University of Padova

Department of Mathematics "Tullio Levi-Civita"

Master Thesis in Data Science

Overtourism and AI:

A model for sustainable development

Supervisor

Prof. Tomaso Erseghe

University of Padova

Master Candidate

Gabriele Carbone

Student ID

2011649

Academic Year

2021-2022

"As this wave from memories flows in, the city soaks it up like a sponge and expands. A description of the city as it is today should contain all the city's past. The city, however, does not tell its past, but contains it like the lines of a hand, written in the corners of the streets, the gratings of the windows, the banisters of the steps, the antennae of the lightning rods, the poles of the flags, every segment marked in turn with scratches, indentations, scrolls."

— Italo Calvino, Invisible Cities

Abstract

The project outlines the construction of an artificial intelligence (AI) that - using data from crowd-source portals - detect potential alternative localities to the most famous tourist destinations. The goal is double: to **reduce over-tourism** inside the bigger cities and to **promote nearby rural areas**.

In the *introduction*, we present the main ideas and key concepts of the project. We also explain the main assumptions of the ideas and how they can impact the final results.

In the second chapter, **"Technical Part"**, we present the codes used for each step along with a detailed description. Starting from freely available dataset, we create a pipeline with a series of manipulations that allow us to create artificial intelligence. Based on that, we first try to use simple models and then we slowly progress towards more complex ones like ensemble models.

The goal is to find the most interesting areas inside the region analyzed. This can be done by using clustering analysis on the best-perfoming zones.

In the third chapter, **"Social Part"**, we explain the different criteria for the analysis of the outputted clusters from the technical part. By considering the different stakeholders point of views and presenting the work under different lights, it's possible to slowly discard non-interesting candidates. The goal of the analysis is to choose the at most 2-3 candidates to promote to the general public, especially outbound tourists.

In the fourth chapter, **"Case Studies"**, we take three hypotetical scenarios: Venice, Amsterdam, and Barcelona. We apply a streamlined version of the pipeline built for the social part to each of them and we briefly discuss the results. For Venice, the final candidates are: San Donà del Piave, Mirano, and Conegliano. For Amsterdam, they are Gouda and Heemskerk. For Barcelona, they are: Matarò and Villanova y Geltrù.

For each of them, we analyze the maps outputted from the pipeline, the major trends, and the assumptions underlying the best performing model.

Finally, the *conclusion* wrap up all the work and highlight the most important points to the reader.

Contents

Listing of figures

Listing of tables

Listing of acronyms

AI ..... Artificial Intelligence
EFB ..... Exclusive Feature Bundling
DT ..... Decision Tree
EDA ..... Exploratory Data Analysis
GOSS ..... Gradient-based One-Side Sampling
ML ..... Machine Learning
PoC ..... Proof of Concept
PoI ..... Point of Interest
RF ..... Random Forest

1 Introduction

1.1 The Problem

Many of the biggest cities in Europe suffer from the **overtourism** problem [1]. Even in Italy, it's easily visible in cities like Turin, Rome, Milan, and Naples. Some of them started to take active measure to limit the phenomenon [2]: in Venice, starting from October 2022, it will be mandatory to book the visit to the city [3] while in Florence, a portion of the Uffizi collection has been re-distributed to their original towns to promote rural tourism [4].

Essentially, overtourism can be defined as *"the impact of tourism on a destination, or parts thereof, that excessively influences perceived quality of life of citizens and/or quality of visitor experiences in a negative way"* [5]. Starting from the definition, it's already possible to assess the breadth of the problem: it involves the social, economical, cultural and environmental spheres [6]. The phenomenon is widely spread: around one every four tourist felt that their destination had been "overcrowded" and for almost 10% of them the trip was negatively impacted by too many people [7].

With the raise of the cultural and environmental awareness, many governments are taking measures - sometimes extreme ones like in Barcelona, Venice and Amsterdam [8] - to reduce the phenomenon. There are several factors to take into account: the recent changes in the tourism industry due to the Covid-19 [9], the gas/oil market fluctuations and the increasing efficiency of the logistic/transport industry.

The travel sector is lagging behind when it comes to the sustainability goals [10], especially when it comes to the cruise ship travel and the coastal cities. But, with the advent of the digital innovation, it may possible to optimize and track several key resources [11].

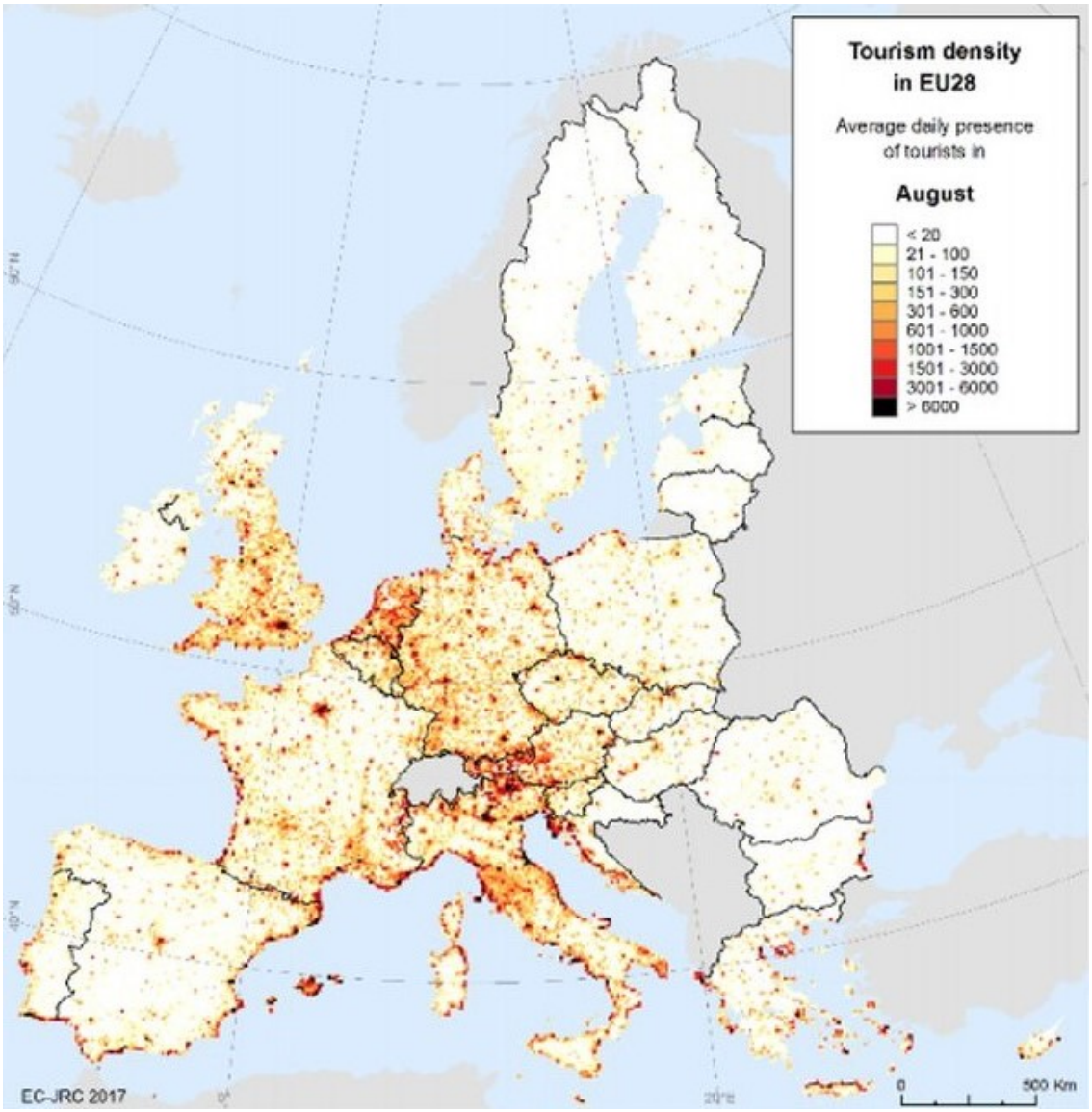

Figure 1.1: Tourism Pressure in Europe

Also, it's worth noticing that there is a raise in a more "conscious" type of travel, especially among the new generation, and the consequent born of specialized travel agencies.

On the other side, there is a wealth of "hidden treasures" near the affected cities: it's only natural that smaller cities developed in a similar way to the main one, just on a different scale.

The **economical and social promotion of rural areas** is an evergreen challenge to every government. To solve the problem, obliviously, is not easy, but it can be done by integrating them inside wealthy realities.

It goes without saying that the correct input can improve the life of the residents in several ways, not just the social/economical one.

1.2 The Approach

In order to understand the problem from different points of view, a deep research on the scientific literature and the divulgative literature was the first step. Just recording a series of fact didn't seem to deliver the full picture, so we created a graph for each activity/stakeholder/goal and used the edges to describe the relationships between each node. The resulting image give back a clear picture of the phenomenon:

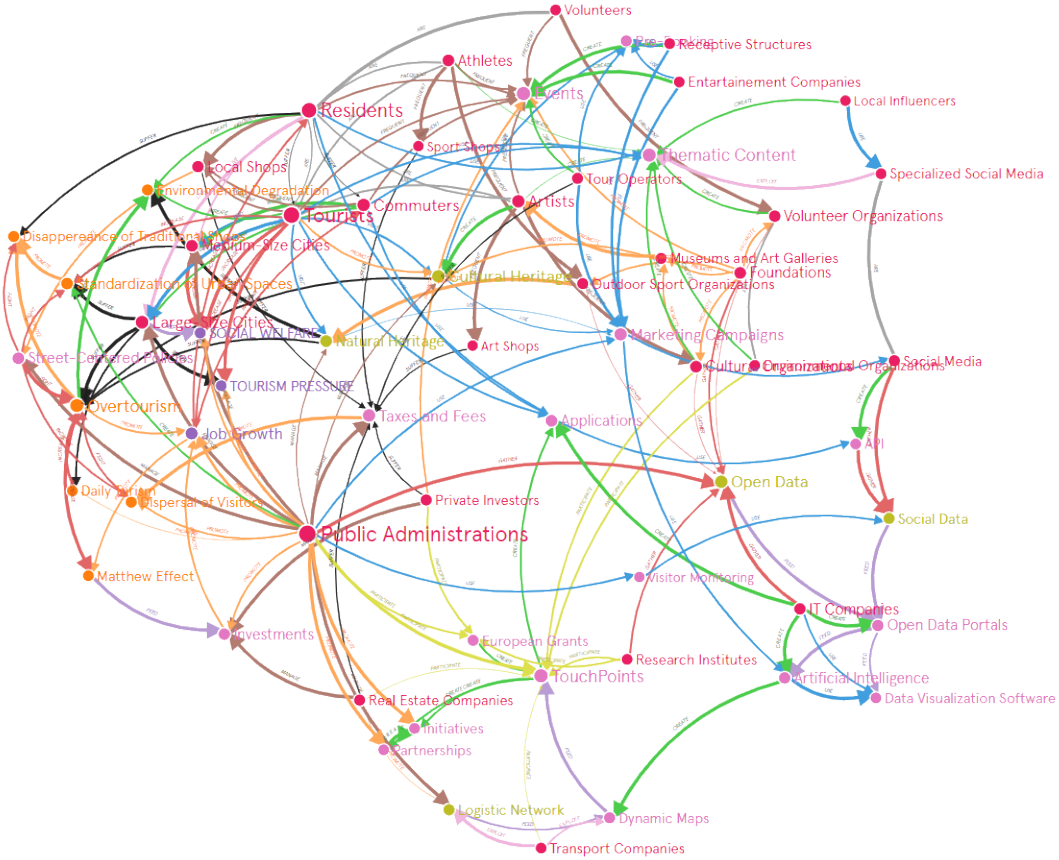

Figure 1.2: Exploratory Graph

Building and reading the graph was the first approach used in the project to be sure that important elements weren't left out.

The node's dimension is proportional to his degree, the edge's thickness is proportional to the estimated strength of the relation, and the color of each element show the different category.

We can immediately see that *the public administrations and the tourists play central roles*, but we can also notice that activities (nodes in light pink) like regular touch points and thematic content may be extremely helpful.

Finally, we can spot – in the top-left side of the graph – a particularly crowded area: it comprises elements regarding the different size of the cities and social goals like welfare. It makes sense to have such a dense elements

group thinking about how correlated are to each other.

The graph is freely accessible at the link:

`https://legacy.graphcommons.com/graphs/2e67f653-29a9-4ca4-b6d1-b1e4381379fc`

1.3 The Solution

The project revolves around the idea of **"channeling" the excessive tourism from the bigger cities to the near - rural - towns**. This way, the main city is somewhat relieved of a portion of that pressure, and the near towns have new economic inputs and the opportunity to increase the welfare of the residents.

To create a proof of concept (PoC), a **mixed approach** is required: the technical part will use Data Science to ingest data and to select a list of potential "candidates" areas, while the social part will use methods typically belonging to Business Administration for selecting the most promising cities.

The evaluation will be done by taking into account **several factors**: security, accessibility, logistics, culture, entertainment industry, and nature. The areas that perform better in these areas will be first selected by the AI and then evaluated in the second part of the process.

The end results it's a **portfolio of 2-4 cities** that have high - unexpressed - tourism potential with a list of the main attractions to promote. In a practical scenario, these results should be shared with key stakeholders like travel agencies, public administrations and local organizations in order to start a conversation about the potential collaborations and activities to fund.

1.4 Key Assumptions

Assumption #1: IF a city is rich from a cultural and environmental standpoint THEN it's probable that the towns near it will be rich, to some degree, in the same way

Assumption #2: IF a town/city has many cultural and environmental point of interest THEN it has high touristic potential

Assumption #3: IF all the major obstacles (security, accessibility, transport) are removed AND the nearby town is correctly promoted by the central city THEN more tourists will visits the nearby town

Assumption #4: At least a portion of the tourists treasure the city they're visiting enough so that they are open to explore other parts of the area to preserve the cultural and natural heritage

1.5 Key Ideas

Crowd-gathered Data: The project uses data from OpenStreetMap, a community of people that create, share and keep updated geo-spatial linked information from the entire planet.
The dataset are available to download and free to use.

Artificial Intelligence: AI will be used to detect high-potential areas that can are ready to receive additional touristic fluxes. But it's not enough: to effectively discern between a good candidate and a bad one, there must be a manual verification that takes into account historical precedents, reviews, distance from the main city, feasibility of the travel and other region-specific problems.

Starting Point: The project itself is only a starting point for a conversation that should happen between different stakeholders (public administration, local organizations, and travel agencies for example) in order to create a well-coordinate effort - and a receiving system - for the new touristic flux.

2 Technical Part

2.1 General Approach

The approach was mainly developed using **freely accessible datasets and tools**.

The first problem we encounter is the fact that it's difficult to find comparable information about every single point in a region. Also, it's even more difficult to find data about different social aspect of every point in a region.

The only data source that is both curated and freely available is, to this day, *OpenStreetMap*. It follows that the majority of the technical part was developed based on the data available.

Most of the technical part, in fact, is focus on extrapolating the right kind of data from the raw datasets and visualizing it. The data science part is present, much like the real-world scenarios, only in the last steps of the workflow.

The code is freely accessible at the link:
`https://github.com/gabrielecarbone/overtourism`

2.2 Importing the Dataset

The first step is, obliviously, importing the dataset.

It's easier said than done: while dealing with geo-spatial data it's not strange to have different gigabytes of data at once, even for small regions. In order to avoid excessive loads on the (freely available) servers, the OpenStreetMap team has put in place several types of download limits.

In the end, to access the data is necessary to use **third-party organization** that bulk-download these data and make them readily available online.

The most famous source is GeoFabrik: `https://download.geofabrik.de`

After selecting the region of interest, it's possible to choose between several different format. The most flexible type of download is the ".shp.zip" type: it allows you to download the information separated in several - smaller - files that you can read one by one using any programming language you prefer.

GeoFabrik divides the information about an area in several - smaller - categories:

Name	Brief Description
Buildings	Contains information about every kind of building not included in the following categories.
Landuse	Contains information, mainly areas, about the use of certain land types (grass, residential, commercial, and so on).
Natural	Contains information about everything related to the natural and biological world, excluded the marine one.
Places	Contains information about the type of areas contained in the dataset. For example, you can find out if an area is an island or a farm or an hamlet.
Places of Worship	Everything related to religion (churches, sanctuaries, monasteries)
Places of Interest	Very broad category containing all the usual place of interest. In this files, we can also find information related to the tourism sector.
Railways	Information about train stations, stops and level passages.
Roads	Information about primary and secondary streets.
Traffic	Usually, it contains additional roads information like motorway junctions, traffic signals or crossings.
Transport	Everything related to the public transport - excluded the railways.
Water	Geographical information about water, wetlands, glaciers and riverbanks.
Waterways	Specific information regarding water bodies that can be used for navigation

Table 2.1: GeoFabrik DataSet Division by Categories

Depending on the area, the relative dimension of each dataset can vary a lot, with some files being still to big to be manipulated on a normal computer.

Finally, unzipping the download, we'll find several data types:

- **.cpg:** Files with this extension contains, usually, raster images. They are used to add satellite photos to the map visualizations.
- **.dbf:** This is one of the most important files. It contains information about the point of interest of the area downloaded (streets, statues, houses, parks, and so on) together with the personal classification system of OpenStreetMap ("fclass").
- **.prj:** It's a simple text file that contains the coordinate system and the projection information.
- **.shp:** ShapeFiles are used to describe vectors (points, lines and areas) using coordinates. Together with the .dbf files, they are the most important for our analysis.

- **.shx:** File containing shapes or fonts compiled by AutoCAD.

During the project, we will use only the .dbf and the .shp files: we will leave the project part to the data visualization online software, Carto.

This allows us to work with fewer files and delegate the project part to a second moment, when the files will be already cleaned and easier to manipulate (given the average dimension of the file).

To work with the files, we will need a couple of **standard libraries**, like Os and GeoPandas (a custom version of Pandas, specifically designed to handle geo-spatial data), plus a dedicated library called "simpledbf". The former one is essential to open the bigger .dbf files because it's optimized to do so, the integrated GeoPandas function require too many resources.

```
import os
import geopandas as gpd
from simpledbf import Dbf5
```

Next, we simply import the file list:

```
input_path = '../000-download_dbf/'
file_list = os.listdir(input_path)
```

And, we read each .dbf file in the uncompressed folder:

```
df_dict = {}

for file_name in file_list:
    if file_name[-4:] == '.dbf':
        dbf_path = input_path + '/' + file_name
        dbf = Dbf5(dbf_path)
        df = dbf.to_dataframe()
        df_dict[file_name] = df
        print('Importing DataFrame: ', file_name[:-4], ' of shape ', df.shape)
```

After being read, a new entry to the dictionary is added using, as the key, the name of the file and, as the value, the information contained in it.

We also add a "print" statement to check on the execution of the cell, because the biggest files can require several minutes to read and import.

2.3 Adding the Coordinates

At this point, we've imported only the point of interested *without* the coordinates.

To import them, we need to use GeoPandas:

```
for df_name, df in df_dict.items():

    print('Starting finding coordinates for ', df_name[:-4], ' of shape ', df.shape)

    shp_path = input_path + '/' + df_name[:-4] + '.shp'
    shp_series = gpd.read_file(shp_path)

    centroids = shp_series.centroid

    df['lon'] = centroids.geometry.x
    df['lat'] = centroids.geometry.y

    print('Added coordinates to ', df_name[:-4], ', now shape is ', df.shape )
```

In this short piece of code, we can see that we first import the files by selecting only the names that ends with ".shp", then we read them with GeoPandas.

At this point, we **compute the centroid** of every element present.

This makes easier to compute a couple of steps down the workflow. For example, it solves the problem to assign an area to a cell of the map if the area overlap many cells.

In general, there are three data types inside OpenStreetMaps databases: points, lines and areas. Using the centroid function reduce everything to the point data type.

Finally, we add the longitude and the latitude of the centroid to the already existing dataframe.

This process is done for every dataframe present in the dictionary previously created.

As always, there are a couple of print statements to check on the status of the execution.

We conclude this section by saving the dataframes into .csv files:

```
    os.makedirs('coordinates/', exist_ok=True)

    for df_name, df in df_dict.items():
        file_name = 'coordinates_' + df_name[:-4] + '.csv'
        df.to_csv('coordinates/' + file_name)
```

The code is easy to follow: we first check that the destination folder exists (and, if not, we create it). After that, we create the new filename - which is essentially the same name of the other files with the extension being ".csv" instead of ".shp" or ".dbf" - and then we export them using the integrated GeoPandas method ".to_csv"

2.4 Clean the Dataset

This is the crucial step for the entire technical part: cleaning the datasets the right way means to be able to manipulate them.

First of all, it's necessary to **assess the information available now**. At this point, the dataframes contain the following features:

Feature	Description
osm_id	A unique identifier assigned to each observation
code	The category code
fclass	The main category of the observation
name	The name of the observation (not always present)
type	The specific category of the observation (not always present)
lon	Longitude of the observation
lan	Latitude of the observation

Table 2.2: Dataset Features before the Cleaning Step

The **key information**, for this step, is the second feature - code -, because with a single number supply both general and particular categorization of the observation.

From the OpenStreetMap guidelines [12], we can see that the first digit already carry general-type information:

Digit	Description
1	Places
2	Point of Interest
3	Places of Worship
4	not-used
5	Roads
6	Railway Stations
7	Landuse
8	Waterways
9	not-used

Table 2.3: Meaning of the First Digit in the OpenStreetMap Classification System

GeoFabrik does not offer a file containing the code and the relative description for their system. A .csv file was manually redacted for the project by starting with the documentation available online.

After that, it was assigned a macro-category to each code: accessibility, art, entertainment, security, nature, sport, tourism, transport.

By assigning these macro-category directly to the code, we can indirectly classify each observation in the database.

2.4.1 Creation of the Filter DataBase

We start by importing the .csv file containing all the GeoFabrik categories into a variable called "filter_df".

```
filter_df = pd.read_csv('categories.csv', sep = ';', encoding = "ISO-8859-1")
```

Here, it's important to notice the additional parameter to the ".read_csv" function: both the ";" separator and the ISO-8859-1 are typical setting of .csv created with Microsoft Excel.

```
filter_df = filter_df.loc[~filter_df.SELECT.isna(),
                          ['SELECT', 'Geometry Code', 'Layer', 'fclass']]

filter_df['Geometry Code'] = filter_df['Geometry Code'].str.strip('0')
```

We remove, from the filter dataframe, rows that may be void and we select only the significant features.

We also remove the leading and trailing zeroes from the "Geometry Code" feature. This step will be crucial in the next steps, because it allows us to discriminate between different codes based on their length.

```
granularity = 4 - filter_df['Geometry Code'].str.len()

multiplier = 10 ** granularity
```

We can now **create a new variable – granularity** --, based on the length of the geometry code. The variable will have maximum value of 3 for code of length 1 and minimum value of 0 for code of length 4. Remember that the longest codes are composed of 4 digits.

This means that the granularity will be higher for codes that are less "specific" and lower for codes that are more specific.

Combined with the "multiplier" variable, defined immediately below, we can obtain a list of number that – multiplied by the geometry code – give us a 4 digit code. So, for example, 44 will become 4400 because it will have a granularity of 2 and a multiplier of $10^2 = 100$.

FOCUS: Why having the same length is important?

To ensure consistency with the geometry codes.
These kind of manipulations guarantee us that all the code will have 4 digits and that every single value between them will be matched. Using directly the codes of the documentation leaded to a very poor match because most of the observations were in the middle of two categories and couldn't be matched with either one.
Following the computations, the next piece of code will create an interval for each geometry code. The interval will vary in span, depending on the granularity of the source code.

```
filter_df['Geometry Code'] = filter_df['Geometry Code'].astype(int)
codes = filter_df['Geometry Code']

filter_df['low_threshold'] = multiplier * codes
filter_df['high_threshold'] = multiplier * codes + (multiplier - 1)
```

First, we enforce the integer type over the feature. After that, we create the low threshold of the interval using simply the computation seen a couple of paragraphs ago. After that, for the high threshold, we use exactly the same computation but we add a quantity based on the granularity in order to reach the **next** low threshold.

Continuing the previous example of 44, the low threshold would be 4400, while the high threshold would be $4400 + 100 - 1 = 4499$. For a geometry code of 3, the low threshold would be 3000 while the high threshold would be $3000 + 1000 - 1 = 3999$ and so on.

```
filter_df['fclass'] = filter_df['fclass'].fillna(filter_df['Layer'])

filter_df['granularity'] = granularity

filter_df = filter_df.sort_values(by=['granularity'])
```

We fill the NaN of the "fclass" feature by using the layer feature and we order the dataframe by granularity: the first rows will be the more specific ones (like 2601) and the last rows will be the more general (like 3).

The order_by command is important because it will influence the behaviour of the search function of the next steps. By putting the more specific categories first, and taking the first match, we ensure that the algorithm search first for the more specific categories and then for the more general ones.

```
filter_df_to_dict = filter_df[['fclass', 'low_threshold', 'high_threshold']]

filter_df_to_dict.drop_duplicates(subset = ['fclass'], inplace = True)
```

Next, we create a new variable that contains only the classes and the thresholds. We also remove occasional duplicates.

```
filter_df_to_dict.index = filter_df_to_dict.fclass

filter_df_to_dict.drop('fclass', axis = 1, inplace = True)

filter_dict = filter_df_to_dict.to_dict('index')
```

We set the "fclass" feature as the index and drop it from the columns.
Finally, we create a dictionary using the rows by specifying the "index" in the .to_dict method.

2.4.2 Applying the Filter DataSet to the GeoFabrik DataSets

We start by importing the file path that contains all the dataframes with the coordinates and the .csv file created to filter the observations.

```
path = '../010-add_coordinates/coordinates/'

file_list = os.listdir(path)
```

Next, using a for cycle, we iterate through all the datasets in the folder and we assign them the class.

```
filtered_df_dict = {}    # Initialize Dictionary

for df_name, df in df_dict.items():

    df['fclass'] = 0     # Create the fclass Feature

    for fclass, value_dict in filter_dict.items():

        df.loc[(df.code >= value_dict['low_threshold']) &
               (df.code <= value_dict['high_threshold']), 'fclass'] = fclass

        final_df = df.merge(filter_df[['fclass', 'SELECT']],
                            on = 'fclass', how = 'left')

        final_df = final_df.drop_duplicates()

        final_df = final_df[final_df.fclass != 0]

    filtered_df_dict[df_name] = final_df
```

In the first cycle, using the ".items()" method, we can extract at the same time both the key and the values of a dictionary. We create, for each dataset, the "fclass" column.

For each dataset, we use the same method/strategy to extract the "fclass" and the corresponding value.

The first command, inside the second "for" loop, essentially take all the observations that fall inside that fclass threshold and assign them the corresponding fclass. For example, all the observation with fclasses like 2401, 2406, 2421 will have their fclass value changed from 0 to – respectively – "hotel", "chalet", and "shelter".

The second command, the merge, is used to **add the macro-category** – the feature "SELECT" on the filter_df – to the dataframe elaborated. By selecting the left join, we ensure that no observation is lost in the process.

As always, we drop the duplicates and the values that didn't matched. We repeat this process for every class on the filter_df and we end it by entering the dataframe name and their values to the initialized dictionary: filtered_df_dict.

```
for df_name in filtered_df_dict.keys():
    print('DATAFRAME NAME: ', df_name)
    print('ORIGINAL DATAFRAME SIZE: ', df_dict[df_name].shape)
    print('FILTERED DATAFRAME SIZE: ', filtered_df_dict[df_name].shape)
```

To **check the output** of the previous chunk of code, we can print the original size and their filtered one. If the code works, most of them should have less rows and one column more ("SELECT", the macro-category).

2.4.3 Removing the Observations without Name

It's also possible, altough not recommended, to remove the observations without a proper description/name.

```
completed_df_dict = {}

for df_name, df in filtered_df_dict.items():
        completed_df_dict[df_name] = df[~df.name.isna()]
        completed_df_dict[df_name] = df
```

This piece of code is relatively easy to explain: it just take each – filtered – dataframe, remove all the observation that return "True" to the .isna() method and create a new key and value in the initialized dictionary: "completed_df_dict".

2.4.4 Exporting the Cleaned DataSets

Finally, we can export the datasets with the same method used while we added the coordinates:

```
os.makedirs('filtered/', exist_ok=True)

for df_name, df in completed_df_dict.items():

    file_name = 'filtered_' + df_name[:-4] + '.csv'
    df.to_csv('filtered/' + file_name)
```

2.5 Feature Engineering

For the **central part** of the technical section, we will discuss each step in depth. The first sections will be dedicated to the exploration of the dataset through the Exploratory Data Analysis (EDA) and the Feature Engineering. After that, there will be a section dedicated to the building of the first models and then a possible expansion of them.

2.5.1 Merging the Datasets

Linked to the last step of the cleaning part, the first thing is to **merge** all the cleaned dataset:

```
path = '../020-clean_dbs/filtered/'

file_list = os.listdir(path)
columns_to_keep = ['osm_id', 'code', 'fclass', 'name', 'lon', 'lat', 'SELECT']
dfs = []

for filtered_df in file_list:
    df = pd.read_csv(path + filtered_df)
    dfs.append(df[columns_to_keep])
```

We select only some of the columns to reduce, whenever possible, the dimensions. Even filtered, the dataset for Venice – for example – was more than 2,5 millions of observations.

In this case, we are directly **removing the index feature**, which is simply a progressing number starting from 0.

```
merged_df = pd.concat(dfs)
merged_df.to_csv('0. merged_df.csv')
```

Having all the dataframe in a list, we can simply cast the function "pd.concat" and pass the list as parameter to vertically merge all the dataframe inside the list.

We can then export the final list into a .csv file.

FOCUS: Why constantly exporting and importing the dataframes?

This practice allows the workflow to be broken in several notebooks and the results to be saved several times. In this way, it's possible to increase the interactivity of the project because it's easier to export the results at a pre-determined step in the process and it's also easier to edit that step.
On *GitHub*, it's possible to see a very **similar way** to organize the files because it allows many different users to edit the project without having to deal with a single, incredibly large, piece of code.

2.5.2 Setting the Parameters

After the re-importing of the dataframe, we select the area we want to analyze:

```
lon_min, lon_max = 11.5, 13
lat_min, lat_max = 45, 46

df = df.loc[(df.lon >= lon_min) & (df.lon <= lon_max)]
df = df.loc[(df.lat >= lat_min) & (df.lat <= lat_max)]
```

Essentially, in this passage we "cut" a rectangle on the area we are analyzing by removing all the observations that didn't have the coordinates inside the parameters we defined.

