Report: The Battle of Neighborhoods

Kenyi

April 2021

Abstract

I currently live in Guadadalaja, Jalisco, México. I'm looking for a job as Data Scientist but my chances of finding a job increase if a look in a big City. Besides, I would like to find a similar city like Guadalajara. For this proposal, it will be necessary to consider the population density of each city, the services and, the average income by state. In the present job, we use different tools of data processing and models to solve this problem.

1 Introduction

In the present document we show the posible answer to the question:
Were is the best state in Mexico to looking for a job?

For this quest I'm looking for a place very similar to Guadalajara, Jalisco, Mex. (My current live place). The main characteristics of this question are

1. Average dairy income for each city in México.
2. Population size, because this determine a better life quality.
3. Venues around 5 km to the downtown.

The data will be collected from different sources like Wikipedia, the official page of the Mexican Government and, information from Foursquare.

The used methodology was CRISP-DM and the used tools include numpy, Folium, pandas and sklearn.

2 The problem

the detailed problem could be seen as a clustering problem in the first approach, we try to identify similar cities to Guadalajara. The derivation will be similar to the third assignment of this course with a lite variation in the data. In this case, we shall aggregate information about population size and average daily income for each city.

The answer to the question will be a city like Guadalajara with the best daily income average.

3 Data

3.1 Data Understanding

The information to solve this problem is available in many sources. We will take the info from the 2020 Census made by INEGI (The National Institute of Statistics and Geography), this data is placed in Wikipedia and the official page of the government. The geospatial data and venue will be taken from Foursquare.

The census information from wikipedia is in a html table.

3.2 Data Preparation

Rank	City	Municipality	State	Geo. coordinates	2020 Census	2010 Census	Change
1	***Mexico City***	16 boroughs‡	Mexico City	19°25′57″N 99°07′52″W	9,209,944	8,851,080	+4.05%
2	*Tijuana*	Tijuana	Baja California	32°32′05″N 117°02′37″W	1,810,645	1,300,983	+39.18%
3	*Ecatepec*	Ecatepec	State of Mexico	19°36′35″N 99°03′36″W	1,643,623	1,655,015	−0.69%
4	*León*	León	Guanajuato	21°07′11″N 101°40′50″W	1,579,803	1,238,962	+27.51%
5	***Puebla***	Puebla	Puebla	19°02′43″N 98°11′51″W	1,542,232	1,434,062	+7.54%
6	*Ciudad Juárez*	Juárez	Chihuahua	31°44′22″N 106°29′13″W	1,501,551	1,321,004	+13.67%
7	***Guadalajara***	Guadalajara	Jalisco	20°40′35″N 103°20′32″W	1,385,621	1,495,182	−7.33%
8	Zapopan	Zapopan	Jalisco	20°43′14″N 103°23′16″W	1,257,547	1,142,483	+10.07%
9	***Monterrey***	Monterrey	Nuevo León	25°40′17″N 100°18′31″W	1,142,952	1,135,512	+0.66%
10	Ciudad Nezahualcóyotl	Nezahualcóyotl	State of Mexico	19°24′00″N 98°59′20″W	1,072,676	1,104,585	−2.89%
11	**Chihuahua**	Chihuahua	Chihuahua	28°38′07″N 106°05′20″W	925,762	809,232	+14.40%
12	***Mérida***	Mérida	Yucatán	20°58′04″N 89°37′18″W	921,771	777,615	+18.54%
13	Naucalpan	Naucalpan, Huixquilucan‡	State of Mexico	19°28′31″N 99°14′16″W	911,168	913,681	−0.28%
14	*Cancún*	Benito Juárez	Quintana Roo	21°09′38″N 86°50′51″W	888,797	628,306	+41.46%
15	***Saltillo***	Saltillo	Coahuila	25°26′00″N 101°00′00″W	864,431	709,671	+21.81%
16	***Aguascalientes***	Aguascalientes	Aguascalientes	21°52′51″N 102°17′46″W	863,893	722,250	+19.61%
17	***Hermosillo***	Hermosillo	Sonora	29°05′56″N 110°57′15″W	855,563	715,061	+19.65%
18	**Mexicali**	Mexicali	Baja California	32°39′48″N 115°28′04″W	854,186	689,775	+23.84%
19	***San Luis Potosí***	San Luis Potosí	San Luis Potosí	22°08′59″N 100°58′30″W	845,941	722,772	+17.04%
20	***Culiacán***	Culiacán	Sinaloa	24°47′31″N 107°23′53″W	808,416	675,773	+19.63%
21	***Querétaro***	Querétaro	Querétaro	20°35′17″N 100°23′17″W	794,789	626,495	+26.86%
22	***Morelia***	Morelia	Michoacán	19°42′08″N 101°11′08″W	743,275	597,511	+24.40%
23	Chimalhuacán	Chimalhuacán	State of Mexico	19°26′15″N 98°57′15″W	703,215	612,383	+14.83%
24	*Reynosa*	Reynosa	Tamaulipas	26°05′32″N 98°16′40″W	691,557	589,466	+17.32%
25	Torreón	Torreón	Coahuila	25°32′40″N 103°26′30″W	690,193	608,836	+13.36%
26	Tlalnepantla	Tlalnepantla	State of Mexico	19°32′12″N 99°11′41″W	658,907	653,410	+0.84%

We use the panda's tool read_html that allows us to parse Html table to a pandas dataframe.

