Report: The Battle of Neighborhoods

Coursera Capstone

Kenyi Josué Ramírez

Data Sceince Course: Applied Data Science Capstone

Overview

1. Problem Description
2. Data
3. Modeling
4. Answer and Conclusions

The Problem

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.

The Problem

The main characteristics of this question are:

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

Proposal of solution

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.

Proposal of solution

The main tools for this Data Analysis will be:

Tools

1. Pandas.
2. KMeans from Sklearn.
3. Numpy.

Sources

Sources

1. Mexican Government Page
2. Wikipedia
3. https://foursquare.com/

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.

Sources

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′18″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%

Figure: HTML Table from Wikipedia.

Sources

kenrp8 Update INEGI_Exporta_20210408010759.csv Latest commit 0ccce53 10 hours ago History

1 contributor

40 lines (40 sloc) | 1.21 KB

Raw Blame

Search this file…

Entidad federativa	Proporción de cotizantes (%)	Nominal - Pesos diarios -	Variación real (%) Interanual	Con respecto a diciembre anterior
Promedio Nacional	100.0	353.83	0.9	4.5
San Luis Potosí	2.2	354.29	3.6	5.5
Coahuila de Zaragoza	3.9	355.36	3.3	5.9
Aguascalientes	1.6	334.10	3.0	6.2
Tamaulipas	3.3	334.16	2.4	5.0
Zacatecas	0.9	323.75	2.2	4.4
Guanajuato	4.9	304.35	2.2	6.2
Tlaxcala	0.5	298.02	2.2	3.8
Querétaro	2.9	400.39	2.0	5.2
Chihuahua	4.4	338.17	1.8	4.6
Baja California	4.3	349.60	1.5	3.4
Nuevo León	8.0	396.64	1.3	5.0
Sonora	3.1	306.12	1.2	4.2
Colima	0.7	298.20	1.2	4.3
Oaxaca	1.1	280.86	1.1	3.5
Jalisco	8.8	346.23	0.9	2.8
Ciudad de México	17.1	448.19	0.9	4.9

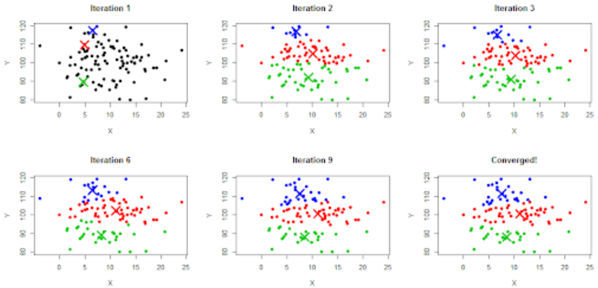
Figure: CVS sample (Is saved in GitHub as CSV file).

Sources

The preprocessing of this data was made using pandas to create a unique table for clustering process.

Clustering Model

We use K-Mean algorithm for clustering. In our data each variable represent a dimension in the space and we cosider each sample of data as a part of a data cloud (Cluster). This represent similarities between samples.



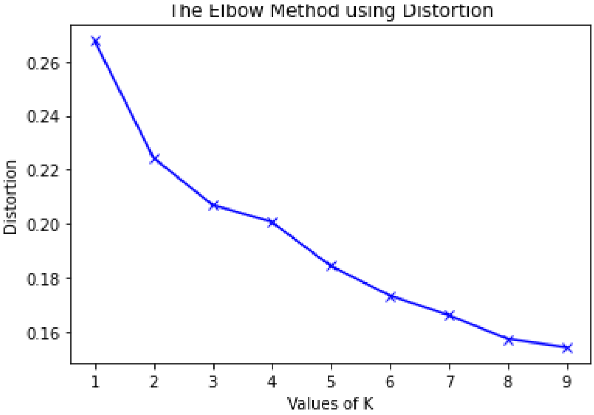
Clustering Model

	City	Nominal daily income	2020 Census	Zoo Exhibit	Accessories Store	Airport	Airport Lounge	Airport Terminal	American Restaurant	Amp
0	Acapulco	0.134864	0.056454	0.0	0.0	0.0	0.0	0.0	0.01	0.0
1	Aguascalientes	0.392071	0.079105	0.0	0.0	0.0	0.0	0.0	0.00	0.0
2	Buenavista	0.436511	0.007703	0.0	0.0	0.0	0.0	0.0	0.00	0.0
3	Cabo San Lucas	0.301220	0.006149	0.0	0.0	0.0	0.0	0.0	0.00	0.0
4	Campeche	0.735280	0.011327	0.0	0.0	0.0	0.0	0.0	0.00	0.0

Figure: Data table before the clustering.

Clustering Model

We use the elbow method to choose the best k-value for the model. In this case we take k=5.



Clustering Model

We proceed to fit the model and we get this segmentation:

Clustering Model

We see that Guadalajara lies in the zero cluster, that is the set of similar cities to Guadalajara.

```
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
```

Clustering Model

Finally we can choose the city in the first cluster with the highest Nominal Dairy income, in this case that corresponds to Querétaro City.

```
[34]  1 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

Clustering Model

This conclusion makes sense because it is a growing city with a lot of new jobs and it is very similar to Guadalajara in the kind of venues[1].

[1] https://www.diariodequeretaro.com.mx/local/imparable-mancha-urbana-crece-2.9-5127464.html

References

Elbow Method for optimal value of k in KMeans
https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

Panda Documentation
https://pandas.pydata.org/docs/

SKlearn KMeans documentation
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

INEGI
https://www.inegi.org.mx/

The End