This is a **crucial passage** because, in many ways, analyzing even a fraction of the GeoFabrik dataset would be too much without professional-grade computers/servers.

```
n_rows, n_columns = 200, 200

lat_min, lat_max = df.lat.min(), df.lat.max()  # rows
lon_min, lon_max = df.lon.min(), df.lon.max()  # columns
```

To make the computation and the comparisons easier, we **divide the grid into cells**. For most cases, a number of cells between 100 and 300 is enough.

By knowing the wanted minimum/maximum of the coordinates and the total number of cells, it's possible to compute the same information in relation to the dataset. This way, we can further restrict the area of our search by removing all the space without observations.

In this phase, we also compute a couple of useful statistics on the coordinates: the difference between the minimum and the maximum, and the step size.

```
diff_lat = lat_max - lat_min
diff_lon = lon_max - lon_min

step_lat = diff_lat / n_rows
step_lon = diff_lon / n_columns
```

We also prepare a variable with all the macro-categories present and choose the starting coordinates:

```
categories = df['SELECT'].drop_duplicates().to_list()

starting_point_lat, starting_point_lon = lat_min, lon_min
```

The starting coordinates are the first elements we need to set in order to choose a direction for the future loops. In fact, they regulate the next chunk of code:

```
# Setting latitude vertices of the cell
bbox_lat_min = starting_point_lat
```

```
bbox_lat_max = starting_point_lat + step_lat

# Setting longitude vertices of the cell
bbox_lon_min = starting_point_lon
bbox_lon_max = starting_point_lon + step_lon
```

2.5.3 From Table to Grid

In this step, essentially, we are **defining the vertices of the first cell** to analyze in the "for" loops. We begin with the starting coordinate and then add the step-size.

Then, we transform our dataset of points into a grid with values based on their macro-category. To do so, we need to iterate first through the rows and then through the cells. This is the first loop:

```
for row in range(n_rows):
    bbox_lon_max = starting_point_lon
    bbox_lon_min = starting_point_lon + step_lon

    row_df = df.loc[(df.lat > bbox_lat_min) & (df.lat <= bbox_lat_max), :]
```

As we can see, we reset the starting point to the starting longitude to each iteration, and we take only the subset of observations that are inside that row. This is the second loop, dedicated to the columns:

```
for column in range(n_columns):
    cell_df = row_df.loc[(row_df.lon > bbox_lon_min) &
                         (row_df.lon <= bbox_lon_max), :]

    cell_name = 'R' + str(row + 1) + 'C' + str(column + 1)
    cell_values = [row + 1, column +1,
                   bbox_lat_min, bbox_lat_max,
                   bbox_lon_min, bbox_lon_max]
```

Much like before, from the "row" select we select only the column that we want, creating a cell. We save the coordinates and the row/column number inside a list. We proceed by counting the values inside each cell, for each macro-category:

```
for category in categories:
    try:
        count = cell_df.SELECT.value_counts()[category]
        cell_values.append(count)

    except:
        cell_values.append(0)
```

In this chunk of code, we need to use **control flow** because there may be a situation where there is not a single observation that fall inside a category. This is a real risk especially for macro-categories that usually have less observations like "accessibility" and "security".

We close the column loop:

```
result_log[cell_name] = cell_values

bbox_lon_min = bbox_lon_max
bbox_lon_max += step_lon
```

We create a new entry in the result dictionary by using the cell name and values defined previously. We also advance of on step in the process.

```
bbox_lat_min = bbox_lat_max
bbox_lat_max += step_lat
```

Similarly, we close the row loop by advancing of one step. This is the full cycle:

```
for row in range(n_rows):
    # reset starting point of the column axis
    bbox_lon_max = starting_point_lon
    bbox_lon_min = starting_point_lon + step_lon

    row_df = df.loc[(df.lat > bbox_lat_min) &
                    (df.lat <= bbox_lat_max), :]

    for column in range(n_columns):
        cell_df = row_df.loc[(row_df.lon > bbox_lon_min) &
                             (row_df.lon <= bbox_lon_max), :]

        cell_name = 'R' + str(row + 1) + 'C' + str(column + 1)
        cell_values = [row + 1, column +1,
                       bbox_lat_min, bbox_lat_max,
                       bbox_lon_min, bbox_lon_max]

        for category in categories:
            try:
                count = cell_df.SELECT.value_counts()[category]
                cell_values.append(count)

            except:
                cell_values.append(0)
```

```
            result_log[cell_name] = cell_values

            bbox_lon_min = bbox_lon_max
            bbox_lon_max += step_lon

        bbox_lat_min = bbox_lat_max
        bbox_lat_max += step_lat
```

Finally, we import the results into a Pandas DataFrame:

```
result_df = pd.DataFrame(result_log).T
result_df.columns = ['row', 'column',
                     'lat_min', 'lat_max',
                     'lon_min', 'lon_max'] + categories
```

We rename all the columns at once by using a list containing the coordinate data and the name of all the categories.

As always, we export the dataframe into a .csv file, while removing the cells that have 0 in every class:

```
filtered_df = result_df[result_df[categories].T.sum(0) != 0]
filtered_df.to_csv('1. filtered_df.csv')
```

2.5.4 Exploratory Data Analysis

We import the filtered dataframe and compute basic statistics on it:

```
df = pd.read_csv('1.filtered_df.csv')

df.describe()
```

The describe method compute and return a list of statistics for each feature of the dataset:

Statistic	Description
count	# of non-NaN values
mean	The mean of the feature's values
std	The standard deviation of the feature's values
min	The lowest value of the feature
25%	The first quartile's value
50%	The median value
75%	The third quartile's value
max	The highest value of the feature

Table 2.4: Statistics computed by the Pandas "describe" method

Then, we begin by assessing the quality of the data by visualizing it. We start by initializing a NumPy matrix:

```
viz_matrix = np.zeros([int(df.row.max()), int(df.column.max())])
values = df.iloc[:, 6:].sum(axis = 1).to_list()

rows = df.row.to_list()
columns = df.column.to_list()
```

The dimension of the matrix initialized is directly proportional with the number of cells we set as a parameter. It's clear that, *the higher the number of the cell the higher the computational complexity* and the lower will be the average number of values inside each cell (the same observations will be split in more cells).

We also take the **sum of the aggregate values** in the macro-categories, for each cell. This way, we can see how many point of interest there are inside each cell of the region analyzed.

```
for value, row, column in zip(values, rows, columns):
    viz_matrix[int(row) - 1, int(column) - 1] = value
```

We add each value to the corresponding cell. Next, we normalize the matrix to reduce the noise provoked by the outliers. Furthermore, we outright remove the top and bottom 1%.

The normalization procedure consist in two operations:

- Compute the mean and the variance of the entire dataset
- Remove the mean from every observation in the dataset
- Divide the result by the variance

There are several reasons to do so, but the most important one is that it makes easier for comparisons between different dataset. The most common mathematical form that the standardization operation has is the following:

$$Z = \frac{X - \mu}{\sigma} \tag{2.1}$$

Where Z is the standardized form, X is the array of values to standardize, μ is the mean of the observations, and σ is the the standard deviation of the observations.

Before normalization, it's a common practice to remove the outliers. There are several methods and thresholds, but we use one of the easiest: the 1% and 99% percentiles. This means that every observation that has values higher than 99% of the other observations gets "lowered" to the 99% percentile value. The contrary is true for the 1%.

The quantile of a distribution is defined as [13]:

$$Q(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\}, \quad with \ 0 < p < 1 \tag{2.2}$$

Where p is the probability we want to obtain (from 0 to 1) and $F(x)$ is the distribution function. Essentially, the **quantile** is the value for which the probability to extract the same value or higher from the distribution $F(x)$ is equivalent to p.

This step is fundamental for this use case: given that we will analyze international tourist destinations, it's probable that we will have extreme outliers with regard to the different kind of features and they can jeopardize

the process. Often, if we skip the step, we will create maps that will be *very* colored in the correspondence of the outlier and almost transparent/blank in every other area.

```
# Standardize Matrix
viz_matrix_normalized = (viz_matrix - viz_matrix.mean()) /viz_matrix.std()

# Compute Outliers
normalization_intensity = 0.01
low = np.percentile(viz_matrix_normalized,
                    normalization_intensity * 100)
high = np.percentile(viz_matrix_normalized,
                     (1 - normalization_intensity) * 100)

# Remove Outliers
viz_matrix_normalized = np.where(viz_matrix_normalized < low,
                                 low, viz_matrix_normalized)
viz_matrix_normalized = np.where(viz_matrix_normalized > high,
                                 high, viz_matrix_normalized)
```

Finally, we plot the matrix using MatPlotLib and SeaBorn. This is the final result for the Venice case study, lighter points indicates areas with more point of interest:

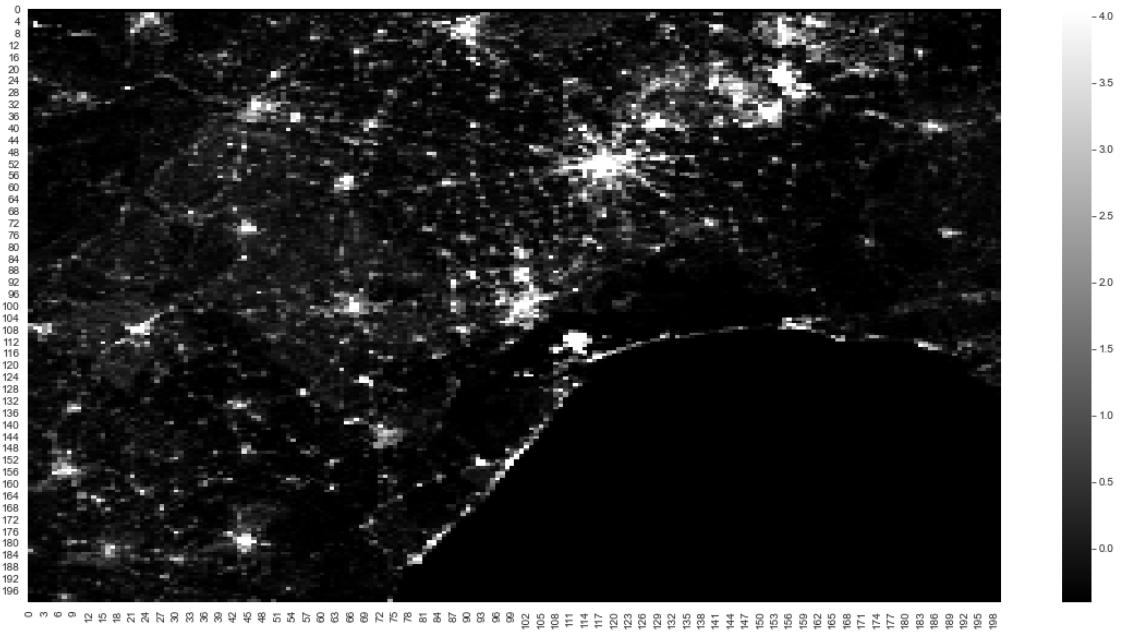

Figure 2.1: Preliminary Visualization of the Aggregated Values

The code to obtain this result is fairly simple, but depending on the manipulation in the previous steps, it may be necessary to change the "k" parameter in the np.rot90 function.

```
fig, ax = plt.subplots(figsize=(20,10))

cmap = sns.cubehelix_palette(light = 0,
                             dark = 1,
                             hue = 0,
                             as_cmap=True)

sns.heatmap(np.rot90(viz_matrix_normalized, k = 3), cmap = cmap)
```

2.5.5 Exporting the Grids

Having all the data in the **correct shape**, alias the grid, we need to correctly export it in order to visualize the information using online tools like Carto.

These services deals with very *specific data types*, the kind we have seen in the very first table of the document. The way it works, the service will try to find the table that contains the "geometry" first and then it will try to understand what kind of data types are the other features (quantitative or qualitative variables).

So, it makes sense to dedicate a final step to the correct elaboration of the output. We start by loading the data and the necessary packages:

```
import geopandas as gpd
from shapely.geometry import Polygon

gdf = gpd.read_file('1.filtered_df.csv')
```

It's important to notice that, in this step, both the packages and – subsequently – *the syntax change slightly*: we loaded GeoPandas and not Pandas, together with a specific function. Once loaded, we normalize the features:

```
columns_to_normalize = ['nature', 'entertainment',
                        'transports', 'art', 'sport',
                        'tourism', 'security', 'accessibility']

for column in columns_to_normalize:
    gdf[column] = [float(x) for x in gdf[column]]

    # normalization
    gdf[column] = (gdf[column] - gdf[column].mean())
    gdf[column] = gdf[column] / gdf[column].std()

    # removing outliers
    low = gdf[column].quantile(0.01)
```

```
    gdf.loc[gdf[column] < low, column] = low

    high = gdf[column].quantile(0.99)
    gdf.loc[gdf[column] > high, column] = high
```

Then, we create a list for each coordinate of the cell vertices:

```
lat_min_list = [float(x) for x in gdf.lat_min.to_list()]
lat_max_list = [float(x) for x in gdf.lat_max.to_list()]

lon_min_list = [float(x) for x in gdf.lon_min.to_list()]
lon_max_list = [float(x) for x in gdf.lon_max.to_list()]
```

And we build the polygons by iterating over each cell:

```
polygons = []
zipped_lists = zip(lat_min_list, lat_max_list,
                   lon_min_list, lon_max_list)

for lat_min, lat_max, lon_min, lon_max in zipped_list:
    polygons.append(Polygon([(lon_max, lat_max), (lon_min, lat_max),
                             (lon_min, lat_min), (lon_max, lat_min)]))
```

Finally, we save the polygons into a new feature called "geometry" and export it:

```
gdf['geometry'] = polygons

exporting_df = gdf.drop(['geometry', 'row', 'column',
                         'lat_min', 'lat_max',
                         'lon_min', 'lon_max'], axis =1)

exporting_df.to_csv('3.final_map_converted.csv')
```

2.6 Training the Models

Coming to the main part, we *first create a model with the dataset as is*, then we try to expand it using several different methods. In this phase, we can use also computational-complex models because we have reduced the dimensionality to a grid. To make a comparison, in the Venice case study, we start with around 2.5 millions observations and we end – in this step – with about **30.000 observations**.

We start, as always, by importing the dataset and slicing it, we get rid of the coordinates values and keep only the row and cell number (stored in the index, here called "Unnamed: 0" by default):

```
df = pd.read_csv('1.filtered_df.csv')
df.index = df['Unnamed: 0']
df.drop(['Unnamed: 0', 'row', 'column',
         'lat_min', 'lat_max', 'lon_min', 'lon_max'],
         axis = 1, inplace = True)
```

At this point the dataset is composed of 8 columns – for each macro-category – and one row for each cell.

Next, we import the data science library, PyCaret:

```
from pycaret.regression import *

exp = setup(df, target = 'tourism')
```

The syntax, here, is extremely simple due to the advancement of the library. PyCaret is a **specific-purpose package** that helps doing "Auto-ML" by reducing at its minimum the number of commands and parameters needed.

In this step, we set up the experiment, which essentially means that we are loading the dataset into PyCaret parser. The output of this function is a *grid with the characteristic of the dataset*, for example: number of numeric features, data shape, and the target type. In the table, there also will be the **standard hyper-parameters** that will be used for the machine learning part: k-fold, train/test shape, number of CPUs to use.

```
best = compare_models()
```

Using this simple command, PyCaret will take the dataframe passed in the experiment and will test different types of models on it while returning different metrics:

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.3841	1.1876	1.0847	0.4800	0.3405	0.7429	0.6870
gbr	Gradient Boosting Regressor	0.3828	1.2293	1.1052	0.4613	0.3326	0.7054	0.5920
et	Extra Trees Regressor	0.3921	1.2241	1.1019	0.4542	0.3720	0.7713	0.5540
rf	Random Forest Regressor	0.3946	1.2745	1.1226	0.4414	0.3646	0.7486	0.3640
br	Bayesian Ridge	0.3855	1.3489	1.1470	0.4345	0.3488	0.7601	0.1680
lr	Linear Regression	0.3855	1.3495	1.1473	0.4342	0.3489	0.7603	0.7960
ridge	Ridge Regression	0.3855	1.3495	1.1473	0.4342	0.3489	0.7603	0.0230
lar	Least Angle Regression	0.3855	1.3495	1.1473	0.4342	0.3489	0.7603	0.1840
en	Elastic Net	0.4009	1.3985	1.1714	0.4200	0.3572	0.7635	0.0700
knn	K Neighbors Regressor	0.3847	1.3806	1.1658	0.4106	0.3737	0.7620	1.2130
omp	Orthogonal Matching Pursuit	0.4326	1.4372	1.1820	0.4046	0.3578	0.7815	0.1700
lasso	Lasso Regression	0.4026	1.4514	1.1935	0.4014	0.3598	0.7571	0.2280
huber	Huber Regressor	0.2968	1.6062	1.2538	0.3436	0.3380	0.8483	1.9650
dt	Decision Tree Regressor	0.4462	2.0935	1.4372	0.1010	0.4255	0.9742	0.1300
llar	Lasso Least Angle Regression	0.5424	2.4729	1.5538	-0.0004	0.4127	0.7611	0.2240
dummy	Dummy Regressor	0.5424	2.4729	1.5538	-0.0004	0.4127	0.7611	0.0560
ada	AdaBoost Regressor	1.2685	3.7319	1.9018	-0.6433	0.7359	1.2168	0.5060
par	Passive Aggressive Regressor	1.4143	26.3586	3.4519	-8.4906	0.6076	2.2509	0.2680

Figure 2.2: Output of the "compare_models" function

We have a model for each row and a metric for each column. The metrics computed depend on the type of problem at hand (regression vs classification), but – in general – they help greatly in the initial phases of modelling.

It's also interesting to explore the **visual aids** used: we can see – on the last column – the time required to compute the model over a light-grey background while the best value for each metric is highlighted in yellow. This makes the comparisons between the models very easy.

In the last version of PyCaret, it's also possible to create an online dashboard for a model with a simple command:

```
dashboard(best)
```

And this is one of the views:

Figure 2.3: Online Dashboard's Tab visualizing the Model's Statistics

Finally, we tune the model using the integrated function:

```
tuned_model = tune_model(best)
```

This command allow us to **tune the hyperparameters** of the model by exploring different combination and

performing a k-fold on each one of them [14]. Of course, it's extremely computationally-intensive, so the advice is to only use it on the simpler models.

One way around that is to pass a custom grid as a parameter.

This conclude the first part on the initial model, now we work on extending it by using several different methods.

2.7 Extending the Model

On this section, we will work to **increase the accuracy** (alias reducing the MAE/MSE) of the best model (or models). We will first explore expanding the features by computing their polynomial combination (only to the second degree, to avoid overfitting) and interaction effects. We follow by building ensemble models using the most common techniques: bagged learning, boosting learning, blender learning, stacking learning. Finally, we will create *new* features by taking into consideration not only the values inside the cell, but also the values of the near cells (extending the dataframe by a factor of 8).

2.7.1 Interaction Effects and Polynomial Features

Interaction effects and polynomial features can be computed by using the **product between two features**. Formally, an interaction feature Z is created when it's multiplied by two features belonging to the dataset, X and Y:

$$z_i = x_i \cdot y_i \ \ with \ \ i \in N \wedge X, Y \in F \tag{2.3}$$

where F is the set of all the features, N is the sample size, i is the index of the observation, x_i is the observation's value for feature X, and y_i is the observation's value for feature Y

If $X = Y$ we compute a second-degree polynomial feature:

$$z_i = x_i \cdot y_i = x_i \cdot x_i = x_i^2 \ \ with \ \ i \in N \wedge X \in F \tag{2.4}$$

Essentially, if we take the **cartesian product** of the set of features we will produce a matrix that will have as off-diagonal elements the interaction features and as diagonal elements the second-degree polynomial features.

We can notice that we *compute two times the interaction effects*: this fact raises from the commutative property ($a \cdot b = b \cdot a$) so, if we iterate over the same set two times using the cartesian product, we will produce two triangular matrix with identical values on specular positions.

To realize this technically, we start by trying to **extract more information** from the feature that we have, meaning that we will slightly increase the fitting to see if the results improve:

```
expanded_df = df.copy()
columns = expanded_df.columns.to_list()
columns.remove('tourism')
```

First, we copy the dataframe, then we create a list with all the columns (necessary to iterate over in the next passage) and we remove the target variable, otherwise we will create interaction features with it.

```
for a in columns:
    for b in columns:
        expanded_df[a + ' x ' + b ] = expanded_df[a] * expanded_df[b]
```

Next, we set up a double loop where we create a new feature – composed by the product of two source features – for each iteration.

In this chunk of code, we are creating at the same time the polynomial features (when $a = b$) and the interaction effects (when $a \neq b$).

Then, we simply follow the same pipeline as before:

```
expanded_df['tourism'] = df.tourism
exp = setup(expanded_df, target = 'tourism')
best_expanded = compare_models()
```

2.7.2 Ensemble Models

The computation of the ensemble models is as easy as the past steps:

```
# Compute Bagged Model
bagged = ensemble_model(best, method = 'Bagging')

# Compute Boosting Model
boosting = ensemble_model(best, method = 'Boosting')

# Compute Blender Model
blender = blend_models(bests) # more than one model

# Compute Stacking Model
stacking = stack_models(bests) # more than one model
```

All these models respect different assumptions and may improve the results or not based on the context.

2.7.3 Adding neighboring Cells

The assumption, for this step, is that the touristic potential of a cell is influenced not only by the point of interest in that cell, but also from the neighboring cells.

We import the dataframe:

```
df = pd.read_csv('1. complete_df.csv')

df.set_index('Unnamed: 0', inplace = True)
```

And we initialize the variable that we will use to iterate on:

```
cardinal_directions = ["N", "S", "E", "W"]
diagonal_directions = ["NW", "NE", "SW", "SE"]

categories = ['nature', entertainment', 'transports', 'art',
              'sport', 'tourism', 'security', 'accessibility']
```

We have **8 total directions**: 4 cardinal and 4 diagonal. We also need to account for the fact that there are 4 cells, at the vertices of the grid, that don't have 5 of that directions.

The best way to tackle this problem is to break it down into smaller parts. For this reason, we start by defining a function for the cardinal feature and by exploring every direction:

```
def add_cardinal_features(source_df, categories_names, direction: str):

    new_categories_names = []
    df = source_df.copy()
    df[['row', 'column']] = df[['row', 'column']].astype(int)
```

We can see that the function takes in **4 arguments**: the source dataframe, the categories names and the direction we need to process. After that, we create a new copy of the source dataframe and enforce the integer type on the row and column features. This is necessary because we will use mathematical operations over these features and Pandas raise errors if there are text variables involved.

```
    # Add the north cell to the dataframe
    if direction == 'N':
        for category_name in categories_names:
            new_categories_names.append(category_name + '_north')

        df = df.loc[df.row > 1, ['row', 'column'] + categories_names]
        df['row'] = df['row'] - 1
```

The rest of the function is **specific** – but similar – to the direction selected. In each of them, we create a new category name for the feature that we are computing, then we slice the dataset accordingly. For example, in the north direction we manipulate the dataframe by removing 1 from the row feature, essentially moving all the dataframe down by one row.

This is because the north cell of a cell is contained in the upper row. If we extended this reasoning to every cell, then we just need to switch everything by one.

```
    # Add the south-east cell to the dataframe
    if direction == 'SE':
        for category_name in categories_names:
            new_categories_names.append(category_name + '_south_east')
```

```
    df = df.loc[(df.row < df.row.max()) &
                (df.column < df.column.max()),
                ['row', 'column'] + categories]

    df['row'] = df['row'] + 1
    df['column'] = df['column'] + 1
```

A similar line of work was followed by the other directions (switching by 1 either rows or columns, or both). Then, the function ends with:

```
    df[['row', 'column']] = df[['row', 'column']].astype(str)
    df.columns = ['row', 'column'] + new_categories_names
    df.index = 'R' + df.row + 'C' + df.column

    df = df[new_categories_names]

    return df
```

We reconvert the row and column features to the string type and add the new category names created. Finally, we return *only* the newly computed features.

Using these functions in the process is easy and straightforward:

```
extended_df = df.copy()

new_df = add_cardinal_features(df, categories, direction = 'W')
extended_df = pd.concat([extended_df, new_df], axis = 1)

new_df = add_diagonal_features(df, categories, direction = 'SE')
extended_df = pd.concat([extended_df, new_df], axis = 1)

# Continue for all 8 directions
...
```

And, we just need to complete the table by filling it with 0 whenever there are null values:

```
extended_df.fillna(value = 0, inplace = True)
```

We can already **use the pipeline** we have seen before with PyCaret, but first we need to drop all the features directly related with the target variable (e.g. "tourism_north", "tourism_south_east"):

```
extended_df.drop(list(df.filter(regex = 'tourism_')),
                 axis = 1, inplace = True)

exp = setup(extended_df, target = 'tourism')
best = compare_models()
```

Generally, this step works pretty well because it's based on a *solid assumption* but it's important to notice that it's also **extremely size-dependent**. If we choose an area too big with few cells, it's possible that some of this interaction between neighbours will be lost because the area of each is so big that it's its own "touristic" environment per se.

2.7.4 Computing the Final Predictions

We compute the final prediction by **splitting the dataset into three parts** and by creating independent model for each one of them, using the other two parts as training/test dataset.

We start by loading the dataset:

```
df = pd.read_csv('5.extended_df.csv')
df.index = df['Unnamed: 0']

df.drop(['Unnamed: 0', 'row', 'column',
         'lat_min', 'lat_max', 'lon_min', 'lon_max'],
         axis = 1, inplace = True)

df.drop(list(df.filter(regex = 'tourism_')),axis = 1, inplace = True)
```

We drop both the coordinates variables and the tourism-related variable, leaving only the original "tourism" feature to be predicted on.

```
splits = 3
splitted_df_list = np.array_split(df, splits)

prediction_df_list = []
```

We define the number of splits and use the .array_split function from the NumPy package to create three separate list containing the indices of the source dataframe. We also initialize the list that will contain the predictions.

We start the loop by iterating over each element of the indices list and executing the division between the training/test dataframe and the validation dataframe:

```
for splitted_df in splitted_df_list:
    training_df = df.drop(splitted_df.index)
    prediction_df = splitted_df
```

We continue in the loop by following the PyCaret pipeline and using the "predict_model" function to generate the predictions over the validation set:

```
    exp = setup(training_df, target = 'tourism')
    best = compare_models()
    predictions = predict_model(best, data = prediction_df)
```

Here, we are displaying the performance table referred to *last* training_df. We complete the loop by re-setting the index (PyCaret doens't keep it when performing the predictions) and appending the final results to the dedicated list.

```
predictions = predictions[['prediction_label']]
predictions.index = prediction_df.index
prediction_df_list.append(predictions)
```

Finally, we merge all the predictions into a dataframe and export them to a .csv:

```
predictions_concat = pd.concat(prediction_df_list)

export_df = pd.read_csv('final_map_converted.csv')
```

2.8 Model Visualizations

Having the final maps, we can create the interactive visualizations. To do so, we'll rely on an online service called Carto: a geo-spatial data platform.

Once registered, the process to import the file into the service is fairly easy to follow:

1. Take the file containing the data for the interactive visualizations ("final_map_converted.csv")
2. Go to the maps section
3. Click on the "New Map" blue button on the top right
4. Click on the "Add source from..." blue button on the low left
5. Go to the "Import file" tab
6. Add the file and click on "Continue"
7. On this step, we need to choose a location to host our file, both organization_data/shared and organization_data/shared_us are ok
8. Finally, we click on "Add source"

After this procedure, we can **build different layers**, depending on the data type. For this project, we will use only 2D visualizations because they're the easiest to read and interpret.

We can use this service to first explore the layers composing the grid dataset: entertainment, sport, nature, art, accessibility, transport, security.

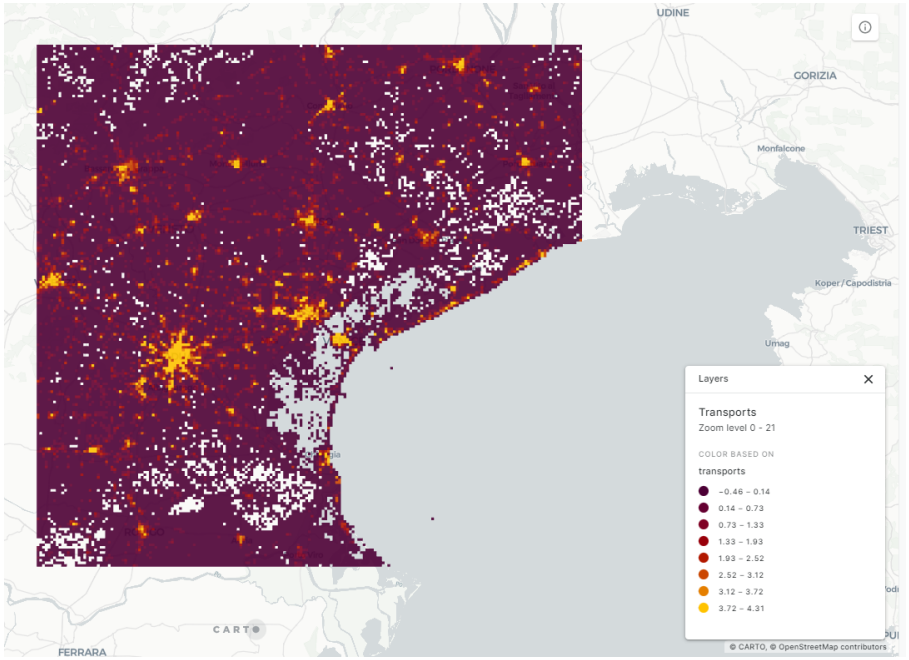

Figure 2.4: Interactive Map showing the Density of Transports in the Venice's Case Study

After that, we can plot the actual tourism and the predicted tourism:

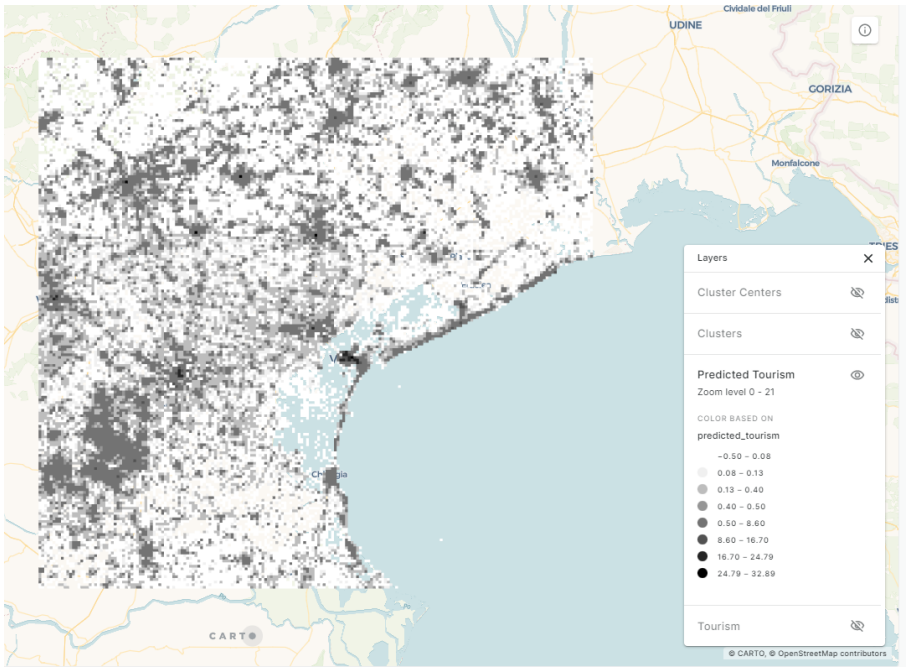

Figure 2.5: Interactive Map showing the Predicted Tourism in the Venice's Case Study

We can clearly see that the dataset become much more "readable" compared to before. It's also easier to spot strange correlations or outliers.

