Food Venues Analysis in Tunis and Vienna

Introduction

Tunis has always been one of the most important cities of North Africa. The cuisine of Tunisia, is a blend of Mediterranean and Berber cuisines. Its distinctive spiciness comes from the many civilizations which have ruled the land now known as Tunisia: Romans, Vandals, Byzantines, Arabs, Spanish, Turkish, Italians (Sicilians), French, and the native Punics-Berber people.

On the other hand, **Vienna** has been the capital of Austria for more than a thousand years. It became the cultural centre of the nation and developed its own regional cuisine. Viennese cuisine is best known for its pastries, but it includes a wide range of other unique dishes.

In this project, we will **analyse the the distribution of food venues in these two capitals : Tunis and Vienna.** We are interested in comparing the differences between the cities in their food offer.

This project will help to answer the following questions:

- Which city provides easier access to food ?
- In which city, a new food store has less competitors ?
- Where in the city, one should launch his food business ?

The target of this work is both the customer and the investor.

For the customer, we will see which city provides easier access to food venues with a better spatial distribution of the venues.

For the investor, we will compare the offer of both cities and thus the food market “competitiveness” and we will see in which place we have more opportunity to get a successful food investment.

Data

We will get our data essentially from **Foursquare** using their Places API.

Foursquare API allows us to search for a specific type of venues, to explore a particular venue, to explore a Foursquare user, to explore a geographical location, and to get trending venues around a location.

The access to the API is provided through the different endpoints available. The following endpoints can be helpful for this specific project:

- **Search** : Search for venues
- **Explore**: Get venue recommendations
- **Trending**: Get trending venues

We will be also using the **Geopy** library to convert addresses to latitude/longitude coordinates in order to use them in Foursquare API.

Methodology

For each city, we will divide the work into 3 parts:

1. Import Data

- Use Geopy to get lon/lat coordinates of the city
- Use **Foursquare API** to explore venues in the city. **Explore** Endpoint is used while specifying **food** as a section. We will retrieve the maximum of 2500 venues within a 5km range from the city center.

Geographical coordinates of the cities:

First, using Geopy we will get the lon/lat coordinates of Tunis and Vienna. In this part, we can notice that for the case of Tunis when we enter as address : “Tunis, TN”, we get a false coordinates pointing to another city.

In order to correct this misbehavior, I used the address of a known monument of the city which is the municipal theater of Tunis.

2. Clean Dataset

- **JSON_normalize** is used to parse the data and convert it to a **pandas** dataframe.

- We get a dataframe containing 22 columns for each city. We are going to use the columns : venue.name, venue.location.lat, venue.location.lng
- We visualize the venues we get using **Folium** library to make sure the data is correct.

Food venues Dataframe:

Then using the explore endpoint of Foursquare API, I've got the list of the food venues within 5km from the center of both cities. Then, I have put all this information in a dedicated data-frame.

	referralId	reasons.count	reasons.items	venue.id	venue.name	venue.location.address	venue.location.lat
0	e-3-4c828d86d8086dcb96b57752-1	0	[{'summary': 'This spot is popular', 'type': '...	4c828d86d8086dcb96b57752	Wetter	Payergasse 13	48.213652
1	e-3-5b0d16809b047300398e59db-2	0	[{'summary': 'This spot is popular', 'type': '...	5b0d16809b047300398e59db	The Pelican Coffee Company	Pelikangasse 4	48.215774
2	e-3-4b5afafbf964a52038dd28e3-3	0	[{'summary': 'This spot is popular', 'type': '...	4b5afafbf964a52038dd28e3	Hitomi Sushi	Josefstädterstraße 53	48.210085
3	e-3-4bce153c29d4b713fdeaa7dc-4	0	[{'summary': 'This spot is popular', 'type': '...	4bce153c29d4b713fdeaa7dc	Konoba	Lerchenfelder Str. 66-68	48.206923
4	e-3-4be2c0201dd22d7ffeb094bd-5	0	[{'summary': 'This spot is popular', 'type': '...	4be2c0201dd22d7ffeb094bd	PARS	Lerchenfelder Strasse 149	48.208169
...	...	...	...	...	...	...	...

Image 1 Pandas Dataframe of food venues

Map of food venues:

Using Folium library, we will plot the map of the food venues in both cities.

