Social development clustering: using data science to identify vulnerable neighborhoods in Mexico City

Juan Bazaldúa, BS Sustainable Development Engineering

Monterrey Institute of Technology and Higher Education, Carlos Lazo 100, Santa Fe, Álvaro Obregón, 01389 Mexico City, Mexico

April 11, 2021

Abstract

Mexico City needs to address environmental threats, resource availability, inequality, technology access, and access to entertainment to become a prosperous, resilient place to live. To help Mexico City's government pinpoint the main focus areas of their social development and resilience agenda for the 2021 – 2030 period, we performed an analysis on 10 variables of social development at a neighborhood level. We used a data science methodology to obtain information on 1076 neighborhoods through open data documents and 2 APIs, and consolidated the information in a single dataset made available to the public in the annexes of this document. We used the k-means clustering algorithm to classify the neighborhoods according to the gathered data, and found that 81% of neighborhoods have a stable social development status, only 44% of Mexico City's stable population lives in those neighborhoods, while 42% of the population has lagging social development and 14% of the population has urgent social needs. Additionally, we found that four areas that the government could focus on to increase resilience and foster prosperity in Mexico City are 1) access to healthcare, 2) bettering house quality, 3) providing free Wi-Fi access at a neighborhood level, and 4) providing access to entertainment at a neighborhood level. The intention of this paper is that the insights gathered through this research are used by policy makers to make Mexico City a more resilient and equitable place to live.

Keywords: social development, resilience, data analysis, clustering algorithm, geographical data, dataframe, machine learning.

Introduction

Cities occupy just 3% of Earth's land, however, they account for 80% of the world's GDP, around 60 to 80% of its energy consumption and 70% of its carbon emissions (UNDP, 2021). Additionally, it is estimated that 68% of the world's population will live in urban areas by 2050. Hence, addressing the necessities of people living in cities will increasingly become more complex and more important. According to the World Economic Forum (WEF, 2018), cities will face 5 main challenges in the 21st century, which we list below.

1. Environmental threats: because infrastructure is not built to withstand the increasing extreme weather events linked to climate change, such threats often result in financial loss and death.
2. Resource availability: cities need resources such as water, food, and energy to be livable, but with a growing population, urban areas must plan to become self-sustainable to avoid destroying neighboring natural ecosystems while extracting resources.
3. Inequality: the gap regarding the provision of basic resources and resilience to environmental threats between urban super-rich and poor is widening, which can destabilize society and upend any benefits of urban development if left unchecked.
4. Technology: the only way to develop resilient and environmental-friendly infrastructure that is livable and connects citizens is to use technology and smart planning within cities.
5. Governance: the general objectives of urban governance in the future should address issues of equity, liveability and sustainability for all citizens, which will require constant innovation to meet them.

In comparison with the global panorama, Mexico faces an even more complex situation, since nearly 80% of the country's population already lives in urban settlements (United Nations Department of Economic and Social Affairs Population Division, 2019). The problem with these cities is that they have grown in a disproportionate and inequitable manner, and without a long term vision that provides their inhabitants with decent living conditions (Centro Para el Futuro de las Ciudades del Tecnológico de Monterrey, 2019). In addition, Mexican experts on business and social transformation believe that a key strategy to foster prosperity in Mexico is to turn cities into sustainable, fun, and innovative hubs to attract and retain talent in the country (Sobrino, Garrocho, Graizbord, Brambila & Aguilar, 2015).

Mexico City exemplifies the need to address all the previously mentioned issues. Mexico City is the largest city in the Western Hemisphere, with more than 20 million inhabitants. However, a large percentage of this population lives in extremely vulnerable conditions: there is proliferating informal employment, a lag in infrastructure, strong social inequality, and the city's geographical conditions make it highly susceptible to seismic hazards and flooding (Resilient Cities Network, 2016). Aware of these challenges, Mexico City's government has created a resiliency strategy focusing on regional coordination, resource availability, mobility and innovation. Although this plan encompasses a global vision to create an equitable society, its success will require the engagement of multiple stakeholders and researchers, as well as the use of modern digital technologies (CDMX Resilience Office, 2016).

Recently, a revision of Mexico City's resilience strategy proved that its implementation still requires granularization and incorporating learning systems that allow the government to take dynamic

responses to how the city changes (Urban Sustainability Exchange, 2020). For this reason, the purpose of this project will be to categorize Mexico City's neighborhoods based on their average population and their performance in 5 main topics: resilience to environmental threats, resource availability, inequality, technology access, and fun. We will perform this analysis by collecting data on all of these topics and use the k-means algorithm to cluster them into 3 types of neighborhood, which will classify the urgency to improve them. By performing this analysis, we hope to provide open access, reliable information on the progress of Mexico City's resilience agenda and point out the neighborhoods that Mexico City's government should prioritize in terms of social development programs.

Data description

Measuring Mexico City's performance on the selected topics requires complex information that comes from multiple sources. For convenience, Table 1 summarizes all indicators selected to evaluate each topic, a description of what the indicator measures and its scale, as well as the data source where it is taken from. All data sources are cited in the references of this document.