Each row of the dataset corresponds with a square/cell of the visualization.

The map with every layers for the Venice case study is freely accessible at the link:
`https://pinea.app.carto.com/map/810354db-9994-4bbf-905a-70da8d32a010`

2.8.1 Clustering

Having the tourism, both actual and predicted, as a grid is certainly helpful. But, to choose the right candidates, we need to understand and **define the areas of major interest**.

It's possible to use unsupervised ML, like clustering, to assign a cluster to each point in the map.

We start by importing the predictions:

```
prediction_df = pd.read_csv('../030-training_models/7.final_prediction.csv')
prediction_df.set_index('field_1', inplace = True)
prediction_df = prediction_df[['predicted_tourism', 'geometry']]
```

The problem with this dataset is that contains the coordinates as polygons, but to do machine learning over that we need to have them splitted. The easiest solution is to directly import the coordinates from a dataset that have them:

```
gdf = pd.read_csv('../030-training_models/1.filtered_df.csv')
gdf.set_index('Unnamed: 0', inplace = True)
gdf = gdf[['lat_min', 'lat_max', 'lon_min', 'lon_max']]

# Merge the dataset with the final prediction and the one with the splitted coordinates
df = pd.concat([prediction_df, gdf], axis = 1)
```

We then filter out the low-scoring cells:

```
filtered_df = df.loc[df.predicted_tourism > df.predicted_tourism.quantile(0.85)]
```

And we perform a KMeans:

```
from sklearn.cluster import KMeans
kmeans_predictor = Kmeans(n_clusters=20)
filter_df_coordinates = filtered_df[['lat_min', 'lat_max', 'lon_min', 'lon_max']]

kmeans = kmeans_predictor.fit_predict(filter_df_coordinates)
```

As per sklearn documentation [15], the KMeans uses a within-cluster sum-of-squares criterion to choose the centroids:

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2) \tag{2.5}$$

where:

- n is the number of samples
- μ_j is the mean of the samples in the cluster
- C is the desired number of disjoint clusters
- x_i is the observation's value

Essentially, it tries to find the point that *minimize the distance between all the point assigned to the cluster C*.
Finally, we export the dataset with the polygon coordinates and the cluster assignment:

```
polygon_coordinates = filtered_df['geometry']
cluster_assignments = pd.Series(kmeans + 1, index = filtered_df.index, name = 'cluster')

exporting_df = pd.concat(polygon_coordinates, cluster_assignments], axis = 1)
exporting_df.to_csv('1.clustered_df.csv')
```

We can also compute the centroids for each cluster, to make easier visualizations:

```
coordinates = filter_df_coordinates
cluster_coordinates_list = kmeans_predictor.fit(coordinates]).cluster_centers_
cluster_coordinates = pd.DataFrame(cluster_coordinates_list)

cluster_coordinates.columns = ['lat_min', 'lat_max', 'lon_min', 'lon_max']

#Start from 1 and not from 0
cluster_coordinates.index = cluster_coordinates.index + 1
```

And we export the average between the maximum/minimum of each coordinate for each centroid:

```
lat_mins = cluster_coordinates['lat_min']
lat_maxs = cluster_coordinates['lat_max']

lon_mins = cluster_coordinates['lon_min']
lon_maxs = cluster_coordinates['lon_max']
cluster_coordinates['lat'] = (lat mins + lat_maxs) / 2
cluster_coordinates['lon'] = (lon_mins + lon_maxs) / 2
cluster_coordinates = cluster_coordinates[['lat', 'lon']]

cluster_coordinates.to_csv('1.cluster_coordinates.csv')
```

This is an example of the final results:

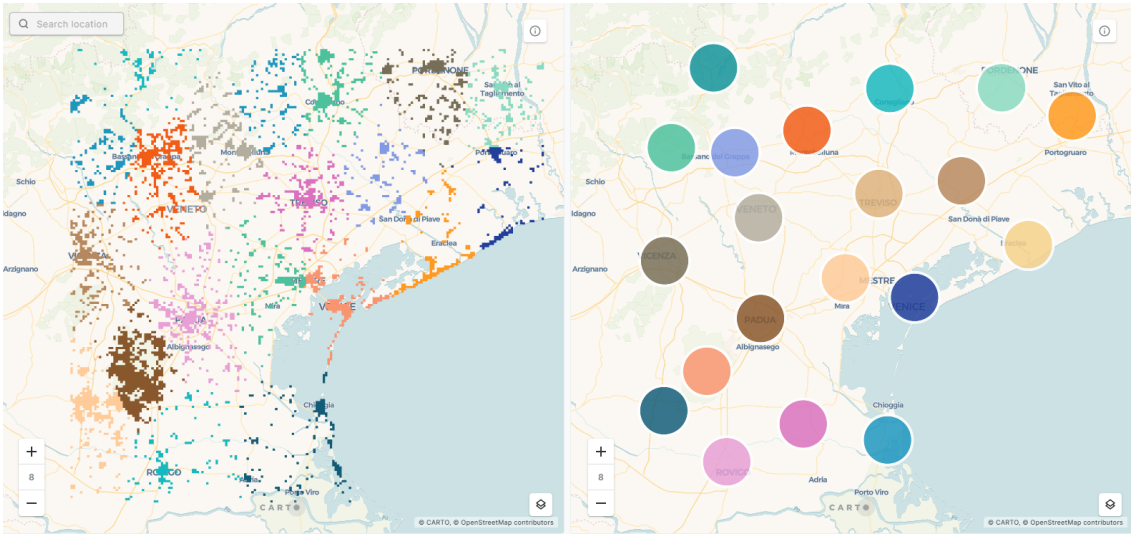

Figure 2.6: Interactive Map showing the Clustering Assignment, on the right, and the Centroids, on the left

The map is freely accessible at the link:
`https://pinea.app.carto.com/map/fee78425-981f-4226-8011-6523803b70f0`

2.9 Summary of the Manipulations

Starting from a **list of different point of interest**, we first removed the observations that were too far away from the city of interest, resulting into a **square grid**. Then, we sliced the grid into smaller cell and we assigned each observation to a cell.

Finally, we built some models using both **simple** and **ensemble** models. The results were much better when the model considered also the neighboring cells.

To visualize the results, we used Carto and **clustering methods** to define potential areas of influence. The final output, then, is a series of interactive infographics about the area of interest.

3
Social Part

3.1 General Approach

The second part of the project is *less technical* and involves a **manual "screening"** of the clusters detected by the AI. This section is extremely important, because even though the computers can help us greatly reduce the time to detect high-potential areas, it's also true that they still not are "hard AI": they do not properly think.

Following from this fact, it's simple to understand why *the final choice must be done by a human and not a machine*.

The part take as input the last step of the technical part (the interactive maps) and go all the way until the final check for feasibility. During the process, additional – local – filters will be applied.

After the first two steps, we take a more **social-oriented approach** and look for a match between the point of interest in the cluster and the potential tourists. Once found a significant niche, the final product will be created: a series of suggested tours coupled with story-telling elements, to create a suggestive and immersive experience for the potential tourist.

Finally, the process will be concluded with the presentation to the stakeholders with both qualitative and quantitative data.

3.2 Data Visualization Analysis

In this phase, we carefully **analyze the base layers** and then we proceed to examine the differences between the predicted tourism and the actual tourism.

Then, we analyze the cluster – both the distribution of the geographical space and the centroids – to assess the potential of each cluster. In this phase, we also remove all the already famous localities because, by growing *their*

tourist flux, we would only move the problem from a bigger city to a smaller one.

For the analysis, we will use a simple **guideline**:

1. Declare what basic category the visualization represent
2. Describe the legend: lowest/highest value, bins, colors
3. Describe the *highest* scoring areas and trace reasons for that
4. Describe the *lowest* scoring areas and trace reasons for that
5. Assess potential interactions between the areas analyzed
6. Assess potential interactions with other layers

By following the outline, we can be sure to pick up any important characteristics of the map and, by repeating it for each layer, we can start to understand the complex networks that are often born in European grounds.

3.3 Adding Contextual Filters

The filter phase is something that can be extremely variable given different sets of conditions. In this project, we will use an additional filter: we will only consider areas that are reachable, by public transport, in under 75 minutes from the city centre.

This way, it will be possible, for a future tourist, to visit the alternative areas into a **one-day trip** without losing too much time travelling.

But this is only an example: different stakeholders may put on the table different requirements. Especially in the European Union, where many of the new initiatives are sponsored by public grants, there may be additional filters to implement.

We can use a well-tested schema to share with the stakeholders:

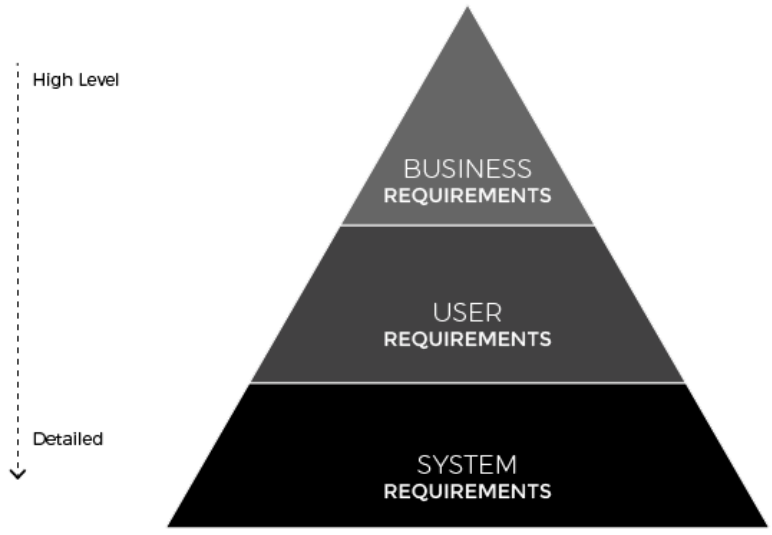

Figure 3.1: Requirement Pyramid Template

We can notice that at the top of the pyramid there are the **business requirements**: they are the reason of the project and – usually – the less specific ones. In this scenario, they can be anything from the tourist flow increase to the reduction of trash in the city. These requirements can be discussed with the stakeholders and adjusted for the lower levels.

At the second level, we find the **user requirements**. This are usability requirements that can be established and tested by professionals like UX designers and developers. In this section we find all the details we need to make a usable system: compatibility of the program with other people and systems, accessibility of the initiatives, expectation management or learnability, and so on.

Finally, we find the **systems requirements**: these are usually less negotiable in the short term, because they come from a long legacy and they require many resources to change. It's still useful to discuss and find all of them because they define the environment where we operate. The main pro is that they're easily declarable and very specific.

While doing this process, we will usually find that some of the clusters are not suitable – at least in the short term – for incremented tourist fluxes or new initiatives: this is the first step to reduce the number of candidates. As a general rule, they shouldn't be more than 5.

3.4 Evaluate Point of Interests

This is **the central part**: we go area by area and see what are the most promoted points of Interest or where people go often. In this phase, it's also possible to interview local people and doing some scouting activity to gather data from different type of sources.

This evaluation will maintain the 7 base categories to make easy comparisons with other scenarios. Like for the data visualization analysis, we follow a simple outline here:

1. General description of the point of interest
2. Information about accessibility: opening hours, costs, logistics, presence of barriers, and so on
3. Information about events, initiatives and organizations that manage the point of interest
4. Information about recent trends that involve the point of interest

If repeated for every major point, the resulting table should be able to help understand why people are – or should – visiting that place.

3.5 Evaluate Target Market

Once the natural and cultural heritage of each candidate is clear, it's possible to understand what **kind of activities** and which **kind of people** may be interested.

We know that the tourism sector is on the rise along with its destructive force [16], especially in Italy [17]. We also notice that there is a underlying standardization of the urban space in the bigger cities [18].

There are *several ways* to identify a target market, but the easiest one is to start from the existing local tourism and expand it by changing a characteristics at time. We can use a slightly edited version of the Ansoff Matrix:

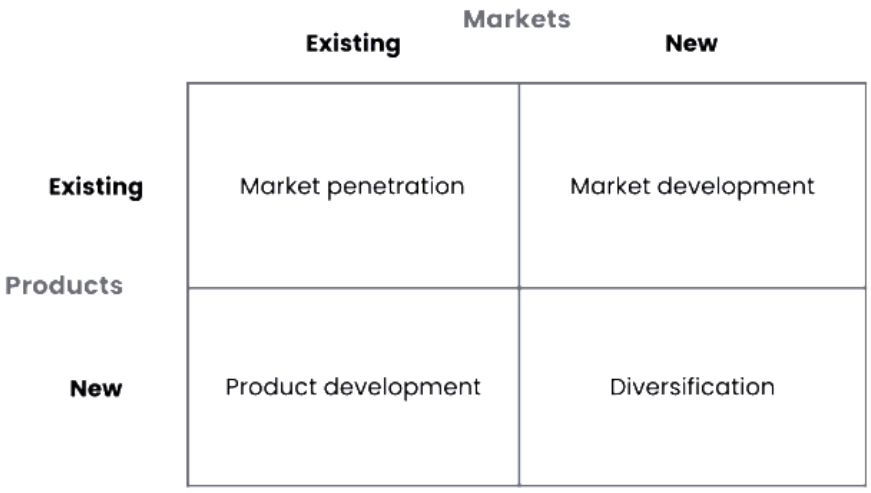

Figure 3.2: Ansoff Matrix Template

In the matrix, on the rows we describe the condition of the market while on the columns we describe the condition of the products/services. Both the market and the services can be either "existing" or "new", relative to the conditions at the moment of the analysis.

The resulting table has 4 cells:

- **Market Penetration:** we are in this scenario when we are talking about points of interest that are already promoted and visited. In this case, the new activities should be designed to increment the number of tourists to the same points of interest.
- **Product Development:** this is the first scenario we consider outside the existing one. Here, we are thinking about different points of interests or activities that can appeal to the already established tourist flux.
- **Market Development:** opposite to the previous case, here we are promoting the same points of interests and activities to different tourists. For example, if we are analyzing an area that is particularly rich from the natural point of view, we might promote it not only to nature-loving tourist but also to people that love outdoor sports (trekking, hiking, trailing, biking, and so on).
- **Diversification:** in this scenario we should try and think to a paradigm change because we are talking about the promotion of different points of interest – compared to the popular ones - sponsored to a different tourist type.

By slowly taking into account each possible scenario, it's possible to conduct more focused discussion with all the stakeholders involved.

3.6 Tour Development

For the tour development section, we will use non-conventional methods to generate ideas. Specifically, we will use the 5 stages of the Design Thinking Process:

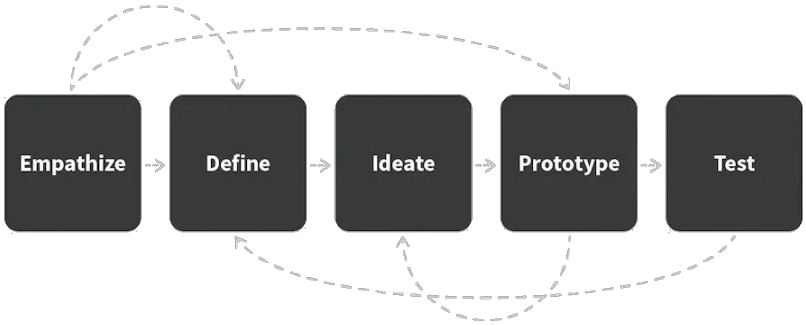

Figure 3.3: 5 Stages of the Design Thinking Process

In the first stage of the process – **empathize** – we should try and put ourselves in the potential tourists' shoes to understand their expectations, emotions and needs. A common practice for this is to create user stories while highlighting every step, obstacle and success for them. The second stage, **"define"**, follows naturally from the first one. At this point, we should "translate" the user stories into requirements and short specifications.

"Ideate" and **"prototype"** are the so-called "divergent phases" were we try and generate as many ideas and prototypes possible. The important thing to remember is that we must always follow the principles and expectations defined in the first two stages.

We end the process by **testing** the prototypes with a sample audience. Theoretically, the design thinking process provides another step after testing: "iterate". This is because, in this branch of knowledge, learning never stops (it's also the reason of the many light grey arrows in the image). So we should always be at least in one of these stages.

While developing the tour, we must take into accounts two points of view. The first one is related to the final user, the tourist, that wants to see something out of routine. The second one is related to the residents: after the initial enthusiasm for a new economic growth, local people will suffer the degenerative effects of a massive flux of people and will start to resent them [19].

3.7 Story-Telling

With all the information gathered, we can also try and build stories about the alternative candidates area. These may use different kind of sources like folklore, artifacts or everyday life.

The important thing is to **generate interest** into the reader while empowering the natural, and artificial, heritage of the location.

We will follow, very *loosely*, the 7 elements of Aristotle's storytelling:

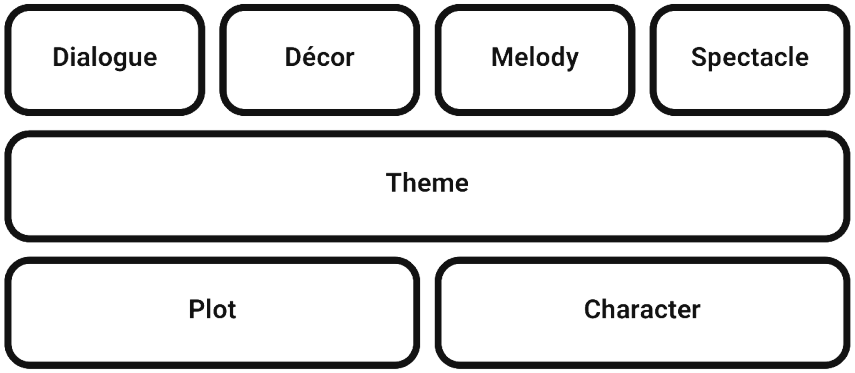

Figure 3.4: 7 Elements of Aristotle's Storytelling

In the schema, there are 7 elements:

- **Dialogue:** the communication used by every characters in different settings. By changing this element, we can set a different tone to the story.
- **Décor:** the physical setting of the story. It can be used combined with the tours to let people immerse themselves into a new experience.
- **Melody:** it's a very particular element. The melody elements comprises everything that amplify the emotions of the characters to the readers. Aristotle thought that every story needs a "chorus" to magnify the facts.
- **Spectacle:** everything regarding plot twists and major moments. They are the "wow" moment of the story.
- **Theme:** the ending goal of the entire story.
- **Plot:** this is the most obvious element. It comprises everything that let the story move forward: obstacles, change in people, goals, and so on.
- **Character:** a brief description of the main characters. This may include: emotions, goals, traumas, abilities, and personality traits.

By combining material experiences (the tours) with immaterial experiences (stories) we can create a truly immersive adventure for everyone. It will also help the stakeholders to promote every aspect of interest or artifact present in the city.

3.8 Checking Sustainability

The final step of the social part (and of the whole project) is to check the feasibility of the tour, initiative and promotional campaigns.

Feasibility is a **broad term** that includes different layers. One of the most accredited framework is the "sustainability pyramid":

Figure 3.5: Sustainability Pyramid

As we can see, there are three main factors that constitute a *sustainable* development. The first one is **economical stability** because, without a constant cash flow, we cannot provide the services and products that should create the development. We can think economic stability as an hygienic factor: we just need to reach a threshold and then we can focus the efforts on the other types of sustainability.

Obliviously, this reasoning is only valid when dealing with public stakeholders and not with private businesses, but there may be a sweet spot where both of them could find a compromise and a way to collaborate.

The second part – **societal responsibility** – involves checking for everything that may be unfair to some part of the populations: obstacles for disabled people, bias against certain ethnicities, discrimination against a gender, and so on. While the economic stability is relatively easy to compute and track, this kind of sustainability may be much more difficult to manage. There are several risks involving hidden minorities or trade-off between costs and usability, plus it's difficult (and very subjective) to track.

The third part is **environmental protection**, but we must read it not from the natural point of view, but also from the cultural. In fact, "environmental" in this setting means "everything that involves the environment were we move". So, this includes theatres, monuments, squares, and so on. This is where most of the cities that suffer from overtourism fail: too many people are difficult to manage and control, and the result is that several art pieces are being destroyed right now.

This is the last filter stage: ideally, in this phase we should finish with no more than 3 potential candidates.

3.9 Final Output

The final output of the project is a **complete package** that includes: numerical dataset, several visualizations, manual evaluations of the most important points and tourism fluxes, and a couple of practical proposal (tours and stories) with a preliminary sustainability analysis done.

This is a starting point.

With the project's output, the stakeholders should decide which project/tours/narratives fund and promote (it may be all 3 proposal, or maybe just 1). But, surely, during a discussion with so many inputs, we can expect the generation of new point of views and ideas. **That's the breeding ground for change.**

4

Case Studies

4.1 Selection Criteria

Coming to the practical part of the project, we choose 3 possible scenarios: **Venice**, **Barcelona** and **Amsterdam**. These cities were the ones that came up most often in the overtourism scientific literature and all of them have many characteristics in common:

- They are *coastal* cities: so they suffer specifically from the cruise tourism
- They are *European* cities: this means that the dataset and the activities that can be performed respect similar guidelines, especially regarding social welfare
- All of them have already put in place *restrictions* to the tourism. This means that they are already aware and pro-active in battling it
- They are commercial and historical **hubs**, so they had a strong influence in the surrounding territory in the past

Let's start with the first one.

4.2 Italy - Venice

Being a city built on a group of island, Venice it's certainly one of the most suggestive places in the world. Given this fame, the city has seen a **constant growth** in the number of arrivals - in 2019 they were almost 6 millions [20] - and an increasingly difficulty in managing all these tourists.

The city authorities have put in place a **limitation** for the future travellers: they must book their day-trip in advance [21], with a maximum capacity of 40,000 - 50,000 people per day. It's important to know that just 6% of the travellers are day-trippers and that only 50,000 - 60,000 live in the historic centre.

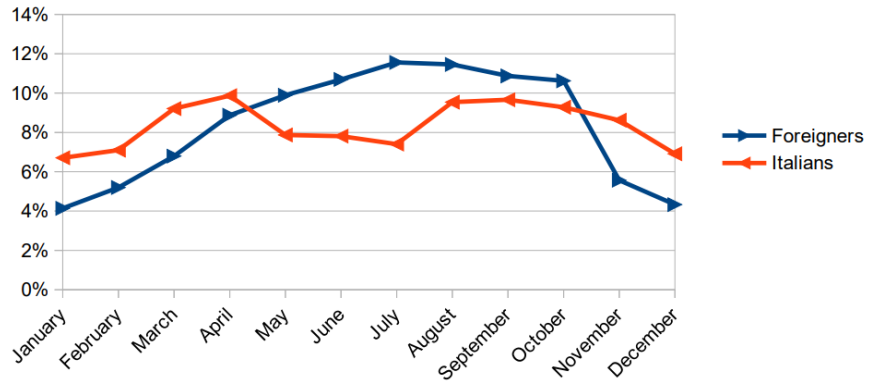

Figure 4.1: Proportion of Italians and Foreigners during the Year

The average stay is around 2.5 nights and less than 15% of the tourists are Italians. The number changes if we look at the *metropolitan area* of Venice: the average stay grow to almost 4 nights and the number of foreigners "shrink" to 75%. We can see that the proportion between foreign and Italian tourist change during the summer with the Italians being less likely to visit the city.

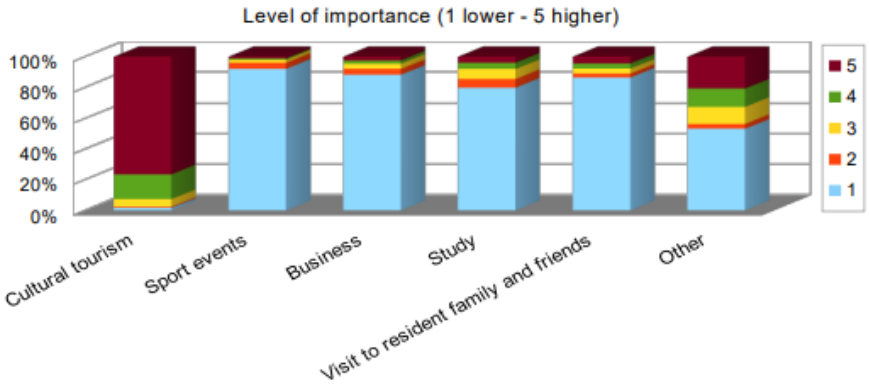

Figure 4.2: Purpose of the Visit, scored by Importance

It's interesting to notice that, almost always, people that uses hotels tend to stay less time compared to other type of accommodations (B&Bs, hostels, campsites, for example). The visitors age is equally distributed between the three categories 26-45 y.o, 46-60 y.o and 60+ y.o. The same thing is valid for gender: there are slightly more male visitors (54%). 76% of the tourists travelled with a family of less than 4 children, around half of them had only one child.

Most of the tourists possessed **advanced education levels** (bachelor or higher): 80%. Also, just 5% of them used a travel agency to book the visit and the stay, most of them (82%) used online channels and autonomously organized the trip (97%). The majority used sites like Booking and AirBnB.

Looking at the reason for the visit, we can see that most of the people visit Venice for **cultural** reasons and not for sport/business/study/family related reasons. Around 22% of the tourists visit at least one other major art city in Italy (Florence, Milan, Naples, Rome) while the remaining part refrain to do so [20].

4.2.1 Analysis

We download the dataset from GeoFabrik and load it on the notebook pipeline. The resulting visualizations are extremely informative, we can evaluate some of them directly on the paper.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.3841	1.1876	1.0847	0.4800	0.3405	0.7429	0.6870
gbr	Gradient Boosting Regressor	0.3828	1.2293	1.1052	0.4613	0.3326	0.7054	0.5920
et	Extra Trees Regressor	0.3921	1.2241	1.1019	0.4542	0.3720	0.7713	0.5540
rf	Random Forest Regressor	0.3946	1.2745	1.1226	0.4414	0.3646	0.7486	0.3640
br	Bayesian Ridge	0.3855	1.3489	1.1470	0.4345	0.3488	0.7601	0.1680
lr	Linear Regression	0.3855	1.3495	1.1473	0.4342	0.3489	0.7603	0.7960
ridge	Ridge Regression	0.3855	1.3495	1.1473	0.4342	0.3489	0.7603	0.0230
lar	Least Angle Regression	0.3855	1.3495	1.1473	0.4342	0.3489	0.7603	0.1840
en	Elastic Net	0.4009	1.3985	1.1714	0.4200	0.3572	0.7635	0.0700
knn	K Neighbors Regressor	0.3847	1.3806	1.1658	0.4106	0.3737	0.7620	1.2130
omp	Orthogonal Matching Pursuit	0.4326	1.4372	1.1820	0.4046	0.3578	0.7815	0.1700
lasso	Lasso Regression	0.4026	1.4514	1.1935	0.4014	0.3598	0.7571	0.2280
huber	Huber Regressor	0.2968	1.6062	1.2538	0.3436	0.3380	0.8483	1.9650
dt	Decision Tree Regressor	0.4462	2.0935	1.4372	0.1010	0.4255	0.9742	0.1300
llar	Lasso Least Angle Regression	0.5424	2.4729	1.5538	-0.0004	0.4127	0.7611	0.2240
dummy	Dummy Regressor	0.5424	2.4729	1.5538	-0.0004	0.4127	0.7611	0.0560
ada	AdaBoost Regressor	1.2685	3.7319	1.9018	-0.6433	0.7359	1.2168	0.5060
par	Passive Aggressive Regressor	1.4143	26.3586	3.4519	-8.4906	0.6076	2.2509	0.2680

Figure 4.3: Models' Performance for Venice's Case Study

The model used is the **Light Gradient Boosting Machine**. This algorithm uses two novel techniques compared to the classic Gradient Boosting Decision Trees: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

The GOSS uses the variance gain V at a splitting feature j for threshold value d to split the data in left and right child nodes.

$$V_{j|O}(d) = \frac{1}{n_O}\left(\frac{\sum_{\{x_i \in O\, x_{ij} \leq d\}} g_i^2}{n^j_{l|O}(d)} + \frac{\sum_{\{x_i \in O\, x_{ij} > d\}} g_i^2}{n^j_{r|O}(d)}\right) \tag{4.1}$$

where:

- j is the splitting feature of the decision tree's node
- O is the training dataset on a fixed node of the decision tree
- d is the splitting point
- n_O is the number of instances from the training set I that belongs to O
- x_i is the training set instance
- x_ij is the instance's value for the splitting feature j
- g_i is the negative gradient of the loss function
- $n^j_{l|O}$ is the number of instances that belongs to O and are lower than or equal threshold value d for feature j
- $n^j_{r|O}$ is the number of instances that belongs to O and are higher than threshold value d for feature j

Then, it ranks the training instances according to their gradients and keep just the top-n observations. The remaining observations will go to the next iteration where the variance will be computed again and the process will be repeated until there are no more observations left. The key advantage of this method is that *compute V on smaller instance subsets*, thus being much more faster compared to the traditional method.

