Capstone Project Report

Identifying similar districts in different Peruvian provinces

Arturo J. Miguel de Priego

1. Introduction

Peru is a very varied country in culture and commerce. The objective of this project is to determine similarities between Peruvian districts of both the same province and between districts of different provinces. For this, **FourSquare** data will be used as an indicator of commercial similarity to obtain clusters of different degrees of grouping, and compare the results with classifications made by national organizations such as INEI.

We will use the codes provided in the laboratories of the course to create functions for the purpose of modularizing the programming tasks. The results of this project could be used to develop and implement government programs adapted to regions with similar needs.

2. Data

We take Peruvian population from the *Instituto Nacional de Estadística e Informática (INEI)*, https://www.inei.gob.pe/media/MenuRecursivo/indices_tematicos/cuadro001_1.xls. Data for 2015 was used. The name of the boroughs (districts) are combined with their respective province name and country name (Peru in this case) to obtain the coordinates using **geopy**. These coordinates are used to make venue requests in FourSquare. The number of venues retrieved for each borough is used as an indicator of borough prosperity. Four provinces were selected to analyze its boroughs data: Arequipa, Chincha, Cusco and Puno.

3. Methodology

First, exploratory data analysis was made with the help of a Python notebook. An interactive routine was developed to select the Peruvian province containing the boroughs. Then, data was analyzed to compare correlations between populations and number of venues.

3.1 Getting data

The python notebook is located in https://github.com/ArturoMigueldePriego/Data-Science-with-IBM/blob/master/CapstoneProject.ipynb. First, Peruvian population data was retrieved from INEI. Cleansing of data included removing NaN rows and changing original names in order to find coordinates with **geopy**.

Next, the data was used to get boroughs coordinates. Delays were inserted as the geopy method stopped after a number de consecutive calls. When results were null, the search string was added to a list for later manual retrieving.

Then, the following issues were resolved:

- Internet connection was slow when looking for data that already exists. To get the data another request was executed.
- Some names were different, for example HUANCA SANCOS is the real name but HUANCASANCOS was registered in geopy database.
- A name was written without a final O in the INEI file.

- In two cases NASCA must be used instead of NAZCA.
- A name was abbreviated in the INEI file. A complete name was used in the request.
- An address containing the word VEINTISIETE (twenty seven) was replaced with the number 27.
- A large name was replaced by a shorter name.
- Also, a misleading address was detected when getting locations for Chincha province. It was fixed with the right address.

All changes were edited manually and saved to the **boroughs_fixed.csv** file. After cleansing the data, a interactive routine was developed to get venues and build their clusters. A lot of code was used from the third lab of the course. In figure 1 the user interface for selecting boroughs is showed.

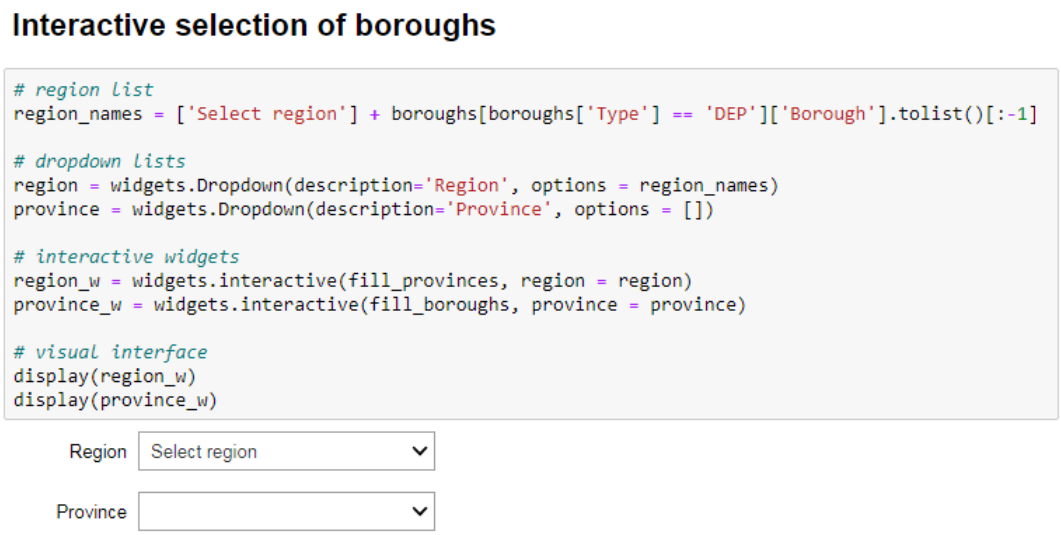

Figure 1. User interface to select boroughs.

