LIMA CITY OVERVIEW – SIMILARITIES AND DIFFERENCES IN LIMA RESTAURANTS

1. INTRODUCTION

1.1. Background

Lima is the capital of Peru, and this city is mainly divided into five zones: north, mid, modern, south and east. Each of these areas has different features, whether in the economic, social and cultural aspects. For instance, while the mid-town Lima is the most traditional part of the city (where the main government buildings are located), the business center and the corporate buildings are in modern districts. Hence, considering only the recreational aspect, it would be very useful to know the most popular types of venues within each of these five areas. Moreover, it might be relevant to include an assessing of how their similarities and differences could be connected. This will be convenient to generate information for the business sector based on a market research as any market study has an overview of the features as starting point over a target; in this case, over a specific sort of venue. Therefore, the two sort of questions that this study tries to solve could be "which type of venue *X* is better in the north zone rather than in the east of the city?", and "are the venues in the south more similar to the north than the center to the east?". As Peruvians well-known for their good cuisine, we are going to evaluate only restaurants as our main venue.

1.2. Problem

This study will list and visualize the different districts that make up each of the five areas of Lima. With this information, we will determine which is the most frequent category of a restaurant in each area and their level of acceptance. Moreover, we will assess the similarities to one area of the city with another based on this information, given some insights from it.

2. DATA ACQUISITION AND CLEANING

2.1. Data Sources

To find the districts of Lima city, we will scrap the references of Lima in its Wikipedia website (Wikipedia Lima Web Article). Likewise, to segment each of these districts by their corresponding area, the same process will be executed from the following link also from Wikipedia (Wikipedia Lima Districts Chart).

- Wikipedia Lima Web Article: https://es.wikipedia.org/wiki/Lima_Metropolitana
- Wikipedia Lima Districts Chart: https://es.wikipedia.org/wiki/Categor%C3%ADa:Distritos_de_Lima

2.2. Data Cleaning

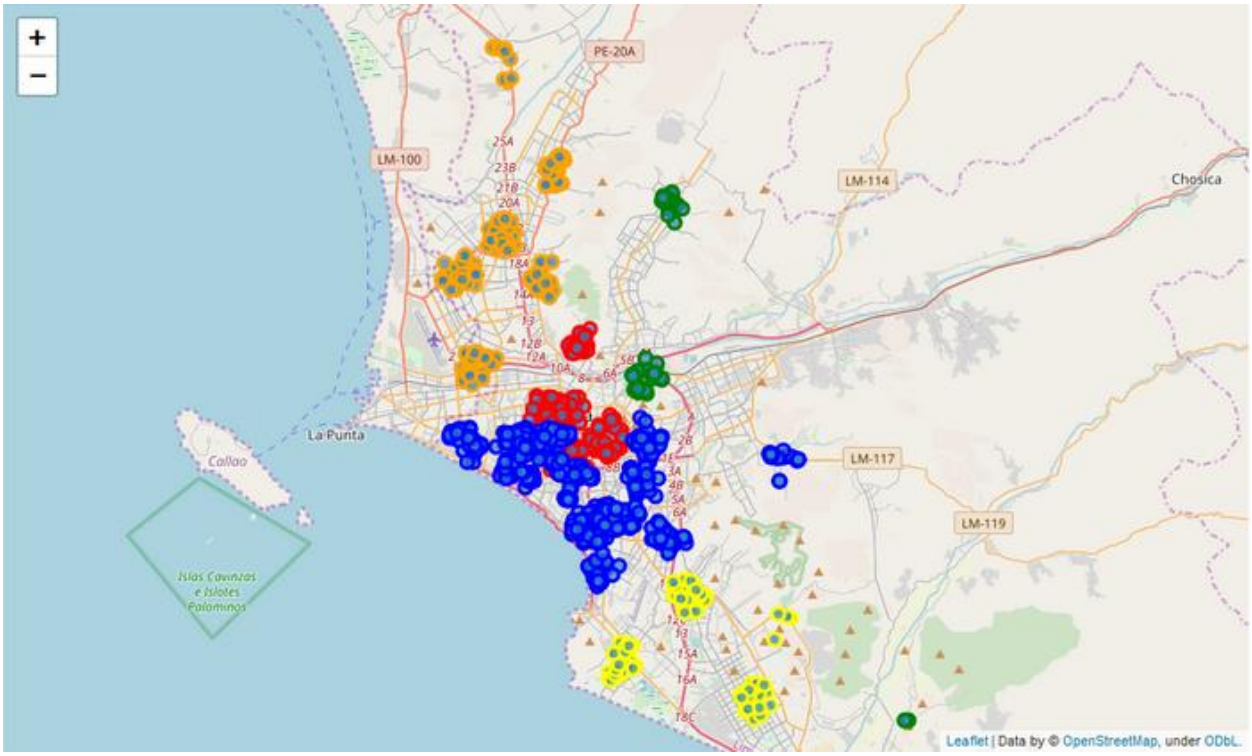
Data scraped from these two URL sources will be useful in order to generate a main dataframe. The process started defining a "Lima" prefix in each district's name to get coordinates information from the Wikipedia Lima Web Article's link. Later, we use Foursquare API to get venues information (such as latitude, longitude, venue name, ID, and so on) only for the "Restaurant" category. In the meantime, the Wikipedia Lima Districts Chart's link will be used to get the segmentation of each district according to these five areas or sectors. Finally, we joint this two information in one general dataframe called "*main_df*".