Table 1. Evaluation areas, indicators, and data sources

Evaluation area	*Indicators*	*Description*	*Data source*
Demographics	Population	Total population of permanent residents of a neighborhood, measured in natural numbers.	SIDESO (2010)
Resilience to environmental threats	Housing Quality Index	Indicator that measures the overall material, quality, and available space per inhabitant of the houses in one neighborhood, measured from 0 to 1, where 1 means the best resilience.	SIDESO (2010)
Resource availability	Sanitary Index	Indicator that measures the availability of drinking water, sewage systems and toilets in the houses of a neighborhood, measured from 0 to 1, where 1 means the best access to water and sanitation.	SIDESO (2010)
	Energy Access Index	Indicator that measures the overall access to Mexico City's electric system of the houses in one neighborhood, measured from 0 to 1, where 1 means the best access to electricity.	SIDESO (2010)
Inequality	Education Gap Index	Indicator that measures the overall literacy and school level of the inhabitants of a neighborhood,	SIDESO (2010)

		measured from 0 to 1, where 1 means the greatest literacy and school level.	
	Health Access Index	Indicator that measures the overall access to social security and health services of the inhabitants of a neighborhood, measured from 0 to 1, where 1 means the best access to social security and health.	SIDESO (2010)
	Social Development Index	Indicator that measures the overall performance of a neighborhood regarding the housing quality index, health access index, education gap index, durable goods index, sanitary index, and energy access index. It is measured from 0 to 1, where 1 means the greatest social development.	SIDESO (2010)
Technology access	Durable Goods Index	Indicator that measures how many of the basic home technology devices (telephone, television, refrigerator and washing machine) that the homes in a neighborhood have, measured from 0 to 1, where 1 means the most devices.	SIDESO (2010)
	Wifi Access Score	Indicator that measures the free Wi-Fi access points per capita within a neighborhood. This variable is normalized by dividing the access points per capita of each neighborhood by the maximum value of the variable within the dataset to be measured from 0 to 1, where 1 means a better access to Wi-Fi.	ADIP (2021) – A ADIP (2021) – B
Fun	Fun Score	Indicator that measures the venues per capita within a neighborhood. This variable is normalized by dividing the venues per capita of each neighborhood by the maximum value of the variable	Foursquare API

		within the dataset to be measured from 0 to 1, where 1 means a better access to entertainment.	

Methodology

The methodology used for the data analysis presented in this paper follows 5 main steps, presented in order in Figure 1. In this section, we will describe what was done in each of the steps of our methodology, justify the parameters with which we acquired some of the data, and explain our machine learning model.

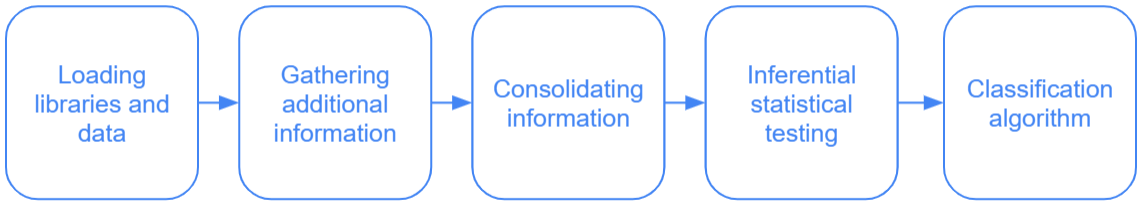

Figure 1. Methodology for data gathering and analysis. Source: Made by author.

Loading libraries and data

The objective of this section was to set up a Jupyter Notebook with a Python 3.8 environment, load the libraries necessary for our data analysis, and import the datasets available for download online. To set up the environment, we used Jupyter Notebook version 6.1.4, initialized through Anaconda Navigator version 4.9.2. Once the notebook and python environments were created, we imported the libraries listed in Table 2, which provides additional information on the way we used each library.

Table 2. Libraries used for data analysis

Library	*Use in analysis*
Pandas	Importing and exporting files in Excel and CSV format to Python, and handling them in a dataframe format.
Numpy	Performing mathematical operations with arrays and dataframes.
Geopy Nominatim	Searching for coordinates of places based on their name in string format.
Folium	Creating interactive geographical maps.
Requests	Making requests of information to the Foursquare API.
Json	Handling data extraction from the Foursquare API, translating it from Json format to strings compatible with dataframe format.
Plotly Express	Creating interactive data visualizations from statistical analysis.
Plotly Graphical Objects	Combining data visualizations created through the Plotly Express library.
Scikit-learn Cluster Kmeans	Performing the k-means clustering machine learning algorithm to automatically group similar neighborhoods into sets.

Once the libraries were loaded into the Notebook, we proceeded to import the data available for download in Mexico City's open data website. Each of these datasets was renamed since their original download name is in Spanish. Table 3 describes each dataset and lists their source.

Table 3. Datasets downloaded from Mexico City's open data portal		
Dataset name	Source file	Description
mexico_neighborhoods	sdi_cdmx.xlsx	Pandas dataframe containing the population all scores for the components measured in the social development index (housing quality index, health access index, education gap index, durable goods index, sanitary index, and energy access index) per neighborhood. The dataframe is composed of 1,473 rows and 10 columns, and it is available for download under the title "Índice de Desarrollo Social por colonia, barrio del Ciudad de México 2010" at http://www.sideso.cdmx.gob.mx/index.php?id=551.
wifi_c5	wific5_cdmx.csv	Pandas dataframe containing all coordinates for C5 street posts, which provide citizens with free wifi and surveillance infrastructure. The dataframe is composed of 13,694 rows and 12 columns, and it is available for download at https://datos.cdmx.gob.mx/dataset/ubicacion-acceso-gratuito-internet-wifi-c5/resource/7cc6ff61-6178-4ba6-b39f-1f9eb3def57b.
wifi_barrio	wifibarrio_cdmx.csv	Pandas dataframe containing all coordinates for Wifi de Barrio street posts, which provide citizens with free wifi at a neighborhood level. The dataframe is composed of 3,000 rows and 11 columns, and it is available for download at https://datos.cdmx.gob.mx/dataset/wifi-de-barrio/resource/7bb21e65-d6d1-44c3-8db0-ee91012ac2d9.