When a user selects a borough the application retrieve data for venues and execute the cluster routines as shown in figure 2. Retrieved data is saved for future requests, as shown in figure 3.

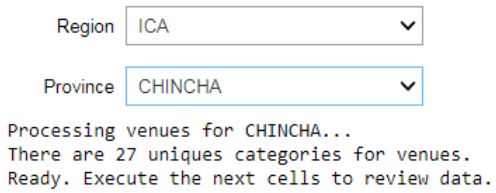

Figure 2. Selecting a borough the first time.

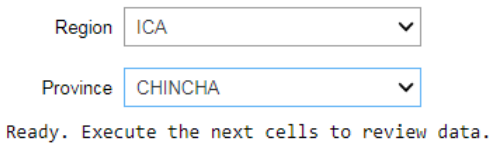

Figure 3. Selecting a borough next time.

Next, tables and maps are visualized by executing the next cells. For example, **province_df** provides the main data for the boroughs of the selected province (figure 4). Boroughs without venues are also showed.

```
province_df
```

	Borough	Population	Latitude	Longitude	Venues	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
1	CHINCHA ALTA	63671.0	-13.417488	-76.132573	15	Plaza: 13.33 %	Ice Cream Shop: 13.33 %	BBQ Joint: 6.67 %	Furniture / Home Store: 6.67 %	Shopping Mall: 6.67 %
2	ALTO LARAN	7387.0	-13.442310	-76.082833	0					
3	CHAVIN	1417.0	-13.077080	-75.912989	0					
4	CHINCHA BAJA	12323.0	-13.459021	-76.161693	5	Beach: 20.0 %	Tourist Information Center: 20.0 %	Grocery Store: 20.0 %	History Museum: 20.0 %	Park: 20.0 %
5	EL CARMEN	13296.0	-13.499838	-76.057515	4	Campground: 25.0 %	Plaza: 25.0 %	Peruvian Restaurant: 25.0 %	Bar: 25.0 %	BBQ Joint: 0.0 %
6	GROCIO PRADO	24049.0	-13.398126	-76.156446	4	BBQ Joint: 25.0 %	Sandwich Place: 25.0 %	Park: 25.0 %	Neighborhood: 25.0 %	Ice Cream Shop: 0.0 %
7	PUEBLO NUEVO	61078.0	-13.404673	-76.127199	6	Plaza: 33.33 %	Business Service: 16.67 %	Soccer Stadium: 16.67 %	Seafood Restaurant: 16.67 %	Furniture / Home Store: 16.67 %
8	SAN JUAN DE YANAC	316.0	-13.210952	-75.786875	0					
9	SAN PEDRO DE HUACARPANA	1660.0	-13.048914	-75.647898	0					
10	SUNAMPE	27496.0	-13.427280	-76.164259	4	Fried Chicken Joint: 75.0 %	Restaurant: 25.0 %	BBQ Joint: 0.0 %	Ice Cream Shop: 0.0 %	Tourist Information Center: 0.0 %
11	TAMBO DE MORA	4990.0	-13.458413	-76.182643	1	Plaza: 100.0 %	BBQ Joint: 0.0 %	Ice Cream Shop: 0.0 %	Tourist Information Center: 0.0 %	Soccer Stadium: 0.0 %

Figure 4. Boroughs and venues summary

The map of borough coordinates is saved in **borough_map** (figure 5).

```
borough_map
```

