WHERE IN MEXICO IS PEOPLE MORE PRONE TO DEVELOP CLINICAL DEPRESSION

Jose NERI, PhD.
December 17th, 2019

1.INTRODUCTION

Mental health in Mexico is recognized as one of the main unresolved issues within the government's health policy agenda. Recently, the subject has been gaining relevance due to the increasing prevalence of mental disorders such as depression and its correlation with other issues within the country, such as violence, inequality, and poverty. In fact, depression is quickly becoming the leading cause of disability, work disability, morbidity and mortality in Mexico. Since depression is characterized by negative changes in mood, cognition, and behavior, it does not only affect the individual's physical health but also their professional development. Therefore, decreasing their economic potential productivity. Moreover, discrimination might further prevent patients from accessing the labor market, lowering their household income and making their reinsertion into society more difficult.

It exists different types of depression, each one of them having its own symptoms and causes. For this reason, this project will only concern to clinical depression also known as depressive disorder. Clinical depression is characterized by a persistent feeling of sadness or a lack of interest in outside stimuli. The unipolar connotes a difference between major depression and bipolar depression, which refers to an oscillating state between depression and mania. Instead, unipolar depression is solely focused on the "lows," or the negative emotions and symptoms that you may have experienced.

Although there are multiple risk factors for developing depression, in 2002 a Mexican National Comorbidity Survey employed a comprehensive interview developed by the World Health Organization (WHO), to assess the epidemiological profile of mental disorders in Mexico. Among their results, they found out that several factors increased any individual's risk for depressive disorders including young age, low education, history of trauma, low socioeconomic status, negative life experiences, such as ethnic or sexual discrimination, loss of a loved one, divorce and many more.

Assuming that it is true that these factors increase the risk for depression. It could be implied that if these factors have a different level of prevalence among Mexico's top 100 municipalities, then it would be possible to group all of these municipalities in 3 clusters. Cluster A: "Municipalities with low level of risk factors that can lead to depression". Cluster B:"Municipalities with mild level of risk factors that can lead to depression. Cluster C:"Municipalities with high level of risk factors that can lead to depression".

In overall, this project is a general study targeting a multiple audience. From any individual or family looking for the place in Mexico with the best chances to live a mentally healthy life. Up to an entrepreneur seeking for the place in Mexico where their employees will be more productive along the time, as it has been demonstrated that depression is antagonistically correlated to work productivity.

2.DATA COLLECTION

In order to compare Mexico's top 100 cities it will be necessary to gather as much information about each municipality. To do so, the first step is to know which information would be helpful to the project and then, to choose the right source of data that will provide such information.

There are a number of causes that may increase the chance of depression namely: Abuse, Medications, Conflicts, Death or Loss, Genetics, Stressful Events, Social Vulnerability, Serious Illnesses, Substance Abuse. Each one of these causes can be linked to factors that can increase the chance of depression. For example, the main cause is "Stressful Events" but the factor could be a "divorce" or a "job loss". In the light of this, it must be gathered as much information about factors related to the causes described before.

Mexico's National Institute of Geography and Statistics (INEGI) performs regular surveys around the country around different subjects. Fortunately, their data is stored electronically so that anyone can have access to it either through tables or by doing an URL query through their API. By doing a research on the available information provided by INEGI, it was possible to create a table containing the different causes for depression and the different factors linked to them that are available from INEGI (Table 1).

Moreover, some factors that trigger depression is substance abuse and lack of leisure activities to reduce stress. So it will be used the Foursquare API to get for each municipality and within a 20km radius, the number of venues related to bars & liquor stores. Evidently, these establishments cannot be blamed for the amount and kind of substances their customers consume. However, they can be compared to the number of leisure & recreational related venues such as parks and gym in the same radius. In this way, the idea is to have for each municipality, a ratio

ratio=(Category A Venues/Category B Venues)

Category A= Nightlife Spot-Number (Bar, Brasseur, Bar Lounge, Nightclub, Stripclub)
Category B= Outdoors & Recreation (Sports and track and field, swimming areas, bay, beaches, botanical gardens, bike paths, camping zones, parks, dog parks, fountains, gardens, farms, fields, fishing spots,hills, lakes, mountains, natural parks, play gardens, pools etc)

The hypothesis behind this ratio is that a municipality will have an increased risk of depression if it is easier to have access to alcohol, for example, than it is to have access to a park or gym. Again, the main idea is to compare the different municipalities around this ratio, not to judge the use or misuse of these venues by their customers.