EFB, on the other hand, uses a different assumption: that high-dimensional data is usually very sparse, thus the feature can be "bundled" together without significant loss of accuracy while simultaneously improving speed [22].

Regarding the feature importance plot, on the x-axis we can find the relative importance of the variable, while on the y-axis we can find the name of the variable. It's also clear the connection between Venice's transport and tourist: there are 4 transport-related features between the top-10 most important ones, with the value inside the cell analyzed being the most important one.

We can also understand the importance of the **nature** (2 features) and **entertainment** (2 features). Finally, we see a strong correlation with accessibility but, given the nature of the feature, it's more probable that fluxes of tourism *cause* accessibility.

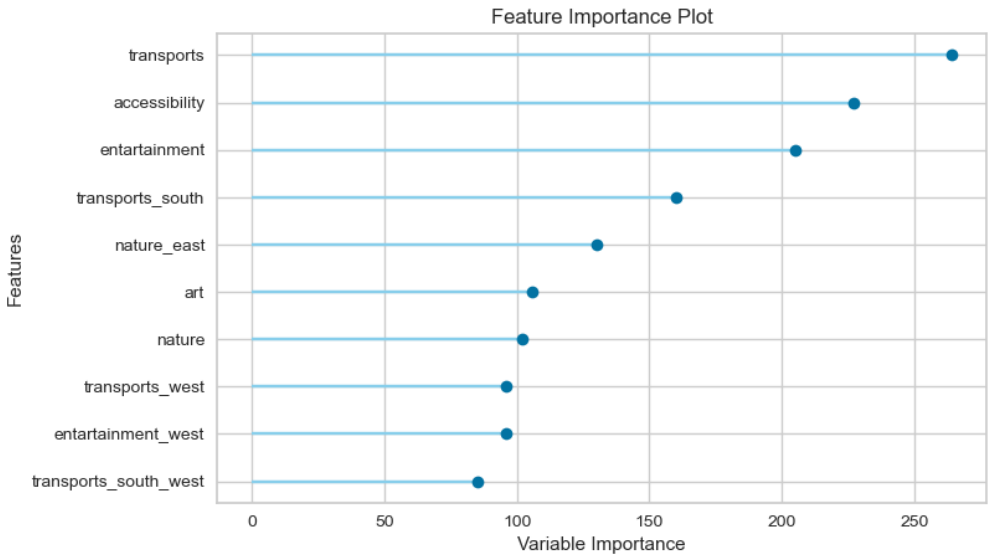

Figure 4.4: Venice's LightGBM Feature Importance Plot

Looking at the actual tourist points in the map, we notice that Venice and Padua are the two main hubs.

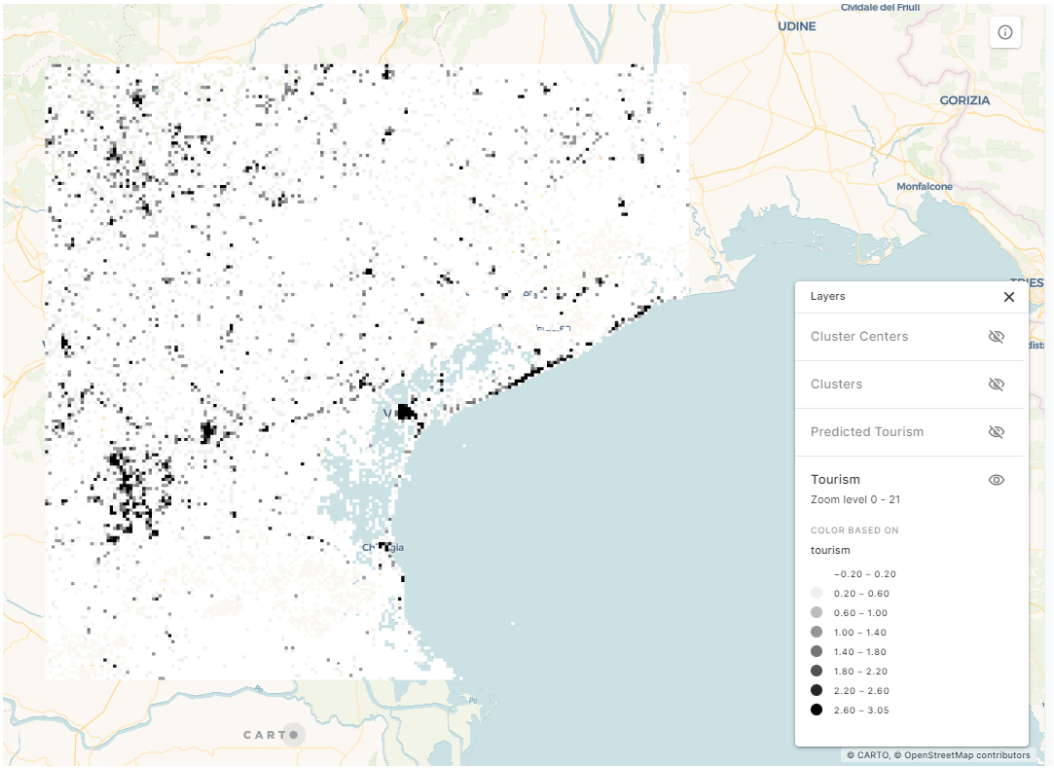

Figure 4.5: Actual tourism of Venice and its Surroundings

But there also a couple of other areas that attract attention: the Colli Euganei Regional Park and all the Venetian east coastline towards Trieste.

A couple of very small points reach the highest concentration of tourism: Treviso and Bassano del Grappa, for example.

The whitest dots have a negative value of -0.20, while the darkest dots are more than 2 standard deviations away from the mean. This means that the distribution is right-skewed and that there are few black dots and many white dots: there are **few optimal tourist spots** a many more that are not so "touristic".

Many of the darkest regions, alias where there is more tourism, follow straight lines, this might indicate a connection with some type of infrastructure. We can verify that information by using the transport map:

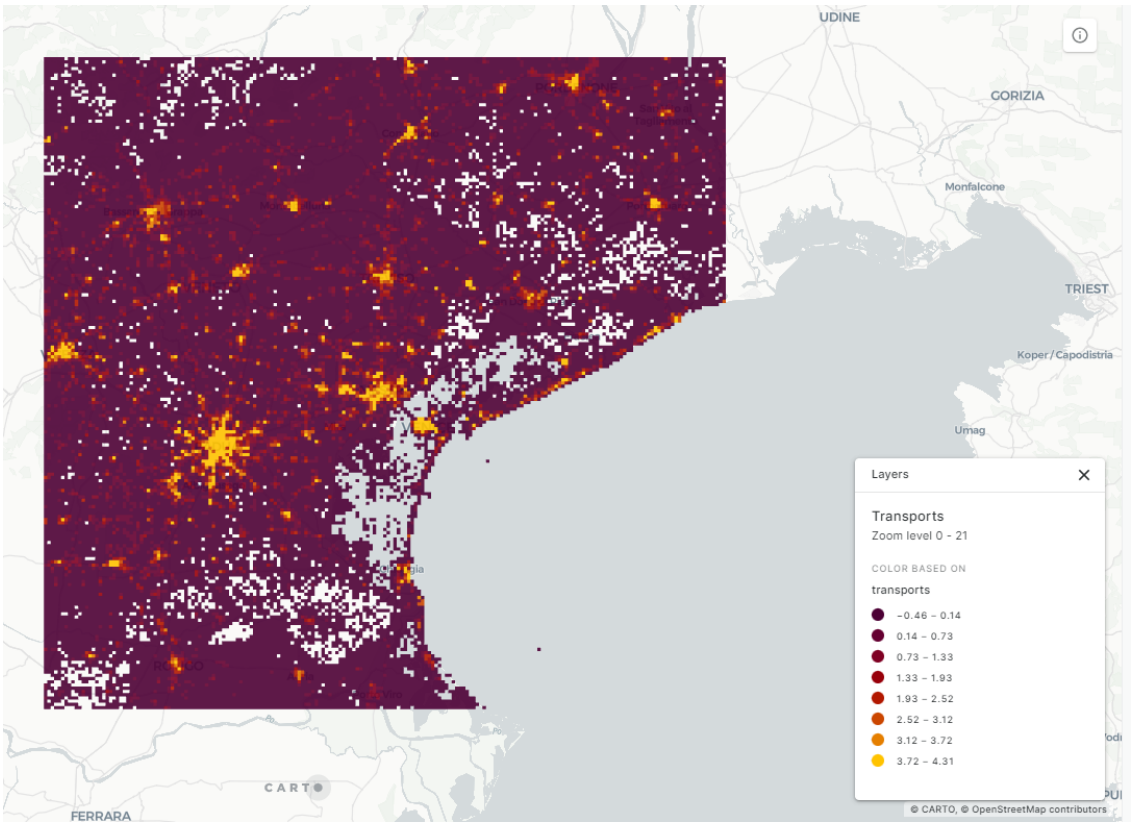

Figure 4.6: Transport Map of Venice and its Surroundings

Here, we have a **distribution similar** to the actual tourism map: many darkest points – with a minimum of -0.46 – and few yellow hubs – with a maximum of 4.31. We can see that there are few areas that are extremely well connected and other areas left disconnected from the transport network.

By using another layers, we can **switch the point of view**: in this map we can spot two huge yellow dots (Padua and Venice) together with several smaller dots (Treviso, Castelfranco Veneto, Vicenza, San Donà di Piave, and so on).

We can clearly notice a sort of transport network, mainly composed of the railway and highway systems, were people can commute to and from the main hub of the region.

From a cultural prospective, as we have seen it's one of the most important, we notice that there are just few darkest points: Venice, Padua, Vicenza, and Treviso.

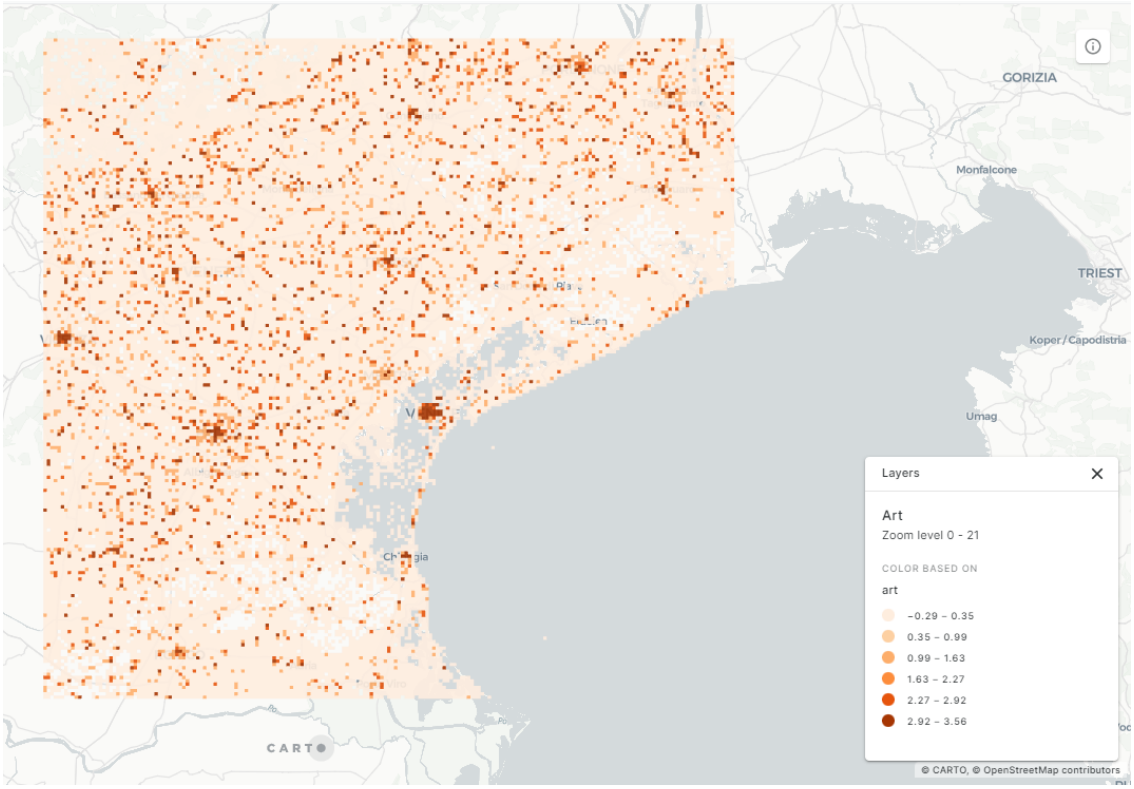

Figure 4.7: Cultural Map of Venice and its Surroundings

Plus, there is a variety of smaller point distributed all across the region. This may be an opportunity for the next steps, because we know that the majority of the tourist go to Venice for cultural purposes. We also note that the area corresponding to the Alps (north-west) and the a portion of the extreme east area (near Marano Lagoon) have a lot **less cultural points**.

In the map, the highest point (dark orange/brown) has a value of 3.58 while the lowest point has a value of -0.29: compared to the transport map, the culture map seems to have less variance given the lower maximum (3.58 vs 4.31) and higher minimum (-0.29 vs -0.46). The 6 colors are assigned based on their value.

It's important to notice that *the value-color pairings are assigned case-by-case*. Even tough the standardization and outlier removal steps helped greatly, there are still scenarios were it's difficult to manipulate the data.

In this case, probably, with more fine tuning we could have obtained a more readable map (like the others). The data also tells a story about what different regions consider "culture": what is culture for an Italian may not

be culture for a Spanish, and vice-versa. Thus, the difficulties in manipulating data while retaining a certain degree of comparability.

Finally, we take a look at the natural heritage map:

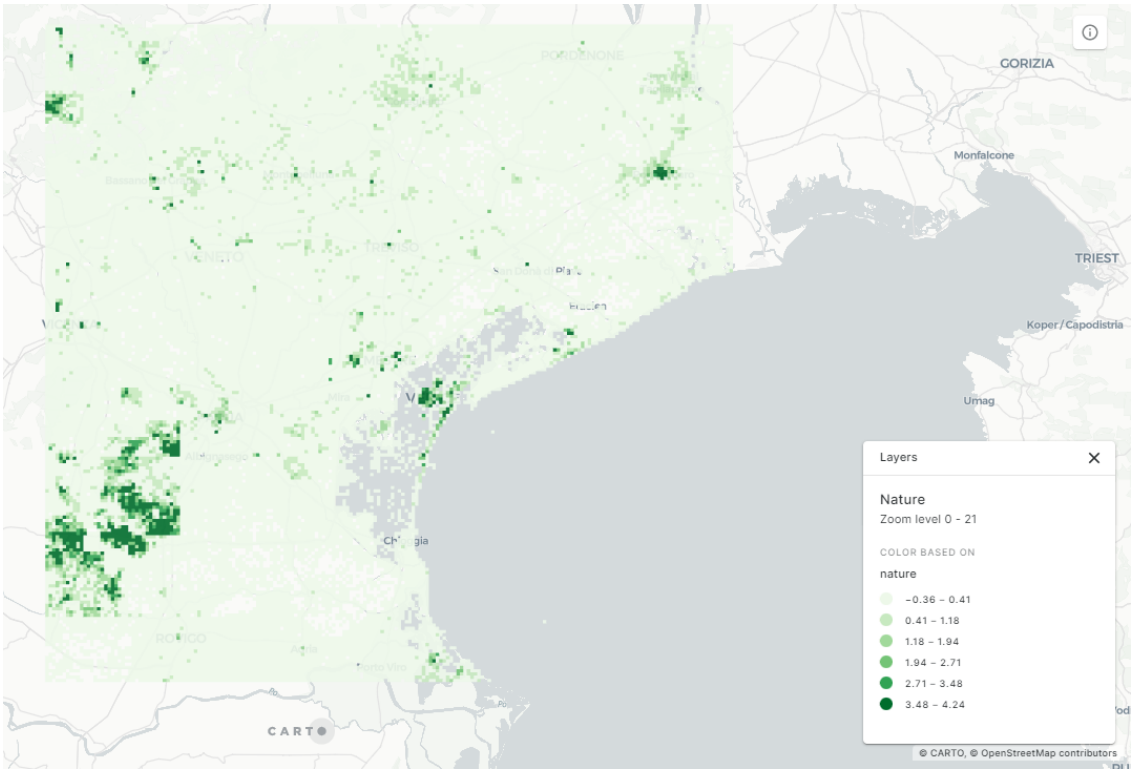

Figure 4.8: Nature Map of Venice and its Surroundings

In this map, the darker the shade of green the more naturalistic point of interest in the area. We have a minimum of -0.36 and a maximum of 4.24.

Again, we can see that the "king" of the region is the **Regional Park of Colli Euganei** but there are several other areas of interest. For example: just outside Venice we can find several naturalistic spots, the same apply for Jesolo and Pordenone.

It's also interesting to notice the shape of the darker areas around Colli Euganei: on the east side, we find a very sharp contrast between high-scoring and low-scoring cells, almost like a line that clearly divides the park with everything else. But the same it's not true on the other sides: we notice a slow decrease in the scores until reaching the lightest color, alias the lower class.

Predicting tourism with this indicators will not be easy. These plots tell us a story about a *region that developed very differently* based on the local scenario and history. There are some areas that extremely suited for cultural tourism and others more suited for entertainment. While looking at these maps we should always remember that they are generated using a "quantitative" process: all the monuments are treated the same, all the streets are treated the same, all the parks are treated the same, and so on.

But we know that tourism doesn't work like that.
Made this brief premise, we can analyze the predicted tourism:

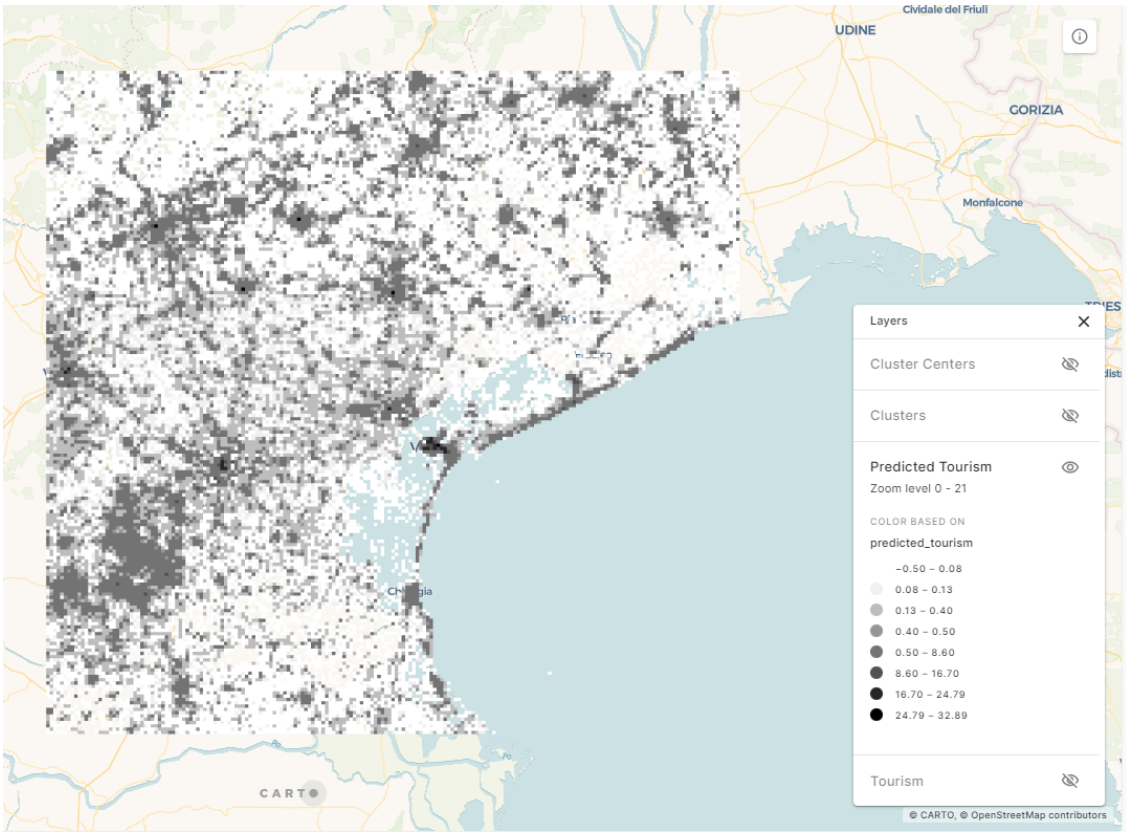

Figure 4.9: Predicted Tourism for Venice and its Surroundings

In this map we use **8 bins** instead of the usual 6, to assess better the differences between the predictions. The results are that the minimum is computed at -0.50 and the maximum at a value of 32.89. This is due the fact that there are **huge outliers** like Venice and Padua that complicate the overall computation.

It's clear that the predictions have **less variance** compared to the actual values, meaning that we have more grey areas and less white/black cells. This is due the fact that the model was tuned to avoid over/undershooting and the top/lower 1% of the prediction were polished to the threshold values.

We notice the effect of the different features on the final score. For example, there is a grey area exactly in correspondence of the Regional Park, so we can infer that the "nature" feature played a major role in determining the color. A similar effect can be seen with the transports: looking at Padua is clear the connection. The "grey" lines that we spot going to north (towards Castelfranco Veneto) and north-west (towards Vicenza) from Padua are probably of that color because of their numerous transport connection like railways, bus stations, and highways.

There is also a *link between the culture and the prediction* the north-west and extreme east areas: in the culture map they were shown as generally uninteresting, so the predictions for them are generally on the lowers part of the range (lighter areas).

As always, both Venice and Padua stands out, but there are many grey areas that we can explore using clustering:

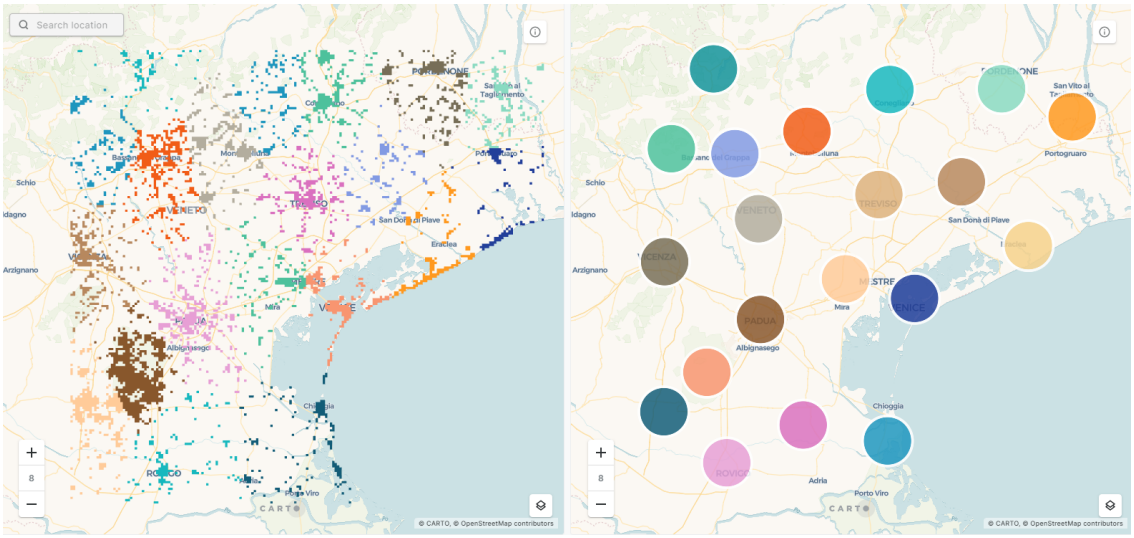

Figure 4.10: Cluster Analysis for Venice's Case Study

On the left map, there are the cells that scored in the **top 15%** . Then, using a clustering algorithm, they were assigned to 20 different clusters (recognizable by the color). On the right side, there are the centroids, one for each cluster. Please notice that the colors between left and right images **don't** match: it was not possible to do so in Carto.

We can identify several interesting clusters: Chioggia, San Donà del Piave, Mirano, Cavarzere, Abano Terme, Pozzonovo, Montebelluna, and Conigliano. All of them are not as notorious as Jesolo, Treviso, Castelfranco Veneto and Padua, but they still score high in several different parameters.

It's also interesting to notice that they are almost equally distributed across the map, we can see an average distance of about 20-30km between each centroid. It's a distance equivalent to about 30/40 minutes of travel time (by car or public transport).

With this knowledge, let's apply the filter.

4.2.2 Filtering

Let's **compute the distance** for each relevant cluster from Venice Santa Lucia using public transport. We use Google Maps and a simulation for a working day with the starting time at 7:00 AM.

Using the table, and considering the time limit of 60 minutes, we can already exclude many of the candidates: **Chioggia**, **Cavarzere**, **Pozzonovo** (it's not possible to reach it without car/bike), and **Montebelluna**.

It makes sense to use public transport at this step because many of the tourist are foreigners and probably don't have access to private means of transport. Plus, using public transports also promote slow and green travel.

We will take just the four nearest cities: San Donà del Piave, Mirano, Abano Terme and Conegliano.

Destination	Travel Time
Chioggia	140 min
San Donà di Piave	46 min
Mirano	39 min
Cavarzere	97 min
Abano Terme	43 min
Pozzonovo	ND
Montebelluna	89 min
Conigliano	50 min

Table 4.1: Travel Time from Venice, using Public Transport, for each Destination

4.2.3 Evaluation

In this step, we are going to evaluate each of the **main candidates** from the step before. The evaluation comprehends: a brief description of the city, the points of interest list, and a couple of possible tourism expansions using the Ansoff Matrix.

Candidate #1: San Donà del Piave

Looking at the first alternative location, we can first look at the most common portals to gather the most valuable point of interests:

- The city's touristic website
- TripAdvisor
- Google Maps

San Donà del Piave has about **40,000 inhabitants** and it's widely recognize as one of the most important cities of Basso Piave: the eastern part of the Metropolitan Area of Venice that takes its name from the local river. Frequently, it's used as point of connection between Venice and other local destinations like Jesolo and Eraclea, famous for their seaside resorts, thanks to the railway and highway network.

San Donà is an ideal place for a day-trip with wide streets and spacious squares. Also, there are several naturalistic points of interest like the river park, many parks, and sport/leisure-dedicated buildings.

Note: Most of the descriptions in the tables of all case studies are directly took from different sources, which are generally difficult to verify – especially in scientific literature. This is due the fact that many of these attractions are not actively promoted by the public administration responsible.

PoI	Description
Fluvial Park	A destination throughout the year for citizens, who reach it with extreme ease being a few hundred meters from the center, as well as for tourists and river fishing enthusiasts, over the years it has become a naturalistic island of great beauty. The environment presents glimpses of considerable charm, beautiful clearings and a wooded heritage of particular value [23].
Land Reclamation Museum	Displays objects, pictures and scale models, organized according to chronological and typological criteria, which present the great reclamation works carried out in the area of San Donà from Antiquity till nowadays. The main feature of this museum is its close relation with the land (ecomuseum), as it is focused on the identity of the place, largely based on local participation and aims at enhancing the welfare and development of local communities[24].
Villa Ancillotto	The history of Villa Ancilotto-Marcato takes us back to the 17th century when the noble Sandi asked for a mapping of his properties. A series of buildings appear on that map, including the Hosteria della Crosetta. This at the time was a well-known hotel-pension that stood on the intersection of four roads, near the bridge built to cross a branch of the Brentella canal, and was equipped with both an oven for cooking bread independently and a shop for the sale of meat [25].
Park of Sculpture in Architecture	The Park of Sculpture in Architecture is in the city of San Donà di Piave, internationally known thanks to scholars and lovers of art and architecture. Since 1991, it has housed sculptures by internationally renowned contemporary artists and architects, exhibited in a real and unique open-air museum which can be visited at any time of day or night [26].
Duomo of Blessed Virgin Mary	The first church of San Donà di Piave of which memory is preserved dates back to the last years of the fifteenth century, but wars and floods have erased its traces. Between 1838 and 1841 a new church was built that too almost completely got destroyed during the First World War. The reconstruction that enshrines its current image takes place on a project by Professor Giuseppe Torres of Venice between 1919 and 1923.
Vittoria Bridge	The bridge has undergone two reconstructions due to the conflicts of the First and Second World War. The Ponte della Vittoria becomes the protagonist every 7 August: it is in fact here that the mayors of Musile and San Donà meet to renew the "friendship pact", on the day of San Donato. According to tradition, the whole territory that now corresponds to the two municipalities was a single village, which bore the name of San Donato because it developed near the chapel dedicated to the saint (now in Musile). Following a flood, the Piave changed its course and divided the country into two.

Table 4.2: San Donà del Piave's main Points of Interest

Market Development – Looking at the potential market for the city, we can found several combinations that may works. For example, thanks to alternative activities like the Vittoria Bridge walk and the Fluvial Park, San Donà seems perfect for a one-day trip for families.

Figure 4.11: San Donà del Piave's Park of Sculpture in Architecture

Market Development – Also, we can target, developing a new market, landscape and history enthusiast using the Landmark Reclamation Museum, Villa Ancillotto and the Sculpture in Architecture Museum.

Candidate #2: Mirano

Mirano is an active **commercial center** and home to many artisanal and industrial activities.

The weekly market on Mondays is flourishing and attracts many people from the surrounding area. Even more visitors are attracted to the center of Mirano by the ancient Fiera di San Matteo: the first edition was authorized by the Veneto Senate with a decree of 6 September 1477 and since then, except for wartime suspensions, it has always been held regularly.

In fact, although this country has a **heterogeneous nature**, rich in ferments and in constant evolution, it does not forget its most ancient traditions that also reverberate in the initiatives of associations and committees, between Piazza Martiri and the hamlets, which are occasions for social life and a sign of well-being and the pleasure of being together.

We can directly use the city's website accurate descriptions [27] for the main points:

Market Penetration – It's clear that one of the strongest point of Mirano are the mansions, so it's possible to build itineraries to pass all of them and attract architecture and art enthusiast.

