# Explaining the levels of employment in the city of Cancun, Mexico Through Linear Regression and k-means cluster analysis

Adolfo Camacho Yague

March 6, 2020

# 1. Introduction

## 1.1 Background

The city of Cancun, Mexico is one of the most important sites for tourism in the country. The city is located in the state of Quintana Roo, the economic output of the city is greater than the capital city of the state Chetumal. The main attractions of the city are the archaeological of mayan culture sites such as Xcaret, Yamil Lu'um or Coba. Or natural sites like the mangrove swamp od Nichupte and the Cenotes which are natural sources of underground water

## 1.2 Problem

The main goal of this project is to explain the factors that contribute to the total number of people employed in the city. As a first approach the data related to the city, its venues, was downloaded, resulting in a query of 125 locations. It clear that this data is insufficient to describe the levels. It was decided to compare this set of information with the source of the Mexican government, the INEGI

## 1.3 Interest

This project could be of interest of the government officials and the people of Cancun, describing other variables and how they affect the level of employment in the city could be useful to improve or develop new government policies

# 2. Data acquisition and cleaning

## 2.1 Data Sources

As described above, the information stored in FourSquare was compared against the official source of the Mexican agency, available here . Querying the information only for the city

Besides the target and independent variables, the postal code and the latitude and longitude of each AGEB, this for the construction of various maps

## 2.2 Data Cleaning

5 different datasets were downloaded, one for each independent variables of the models; total population, holders of BA degrees 25 years and above, people with employment, no vacant housing, people with government health insurance. The datasets consist in only two variables, the amount of people with the trait of the variable and the key value of the zone inside the city, AGEB, that is the division for urban population for the census of the Mexican government

The coordinates of this AGEBs were retrieved for another source of the government, from the same agency INEGI, available here . The queries in this site are presented as lists, these lists were segmented in rows for every location, the data was transposed and only the useful information were filtered

For the purpose of having only one dataframe, all the sources were merged, all the rows with missing values were drop from the final table of data

After the visualization of the maps, a k-means analysis was conducted, the label for the variable was decomposed into dummy variables to conduct a second linear regression

## 2.3 Feature Selection

Afterwards cleaning the two datasets, only 20 rows were retrieved for the query of FourSquare and 278 rows for the data from the INEGI databases

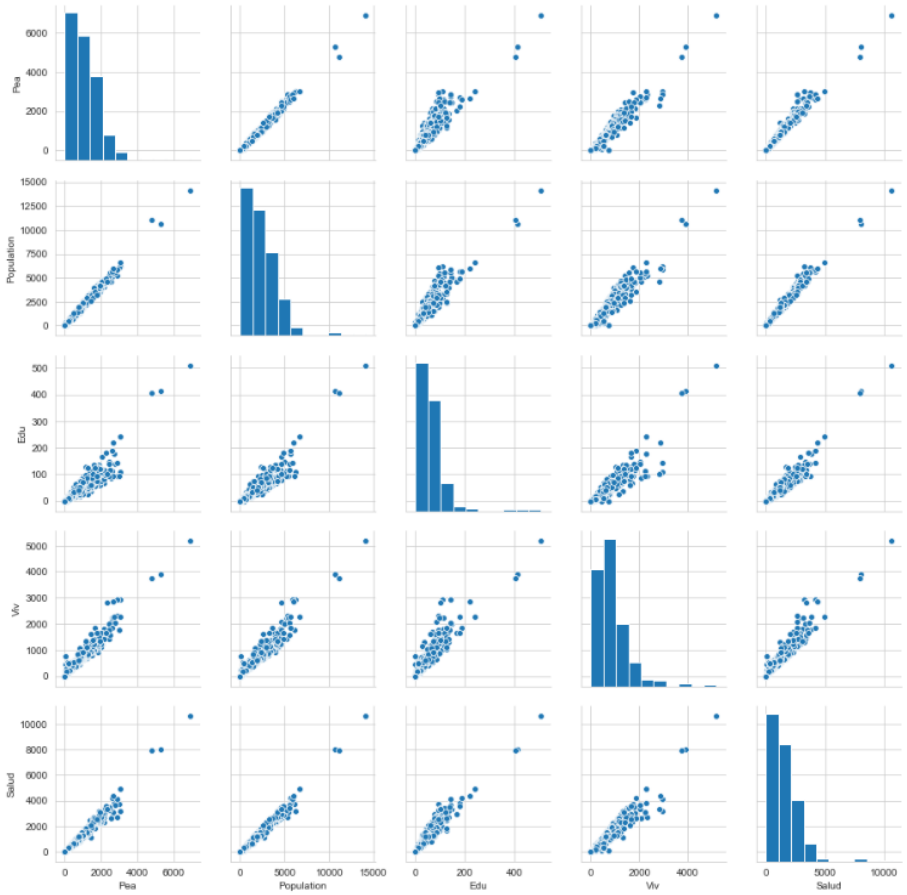
| | Ageb | Pea | Population | Longitud | Latitud | Código Postal | Clave_Ageb | Edu | Viv | Salud |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2300500010084 | 1071 | 2170.0 | -86.833074 | 21.160870 | 77516.0 | 2300500010084 | 45.0 | 947.0 | 1272 |
| 1 | 2300500010188 | 1094 | 2247.0 | -86.834151 | 21.169075 | 77560.0 | 2300500010188 | 48.0 | 985.0 | 1213 |
| 2 | 2300500010192 | 1417 | 2721.0 | -86.826575 | 21.173982 | 77535.0 | 2300500010192 | 35.0 | 1302.0 | 1128 |
| 3 | 2300500010205 | 1275 | 2427.0 | -86.820003 | 21.177392 | 77524.0 | 2300500010205 | 26.0 | 1159.0 | 1351 |
| 4 | 230050001021A | 2622 | 5150.0 | -86.836347 | 21.176514 | 97510.0 | 230050001021A | 112.0 | 2223.0 | 2763 |