Cause for Depression	Description	Linked Factors to Query from INEGI's API
Abuse	Past physical, sexual, or emotional abuse can increase the vulnerability to clinical depression later in life.	People admitted for breach of trust to state prisons (common law)
		People admitted sexual abuse to state prisons (common law)
		Discrimination
Conflicts	Depression in someone who has the biological vulnerability to develop depression may result from personal conflicts or disputes with family members or friends	Workplace Conflicts
		Persons admitted for breach of obligations of family assistance to state prisons (common law)
		People admitted for other crimes against the family to state prisons (common law)
		People admitted gender-based violence in all its different modalities to domestic violence to state prisons (common law)
		Persons admitted for family violence to state prisons (common law)
Death or Loss	Sadness or grief from the death or loss of a loved one, though natural, may increase the risk of depression.	Natural Deaths
		Deaths by Homicide
		Deaths by Suicide
		Deaths of mother's giving birth (Rate for 100 000 newborns)
Genetics	A family history of depression may increase the risk.	Rate of recently diagnosed people with depression
		Percentage of family members above 7 that claim have felt depression once in their lives
Stressful Events	Events such as moving, losing a job or income, getting divorced, retiring or any other stressful life events.	Number of Divorces
		Percentage of Retired Employees
		People migrating out of the region
		People migrating inside of the region
Social Vulnerability	Vulnerability to multiple stressors and shocks, including abuse, social exclusion and natural hazards	Percentage of homes lacking a proper roof
		Percentage of people non affiliated to social security
		Population with access to water
		Percentage of population between 14-25 with access to school
		Percentage of population between 14-25 that has to go out from their region to go to school
Serious Illnesses	Sometimes depression co-exists with a major illness or may be triggered by another medical condition.	Rate of recently diagnosed people with anorexia or bulimia
		Population mortality breast cancer
		Population mortality prostate cancer
		Population mortality diabetes

Table 1.Causes and Factors increasing the chance of developing clinical depression

3. METHODOLOGY

In total, it was gathered for each one of Mexico's 100 most populated cities 29 features or risk factors increasing the chance of developing clinical depression. Among these risk factors 28 came from INEGI either through a table or through their API, the last risk factor was obtained from Foursquare API by calculating a ratio between Category A venues and Category B venues described in the "Data Collection" section of this paper.

All data was tabulated as follows:

Municipality Name	Risk factor A	Risk Factor B	Risk Factor C	...	Risk Factor 29th
Name 1	Value	Value	Value	Value	Value
Name 2	Value	Value	Value	Value	Value
Name 3	Value	Value	Value	Value	Value
Name 4	Value	Value	Value	Value	Value
Name 5	Value	Value	Value	Value	Value
...	Value	Value	Value	Value	Value
Name 115	Value	Value	Value	Value	Value

Table 2. Municipalities and their risk Factors increasing the chance of developing clinical depression

Following the Mexican National Comorbidity Survey,where it was shown that these factors increase any individual's risk for depressive disorders. The main objective of this study is to determine how many of these factors are affecting Mexico's 115 most populated municipalities and then, cluster the municipalities in 3 groups. Cluster A: "Municipalities with low level of risk factors that can lead to depression". Cluster B:"Municipalities with mild level of risk factors that can lead to depression. Cluster C:"Municipalities with high level of risk factors that can lead to depression". However, to be able to cluster municipalities it must be shown first, that risk factors' value is different enough within each municipality. In other words, it must be proved that there is enough variance between the risk factors' value for each municipality.

The variance test was done through the Principal Component Analysis of the 29th risk factors along the 115 municipalities (Figure 1). Results show that PC1, is the highest value Principal Component and explains almost 20% of the variance. If the PCA analysis was meant to reduce the number of dimensions before clustering, it could be said that it would be needed at least 17 risk factors to explain 90% of variance along the dataset, making it unwise to reduce the dimensions to 2 or 3 as no data clusters can be seen from the scatter plot between PC1 and PC2 (Figure 2). However, for the purpose of this study, the PCA analysis was a great way to show that in most of the risk factors, expressed in here from PC1 to PC28, there is enough variance between municipalities. Which means that risk factors' values are different enough between each municipality and therefore, it is possible to find for example, municipalities with low values of risk factor A, and municipalities with high values of risk factor A. Hence, it is possible to cluster municipalities in at least 2 groups.

