Placing a Café
In Arica

By: Cristian Faúndez L.

1. Introduction and business problem

1.1 Background:

Arica is a small city located at the North border of Chile, it has a estimated population of 202.131 hab. My brother is a No1 Coffee fan and in the last year he has been investigating about it in the region and found out that there are a very small amount of café in the city and almost none of them sells real quality Coffee. I decided to help him in his goal and apply all the knowledge that I had received in the IBM Data Science Course to find the perfect location to place a Café.

1.2 Problem to solve

The problem to solve is to find areas that fulfill the next conditions with the idea of find the best place to locate a Cafe:

- Be highly populated, more people mean more possible clients
- Low amount of cafe nearby, to fulfill an empty place and avoid competition

1.3 Target

The target is to help people who wants to start their own Cafe to find the best place in Arica, Chile to locate their shop, They will really care about this because is well known that the income that this kind of shop receive are highly related to the place in which they are located. So, this work will at least find the best place based on statistics and may increase the possible revenue that any owner could have.

1.4 Interest

This work will be of high interest for those people who wants to start its own Cafe in the region, and it may be of interest for people who want to make the same research in their city.

2. Data

In chile the biggest source of demographic information is the National Census that is held by the “Instituto Nacional de Estadisticas” (INE), In this case the 2002 census will be used because is the one that’s has the highest accuracy.

2.1 Data Requirements

The relevant data needed is the following:

- City separated by sectors, unfortunately unlike the US in Chile the INE does not separate the Cities by Neighborhoods or Postal code. The most similar to that are the Districts (Clusters of highly populated areas in each city).
- Geographical References to place each district.
- Population of each District.
- Amount of Cafe near each District.

2.2 Data Sources.

The INE has all the information that is needed but it isn’t highly detailed so there is a lack of information in some cases, as an habitant of the city I will use the INE data and fill the empty space with what I know about the city, for Ex. The districts don’t have the Latitude and longitude but there is a map with the borders of each one and their names are based on relevant places in each district so it will be easy to fulfill those empty spaces, I will locate the relevant places on Google Maps, Wikipedia and fuse the data with each district. The population of each district is found in this same data source.

The Cafe data will be obtained from the Foursquare API.

Example of INE District information:

Distrito censal	Población (2002)	Superficie (km²)
Puerto	2 744	1,2
Regimiento	3 880	0,7
Chinchorro	12 816	13,3
San José	13 216	1,2
Población Chile	9 086	17,3

Example of District map: Green Lines represents each district area.

2.3 Data Cleaning

The data was obtained from different sources, the districts where taken from INE CENSUS but this source didn't have the location of each District so the data was combined with longitude and latitude handly obtained from Wikipedia and Google Maps, It is an estimated value based on the center of the whole area that it covers, the column labels where translated to the English language in order to be more comprehensible to the international readers, The income data was obtained from an INE graph because there wasn´t numeric data obtainable from other source, the income was write In Low,Mid-low,Mid,Mid-High and High, for data analysis purposes it was converted into a scale where Low=0 and it increase +0.25 per level until it reach High=1.

2.4 Feature Selection

All the data will be used in exception of the area column that will be deleted because it doesn´t contribute to the analysis goal. The area column was dropped.

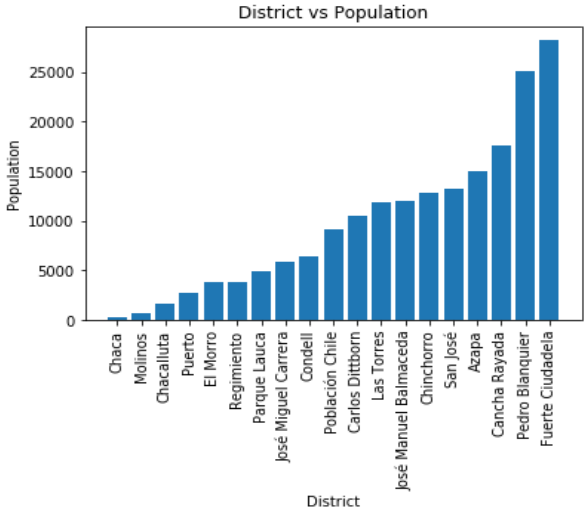
Example of final data:

	District	Population	Latitude	Longitude	Income
0	Puerto	2744	-18.476137	-70.320458	0.50
1	Regimiento	3880	-18.472886	-70.312322	0.50
2	Chinchorro	12816	-18.465493	-70.302521	0.75
3	San José	13216	-18.470838	-70.292194	0.50
4	Población Chile	9086	-18.473893	-70.285912	0.25

3. Methodology

3.1 Relation between District and Population

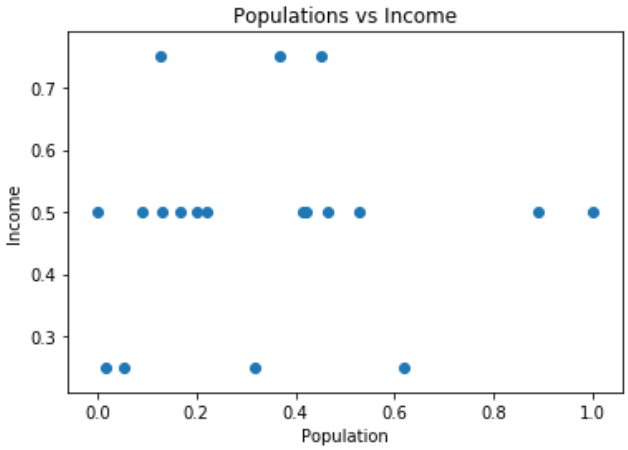
A bar graph is used to see the number of habitants per district in order to notice the more relevant districts.



3.2 Relation between Population and Income

A scatter plot was used to observe if there is any relation between the number of habitants per district and the mean income of each one, the population was normalized in order to match the income scale.

	Population	Income
0	0.090081	0.50
1	0.130672	0.50
2	0.449975	0.75
3	0.464268	0.50
4	0.316694	0.25



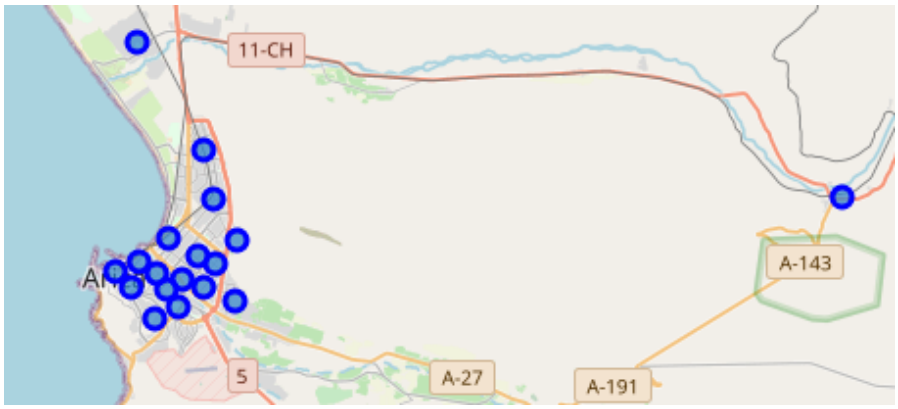
There is no actual relation between the population and the income of each district showing that every case must be viewed separately one from the other.

Districts with an income lower than the 0.50 will be dropped from the data because the basic idea is that people has the enough conditions to go to café, larger population doesn't mean that people will actually buy a café in your place.

3.3 Districts too far from the city

Map of the districts

The Geopy api is used to find the specific location of Arica and Folium is used to find and mark in the map all the districts locations.

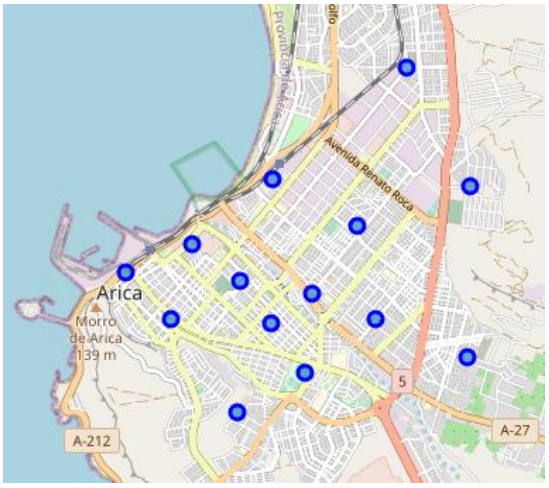


Those that are too far away from the city will be dropped because there is no reason to have a Cafe in places that are too far away from the working center of the city.