After loading the data, we renamed each column in the mexico_neighborhoods dataframe to the name of each feature in English, since the original column names were in Spanish. We also combined the wifi_c5 and wifi_barrio dataframes into a unified dataframe called mexico_wifi, containing all coordinates of all free wifi posts in the city, and gave each post an ID number according to

the type of infrastructure they belong to ('c' for C5 and 'b' for Barrio). The mexico_wifi dataframe is composed of 16,694 rows and 3 columns.

Gathering additional information

In this section, we obtained the geographical coordinates of the neighborhoods contained in the mexico_neighborhoods dataframe and the venues in each neighborhood. This step was of essential importance, since the presentation of the results relies on the geographical coordinates of each neighborhood, and the fun score is measured in terms of access to entertaining venues on a neighborhood level.

To obtain the geographical coordinates of each neighborhood, we defined a function using the Geopy Nominatim library which wrote the full address of each neighborhood using the format “Neighborhood, Borough, Mexico City, Mexico”. After applying the function to each row in the mexico_neighborhoods dataframe it returned the coordinates for 1080 neighborhoods, which is around 73% of the original neighborhoods. The coordinates for those neighborhoods were stored in two new columns in the mexico_neighborhoods dataframe, named “Latitude” and “Longitude”. Then, we plotted the remaining neighborhoods into a map centered around Mexico City to visualize the area covered by our analysis. This map can be seen in Figure 2.

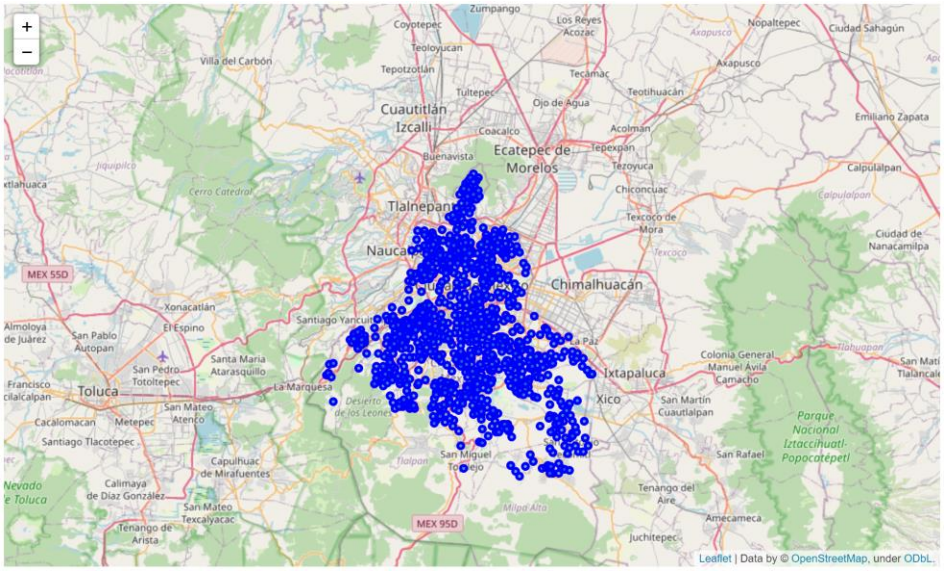

Figure 2. Initial visualization of neighborhoods in Mexico City. Source: Made by author.

After obtaining the geographical coordinates of each neighborhood, we had the necessary information to make requests to the Foursquare API. To make the connection with the Foursquare API we used the Requests library and divided the mexico_neighborhoods dataframe into 3 sub-dataframes, since the Foursquare API can only handle 500 calls per user each hour. We defined a function that searches for all venues located in a 1 km radius around the center of each neighborhood, which returned a dataframe containing the neighborhood, neighborhood latitude, neighborhood longitude, venue latitude, venue longitude, name of the venue and venue category for all venues found around

each neighborhood. The resulting dataframes were combined into a single one, named mexico_venues and composed of 66,609 rows and 7 columns.

Subsequently, we validated that the venues in mexico_venues did provide entertainment to Mexico City's Inhabitants by displaying the venue categories list. We found that most of the venues retrieved by the Foursquare API are restaurants, museums, concert halls, stores, or sports centers where citizens can do recreational activities. This confirms that these venues fall into the "fun" category, and therefore useful to measure how much access to entertainment do citizens have.

Consolidating information

In this section, we consolidated all useful information into a single dataframe called mexico_data. To do so, we counted the number of venues and the number of wi-fi access points in each neighborhood, and then assigned a score based on a normalized scale of the per capita count of both variables for each neighborhood.