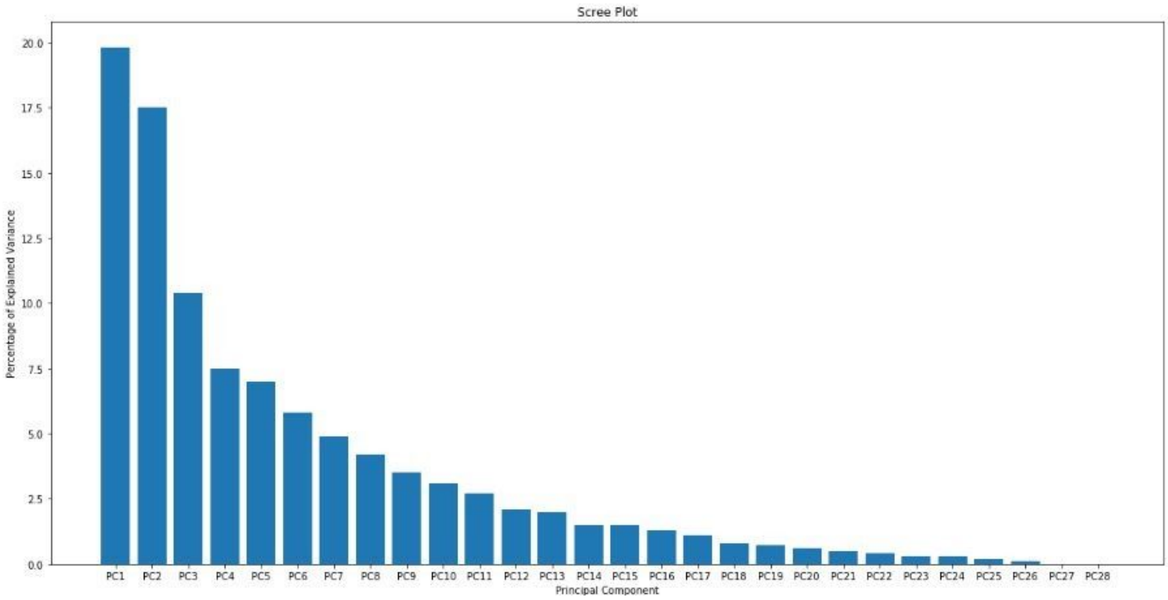

Figure 1. Scree Plot of the Principal Component Analysis done over the risk Factors increasing the chance of developing clinical depression in Mexico's top 100 municipalities. [y_{axis}=%of explained variance, x_{axis}=Principal Component]

