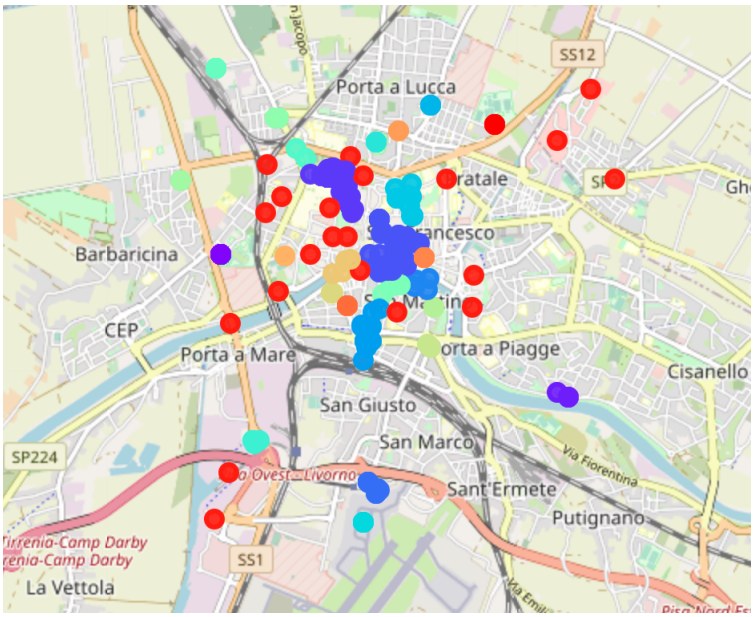

Figure 4.11: Application of DBSCAN with epsilon=0.2 and min_samples=2.

OPTICS

Let us now pass to the application of the OPTICS algorithm. We run the first test with min_samples=5, as done previously with DBSCAN, and default distance metric=minkowski. In table 4.3 it is possible to look at the evaluation metrics obtained.

OPTICS 1 metrics	
Parameters	min_samples=5
N. Clusters	28
Biggest cluster dimension	163
Silhouette	0.7110687027019293
Mean cluster dimension	24
Median cluster dimension	9
% Outliers	12.708018154311649

Table 4.3: Metrics from evaluation of OPTICS 1.

If we compare the number of clusters with the same metric from the first application of DBSCAN, we can notice a great improvement, since the division in clusters of our

data is more defined. Regarding the percentage of outliers there is still a lot of room for improvement, but if we focus on the other metrics, namely the biggest cluster dimension, the mean and the median cluster dimension we can see a clear improvement, while the silhouette is almost very similar. In figure 4.12 you can see the OPTICS division on the map.

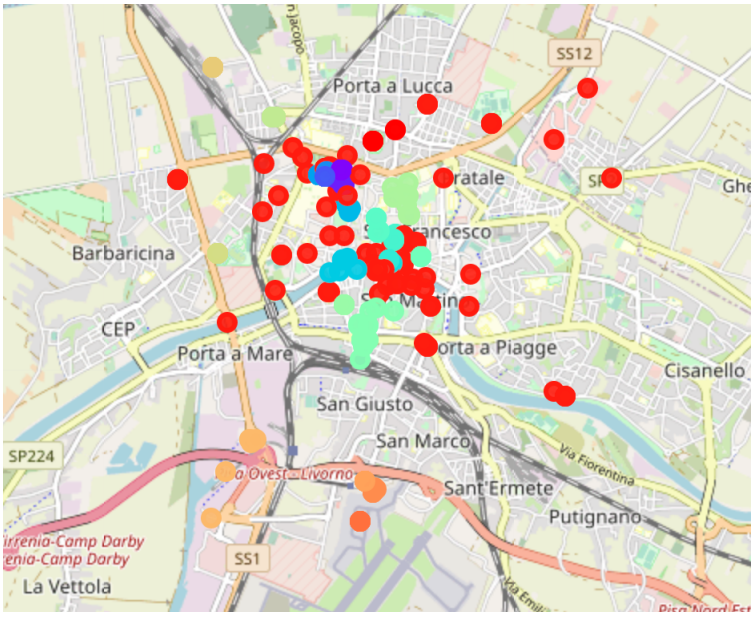

Figure 4.12: Application of OPTICS with min_samples=5.

Clusters definition is still approximative, so, in order to improve the overall results, we notice that we need a smaller number of samples in a neighborhood for a point to be considered as a core point. Therefore, we try the same algorithm changing parameters: min_spamples=3 and min_cluster_size=2, which is the minimum number of samples in an OPTICS cluster, while the default distance metric is still the minkowski.

