Recommender System Based on Millennial Preferences in the Real Estate Sector in Mexico's Capital

Alec García Barba

April 2020

1. Introduction

Background

Given that the 'millennial' generation is known, among many other aspects, to be a generation that prefers renting over buying, and that it is slowly moving towards an age where marrying, having kids and settling are the expected steps for this age range that varies from 1980s to 1996, it is inevitable to think where they want these steps to be taken.

A study realized by the company *iCasas* [1] *showed that* there are approximately 30 million people in Mexico within the generation Y. And Mexico's capital is where 80% of young adults rent, and 20% of them chooses to buy its first home. The second in list is the State of Mexico, which is directly at north of the capital, also called CDMX, with almost five million young adults renting or buying. This region is by far the best suit to implement a data science solution to this market needs.

Problem Description

Selling houses and departments to this generation can be tough. This technology and freedom driven generation seems to be unreadable to the real estate sector, making the real estate business have a hard time renting and selling to this niche. Factors like a big garden, 3 bathrooms, or 3 story houses are not as relevant to potential customers as they used to be in other generations, like boomers, for example. A new approach to real estate's sales strategy must be reached.

Business Opportunity

I believe that data can generate a profile that fits a large portion of the financially stable person, between the ages 25 – 37 approximately, and generate a list of neighborhoods that match this profile interests, as well as benefit its lifestyle. With this segmentation, a recommender system can be applied to the neighborhoods and finally deliver a set of neighborhoods which can be potentially profitable to the sector.

This project aims to find the best neighborhoods for this particular generation section profile, in order to seek business opportunities in the real state sector.

Previous analysis

A survey conducted by NAR[2] showed that 62% of millennials would rather have easy access to restaurants and shops, as well as shorter commutes than a bigger home. This survey also showed that an overwhelming 88% of the sample population believes that having amenities within walking distance increased quality of life. 70% also said that walkability, short commutes and proximity are important factors when deciding where to live. On the contrary, the majority of older generations consider that they have no problem with a longer commute and driving to amenities, if it means living in a single-family, detached home.

2. Data Acquisition and cleaning

Acquisition

At least two different datasets will be used. The first one will be a public dataset provided by government, containing all the geographical coordinates, neighborhoods and Boroughs of the city[3]. This data is vital to the project. The second dataset will be obtained via Foursquare, containing all the nearby venues of each neighborhood, obtained based on the coordinates of the first dataset. I believe these two sets will provide sufficient information about all the amenities desired by our profiled buyer, considering transportation, amusement (cultural and social), restaurants and a general sense of living costs near the neighborhood.

In case these two datasets were insufficient to compare and analyze each neighborhood, two additional datasets will be added to our base. One containing information regarding crime (by neighborhood and Borough)[4] , and one containing information regarding the position of every choice of public transportation.[5] This will be to obtain new insights in security, and a complete scheme of transportation options that would benefit potential customers.

Cleaning

The information of the first dataset will be filtered severely. Columns regarding the entity number, postal code, neighborhood ID, etc. will be dropped. The information of this dataset will only be the neighborhood, its Geolocation and geo shape, and the Borough.

Given the size of the capital, and the limitation on my Foursquare developer account, this set will also be filtered to the top 3 boroughs in CDMX, based on population, commuting and general popularity, these boroughs are: *Coyoacán,*

Miguel Hidalgo, and Álvaro Obregón. I will obtain nearby venues via the Foursquare API and focus primarily on public transport, then on social and cultural venues.

In case of merging security and complete transportation datasets, the security dataset would be grouped by neighborhood and counted, in order to have a numeric value representing crime in each neighborhood. Lastly, the public transport system dataset would be merged by its coordinates, and using geo shape information, each station would be assigned to a neighborhood. This is a last resource in case the foursquare API could not deliver public transportation venues as accurate as desired. For this report, only transport commutes were considered.