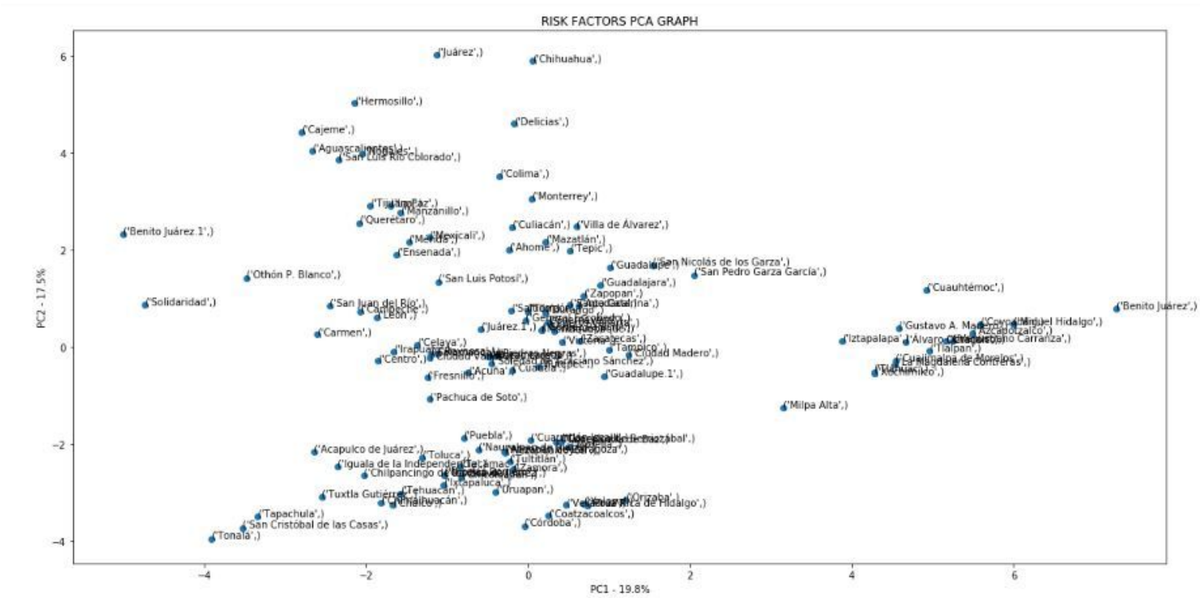

Figure 2. Scatter plot of the Principal Components with most variance PC1 and PC2. [y_{axis}=PC2, x_{axis}=PC1]

Before clustering, it was in interesting exercise to express the PCA loading scores in terms of risk factors in order to know which risk factors induce to more variance between the 115 municipalities (Table 3; Figure 3). For example, PC1, which has the highest variance among the principal components, is actually the variance of the amount of people migrating inside the State where the municipality is located. This implies that the values for this risk factor between all municipalities were the most different when compared to their mean. This also implies that risk factors with higher loading scores will be the key to find significative differences between municipalities.

PCA Loading Scores for PC1	
People migrating inside the State	0.34520
Total Mortality by breast cancer	0.314364
Percentage of population between 14-25 that has to go out from their region to go to school	0.308595
Percentage of homes with internet access	0.282475
Total Mortality by Prostate Cancer	0.273816

Table 3. Top 5 Loading Scores for PC1 as a result from the Principal Component Analysis. Results show which risk factors have more variance among the different municipalities.

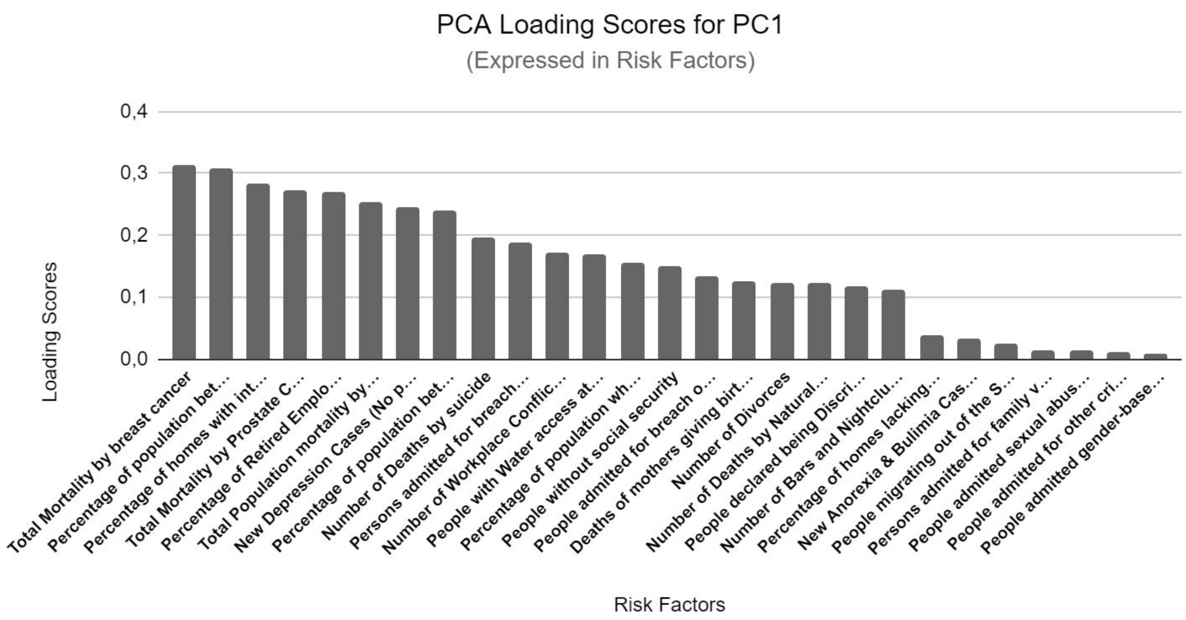

Figure 3. Bar Chart of the Loading Scores for PC1 resulting from the Principal Component Analysis. Results show which risk factors have more variance among the different municipalities. [y_{axis}=Loading Score, x_{axis}=Risk Factors]

Now that it is certain that there exist enough variation between municipalities for most risk factors(features), it is possible to proceed to the next step, which is to cluster municipalities upon their risk factors presence.