OPTICS 2 metrics	
Parameters	min_samples=3, min_cluster_size=2
N. Clusters	49
Biggest cluster dimension	163
Silhouette	0.8040413741648764
Mean cluster dimension	13
Median cluster dimension	4.5
% Outliers	7.866868381240544

Table 4.4: Metrics from evaluation of OPTICS 2.

Here we can discern that the number of clusters has increased from 28 to 49, meaning that the precision has clearly improved and that the parameters are more suitable for an analysis of this type, that is, in a relatively small space and which contains a vast constellation of small clusters. In figure 4.13 the map with the clustering applied is shown.

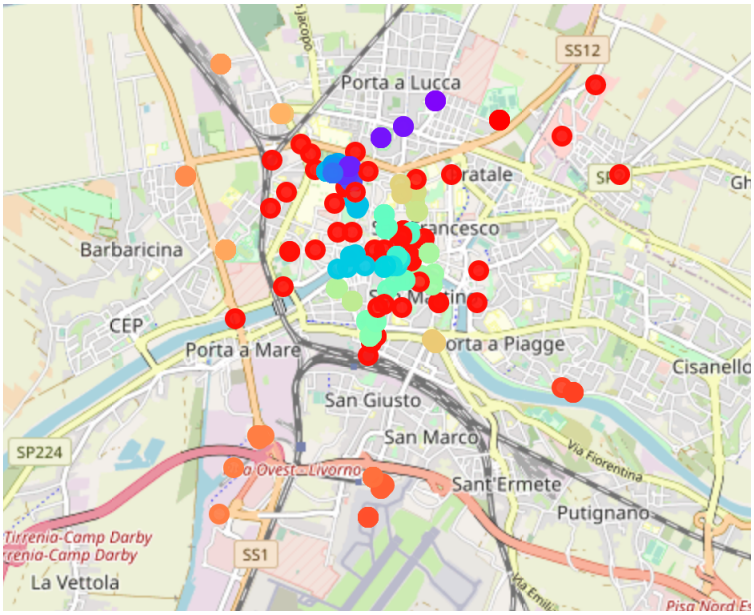

Figure 4.13: Application of OPTICS with min_samples=3 and min_cluster_size=2.

Bisecting K-Means

A Bisecting K-Means algorithm based on our needs has been written following the basic idea of splitting the set of points into two clusters, select one of these clusters to split and continuing until K clusters have been produced. The parameters used have been initialized in the following way:

- K=2, that is the number of clusters in which the set is split;
- min_cluster_size=1, that is the minimum size of a cluster;
- min_split_size=3, that is the minimum size of a cluster to be split;
- max_distance_thr=100/1000 (kilometers), that is the maximum distance to consider for a point to be in a cluster;
- max_iter=300, that is the maximum number of iterations of the splitting procedure;
- metric="haversine", that is the metric used to calculate the distance among the data points.

The results of the application of this algorithm using the parameters shown are reported in table 4.5.

Bisecting K-Means metrics	
N. Clusters	53
Biggest cluster dimension	163
Silhouette	0.803241569685739
Mean cluster dimension	12
Median cluster dimension	4
% Outliers	2.4205748865355523

Table 4.5: Metrics from evaluation of Bisecting K-Means.

Having used parameters that allow a very detailed subdivision, we have obtained a much more defined capillarity (as it is possible to state observing the clustering division

on the map in figure 4.14), which can be seen from the percentage of outliers, which has dropped to 2.4; obviously also the number of total clusters produced has slightly increased, thus including other data that had previously been left out.