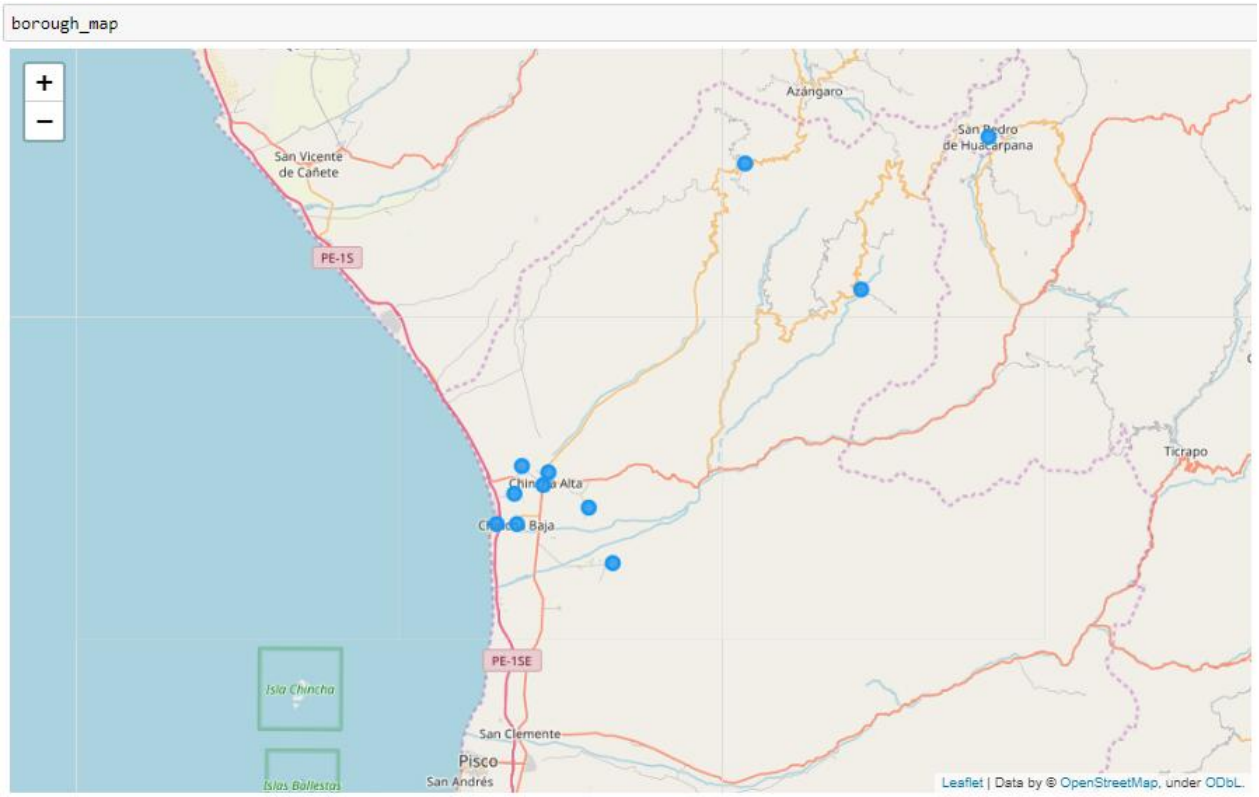

Figure 5. Boroughs map.

All venues in each borough are recorded in **venues_df** (figure 6).

```
venues_df
```