Because the information and the reach of foursquare it is different, and the site it is not widely used in the city. Moreover, the information from INEGI has not divide the information into Neighborhood, it was not possible to conduct an analysis like the one presented in the course

| 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue | Neighbourhood | Postal_Code | lat | long | PostCode |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mexican Restaurant | Convenience Store | Hotel | Fast Food Restaurant | Lighthouse | Breakfast Spot | City | Pizza Place | Los Cedros, San Martín Caballero, Nuevo Amanec... | 77527 | 21.1924 | -86.8346 | 77527 |
| Seafood Restaurant | Pizza Place | Convenience Store | Restaurant | Gym / Fitness Center | Pharmacy | Italian Restaurant | Gym | Los Santos, Privadas Sacbe, Paseos Chac Mool, ... | 77518 | 21.1634 | -86.8747 | 77518 |
| onvenience Store | Pizza Place | Seafood Restaurant | Pharmacy | Gym / Fitness Center | Grocery Store | Gym | Athletics & Sports | Paseos del Sol (Supermanzana 205), Paseos del ... | 77519 | 21.1620 | -86.8810 | 77519 |
| Italian Restaurant | Café | Sushi Restaurant | Coffee Shop | Juice Bar | Ice Cream Shop | Seafood Restaurant | Burger Joint | Residencial Caracol, Supermanzana 313, Privada... | 77533 | 21.1391 | -86.8496 | 77533 |
| Mexican Restaurant | Convenience Store | Brewery | Burrito Place | Paintball Field | Seafood Restaurant | Fast Food Restaurant | Rental Car Location | Residencial Cumbres, Doctores, Supermanzana 29... | 77560 | 21.0827 | -86.8572 | 77560 |