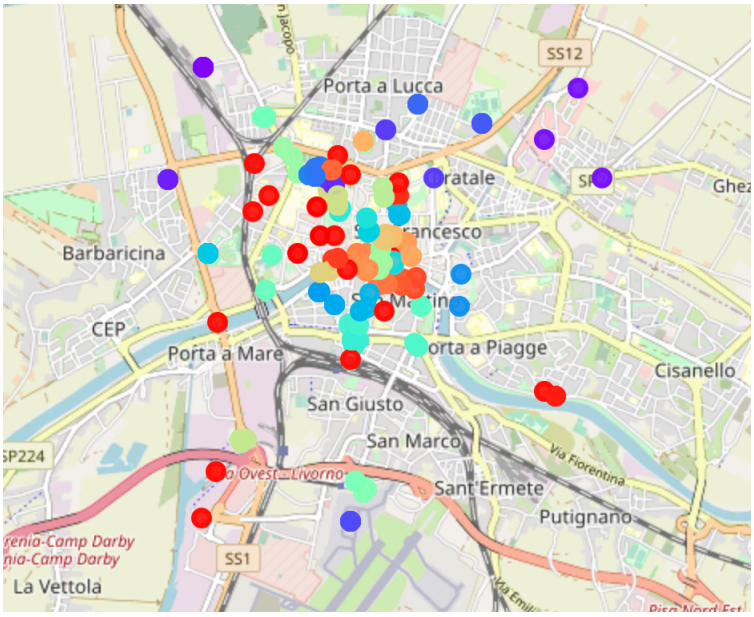

Figure 4.14: Application of Bisecting K-Means.

The other metrics, in general, have remained fairly stable compared to the last analysis which presented improvements, so we can consider ourselves satisfied with this first application and continue with the final evaluation of the best results obtained through the clustering analysis.

Clustering Algorithms comparison

In the table above, all the results of the applications of the clustering algorithms, which we have already seen previously, have been grouped in order to be able to evaluate them as a whole. Among all, we can observe that the two best sets of results were recorded in the case of using Optics with the second set of parameters tested and Bisecting K-means, since both have a very precise and fundamental granularity for the type of study and also a smaller percentage of outliers and a fairly balanced cluster size. For these reasons we decided to continue the analysis in a bifurcated way, on the one hand with the subdivision generated by Optics 2 and on the other hand with that produced by Bisecting K-Means.

Algorithm	N. Clusters	Perc. Outliers	Mean (cluster dimension)	Median (cluster dimension)	Biggest Cluster Dimension	Silhouette
Dbscan1	15	10.29	47	13	308	0.69
Dbscan2	29	3.18	23	4	308	0.73
Optics1	28	12.7	24	9	163	0.71
Optics2	49	7.87	13	4.5	163	0.8
Bisecting K-Means	53	2.4	12	4	163	0.8

Table 4.6: Clustering comparison.

4.3.4 ODMatrix

To observe the results of the project section on the creation of an ODMatrix, it was chosen to use a Heat Map, able to show the various displacements with a more or less intense color gradation: in the cells in which more movements are reported, the color will be more intense, while in those with less or without movements the color will be more tenuous or absent.

The matrix produced by the clustering application of Optics 2 (see section 4.3.3) is shown in figure 4.15.