3.4 Final Data

Resulting data frame after applying the drop of districts with an income lower 0.5 and those who where too far away from the center of the city.

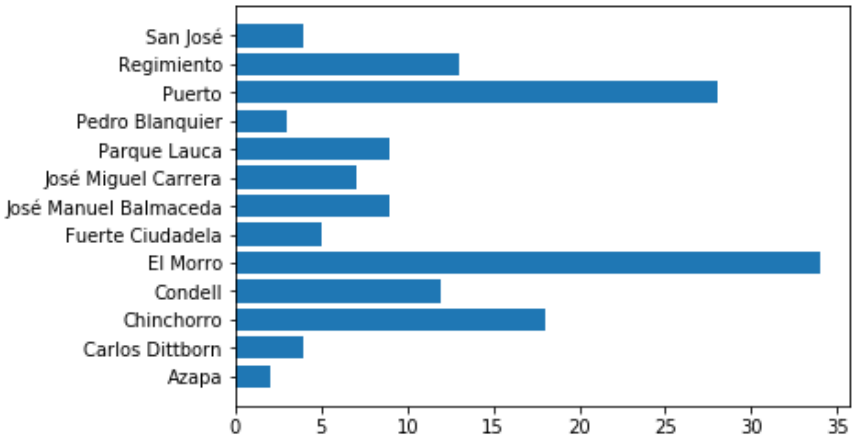
	District	Population	Latitude	Longitude	Income
0	Puerto	2744	-18.476137	-70.320458	0.50
1	Regimiento	3880	-18.472886	-70.312322	0.50
2	Chinchorro	12816	-18.465493	-70.302521	0.75
3	San José	13216	-18.470838	-70.292194	0.50
4	Azapa	14991	-18.466193	-70.278426	0.50
5	José Manuel Balmaceda	11984	-18.481534	-70.289966	0.50
6	Carlos Dittborn	10525	-18.487754	-70.298611	0.75
7	Parque Lauca	4934	-18.478699	-70.297719	0.50
8	José Miguel Carrera	5836	-18.482093	-70.302790	0.50
9	Condell	6358	-18.477089	-70.306425	0.50
10	Fuerte Ciudadela	28209	-18.492236	-70.306770	0.50
11	El Morro	3826	-18.481568	-70.314910	0.75
12	Pedro Blanquier	25131	-18.452575	-70.286289	0.50
13	Las Torres	11878	-18.485908	-70.278773	0.50



3.5 Venues per district

Foursquare api was used to find the amount of venues close to each district.

District	District Latitude
Azapa	2
Carlos Dittborn	4
Chinchorro	18
Condell	12
El Morro	34
Fuerte Ciudadela	5
José Manuel Balmaceda	9
José Miguel Carrera	7
Parque Lauca	9
Pedro Blanquier	3
Puerto	28
Regimiento	13
San José	4

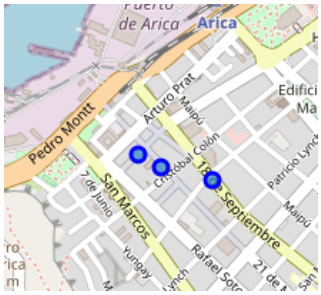


There are venues that has a low amount of venues, I make the supposition that these zones are low economic centers, so a larger amount of venues will be privileged.

3.6 Number of Café near each district

The information of venues was used to found only the venues that are in the categories of Café or Coffee Shop (These are the categories that Foursquare use to call the kind of store that we are looking for).