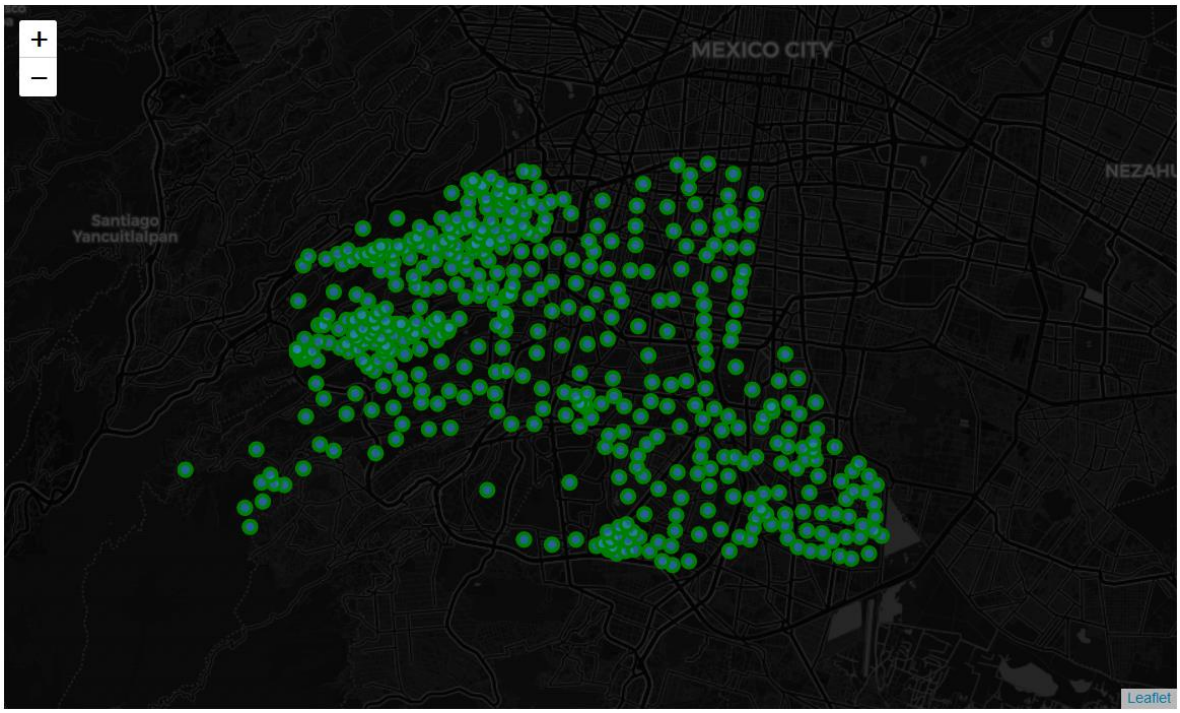
3. Methodology

The methodology for this process is really simple. Assign weights to each general type of venue, based on millennial preferences and trends, and obtain a numerical value which we can use to compare neighborhoods between them. The real technical difficulty is in the data cleaning and data wrangling to obtain a dataset that can be applied to a recommender system.

The first step was to obtain each neighborhood's centroid coordinates, the borough and name:

	COLONIA	Geo Point	ALCALDIA
0	LOMAS DE CHAPULTEPEC	19.4228411174,-99.2157935754	MIGUEL HIDALGO
1	LOMAS DE REFORMA (LOMAS DE CHAPULTEPEC)	19.4106158914,-99.2262487268	MIGUEL HIDALGO
2	DEL BOSQUE (POLANCO)	19.4342189235,-99.2094037513	MIGUEL HIDALGO
3	PEDREGAL DE SANTA URSULA I	19.314862237,-99.1477954505	COYOACAN
4	AJUSCO I	19.324571116,-99.1561602234	COYOACAN

The following map shows that there are a surprisingly lot of neighborhoods in Mexico City



This was a big factor in considering cutting the number of boroughs that would be analyzed. At the end, only three boroughs were put in this recommender system, for practical purposes: Coyoacán, Miguel Hidalgo and Alvaro Obregón.