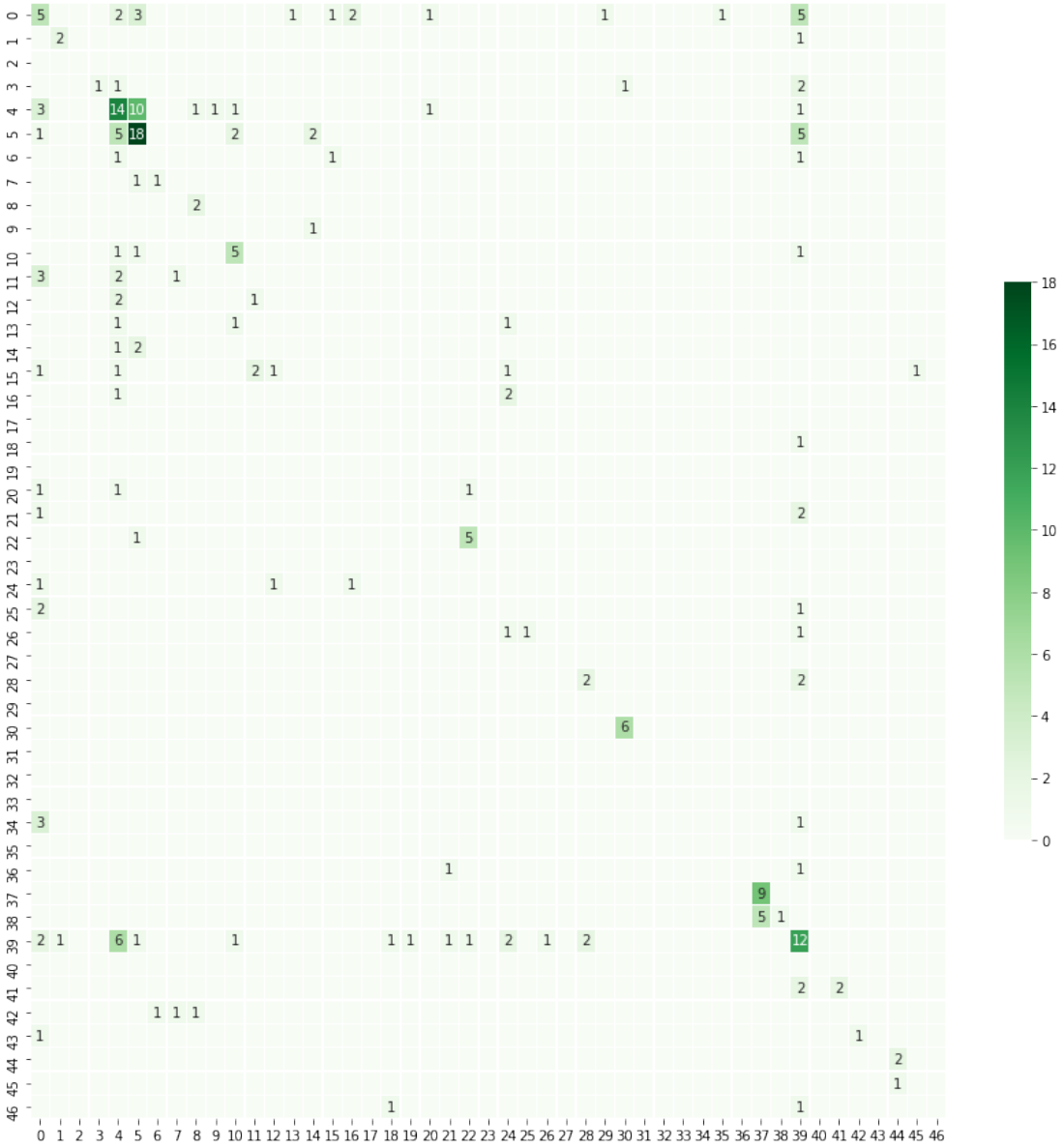

Figure 4.15: Heat Map showing the matrix realized from the Optics 2 clustering application.

It is easy to observe that in all the possible combinable displacements among the 46 clusters produced by Optics 2, there are few that have a significant number to take into consideration.