To count the number of venues in each neighborhood, we used the *.groupby*() method combined with the *.count()* method of the Pandas library, which resulted in a dataframe that aggregated a count of all venues per neighborhood. We stored that information in a new dataframe called venues_grouped, and proceeded to combine it the information in mexico_neighborhoods in a new dataframe called mexico_data. Because of the way the Foursquare API works, all neighborhoods which did not return any venues in a 1 km radius were deleted from mexico_data. Hence, mexico_data included 1,076 neighborhoods, which is still roughly 73% of all neighborhoods in Mexico City. The venues per neighborhood count was stored in a new column in mexico_data called "Venue Count", then the count was divided by the population of each neighborhood to create the "Venues Per Capita" column, and finally, that column was normalized dividing all of its values by the maximum value to create the "Fun Score" column.

To count the wi-fi access points per capita, we defined a function to calculate the distance between two latitude and longitude coordinates based on the Haversine Formula, described by Veness (2020) in the article published in Movable Type Scripts. Afterwards, we used a double for loop to calculate the distance between each neighborhood and each wi-fi access point in mexico_wifi, and a conditional that increased the count of each neighborhood for each distance that was 1 km or less. The count was stored in a new column in mexico_data called "Wifi Post Count", then the count was divided by the population of each neighborhood to create the "Wifi Posts Per Capita" column, and finally, that column was normalized dividing all of its values by the maximum value to create the "Wifi Access Score" column.

After performing these processes, we consolidated all information required to perform the classification algorithm. The data consolidation process can be seen in a graphic summary in Figure 3.

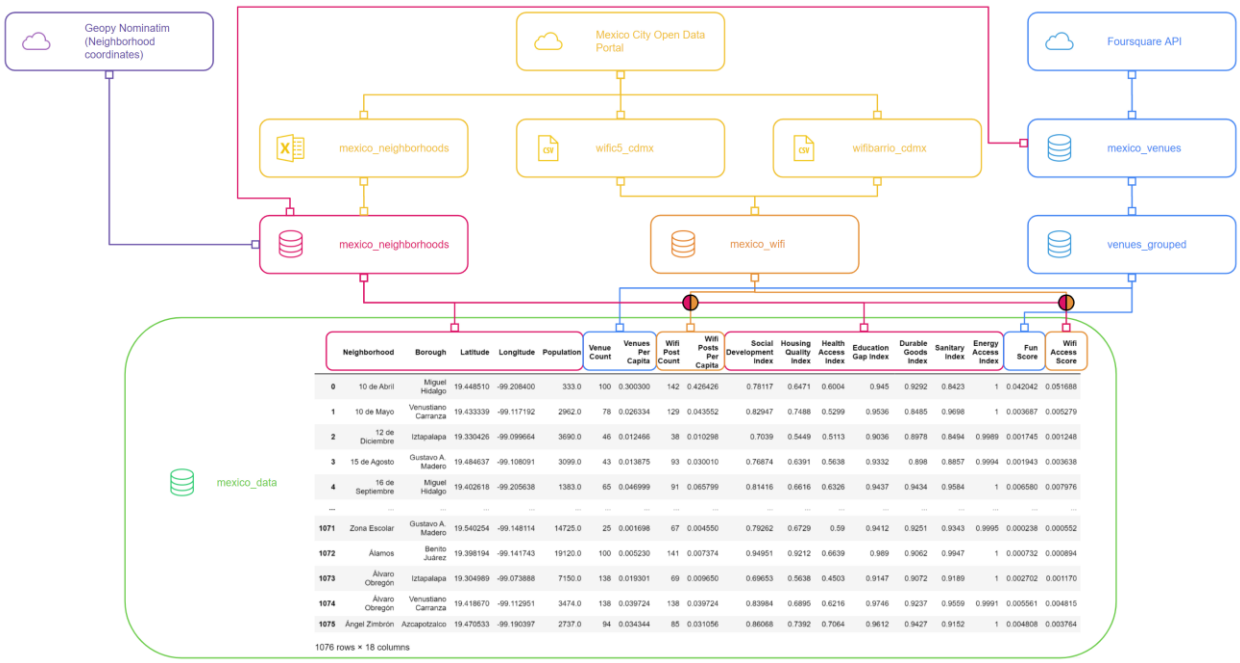

Figure 3. Data consolidation process. Source: Made by author.

Inferential statistical testing

To understand the repercussions on the statistical confidence of our analysis after the reduction of data which resulted from the use of Geopy Nominatim, the Foursquare API and the data consolidation, we calculated the margin of error under two scenarios using the number of neighborhoods and the total population that lives in the remaining neighborhoods. The results of these calculations pointed out that the data reduction produced a minimal statistical margin of error, which can be seen in Table 4.

Table 4. Statistical confidence of the analysis

Variable	*Original value*	*Value after data reduction*	*Scenario 1 margin of error (99% statistical confidence)*	*Scenario 2 margin of error (95% statistical confidence)*
Population (total permanent residents)	8,610,663	6,505,146	< 1%	< 1%
Number of neighborhoods	1,473	1,076	2%	2%

After confirming the statistical accuracy that our analysis could yield, we obtained the main statistical measurements of the 10 following variables: 1) Population, 2) Social Development Index, 3) Housing Quality Index, 4) Health Access Index, 5) Education Gap Index, 6) Durable Goods Index, 7) Sanitary Index, 8) Energy Access Index, 9) Fun Score, 10) Wifi Access Score. These measurements are presented in Table 5.

