Data Science Capstone Project

Destination Cities - Vacation Travel Planning

June 9th, 2020

Tim Krompinger

1. Introduction

1.1 Background

On a monthly basis hotel and travel sites experience millions of users visiting their platforms. Yearly revenue for these sites can be in the billions of dollars as most US adults now prefer to book their travel plans online. However, with so many site options consumers are also spending much more time switching between sites to research possible travel plans, in many cases spending weeks prior to booking their destination.

1.2 Problem

Hotel and Travel sites have a need to encourage users to book travel destinations but when making decisions about vacation plans, individuals may not know specifically where they want to travel next or they may want to visit a new location but the location they desire is over their budget causing those users to leave the site without booking their next trip. In these cases, can we automatically provide suggestions for similar travel destinations based on venues and attributes by leveraging the Foursquare location data?

1.3 Interest

By providing location recommendations the traveler who may not know where they want to visit next but knows the locations they enjoyed previously can narrow their destination options to those which are similar to locations they have already enjoyed. In the case of the individual who may have a location in mind that exceeds their budget, providing similar destinations can allow the individual to seek a similar experience in a different location that may fit in their budget. This application of the Foursquare data can be useful in search logic for travel or hotel sites to solve the decision problems facing travelers and in turn increase booking rates.

2. Data Acquisition and Cleansing

2.1 Data Sources

To solve this problem and execute an example of this functionality the following data was used:

- My own list of possible vacation locations was predetermined as a sample set. These locations include specific areas of possible destination cities that may be of interest to visit on a vacation. A total of 14 U.S. locations have been included in the sample list which is composed of the following:

  "Savannah Historic District,Savannah,Georgia"

  "Downtown Charleston,Charleston,South Carolina"

  "Southwest Orlando,Orlando,Florida"

  "Garden District,New Orleans,Louisiana"

  "Fisherman's Wharf,San Francisco,California"

  "Beacon Hill,Boston,Massachusetts"

  "Bell Rock,Sedona,Arizona"

  "Downtown,Key West,Florida"

  "The Loop,Chicago,Illinois"

  "Downtown,Houston,Texas"

  "Downtown,Nashville,Tennessee"

  "Union Station,Denver,Colorado"

  "Downtown,Asheville,North Carolina"

  "Downtown,St. Augustine,Florida"

- The latitude and longitude coordinates have been determined for each of these locations to build proper associations that allow us to call and analyze location data. A Nominatim geocoder tool was used to establish these coordinates.

- The Foursquare API was used to return location data from the Foursquare database to find venue information for the indicated area in the destination cities outlined above. This information includes such venues as:

  "Concert Hall"

  "Tea Room"

  "Bar"

  "Burger Joint"

  "Historic Site"

ect.

The venue information was analyzed across all locations and used to determine similarities between the possible destinations. For example, destinations with a large number of historic sites may be similar or locations with a high number of concert halls or music venues. Additional information regarding the complete analysis of this section is expressed in the Methodology section below.