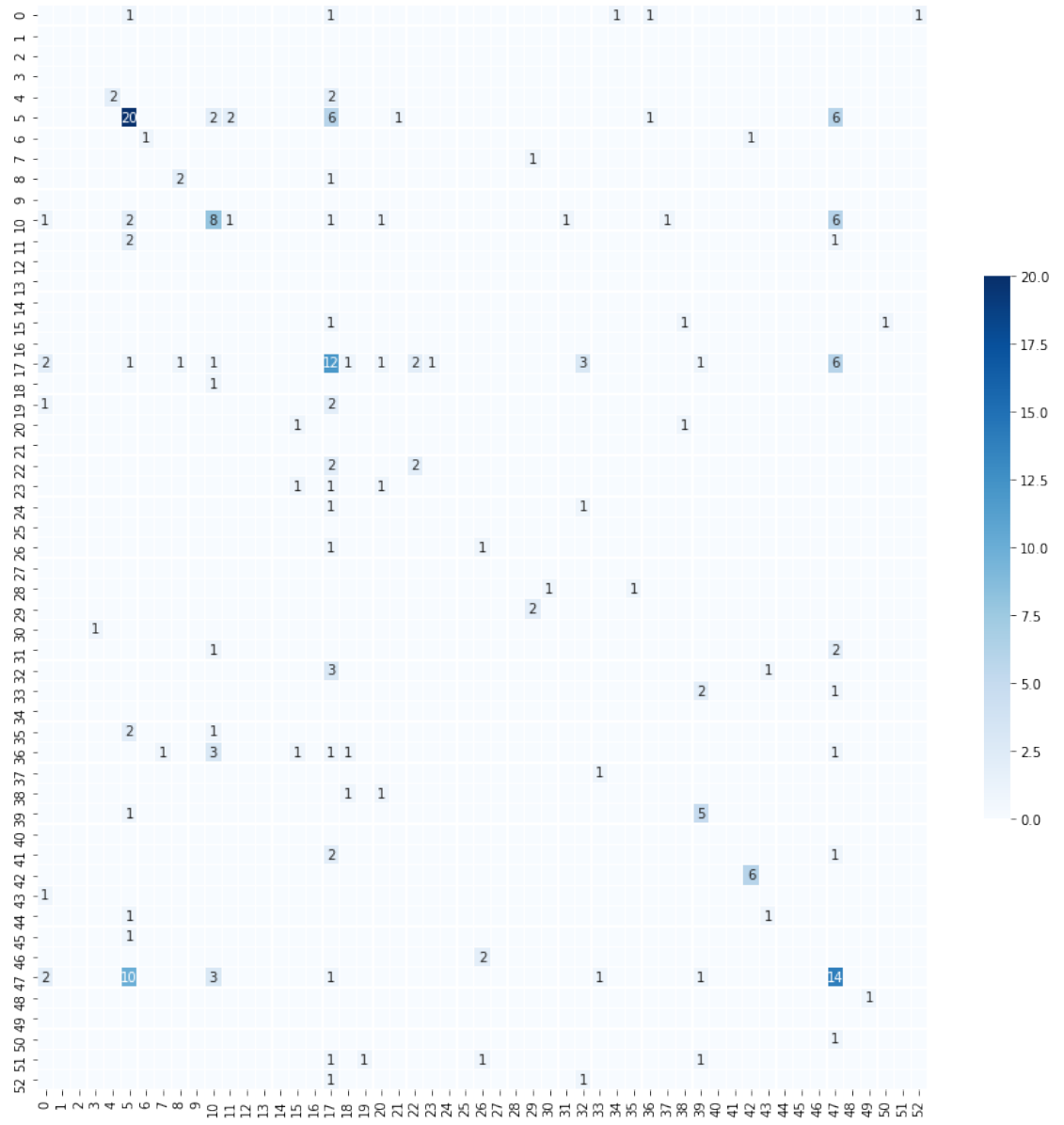

Figure 4.16: Heat Map showing the matrix realized from the Bisecting K-Means clustering application.

In figure 4.16 it is possible instead to notice the Heat Map produced by the Bisecting K-Means application, which presents a result very similar to the one shown previously: there are few points of interest and for the most part they are those in which a “non-movement” occurred.

4.3.5 Pattern Mining

To apply the PrefixSpan algorithm, it is necessary to have a dataframe with the labels obtained from clustering analysis, because we want to see which are the clusters most visited by festival visitors. For this reason, also in this case we perform pattern mining both on the dataframe obtained by adding the Optics 2 labels, and on the one with the Bisecting K-Means labels.

In the first case, the results from applying PrefixSpan with frequency=3 on Optics 2 labels are the following:

```
[(7, [38, 38, 38]),
 (7, [3, 3, 3]),
 (4, [38, 38, 38, 38]),
 (4, [4, 4, 4]),
 (4, [3, 4, 4]),
 (4, [3, 3, 4]),
 (3, [3, 4, 4, 4]),
 (3, [3, 3, 4, 4]),
 (3, [3, 3, 3, 3])]
```

It is possible to notice that predictably the patterns with the highest frequency are also those that do not present real movements, but are those in which users have tweeted several times in the same place (we called them “non-movements”). Excluding those, we can see that many people frequently made a movement from cluster 3 to cluster 4, so we show it on the Folium map in figure 4.17 by drawing it with a red arrow.