PoI	Description
Villa Belvedere Park	Built by the Bollani family in the sixteenth century, it is one of the oldest manor houses, located in this place because it was initially built as a farm that managed some whirlpools located along the various waterways in the area, such as the Muson.
Duomo di Mirano	The Cathedral is dedicated to San Michele Arcangelo. The current aspect is the result of a late seventeenth-century renovation, mentioned in a plaque placed inside. Externally it has an essential style, while inside it is richly decorated with elegant pilasters and stuccos in imitation of the festoons that join the capitals together. The recent restoration, which took place in 2022, gives the opportunity to admire each of its decorations with renewed splendor.
Il Castelletto	Vast complex with four rooms (two, intentionally, left open to the sky) next to which stands the five-storey octagonal tower on the remains of the base, the fake ruin, where a pointed arched window opens. The underground structures are characterized by barrel or cross vaults that divide the architectural complex into several parts. The artifice of this underground complex was originally made even more wonderful by stalactites and stalagmites, which today are no longer preserved but which remain in memories and which are believed to be true.
Villa Giustinian – Morosini Park	With a neo-Palladian imprint, this villa, from the point of view architectural, it is perhaps the most beautiful of the municipal villas of Mirano and belongs to a complex that originally provided for two Barchesse to the north, of which today it retains only one, used as the seat of cultural events. The complex is surrounded by a large English park, with a lawn area surrounded by plants on three sides and shrubs on the side of Viale Mariutto, for an area of approximately 3,5 hectares.
Villa Tiepolo	It is located about one kilometer from the center of Mirano, and is a typical example of a Venetian villa. Very well known are the fresco decorations of the rooms on the ground floor, the stair front and the hall on the main floor which, following the purchase by a French collector, were torn up in 1906 and destined to expatriate to France. Thanks to the intervention of the Municipality of Venice, they were recovered at the border and are now located in Ca 'Rezzonico.
Church of the Nativity of Mary	Of fifteenth-century origin, it was completely transformed in the eighteenth century. The Church has an orientation from east to west, with a plan irregular. The construction of two large chapels created a sort of transept, thus transforming the basilica plan into a Latin cross. The range of colors, the various shades of gold, the iconographic choices and the figurative composition recall the paintings of the Tiepolo.

Table 4.3: Mirano's main Points of Interest

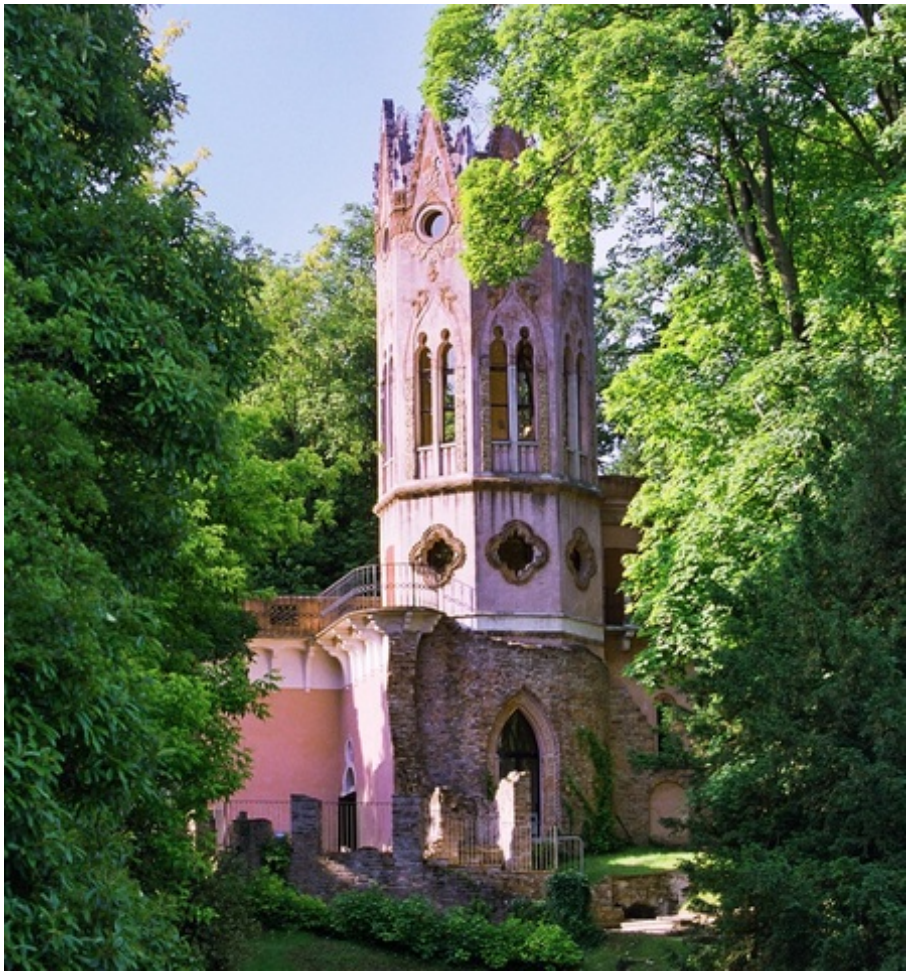

Figure 4.12: Mirano's Castelletto

Service Development – Also, many of the mansions have parks and large spaces, so they are very good to host events for families and young people. Based on the type of event, it's possible to apply a sort of "auto-selection" of the target.

Candidate #3: Conegliano

Conegliano is a small city, just a little more than 36 km^2, of about 35,000 inhabitants. The territory expands between the **UNESCO World Heritage Sites** of "Le Colline del Prosecco" and the Venetian plains. The historic centre is on an hillside and in the last years the city expanded towards the plains. It's a point of connection between the province of Treviso and Belluno, thanks to the highway.

Conegliano is a lovely little town. It's **not unknown to tourists** but it has the 'misfortune' to be close to many other attractive and more significant towns, meaning that while tourists teem around Venice, and some explore Treviso, Conegliano maintains a mellow and peaceful atmosphere.

PoI	Description
Prosecco Hills	The Hills of the Prosecco of Conegliano and Valdobbiadene Unesco Heritage: on 7 July 2019 the UNESCO Assembly recognizes the World Heritage "The Hills of the Prosecco of Conegliano and Valdobbiadene" and inscribes them in the list of cultural landscapes to be protected. This area is characterized by a particular geomorphological conformation, called hogback, consisting of a series of spiky and steep reliefs extended in an east-west direction and interspersed with small valleys parallel to each other. In this difficult environment, man has been able to adapt over the centuries, shaping steep slopes and perfecting his agricultural technique.
Conegliano's Castle	The structure, which houses the City Museum, is the result of a series of renovations and reconstructions. Of the original Scaligera foundation, the deep splayed loopholes remain at the bottom. The castle was partly rebuilt after the collapse of 1491, and was raised to its present height in 1847-55 with the antistorico crowning of Ghibelline battlements.
Duomo and the Hall of the Beaten	In the large district 'Contradal Grande' there is the city Cathedral, whose construction was commenced by the Battuti family in 1345. It constitutes the religious and artistic heart of the city. The interior of the Sala dei Battuti, with its rectangular layout, has a beautiful wooden ceiling and a cycle of frescoes made for the most part by Francesco da Milano.
Brolo	The "brolo" was the botanical garden of the Convent of San Francesco. The Brolo of the Convent of San Francesco is located on the hill behind the ancient convent, built by the Franciscan friars at the beginning of the fifteenth century. It is bordered by Calle Madonna della Neve and a stretch of the ruins from the fourteenth-century walls of Conegliano.
Jewish Cemetery	The Jewish cemetery, built in 1545 on the hill known as the "Cabalan", is one of the most evocative and panoramic places in the city. It stands as a precious testimony of the presence of the integrated Jewish community and it still retains much of its dignity and assorted beauty. There are about 130 tombstones, most of them facing east, in the direction of Jerusalem.
Church of the Madonna della Neve	In the middle of the Calle della Madonna della Neve is the homonymous oratory, for which the first written account of historical information dates back to 1544 and in which it appears as a branch of the parish church of San Leonardo in Castello. This church is located close to the Carrarese walls, at a gate-tower, which originally allowed access from the village to the last stretch of the ancient road, which led to the fortress, known as "Porta della Castagnera".

Table 4.4: Conegliano's main Points of Interest

Visitors can potter around the small centre, climb to the castle, and sit enjoying local Prosecco and food specialities, like Treviso radicchio, in a relaxed and **uncrowded** environment.

Diversification – Conegliano is undoubtedly a place rich in historical and religious aura. With its numerous churches and sacred monuments, it's the ideal location for spiritual people. A short itinerary passing between the Brolo, the Church of the Madonna della Neve and the Jewish Cementery would be more that an experience for a one-day trip outside the overcrowded Venice.

Figure 4.13: Conegliano's Prosecco Hills

Market Development – But, the city has to offer much more: with its hills and local cuisine it can be a fantastic experience for wine and food enthusiast. It's the perfect place to lay back and relax a bit, looking at the typical Venetian/Italian landscape that surround the city.

4.2.4 Proposal Development

We will now proceed to the actual proposal development. In this step, we will create a **brief tour** for each candidate and we will search for **folklore elements** that may enrich the tourists' visit.

San Donà del Piave - Itineraries and Stories

We start the tour by taking the train at 9:05 AM from Venice Santa Lucia and going right to the centre of the city at 9:55 AM. Immediately outside the station, we can observe the main street of San Donà, via Noventa, that will take us to the first stop of the tour: Villa Ancillotto. We can easily take one hour to appreciate the mansion and the surrounding park, then we can walk for about 20 minutes and visit the second park of the city: the Park of Sculpture in Architecture. Again, the advice is to just immerse themselves in the out-worldly experience and let the time pass.

When ready, we can go to the **central part** of the city for a quick launch with local products (the specialty is river fish) and start the second part of the tour by visiting the Duomo for a quick coffee in the main square of the town.

The next stop will be something very peculiar: the Land Reclamation Museum. Finally, we complete the tour by observing first-hand the effect of the nature on the landscape by visiting the beautiful Fluvial Park and the adjacent Vittoria Bridge, then we can return to the station to take the train for Venice Santa Lucia of the 7:43 PM just in time for a dinner in "La Serenissima".

When walking alongside the Piave, we suggest the traveler to be careful of the river bed: in these lands **hydromancy** was practiced. For starting it, just 3 stones were required: one round, one squared, and one triangular. The stones were thrown in the water in the same sequence and, based on the water circles formed around the stone, the belief say that it was possible to predict the future.

These rituals were often practiced by the "zobie" – the Venetian term for "witches" – that has the same etymological root of Thursday. Based on the legends, these witches met in secret to perform the "sabba" where magical practices and enchantments were cast around a bonfire[28].

Mirano - Itineraries

We start the tour by taking the bus at 9:04 AM from Venice Santa Lucia and slightly outside the main part of the city at 9:59 AM. After a 10 minutes walk we meet the first stop of the tour: the Church of the Nativity of Mary. The architecture per se is not huge, it's possible to fully view the religious landmark in about 30 minutes. After another 10 minutes walk, we meet the first mansion of the tour: Villa Tiepolo, to visit it the suggested time is around 30-60 minutes.

Compared to San Donà del Piave, this tour can be completed in about **half a day**, so it may be used in different situation compared to the first one. Plus, it's easily extendibile by visiting the many parks present in a 30 minute walk in the Orgnano/Macello area.

We continue the **second part** of the tour by visiting the Villa Belvedere Park that includes also the Castelletto of Mirano and the mansion Villa Giustinian Morosini. A couple of hours for all these point of interest and to fully appreciate the environment is a must here.

Finally, we have time for a quick view of the main square – the Duomo of Mirano – before taking the bus from the opposite part of the city and be again in Venice Santa Lucia before 3 PM.

Conegliano - Itineraries

Like with San Donà del Piave, we take the train at 9:01 AM to reach the centre of the city just before 10 AM. After a 10 minute walk we arrive at the first stop: the Duomo with the Hall of the Beaten. After an one-hour visit, we can walk 200 meters and find the second stop of the tour the Brolo inside the Convento of San Francesco. This is a smart step because from there we can directly reacy the Church of Madonna della Neve and Conegliano's Castle.

To finish this **first part** will take about a couple of hours, so the advice is to exploit the opportunity to visit Villa Gera and its outstanding view of the city before going for a quick bite in the city centre.

The **second part** of the tour will be a gradual "escape" from the city: visiting an extremely powerful memorial – the Jewish Cemetery – may give rise to mixed emotions, but they will be tuned in by the visit to the Prosecco Hills in a environment were you can find unique scents, views, and tastes.

To reach the Prosecco Hills it's strongly recommended to book a public transport *before* coming to Conegliano. The distance is about 30 chilometers, covered in around 40-45 minutes. If the timing is correct, and you start going by 3:30 PM, you should have the chance to taste a superb appetizer together with the UNESCO Prosecco. Once the tasting experience ends, you can visit many local structures that are used to make Prosecco but, overall, you should start returning by 7PM so you can reach Venice Santa Lucia by 8:30 PM.

In these territories you will have the opportunity to appreciate the born of the Prosecco DOCG with the first **Italian Enological School** founded in Conegliano by **Antonio Carpenè**; a Mazzinian, participated in some important battles of the Risorgimento. It was a positivist and progressive scientist, he had contacts with Robert Koch and Louis Pasteur. The latter wrote to him inviting him to deepen the important research on the effects of sulphurous acid on the ferments of wine and beer.

4.2.5 Sustainability

Finally, we can evaluate the three alternatives based on the sustainability criteria:

- **Environmental Protection:** All of the tours and stories used public transports and empowered naturalistic landscapes. But the same places will still suffer from overtourism, like every other city, so it's imperative to assess *before* the promotional activities the tourist pressure that each alternative locality can bear.

  A couple of cities, like Mirano and San Donà del Piave, are extremely green and pro-bike environments, so they may also provide an additional push to the "green side" of the tourists.

- **Societal Responsibility:** Most of these territories would benefit from an increase tourism flux, and Venice would be somewhat relieved of a portion of the pressure.

  It's important to remember that many steps of these tours may not be easily accessible to everyone given cost barriers (Prosecco Hills) or physical barriers (like climbs, steps, or similar). This was in part accounted for: all the physical accessibility barriers provided a negative score to the AI.

- **Economic Stability:** The economic result of the promotional activities should be almost surely positive, if we consider the double-faced social impact generated by moving tourist to the rural part of the region.

  It's also true that many of the tourist went to Venice specifically to appreciate its attractions, but more than 20% of the people visited another major city in the end. So this means that some people **are** willing to travel additional time to see more interesting points. It makes sense, from a marketing point of view, to exploit the cultural sensibility of the majority of the tourist: most of the are highly educated (around 80%) so it may seems logical to them.

  Finally, all the tours don't require additional capital expenses from a logistical point of view because they use in-place public transports, although they may benefit from additional means.

In theory, the output of each of these phases should be discussed with one or more stakeholders to assess it from a qualitative point of view. After this step, there should be a series of practical brainstorming and decision-making sessions were new promotional ideas are created and validated.

The interactive maps for the case study are available at the links:
`https://pinea.app.carto.com/map/810354db-9994-4bbf-905a-70da8d32a010` and
`https://pinea.app.carto.com/map/fee78425-981f-4226-8011-6523803b70f0`

4.3 Netherlands - Amsterdam

Amsterdam is the capital – and the most populous – city of Netherlands. Its population is reaching a million people and, considering the urban area, the residents are more than 1.5 millions.

Amsterdam was born as a small fishing village around the 12th century. With the advent of the Dutch Golden Age – during the 17th century – the city will become a major world port, using the Netherlands as economic powerhouse. In the last two centuries, the city expanded and included many new neighborhoods [29].

The travel industry in Netherlands has a huge impact: the market size was almost 90 billions euros in 2019 [30]. More than half of the revenue come from the **domestic tourism expenditure**. We can also notice that the sector was growing every year at a steady pace until 2019: the COVID's year.

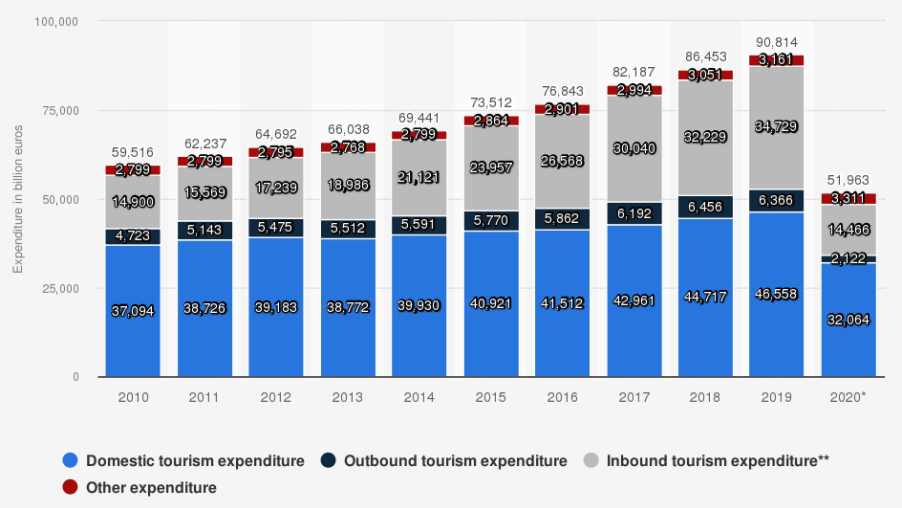

Figure 4.14: Tourism Expenditure in the Netherlands from 2010 to 2020, by Category

Compared to other metropolis, Amsterdam is a relatively small city and most of the travellers just stay in the inner part of it. But, for example, the locals have a much different visiting pattern: they're are slightly more spread across the map and have an higher probability to go and visit the outskirt of the city [31].

Considering that about half of the visitors are day-trippers it makes sense to visit just the main part of the territory. The situation is critical even for the **over-nighters**: 10 millions, around 12 for each resident. The total is around 21 millions nights [32].

This depict a clear picture: there is a portion of tourist that visit Amsterdam during a single day while there is another that stays for 3 days on average. Their trip purpose are associated with substantial differences in total daily expenditure but the activities undertaken are not limited to their initial trip purpose [33].

Rouwendal [33] gave us a small drill-down on the **purposes**: 53% of the tourists indicated that they had come to Amsterdam to see the old city and the canals; 3% to visit a special event or exhibition; 3% for the use of cannabis; 10% of the tourists had mixed purposes; and 31% had another travel purpose.

It's also interesting to note that, from 2019, landlords of entire homes in the capital may rent out their house to tourists for a maximum of thirty days a year. This was proposed by the Mayor and Municipal Executive of Amsterdam at the beginning of this year [32].

This was not the only policy put in place to fight overtourism: there is also the "no, unless" hotel policy. Substantially, new hotels cannot open in the city unless they add something special.

It's clear that *overtourism is tackled very seriously* in the "Venice of the North" because historically, Dutch people are very sensible about the impact that foreigners can have both materially and immaterially [34].

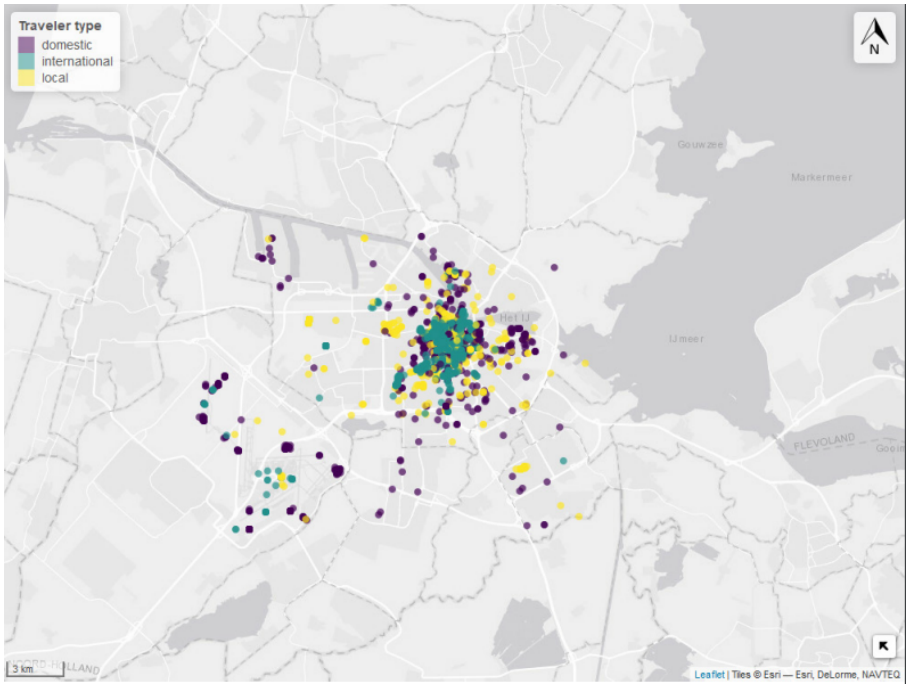

Figure 4.15: Geo-referenced Flickr posts from Locals and Tourists

4.3.1 Analysis

We download the datasets from GeoFabrik and load it on the notebook pipeline. Unfortunately, there wasn't a single region available sufficiently large to cover for the neighboring territories of the Metropolitan City of Amsterdam. So we had to develop a new solution: manually download all the areas of interest and use a software – called "Bulk Rename Utility" – to extract and rename the file automatically.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	0.3282	2.6542	1.4794	0.3350	0.3174	0.7001	5.5510
br	Bayesian Ridge	0.3570	2.6892	1.4918	0.3214	0.3169	0.7825	0.1380
ridge	Ridge Regression	0.3585	2.6942	1.4943	0.3183	0.3180	0.7865	0.1040
lr	Linear Regression	0.3587	2.6953	1.4950	0.3173	0.3180	0.7865	1.5670
en	Elastic Net	0.3488	2.7457	1.5073	0.3103	0.3211	0.7141	0.0770
omp	Orthogonal Matching Pursuit	0.3334	2.7521	1.5084	0.3096	0.3170	0.7616	0.0660
lightgbm	Light Gradient Boosting Machine	0.3266	2.7333	1.5080	0.3027	0.3138	0.7563	0.4990
lasso	Lasso Regression	0.3495	2.8056	1.5275	0.2904	0.3238	0.6956	0.4080
gbr	Gradient Boosting Regressor	0.3192	2.7341	1.5210	0.2644	0.2993	0.6802	1.7670
huber	Huber Regressor	0.2774	2.9613	1.5715	0.2493	0.2975	0.8039	1.8570
lar	Least Angle Regression	0.4300	2.9714	1.5887	0.2159	0.3542	0.8923	0.0740
rf	Random Forest Regressor	0.3313	2.9463	1.5814	0.2157	0.3161	0.7307	4.9650
knn	K Neighbors Regressor	0.3406	3.1370	1.6419	0.1605	0.3515	0.7503	2.5940
par	Passive Aggressive Regressor	0.5190	3.4656	1.7507	0.0155	0.4161	1.0696	0.0990
llar	Lasso Least Angle Regression	0.5149	3.6022	1.7792	-0.0007	0.4012	0.7948	0.0620
dummy	Dummy Regressor	0.5149	3.6022	1.7792	-0.0007	0.4012	0.7948	0.0630
dt	Decision Tree Regressor	0.3999	5.5709	2.2651	-0.9675	0.4024	1.0603	0.2000
ada	AdaBoost Regressor	1.6999	6.7517	2.5470	-1.5119	0.9101	1.5792	2.7570

Figure 4.16: Models' Performance for Amsterdam's Case Study

Using this **modified workflow**, it's possible to use all the files directly as input. Their prefix isn't so important: it just needs to be different from the other areas to avoid the default overwriting. After that, the pipeline will automatically detect the files based on their extension and merge all of them into a single database.

The pipeline chose the Extra Tree model as the best one based on its performance on MSE and R^2. Extra Tree models have a set of unique features compared to the standard Random Forest [35]:

- Like RF, it creates many decision trees, but the sampling for each tree is random and without replacement. This means that each tree has a unique dataset
- The feature from each sample are also selected randomly
- Finally, the splitting point for the tree isn't chosen based on the locally optimal value using Gini but it's selected randomly

These characteristics grants the model **more randomicity** and less correlation between the trees compared to

the Random Forest algorithm. The predictions are computed as [36]:

$$\hat{y}(x) = \sum_{i_1=0}^{N} \cdots \sum_{i_n=0}^{N} I_{(i_1,\ldots,i_n)}(x) \sum_{X \subset \{x_1,\ldots,x_n\}} \lambda^{X}_{(i_1,\ldots,i_n)} \prod_{x_j \in X} x_j \tag{4.2}$$

where:

- $\hat{y}(x)$ is the predicted value for observation x
- N is the sample size
- i is the observation index, going from 1 to N
- $\lambda_{(i_1,\ldots,i_n)}$ is a real-valued parameter that depends on the sample inputs x^i and outputs y^i
- $I_{(i_1,\ldots,i_n)}(x)$ is the characteristic function of an hyper-interval

Resulting in a map with slightly more concentrated predictions compared to the actual ones:

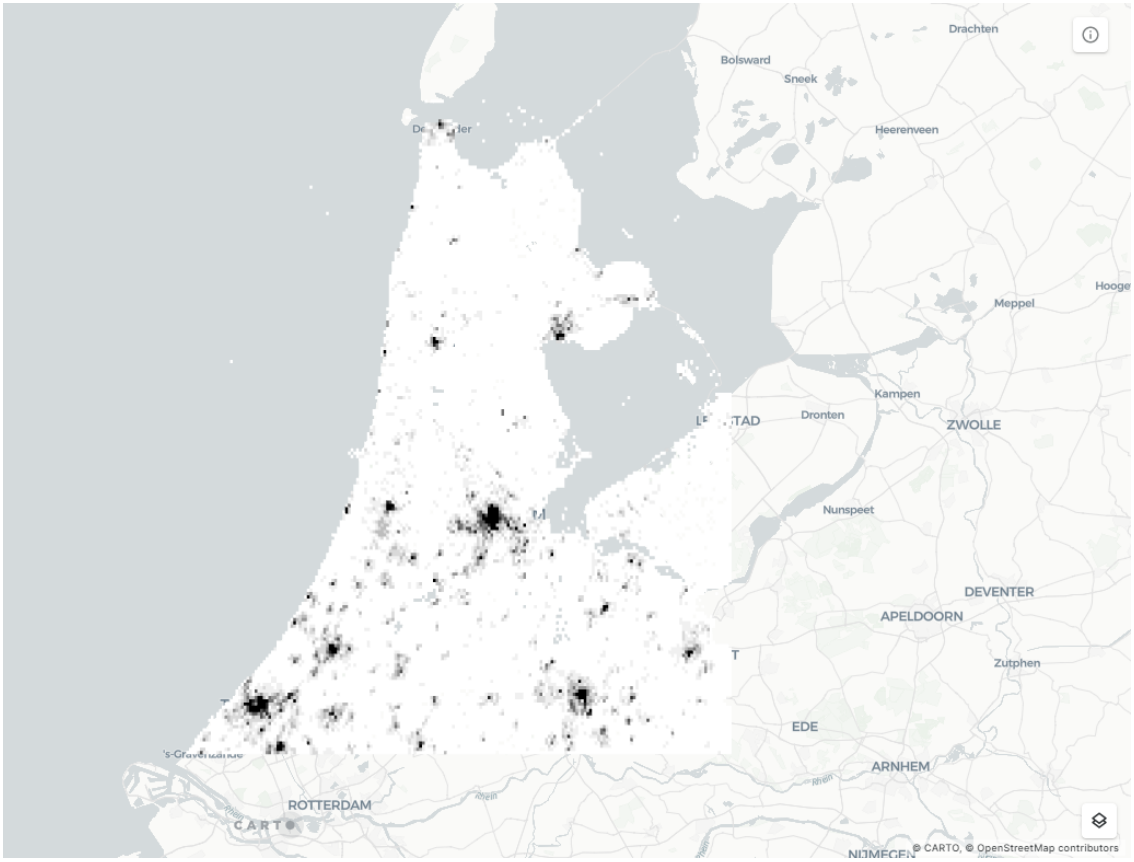

Figure 4.17: Predicted Tourism of Amsterdam and its Surroundings

Like the previous case study, the prediction were computed based on the *7 features plus the data from the neighboring cells*. We can spot Amsterdam – the black group at the centre of the map – together with other major localities like Utrecht and the Hague.

Compared to Venice, we see **higher concentration on a small set** of cities instead of a well-balanced network of nation-renowed localities. This is true especially observing the northern part of the Dutch nation: there are almost no point were the AI predict tourism potential.

Looking at the *actual* tourism values, the map explains a bit of what is happening:

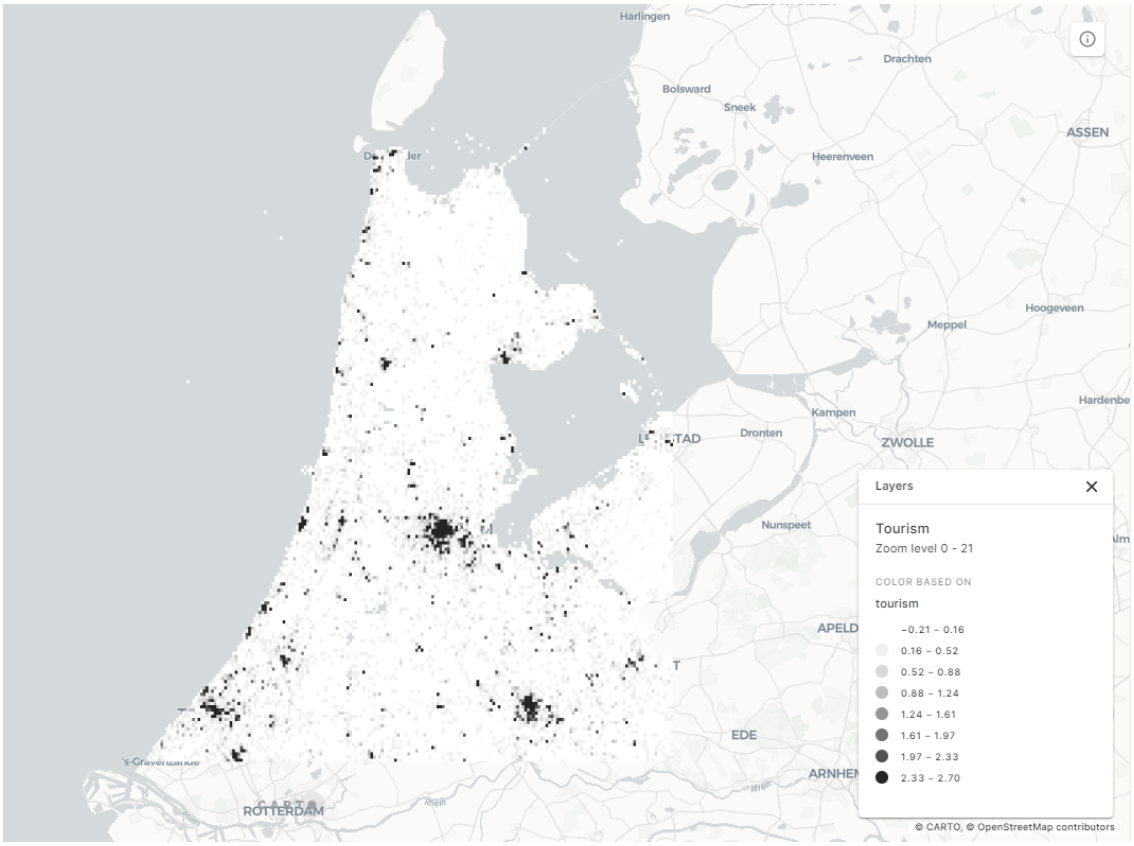

Figure 4.18: Actual Tourism of Amsterdam and its Surroundings

There are several small spots – especially in the southern regions – that the model didn't pick. This is probably given by the outlier removal and by assumptions that are too general. It's also interesting to note that the model empowered more the southern part of Amsterdam compared to the actual tourism.

