

# Report: The Battle of Neighborhoods

Kenyi

April 2021

## Abstract

I currently live in Guadalajara, Jalisco, México. I'm looking for a job as Data Scientist but my chances of finding a job increase if I 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 possible answer to the question:

**Where is the best state in Mexico to look 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 daily income for each city in México.
2. Population size, because this determines 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 little 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	<span><span><span><span>19°25′57″N</span> <span>99°10′32″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>19.4325°N 99.1756°W</span><span><span>﻿</span> / <span>19.4325; -99.1756</span></span></span></span>	9,209,944	8,851,080	+4.05%
2	Tijuana	Tijuana	Baja California	<span><span><span><span>32°32′05″N</span> <span>117°02′37″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>32.5347°N 117.0436°W</span><span><span>﻿</span> / <span>32.5347; -117.0436</span></span></span></span>	1,810,645	1,390,983	+30.18%
3	Ecatepec	Ecatepec	State of Mexico	<span><span><span><span>19°38′05″N</span> <span>99°52′38″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>19.6347°N 99.8772°W</span><span><span>﻿</span> / <span>19.6347; -99.8772</span></span></span></span>	1,643,623	1,655,015	-0.69%
4	León	León	Guanajuato	<span><span><span><span>21°07′17″N</span> <span>101°40′50″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>21.1214°N 101.6806°W</span><span><span>﻿</span> / <span>21.1214; -101.6806</span></span></span></span>	1,579,903	1,238,962	+27.51%
5	Puebla	Puebla	Puebla	<span><span><span><span>19°02′47″N</span> <span>98°15′51″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>19.0464°N 98.2642°W</span><span><span>﻿</span> / <span>19.0464; -98.2642</span></span></span></span>	1,542,232	1,434,062	+7.54%
6	Ciudad Juárez	Juárez	Chihuahua	<span><span><span><span>31°44′22″N</span> <span>106°29′13″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>31.7394°N 106.4869°W</span><span><span>﻿</span> / <span>31.7394; -106.4869</span></span></span></span>	1,501,551	1,321,004	+13.67%
7	Guadalajara	Guadalajara	Jalisco	<span><span><span><span>20°40′35″N</span> <span>103°20′32″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>20.6764°N 103.3422°W</span><span><span>﻿</span> / <span>20.6764; -103.3422</span></span></span></span>	1,385,621	1,495,182	-7.33%
8	Zapopan	Zapopan	Jalisco	<span><span><span><span>20°43′14″N</span> <span>103°23′18″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>20.7206°N 103.3883°W</span><span><span>﻿</span> / <span>20.7206; -103.3883</span></span></span></span>	1,257,547	1,142,483	+10.07%
9	Monterrey	Monterrey	Nuevo León	<span><span><span><span>25°40′17″N</span> <span>100°18′31″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>25.6714°N 100.3083°W</span><span><span>﻿</span> / <span>25.6714; -100.3083</span></span></span></span>	1,142,952	1,135,512	+0.66%
10	Ciudad Nezahualcóyotl	Nezahualcóyotl	State of Mexico	<span><span><span><span>19°24′00″N</span> <span>98°59′20″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>19.4000°N 98.9889°W</span><span><span>﻿</span> / <span>19.4000; -98.9889</span></span></span></span>	1,072,676	1,104,585	-2.89%
11	Chihuahua	Chihuahua	Chihuahua	<span><span><span><span>28°38′07″N</span> <span>106°55′20″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>28.6353°N 106.9222°W</span><span><span>﻿</span> / <span>28.6353; -106.9222</span></span></span></span>	925,762	809,232	+14.40%
12	Mérida	Mérida	Yucatán	<span><span><span><span>20°58′04″N</span> <span>89°37′18″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>20.9678°N 89.6217°W</span><span><span>﻿</span> / <span>20.9678; -89.6217</span></span></span></span>	921,771	777,615	+18.54%
13	Naucalpan	Naucalpan, Huixquilucan	State of Mexico	<span><span><span><span>19°28′31″N</span> <span>99°14′18″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>19.4750°N 99.2383°W</span><span><span>﻿</span> / <span>19.4750; -99.2383</span></span></span></span>	911,168	913,681	-0.28%
14	Cancún	Benito Juárez	Quintana Roo	<span><span><span><span>21°09′38″N</span> <span>86°50′51″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>21.1606°N 86.8475°W</span><span><span>﻿</span> / <span>21.1606; -86.8475</span></span></span></span>	888,797	628,306	+41.46%
15	Saltillo	Saltillo	Coahuila	<span><span><span><span>25°28′00″N</span> <span>101°00′00″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>25.4667°N 101.0000°W</span><span><span>﻿</span> / <span>25.4667; -101.0000</span></span></span></span>	864,431	709,671	+21.81%
16	Aguascalientes	Aguascalientes	Aguascalientes	<span><span><span><span>21°52′51″N</span> <span>102°17′48″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>21.8808°N 102.2967°W</span><span><span>﻿</span> / <span>21.8808; -102.2967</span></span></span></span>	863,893	722,250	+19.61%
17	Hermosillo	Hermosillo	Sonora	<span><span><span><span>29°05′56″N</span> <span>110°57′15″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>29.0989°N 110.9542°W</span><span><span>﻿</span> / <span>29.0989; -110.9542</span></span></span></span>	855,563	715,061	+19.65%
18	Mexicali	Mexicali	Baja California	<span><span><span><span>32°39′48″N</span> <span>115°03′04″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>32.6633°N 115.0511°W</span><span><span>﻿</span> / <span>32.6633; -115.0511</span></span></span></span>	854,186	689,775	+23.84%
19	San Luis Potosí	San Luis Potosí	San Luis Potosí	<span><span><span><span>22°08′50″N</span> <span>100°56′30″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>22.1472°N 100.9417°W</span><span><span>﻿</span> / <span>22.1472; -100.9417</span></span></span></span>	845,941	722,772	+17.04%
20	Culiacán	Culiacán	Sinaloa	<span><span><span><span>24°47′31″N</span> <span>107°23′52″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>24.7919°N 107.3978°W</span><span><span>﻿</span> / <span>24.7919; -107.3978</span></span></span></span>	808,416	675,773	+19.63%
21	Querétaro	Querétaro	Querétaro	<span><span><span><span>20°35′17″N</span> <span>100°23′17″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>20.5881°N 100.3881°W</span><span><span>﻿</span> / <span>20.5881; -100.3881</span></span></span></span>	794,789	626,495	+26.86%
22	Morelia	Morelia	Michoacán	<span><span><span><span>19°42′08″N</span> <span>101°11′08″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>19.7022°N 101.1856°W</span><span><span>﻿</span> / <span>19.7022; -101.1856</span></span></span></span>	743,275	597,511	+24.40%
23	Chimalhuacán	Chimalhuacán	State of Mexico	<span><span><span><span>19°28′15″N</span> <span>98°57′15″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>19.4708°N 98.9542°W</span><span><span>﻿</span> / <span>19.4708; -98.9542</span></span></span></span>	703,215	612,383	+14.83%
24	Reynosa	Reynosa	Tamaulipas	<span><span><span><span>28°08′32″N</span> <span>98°18′40″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>28.1422°N 98.3111°W</span><span><span>﻿</span> / <span>28.1422; -98.3111</span></span></span></span>	691,557	589,466	+17.32%
25	Torreón	Torreón	Coahuila	<span><span><span><span>25°32′40″N</span> <span>103°26′30″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>25.5444°N 103.4417°W</span><span><span>﻿</span> / <span>25.5444; -103.4417</span></span></span></span>	690,193	608,836	+13.36%
26	Tlalnepantla	Tlalnepantla	State of Mexico	<span><span><span><span>19°32′12″N</span> <span>99°11′41″W</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>19.5367°N 99.1947°W</span><span><span>﻿</span> / <span>19.5367; -99.1947</span></span></span></span>	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_csv` 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.