	Borough	Latitude	Longitude	Venue	Latitude	Longitude	Category
0	CHINCHA ALTA	-13.417488	-76.132573	Raspadilla Zambo	-13.420301	-76.130650	Ice Cream Shop
1	CHINCHA ALTA	-13.417488	-76.132573	Hotel Princess	-13.418101	-76.133221	Hotel
2	CHINCHA ALTA	-13.417488	-76.132573	Bodega Tabernero	-13.418227	-76.139225	Winery
3	CHINCHA ALTA	-13.417488	-76.132573	Mercado De Chincha	-13.416695	-76.136664	Market
4	CHINCHA ALTA	-13.417488	-76.132573	Chifa Continental	-13.420677	-76.133951	Chinese Restaurant
5	CHINCHA ALTA	-13.417488	-76.132573	Plaza de Armas	-13.417555	-76.132638	Plaza
6	CHINCHA ALTA	-13.417488	-76.132573	Sodimac Chincha	-13.413394	-76.129183	Furniture / Home Store
7	CHINCHA ALTA	-13.417488	-76.132573	Tottus Chincha	-13.419022	-76.135901	Shopping Mall
8	CHINCHA ALTA	-13.417488	-76.132573	Plaza de armas de Chincha	-13.417470	-76.132615	Plaza
9	CHINCHA ALTA	-13.417488	-76.132573	Heladeria Don Giusseppi	-13.418131	-76.132502	Ice Cream Shop
10	CHINCHA ALTA	-13.417488	-76.132573	House gym chincha	-13.420436	-76.134637	Gym / Fitness Center
11	CHINCHA ALTA	-13.417488	-76.132573	plazaVea	-13.416401	-76.141662	Grocery Store
12	CHINCHA ALTA	-13.417488	-76.132573	Norky's	-13.415924	-76.141622	BBQ Joint
13	CHINCHA ALTA	-13.417488	-76.132573	Virgen del Carmen Inversión Textil SAC - Vircatex	-13.424407	-76.130118	Clothing Store
14	CHINCHA ALTA	-13.417488	-76.132573	Maestro Chincha	-13.424799	-76.137700	Department Store
15	CHINCHA BAJA	-13.459021	-76.161693	Chincha Baja	-13.458778	-76.161490	Park

Figure 6. Venues details.

The **map_with_venues** method is useful to show the relative number of venues. Boroughs without venues are marked in red (figure 7).

```
map_with_venues(province_df)
```

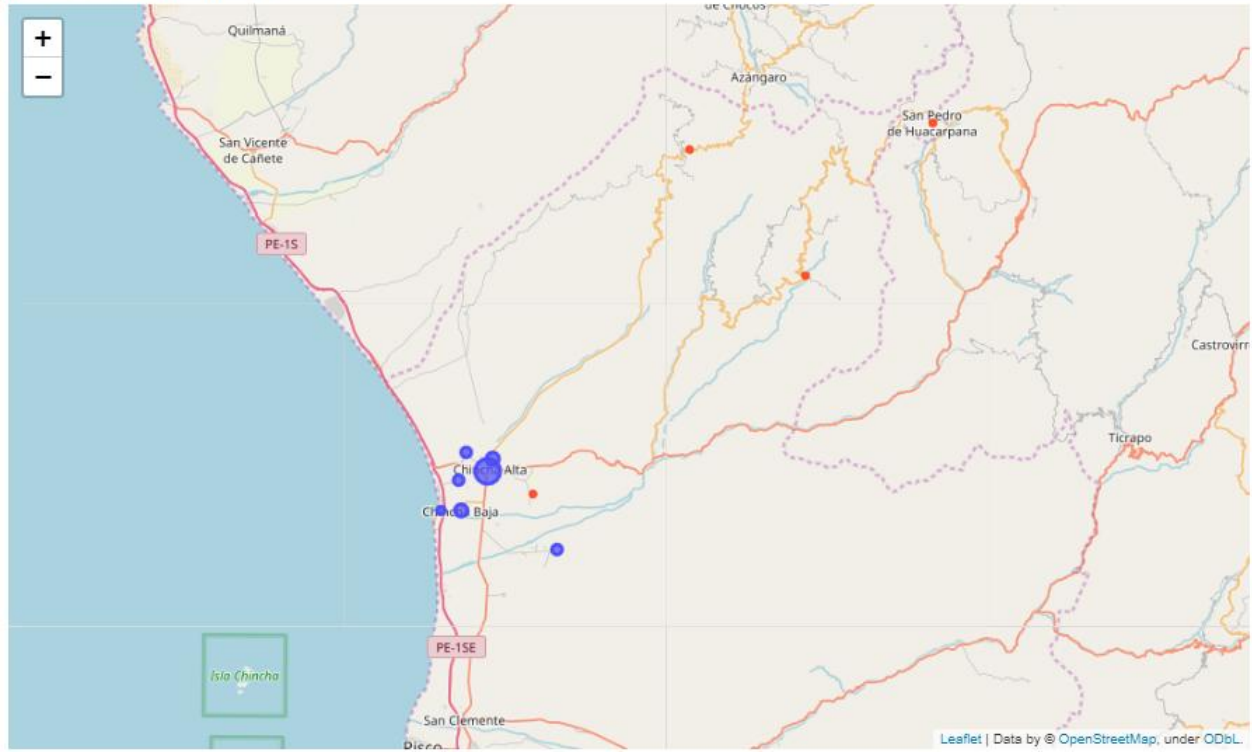

Figure 7. Visual representation of number of venues in each borough.