	Table 5. Main statistical measurements of social development variables									
	Population	*Social Development Index*	*Housing Quality Index*	*Health Access Index*	*Eduaction Gap Index*	*Durable Goods Index*	*Sanitary Index*	*Energy Access Index*	*Fun Score*	*Wifi Access Score*
count	1076	1076	1076	1076	1076	1076	1076	1076	1076	1076
mean	6045.674721	0.812475	0.701933	0.581607	0.945274	0.910226	0.898349	0.999553	0.00785	0.007473
std	7952.445355	0.102414	0.15372	0.088232	0.036929	0.045716	0.111879	0.002029	0.0449	0.049387
min	4	0.54906	0.3841	0.2066	0.7815	0.5685	0.3213	0.9396	2.9E-05	0
25%	1769.75	0.736195	0.57795	0.52715	0.922675	0.890675	0.85265	0.9996	0.00092	0.001034
50%	3624.5	0.808555	0.68385	0.5921	0.9479	0.917	0.93425	1	0.00251	0.00221
75%	7141.75	0.900312	0.83925	0.64085	0.975025	0.9391	0.9846	1	0.00556	0.004985
max	93364	0.99992	0.9932	1	1	1	1	1	1	1

To better understand these measurements, we produced boxplot visualizations with the Plotly Express library, which we can see in Figures 4 and 5.

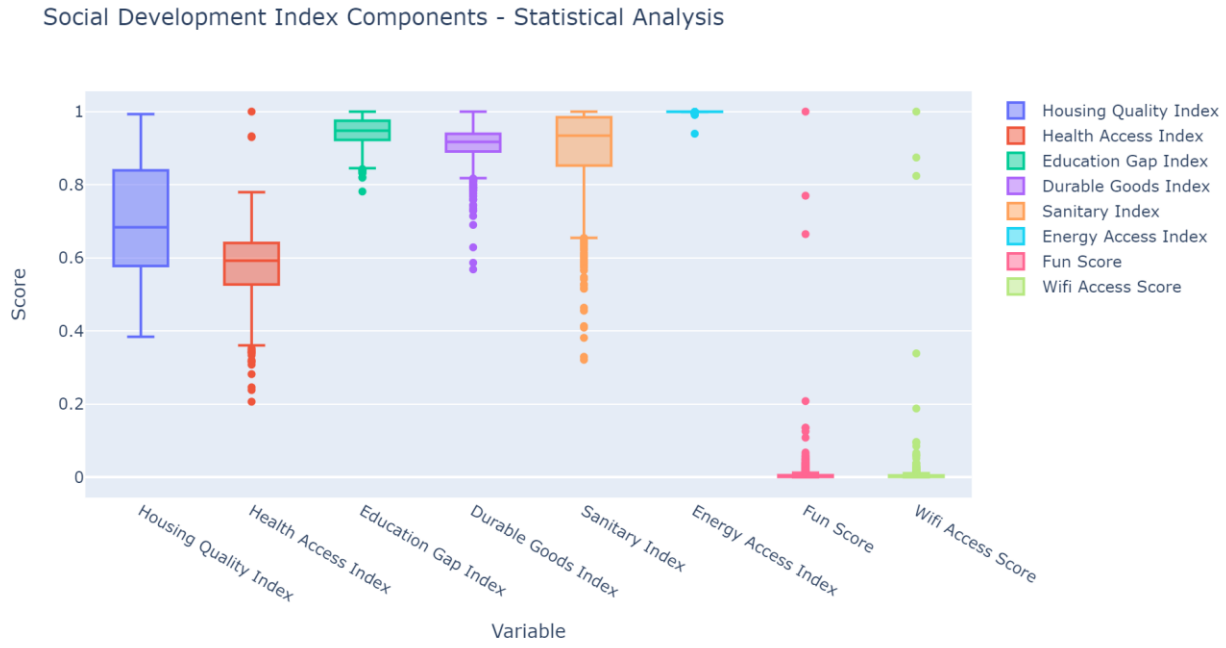

Figure 4. Statistical analysis of social development variables. Source: Made by author.

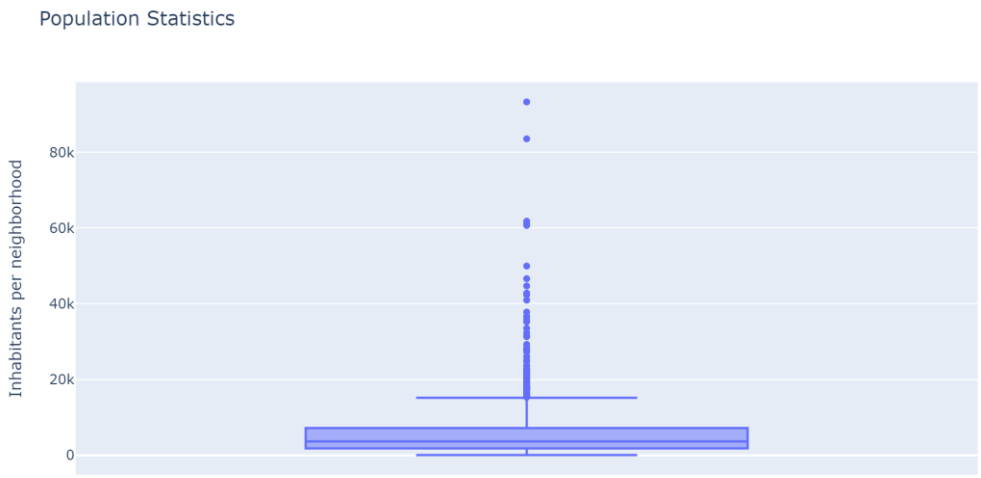

Figure 5. Statistical analysis of the population in each neighborhood. Source: Made by author.

