Predicting potential travel locations

Jaime Velazquez

12 August 2020

1. Introduction

1.1 Background

Tourism without a doubt one of the biggest business ventures that many global markets partake in. It is vital to many economies around the world, such as Mexico, Spain, and Jamaica. In Jamaica alone, over 30% of the country's total employment is based on the travel industry alone. Apart from being very important to the global economy, traveling is understood to be one of the more prominent leisure activities in people's lives. People look forward to and even dedicate their lives to traveling the world and experiencing new places, food, and cultures.

1.2 Problem

One issue some people tend to have when traveling is deciding where to go. This seems to be more common with people who have traveled to a few places but aren't sure where to go next. At that point they would have already decided what they like and don't like about their travel experiences, and what types of places resonated with them. For example, Puerto Vallarta is vastly different from London, England. This project aims to use the preferences and experiences of travelers to help determine what their best match for their next travel trip would be.

1.3 Interest

Companies in the travel industry already collect data on their users to build their profiles and preferences. A company like Airbnb who specializes in reservations also has a feature that allows the user to explore the surrounding area of their stay. Once a trip is complete, Airbnb could use my tool to recommend other cities to visit near places of interest and offer reservations there.

2. Data Acquisition & Cleaning

2.1 Data Sources

Initially, I wanted to focus on cities that are tourism hotspots around the world. However, no such dataset or database was available on such a large scale. I decided to just include cities in Europe using a Wikipedia page[1] that had a list of urban areas with a population of 500,000 or

[1] Wikipedia page: https://en.wikipedia.org/wiki/List_of_urban_areas_in_the_European_Union

more. I chose this page because I was unable to find a list of cities within Europe that were tourism hotspots that contained more than 10 cities on it. I wanted a larger list to be able to provide a wider range of location recommendations to the user. In total, I used 79 cities and municipalities across Europe. In order to get the location data for each city, I used the Nominatim Python package from Geopy to get the longitude and latitude. Finally, I used the Foursquare API to get the venue information for the first 100 venues within a 500 meter radius from each city's coordinates.

2.1 Data Cleaning

Cleaning the data was relatively easy as the data scraped from the Wikipedia page was already in a list format. I converted that information into a Pandas DataFrame, and then cleaned up each name of the city by getting rid of things like parentheses to comply with the input standards to use Nominatim. Finally, adding the coordinate data for each city was simple and led to a complete dataframe.

3. Methodology

3.1 Exploratory Data Analysis

To begin, I wanted to see what kind of data Foursquare provided for each city. Before exploring each city, I wanted to explore just one so I created an API call request for the city of Rome, Italy. The results contained the information in json format, where the actual info we needed was nested within the json itself. I created a simple function that extracted the information we needed for each venue. I extracted the name of the venue, the category, the latitude, and the longitude, and created a DataFrame using this information. Using this DataFrame, I was able to identify which types of venues were more common in Rome. I created a visualization of a map of Rome with the listed venues superimposed (Figure 1).