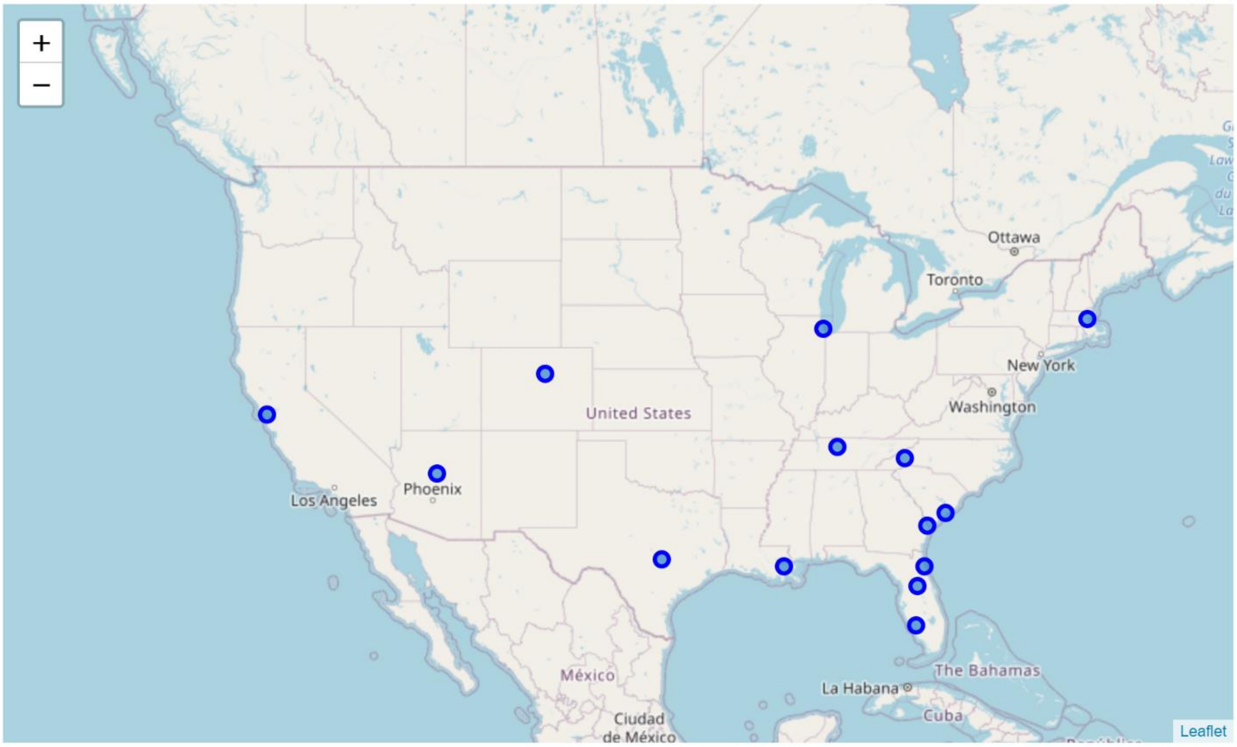
3. Methodology

3.1 Exploratory Data Analysis

I started with my own list of possible destination cities for a vacation which intentionally included specific areas in each city to replicate a traveler who would be staying in a hotel or rental within a specific area of the city and therefore would tie their experience to the area they visited. This data was merged with latitude and longitude information from a Nominatim geocoder to render a dataframe with the main components City, State, Latitude, and Longitude sorted by State:

	State	City	Latitude	Longitude
0	Arizona	Sedona	34.831453	-111.775264
1	California	San Francisco	37.809167	-122.416599
2	Colorado	Denver	39.753630	-105.000748
3	Florida	Orlando	28.876887	-81.695584
4	Florida	Key West	26.642532	-81.862867
5	Florida	St. Augustine	29.904286	-81.319455
6	Georgia	Savannah	32.072732	-81.093158
7	Illinois	Chicago	41.881609	-87.629457
8	Louisiana	New Orleans	29.929605	-90.084388
9	Massachusetts	Boston	42.358708	-71.067829
10	North Carolina	Asheville	35.593791	-82.556748
11	South Carolina	Charleston	32.777847	-79.965938
12	Tennessee	Nashville	36.163366	-86.783091
13	Texas	Houston	30.265002	-97.739304

A folium map was created with the Destination Cities superimposed based on the geographical coordinates rendered in the list above to visualize the possible destination cities that could be visited on a vacation:



Then the foursquare api was called to render the venues surrounding each of the coordinates indicated for the destination cities with a limit of 100 venues and a radius of 8,000 meter. The first city in the list was reviewed in detail regarding the information returned in the foursquare api json response. A function was then created to create a list of all venues and relevant information for every Destination City proposed for a total of 593 venues. The head for this list is as follows:

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Sedona	34.831453	-111.775264	Mystic Trailhead	34.833255	-111.775751	Trail
1	Sedona	34.831453	-111.775264	Mountaintop Therapy	34.831454	-111.771963	Massage Studio
2	Sedona	34.831453	-111.775264	Sedona Fire District 9/11 Memorial	34.834600	-111.776480	Memorial Site
3	San Francisco	37.809167	-122.416599	Musée Mécanique	37.809333	-122.415952	Museum
4	San Francisco	37.809167	-122.416599	The Baked Bear	37.807447	-122.417310	Ice Cream Shop

The total number of venues for each destination city were analyzed to assess the dataset. It became clear in analysis that 2 of the cities (Chicago and Houston) returned the maximum number of venues based on the parameter limits set for the Foursquare API call. While the remaining cities varied in the total number of venues returned. These results are based on the latitude and longitude information selected for each city and would vary with each area selected with the given city for analysis. Once again, the areas were intentionally selected to mimic areas of interest within each city for a possible hotel or rental resulting in a specific impression of the destination based on this stay and surrounding venues. The venue summation information for each city is shown here:

	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
City						
Asheville	40	40	40	40	40	40
Boston	37	37	37	37	37	37
Charleston	15	15	15	15	15	15
Chicago	100	100	100	100	100	100
Denver	71	71	71	71	71	71
Houston	100	100	100	100	100	100
Key West	7	7	7	7	7	7
Nashville	53	53	53	53	53	53
New Orleans	42	42	42	42	42	42
Orlando	2	2	2	2	2	2
San Francisco	45	45	45	45	45	45
Savannah	47	47	47	47	47	47
Sedona	3	3	3	3	3	3
St. Augustine	31	31	31	31	31	31

3.2 K-means Clustering

In total 180 unique venue categories were identified in this dataset. The top ten venue categories for each of these cities were then rendered as seen in the head of the following table:

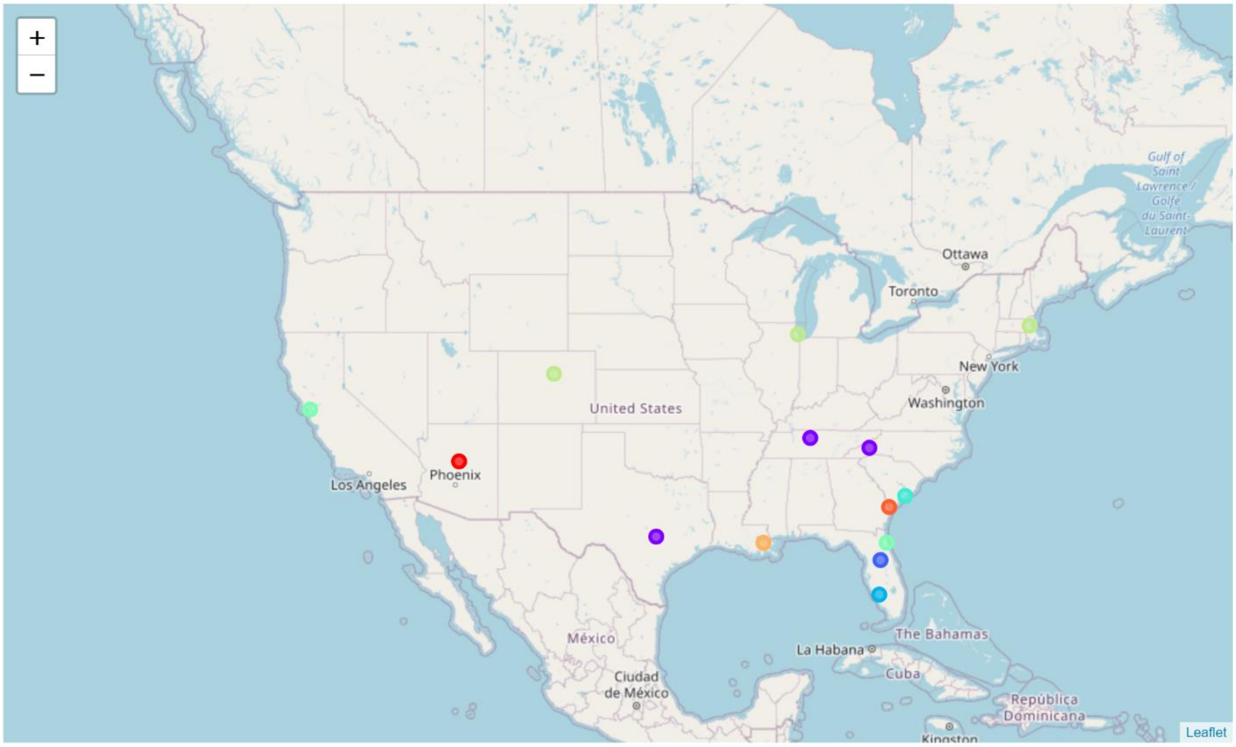
	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Asheville	Hotel	Brewery	Wine Bar	Bar	Coffee Shop	Dessert Shop	Cocktail Bar	Spa	French Restaurant	Chocolate Shop
1	Boston	Pizza Place	Italian Restaurant	Hotel	French Restaurant	Hotel Bar	Plaza	Park	Outdoor Sculpture	Other Repair Shop	Optical Shop
2	Charleston	Hotel	Boat or Ferry	Harbor / Marina	Hotel Bar	Tourist Information Center	Breakfast Spot	Boat Rental	Bar	Kitchen Supply Store	Sporting Goods Shop
3	Chicago	Coffee Shop	Theater	Hotel	Bakery	Middle Eastern Restaurant	Snack Place	Shoe Store	Concert Hall	Museum	Sandwich Place
4	Denver	Hotel	Coffee Shop	Restaurant	American Restaurant	Mexican Restaurant	Cocktail Bar	Pizza Place	Sushi Restaurant	Gym	New American Restaurant