From these statistical tests, we took the following insights:

1. Most neighborhoods have a high energy access index, and except for a few outliers, we could say that energy access is equitable in Mexico City.
2. In regards of access to entertainment and wifi, most neighborhoods have a score near to zero, which means that is a main issue in Mexico City.
3. Out of the original components of the social development index, health access is the aspect in which most neighborhoods perform the worst. This means health access in Mexico City is very difficult.
4. The housing quality index has the highest variability and a low mean, which means buildings in Mexico City need more maintenance, regulation and procurement when being erected.
5. Although it has a higher mean, the sanitary index also has a high variability, which means access to water and sanitation services is not equitable in Mexico City.
6. The durable goods index has a high mean and a relatively low variability, which could suggest a good access to technology, however, there are a lot of outliers with a much lower access. This means there are certain neighborhoods with a clear lag in technology access, further deepening the disparities between neighborhoods.
7. While the education gap has a high mean and a low variability, we must not oversee that Mexico City is one of the most densely populated cities in the world. This could point out that a great amount of people could have an education gap if it turns out that highy populated neighborhoods have lower education access.

Classification algorithm

Following the data consolidation and the exploratory analysis of the information gathered, we chose a classification algorithm to perform the main component of the analysis. According to Dabbura (2018), clustering means grouping things which are similar or have features in common, and the k-means clustering algorithm is a machine learning technique which can be applied to data that has a smaller number of dimensions, is numeric, and is continuous. Since the problem we want to solve deals with

classifying neighborhoods according to their similarities and differences with respect to 10 social development variables, and those variables are numeric and continuous, we chose to use the k-means clustering algorithm.

To set up the parameters of our k-means clustering algorithm, we took into consideration the objective of the research presented in this paper, which is to classify the neighborhoods of Mexico City into 3 categories, namely: a) neighborhoods with urgent social necessities, b) neighborhoods lagging from the best social development, and c) neighborhoods with stable social development. To fulfill such purposes, we used the Scikit-learn Cluster Kmeans library to import a pre-made function of the k-means clustering algorithm, which we configured by setting the number of clusters to 3, created a dataframe containing only the 10 variables mentioned in the inferential statistical testing section, and fit the model to that dataframe. The code of this algorithm can be seen in Figure 6.

```
In [36]:
# Seting the number of clusters:
kclusters = 3

# Creating the new dataframe:
mexico_data_clustering = mexico_data.drop(['Neighborhood','Borough','Latitude','Longitude','Venue Count',
                                           'Wifi Post Count','Venues Per Capita','Wifi Posts Per Capita'], axis=1)

# Running the k-means clustering algorithm:
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(mexico_data_clustering)

# Checking the cluster labels generated for the first 10 rows in the dataframe:
kmeans.labels_[0:10]

Out[36]: array([0, 0, 0, 0, 0, 0, 1, 0, 0, 0])
```

Figure 6. K-means clustering algorithm configured with Scikit-learn. Source: Made by author.

As we can see, the output of the clustering algorithm is an array of labels. To finish the data analysis, we stored these labels in a new column in mexico_data called “Cluster Labels”.

Results

To understand the results of our k-means clustering, we calculated the mean of the 10 social development variables analyzed for each cluster and plotted them in a comparison matrix, which can be seen in Table 6.

Table 6. Cluster social development comparison matrix

Cluster	*0*	*1*	*2*
Population	3,276.48	14,724.61	46,936.35
Social Development Index	0.812185	0.817132	0.782281
Housing Quality Index	0.699238	0.716843	0.682275
Health Access Index	0.582652	0.579997	0.55086
Education Gap Index	0.945712	0.945184	0.927045
Durable Goods Index	0.911917	0.906439	0.87134
Sanitary Index	0.897026	0.90444	0.89999
Energy Access Index	0.999537	0.999632	0.9995
Fun Score	0.009524	0.000714	0.000253
Wifi Access Score	0.009062	0.000724	0.000309

From the comparison matrix, we can interpret our results in the following way:

- Cluster 0 has the lowest average population, the highest score for 5 variables, a middle score for 3 variables and the worst score for 1 variable.
- Cluster 1 has a population close to the median of the full dataset, the highest score for 4 variables, and a middle score for 5 variables.
- Cluster 2 has the highest population, a middle score for 1 variable and the worst score for 8 variables.

Therefore, we will label the clusters in the following way:

- Cluster 0 - Stable Social Development.
- Cluster 1 - Lagging Social Development.
- Cluster 2 - Urgent Social Needs.

To better understand the results, we analyzed the total number of neighborhoods in each cluster, and the population percentage of each cluster, which can be seen in Table 7.

Table 7. Distribution of neighborhoods and population under each cluster

Cluster Labels	*Meaning*	*Total neighborhoods*	*Proportion of Neighborhoods*	*Total Population*	*Proportion of Population*
0	Stable Social Development	872	81.04%	2,857,090	43.92%
1	Lagging Social Development	184	17.1%	2,709,329	41.65%
2	Urgent Social Needs	20	1.86%	938,727	14.43%

A visualization of the neighborhoods colored by cluster label is presented in Figure7, which fulfills the objective intended for the analysis in the introduction section.