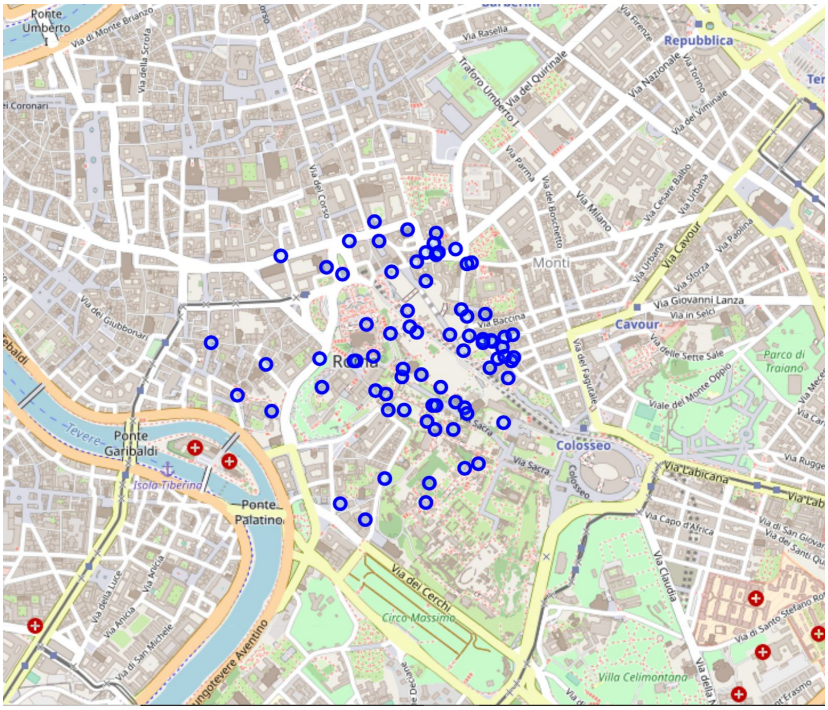

Figure 1: Map of Rome, Italy with venues superimposed

It's clear to see a large cluster of venues near the central part of Rome. But what are these venues exactly? I took a closer look at the venues' categories to get a better understanding of what goes on within the city. In order to do this, I used the DataFrame containing Rome's venue data and created a bar graph to illustrate which types of venues were most common.

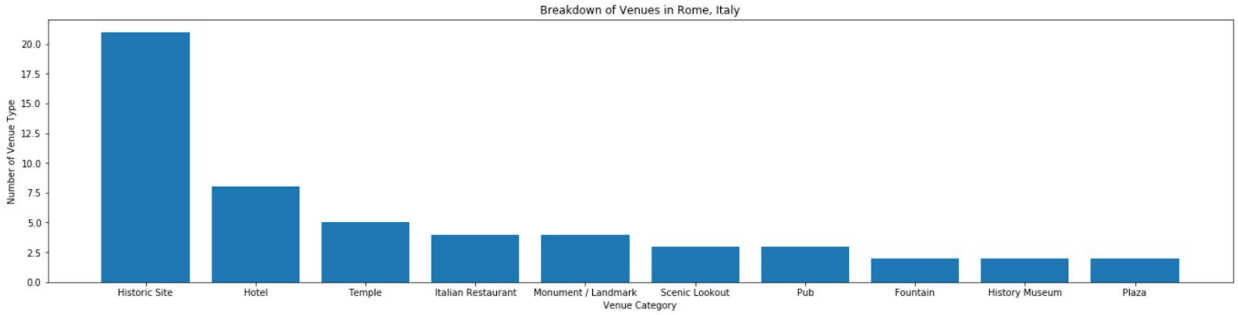

Figure 2: Bar plot of the breakdown of venues within Rome

By looking at this chart (Figure 2), I was able to clearly identify what makes Rome so unique compared to many places around the world. The city is crawling with historical sites all clustered around in a tight-knit area. After realizing this, I remembered that on my trip, I would wander aimlessly around the city center and find monument after monument and historical site after historical site just by stumbling upon them. This is one of the reasons why I loved Rome, being able to be lost in a sea of history. Upon realizing this, I made note of all the venue types that

made me fall in love with Rome, and began to search for the next best place for me to travel. Before I could make any attempt at figuring that out, I had to repeat the whole process above with all of the cities within our DataFrame that contains the cities in Europe we will have access to explore. This was done quickly using a function that quickly repeats the process, and fills a new DataFrame with all of the information needed for each city. We now had a DataFrame full of specific venues from cities all around Europe. We can take a look and see all of the unique venue categories and see that there are nearly 350 kinds. This is great as the many kinds of venues allows for cities to be unique and distinguishable from others. Figure 3 shows us the top 10 most common venue types within Aachen, Amsterdam, Antwerp, Athens, and Barcelona.

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Aachen	Bakery	Drugstore	Café	Plaza	Ice Cream Shop	Sushi Restaurant	Bar	Coffee Shop	German Restaurant	Cocktail Bar
1	Amsterdam	Bar	Hotel	Coffee Shop	Marijuana Dispensary	Italian Restaurant	Pub	Cocktail Bar	Bakery	Cheese Shop	Museum
2	Antwerp	Bar	Coffee Shop	Plaza	Cocktail Bar	Restaurant	Italian Restaurant	Bookstore	Belgian Restaurant	French Restaurant	Sushi Restaurant
3	Athens	Bar	Café	Coffee Shop	Theater	Hotel	Falafel Restaurant	Greek Restaurant	Bookstore	Dessert Shop	Kafenio
4	Barcelona	Tapas Restaurant	Plaza	Ice Cream Shop	Bar	Spanish Restaurant	Mediterranean Restaurant	Italian Restaurant	Hotel	Cocktail Bar	Pizza Place

Figure 3: Table containing most common types of venues for first 5 cities listed alphabetically

3.2 Machine Learning

Now that we have all of the information we need, we're almost ready to make use of some machine learning techniques. One major machine learning technique that has seen increased use over the past decade is the recommender system. A recommender system, or a recommendation system (sometimes replacing 'system' with a synonym such as platform or engine), is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item. Companies like Netflix and Hulu use this system to create a list of options for the user to watch. This is essentially what I hope to accomplish with a list of cities to visit rather than films.