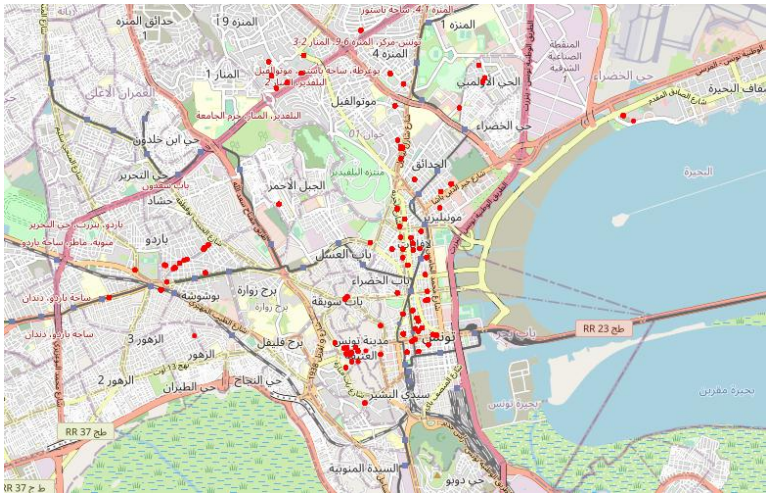

Image 2 Food Venues in Tunis

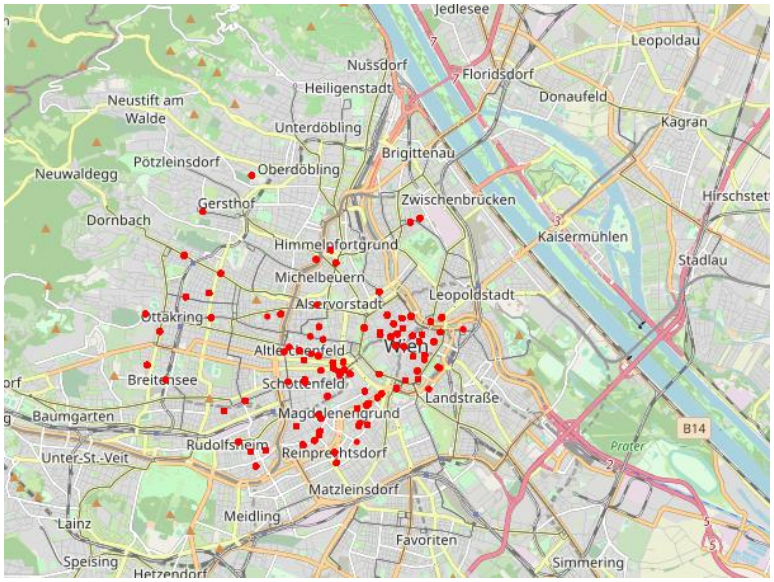

Image 3 Food Venues in Vienna

3. Make Clusters

- First we need to get the optimal number of clusters in each city. Two methods are employed: **Elbow curve** and **Silhouette score**.
- We cluster the venues using **K-means** and the spatial coordinates of the venues
- We display information about each cluster and we visualize it directly on the map.

Calculating the optimal number of clusters:

Choosing the optimal number of clusters is not an easy task as we have to make a compromise between the numbers of clusters and the sum of squared errors or the silhouette score.

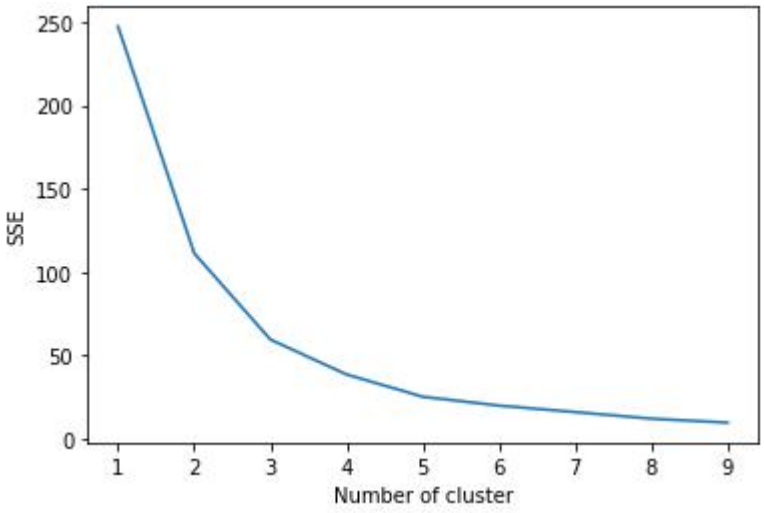

Image 4 Elbow Method

Using the **elbow method**, if the line chart resembles an arm, then the “elbow” (the point of inflection on the curve) is a good indication that the underlying model fits best at that point. In our case, the point of inflection happens when **k=3**.

```
For n_clusters=2, The Silhouette Coefficient is 0.514838021006832
For n_clusters=3, The Silhouette Coefficient is 0.5915796297089859
For n_clusters=4, The Silhouette Coefficient is 0.567363423484236
For n_clusters=5, The Silhouette Coefficient is 0.5947678468448009
For n_clusters=6, The Silhouette Coefficient is 0.6000634980279541
For n_clusters=7, The Silhouette Coefficient is 0.6008165752984389
For n_clusters=8, The Silhouette Coefficient is 0.6180266737545652
For n_clusters=9, The Silhouette Coefficient is 0.5780224289728945
For n_clusters=10, The Silhouette Coefficient is 0.6299583987983361
```

Image 5 Silhouette coefficient scores

On the other hand a higher **Silhouette Coefficient score** relates to a model with better-defined clusters. We cannot rely on this method with our dataset as we get increasing scores as the number of clusters k increases. Yet, we can see that k=3 is better than k=4.