# 3. Exploratory analysis

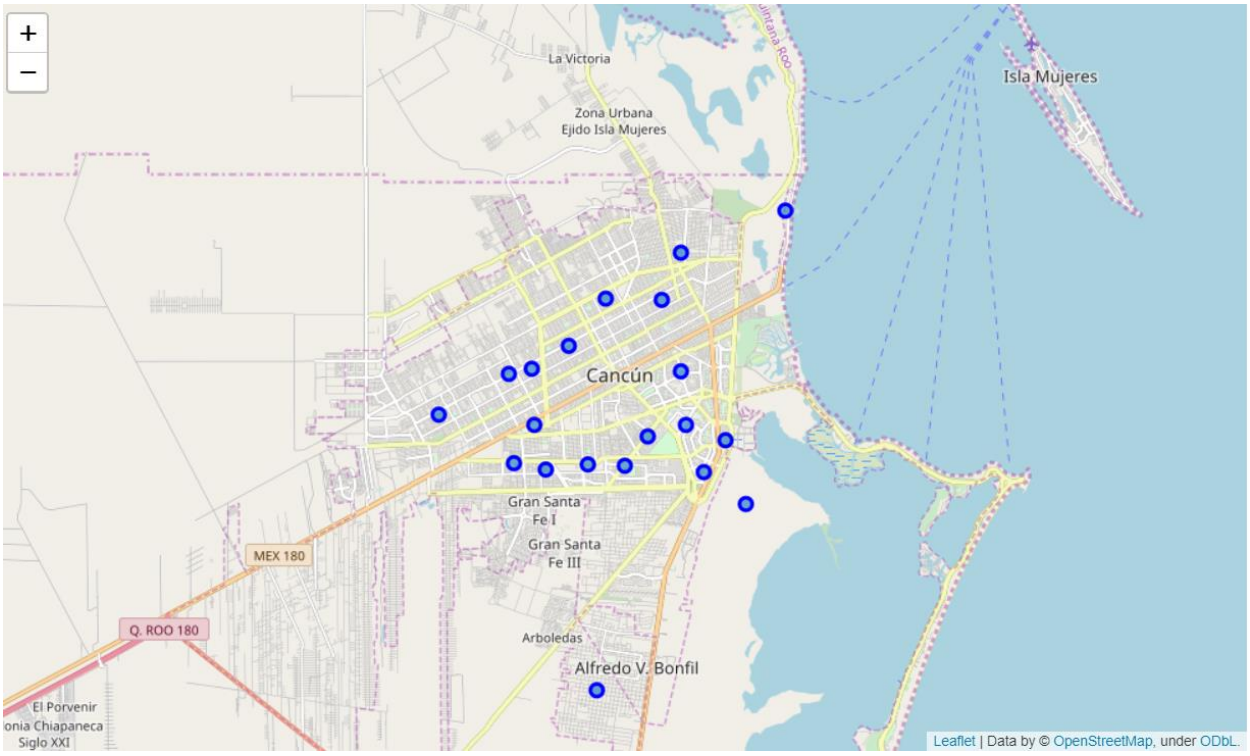
## 3.1 Correlations Between target and independent variables



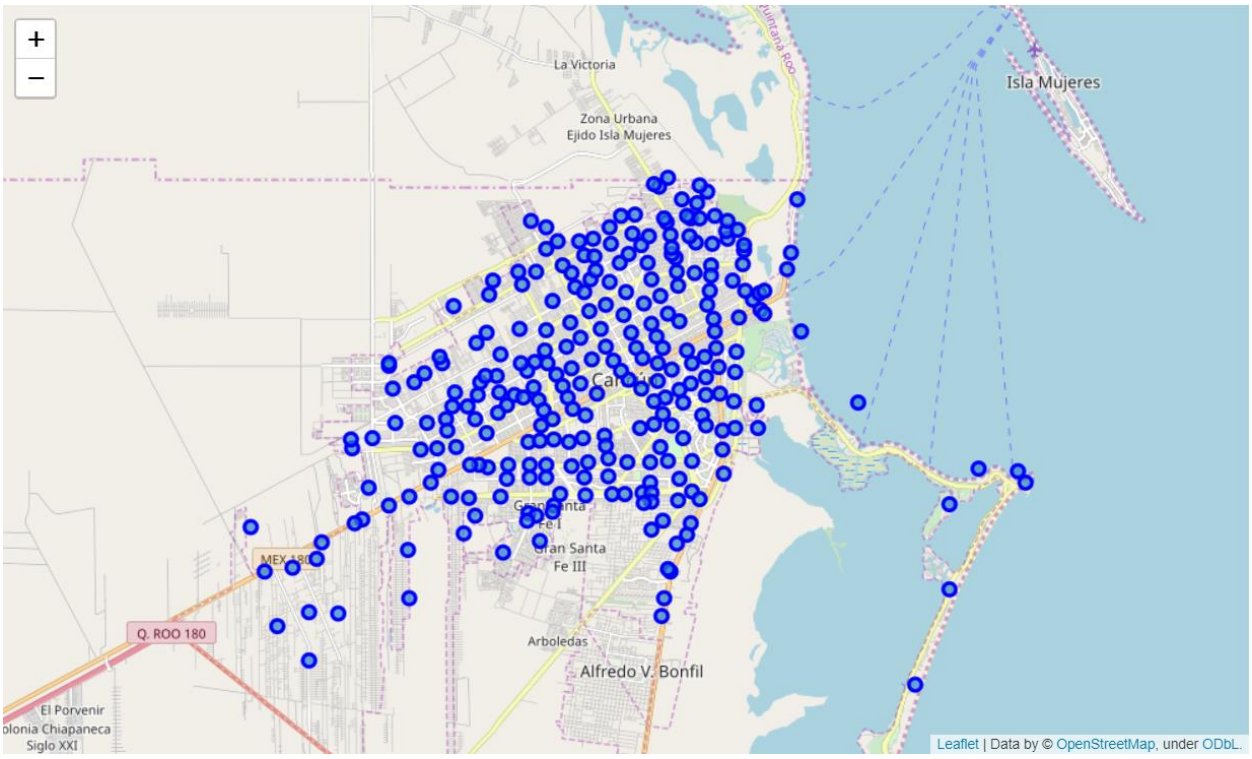
Every variable is direct related to everyone else, the same three point are out layers in every cross, these three AGEBs are the most populated, that is why they reach the highest levels in the graphs (all the variables are expressed in number of people)