To begin creating this system, we must one-hot encode the DataFrame we have created that contains the information on the venues in each city. Then, we must group this list by city and take the mean of each row. Finally, we have a DataFrame that has the percentage of each type of venue for each city in Europe. The reason we need to do this is these percentages act as weights that define the city. That is, the higher the percentage is for a specific category of venue, the more venues of that type the city has. For example, as we have seen earlier, the venue most common within the center of Rome is the 'Historical Site'; therefore, the 'Historical Site' column in the 'Rome' row of the DataFrame will have the highest percentage of that row. Now, we need to create another DataFrame that contains the preferences of the user. This DataFrame

will contain similar information to that of the city DataFrame, with the key difference being an integer value of ‘1’ or ‘0’ for each category. ‘1’ will be placed in categories where the user enjoys this type of venue, and would be interested in exploring other places of this type. All other categories will be left as ‘0’. Finally, we have a user matrix and our city matrix. The last is simply performing the dot product of these two matrices, which will output a new matrix with a score for each city. The higher the score, the stronger the recommendation to travel to that city based on the preferences.

4. Results

As an example of using this methodology, the user matrix I created contained ‘1’s within the 'Historic Site', 'Monument / Landmark', 'Temple', 'Plaza', 'Church', and 'Art Gallery' categories. These types of venues were most prevalent within Rome. Upon performing the dot product with the city matrix, these were the following results.

	Score	City	Country	Latitude	Longitude
0	0.440000	Rome	Italy	41.8933	12.4829
1	0.333333	Ruhr	Germany	51.5175	7.14392
2	0.227273	Palermo	Italy	38.1112	13.3524
3	0.176471	Lyon	France	45.7578	4.83201
4	0.156250	Hamburg	Germany	53.5438	10.0099
5	0.153846	Catania	Italy	37.5022	15.0874
6	0.150000	Genoa	Italy	44.4073	8.93386
7	0.150000	Bologna	Italy	44.4937	11.343
8	0.150000	Florence	Italy	43.7699	11.2556
9	0.140000	Seville	Spain	37.3886	-5.99534
10	0.130000	Kraków	Poland	50.0619	19.9369

Figure 3: Table containing scores for top 10 recommended cities to visit.

As shown in Figure 3, Rome comes out on top as the best place to visit. However, in our case we based our preferences off of the venues the city has to offer, so we can scratch it off the list. The rest of the list contains the top 10 best places for us to visit. Notably, A majority of the places within the top 10 are other cities in Italy, which makes sense considering they have many similar venues. However, we were able to create an interesting list that contains cities in other countries such as France, Spain, Poland, and Germany. The table also contains the coordinates of these cities, in case the user would like to explore the venues in a city using the Foursquare API.

5. Discussion

This tool I created can be seen as preliminary in the effort to create a fleshed out application that can give users recommendations on places to travel to and eventually, specific venues that the user might enjoy. To begin updating this tool, it's necessary to create a more profound algorithm to create the recommender system. Currently, the algorithm used is effective but primitive, and can definitely be expanded upon using better and newer methods of this technology. The algorithm could be improved simply by providing extra user information. Currently, it is limited to only one user's preferences. However, with multiple user's information, the system in place could grow from a content-based system to a hybrid between content-based and collaborative, which uses other user's preferences to help make recommendations that the user might be interested in. I would also expand on the tool to provide recommendations on specific venues to visit. The Foursquare API provides a rating system for each of the venues within their database. However, these ratings are provided through premium API calls which come at a price. If a company has the resources to use this tool, or finds a way to get this information through other means, the ratings for each venue would be integral to creating a recommendation system for specific venues in a city, based on these ratings. One issue with the results achieved above would be that the first city that is recommended to visit is Ruhr, Germany, which is more of a region or district that contains several cities. This is more so an issue with the data used, which was solely based on Urban areas over 500,000 inhabitants, with no guarantee of being cities, specifically. However, with a better dataset, the algorithm would be able to provide more accurate results.

6. Conclusion

In this study, I combined Foursquare API data and data on European cities to create a tool that creates recommendations on where to travel based on a user's travel preferences. This was done using the machine learning technique called a recommender system that is used by many major companies such as Netflix and Hulu when creating movie/show recommendations for their users. When fully fleshed out and implemented, this tool would be a great asset to any company that specializes in travel services.