2.3. Feature Selection

After data cleaning, the main dataframe looks like the following picture, including all the features:

	Localidad	Localidad Latitude	Localidad Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue ID	Sector	sector_color
166	Lima, Rímac	-12.0203	-77.03546	Manka Mitu (Olla De Barro)	-12.02367	-77.03249	Restaurant	5a89d47f178a2a50af879402	Centro de Lima	red
167	Lima, Rímac	-12.0203	-77.03546	Purito Catacaos	-12.02434	-77.03626	Seafood Restaurant	527697e911d2f813ded3b722	Centro de Lima	red
168	Lima, Rímac	-12.0203	-77.03546	Chifa Hoo Wa	-12.02666	-77.03402	Chinese Restaurant	4c6842cd73851b8d4721ba39	Centro de Lima	red
169	Lima, Rímac	-12.0203	-77.03546	Chifa Yon Fu	-12.02733	-77.03333	Cantonese Restaurant	5691d1d938faabbd111febb0	Centro de Lima	red
170	Lima, Rímac	-12.0203	-77.03546	El Rinconcito	-12.02807	-77.03404	Seafood Restaurant	51d5c205498ec498fc1e7a9c	Centro de Lima	red
171	Lima, Rímac	-12.0203	-77.03546	El Loco Pez	-12.02324	-77.03147	Seafood Restaurant	5183eeea498ef05a2d4668e7	Centro de Lima	red
172	Lima, Rímac	-12.0203	-77.03546	Restaurant de Marina y Jcarlos	-12.02623	-77.03559	Peruvian Restaurant	51d1c81b498e991727206311	Centro de Lima	red

Using the Folium Map, we can visualize each restaurant venue by sector as follows:



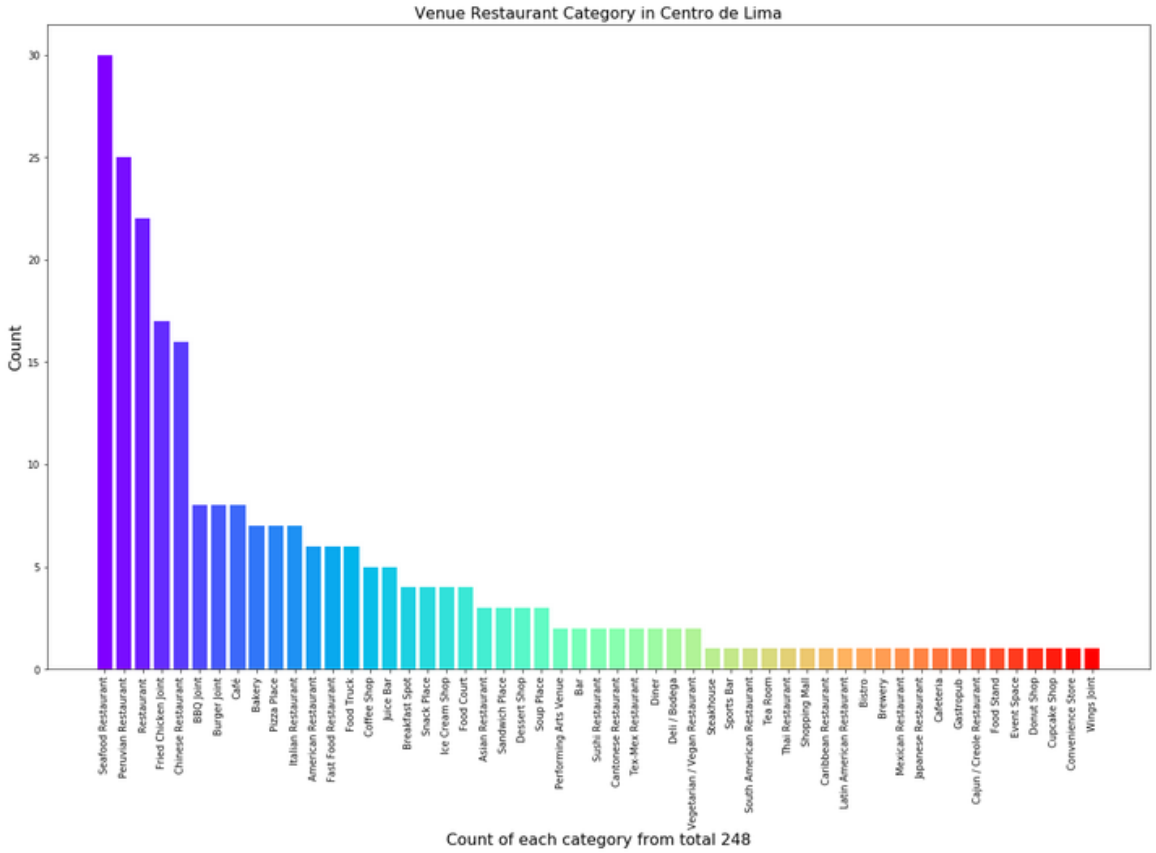
Where:

- North: orange.
- Center (Mid): red.
- East: green.
- Modern (residential): blue
- South: yellow

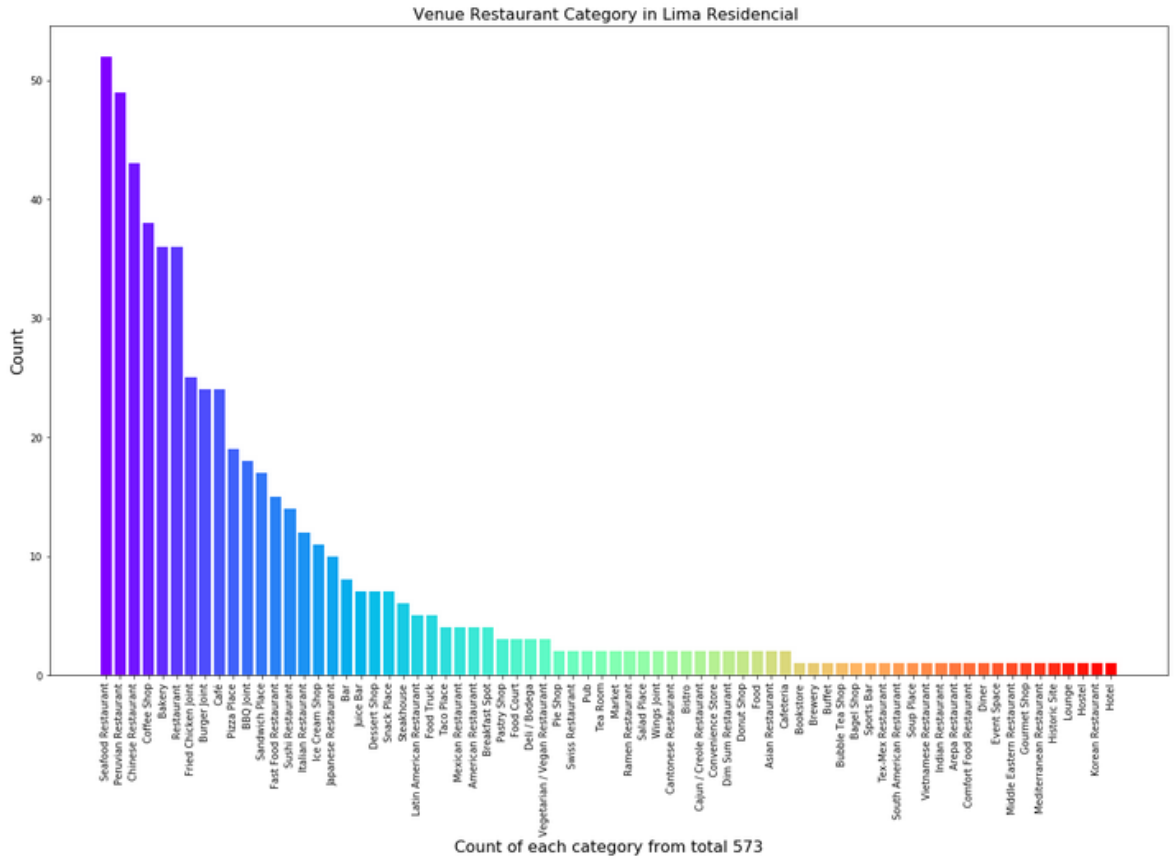
3. METHODOLOGY

3.1. Which type of restaurant is better by sector and in the entire Lima city?

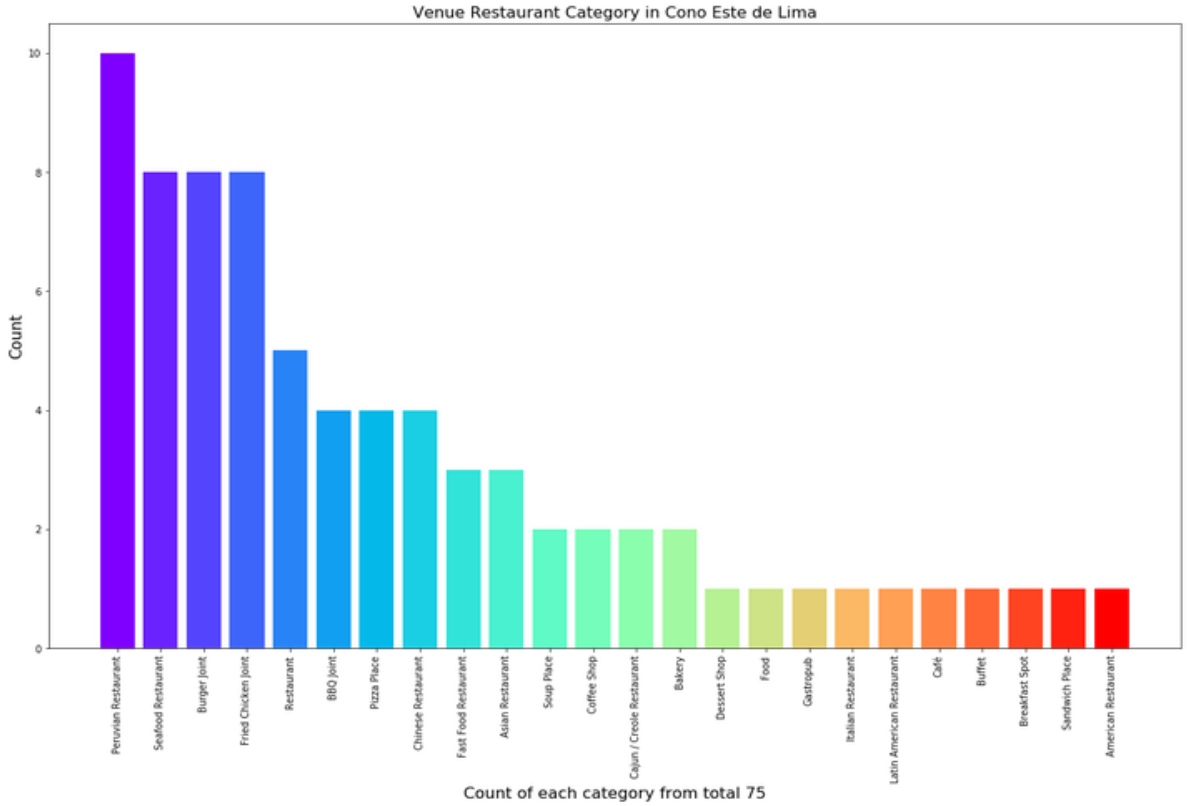
Our first goal is to find which type of restaurant is more popular in each Lima sector (and also in the whole city). In order to do this, we only count each of them by sector. Later, we can get two different