Also, it is seen that in the histograms, every variable has the highest values in the first or the second bind, describing a gamma behavior

Points described by Foursquare



Points described by INEGI




# 4. Predicting Modelling

## 4.1 Linear Regression

A first linear regression was conducted, using the people with employment as the target variable and the independent variables described before. Resulting in the next equation;

| | Coeff |
|---|---|
| Edu | -1.367030 |
| Viv | 0.219954 |
| Salud | 0.146979 |
| Population | 0.338602 |

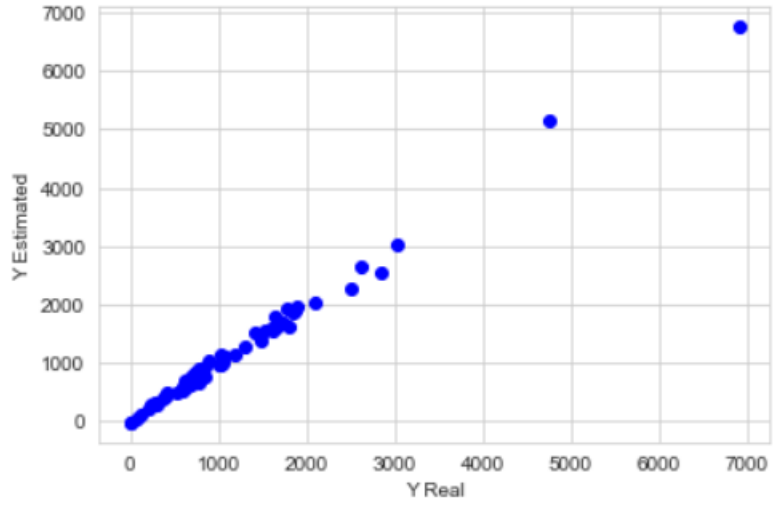
$$\hat{y} = -28.334 - 1.36\,Education + 0.219\,Housing + 0.146\,Health + 0.3386\,TotPop$$

$\hat{y}$ Prediction of people with employment in the AGEB

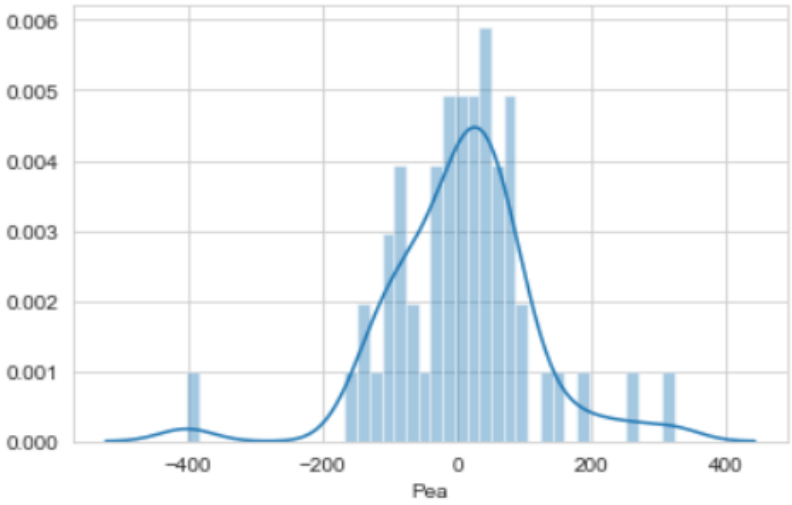
| | |
|---|---|
| Education | Holders of BA with 25 years old or above |
| Housing | No vacant houses in the zone (AGEB) |
| Health | People Health Insured (Government) |
| TotPop | Total Population |