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
13	Puerto	-18.476137	-70.320458	Benaki	-18.477841	-70.319844	Café
18	Puerto	-18.476137	-70.320458	Café del Mar	-18.478226	-70.319157	Coffee Shop
142	El Morro	-18.481568	-70.314910	Café Latino	-18.478584	-70.317591	Café



There are 3 cafés, two of the in the district “Puerto” and one in “El Morro” and how it was predicted these are the Districts with the biggest amount of venues, revealing that there is an actual correlation between the economic centers and amount of Café in the city (It isn’t linear).

3.7 Machine learning

There is no need of Machine Learning techniques in this analysis because we are looking for a specific place that fulfill certain conditions, and there in no need of predictiveness because we have all the information that we need to perform a manual analysis of the situation, although the system could be converted to an classification type algorithm.

4. Results.

4.1 What is considered good?

All the places where:

a) The Income is higher than 0.5, being 0.75 the highest priority.

b) The number of venues is larger than 10, the more the best.

c) There is a low number of cafés, being 0 the best and 2 the worst.

d) The number of habitants is the biggest possible, over 10k is a good amount (No priority).

4.2 The best result.

The district that best fit the requirements is the “Chinchorro” district. The reasons are explained using the same punctuation than the previous section.

a) Is one of the three districts with the highest income of 0.75, being these Carlos Dittborn, El Morro and Chinchorro.

	District	Population	Latitude	Longitude	Income
2	Chinchorro	12816	-18.465493	-70.302521	0.75
7	Carlos Dittborn	10525	-18.487754	-70.298611	0.75
13	El Morro	3826	-18.481568	-70.314910	0.75

b) Is one of the five districts with the highest number of venues, having 18 which is greater than Carlos Dittborn (4) and lower than Chinchorro (34).

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
2	Chinchorro	20	20	20	20	20	20
3	Condell	11	11	11	11	11	11
4	El Morro	33	33	33	33	33	33
10	Puerto	28	28	28	28	28	28
11	Regimiento	13	13	13	13	13	13

c) Is one of the 12 districts with 0 cafe. Is lower than Puerto (2) and Chinchorro(1).

d) Is one of the 12 districts with population over 10k. being higher than Morro and Carlos Dittborn.

	District	Population	Latitude	Longitude	Income
2	Chinchorro	12816	-18.465493	-70.302521	0.75
3	San José	13216	-18.470838	-70.292194	0.50
5	Azapa	14991	-18.466193	-70.278426	0.50
6	José Manuel Balmaceda	11984	-18.481534	-70.289966	0.50
7	Carlos Dittborn	10525	-18.487754	-70.298611	0.75
11	Fuerte Ciudadela	28209	-18.492236	-70.306770	0.50
16	Pedro Blanquier	25131	-18.452575	-70.286289	0.50
18	Las Torres	11878	-18.485908	-70.278773	0.50

According to this is easy to see that Chincorro is an economic center, having 18 venues which represent a good presence of public in this place and has an income of 0.75 one of the highest in the city. The amount of 10k people and 0 café represents that there may be a need of having a Café in this place because there is enough people, income and store presence to have one.

So according to all the previously said Chinchorro is the District that best match the conditions that has been propose.

5 Discussion.

According to the results there are a very good amount of places to locate a Café but because of the conditions that were placed it only ends showing one relevant place that is the Chinchorro districts, but there could be a possibility that the best place could be el Morro or Puerto because they already have Café meaning that there is a very good place to locate one, and because the amount of café is very low in general, the competition that could exist is very low. Is highly recommended to repeat the process in the future with more relevant data and use apis that actually work for the place In which you are located at

6. Conclusion

Although according to the results the best place to locate a Café is the Chinchorro District there is not enough data or information to make a precise decision. The foursquare api doesn't have all the Café of the zone registered and the lack of information of the INE gives a big space to misinterpretations, This data could be relevant to discard some places but for the final decision it would be necessary to make a more deepest research including data like number of Tourist/public and how probably is that a person that visit the place bought a coffee, With all that extra information the real results could be a more efficient method to choose a place where to locate a Café. But now for obvious motives it is impossible to go out on quarantine , I expect that this work could help people decide where to locate a café but I will definitely end improving this data analysis in order to give a great work that could help the whole community.