Like in the Venice case, there seems to be a strong *link between the accessibility feature and the tourism one*. In this situation, though, art is much more important: it possess 5 of the top 10 most important features. This fact resembles well the identikit of the average Amsterdam tourist: they come to visit art-related attractions like the old canals.

For the first time, we also notice the security feature show up – three times – in the plot. Being an "hygienic" feature, it should be treated similarly to the accessibility one: it's probably an an effect of tourism, and not a cause. Surprisingly, only one feature related to entertainment a no feature related to nature at all.

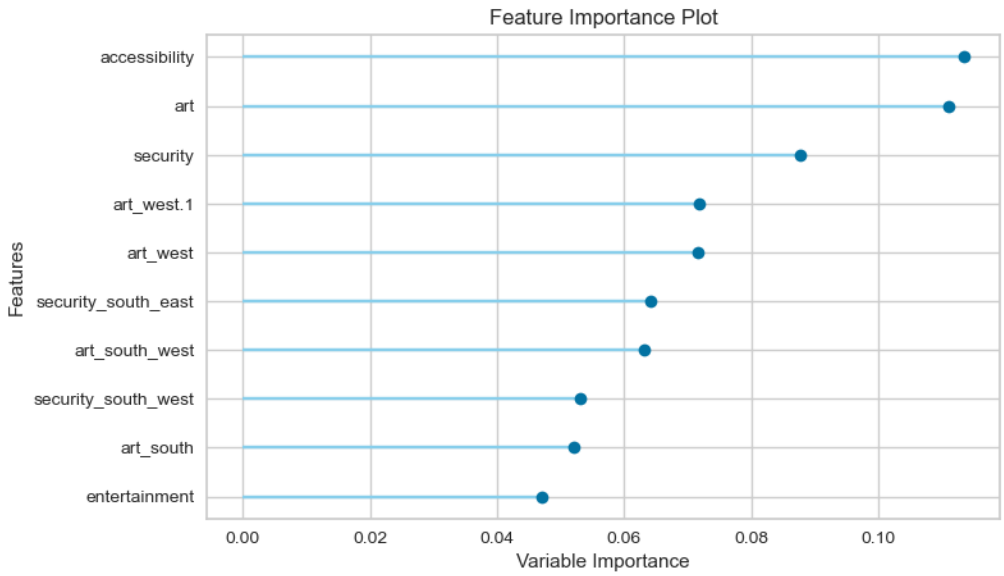

Figure 4.19: Amsterdam's ExtraTree Feature Importance Plot

We can understand better the rating by analyzing the single layers, starting from the transport:

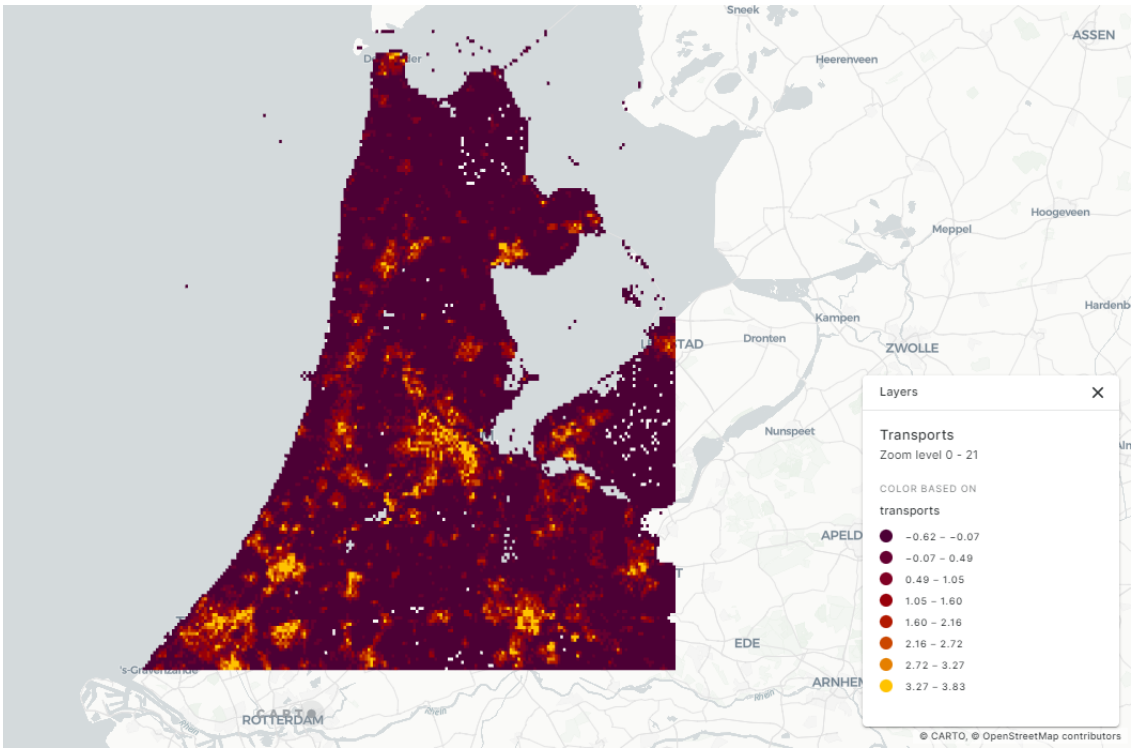

Figure 4.20: Transport Map of Amsterdam and its Surroundings

The yellow regions – the ones that indicate a better transport network – are present mainly in the most important cities, while **large chunks** of the map, marked with purple, are left without almost nothing. Even to the naked eye it's possible to understand the **strong correlation** that tourism and transport carry: there are just a couple of areas in the north that reach the upper classes with regard to the transport and most of them have no significant points of interest.

The only type of relevant points of interest in the north are the natural/environmental ones:

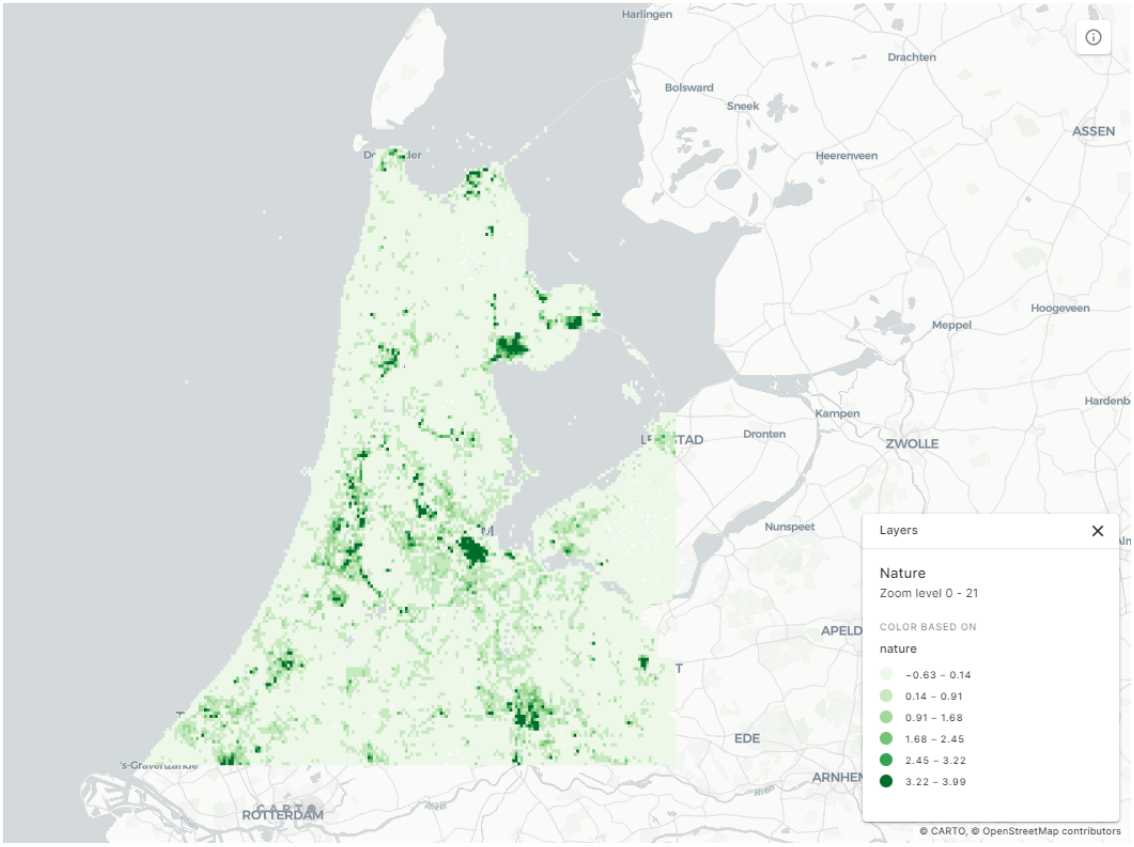

Figure 4.21: Nature Map of Amsterdam and its Surroundings

We can spot a relatively strong correlation in the northern area between transport and natural points of interest. The main attraction in that area are the Geestmerambacht, a major lake surrounded by a forest, and Hoorn: an area surrounded by cultivated fields. The second one is clearly of industrial interest and not of a tourist one.

Around the **centre of the map** we can notice that to the west of Amsterdam there are several "green" points and the same is true for Utrecht and the Hague. In the southern parts there are many smaller groups that may be interesting to explore after the Filtering step.

Finally, knowing the strong social welfare policies, we can appreciate the much higher rate of accessibility compared to Venice:

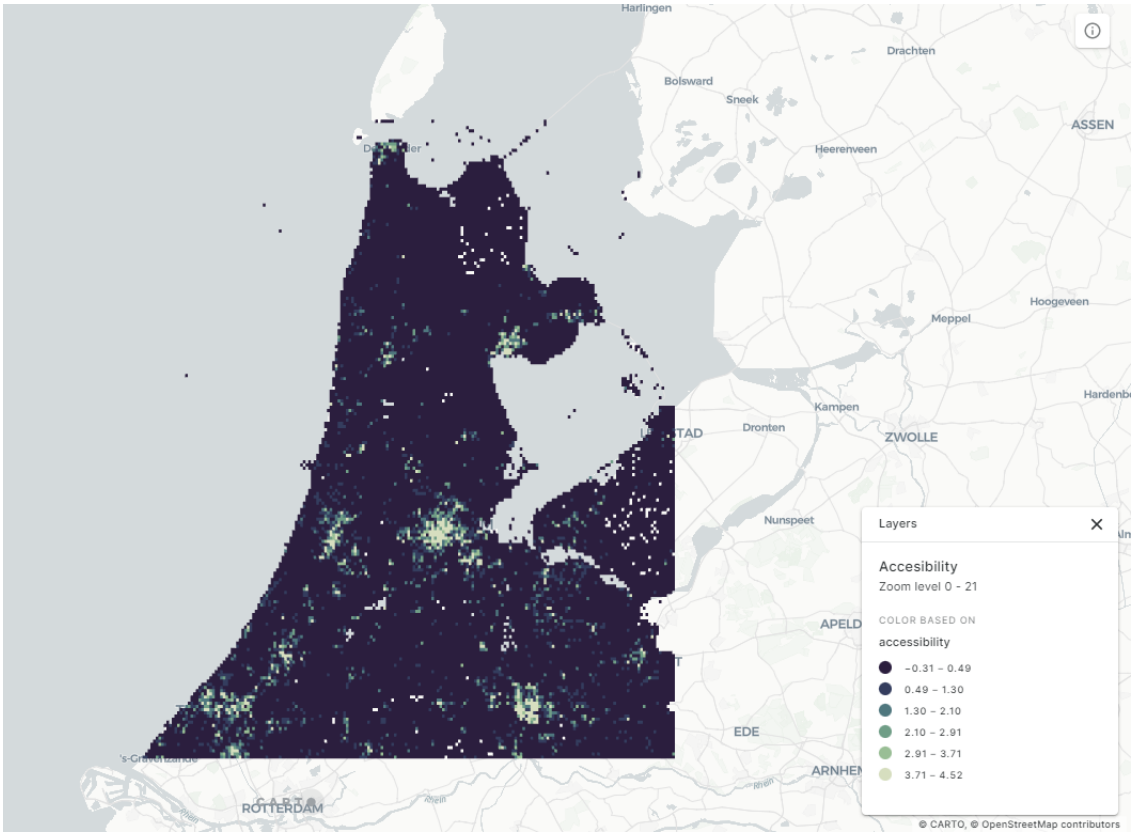

Figure 4.22: Accessibility Map of Amsterdam and its Surroundings

It's also important to remember that Netherlands is the **flattest country** in Europe, so it must be certainly simpler to be accessible compared to a geologically complex region like Veneto. Most of the touristic areas obtain generally good ratings, considering that through the manipulation we essentially count the anthropological features that create a more accessible environment but in many cases – having flat surfaces – they may not be needed, so with this regard the accessibility of this area may be severely underestimated.

Comparing the sizes, it's also clearer that the Dutch government worked more on creating "accessible environments" and not simple "excellencies" that are the extra-ordinary and not the ordinary. But we still need to keep in mind that, even tough the situation is overall better, the majority of the map is still dark blue.

Finally, going into the **cluster analysis**, we can see that the AI had some difficulties in finding 20 of them because there are a couple of cluster really near each other (both on Amsterdam) but there are still interesting results to analyze.

For example, if we remove, as always, the clusters near the most important cities (Amsterdam, Utrecht and the Hague) we can still pick out different potential candidates: Alkmaar and Hoorn are still potential candidates, but we know that they score higher because of industrial interest for the environmental resources that developed the transport network.

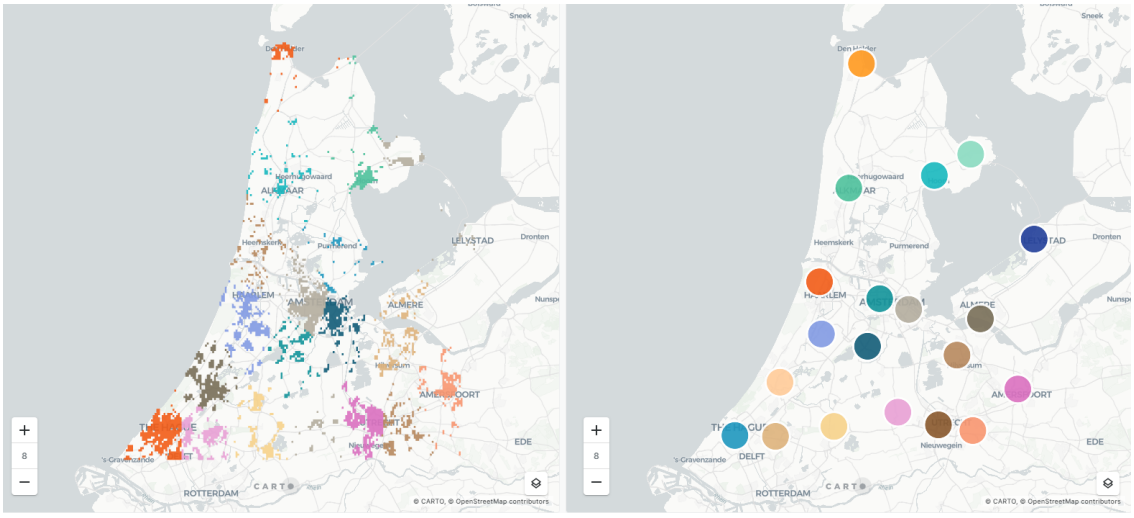

Figure 4.23: Cluster Analysis for Amsterdam's Case Study

Other possibile candidates, in the south-west extremities, are 's-Gravenzande, Gouda and Delft while in the north-west we have Worden, Amersfoort and Hilversum. Finally, in the northern part we have Den Helder and Heemskerk.

4.3.2 Filtering

Like in the previous case study, we will use Google Maps estimates to assess the trip time from Amsterdam main station:

Destination	Travel Time
's-Gravenzande	123 min
Gouda	50 min
Delft	60 min
Woerden	37 min
Amersfoort	34 min
Hilversum	20 min
Den Helder	77 min
Heemskerk	40 min

Table 4.5: Travel Time from Amsterdam, using Public Transport, for each Destination

Almost every destination match the filter of "maximum 60 minutes of travel time using public transport". Given their relative importance and their distance to each other, we will select one city from the south-west region (**Gouda**), one from the south-east (**Amersfoot**) and one from the north (**Heemskerk**).

Curiously, transport in the south-west area is slower than expected – especially for the rural ones – and it has been shown very difficult to do more than 60km/h.

4.3.3 Evaluation

Like in the previous case study, we are going to evaluate the main candidates: Gouda, Amersfoot, and Heemskerk. We expect to understand more the **peculiar culture** of each of these cities and to exploit it for tourism purposes.

Candidate #1: Gouda

Gouda is a **medium-sized town** of about 75,000 inhabitants. It's located in South Holland and it's just 18 km^2, about half the size of Conegliano. It's notoriously famous for the **Gouda cheese**, produced in the plains around the city and sold in the historic market. The economy is mainly based on leisure and retail, while the offices are generally located on the outskirts.

The city has seen a **tremendous growth** after the World War II: it nearly tripled in size during the aftermath.

The Gouda Cheese & Craft Market attract 60,000 tourists each year [37]. The ratio between the residents and the tourist is still under 1, so this means that there is still a lot of space before going into overtourism.

Gouda is **well connected** to the major cities – The Hague, Rotterdam, Utrecht, Amsterdams, and Alphen aan den Rijn/Leiden – thanks to its tow railway stations: Gouda and Gouda Goverwelle. It also lies alongside two highways: the A12 and the A20.

PoI	Description
Old City Hall and Market Square	The city hall (Stadhuis) in Gouda is the oldest gothic City Hall in the Netherlands and was built in 1450. It is located at the main square of the city, the Markt, and is one of the best known monuments of Gouda. The Market host the Gouda Cheese & Craft Market every Thursday from 10:00 to 13:00.
The Waag	De Goudse Waag is a historic building dating from 1668 which houses a museum that can tell you all about the history of De Goudse Waag and the products that were weighed and traded here. We also offer various options for welcoming small and large groups in conjunction with a visit to the museum. A museum shop with delicious real Dutch cheeses and a variety of delightful traditional Dutch souvenirs can be found on the ground floor of De Goudse Waag [38].
The Sinking Street	Along the Turfmarkt, there is a sinking street: annually the soil in Gouda drops by three millimetres and the canal is reaching the side of the canal.

PoI	Description
Saint John's Church	Houses a traditional carillon of 50 bells. The carillon was installed in 1676, with 37 bells cast by Hemony (Netherlands) – 16 bells remain from this original work. In 1966, the Royal Eijsbouts Bell Foundry enlarged the carillon with 33 bells, bringing the total to 50 bells. The Eijsbouts bells were given names of people relating to the history of the carillon; the smallest bell, however, is the name of a "Ros Beyaard" – the name of a famous horse in Dutch literature, and a synonym of 'carillon' [39].
Verzetsmuseum	Permanent exhibit of the museum recreates the atmosphere of the streets of Amsterdam during the German occupation of the WWII. Big photographs, old posters, objects, films and sounds from that horrible time, help to recreate the scene. This is an exhibition about the everyday life during that time, but also about exceptional historical events, resistance of the population against the Nazis and heroism [40].
Waaier Locks	Lock complex from 1856 with a lock and a discharge lock in the Hollandsche IJssel, between the river and the canalised part. The lock is special because it was the first time a fan lock was used in a waterway. The manually operated fan doors in the Waaiersluis near Gouda are still in operation and are also used a few times.
Goudse Schouwburg	The Goudse Schouwburg is a theater in the Dutch city of Gouda . The current theater on the Boelekade from 1992 was designed by AGS Architects and Planners, J. Paré from Heerlen. In 1994 the theater received the Architecture Award from the city of Gouda, a public award. In 2016 and 2017 it won the election of "The most hospitable theater in the Netherlands".
Van Bergen IJzendoorn-park	It was donated by the Mayor of Gouda – Albertus Adrianus van Bergen IJzendoorn – to the city. In 1897 Hendrik Copijn drew up a plan for the park in the English landscape style: characteristic of this style are the winding paths, ponds and bat bridges. At the entrance of the park on the station side, two pillars were placed with the lions on them, originating from the Kleiwegspoort, which was demolished in 1843.
Natuurspeeltuin	A park-like nature playground in Gouda. This playground with height differences, nice hiding places, space to build huts, a climbing wall, cable car, slides and tree houses. A fantastic place for children from 2 to about 12 years old and their parents. Enjoy playing, moving and enjoying nature.

Table 4.6: Gouda's main Points of Interest

Market Development – Gouda benefits from a strong tourist flux coming to visit the cheese attractions but, as we've seen, there are many different point of interest throughout the city starting from the sinking street and going all the way through a World War II museum.

Figure 4.24: Gouda's Sinking Street

Diversification – Looking at the naturalistic point of view, Gouda seems perfect for a fun, family-sized, day-trip. Starting from the Van Bergen park and then going to one of the biggest "natural" playground: the Natuur-speeltuin. So, maybe, it's possible to promote these attractions to the non-cheese enthusiasts.

Candidate #2: Amersfoot

Amersfoort is one of Holland's main transport hubs. It's a **railway junction** with its three stations: Amersfoort Centraal, Schothorst, and Vathorst. The city hosted the pentathlon events of 1928 Summer Olympics. Like Gouda, the town is served by two highways: the A1 and the A28.

The city is **fifth in the country by population**: it counts about 150,000 inhabitants in just 63 km^2. It was founded by hunter gatherers during the Mesolitic period: to this day some traces are still found by the archaeologists. During the Middle Ages, it was an important textile centre, counting many breweries.

The historic centre of the city, exhibiting mainly medieval artifacts, can offer many entertaining visits and views. There is also a forest accessible from both west and south. The town celebrated its 750th anniversary in 2009.

Bonus fact: Amersfoort is the *greenest* city in the Netherlands and also hosts one of the tallest medieval church towers (the Onze-Lieve-Vrouwentoren).

PoI	Description
Onze Lieve Vrouwetoren	With its 98-meter height, the 'Onze Lieve Vrouwetoren' is the second highest tower in the Netherlands. There are regular tours during the summer months where you can discover everything about the history and legends surrounding the tower. You can climb the 364 steps on a guided tour, from July to mid-September. The climb on the former cathedral spire will reward you with an amazing view.
Koppelpoort	Probably the best preserved city gate in the Netherlands. It is the northern city gate of the second wall and a dual water and land entrance to the city, built over the river Eem, which begins under the gate.
Sk8Park	The Sk8Park Vathorst in Amersfoort is the largest skate park in Europe. You can try inline skating, skate boarding and ride a BMX bike.
Park Randenbroek	The park has a wide diversity of trees, plants and animals, including a colony of herons.
Mondriaanhuis	Situated in the house of birth of the famous artist Piet Mondriaan, this museum is completely dedicated to his work and includes a reconstruction of his 1920s Paris studio. The museum has English descriptions of objects and artefacts, English brochures and guided tours in English. Please make a reservation for a tour in English at least four days in advance.
Latin American Art Museum of Amersfoort	The first museum in the Netherlands dedicated to Spanish, Mexican, Caribbean, Central and South American contemporary art. English and Spanish descriptions of objects and artefacts are present. Guided tours in English and Spanish possible. For a guided tour in English or Spanish for groups consisting of more than ten people please make a reservation.
Living History	Actors re-enact professions and people from the rich history of Amersfoort, letting you experience times gone by. You can see re-enactment on Burgerweeshuis (a former orphanage), Mannenzaal (a nursing home), and in the city centre on Sunday.

Table 4.7: Amersfoort's main Points of Interest

Market Penetration – Sport is one of the major tourist enhancer. It's main perk is that, usually, it's *not* seasonal like the majority of other tourism types. So, by developing more this market, Amersfoort can become the reference for skateboarding competitions and events. This would require building additional structures and services but, generally, the benefits outweigh the cons, especially in the long term.

Figure 4.25: Amersfoort's Koopelpoort

Diversification – Amersfoort is extremely well connected and, like most of Holland, is mainly flat. This opens the space to easily transfer tourist from Amsterdam and other neighboring cities, like Utrecht. Many of the PoI are of artistic/historical interest (Onze Lieve Vrouwetoren and Koppelpoort) so it may be interesting to develop an half-day trip for this location. For this reason, this location may be ideal for visitors that want to gain a more comprehensive point of view about the Dutch culture without sacrificing too much time.

Candidate #3: Heemskerk

Heemskerk is a small town in the Kennemerland region in North Holland, Netherlands. Heemskerk *does not rate high as a tourist destination*, but it's definitely worth visiting. Especially the coastal dune area is extremely well-equipped and not very well known compared to other areas along the Dutch coast. Heemskerk was also the birthplace of 16th century painter Maerten van Heemskerck (1498-1574), one of the most important Renaissance painters in the Netherlands.

The traces of the city's name are dated around the 11th century. During the last part of the Middle Ages, it's main function was a **ceremonial** one: Heemskerk was, in fact, the site where the counts of Holland were inaugurated. To celebrate the occurrence, six castles were built as a defensive measure against the West-Frisians.

PoI	Description
Lunettenlinie	This defence line, running from Wijk aan Zee to Heemskerk was built in 1799. The Netherlands at the time were under French rule, and a combined English-Russian force trying to invade the Batavian Republic, was beaten off near Castricum. It was then decided to build a series of small fortifications known as lunettes to keep future invaders away. The defence line, however, was never put to use, but a number of the lunettes are still preserved. Finding them, however, needs a little bit of internet searching.

PoI	Description
Slot Assumburg	The castle originally dates from the 13th century and was rebuilt in 1546. Since 1933 it is in use as a youth hostel. The castle itself is not open to visitors, but the beautiful French classicist gardens are worth a little detour.
Noordhollands Duinreservaat	The coastal dune area west of Heemskerk is managed by the Provincial Waterworks of North Holland (PWN). It is a beautiful area for walking and cycling which stretches all the way from Wijk aan Zee to Schoorl.
Tuinen en Duinen route	A 9.6 km walking route will lead you through the most attractive parts of Heemskerk. Especially worthwhile in Spring, when the flower fields are in bloom.
Château Marquette	The enchanting Marquette hotel in Heemskerk is situated on a sprawling 13th century estate. Surrounded by lush greenery, the hotel's stunning structure dates back to 1225, and includes a castle that houses a restaurant and events facilities.
Dorpskerk	This tower was built in the 13th century from so-called monasteries (large stones). The substructure of the tower belongs to the Romanesque architectural style. The superstructure belongs to the Flemish art Gothic. The spire, built entirely of stone, is a rarity in the Netherlands. In 1585 the spire collapsed due to a lightning strike. In 1628 the tower was rebuilt with a different type of stone, which can still be seen.
Strand Wijk aan Zee	The beach, 4km wide, has received 11 times the international Blue Flag environmental award that guarantees excellent quality bathing water and a well-kept beach. The beach provides several family-friendly activities like surfing, paraskiing and kite flying.

Table 4.8: Heemskerk's main Points of Interest

Product Development – Heemskerk seems very "vertical" on two types of tourism: sport and artistic/historical. But it may benefit from other types of attraction, especially regarding the food & beverage section. Being between the sea and and a very rich culinary tradition from the rest of Netherlands it may be disappoint for a visitor to discover that there is nothing special or traditional from this point of view.

Market Penetration – Giving the abundance of natural attractions like beautiful walks and beaches, Heemskerk may complement very well a summer/spring trip to Amsterdam. The day-trip can be used to break from the big city and explore a variety of sports: both outdoor and aquatic. One of the best selling points is that most of the attraction are suitable for families, so larger groups can be "moved" from Amsterdam to Heemskerk.

Figure 4.26: Heemskerk's Slot Assumburg

4.3.4 Proposal Development

Like with the previous case study, we will create a brief tour and we will find out if there are interesting folkloric elements. We will refrain from developing the entire storytelling process due to space, time, and skill constraints.

Gouda - Itineraries and Stories

Starting from the Amsterdam Central Station, we take the train at 9:19 to arrive by 10:10 directly in the hearth of Gouda. From there, we can immediately admire the mayor's park: the Van Bergen IJzendoornkpark. Having seen the 4-5 point of interest present (different statues and architectures) we can appreciate the "Sinking Street" on the district Turfmarkt. Around 12:00 we should be ready to visit centre with many of the most iconic city's attractions: the Town Hall, the Cheese Market, the Saint John Church and the Goudse Waag complex. Of course, in this occasion we can taste the different **culinary specialties** that Gouda has to offer.

After a not-so-light lunch, by 17:00 we are ready for a **cultural moment** in the World War II resistance museum: the Vierzetsmuseum. With a brief walk, about 20-30 minutes, we can enjoy a couple of extremely peculiar views: the Water Locks and the natural playground Het Eiland. If we are on time we can do so with a spectacular view created by the sunset.

While returning to the station – another 20 minute walk – we can visit the last attraction: the Schouwburgplein, the most important theater of the city. Finally, at 19:09 we take the train and we are back to Amsterdam just 5 minutes after 20.

Folklore is a central part to Gouda: we have several stories talking about the famous cheese's origin, but there is also a small locality just outside town – Oudewater – well known to be the "city of witches" with the

Heksenwaag: the house of the witches' weight.

Centuries ago, it was believed that a woman to be a witch and use a broom had to be very light, practically weightless; so in order to be acquitted of the charge of witchcraft, it was necessary to prove that her weight corresponded to the proportions of his own body. In the past there were many innocent people accused of witchcraft throughout Europe, some of these (at least, those who could afford the trip) went to countries like Oudewater to avoid being burned at the stake [41].