results: a bar plot and a representative percentage of the most common restaurant type based on the total restaurants in that sector. In that sense, we get:



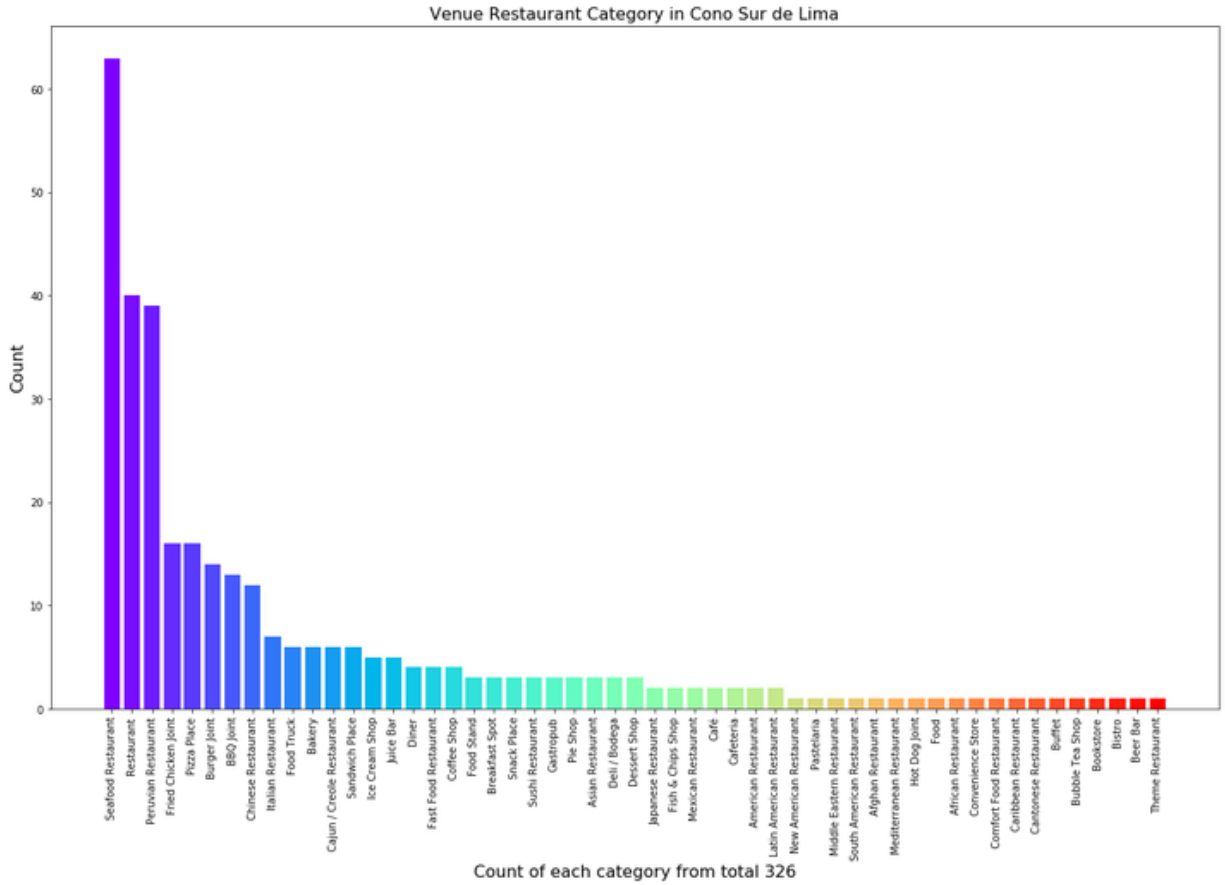

```
General Info Top 1 venues summary:
                                        cant
Sector         Venue Category
Centro de Lima Seafood Restaurant  12.10% | Total: 248
********************************************************************************************************
```