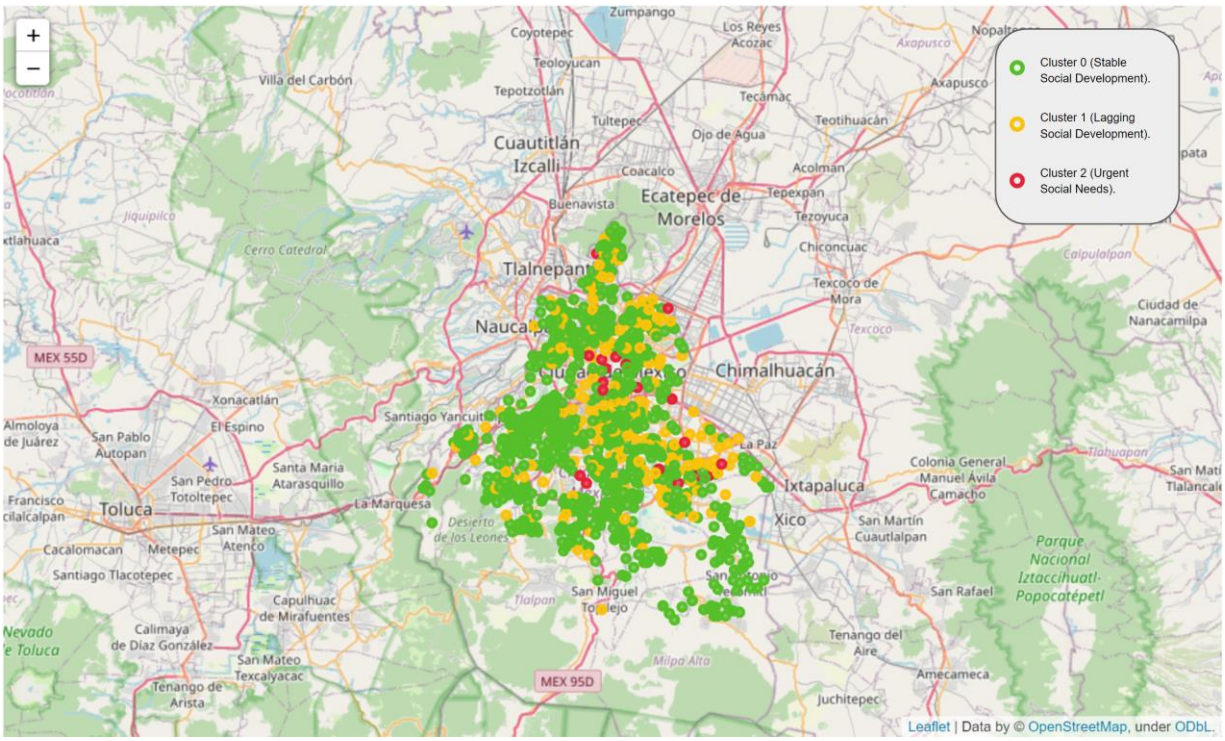

Figure 7. Neighborhoods represented by cluster label. Source: Made by author.

Discussion

As we can observe in Figure 7 and Table 7, a vast majority of neighborhoods have stable social development. However, our statistical analysis points out that while 81% of neighborhoods have a stable social development status, only 44% of Mexico City's stable population lives in those neighborhoods, while 42% of the population has lagging social development and 14% of the population has urgent social needs.

We can also see that while there are only 20 neighborhoods with urgent social needs, they have the largest average population, close to 47,000 people per neighborhood. This fact alone hinders access to sanitation, education, technology, and health in great measure.

The results presented in Table 6 show that access to entertainment and free Wi-Fi at a neighborhood level are close to none. This could be a factor contributing to population displacement and high traffic in the city, since people must move around to get to areas where they can enjoy leisure time, or access the internet for free, which is a major need exacerbated by the COVID-19 global pandemic.

In addition, neighborhoods in Mexico City lack health access and housing quality, which have a mean score of less than 0.58 and 0.72 out of 1 in all clusters. These are, by far, the lowest mean scores for any indicator measured in the social development index census. The fact that Mexico City is one of the most densely populated places in the world and that it is located in a highly seismic region with a yearly increase of climate risks, indicates that housing quality should be a top priority for the government, since the low level of resilience from houses only makes people more vulnerable and

increases the need for health access. If both aspects of social development are left unchecked, a combined catastrophe of infrastructure and health, could occur, further contributing to expenses that the government could avoid by spending in preventive measures that benefit their population.

It is important to note that the scores for all indicators of social development, except for the Wi-Fi access score and the fun score, were extracted from the last social development index census, published by Mexico City's government in 2011. According to World Population Review (2021), Mexico City's Metropolitan Area had a global population (composed of permanent residents and floating population) of 20.37 million in 2011, which grew to 21.92 million in 2021. If we considered no factors contributed to bettering any of the social development indicators published in 2011, and applied the percentages of population living in each cluster shown in the summary table, we could calculate that around 11.42 million people lived with lagging social development or urgent social needs in 2011, number that could have increased to 12.92 million in 2021. Therefore, we can see how population dynamics could point out to further revision of our calculations, and even further revision of the social development plans in Mexico City.