The borough clustering is made it by the **get_borough_data** method. It returns a summary of boroughs with venues (figure 8) and a cluster map (figure 9). The number of top venues and the number of cluster can be adjusted.

```
boroughs_with_venues_df, boroughs_with_venues_map = get_borough_data(5, 5)
```

```
boroughs_with_venues_df
```

	Borough	Population	Latitude	Longitude	Venues	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Labels	1st Most Common Venue	2nd Most Common Venue	3 C
1	CHINCHA ALTA	63671.0	-13.417488	-76.132573	15	Plaza: 13.33 %	Ice Cream Shop: 13.33 %	BBQ Joint: 6.67 %	Furniture / Home Store: 6.67 %	Shopping Mall: 6.67 %	1	Plaza	Ice Cream Shop	
4	CHINCHA BAJA	12323.0	-13.459021	-76.161693	5	Beach: 20.0 %	Tourist Information Center: 20.0 %	Grocery Store: 20.0 %	History Museum: 20.0 %	Park: 20.0 %	0	Beach	Park	
5	EL CARMEN	13296.0	-13.499838	-76.057515	4	Campground: 25.0 %	Plaza: 25.0 %	Peruvian Restaurant: 25.0 %	Bar: 25.0 %	BBQ Joint: 0.0 %	1	Bar	Campground	
6	GROCIO PRADO	24049.0	-13.398126	-76.156446	4	BBQ Joint: 25.0 %	Sandwich Place: 25.0 %	Park: 25.0 %	Neighborhood: 25.0 %	Ice Cream Shop: 0.0 %	4	BBQ Joint	Sandwich Place	
7	PUEBLO NUEVO	61078.0	-13.404673	-76.127199	6	Plaza: 33.33 %	Business Service: 16.67 %	Soccer Stadium: 16.67 %	Seafood Restaurant: 16.67 %	Furniture / Home Store: 16.67 %	1	Plaza	Soccer Stadium	Re
10	SUNAMPE	27496.0	-13.427280	-76.164259	4	Fried Chicken Joint: 75.0 %	Restaurant: 25.0 %	BBQ Joint: 0.0 %	Ice Cream Shop: 0.0 %	Tourist Information Center: 0.0 %	3	Fried Chicken Joint	Restaurant	
11	TAMBO DE MORA	4990.0	-13.458413	-76.182643	1	Plaza: 100.0 %	BBQ Joint: 0.0 %	Ice Cream Shop: 0.0 %	Tourist Information Center: 0.0 %	Soccer Stadium: 0.0 %	2	Plaza	Winery	

Figure 8. Visual representation of number of venues in each borough.

```
boroughs_with_venues_map
```

Figure 9. Cluster map.

3.1 Analyzing data

For data analysis, the scatter of population and number of venues in each borough is plotted. The correlation coefficient is inserted in the title of the plot (figure 10).

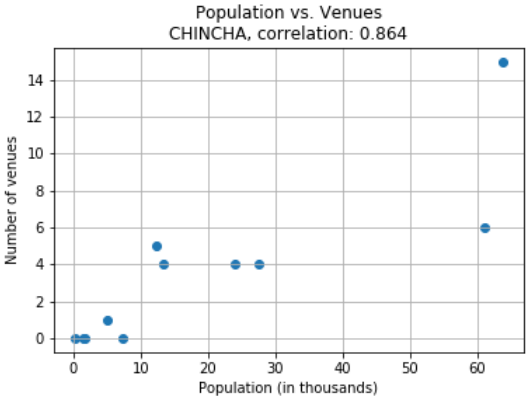

Figure 10. Scatter plot and correlation coefficient for the boroughs of Chincha province.

4. Results

Four provinces were selected to analyze its boroughs data: Arequipa, Chincha, Cusco and Puno. From figure 11, the correlation coefficients indicate some interesting results that must be analyzed considering the characteristics of each borough.