Since there are common venue categories across destination cities, I used a K-means algorithm to cluster the cities for this unsupervised data set to provide recommendations based on venue similarities. Cities were clustered into 9 possible clusters and labels were merged into the dataframe below:

	State	City	Latitude	Longitude	Cluster Labels	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	Arizona	Sedona	34.831453	-111.775264	0	Memorial Site	Massage Studio	Trail	Yoga Studio	Donut Shop	Food Court	Food & Drink Shop	Fast Food Restaurant	Farmers Market	Falafel Restaurant
1	California	San Francisco	37.809167	-122.416599	5	Ice Cream Shop	Tour Provider	Historic Site	Seafood Restaurant	Bike Rental / Bike Share	Gym / Fitness Center	Food Truck	Pharmacy	Hotel	Pizza Place
2	Colorado	Denver	39.753630	-105.000748	6	Hotel	Coffee Shop	Restaurant	American Restaurant	Mexican Restaurant	Cocktail Bar	Pizza Place	Sushi Restaurant	Gym	New American Restaurant
3	Florida	Orlando	28.876887	-81.695584	2	Performing Arts Venue	Italian Restaurant	Dive Bar	Food Truck	Food Court	Food & Drink Shop	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Exhibit
4	Florida	Key West	26.642532	-81.862867	3	Theater	American Restaurant	Convenience Store	Brewery	History Museum	Science Museum	Tea Room	Dive Bar	Fast Food Restaurant	Farmers Market
5	Florida	St. Augustine	29.904286	-81.319455	5	Historic Site	Hotel	Pizza Place	History Museum	Intersection	Fast Food Restaurant	Breakfast Spot	Boutique	Museum	Fried Chicken Joint
6	Georgia	Savannah	32.072732	-81.093158	8	Plaza	Bed & Breakfast	Bookstore	American Restaurant	Museum	Bistro	Coffee Shop	Playground	Pizza Place	Breakfast Spot
7	Illinois	Chicago	41.881609	-87.629457	6	Coffee Shop	Theater	Hotel	Bakery	Middle Eastern Restaurant	Snack Place	Shoe Store	Concert Hall	Museum	Sandwich Place
8	Louisiana	New Orleans	29.929605	-90.084388	7	Public Art	Furniture / Home Store	Historic Site	Coffee Shop	Breakfast Spot	Accessories Store	Neighborhood	Bookstore	Burrito Place	Bus Stop
9	Massachusetts	Boston	42.358708	-71.067829	6	Pizza Place	Italian Restaurant	Hotel	French Restaurant	Hotel Bar	Plaza	Park	Outdoor Sculpture	Other Repair Shop	Optical Shop
10	North Carolina	Asheville	35.593791	-82.556748	1	Hotel	Brewery	Wine Bar	Bar	Coffee Shop	Dessert Shop	Cocktail Bar	Spa	French Restaurant	Chocolate Shop
11	South Carolina	Charleston	32.777847	-79.965938	4	Hotel	Boat or Ferry	Harbor / Marina	Hotel Bar	Tourist Information Center	Breakfast Spot	Boat Rental	Bar	Kitchen Supply Store	Sporting Goods Shop
12	Tennessee	Nashville	36.163366	-86.783091	1	Hotel	Coffee Shop	Steakhouse	Cocktail Bar	Mexican Restaurant	Sushi Restaurant	Concert Hall	Music Venue	Bar	Library
13	Texas	Houston	30.265002	-97.739304	1	Bar	Hotel	Nightclub	Cocktail Bar	American Restaurant	Burger Joint	New American Restaurant	Juice Bar	Grocery Store	Steakhouse

4. Results

The result of the K-means clustering algorithm creates 9 clusters out of the 14 possible destination cities used for this project. These clusters are visualized in the folium map below:



The specific clusters are outlined as follows with references to the most common venues associated to each destination city that were used to define this clustering in the K-means algorithm. These clusters can then be used to provide recommendations for other possible vacation locations based on similarities between a city's venues by suggesting other cities within the same cluster:

Cluster 1:

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
Sedona	Memorial Site	Massage Studio	Trail	Yoga Studio	Donut Shop	Food Court	Food & Drink Shop	Fast Food Restaurant	Farmers Market	Falafel Restaurant

Cluster 2:

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
Asheville	Hotel	Brewery	Wine Bar	Bar	Coffee Shop	Dessert Shop	Cocktail Bar	Spa	French Restaurant	Chocolate Shop
Nashville	Hotel	Coffee Shop	Steakhouse	Cocktail Bar	Mexican Restaurant	Sushi Restaurant	Concert Hall	Music Venue	Bar	Library
Houston	Bar	Hotel	Nightclub	Cocktail Bar	American Restaurant	Burger Joint	New American Restaurant	Juice Bar	Grocery Store	Steakhouse