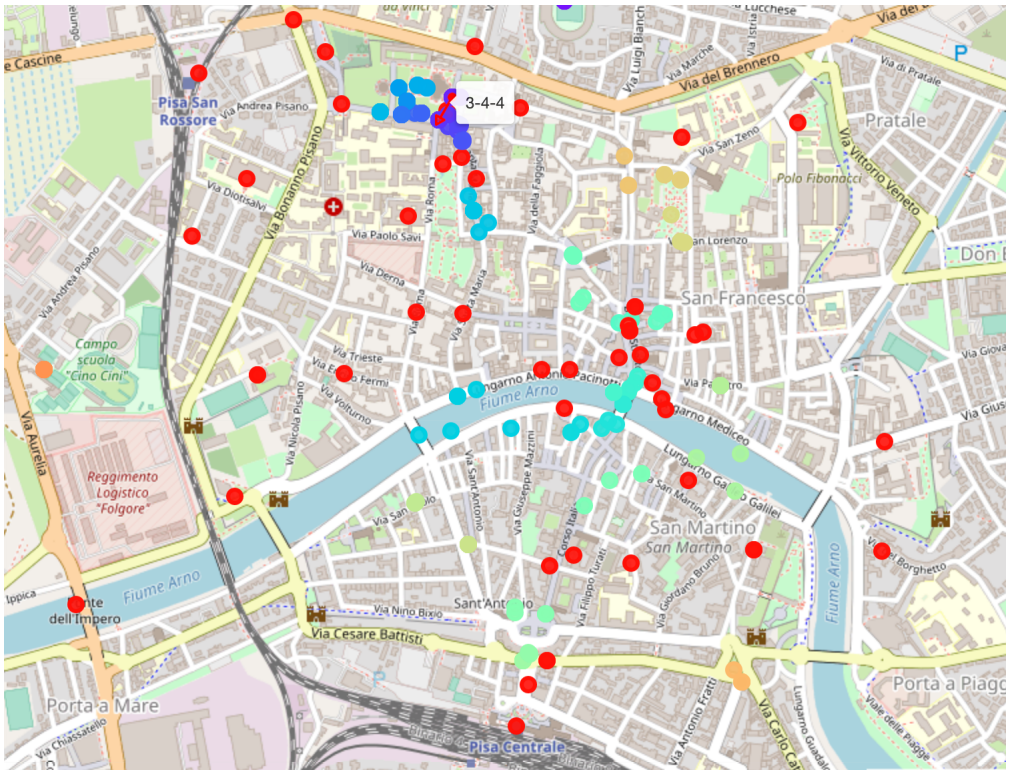

Figure 4.17: Folium map showing one of the most followed paths (3-4) with a minimum frequency of 3 extracted with PrefixSpan on the clusters obtained from Optics 2.

The area where the movement was made is not very informative for the purpose of our study: indeed, there is probably some noise caused by the paths followed by tourists, who are concentrated mainly in the Piazza dei Miracoli area, where the major attractions of the city of Pisa are present.

If we move to the analysis of the pattern mining results on clusters obtained by Bisecting K-Means, we obtain the following most frequent patterns:

```
[(7, [46, 46, 46]),
 (7, [16, 16, 16]),
 (4, [46, 46, 4]),
 (4, [46, 4, 4]),
 (4, [16, 16, 16, 16]),
 (4, [4, 4, 4]),
 (3, [46, 46, 46, 46]),
 (3, [46, 46, 4, 4]),
```

```
(3, [46, 4, 4, 4]),
(3, [35, 9, 46]),
(3, [9, 46, 46]),
(3, [9, 46, 4]),
(3, [9, 9, 9]),
(3, [4, 46, 46])]
```

We can note that, also in this case, the "non-movements" compare in the most frequent patterns, but we can also see some actual movements, such as from cluster 46 to 4 or as from 35 to 9 to 46. We show this latter pattern on the Folium map in figure 4.18.

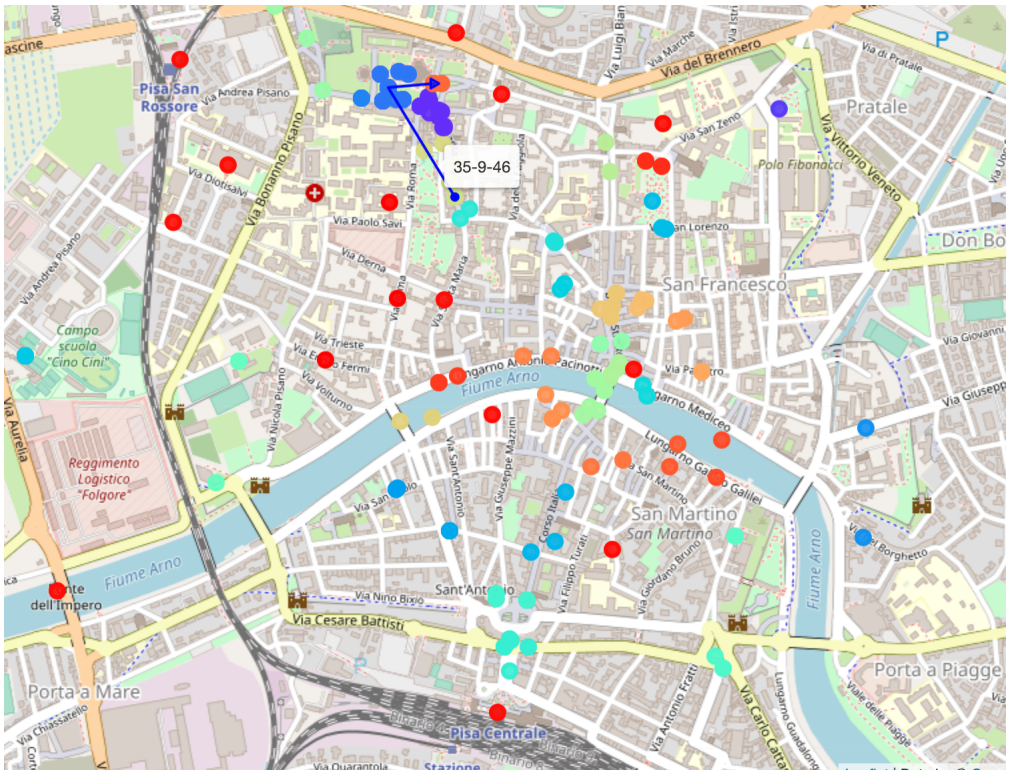

Figure 4.18: Folium map showing one of the most followed paths (35-9-46) with a minimum frequency of 3 extracted with PrefixSpan on the clusters obtained from Bisecting K-Means.