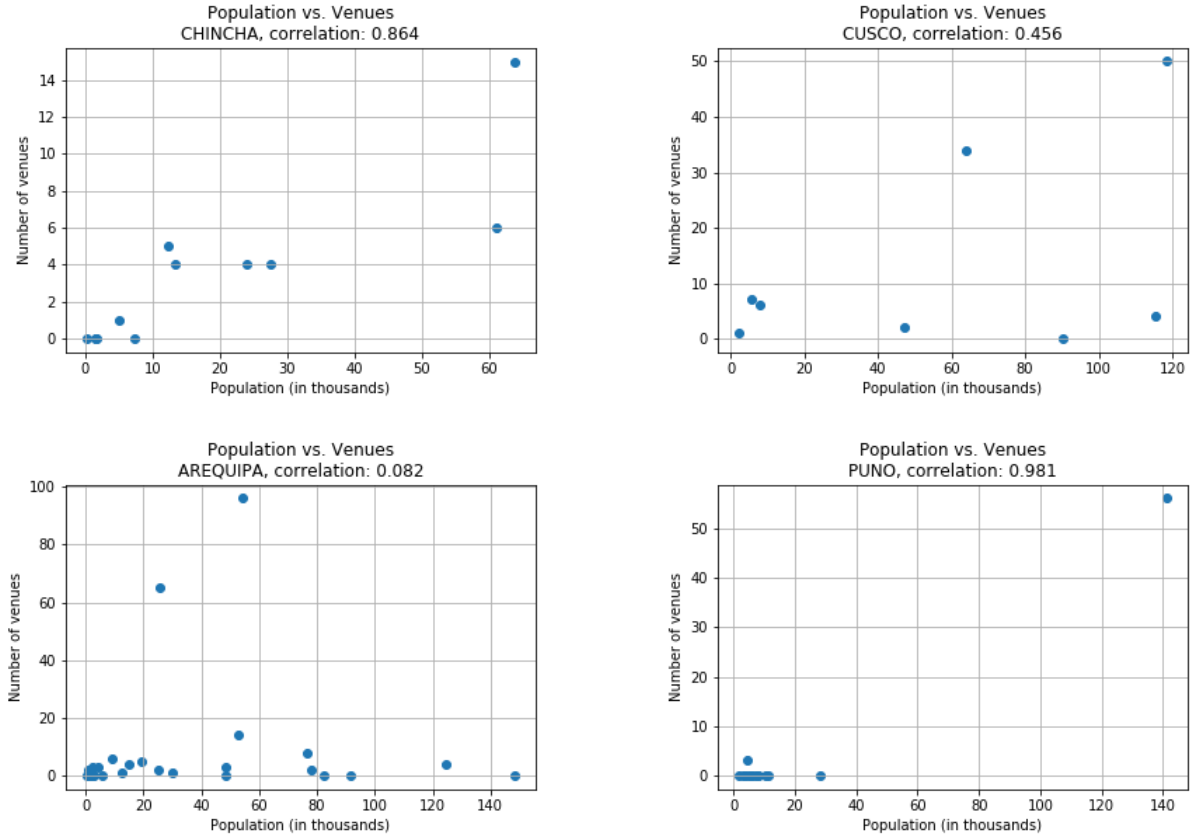

Figure 11. Scatter plots and correlations between populations and number of venues by boroughs

The result for Chincha shows a strong correlation. It means, as a borough has more population then the city has more commerce. This is observed in real life. Although Puno shows a very strong correlation, the data is insufficient to conclude interesting results. This is because few venues were retrieved for only two cities. The other cities have no venues reported.

It is interesting Arequipa shows uncorrelated data, maybe because of their number of boroughs. Cusco shows a weak correlation, as two boroughs concentrate the main part of the venues reported.

5. Discussion

The relative low number of venues reported by FourSquare makes difficult find strong evidence for correlations between population and commerce activities. Many boroughs do not have venue data from FourSquare. In Peru, the telecommunications still are in development in many cities and towns, and reporting venues is also not a common activity.

6. Conclusions

- The main conclusion from this work was the great learning in developing Python programs to retrieve, clean, process, and visualize data.
- With the tool developed, more work can be made to get more borough data and compare many boroughs at macro-regional, regional and country levels.