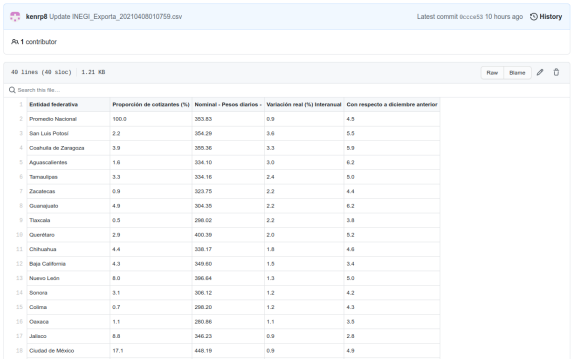
We only need the columns of city names, states and, 2020 census, then we drop anything else.

The coordinates for each city will be taken from a request method with geopy library.

The information about Nominal Dairy Income came from a statistical report in pdf format. This info was parse to CSV format and saved in the GitHub repository of this project. In the notebook, this info is fetched using the read_cvs tool from panda. We drop all columns except Nominal dairy income and city name. The name of cities will be replaced for respective names in the first table.

Finally, we can get venue information using foursquare like in the third assignment and we create dummy variables for each venue.

The final step is to merge all tables using the city name attribute.



kenrp8 Update INEGI_Exporta_20210408010759.csv — Latest commit 8ccce53 10 hours ago — History

1 contributor

40 lines (40 sloc) 1.21 KB — Raw Blame

	Entidad federativa	Proporción de cotizantes (%)	Nominal - Pesos diarios -	Variación real (%) Interanual	Con respecto a diciembre anterior
2	Promedio Nacional	100.0	353.83	0.9	4.5
3	San Luis Potosí	2.2	354.29	3.6	5.5
4	Coahuila de Zaragoza	3.9	355.36	3.3	5.9
5	Aguascalientes	1.6	304.10	3.0	6.2
6	Tamaulipas	3.3	334.16	2.4	5.0
7	Zacatecas	0.9	323.75	2.2	4.4
8	Guanajuato	4.9	304.35	2.2	6.2
9	Tlaxcala	0.5	298.02	2.2	3.8
10	Querétaro	2.9	400.39	2.0	5.2
11	Chihuahua	4.4	338.17	1.8	4.6
12	Baja California	4.3	349.60	1.5	3.4
13	Nuevo León	8.0	396.64	1.3	5.0
14	Sonora	3.1	306.12	1.2	4.2
15	Colima	0.7	298.20	1.2	4.3
16	Oaxaca	1.1	280.86	1.1	3.5
17	Jalisco	8.8	346.23	0.9	2.8
18	Ciudad de México	17.1	448.19	0.9	4.9

4 Modeling

In this case, we use a classic clustering model, we use the K-means algorithm. This allows us to know what cities are similar to Guadalajara. Similar in Venues, Population and Income.

Why these variables? There exist a lot of indicators for solving the main question. We take these indicators because under my observation are the characteristics that I like from my living place.

We use Elbow Method to chooce a good k value. K=5.

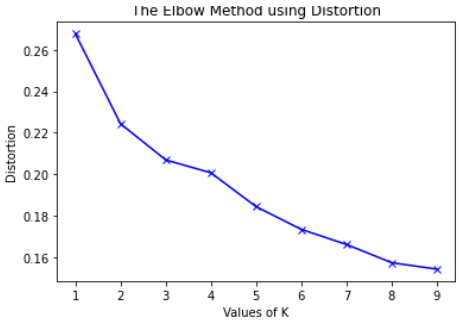


Figure 1: Plot of Elbow Method

5 Evaluation

At the end of the fitting model, we see that Guadalajara lies in cluster zero.

That cluster represents the cities that are similar to Guadalajara. Considering socioeconomic facts about México the result seems satisfactory.


```
[1] clustered_data.loc[clustered_data['Cluster Labels'] == 0, clustered_data.columns[0]]
```

1	Aguascalientes
2	Buenavista
4	Campeche
7	Chalco
9	Chicoloapan
10	Chihuahua
12	Chimalhuacán
13	Ciudad Acuña
14	Ciudad Apodaca
15	Ciudad Benito Juárez
16	Ciudad Juárez
17	Ciudad López Mateos
18	Ciudad Madero
19	Ciudad Nezahualcóyotl
20	Ciudad Nicolás Romero
22	Ciudad Victoria
24	Ciudad del Carmen
25	Coatzacoalcos
27	Cuautitlán Izcalli
28	Cuautla
29	Cuernavaca
32	Ecatepec
33	Ensenada
34	García
35	General Escobedo
36	Guadalajara
37	Guadalupe
42	Ixtapaluca
43	Jiutepec
48	Matamoros

6 Response

Now we have the information to make a decision and solve the main problem. Ordering the elements using the Nominal Daily Income we get the following results:

```
[34] clustered_data[clustered_data['Cluster Labels'] == 0].sort_values("Nominal daily income", as
```

	City	Nominal daily income	2020 Census	Zoo Exhibit	Accessories Store	Airport	Airport Lounge	Airport Terminal	American Restaurant	A
67	Querétaro	0.745298	0.071480	0.0	0.01	0.0	0.0	0.0	0.0	
73	San Juan del Río	0.745298	0.003393	0.0	0.00	0.0	0.0	0.0	0.0	
24	Ciudad del Carmen	0.735280	0.004885	0.0	0.00	0.0	0.0	0.0	0.0	
4	Campeche	0.735280	0.011327	0.0	0.00	0.0	0.0	0.0	0.0	
15	Ciudad Benito Juárez	0.725316	0.017800	0.0	0.00	0.0	0.0	0.0	0.0	

5 rows × 345 columns

We can conclude that the best option is Querétaro, Qro, México.