The same process is applied to each city and we found out that k=3 is the optimal values for both cities.

Results

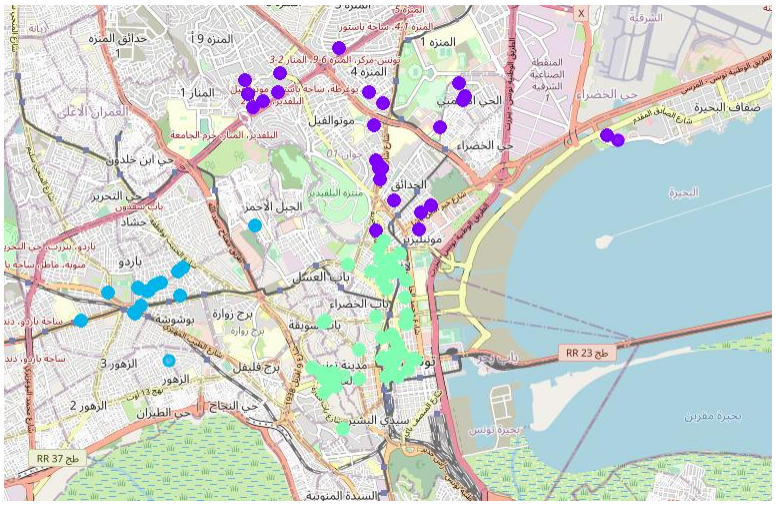

Image 6 Tunis Clusters

Here we can see that we have a big dense cluster in the middle of the capital, and 2 smaller clusters located in the north and the west of the capital's center. We can also notice a gap zones between the clusters where there is no food venues.

So, if you are in Tunis center you will find no problem to eat. However, as soon as you get further from the center, you will get some problems finding a place for eating.

Thus, an investor can profit from the lack of food demand in these locations to open his new restaurant.

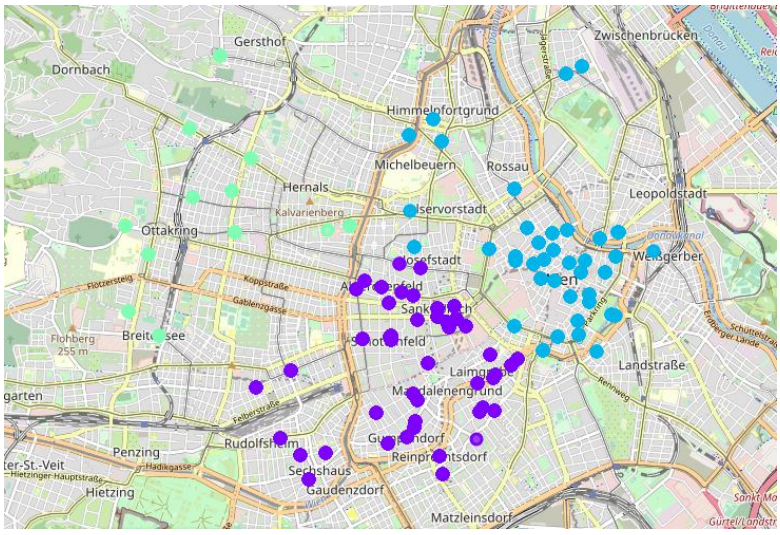

Image 7 Vienna Clusters

Here we can see that we have two big dense clusters in the middle of the Vienna, and a smaller cluster located in the west of the capital's center. In the capital center, food venues are available almost everywhere. So wherever you are, you are almost sure that you will find a food venue nearby.

The only places where you might have problems to find where to eat is in the north or the west of the capital's center. Those places are potential opportunities for investors who want to open new restaurants in Vienna.

Discussion

These are some points/recommendations about the project that I would like to discuss :

- After converting an address to lon/lat coordinates, always check the result on the map to verify if the conversion was correct. If not, try to change the address to some other known close addresses.
- Using Foursquare API, we can easily reach the limit of exploring 2500 venues. This can have impact on the results of our study since the more data we have the more accurate we are.
- Getting the optimal number of clusters is really tricky as we need to compromise between a smaller number and less squared root sum of error.

Even after using the elbow curve method and calculating the silhouette scores, the choice of number of clusters is not very obvious.

Conclusion

In this project, we have used Foursquare API and Geopy to gather information about food venues in both Tunis and Vienna. After scrapping the data and cleaning it, we have created clusters using K-means on the spatial coordinates of the venues.

With the clusters visible to us, we can compare the food offer distribution between the 2 capitals.

As a **consumer viewpoint**, In Tunis, we have a big dense cluster in the center of the city where there are a lot of restaurants. However, there is a lack of food venues as soon as we get out of the Tunis center. In Vienna, food venues are better distributing in the capital and it is much easier to eat wherever you are in the city.

As an **investor viewpoint**, it is more interesting to invest in a new food venue in Tunis, especially in the north/west part of the city. In Vienna, there is much more competition in the food market, but if you have to invest you would better do it in the west part of the city where there is a lack of food venues.