For the clustering, it could be possible to do a k-means clustering directly to the current dataframe including 29 risk factors (features) and 115 cities. However, as it was seen during the PCA analysis, some risk factors have more variance than other risk factors and therefore, if the k-means clustering was done directly on all of the 29 risk factors, the risk factors with more variance would have more weight during clustering. Moreover, values between the different risk factors are very different in scale within each other (Table 4). For example, values for some risk factors are around 0.05 and some others around 150. Which means that the net variance for values around 0.05 are smaller than the one around 150 and this could lead to an underestimation of risk factors having a smaller scale during clustering. Hence, if the dataframe was clustered as it is, not all of the risk factors would be equally represented in the cluster.

Municipality Name	Risk factor A	Risk Factor B	...	Risk Factor 29
Name 1	Value type 1	Value type 2	Value type 3	Value type 4
Name 2	Value type 1	Value type 2	Value type 3	Value type 4
Name 3	Value type 1	Value type 2	Value type 3	Value type 4
Name 4	Value type 1	Value type 2	Value type 3	Value type 4
...	Value type 1	Value type 2	Value type 3	Value type 4
Name 115	Value type 1	Value type 2	Value type 3	Value type 4
	Total Value type 1	**Total Value type 2**	**Total Value type 3**	**Total Value type 4**

Table 4. Municipalities and their risk Factors increasing the chance of developing clinical depression

For this reason, a couple of extra steps are needed for data harmonization in order to make sure every risk factor is equally represented during clustering. The first step consists in taking the sum of all values per risk factor (sum A), then divide the municipality values for each risk factor by sum A and multiply the result by 100. By doing so, all values in the dataframe will be in the percent scale 0-100% minimizing the risk of risk value underestimation (Table 5).

Municipality Name	Risk factor A	Risk Factor B	...	Risk Factor 29
Name 1	Value type (%)	Value type (%)	Value type (%)	Value type (%)
Name 2	Value type (%)	Value type (%)	Value type (%)	Value type (%)
Name 3	Value type (%)	Value type (%)	Value type (%)	Value type (%)
Name 4	Value type (%)	Value type (%)	Value type (%)	Value type (%)
...	Value type (%)	Value type (%)	Value type (%)	Value type (%)
Name 115	Value type (%)	Value type (%)	Value type (%)	Value type (%)
	Total Value (100%)	**Total Value (100%)**	**Total Value (100%)**	**Total Value (100%)**

Table 5. Municipalities and their risk Factors increasing the chance of developing clinical depression

As a reminder, the main objective of this study is to determine in which municipalities an individual is more prone to develop clinical depression. To know this, it is necessary that municipalities are compared by their level of risk factors. Given the fact that some municipalities have for example, high risk factor A and B, but low risk factor C, and that some other municipalities have mild risk factor B and C and low risk factor A. It would be difficult to interpret the cluster since it isn't known which one of the risk factors has more impact in triggering depression. Let's imagine that hypothetically the k-means clustering yields 2 clusters. Cluster 1 has high Risk Factor A,B,C,D. Cluster 2 has high Risk Factor E,F,G,H. Then, how could it be possible to know which cluster represents municipalities where individuals are more prone to develop depression? Is it cluster 1 or 2 ? If for instance, Cluster 1 represents municipalities with high "divorce rate" and Cluster 2 represents municipalities with high "victims of theft rate" then without any further psychological studies, which are far from the intention of this study, it would be difficult to say that people are more prone to develop clinical depression in municipalities with higher divorce rate than they are in municipalities with high theft rate. As a result, clustering the 29 risk factors along the 115 municipalities wouldn't answer the initial question about which municipalities is people more prone to develop clinical depression.