[Update INEGI\\_Ejercicio\\_20210408010758.csv](#)
Latest commit: successful 10 hours ago
[History](#)

R 1 contributor

49 Lines (49 lines) 1.21 KB

Raw Edit

	Entidad federativa	Proporcion de visitantes (%)	Numeral - Paises diarios	Varacion real (%) Intervalo	Con respecto a diciembre anterior
1	Panamá Nacional	100.0	303.83	0.9	4.9
2	San Luis Potosí	2.2	204.29	3.6	5.5
3	Coahuila de Zaragoza	3.9	205.36	3.3	5.9
4	Aguascalientes	1.6	204.16	5.6	4.2
5	Tamaulipas	3.3	204.16	2.4	5.0
6	Zacatecas	0.9	203.75	2.2	4.4
7	Guerrero	4.9	204.35	2.2	6.2
8	Tlaxcala	0.5	206.02	2.2	3.8
9	Quintana Roo	2.9	400.39	2.0	5.2
10	Chihuahua	4.4	306.17	1.6	4.6
11	Baja California	4.3	340.60	1.9	3.4
12	Nuevo León	6.0	306.64	1.3	5.0
13	Baja	3.1	306.12	1.2	4.2
14	Colima	0.7	206.20	1.2	4.9
15	Oaxaca	1.1	200.06	1.1	5.8
16	Jalisco	6.8	340.23	0.9	3.8
17	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 prove several variations of the k parameter in clustering but seven was the best considering the sample size.

## 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.

## Clusters

cluster 0

Guadalajara is in this cluster.

```
In [199]: clustered_data.loc[clustered_data['Cluster Labels'] == 0, clustered_data.columns[0]]
```

```
Out[199]: 1      Aguascalientes
2      Buenavista
7      Chalco
9      Chicoloapan
10     Chihuahua
12     Chimalhuacán
13     Ciudad Acuña
16     Ciudad Juárez
17     Ciudad López Mateos
18     Ciudad Madero
19     Ciudad Nezahualcóyotl
20     Ciudad Nicolás Romero
22     Ciudad Victoria
25     Coatzacoalcos
27     Cuautitlán Izcalli
28     Cuautla
29     Cuernavaca
32     Ecatepec
33     Ensenada
36     Guadalajara
37     Guadalajara
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:

```
clustered_data[clustered_data['Cluster Labels'] == 0].sort_values('Nominal daily income', ascending=False).head()
```

	City	Nominal daily income	2020 Census	Zoo Exhibit	Accessories Store	Airport	Airport Lounge	Airport Terminal	American Restaurant	Amphitheater	Aquarium	Arcade	Arepa Restaurant	Argentinian Restaurant
37	Guadalupe	0.531118	0.028245	0.0	0.0	0.0	0.00	0.00	0.00	0.0	0.0	0.0	0.0	0.00
13	Ciudad Acuña	0.505355	0.001463	0.0	0.0	0.0	0.00	0.00	0.00	0.0	0.0	0.0	0.0	0.00
87	Torreón	0.505355	0.059939	0.0	0.0	0.0	0.01	0.01	0.00	0.0	0.0	0.0	0.0	0.00
70	Saltillo	0.505355	0.079164	0.0	0.0	0.0	0.00	0.00	0.00	0.0	0.0	0.0	0.0	0.01
62	Piedras Negras	0.505355	0.002978	0.0	0.0	0.0	0.00	0.00	0.03	0.0	0.0	0.0	0.0	0.00

We can conclude that the best option is Guadalupe, Monterrey, México.