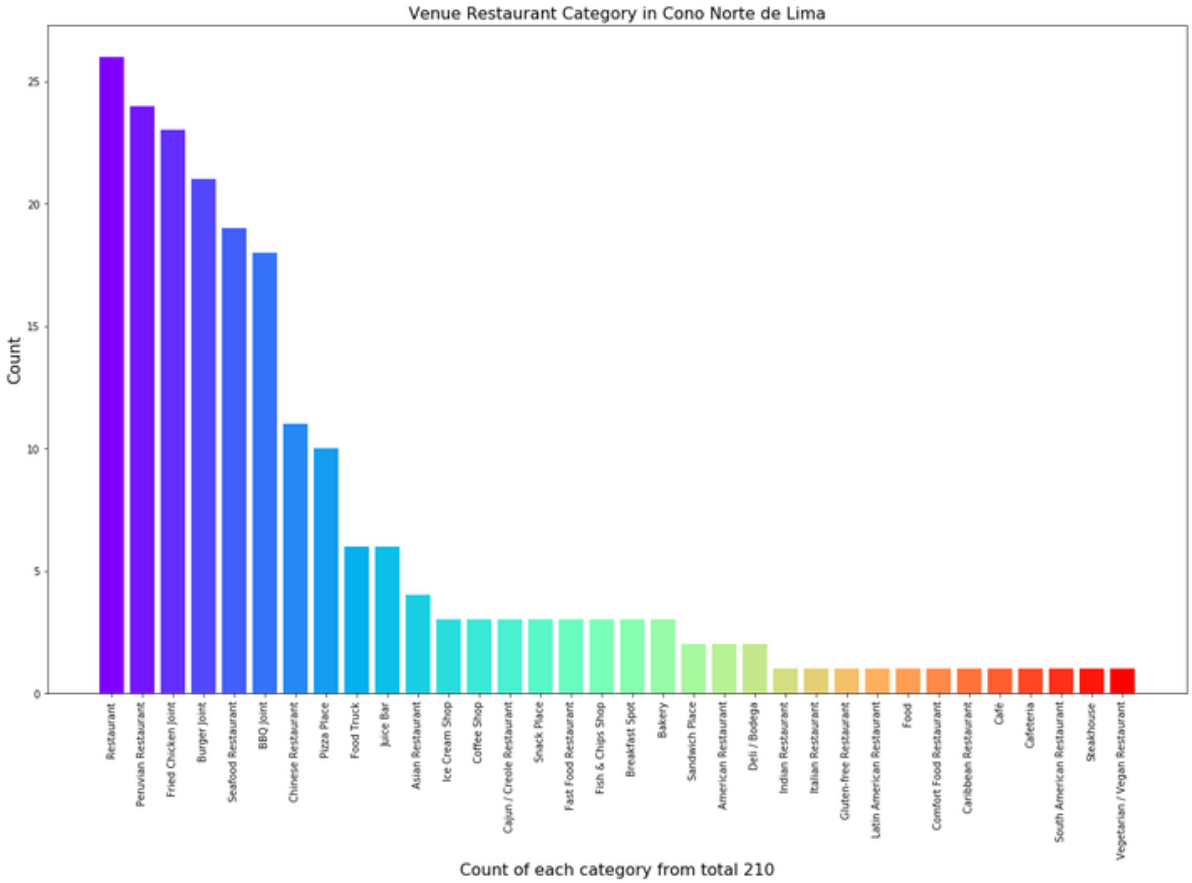

```
General Info Top 1 venues summary:
                                    cant
Sector            Venue Category
Lima Residencial  Seafood Restaurant  9.08% | Total: 573
**********************************************************************************************************
```




```
General Info Top 1 venues summary:
                                    cant
Sector            Venue Category
Cono Este de Lima Peruvian Restaurant  13.33% | Total: 75
**********************************************************************************************************
```




```
General Info Top 1 venues summary:
                                        cant
Sector           Venue Category
Cono Sur de Lima Seafood Restaurant  19.33% | Total: 326
**********************************************************************************************************
```




```
General Info Top 1 venues summary:
                                      cant
Sector             Venue Category
Cono Norte de Lima Restaurant      12.38% | Total: 210
**********************************************************************************************************
```