Amersfoort - Itineraries

We take the train at 9:00 from the Central Station and by 9:34 we're arrived at Amersfoort. Starting from the central train station, we take a pass a couple of streets with **terraced houses** to end in a beautiful park called "Speeltuin Juliana van Stolbergpark" that contains a small skatepark, the city's sport. In five minutes from there, we are at the Latin American Art Museum: the first Museum in the Netherlands dedicated to Spanish, Mexican, Caribbean, Central and South American Contemporary Art.

After a couple of hours, we can take another brief walk to one of the biggest park in Amersfoort – Speeltuin Park Randenbroek – where we can appreciate, especially in spring, the wonderful even venue "Huize Randenbroek". Just in time for lunch, we can reach the centre in about 20 minutes by foot.

Here, we will find all the major attractions: the Mondriaan House, the Onze-Lieve-Vrouwetoren, and the historical centre with his **living history spectacles**. We can finish our visit, by 5:00 while returning to the station. In fact, on our way back we will find the best preserved gate in Holland: the Koppelpoort.

While having time for a small break to taste the culinary specialities, we can take the train at 18:31 and be back to Amsterdam at 19:19.

Heemskerk - Itineraries and Stories

As always, we start at Amsterdam Central Station by taking the 9:13 train and arriving by 9:53 at Heemskerk main station.

For this tour, we will need a **bike** and some training: the path is around 22 kilometers long – spread over the entire day – and it doesn't have any particular climb or descent. Holland is a bike-friendly country, so it's safe for children too.

Starting from the bus stop, we pick our bikes and go directly to the Chateau Marquette event venue to appreciate both the building and the surrounding park. After a 10-minute run on the outskirt of the city, we arrive at the coastal *dune* area with all the characteristic flora and fauna. By 11:00, we should be ready for the longest run of the day: about half an hour in a typical natural Dutch pathway. At the end, we will find the well-known beach Strand Wijk aan Zee that provide a wealth of familiy-friendly activities like paraskiing and kite flying.

After lunch, we're ready to come back with another 30 minute ride along an incredible wee-preserved defense line called "Lunettenlinie". Following the line, we will find ourselves inside the central part of the city, just few moments away from the Dorpskerk: one of the most iconic building.

We finish the tour by making a slight deviation toward the main attraction: the castle Slot Assumburg with its surrounding French-curated gardens. By 18:00 we should be ready to take the bus back to the capital *or* we can extend our stay and visit the Animal Farm too.

Looking at the **folklore**, Heemskerk found its identity in the founder Maarten van Heemskerck. A Dutch portrait and religious painter, who spent most of his career in Haarlem. He was a pupil of Jan van Scorel, and

adopted his teacher's Italian-influenced style. In Heemskerk, there is a statue dedicated to him.

At his death, he left money and land in trust to the orphanage of Haarlem, with interest to be paid yearly to any couple who should be willing to perform the marriage ceremony on the slab of his tomb in the cathedral of Haarlem. It was a **superstition** in Catholic Holland that a marriage so celebrated would secure the peace of the dead within the tomb.

4.3.5 Sustainability

As the final step, we evaluate the three alternatives based on different kinds of sustainability:

- **Environmental Protection:** All of the tours and stories used public transports and empowered naturalistic landscapes, some of them – like the Heemskerk one – directly integrated sport activities like biking and paragliding.

  Given its different nature, surely the Heemskerk tour can be seen as the most dangerous from the environmental point of view: it comprehends different fragile habitats like the dunes and the beach. On the opposite side, Gouda and Ameersfoot don't promote tourism near extremely fragile habitats, but just common trails and parks.

- **Societal Responsibility:** Estimating the social impact is difficult in this case because they are three *very* different cities: Gouda is already a domestic tourist point with a rich folklore, Heemskert has a very strong balance between nature and architecture, while Heemskerk can provide for a different – more natural – kind of experience while still retaining the Dutch culture.

  In most of the tour, the historical and artistic value has been empowered and integrated between "green" moments and tasting moments to provide a 360° view on the cultural richness of the localities so many different stakeholders should benefit from the new visitors.

- **Economic Stability:** Given the growing tourism trend on Amsterdam, certainly it should be possible to move *some* of the visitors to external cities, especially considering the efficiency of the transport network.

  Tourism could also be considered a development vector for regions that still need some reinforcements in one or more layers: for example transport in the northern part of the country. Finally, Holland is well known for its capacity to build smart systems and environments, so it shouldn't be a problem for the government to properly use these new resources to raise the wellness of rural areas.

Given these premises, probably the two most interesting proposals would be **Gouda** and **Heemskerk**. While Gouda is already a known touristic destination at the national levels, it may be elevated to the international level with a stronger public promotion and a stronger identity that delve deep into his folklore.

Similarly, Heemskerk can offer to the visitor very different experiences all in one package: from beach to historical buildings, from nature to sport activities. Given its abundance of historical figures, the city could have a lot to offer to culture-enthusiasts with live re-enactments like in Gouda. All these changes are **fairly inexpensive** and can help solving both the overtourism and the rural development problem at once, so they are definitely worth exploring with more stakeholders.

The interactive maps for the case study are available at the links:
`https://pinea.app.carto.com/map/9fb4da65-3501-4190-99f0-a43982148ab0` and
`https://pinea.app.carto.com/map/82a59fe0-1fb1-47cb-9bf0-6cd3f3a459a2`

4.4 Spain - Barcelona

Barcelona, together with Rome and Paris, is one of the main tourist destinations in Europe. More than **1.5 millions people** live inside the city, while more than 5 millions live in the metropolitan area: this makes Cataluna one of the most densely populated regions in the Mediterranean Sea.

Originally, it was founded as a **Roman City** but during the Middle Ages became the capital of the County of Barcelona. Today, Barcelona is a *transport* hub (with a major port, an extensive motorway network, an airport, and a high-speed rail line) as well as a *commercial* and *cultural* hub with the works of Antoni Gaudì and Lluís Domènech i Montaner declared as UNESCO World Heritages Sites.

The city host more than **10 million visitors annually** (pre-covid levels, at the moment they are reaching again the same levels) [42]. As we can see in the image, most of them are mainly interested in art: Sagrada Familia Basilica, Picasso Museum, and El Borne Centre Cultural are all between the most visited attractions of the city. In fact, each of the top 10 most visited attractions – with the exception of FC Barcelona Museum and the Acquarium – are all historical/art related places.

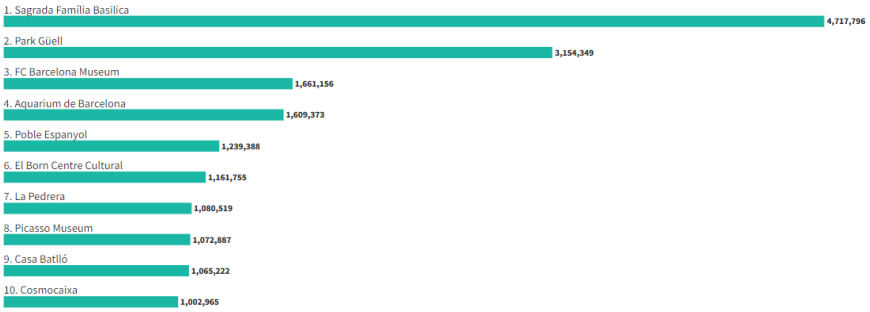

Figure 4.27: Number of Visits in Barcelona City in 2019

Looking at the data, the average tourist is 35-45 years non-Spanish old man that came for a **leisure visit**. Domestic tourism, in fact, make up only about a fourth of the visits, while the other come from international travellers (mainly Americans, French, English and Italians).

The favourite mean of transport is by far the airplane (it takes up more than 80% of travellers) followed by the high-speed train on Madrid-Barcelona trade.

With so many visitors, the city is bound to have some problems: about one out of every four citizens thinks that tourism has not been beneficial while many of the tourist lament too much noise.

The average stay is about **3-4 nights**, but the data has high variance on this statistics. This means that, based on the tourist type, we can have different stay lengths: a domestic tourist will probably stay less time compared to an over-sea tourist.

The **seasonality** is obliviously biased towards the summer, but it's not so sharp like one would expect: the hotel occupancy is just under 1.5 millions stays at its minimum in December, while it's just over 2 millions at its highest in August.

Figure 4.28: Evaluation of Different Aspects by Travel Purpose, from 0 to 10

4.4.1 Analysis

We download the datasets from GeoFabrik and load it on the notebook pipeline. Like in the Amsterdam case study, the Geofabrik region for Barcelona was too small, so we had to download the surrounding regions too and merge them.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	0.4528	1.5210	1.2308	0.3820	0.3884	0.7304	7.2740
gbr	Gradient Boosting Regressor	0.4222	1.5441	1.2380	0.3676	0.3578	0.7381	2.4380
lightgbm	Light Gradient Boosting Machine	0.4231	1.5940	1.2595	0.3569	0.3638	0.7728	0.8080
rf	Random Forest Regressor	0.4570	1.6414	1.2770	0.3410	0.3899	0.7317	4.9010
br	Bayesian Ridge	0.4717	1.8879	1.3707	0.2446	0.3951	0.8294	0.1540
ridge	Ridge Regression	0.4715	1.8893	1.3714	0.2438	0.3956	0.8316	0.1430
lr	Linear Regression	0.4716	1.8902	1.3717	0.2435	0.3957	0.8327	1.8250
omp	Orthogonal Matching Pursuit	0.4663	1.9132	1.3771	0.2431	0.3879	0.8087	0.0560
knn	K Neighbors Regressor	0.4390	1.9179	1.3791	0.2386	0.4165	0.7934	3.6850
en	Elastic Net	0.4814	1.9773	1.4000	0.2195	0.3944	0.8091	0.1040
lasso	Lasso Regression	0.4838	2.0114	1.4110	0.2093	0.3921	0.7972	0.5460
huber	Huber Regressor	0.3179	2.1013	1.4401	0.1764	0.3705	0.9395	1.8320
llar	Lasso Least Angle Regression	0.5783	2.5648	1.5908	-0.0002	0.4344	0.7774	0.0590
dummy	Dummy Regressor	0.5783	2.5648	1.5908	-0.0002	0.4344	0.7774	0.0840
dt	Decision Tree Regressor	0.5230	3.1424	1.7610	-0.2422	0.5000	1.0514	0.2290
ada	AdaBoost Regressor	2.1398	8.7811	2.8732	-2.6456	1.0300	1.7938	4.6710
par	Passive Aggressive Regressor	1.5619	64.5377	7.0102	-25.7017	0.6861	2.9208	0.0890
lar	Least Angle Regression	1.8782	122.2793	5.3597	-53.6597	0.6379	3.1583	0.0710

Figure 4.29: Cross-validation Results for Barcelona Case Study

The pipeline chose – again – the **Extra Tree model** as the best one based on its performance on MSE and R^2, but we can see that a close second is the Gradient Boosting Regressor where the $RMSLE$ is lower and the speed is much higher compared to the best model.

Like other boosting methods, gradient boosting combines weak "learners" into a single strong learner in an iterative fashion. The math behind Gradient Boosting is slightly more complex compared to other models: this is because it's designed to receive any kind of loss function as long is differentiable: this allows the algorithm to be extremely flexible. For example, LightGBM is a particular implementation of this system [43].

The prediction function is slightly more complex compare to the Extra Trees [44]:

$$y_{jm} = \arg\min_{y} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma) \; for \; j = 1, ..., J_m \tag{4.3}$$

where:

- y_{jm} is prediction that minimize the loss for the node j on the tree m
- x_i is the i-th feature of the observation x
- R_{jm} is the residual set for the node j on the tree m
- L is the loss function
- F_{m-1} is the predicted value based on the trees the model already has
- γ is the value to minimize over, the prediction
- m is the m-th tree, comprised between 1 and M
- M is the number of trees that we want to create

Gradient boosting has a couple of assumptions under its belt:

- The phenomenon, object of the study, is describable by a tree model
- Several weak tree models perform better than a single, hyper-optimized, one
- We can procedurally built new tree models on the top of the first one

The **final results are fairly close**, even tough they posses a lower degree of variance, compared to the actual tourism map.

We can see, as always, the black group at the centre of the map, but we can also spot several smaller and greyer groups around the map. First of all, there is a small group, with a very black cell, in the south-west side of the map: it's Vilanova i la Geltrù.

Then, there is a slightly bigger, and greyer group corresponding to Matarò and a belt of interesting cities in the inner part of Barcelona: Ripollet, Sant Cugat del Vallès, la Llagosta, and Mollet del Vallès.

Figure 4.30: Predicted Tourism of Barcelona and its Surroundings

The model use heavily the feature related to entertainment attraction (4 times in the top 10):

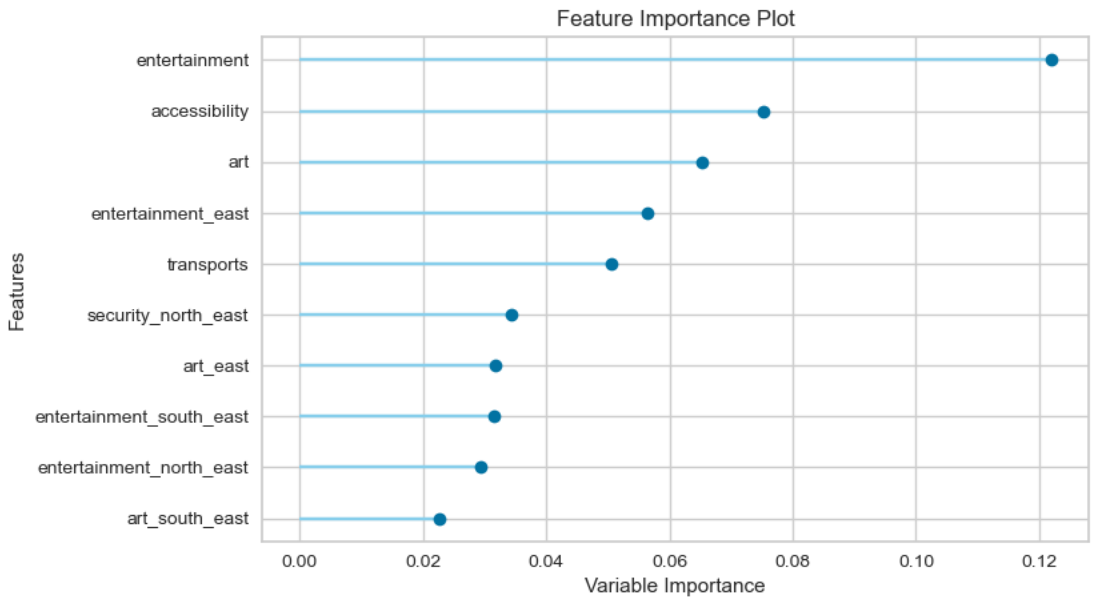

Figure 4.31: Barcelona's Gradient Boosting Feature Importance Plot

As always, we can notice a match with accessibility (and security). Finally, the other important features are *all* related to the cultural world.

It's interesting to analyze this results by **comparing** them with the other two case studies: they tell a story about what the typical tourist want to visit or do in each city. For Barcelona – Europe's night life hub – it's almost granted that many people will seek entertainment, while for Venice the may search for something more sea (nature) or art-related.

Figure 4.32: Actual Tourism of Barcelona and its Surroundings

In the actual tourist map we can see a much **richer picture**: there are several localities in the inner part of the Catalan region that may be potential touristic destinations. The problem is that many of them are not well connected to any of the major transport hub.

In order to create and empower these kind of territories, one of the best strategies is to let the rural tourism "spread" from the central hub – Barcelona – to its surrounding cities, and then repeat the same workflow with the surrounding cities, each time expanding a little more the touristic area and increasing the region touristic capability and synergies.

To understand better the effect of the main features/layers in the computation of the scores, we can analyze them one by one:

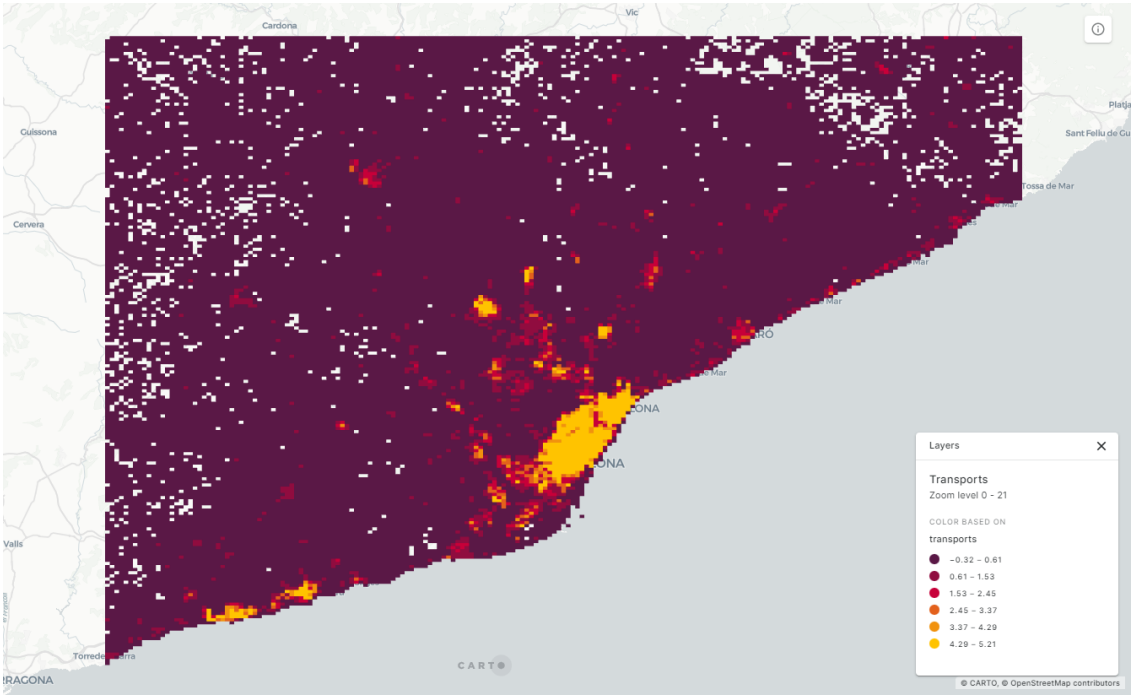

Figure 4.33: Transport Map of Barcelona and its Surroundings

Many of the highly touristic cities also benefit from the large transport network of Barcelona, being directly connected to them. The highest-scoring areas are more than 5 standard deviations away from the average, while the lowest is just -0.32 standard deviations away. We already know that this is a right skewed distribution: the average is much closer to the minimum compared to the maximum, meaning that many points have lower scores with a few strong outliers.

As we can notice from the map, the network is extremely **well developed** around the metropolitan area, but it's less capillary once outside the city. Both Matarò and Vilanova i la Geltrù are well connected, with the former reaching the highest of the 6 level of the graph (in yellow) and the last having a slightly less concentration and slightly bigger area. We will elaborate that part better in the following sections but this scoring is not a coincidence: the Matarò-Barcelona trade is the first (and one of the most trafficked) railway in Spain.

The **inner part** of the city are well connect and there seems to be a main trade along the South-East / North-West axis that is capable of connecting several tens of kilometers inside the country.

On the **extremes of the map**, both on North and West, we can spot several transparent cells: this means that, for that point, the are no relevant markers contained in the original dataset imported from GeoFabrik. This means that no volunteer has ever included a point of interest for that coordinates, it doesn't mean that there is no means of transportation whatsoever.

The transparent cells follow approximately the geographical pattern: in that points there are mountains or hills notoriously difficult to connect with the rest of the transport network.

Next, we are going to analyze the entertainment layer to assess if there are *other* possible destinations with a good concentration of amusement attractions:

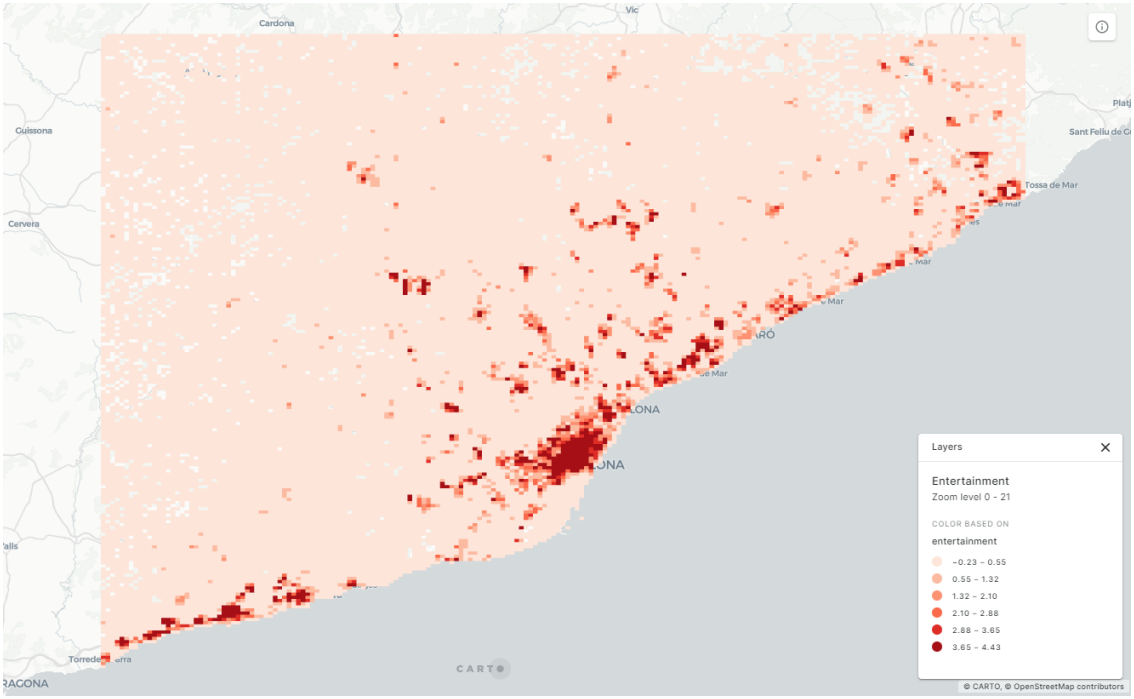

Figure 4.34: Entertainment Map of Barcelona and its Surroundings

Being a world-known entertainment locality, Barcelona score extremely high on this side, but we can see that this trait is common along all the **coastal side** of the region, with many places reaching a deep red (= higher points of interest related to the entertainment industry).

What is most interesting, in this map, is that there are also several inner-land regions that seems to be strong scorer along this feature: the Terrassa area, the Garriga area, and Granollers just to cite a few examples.

This is very important because it may indicate already a sort of developed "rural" tourism, or it may be domestic tourism. Unfortunately, there is not enough historical data to assess the tourism fluxes.

Finally, looking at the **cluster analysis**, we can notice major differences both in size and shape between the clusters. Some of them, the ones closer to Barcelona, then to have more regular shapes and less area, while the ones that are further then to be bigger and less regular.

We also notice that some of them are *very* far away from good candidates: for example the cluster in the upper-left corner is obliviously not feasible due to transport condition. This is a potential weakness of the process: by searching for 20 clusters, there may be some of them constituted by just 4 or 5 interesting cells that are not a good choice for the average tourist. This is one of the reasons why the social part is so important: we can detect this kind of obstacles early on.

Removing the 3 cluster that are in the immediate surrounding of Barcelona, we still have several alternatives ready: Terrassa/Sabadell, Vilanova i la Geitrù, Matarò, Granollers, Sant Andreu de la Barca, Olesa de Montserrat, Canet de Mar, Villafrance del Penedès, and Igualada.

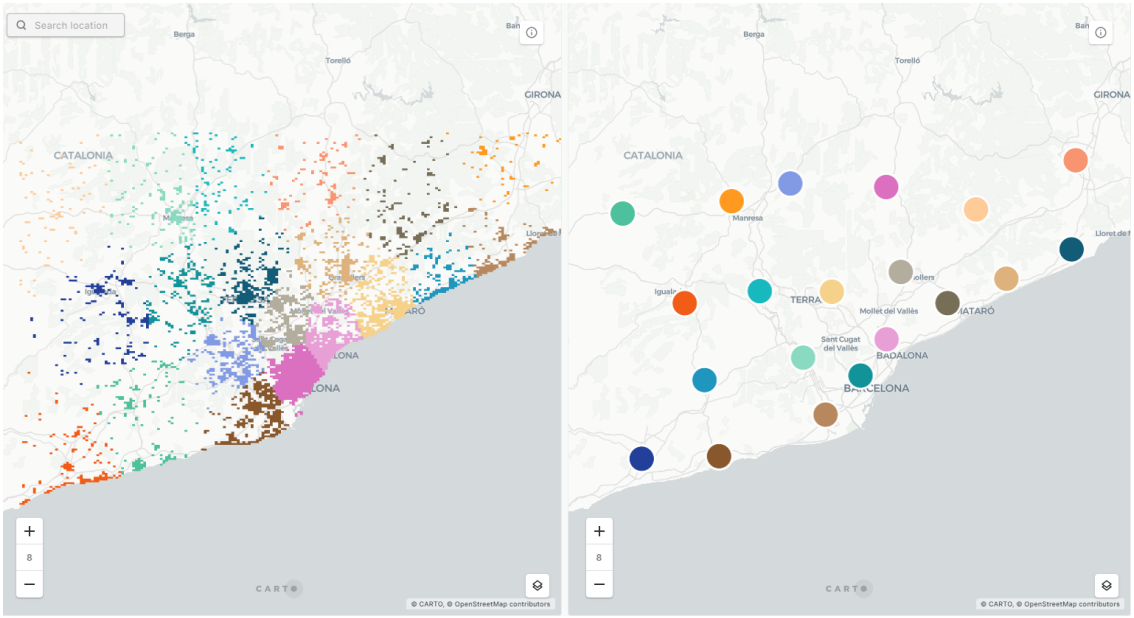

Figure 4.35: Cluster Map of Barcelona and its Surroundings

We can use the contextual filter, alias the travelling distance, to select the 3-4 potential candidates for the next phase.

4.4.2 Filtering

Like in the previous case study, we will use Google Maps estimates to assess the trip time from Barcelona main station:

Destination	Travel Time
Sabadell	44 min
Villanova i la Geitrù	43 min
Matarò	54 min
Granollers	52 min
Sant Andreu de la Barca	44 min
Olesa de Montserrat	62 min
Canet de Mar	66 min
Villafrance del Penedès	60 min
Igualada	123 min

Table 4.9: Travel Time from Barcelona, using Public Transport, for each Destination

As always, the travel time is estimated starting from a comfortable reaching point in the city. In the case of Barcelona, it was the Sants Train Station. Each simulations involved starting the travel at 9:00 and the threshold should be 60 minutes maximum.

We can see the **importance of the transport network** in the Igualada travel time: even tough it's not much further as the crow flies compared to Villafrance del Penedès, it requires more than double the time to reach it (60 minutes vs 123 minutes). As we've discussed in the visualizations before, the Barcelona transport network is strong only in its immediate surrounding, but it suffers extreme loss in efficacy just outside it.

Similar to the Amsterdam case study, we will take the 3 cities as much distant as possible between one and the other, that still respect the 60 minutes limit. In this case, we will use: **Matarò** for the eastern part, **Sabadell** for the northern part, and **Villanova i la Geitrù** for the western part.

4.4.3 Evaluation

We will use this step to deepen our knowledge about the candidates (Matarò, Sabadell, Villanova i la Geitrù) and to discover possible synergies with Barcelona typical tourism, based on night life and culture.

Candidate #1: Matarò

Matarò has a rich history that dates back to the **Roman times**. Nowadays it's an industrial, commercial and tourist hub in Costa de Maresme. The capital of Maresme has more than 127,000 inhabitants and enjoys a privileged location: it is located between the sea and the Serralada Litoral, just **30 minutes north of Barcelona** and very close to the most touristic towns in the area.

Mataró is a city to discover throughout the year, but in summer the activity intensifies. At the end of July the city explodes with the patron saint festivities of Les Santes, declared a **Heritage Festival of National Interest**. Giants and big heads, fireworks, concerts, the Nit Boja and the performances of the towers -a unique practice in the world- are some of the main activities that take place in the city during the festivities.

We can find several historical landmarks whether we walk through the city, like the 15th-century Basilica di Santa Maria, or we take the first railway in the history of Spain, the route Matarò-Barcelona [45].

PoI	Description
Casa Coll i Regàs	A modernist building designed by the Catalan architect Josep Puig i Cadafalch in 1898 and commissioned by the entrepreneur Joaquim Coll i Regas, a major textile manufacturer Mataro (Barcelona). Exponent of the significant elements of decorativism that characterized the modernist movement was declared in 2000 a Cultural Interest, in the category of historical monument.

PoI	Description
Hermitage of St. Simon	A small Spanish parish church located in the east end of the Royal Road, in the faubourg of Havana, in the municipality of Mataró, comarca of Maresme. Dating to 1611, the seaside chapel is well-known along the Catalonia coast. It has a single nave in keeping with ancient seafaring tradition. The Feast Day is 28 October.
Roman villa of Can Llauder	Built in the first imperial period (1st century BC) and remodeled at the beginning of the 3rd century AD, the villa had a rich decor of marble and mosaics, with traces of stucco and paint. It was owned by several wealthy owners who possibly resided in the villa with his family and their slaves. An inscription has been found linking it to Gaius Marius. The villa was used until the Middle Ages when it fell into disrepair and ruin.
Regional Museum of Maresme	The Mataró Museum is a multidisciplinary institution linked to the territory, and whose diverse collections include archaeological materials, natural specimens, historical objects, and artistic and decorative arts. It offers stable research and dissemination programs (exhibitions, publications, school activities, visits, and much more) and houses a natural science documentation center and the archaeological management area of the city of Mataró.
La Beneficiència	Building designed by Puig i Cadafalch in 1894. It combines neo-Gothic elements, such as trigeminal crown windows, with exposed brick imposts and arches and ceramic details that contrast with the smooth finish of the facade. It has been restored by the architect Manuel Brullet and is currently the seat of the Patronat de Cultura.
The First Railway	The nineteen miles by railway between Barcelona and the fishing town of Mataró once served as an important turn in Spain's transportation history. It was the very first railroad in Spain to open. The Maresme line (R1) of Spain's railway system connects Maçanet-Massanes with Molins de Rei, expanding from Gerona to La Selva. Thousands of people use this line to go up and down along the coast of Catalonia, making the R1 among the most utilized in Spain. As a major first in Spain's transit, its source came from a Mataró local who started out as a noble fisherman before he became known all over Spain.