It is described that the intercept is negative, the unemployment in the area is relatively high, therefore these factors are very to describe the employment dynamics in the zone. Something that is very interesting too is that holder with BA degree education are affected negatively for the model, it seems like there is not enough opportunities for qualified people in the city. The most marginal factor that contributes positively to the model is health insurance

Predicted Vs Real



Residual Distribution (Prediction - Real)




Metrics

R – Squared 0.9914

MAE 76.329

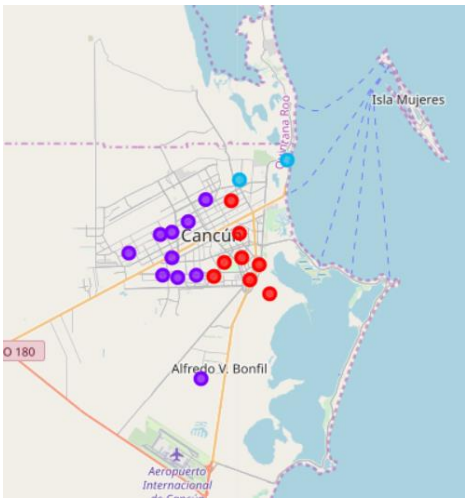
MSE 11417.394

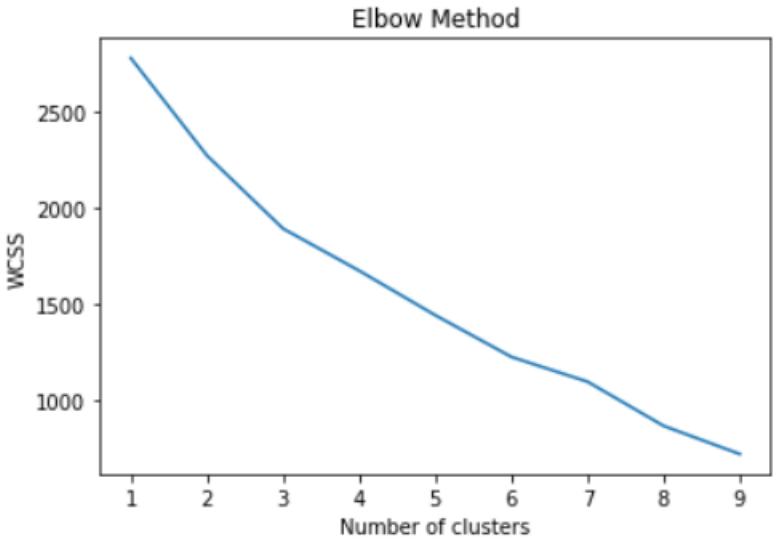
RSME 106.852

## 4.2 K means cluster analysis

A K means cluster analysis was made Foursquare with the independent variables described in the course. In order to select the optimal number of clusters, the best K selected was k=3

WCSS is the sum of all Euclidean distances from the centroids to each point in the data, the criteria for select the best was to find the k that at least grouped 5% of the data and have a reasonable WCCS score





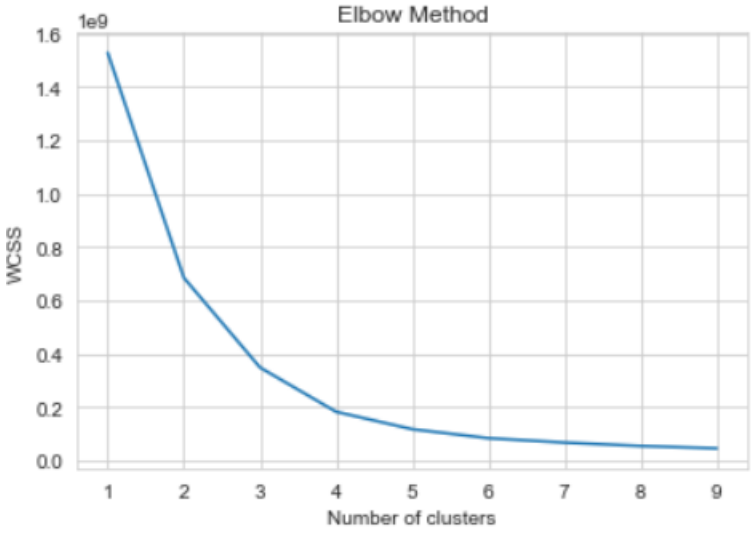

Using the frequencies of each venue, and after applying a scalar method, the 20 point were classified as follows