Cluster 3:

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
Orlando	Performing Arts Venue	Italian Restaurant	Dive Bar	Food Truck	Food Court	Food & Drink Shop	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Exhibit

Cluster 4:

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
Key West	Theater	American Restaurant	Convenience Store	Brewery	History Museum	Science Museum	Tea Room	Dive Bar	Fast Food Restaurant	Farmers Market

Cluster 5:

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
Charleston	Hotel	Boat or Ferry	Harbor / Marina	Hotel Bar	Tourist Information Center	Breakfast Spot	Boat Rental	Bar	Kitchen Supply Store	Sporting Goods Shop

Cluster 6:

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
San Francisco	Ice Cream Shop	Tour Provider	Historic Site	Seafood Restaurant	Bike Rental / Bike Share	Gym / Fitness Center	Food Truck	Pharmacy	Hotel	Pizza Place
St. Augustine	Historic Site	Hotel	Pizza Place	History Museum	Intersection	Fast Food Restaurant	Breakfast Spot	Boutique	Museum	Fried Chicken Joint

Cluster 7:

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
Denver	Hotel	Coffee Shop	Restaurant	American Restaurant	Mexican Restaurant	Cocktail Bar	Pizza Place	Sushi Restaurant	Gym	New American Restaurant
Chicago	Coffee Shop	Theater	Hotel	Bakery	Middle Eastern Restaurant	Snack Place	Shoe Store	Concert Hall	Museum	Sandwich Place
Boston	Pizza Place	Italian Restaurant	Hotel	French Restaurant	Hotel Bar	Plaza	Park	Outdoor Sculpture	Other Repair Shop	Optical Shop

Cluster 8:

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
New Orleans	Public Art	Furniture / Home Store	Historic Site	Coffee Shop	Breakfast Spot	Accessories Store	Neighborhood	Bookstore	Burrito Place	Bus Stop

Cluster 9:

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
Savannah	Plaza	Bed & Breakfast	Bookstore	American Restaurant	Museum	Bistro	Coffee Shop	Playground	Pizza Place	Breakfast Spot

5. Discussion

The study performed in this project was created based on a subset sample list of possible destination cities that I determined for a vacation. This allowed me to proof and test the concepts of clustering cities within a hypothetical travel or hotel site to develop a recommendation engine for additional city options based on location venues. However, in a more robust application this methodology can be extrapolated to cover a much larger scope of available destinations to provide even more clusters and more options within each.

By using a K-means algorithm to provide results for this functionality and solve the problems presented, it is expected that performance in matching similar cities by using common venues from locations data would improve with a larger dataset of available locations. For this increase in scope additional analysis may be required to determine the proper number of clusters to use depending on the size of the data set. To determine the optimal k value in this case a function to dynamically assess the data based on the elbow method and data set size might be implemented to improve results.

In addition to using venue categories for assessment, other variables may be included in the algorithm in a similar fashion to cluster similar destination cities. A process could also be implemented to establish a user selected variable that is most important to them in order to compare locations beyond just similarities in overall venues at a location such as narrowing venue scope to only restaurants or historic locations depending on an individual's travel interests.

6. Conclusion

Based on the results of the methodology implemented on the specified dataset we can determine separate clusters for the destination city list. In application, these clusters allow us to implement functionalities such as a recommendation engine in a hotel or travel site in turn solving the problems initially outlined in this report.

For a traveler who may not know where they want to visit next but knows the locations they enjoyed previously we can narrow their destination options using the results of our K-means cluster algorithm to present them with a list of cities within the same cluster as the one they have already previously enjoyed. As an example from this study, a user who enjoyed visiting Chicago would be recommended to visit Boston or Denver as their next possible destination. In the case of the individual who may have a location in mind that exceeds their budget we can now offer recommendations for other destination options which may provide a similar experience that are more aligned with their budget. For example, if a user had interest in visiting San Francisco within our case study we can suggest to them St. Augustine. On a hotel or travel site these recommendations can be aligned with booking options in these specified areas allowing users to immediately assess their new options, reduce the time spent researching on multiple travel sites, and encourage more immediate travel bookings. By removing the barriers to the decision-making process for their next travel destination we intend to increase booking rates for sites with these implementations.

7. References

- Foursquare API: https://developer.foursquare.com/
- Nominatim geocoder