Although here too the movement has shifted towards a predominantly tourist area, we can still observe an improvement in the definition of movements between the various clusters. This may be an indicator of greater precision brought by the usage of the Bisecting K-Means algorithm in the clusters definition.

5. Conclusions

In this experimental thesis project we have seen two possible types of analysis to be carried out to study cultural events' data, in order to have a greater understanding of the visitors' behaviors with a view to the organization of future events. In particular, we focused on the past data of the Pisa Internet Festival, both with a Business Intelligence analysis and with a Data Mining analysis. The two methods clearly had different implications, so it is difficult to judge which was actually the most appropriate for this type of study. What is certain is that, if we had had better data in the Data Mining analysis part, we could have made a much more consistent estimation.

In this case, the most informative and concrete results were obtained with the first analysis, as they allowed us to see some interesting trends that emerged from the internal indicators, such as factors relating to visitors' engagement over the last few years, or others relating to to the resonance of the festival from the press coverage, or even aspects related to the people employed in the works of organizing the events. At the same time, they were useful for observing the trends in social media accounts' coverage over a year, in order to study the most suitable time to advertise the event and through which channels. For example, it emerged that during the month of October there is a lot of activity on social channels, probably because people tend to talk about it more consistently while attending the festival. It was also highlighted that participation on social media is mainly evident from the most populous cities (in addition to Pisa itself, which obviously covers the first spot because it hosts the festival), such as Rome, Milan, Florence, Naples or Turin. Moreover, the age group with the greatest involvement is the one between 25 and 34 years, with almost parity between the female and male gender. To sum up, it provides undoubtedly interesting information for a more general view of the event, which, through the use of graphs, has given light to aspects that were already present in the extracted data, but which we could not see since they were still in a raw CSV format. It is therefore a much faster type of analysis and, in some ways, essential, which aims to extrapolate those aspects that the data tell of itself in numerical format.

Still in terms of results, the second type of analysis was not informative for the organization of the festival, but only for a problem linked to the data types we had. Indeed, the data provider also had to provide us with data for the years 2017 and 2018, so that we could see characteristics or similarities between the results of different periods, but this was not possible because the data was not readable. This was a big loss, since these other data that we could not exploit and study were much larger than the ones from 2016, where we set the analysis on. In addition, the 2016 data period referred to a week before the festival started, so if we could have analyzed those during the actual festival days, we would certainly have gotten different results. Therefore, although the data we had was few and of bad quality in order to carry out a satisfactory analysis, we still preferred to finish it, in order to develop a prototype that could be used for future analyzes on other events or on the same event but with different data.

Regardless of these problems, however, a very clear result emerged: the methodology with data mining methods works, indeed the most crossed paths have been identified, which are located in the most touristic areas of the city and this suggests that they represent those paths ran across by tourists, who probably use Twitter even more than regular Pisan inhabitants to publish their new experiences and visits during their vacation. This is also confirmed by the fact that, if you try to conduct a semantic analysis on the Tweets' texts, the words that emerge the most are those related to the places visited rather than those related to the Internet Festival. So, despite the small amount of data and their low quality, we still managed to give a valid setting for other future analyzes that can be carried out on other types of cultural events, in such a way as to observe characteristics that from the initial JSON format files we could never have witnessed.

There are certainly many possibilities for future implementations of this project, mainly related to the Data Mining Analysis part. First, retry the analysis on different and more numerous data in order to extract reliable results.

Moreover, to further refine the results, it is recommended to conduct a semantic analysis of the texts of users' Tweets, so as to filter more those users who are classified as tourists and those Tweets that instead contain words explicitly referring to the Internet Festival, using hashtags or the individual events' keywords. There are also many studies

from which it is possible to take inspiration to distinguish the behavior of tourists and inhabitants on the basis of some characteristics [**8**].

A further implementation could concern the matrices' visualization, created after the clustering step: to understand the movements more immediately, a direct graph created with the JavaScript library D3.js[14], which is one of the most used libraries in the Data Visualization's field, could be used, since it allows you to take advantage of advanced layouts to improve the data narration.