The method classified the venues the West size as the less frequent venues, whereas the east size is where it is more common to find residential areas. The two points in the north did not fit the criteria and they were classified in a different group

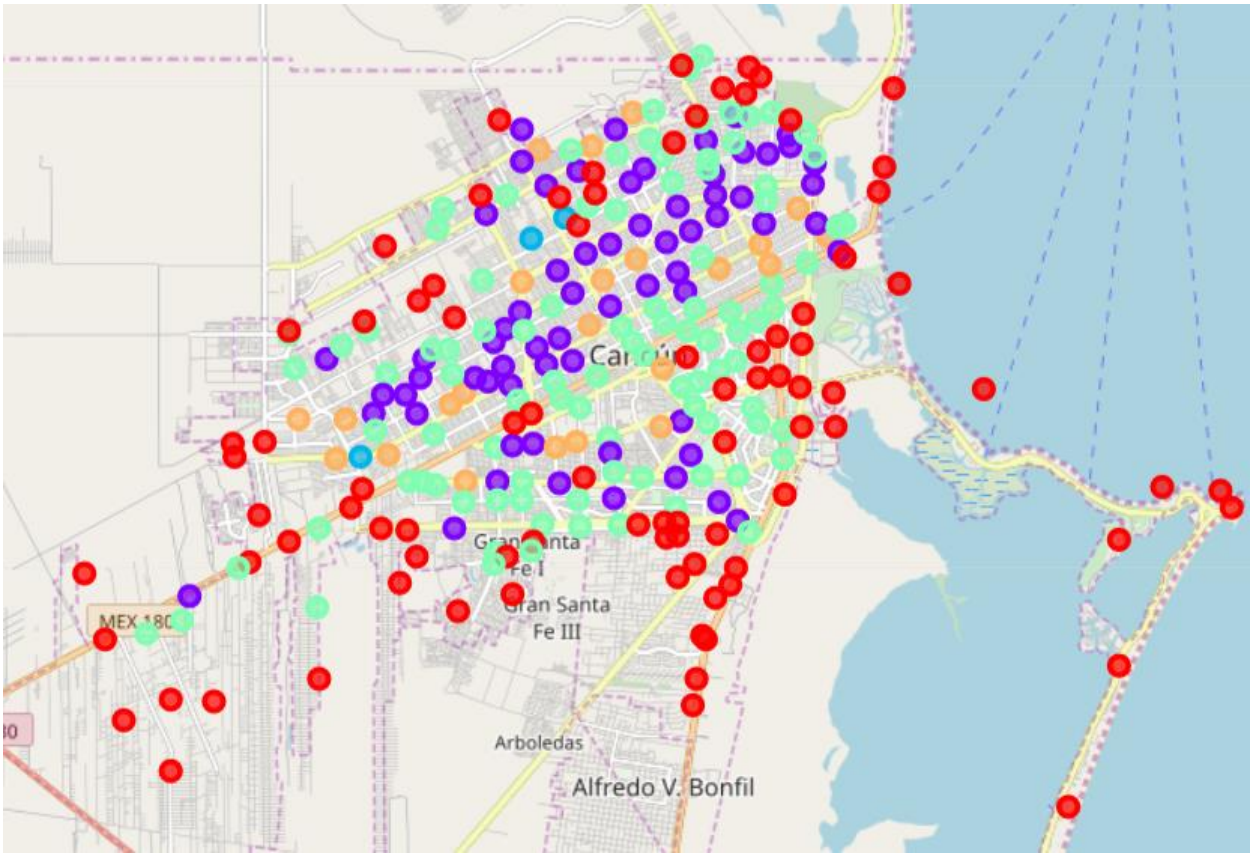
## 4.3 K means cluster analysis , INEGI data

A second cluster analysis was made using the information from INEGI, this data was grouped into 278 rows, each one describing an individual AGEB. The training information for this model was the four independent variables used in the liner regression

Once again an elbow method was run to choose the best k for the model. This time it was decided to used k=5 as the best parameter