Thus, choosing the top 1 type restaurant venues by sector, we can get:

	Sector	Venue Category
0	Centro de Lima	Seafood Restaurant
1	Lima Residencial	Seafood Restaurant
2	Cono Este de Lima	Peruvian Restaurant
3	Cono Norte de Lima	Restaurant
4	Cono Sur de Lima	Seafood Restaurant

3.2. Are the restaurant venues in the south more like the north than the center to the east?

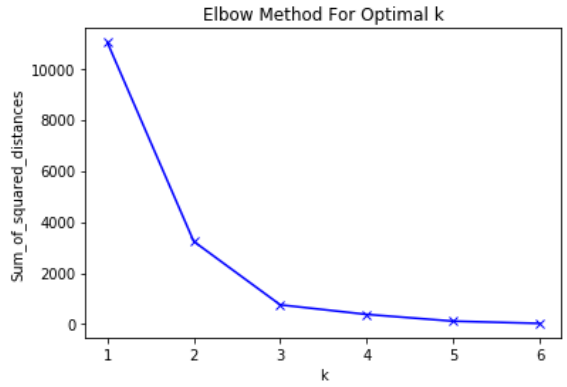
Our second goal is to cluster each of this type of restaurants using an unsupervised machine learning classification model. Why is that? Basically, because we want to get some differences and similarities among the restaurant types without any information about Lima sectors. This is relevant in the sense we can get some venues that are similar to each other independently whether they belong to one sector or not. What if one restaurant from the north is pretty similar to another in the east, given some information data? This is the class of conclusion that we want to achieve.

In order to do this, we use the Foursquare API to get additional information –specifically, we get rating and tips information for each type of restaurant. However, there are some limitations here. Foursquare gives only 50 maximum calls to get this info, thus unfortunately we must select randomly some samples from our restaurant dataframe by type and sector. We select only 5 restaurants by sector, which makes a total of 25 restaurants venues in each Lima zone. Considering that we are going to get 2 data about each restaurant (rating and tips), we achieve the maximum 50 calls allowed.

Using the Foursquare API to get ratings and tips, we get the following dataframe. Note that in some sectors there are less than five restaurants selected. This is because not all the restaurants have ratings *or* tips. We delete the venues without this information in order to avoid give unnecessary conclusions:

		Localidad	Localidad Latitude	Localidad Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue ID	Distrito	Sector	sector_color	rating	tips
Sector														
Centro de Lima	**519**	Lima, Lima	-12.06211	-77.03653	PuertoNortePe	-12.05775	-77.03855	Seafood Restaurant	4daf32884df00ee01d34c00f	Lima	Centro de Lima	red	7.3	52
	1015	Lima, Lince	-12.08657	-77.03665	El Merlín de Cabo Blanco	-12.08807	-77.04715	Seafood Restaurant	4c740c7cff1fb60c5469f2a7	Lince	Centro de Lima	red	7.8	113
	1024	Lima, Lince	-12.08657	-77.03665	Punto Azul	-12.09212	-77.03253	Seafood Restaurant	4ba3d856f964a520d86538e3	Lince	Centro de Lima	red	7.7	42
Cono Este de Lima	**426**	Lima, San Juan de Lurigancho	-11.94875	-76.97791	El Rinconcito de la Panchita	-11.94427	-76.98196	Peruvian Restaurant	50c3867ce4b0e77d29fa8312	San Juan de Lurigancho	Cono Este de Lima	green	0	3
	468	Lima, Santa Anita	-12.22338	-76.84771	Las Leñas	-12.22381	-76.85379	Peruvian Restaurant	4de0094145dd3eae8763f5a1	Santa Anita	Cono Este de Lima	green	6	17
Cono Sur de Lima	**1533**	Lima, San Bartolo	-12.38707	-76.77795	Restaurante Chelulo	-12.38923	-76.77738	Seafood Restaurant	4d6abcf4a27f236aeb5c9666	San Bartolo	Cono Sur de Lima	yellow	5.7	16
	289	Lima, Pucusana	-12.48209	-76.79745	Restaurant Jhony	-12.481	-76.79829	Seafood Restaurant	4ec955356c251306cda4ac3b	Pucusana	Cono Sur de Lima	yellow	0	2
	858	Lima, San Juan de Miraflores	-12.15585	-76.97213	Cevicheria "Dorita"	-12.15469	-76.96853	Seafood Restaurant	50739334c84c42b90bc5beef	San Juan de Miraflores	Cono Sur de Lima	yellow	0	1
Lima Residencial	**134**	Lima, San Miguel	-12.07866	-77.09528	Mi Propiedad Privada	-12.08707	-77.09292	Seafood Restaurant	4bfd89a5b68d0f4709eee857	San Miguel	Lima Residencial	blue	8.7	39
	1227	Lima, San Luis	-12.07236	-76.99589	Mi Cebi-chela	-12.08322	-76.99313	Seafood Restaurant	4d32166db60937040627f3df	San Luis	Lima Residencial	blue	6.5	14
	1235	Lima, San Luis	-12.07236	-76.99589	Los Cantaritos	-12.08189	-76.99329	Seafood Restaurant	4e6d1a3be4cd4bedebeb02b2	San Luis	Lima Residencial	blue	7.2	29
	557	Lima, Barranco	-12.14396	-77.02027	Chanfainita Pale	-12.14679	-77.01442	Seafood Restaurant	4d5d72ee775f8cfac257c2e0	Barranco	Lima Residencial	blue	8.2	28
	1507	Lima, Pueblo Libre	-12.07664	-77.06786	Don Pedrito	-12.08731	-77.07251	Seafood Restaurant	51d851f2498eb4a6d5b55e0e	Pueblo Libre	Lima Residencial	blue	7.2	11