Table 4.10: Matarò's main Points of Interest

Diversification – While Barcelona is mainly famous for the medieval building and attractions, Matarò can offer point of interest in other time periods. For example, we can promote the first hyberic railway – opened in

1848 – or the Roman archeological site of Villa of Can Llauder. The town can also rely on the proficiency of Josep Poig i Cadafalch, a notorious noucentista architect that worked on many of the symbolic buildings of the city.

Figure 4.36: Replica of the First Train on the Route Barcelona-Matarò

Market Penetration – Matarò also offers very similar experiences to Barcelona in terms of coastal activities. It's also worth noting that Matarò has one of the largest biological variety in their waters and, along the promenade, many commercial activities are based on that: shops, restaurants and recreational centre. So, it may be a good opportunity to offer the same kind of activities of "The City of Counts" (alias Barcelona) while keeping the streets less crowded in order to avoid the typical loudness and hyper-activity of the night city.

Candidate #2: Sabadell

Sabadell is currently co-capital of the Vallés Occidental region, together with Tarrasa. It is the **fifth municipality in Catalonia** by population, with 216,520 inhabitants. It obtained the city status in the 19th century thanks to the influence of the textile industry. To this day, Sabadell remains an important industrial hub with metallurgy, chemicals, electrical goods, and leather products [46].

The city was a pioneer in the **Industrial Revolution** in Catalonia within the textile sector and in the mid-19th century it became the most important wool-producing city in Spain; in fact, it was known by the name of "Catalan Manchester" in the second half of the 19th century.

Numerous chimneys and steam rooms can still be seen today, many of them converted into places of social services such as libraries or youth areas . This textile heritage has left a marked industrial character in the city. Over the last decades, Sabadell has been diversifying around the service sector.

There is no agreement among the local historians on the **derivation of the toponym**, some date it back to the Saturday when the weekly market took place and still takes place, others from San Salvador to which a small church was dedicated with the evolution from Salvador to Salvadorell, Salvadell and finally Sabatell. Still others think of the subsequent deformation of the Latin name of the ford vadum or vadellum; in fact there was a ford on the Ripoll river which passes near the city.

PoI	Description
Caixa d'Estalvis	In 1902, and given the growth that Caixa de Sabadell was experiencing, a plot between party walls was acquired in the center of Sabadell, where what should be the headquarters was built, which is known by the name of El Palau de l'Estalvi (El Palacio del Ahorro), becoming a jewel of modernist art. It was officially opened in 1916.
Torre de l'Aigua	A modernist water tank considered one of the city's most emblematic buildings, to the point of being its identifying symbol. It is considered one of the 100 "Elements of the Industrial Heritage of Catalonia " and guided tours can be done on the days of Festa Major, the first weekend of September. It is one of the first constructions made with reinforced concrete in Sabadell (a mixture of gravel, sand and water, in combination with steel bars), which made it possible to erect a slender structure that at the same time had to support a water tank.
Mercat Central	It was during the Festa Major of 1927 when the first stone of what was to be the new Mercat de Sabadell was laid on the esplanade called Camp de la Sang. The work was a project by the local architect Josep Renom, a pioneer in our country in the use of concrete in non-structural areas. In an area of approximately 8,000 m^2 and, after marking the roads, 3,380 square meters of floor space were built, apart from the basements. The New Market was inaugurated in 1930 and operated continuously until the end of 1998, when, seeing that its deterioration prevented the development of normal commercial activity, it was rehabilitated.
Castle of Can Feu	Feu's Tower is a romantic castle in Sabadell (Barcelona). It began as a medieval tower in the XI Century, but then it was destroyed in the XVII century. At the same place, the owners built another building. And finally, in 1881 it was restored by the arquitect August Font under the will of the owner Josep Nicolau d'Olzina, and the romantic laws back then of restoration. The arquitect redecored it with newer towers and built a huge garden arround the castle.

PoI	Description
Parc de Catalunya	The Parque occupies an area of almost 45 hectares. The park has a bike lane, an artificial lake, a jetty (with boats and skates for rent), a miniature train, a skate track, a Biketrial circuit, playgrounds for different ages, an ice rink that works only during the winter, picnic areas, a playground for dogs, gymnastics areas for the elderly, an amphitheater and a wi-fi area.
Casa Duran	This building is the best example of Renaissance architecture in the city due to the many ornamental elements it presents. On the façade you can see a plaque with the date of completion of the building (1606). The house was declared a Historic-Artistic Monument in 1958. In the year 2000 it became municipal property and restoration works began that have lasted eight years and have consisted of the structural consolidation of the building and the rehabilitation of the facades as well as such as the recovery of different spaces making it possible to visit them.

Table 4.11: Sabadell's Main points of Interest

Market Development – Considering the fact that the average tourist is in its late thirties/early forties, we can try and expand the market to the family segment. In fact, Sabadell offer many attraction suitable to families like the Parc of Catalunya and the Castle of Can Feu. The area is also well connected to the main city and there are many local specialties to try.

Figure 4.37: View from Sabadell's Parc de la Catalunya

Product Development – While there are more than enough point of interest, Sabadell seems to struggle in creating a unique file-rouge that can really tell the town history. Several attractions belong to the same time period (the Middle Age), so it make sense to enhance this characteristic through activities and events. By re-creating the unique medieval atmosphere, the city can be more appealing to the tourist eyes.

Candidate #3: Villanova i la Geitrù

Known as "La Habana Chica" in the 19th century, Vilanova i la Geltrú is today one of the main capitals of **popular and traditional culture** in Catalonia. Els Tres Tombs or its carnival, declared a heritage festival of national interest, are some of the festivals that have made the capital of El Garraf a benchmark. Its splendid promenade, where colonial-style buildings coexist harmoniously with fishermen's houses, its Rambla, the city's social and commercial axis, and its wide gastronomic offer of seafood cuisine and xató, make Vilanova a place to go and to visit it again. [47].

It was officially founded in 1274 when King Jaime I granted it the Puebla Charter. In the middle of the 18th century, when King Carlos III allowed Villanova to trade with America, the city experienced a very important economic effervescence. A progress that is not limited to an accumulation of wealth, but rather reports an investment in culture. It is then when the **first recreational societies**, meeting places and entertainment, are founded. At the beginning of the 19th century, the majestic meeting gardens appear, giving it the appearance of a large, cheerful city, due, in large part, to the contact it was acquiring with the island of Cuba. Legend has it that the town was born because the feudal lord of La Geltrú promulgated a law according to which, when a young woman married, she had to spend her first night with him due to the right of seigneur, and many Geltrunenses left, settling nearby. from the sea, in the territories of Cubellas, founding the Villa Nueva de Cubellas. Over time both grew to become one.

Today, the city counts more than 65,000 inhabitants over 65 km^2 and it has become a leisure, cultural and commercial hub thanks to its rich history and peculiar events.

PoI	Description
Museo Romántico Can Papiol	In the 17th century, the Papiol family settled in Vilanova de Cubelles, the current Vilanova i la Geltrú. Its power and properties increased until it became one of the most influential families of the moment. The way to demonstrate this social position was to build a 5-story mansion in the middle of Calle Major, a fact that few families could afford.
La Geltru	The origin of Vilanova i la Geltrú is the village of La Geltrú, documented in the sX, to walk there is to make a leap in time and perhaps in space. You will feel the charm of picturesque corners, deserted streets and cozy squares.

Table 4.12: Villanova y Geitrù's main Points of Interest

PoI	Description
Torre d'Enveja	Large round tower with medieval walls and it's the oldest example of fortification preserved in Vilanova i la Geltrú. The construction is of stone made with small and irregular stones . The openings are framed with stone and some covered loopholes can be seen. In some places it seems that the stones form an "opus spicatum". The dating of this building poses some problems. The type of construction or the rusticity of the upper door suggest that it was made during the first half of the 11th century . Other aspects, such as the slenderness or thickness of the wall, are characteristic of a 13th century watchtower.
Biblioteca Museo Victor Balaguer	The politician and writer Victor Balaguer, as a man of the Catalan Renaixença, was convinced that culture was the basis for the people's progress. For this reason, in 1884, he commissioned the first public building in the country intended to function both as a museum and a library, to be built at Vilanova i la Geltrú, and in which to make his art, book and ethnographic collections available to the public. Currently, the Museum has a collection of more than 8,000 items that include an archaeological and ethnographic collection donated by some illustrious friends of Víctor Balaguer. The highlight being the mummy of a child from ancient Egypt, popularly known as Nesi.
Railway Museum of Catalonya	A valuable heritage facility at the service of people, dedicated to the custody and dissemination of railway culture. Since 1990, the old steam locomotive depot at Vilanova i la Geltrú has hosted one of the most important railway collections in Europe. More than 60 vehicles from all eras, technologies and countries, including 28 steam locomotives from the late 19th century, make up the bulk of the displays at the Museu del Ferrocarril de Catalunya. As well as the technical and historical aspect, the Museum invites you to discover the social and emotional aspects of the world of the train. For this reason, it has been designed as an experience space. And so visitors can go inside the locomotives, travel on the passenger trains and even watch audiovisual projections inside a freight wagon.
Camino de Ronda	A low-difficulty coastal path that links two of the best-known coastal towns of Garraf: Sitges and Vilanova i la Geltrú. Easy to follow as it coincides with the GR 92, this route runs parallel to the Renfe line. The stony coves and beautiful views along the way contrast with some of the monuments visited in the two cities, such as the church of Sitges or the Plaça de la Vila de Vilanova i la Geltrú.

Table 4.13: Villanova y Geitrù's main Points of Interest

Market Development – Similar to Matarò, Villanova has a deep history with trains and railways, in fact it host the Railway Museum of Catalonya. This can be a great selling point and a synergy that can exploited by travel agencies: they can build a sort of "railway" tour that start from Matarò and takes the visitor through Barcelona and Villanova y Geitrù. There are many train-enthusiast – or "railfans" as they call themselves – around the world, just in the US they're are estimated to be around 175.000 [48].

Figure 4.38: The Castle and the Church of La Geltrù

Diversification – Villanova has a strong identity in some very peculiar people, like Victor Balaguer or the Papiol family, that donated to the city something unique. By exploiting this cultural richness and bringing back to life the history of these amazing people, they town can become an "inspirational" hub for many visitors.

There are many stories to tell in this city, and they can appeal to both children and grown up in different ways: looking at the richness of their official touristic website, Villanova seems ready to change and promote them so it may be worth a shot.

4.4.4 Proposal Development

As always, we conclude the case study with the practical part: a one-day possible trip plus folkloric elements. We notice that Spanish cities are particularly rich of the former.

Matarò - Itineraries and Stories

Starting from the Sants Station, we take the train at 9:03 and by 9:54 we're arrived at Matarò Central Station. We

walk toward the historical centre of the city, about 10 minutes in we can already appreciate many of the main attractions: the Regional Museum of Maresme, la Beneficiència, and la Casa Colli i Regàs.

The general advice is to visit immediately the Museum, thus removing the opening hours bottleneck, and then proceeding with the free access point of interest.

By 14:00, we can search for one of the many local restaurant to taste the traditional cousine and, by 15/15:30 we are ready for the next part of the tour. It's important to remember that, in the **Spanish tradition**, people usually eat both lunch and dinner much later compared to central Europe, so many restaurants will have opening hours that will go until the first hours of the afternoon.

After 15-minutes walk, we arrive at the **edge of the city**, to the Hermitage of Saint Simon. We can exploit the opportunity and also take a look at the local flora. By 17:00 we are ready to walk along the beautiful promenade and, while taking a small marine snack, we reach the last attraction of the day: the archaeological site of the Villa Romana de Torre Llauder. The visit is relatively short, so we can comfortably be back at the train station for the 19:00 train.

By 19:48 we're back at the Sants Station.

Matarò's **folklore** is extremely rich: they have a special cult for the Three Magi (also called "the Three Wise Men") and have a very elaborated carnival.

They also host several peculiar fairs like the Tres Tombs and the Saint Ponç Fair that has been done for centuries. Its origins date to the 16th century, when in Spring herbalists took medicine to the sick. Today Sant Ponç fair is celebrated to preserve the antique customs.

As a final note, Matarò also host the international dance festival "Days of Dance".

Sabadell - Itineraries and Stories

Like before, we start from the Sants Station at 9:12 and we arrive at 9:53 at Sabadell Centre. The tour will be one of the shortest one: it's only 5.4 kilometers spread across a day.

From the central station, we head north to the city's symbol: the Torre de l'Aigua. Here, we can get a general view of the **town's identity** and we can also appreciate the elaborated garden "Parc del Tauli". We then walk along Carrer de Vilarrùbias to arrive to one of the biggest park in the city: "Parc de Catalunya". Here, there are several activities for all ages: picnic areas, playground for dogs and children, gymnastic areas and a mini-train.

We can resume our tour by midday, to go and taste the local cousine at the Mercat Central while visiting the **historic centre** of the city. Finally, we visit the jaw-dropping Casa Duran: a masterpiece of traditional Spanish architecture and and an onbliged stop for every art enthusiast.

We can then take our time to go back to the train station, it's only 10 minutes away, but the numerous and well-balanced experiences of this short-day trip have surely left something in the tourist mind.

Sabadell doesn't boast the same level of cultural identity seen in Matarò, and it's difficult to find resources about it. There seems to be some traditional dances left, but it's difficult to tell for sure.

Like previously mentioned, Sabadell has a rich cultural heritage from the Middle Age and it could really make a difference to use and promote it.

Villanova y Geitrù - Itineraries and Stories

As always, we start from the Sants Station at 9:04 and we arrive at 9:44 at Sabadell Centre. This is gonna be the

shortest path at 4.3km of length: less than an hour by foot. The tour will only last half a day, so it may be optimal for people that doesn't stay in Barcelona too long but still want to experience something different.

Just outside the station, we head north toward the Geltru historical neighborhood were we can still see the remains of the Geltru Castle. The second stop in the tour is an **emotional-intensive experience**: the Can Papiol Romanticism Museum. The visit is about 1-2 hours long but it will be something that needs to be "digested". To counter-balance the strong emotions of the last stop, we head for the Torre d'Enveja. Surely, this is a more relaxing experience thanks to the green surroundings of the Parc de la Quadra d'Enveja.

After a short while, we can head back to the station: in fact, we will find both the Biblioteca Museu Victor Balaguer and the Railway Museum. Both of them are built to put the visitor in a first-person point of view and they deliver both the technical and the emotional part of the story at the same time. There are no right duration for these visit, so we can be quite flexible with this regard.

Technically the tour is finished here and we could head back by the first hours of the afternoon (around 16:00), but there is also the well-know **Camino de Ronda** that will lead the sporty tourist to the city of Garraf while passing numerous point of interest, both of cultural and natural value.

Like Matarò, Villanova y Geltrù is rich in **folklore**: there is a local saying that goes like "always have a leg in the air" referring to their festive proclivity.

Also like Matarò, the Carnaval culminates in a week-long debauch of dances, feasts, and processions. All this is done in honour of Sa Maastat el Rei Carnestoltes also known as the "king of the senseless" celebrated for his prodigious sexual prowess and devastating satire. Up to a third of the population participates in Les Comparses, a couples dance in which rival groups hurl hard candies at one another in what is called the Sweet War.

But there are also other festivals, for example in early August the Vilanovins celebrate their Festa Major, dedicated to the city's patron saint (the Virgin of the Snows). Processions begin with a correfoc of ritual devils led by the Ball de Diables de Vilanova i la Geltrú, established in 1832 and one of eight dances of devils in Catalonia with a history of one hundred years or more.

4.4.5 Sustainability

As the final step, we evaluate the three alternatives based on different kinds of sustainability:

- **Environmental Protection:** Like the previous case studies, all the tours require just public transport and walking by foot. Therefore, they may contribute to reduce the carbon footprint of each visitor.

  Contrary to the Holland case study, there all much less incentives towards sportive activities and the majority of tourism can be re-directed, at the moments, towards different cultural and natural alternatives.

  Each of the three cities presented didn't have an amount of green comparable to Barcelona, or even Holland and Italy case studies. So this may be a factor to reason upon with the other stakeholders.

- **Societal Responsibility:** Each of the three cities offer many points of view regarding their historic aspects: going from the medieval city of Sabadell to the festive Vilanova. But they seems to lack venues to promote modern cultural movements and to give voice to many minorities.

  It may be interesting to explore these opportunities because the Spanish cultural was always a melting pot of different identities.

  Looking at the final product - the tours - all of them are accessible by disabled persons and they don't require large sums of money to be accessed. Additionally, they don't even require a bike like in the Netherlands case study.

- **Economic Stability:** Barcelona, together with Paris and Rome, is one of the cities that suffer more of overtourism. So, be creating special identities for surrounding localities and re-directing paying travellers to them it can lower the pressure while retaining the tourists' expectations.

  The new cities are not isolated and the majority of them had their own history and industrial prowess, so they should be able to accommodate for new people without too much effort. The main problem may be the peak season, alias summer, where they can struggle to supply enough hosting capacity.

Given the already **structurally sound identity and folklore**, the best choices to invest at the moment seem to be Matarò and Villanova y Geltrù. Both of them already possess a – local – festive spirit and that's the secret to become attractive to the average Barcelona tourist.

The major problem with regard these cities came with the **green spaces**: if you want to seem them you need to go at the outskirt of the town. Also, given that they have around 5-10 attractions that a tourist may want to see, it may be a good idea to up the game in the information and logistic sector: this means create a dedicated website for tourist (Villanova already has it, but it's from more than 5 years ago and it's not mobile-responsive) and promote a tourist pass for the transport and the major point of interest.

All of these suggestion can be implemented with relatively low budget and short time frames just to taste the terrain and, only if the market respond positively, it's possible to expand the services.

The interactive maps for the case study are available at the links:
`https://pinea.app.carto.com/map/a01b3181-1a98-4833-866c-259e12313948` and
`https://pinea.app.carto.com/map/aac859b6-b1e6-4499-a13d-cd1f01f5da31`

5
Conclusion

After a brief introduction about the born, the assumptions and the key ideas of the **project** we delve into the technical part. Using freely available datasets from GeoFabrik it's possible to gather lot of geo-referenced informations about any European region. These datasets were first converted and then elaborated using domain-specific libraries like PyCaret and GeoPandas.

One of the main steps was the conversion of the dataset into a map of adjustable size (both regarding the area and the number of cells inside it). This allows for more informative visualizations down the line and easier comparisons between different territories.

Regarding the Machine Learning part, the first models built weren't performing much better compared to the trivial model so few extensions were implemented mainly through the use of **ensemble models**. Ensemble models didn't seem to increase the accuracy by a relevant margin, so the only solution was to increase the number of features by including also neighboring cells. This has proven - using k-fold Cross Validation - a much better strategy that could make the difference between the built model and the trivial one.

The social part started were the technical part ended: by analyzing the outputted maps. By adding to the visualizations a couple of key information retrievable from the main touristic offices, the situation became clearer.

Using the maps, we select a few candidates and we filter them by a contextual filter (in the case study was distance by public transport) and began to analyze the **peculiarities** of each candidate. For each candidate, a brief tour and story - whenever possible - was reported. This output, together with all the visualizations, can be used by the relevant stakeholders to make future decisions about tourist management.

The **main findings** of the project are:

1. It's possible to build models that **predict tourism** using information regarding entertainment, sport, accessibility, security, transports, culture, and nature. This means that the 7 features have prediction power. None of the models used the coordinates, or any geo-spatial related, features.

2. The models are much more accurate if they took into account the **neighboring cells**: this may indicate that tourism is a *systemic* phenomenon.
3. For different regions, different models have **different prediction powers**. We can exploit this fact to study the underlying assumption of each model to understand better the tourist phenomenon in a specific region. The same reasoning can be applied by studying the relative importance of each feature for the best perfoming model.
4. Complicated workflows are not necesssary: with a combination of PyCaret, Carto, and GraphCommons it's possible to create a **simple process**. This simplicity can be used to better explain the models and the pipeline to different stakeholders, increasing the communicative capacity of the discipline.
5. Finally, from a socio-economical point of view, it's clear that there are an indefinite number of **locations** that may be extremely interesting tourist destinations, with this project we can identify few of them, but a more extensive development program seems necessary.

5.1 Future Works

The project is just a **PoC** at the moment, but there are several aspects that can be further developed.

For example, given the computational complexity and the dataset dimension, **Neural Networks** were excluded by the possible algorithms but they could offer a whole new level of prediction accuracy.

Another interesting aspect is that we have just used "plain" numbers to describe each cell in the area, but each monument/park/point of interest is unique. **Losing this "uniqueness"** is a major setback for most models because different people will react differently to the same point of interest. Also, each point of interest hold a different type of cultural, historical and geological importance.

Literature is also a relevant bottleneck at this time: overtourism studies are linked to the call for sustainability that has raised on the last decades, so many long-term studies on the subject aren't published yet and a consistent body of work still has to be built. Nevertheless, the studies published all points to similar conclusions regarding the effect of some policies and tourism management.

Finally, for timing reasons **many stakeholders weren't included in the Proof of Concept creation**. So the workflow may still have room for improvement based on the feedback given by the different actors.

References

[1] Statista, "Overtourism in european destinations - statistics facts," December 2020. [Online]. Available: https://www.statista.com/topics/4316/overtourism-in-european-destinations/#topicOverview

[2] R. Stanchev, "The most affected european destinations by over-tourism," 2017. [Online]. Available: https://dspace.uib.es/xmlui/bitstream/handle/11201/148140/Stanchev_Rostislav.pdf

[3] E. Povoledo, "Visiting venice? make a reservation and be ready to pay." *New York Times*, July 2022. [Online]. Available: https://www.nytimes.com/2022/07/01/world/europe/venice-tourism-register-pay.html

[4] J. Buckley, "Italy has a new way to combat overtourism," *CNN*, March 2021. [Online]. Available: https://edition.cnn.com/travel/article/uffizi-diffusi-tuscany-galleries-overtourism/index.html

[5] S. Carvão, K. Koens, and A. Postma, "'overtourism'? – understanding and managing urban tourism growth beyond perceptions," 2018.

[6] Interreg Europe, "A policy brief from the policy learning platform on environment and resource efficiency," December 2020. [Online]. Available: https://www.interregeurope.eu/sites/default/files/inline/Sustainable_Tourism_Strategies_to_counteract_overtourism.pdf

[7] Policy Department for Structural and Cohesion Policies, "Research for tran committee - overtourism: impact and possible policy responses," 2018. [Online]. Available: https://www.europarl.europa.eu/RegData/etudes/STUD/2018/629184/IPOL_STU(2018)629184_EN.pdf

[8] M. J. Andrade, J. P. Costa, and E. Jiménez-Morales, "Challenges for european tourist-city-ports: Strategies for a sustainable coexistence in the cruise post-covid context," *Land*, October 2021. [Online]. Available: https://doi.org/10.3390/land10111269

[9] G. Cosenza, "Pnrr: gli investimenti in borghi e giardini storici e nelle imprese culturali," *Corriere della Sera*, August 2021. [Online]. Available: https://www.ilsole24ore.com/art/pnrr-investimenti-borghi-e-giardini-storici-e-imprese-culturali-AEJElGf

[10] H. Goodwin, "Responsible tourism partnership," February 2022. [Online]. Available: https://responsibletourismpartnership.org/platform-for-change/responsible-tourism-enters-its-3rd-decade/

[11] Eco-Union, "Managing (over)tourism in natural protected areas: learnings from national parks in spain and europe," 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5532456

[12] F. Ramm, "Openstreetmap data in layered gis-format," 2022. [Online]. Available: https://download.geofabrik.de/osm-data-in-gis-formats-free.pdf

[13] R. Hyndman and Y. Fan, "Sample quantiles in statistical packages," *The American Statistician*, 2009. [Online]. Available: http://www.jstor.org/stable/2684934?origin=JSTOR-pdf

[14] M. Ali, "Optimize the model," *PyCaret Official Documentation*. [Online]. Available: https://pycaret.gitbook.io/docs/get-started/functions/optimize

[15] Sci-Kit, *K-Means*, sci-kit learn - version 1.1.3 ed., 2022. [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#k-means

[16] R. R. Canale and R. D. Siano, "Territorial pressure and tourism contribution to gdp: The case of italian regions," *International Journal of Tourism Research*, 2021. [Online]. Available: https://doi.org/10.1002/jtr.2451

[17] A. Mortesen and S. Braithwaite, "The italian gardens hoping to change tourism," *CNN*, July 2021. [Online]. Available: https://edition.cnn.com/travel/article/italy-gardens-botanical-tourism-cmd/index.html

[18] A. Montanari and B. Staniscia, "Rome: a difficult path between tourist pressure and sustainable development," *Rivista di Scienze del Turismo*, 2010. [Online]. Available: https://www.ledonline.it/Rivista-Scienze-Turismo/Allegati/RST-I-2-17-Montanari-Staniscia.pdf

[19] G. Hospers, "Overtourism in european cities: From challenges to coping strategies," 2019. [Online]. Available: http://hdl.handle.net/10419/216242

[20] Assessorato al Turismo, "Yearbook of tourism data," 2019. [Online]. Available: https://www.comune.venezia.it/sites/comune.venezia.it/files/immagini/Turismo/Adt19%20ing%20ver%204%201%202021%281%29.pdf

[21] S. Brady, "Day-trippers to venice must soon pre-book and pay to enter," 2022. [Online]. Available: https://www.lonelyplanet.com/news/venice-introduces-new-booking-system-and-entry-fee-for-visitors

[22] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. [Online]. Available: https://papers.nips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[23] Turismo Venezia, "Parco fluviale di san donà." [Online]. Available: http://www.turismovenezia.it/Jesolo-ed-Eraclea/Parco-fluviale-di-San-Don-150271.html

[24] Dipartimento di Archeologia dell'Università di Padova, "Land reclamation museum - san donà di piave." [Online]. Available: http://www.archeoveneto.it/portale/wp-content/filemaker/stampa_scheda_sintetica_inglese.php

[25] Microturismo delle Venezie, "Villa ancillotto-marcato." [Online]. Available: https://www.microturismodellevenezie.it/scheda/crocetta-del-montello-villa-ancilotto-marcato/

[26] Italian Botanical Heritage, "Park of sculpture in architecture." [Online]. Available: https://luoghi.italianbotanicalheritage.com/en/park-of-sculpture-in-architecture/

[27] Comune di Mirano, "Itinerari turistici," 2022. [Online]. Available: https://www.comune.mirano.ve.it/it/page/itinerari-turistici-205255b1-fd6f-47ef-a304-22805a715b40

[28] Punto e Viaggio, "Leggende del piave: dove nascono gli spiriti della natura e altre storie." [Online]. Available: https://puntoeviaggio.it/leggende-del-piave/

[29] Y. Cassis, *Capitals of Capital: A History of International Financial Centres 1780-2005*. Cambridge University Press, November 2006.

[30] Statista, "Total contribution of travel and tourism to gdp in the netherlands in 2019 and 2020," 2022. [Online]. Available: https://www.statista.com/statistics/810736/travel-tourism-total-gdp-contribution-netherlands/

[31] I. I. Apriani *et al.*, "Final report fairbnb science shop: Monitoring impact and assessing data driven solutions," 2019. [Online]. Available: https://www.wur.nl/upload_mm/8/2/1/26f38e3d-743a-4c78-8833-5fe17b8b974b_Sustainable%20tourism%20in%20Amsterdam%20Final%20Report.pdf

[32] M. Brene *et al.*, "Tourism in amsterdam: Today and tomorrow." Ecorys, October 2018. [Online]. Available: https://news.airbnb.com/wp-content/uploads/sites/4/2020/02/03122018_Tourism-in-Amsterdam_.pdf

[33] R. van Loon and J. Rouwendal, "Travel purpose and expenditure patterns in city tourism: evidence from the amsterdam metropolitan area," *J Cult Econ*, July 2017. [Online]. Available: https://doi.org/10.1007/s10824-017-9293-1

[34] Municipality of Amsterdam, "City in balance 2018-2022: Towards a new equilibrium between quality of life and hospitality," March 2019. [Online]. Available: https://www.etoa.org/wp-content/uploads/2020/08/Amsterdam-City-in-Balance-2018-2022.pdf

[35] "How extra trees classification and regression algorithm works," ArcGIS Pro, 2022. [Online]. Available: https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-extra-tree-classification-and-regression-works.htm

[36] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach Learn*, March 2006. [Online]. Available: https://doi.org/10.1007/s10994-006-6226-1

[37] P. van den Belt, "Beelden goudse kaasmarkt gaan hele wereld over," March 2017. [Online]. Available: https://www.ad.nl/gouda/beelden-goudse-kaasmarkt-gaan-hele-wereld-over~a8234771/

[38] D. G. Waag, "De goudse waag," June 2022. [Online]. Available: http://goudsewaag.nl/en/

[39] Bok Tower Gardens, *Sint Janskerk (Gouda, Netherlands)*, 1966. [Online]. Available: https://cdm16755.contentdm.oclc.org/digital/collection/p16755coll3/id/639/rec/1

[40] Amsterdam.info, "Verzetsmuseum (dutch resistance museum)," 2022. [Online]. Available: https://www.amsterdam.info/museums/verzetsmuseum/

[41] Olanda.cc, "Guida alle città olandesi - gouda," 2022. [Online]. Available: https://www.olanda.cc/gouda.html

[42] Observatori del Turisme de Barcelona, "Key figures 2019," 2019. [Online]. Available: https://www.observatoriturisme.barcelona/en/key-figures-2019

[43] T. Masui, "All you need to know about gradient boosting algorithm," *Towards Data Science*, January 2020. [Online]. Available: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502

[44] J. H. Friedman, "Greed function approximation: A gradient boosting machine," *Ann. Statist. 29*, October 2001. [Online]. Available: https://doi.org/10.1214/aos/1013203451

[45] Spain's Official Tourism Website, "Matarò," 2022. [Online]. Available: https://www.spain.info/en/destination/mataro/

[46] A. Tikkanen *et al.*, "Sabadell - spain," 2010. [Online]. Available: https://www.britannica.com/place/Sabadell

[47] C. O. Website, "Vilanova i la geltrú - guía de municipios," 2022. [Online]. Available: https://www.catalunya.com/vilanova-i-la-geltru-2-1-83073?language=es

[48] S. Chen, "Trains are life for avid 'railfans'," *CNN*, 2009. [Online]. Available: https://edition.cnn.com/2009/TRAVEL/05/08/railfan.train.watching/index.html

Acknowledgments

I would personally like to thank Fondazione San Paolo, in particular *Francesca Gambetta* and *Luca Grbac*, for the support they gave to me during the development of the social part.

I would also like to thank my supervisor, *Prof. Tomaso Erseghe*, for being incredible open and supportive of my idea from the very start until the end.

Finally, I'm grateful to everyone near me during this academic path for the motivational support.