Conclusion

After analyzing 10 indicators regarding social development for 73% of Mexico City's neighborhoods, and performing a k-means clustering algorithm, we could classify them into 3 categories: 0) neighborhoods with stable social development, 1) neighborhoods with lagging social development, and 2) neighborhoods with urgent social needs.

We identified 20 neighborhoods with urgent social needs and 184 neighborhoods with lagging social development, an effort that pin-points the locations that Mexico City's government should prioritize in their social development plans. A list of all neighborhoods, their scores for each social development component, and their category regarding social development is provided in the documentation of this notebook.

Given that approximately 56% of people in Mexico City are currently living with lagging social development or urgent social needs, our analysis suggests that there are four major aspects of social development that Mexico City's government should focus on:

1. Access to healthcare.
2. Bettering house quality.
3. Free wifi access at a neighborhood level.
4. Access to entertainment at a neighborhood level.

Further research is required to increase accuracy of the results, and a second paper investigating social development clustering should use the latest social development index census data, which is yet to be published by the government. Regardless of this observation, our intention is that Mexico City's government, or anyone working on public policy regarding social development, could use our full dataset to focus efforts on reinforcing Mexico City's resilience plan, since we are convinced that collective efforts from all sectors will be the key to address social development in urban areas, and create more equitable and resilient places to live.

References

ADIP (2021) – A. Ubicación de puntos de acceso gratuito a internet WiFi vía infraestructura C5. Retrieved March 20, 2021, from https://datos.cdmx.gob.mx/dataset/ubicacion-acceso-gratuito-internet-wifi-c5

ADIP (2021) – B. Wifi de Barrio. Retrieved March 20, 2021, from https://datos.cdmx.gob.mx/dataset/wifi-de-barrio

CDMX Resilience Office (2016). CDMX Resilience Strategy. Retrieved March 22, 2021, from https://resilientcitiesnetwork.org/downloadable_resources/Network/Mexico-City-Resilience-Strategy-English.pdf

Centro Para el Futuro de las Ciudades del Tecnológico de Monterrey (2019). Planificación urbana y Ordenamiento Territorial. Retrieved March 19, 2021, from https://drive.google.com/file/d/13zq-pHpqR5Y7fmjFte9F8zaUKzUGttjC/view?pli=1

Dabbura, I. (2018). K-means Clustering: Algorithm, Applications, Evaluation Methods, And Drawbacks. Retrieved March 24, 2021, from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

EVALUA DF (2011). Índice De Desarrollo Social De Las Unidades Territoriales Del Distrito Federal Delegación, Colonia Y Manzana. Retrieved March 22, 2021, from http://data.evalua.cdmx.gob.mx/files/indice/ind_inf.pdf

Resilient Cities Network (2016). Mexico City's resilience journey. Retrieved March 22, 2021, from https://resilientcitiesnetwork.org/networks/mexico-city/

SIDESO (2010). Índice de Desarrollo Social de las Unidades Territoriales de la Ciudad de México. Retrieved March 19, 2021, from http://www.sideso.cdmx.gob.mx/index.php?id=551

Sobrino, J., Garrocho, C., Graizbord, B., Brambila, C. & Aguilar, A. (2015). Sustainable cities: a conceptual and operational proposal. Mexico City: Soluciones Integrales Visión Arquitectónica Sustentable S.A. de C.V. https://mexico.unfpa.org/sites/default/files/pub-pdf/Sustainable_cities_eng.pdf

UNDP (2021). Goal 11: Sustainable cities and communities. Retrieved March 19, 2021, from https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-11-sustainable-cities-and-communities.html

United Nations Department of Economic and Social Affairs Population Division (2019). World Urbanization Prospects: The 2018 Revision. New York: United Nations. https://population.un.org/wup/**Publications**/Files/WUP2018-Report.pdf

Urban Sustainability Exchange (2020). City of Mexico Resilience Strategy. Retrieved March 22, 2021, from https://use.metropolis.org/case-studies/cdmx-resilience-strategy#:~:text=The%20Resilience%20Strategy%20is%20a,strengthening%20the%20city's%20a

daptive%20capacity.&text=It%20is%20important%20to%20foster,and%20actions%20to%20build%20resilience.

Veness, C. (2020). Calculate distance, bearing, and more between latitude/longitude points. Retrieved March 24, 2021, from https://www.movable-type.co.uk/scripts/latlong.html

WEF (2018). 5 big challenges facing big cities of the future. Retrieved March 19, 2021, from https://www.weforum.org/agenda/2018/10/the-5-biggest-challenges-cities-will-face-in-the-future/

World Population Review (2021). Mexico City Population 2021. Retrieved March 22, 2021, from https://worldpopulationreview.com/world-cities/mexico-city-population

Annexes

Annex 1 – Documentation of the project

The following Github link contains the Jupyter Notebook used for the development of the research, as well as the sdi_cdmx, wifibarrio_cdmx, wific5_cdmx, mexico_neighborhoods, mexico_venues, mexico_data and mexico_data_full datasets.

Github Link: https://github.com/juanbazaldua/social_development_clusteting

Annex 2 – Jupyter Notebook

The following link provides access to the Jupyter Notebook with all its inputs, outputs and commentaries for anywone who just wishes to see the code used to create this project.

Notebook Link:
https://nbviewer.jupyter.org/github/juanbazaldua/social_development_clusteting/blob/main/Final%20Project.ipynb