Points classified by the model




The red dots seem to described the most far zones in the city, by each independent variables had the following basic statistics. These areas are the less populated, thus they have the lowest amount of each variable

| | Edu | Viv | Population | Cluster Labels |
|---|---|---|---|---|
| count | 89.000000 | 89.000000 | 89.000000 | 89.0 |
| mean | 13.011236 | 243.067416 | 482.741573 | 0.0 |
| std | 12.129504 | 187.021620 | 376.096746 | 0.0 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.0 |
| 25% | 4.000000 | 81.000000 | 163.000000 | 0.0 |
| 50% | 11.000000 | 213.000000 | 433.000000 | 0.0 |
| 75% | 18.000000 | 371.000000 | 781.000000 | 0.0 |
| max | 62.000000 | 747.000000 | 1285.000000 | 0.0 |

The purple dots the are mainly in the north part of the city parallel of the main avenue dividing the city in two part. This group has the following statistics. This cluster is 7 times more populated than the red dots

| | Edu | Viv | Population | Cluster Labels |
|---|---|---|---|---|
| count | 66.000000 | 66.000000 | 66.000000 | 66.0 |
| mean | 85.651515 | 1178.378788 | 3526.878788 | 1.0 |
| std | 22.535778 | 252.390160 | 489.519781 | 0.0 |
| min | 45.000000 | 789.000000 | 2533.000000 | 1.0 |
| 25% | 69.500000 | 977.750000 | 3164.500000 | 1.0 |
| 50% | 83.000000 | 1148.500000 | 3394.500000 | 1.0 |
| 75% | 97.500000 | 1304.250000 | 3927.250000 | 1.0 |
| max | 135.000000 | 1854.000000 | 4510.000000 | 1.0 |

The third cluster is where only 3 AGEBs where grouped, these are the outlayers form the linear regression. These AGEBs are highly populated with a relatively high level of education

| | Edu | Viv | Population | Cluster Labels |
|---|---|---|---|---|
| count | 3.000000 | 3.000000 | 3.000000 | 3.0 |
| mean | 441.333333 | 4270.666667 | 11953.333333 | 2.0 |
| std | 57.873425 | 784.276949 | 1886.872103 | 0.0 |
| min | 404.000000 | 3736.000000 | 10643.000000 | 2.0 |
| 25% | 408.000000 | 3820.500000 | 10872.000000 | 2.0 |
| 50% | 412.000000 | 3905.000000 | 11101.000000 | 2.0 |
| 75% | 460.000000 | 4538.000000 | 12608.500000 | 2.0 |
| max | 508.000000 | 5171.000000 | 14116.000000 | 2.0 |

The fourth cluster analysis are the green dots. They are mainly found in the south of the city, in comparison with the north part of the city they are more populated, but they have more holders of BA degrees

| | Edu | Viv | Population | Cluster Labels |
|---|---|---|---|---|
| count | 97.000000 | 97.000000 | 97.000000 | 97.0 |
| mean | 52.938144 | 743.474227 | 1969.773196 | 3.0 |
| std | 20.439349 | 196.553486 | 455.285339 | 0.0 |
| min | 18.000000 | 379.000000 | 1191.000000 | 3.0 |
| 25% | 39.000000 | 600.000000 | 1535.000000 | 3.0 |
| 50% | 50.000000 | 731.000000 | 1968.000000 | 3.0 |
| 75% | 66.000000 | 846.000000 | 2370.000000 | 3.0 |
| max | 127.000000 | 1346.000000 | 2777.000000 | 3.0 |

The last cluster, the orange dots, they are found also in the north of the city, in comparison with the second cluster they are more populated, and they hold the most people with higher education, even more than the fourth cluster

| | Edu | Viv | Population | Cluster Labels |
|---|---|---|---|---|
| count | 23.000000 | 23.000000 | 23.000000 | 23.0 |
| mean | 131.869565 | 2024.130435 | 5252.695652 | 4.0 |
| std | 45.145527 | 505.868138 | 621.181677 | 0.0 |
| min | 77.000000 | 1222.000000 | 4465.000000 | 4.0 |
| 25% | 97.000000 | 1647.500000 | 4644.000000 | 4.0 |
| 50% | 115.000000 | 2013.000000 | 5181.000000 | 4.0 |
| 75% | 156.000000 | 2258.000000 | 5643.500000 | 4.0 |
| max | 242.000000 | 2936.000000 | 6644.000000 | 4.0 |