To answer the initial question it would be necessary to perform a second operation to the dataframe in order to give an unique meaning to the k-means clustering. To do so, it will be calculated an "Estimated Risk Factor Index" for each municipality, this Index is the sum of all risk factors for each municipality. Since values for each risk factor were previously harmonized within the percentage scale, it is possible to sum directly all risk factors values for each municipality. As a result, each municipality will have an unique "Estimated Risk Factor Index" that represents all risk factors that contribute to depression development. It wouldn't matter anymore if municipality 1 has high risk factor A and low risk factor B and municipality 2 has low risk factor C and high risk factor A. The "Estimated Risk Factor Index" will take into account the contribution of each risk factor in one single figure, making it possible to compare the Index of all 115 municipalities. At the end, it would be clear that if municipality 1 has for example, an Index of 100 (Sum of all risk factors values =100) and municipality 2 has an Index of 50 (Sum of all risk factors values =50), then municipality 1 is certainly a place with higher risk values that may lead for depression, hence, municipality 1 is also a place where people is more prone to develop clinical depression.

4. RESULTS

Once obtained the "Estimated Risk Factor Index" for each municipality, it was possible to perform the k-means clustering. Given the simplicity of the dataframe, it could have been possible to choose as many clusters as desired, in this case, k-means was set to do 3 clusters. Cluster labels were added to the risk factor dataframe as well as their classification description (Table 6) and plotted over the map of Mexico (Figure 4).

Municipality Name	Estimated Risk Factor Index	Cluster Value	Classification	Risk factor A	Risk Factor B	...	Risk Factor 29
Name 1	Value	0	Low Risk	Value	Value	Value	Value
Name 2	Value	1	Mild Risk	Value	Value	Value	Value
Name 3	Value	2	High Risk	Value	Value	Value	Value
Name 4	Value	0	Low Risk	Value	Value	Value	Value
...	Value	1	Mild Risk	Value	Value	Value	Value
Name 115	Value	2	High Risk	Value	Value	Value	Value

Table 6. Municipalities and their risk Factors increasing the chance of developing clinical depression. ***Label 0 = Municipalities with "Estimated Risk Factor" below 20. Also known as "Low Risk Factor Index". Label 1= Municipalities with "Estimated Risk Factor" between 20 and 30. Also known as "Medium Risk Factor Index". Label 2= Municipalities with "Estimated Risk Factor" above 30. Also known as "High Risk Factor Index".***

