Capstone Project Report

Identifying similar districts in different Peruvian provinces

Arturo J. Miguel de Priego

Presentation

- This project presents a solution for getting data from Peruvian boroughs (districts) in order to compare venues of both the same province and between districts of different provinces.
- FourSquare data is used as an indicator of commercial similarity to obtain clusters of different degrees of grouping.
- The results of this project could be used to develop and implement government programs adapted to regions with similar needs.

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.
- Four provinces were selected to analyze its boroughs data: Arequipa, Chincha, Cusco and Puno.

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.

Interactive selection of boroughs

```
# region list
region_names = ['Select region'] + boroughs[boroughs['Type'] == 'DEP']['Borough'].tolist()[:-1]

# dropdown lists
region = widgets.Dropdown(description='Region', options = region_names)
province = widgets.Dropdown(description='Province', options = [])

# interactive widgets
region_w = widgets.interactive(fill_provinces, region = region)
province_w = widgets.interactive(fill_boroughs, province = province)

# visual interface
display(region_w)
display(province_w)
```

Region: Select region

Province:

Region: ICA

Province: CHINCHA

```
Processing venues for CHINCHA...
There are 27 uniques categories for venues.
Ready. Execute the next cells to review data.
```

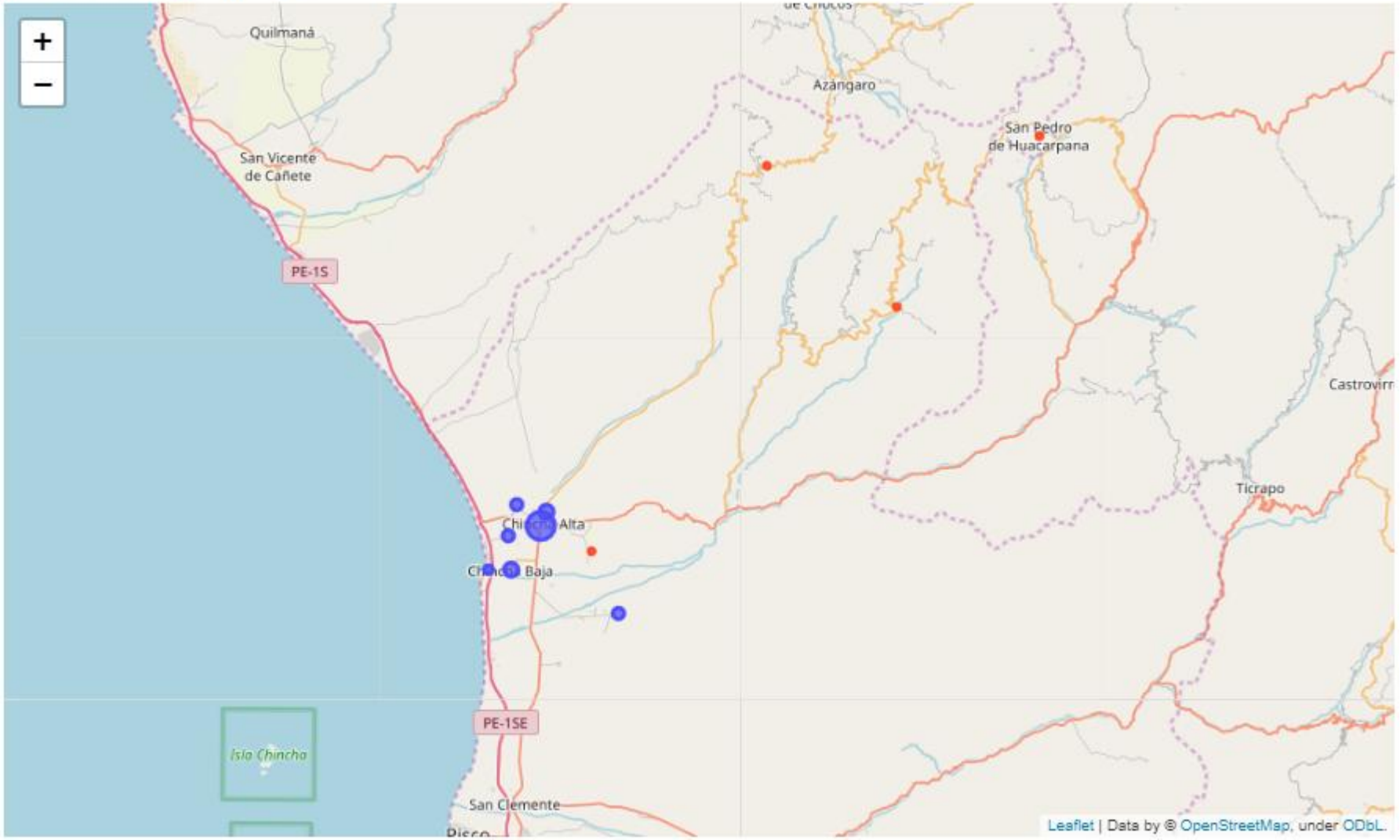
Region: ICA

Province: CHINCHA

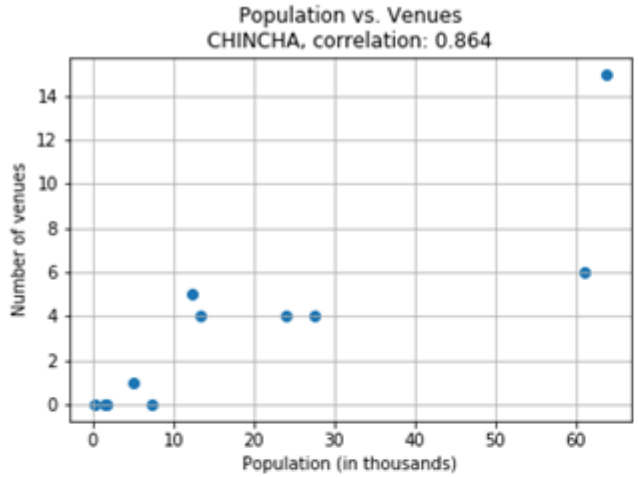
```
Ready. Execute the next cells to review data.
```

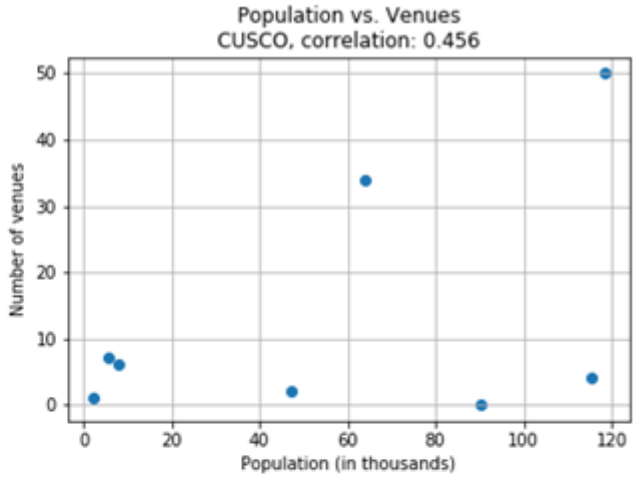
Visual representation of number of venues in each borough

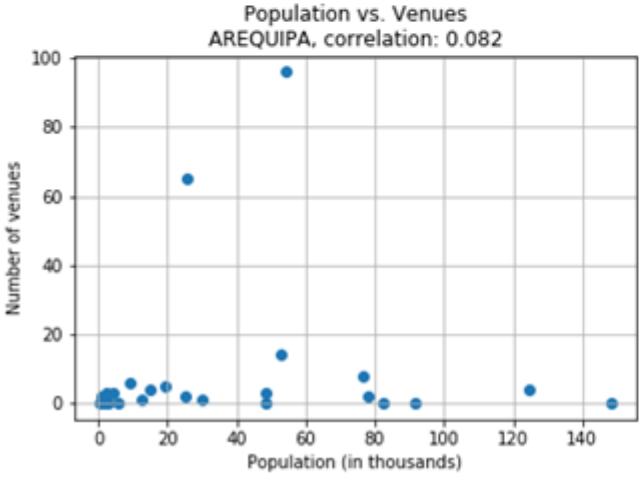
```
map_with_venues(province_df)
```

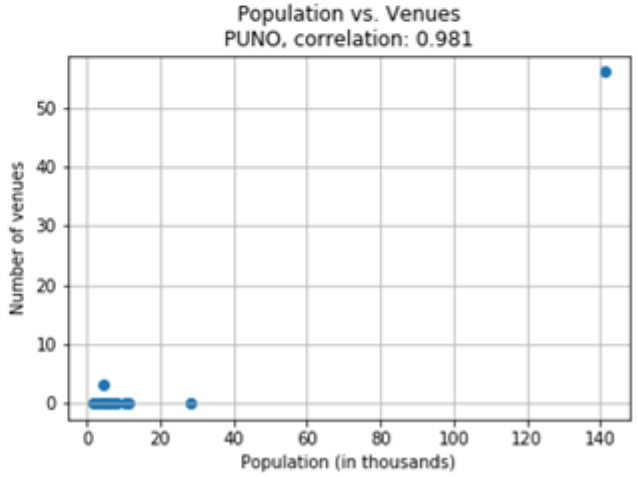


Scatter plots and correlations between populations and number of venues for four Peruvian provinces









Discussion and Conclusions

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