[14]To learn more about this topic, visit the web page at the following link: `https://d3js.org/`

Bibliography

[1] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, *Introduction to Data Mining*, Pearson, 2020.

[2] Alex Burns, *Business Intelligence*, Australian Foresight Institute, March 2003, pag. 2.

[3] R. Guidotti, A. Monreale and S. Rinzivillo, *Learning Data Mining*, 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2018, pp. 361-370.

[4] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, Meichun Hsu, *PrefixSpan: Mining Sequential Patterns by Prefix-Projected Growth.*, Proceedings of the 17th International Conference on Data Engineering, 2001.

[5] Jiajun Liu, Haoran Li, Yong Gao, Hao Yu and Dan Jiang, *A geohash-based index for spatial data management in distributed memory*, 22nd International Conference on Geoinformatics 2014, pp. 1-4.

[6] M. Ankerst, M.M. Breunig, H.-P. Kriegel and J. Sander, *OPTICS: Ordering Points to Identify the Clustering Structure*, SIGMOD Record (ACM Special Interest Group on Management of Data), 1999, vol. 28, no. 2, pp. 49-60.

[7] R. Guidotti, R. Trasarti, M. Nanni, F. Giannotti and D. Pedreschi, *There's A Path For Everyone: A Data-Driven Personal Model Reproducing Mobility Agendas*, 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2017, pp. 303-312.

[8] R. Guidotti, L. Gabrielli, *Recognizing Residents and Tourists with Retail Data Using Shopping Profiles.* In: Guidi, B., Ricci, L., Calafate, C., Gaggi, O., Marquez-Barja, J. (eds) Smart Objects and Technologies for Social Good. GOODTECHS 2017. Lecture

Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 2018, vol 233. Springer, Cham. <https://doi.org/10.1007/978-3-319-76111-4_35>.

[9] R. Trasarti, R. Guidotti, A. Monreale, F. Giannotti. *MyWay: Location prediction via mobility profiling*, Information Systems, Elsevier, March 2017.

[10] R. Guidotti, A. Sassi, M. Berlingerio, A. Pascale and B. Ghaddar, *Social or Green? A Data-Driven Approach for More Enjoyable Carpooling*, 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 2015, pp. 842-847.

[11] Z. Desai, K. Anklesaria and H. Balasubramaniam, *Business Intelligence Visualization Using Deep Learning Based Sentiment Analysis on Amazon Review Data*, 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-7.

[12] P. Runnel, P. Pruulmann-Vengerfeldt, A. Aljas, K. Tampere (ERM), A. De Cesare (FST), M. Cerrai (FST), G. Ferrari (UniPi), C. Gagnaire (DDS), *Data and Impact. Guidelines on how data helps to understand the impact of the CCIs*, D2.2 - Guidelines report, Me-Mind Consortium 2021-2022, 20 December 2021, accessed 7 April 2022, https://www.memind.eu/wp-content/uploads/2022/03/Me-Mind_D2.2_Guidelines_report.pdf.

[13] V. Rohilla, M. S. S. kumar, S. Chakraborty and M. S. Singh, *Data Clustering using Bisecting K-Means*, 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2019, pp. 80-83.

[14] *Data dictionary: Standard v1.1*, Developer Platform Twitter, accessed 7 April 2022, <https://developer.twitter.com/en>.

[15] S. Alberti, *Why is Data Science important in cultural field?*, Data Science in Me-Mind, 21 May 2021, accessed 7 April 2022, <https://www.memind.eu/data-science/>.

[16] *INTERNET FESTIVAL, Evento dedicato al tema della Rete e della Rivoluzione digitale, a Pisa ad ottobre*, Fondazione Sistema Toscana, accessed 7 April 2022, <https://www.fondazionesistematoscana.it/progetto/internet-festival/>.

[17] *Cos è Power BI?*, PowerBI, Microsoft, accessed 7 April 2022, <https://powerbi.microsoft.com/it-it/what-is-power-bi/?&ef_id=EAIaIQobChMIk82-6_r39gIV6o9oCR2iLgmoEAAYASABEgJAU_D_BwE:G:s&OCID=AID2203275_SEM_EAIaIQobChMIk82-6_r39gIV6o9oCR2iLgmoEAAYASABEgJAU_D_BwE:G:s&gclid=EAIaIQobChMIk82-6_r39gIV6o9oCR2iLgmoEAAYASABEgJAU_D_BwE>.