Whit this dataframe, we can apply Kmeans algorithm to find the right number of clusters. We will use only ratings and tips data to do this. Based on the Elbow Method, we compute the right number of clusters assessing:

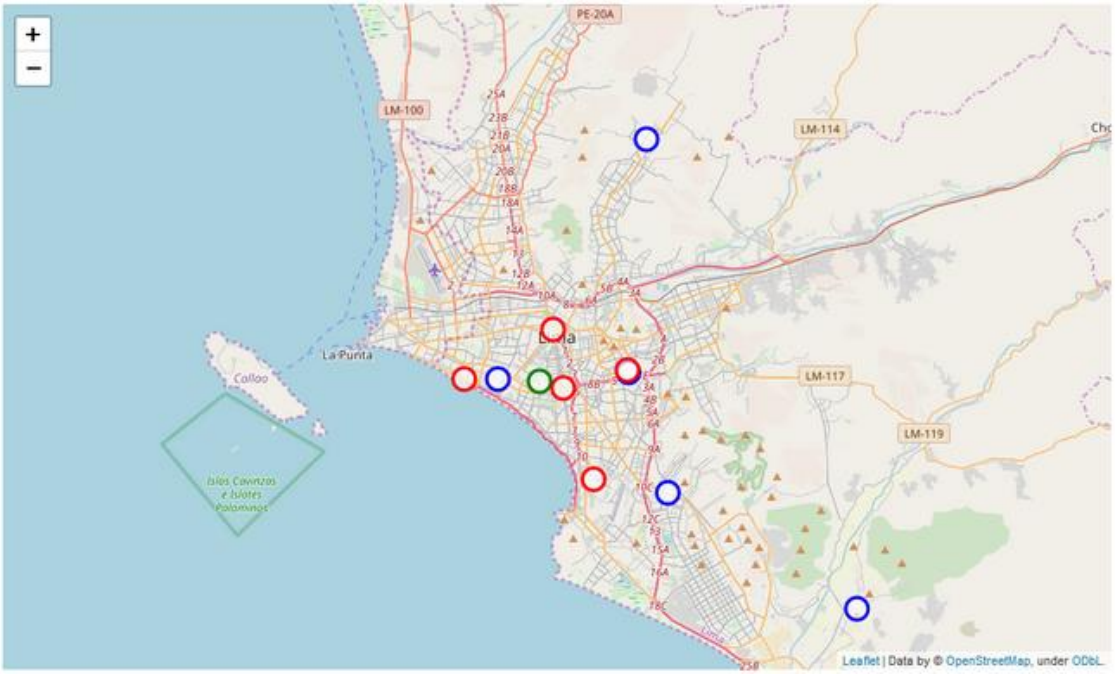


Ideal number of clusters: *3*

Assigning the labels to each venue, we get our modified dataframe:

		Localidad	Localidad Latitude	Localidad Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue ID	Distrito	Sector	sector_color	rating	tips	Labels	label_color
Sector																
Centro de Lima	**519**	Lima, Lima	-12.06211	-77.03653	PuertoNortePe	-12.05775	-77.03855	Seafood Restaurant	4daf32884df00ee01d34c	Lima	Centro de Lima	red	7.3	52	0	red
	1015	Lima, Lince	-12.08657	-77.03665	El Merlín de Cabo Blanco	-12.08807	-77.04715	Seafood Restaurant	4c740c7cff1fb60c5469f2a	Lince	Centro de Lima	red	7.8	113	1	green
	1024	Lima, Lince	-12.08657	-77.03665	Punto Azul	-12.09212	-77.03253	Seafood Restaurant	4ba3d856f964a520d86538e3	Lince	Centro de Lima	red	7.7	42	0	red
Cono Este de Lima	**426**	Lima, San Juan de Lurigancho	-11.94875	-76.97791	El Rinconcito de la Panchita	-11.94427	-76.98196	Peruvian Restaurant	50c3867ce4b0e77d29fa8312	San Juan de Lurigancho	Cono Este de Lima	green	0	3	2	blue
	468	Lima, Santa Anita	-12.22338	-76.84771	Las Leñas	-12.22381	-76.85379	Peruvian Restaurant	4de0094145dd3eae8763f5a1	Santa Anita	Cono Este de Lima	green	6	17	2	blue
Cono Sur de Lima	**1533**	Lima, San Bartolo	-12.38707	-76.77795	Restaurante Chelulo	-12.38923	-76.77738	Seafood Restaurant	4d6abcf4a27f236aeb5c9	San Bartolo	Cono Sur de Lima	yellow	5.7	16	2	blue
	289	Lima, Pucusana	-12.48209	-76.79745	Restaurant Jhony	-12.481	-76.79829	Seafood Restaurant	4ec955356c251306cda4a	Pucusana	Cono Sur de Lima	yellow	0	2	2	blue
	858	Lima, San Juan de Miraflores	-12.15585	-76.97213	Cevicheria "Dorita"	-12.15469	-76.96853	Seafood Restaurant	50739334c84c42b90bc5beef	San Juan de Miraflores	Cono Sur de Lima	yellow	0	1	2	blue
Lima Residencial	**134**	Lima, San Miguel	-12.07866	-77.09528	Mi Propiedad Privada	-12.08707	-77.09292	Seafood Restaurant	4bfd89a5b68d0f4709eee857	San Miguel	Lima Residencial	blue	8.7	39	0	red
	1227	Lima, San Luis	-12.07236	-76.99589	Mi Cebichela	-12.08322	-76.99313	Seafood Restaurant	4d32166db60937040627f3df	San Luis	Lima Residencial	blue	6.5	14	2	blue
	1235	Lima, San Luis	-12.07236	-76.99589	Los Cantaritos	-12.08189	-76.99329	Seafood Restaurant	4e6d1a3be4cd4bedebeb02b2	San Luis	Lima Residencial	blue	7.2	29	0	red
	557	Lima, Barranco	-12.14396	-77.02027	Chanfainita Pale	-12.14679	-77.01442	Seafood Restaurant	4d5d72ee775f8cfac257c2e0	Barranco	Lima Residencial	blue	8.2	28	0	red
	1507	Lima, Pueblo Libre	-12.07664	-77.06786	Don Pedrito	-12.08731	-77.07251	Seafood Restaurant	51d851f2498eb4a6d5b55e0e	Pueblo Libre	Lima Residencial	blue	7.2	11	2	blue

Finally, using the Folium map, we can identify this clusters in Lima:



4. DISCUSSION

As I mentioned before, Lima is a city divided in five areas. We are not including the districts belong to Callao, which is recognized legally as a province. In that sense, Lima is a multicultural city with many features. That is why a study only based on the restaurant venues information might not be a well-structured deep research in order to identify the relationships between differences and similarities of the city by sector, even if the target is focusing only in this sort of venues. Other variables, such as population, economic wealth of citizens, transportation, etc., should be taken into consideration as well. However, this could be an initial point to draw an overview of the city.

It is important to understand the limitations in the application of our Kmeans classification model. In that sense, we use the Elbow method to get the optimum K value, which result in number 3. However, we are only use perhaps a non-representative sample of 5 restaurants by sector because of the maximum calls allowed by Foursquare. Moreover, in that sample of 25 venues, not all of the them were used –some do not have information about rating or tips. Thus, is obvious that, in a deep research, this optimum K value might not be the same. In a real scenario, and based on the previous comment, we must use more restaurant type venues and more variables in order to get a better understanding.

About the data, it is necessary to comprehend that the type "Restaurant" and "Peruvian Restaurant" might include seafood cuisine too. However, although this could reinforce the idea that Seafood Restaurant might be consider as one of the main cuisines in this city, we are not going to use these two categories as a part of our conclusion.

As a final detail, we performed a data analysis through the district information by adding their coordinates and venues information as a general data and visualizing in using a Folium Map. In a deep research, these data can also be used in a dynamically way from some platforms or packages automatically.

5. CONCLUSIONS

Based on the results, we can draw two conclusions.

5.1. Based on the results in 3.1., we can conclude that "Seafood Restaurant" is the most common type of restaurant in Lima, Peru –the relevance of its importance is different among districts, being 12.10% of the total restaurants in mid-town Lima (Centro de Lima), 9.08% in modern Lima (Lima Residencial) and 19.33% in Lima south (Cono Sur). Perhaps, this represent an important supply and demand relationship of this sort of food in the citizens of Lima.

5.2. Based on the results in 3.2., we can conclude effectively that there are some similarities and divergences among restaurants in the same or in a different sector. As we can see in the last Folium map, there are some restaurants that, in the same area, belongs to another label. In the same way, we can identify some venues that are part of the same cluster although they fit in a different sector of the city.