## 4.4 Linear Regression, cluster labels

With the cluster labels, this variable was decomposed in five dummy variables (0,1) , one for each cluster, a second linear regression model was proposed, resulting in the next equation

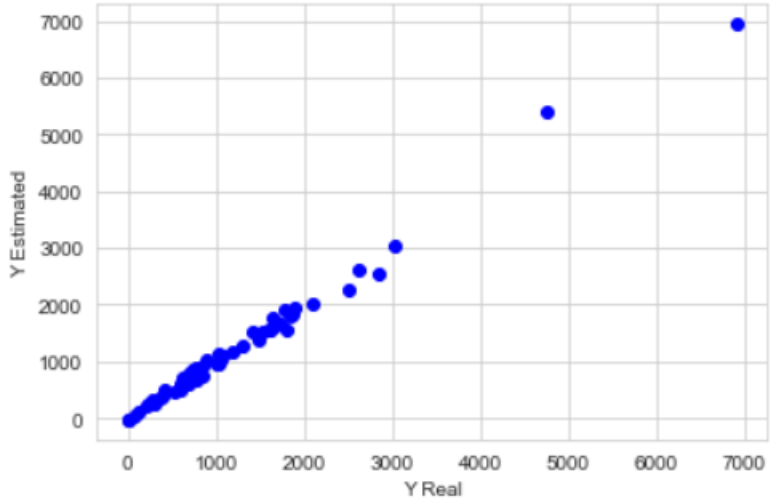
This time the intercept is positive, which means that adding the classification to the model filters the zones in the city that has a higher employment of people or vice versa

| | Coeff |
|---|---|
| Edu | -1.761617 |
| Viv | 0.217771 |
| Salud | 0.132807 |
| Population | 0.348285 |
| G1 | -98.111954 |
| G2 | -81.302152 |
| G3 | 309.624098 |
| G4 | -72.793926 |
| G5 | -57.416066 |

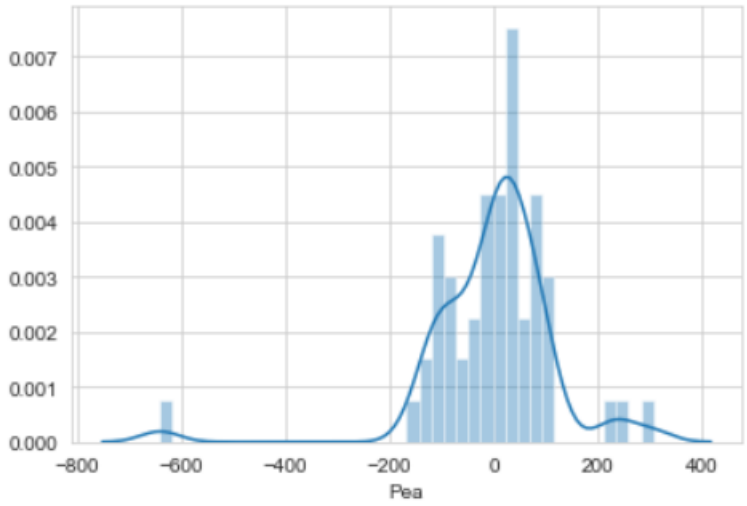
$$\hat{y} = 74.473 - 1.76\,Education + 0.217\,Housing + 0.132\,Health + 0.3482\,TotPop - 98.1\,G1 + 81.3\,G2 + 309.62\,G3 - 72.79\,G4 - 57.41\,G5$$

Every group diminishes the number of predicted number with employment target, except for the third group, which are out layers in the data

Predicted Vs Real




Residual Distribution




R-Square 0.9884
MAE 78.45685308975901
MSE 15444.490939683097
RSME 124.27586628015553

# 4. Conclusions

Based in the second regression we can concluded that the city suffers a systemic problem of unemployment, moreover, the people of the first and the second cluster of AGEBs are affected more severe. These AGEB are the ones that are more far away from the center of the city or they are in the north of the city

Furthermore, the city does not offer many opportunities for people that have a bachelor degree, the AGEBs with more people with this trait they tend to be worse than other clusters

The third cluster, with only three points, seem to be an example of success, in the map they are not near to each other, so the dynamics in these zones could be an example to develop new policies to reduce the unemployment in Cancun