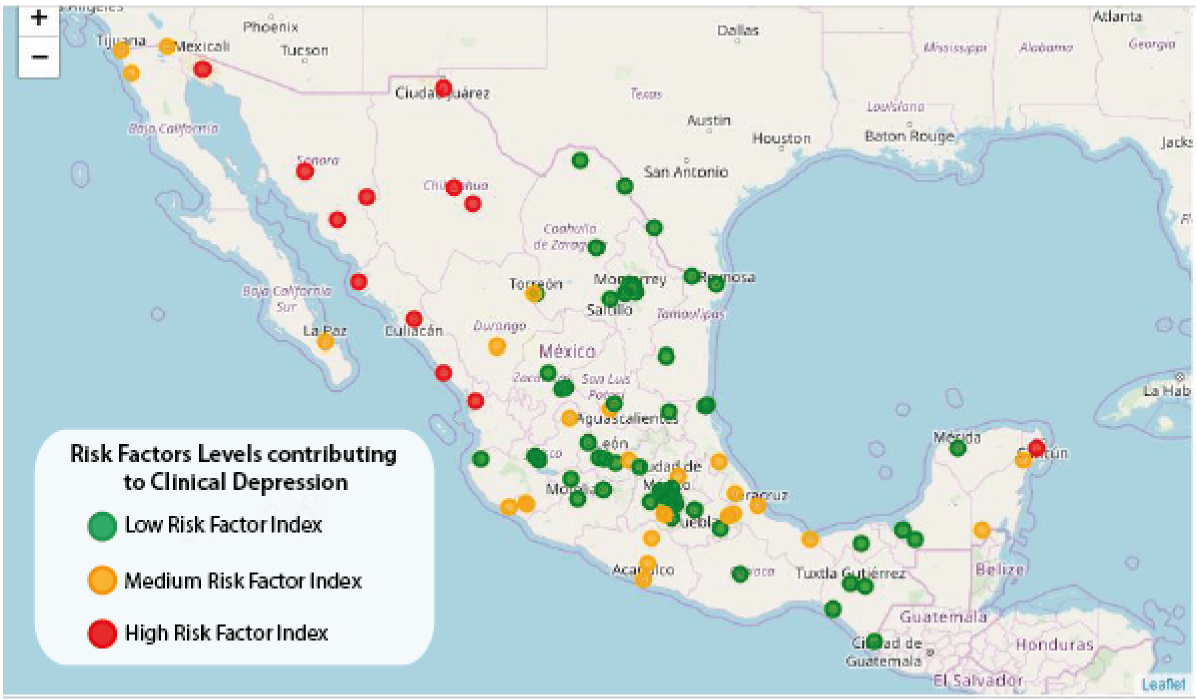

Figure 4. Mexico's top 115 Municipalities clustered upon their risk Factors increasing the chance of developing clinical depression

5. DISCUSSION

From the map it can be seen how Mexico's top 115 Municipalities are clustered upon their risk factors' presence. Although dots are scattered along the country, it is already visible some hot spots particularly in the west side of mexico, where all analyzed municipalities from Chihuahua, Sinaloa and Sonora are classified as places with High Risk Factor Index. Some major municipalities in these states are Ciudad Juarez, Culiacan, Hermosillo, Tepic and Mazatlan. Another interesting hotspot is the state of Quintana Roo, where its main municipalities cover cities like Cancun (high risk factor Index), Chetumal and Playa del Carmen (medium risk factor index). Even though this study is limited to 29 Risk Factors, it is a good background for a further and more comprehensive analysis on the risk factors' values in these municipalities. It could be important to notice, for instance, which are the risk factors that are more present in these municipalities and if they have something in common. Taking a quick look into the dataset, it is visible that in Ensenada, for instance, it is a municipality where there are more bars and nightclubs than outdoor and recreational venues. Or Aguascalientes, another municipality with high risk values located in the center of Mexico, it is a place with an increasing migration out of the municipality. As it can be seen, although this a general study, it is a good beginning to highlight some municipalities with a heavy presence of risk factors increasing the chance to develop depression.

Also, retrieved data comes from year 2018-2019. It could be a good idea to take more years into account and check through historical analysis if there's a trend around any of the 29 studied risk factors. perhaps if the same analysis were replicated in 2015-2016 other results could have been found. Moreover, it would be interesting to put some efforts into trying to find out if there's a correlation between the risk factors. At a first glance, it seems possible that migration out of the state could be rated to higher criminality rates in the municipality. However, there must be other logical correlations that could be explained using any other data exploration techniques in a further study.

6. CONCLUSION

Depression is quickly becoming the leading cause of disability, work disability, morbidity and mortality in Mexico. In 2002 a Mexican National Comorbidity Survey pointed out a correlation between depression with other issues within the country, such as violence, inequality, and poverty. For this reason, there were identified 29 risk factors that increase the chance of developing depression. Furthermore, data concerning these risk factors were queried from INEGI's and Foursquare Databases in order to obtain values for these factors in Mexico's 115 most populated municipalities. After data extraction, exploration and transformation it was possible to group all of these municipalities in 3 clusters. Cluster A: "Municipalities with low level of risk factors that can lead to depression". Cluster B:"Municipalities with medium level of risk factors that can lead to depression. Cluster C:"Municipalities with high level of risk factors that can lead to depression". The resulting cluster labels were plotted in a map of Mexico where the results show that the states of Sonora,Chihuahua and Nayarit and Sinaloa are hotspots were all of their analyzed municipalities are labeled as with High Risk Factor of developing clinical depression. This study is a first and general approach towards a more comprehensive study where historical and psychological data could be added as a mean to understand any correlation between these municipalities and more importantly, if there's a correlation between risk factors or if any risk factor has more weight on depressive symptoms appearance. Finally, even if this is a general approximation, it is already a good start for any individual or family looking for the place in Mexico with the best chances to live a mentally healthy life. As well as for any entrepreneur seeking for the place in Mexico where their employees will be more productive along the time, as it has been demonstrated that depression is antagonistically correlated to work productivity.