Using Foursquare API, all the venues inside a radius of half a km were obtained. An algorithm was applied to merge our neighborhood *pandas'* data frame, to the count of each category type provided by Foursquare, instead of normalizing its value.

	neighborhood	Candy Store	Grocery Store	Playground	Coffee Shop	Restaurant	Pedestrian Plaza	Dessert Shop	Ice Cream Shop	Tea Room	...
0	19 DE MAYO	0	0	0	1	1	0	0	0	0	...
1	1RA VICTORIA	0	0	0	1	1	0	0	0	0	...
2	1RA VICTORIA SECCION BOSQUES	0	0	0	0	0	0	0	0	0	...
3	26 DE JULIO	0	0	0	2	1	0	0	0	0	...
4	2DA JALALPA TEPITO (AMPL)	0	0	0	0	0	0	0	0	0	...

5 rows × 275 columns

Since 275 columns and repeated venue categories are not suitable for a good recommender system, they were manually grouped into 5 general categories:

	neighborhood	Transport	Restaurants	Arts & Entertainment	Night Life	Fitness
0	19 DE MAYO	0	2	0	0	0
1	1RA VICTORIA	0	16	1	1	5
2	1RA VICTORIA SECCION BOSQUES	0	5	0	0	3
3	26 DE JULIO	0	1	0	0	0
4	2DA JALALPA TEPITO (AMPL)	0	2	0	0	2
5	2DA EL PIRUL (AMPL)	2	13	0	0	1

The last step was just to generate a data frame containing the weights our target customer would assign to each category

	Transport	Restaurants	Night Life	Arts & Entertainment	Fitness
0	10.0	8.0	5.0	5.0	5.0

The sum of the multiplication of each venue category count with its corresponding weight is how valuable the neighborhood is to the real estate sector in terms of saleability. A normalized column of this indicator was also calculated.

4. Results

The results of this project can be divided in two. I consider the first result to be the dataset containing a count of restaurants, commutes, night life, health amenities and cultural & entertainment venues. This dataset is a profoundly valuable asset in any real estate company.

	neighborhood	Transport	Restaurants	Arts & Entertainment	Night Life	Fitness	totalWeight
413	TLACOPAC	29	1	0	0	0	298.0
384	SAN PEDRO DE LOS PINOS	29	8	0	0	4	374.0
288	NARVARTE III	23	13	0	1	1	344.0
396	SANTA URSULA COAPA (PBLO)	23	1	0	0	0	238.0
63	CAMPESTRE CHURUBUSCO	21	8	0	0	4	294.0

The second result is the listing of neighborhoods based on different aspects of each neighborhood:

* Commuting facilities

* Restaurant count

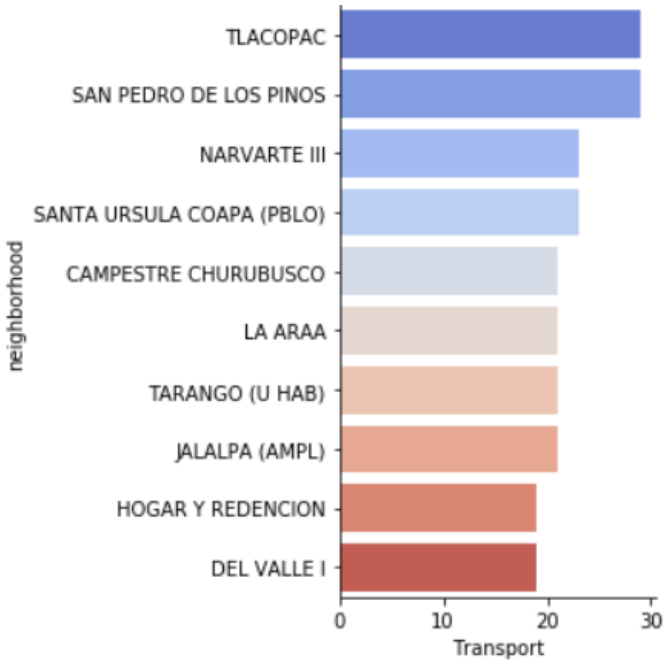
* Fitness lifestyle

* Arts, culture and Entertainment

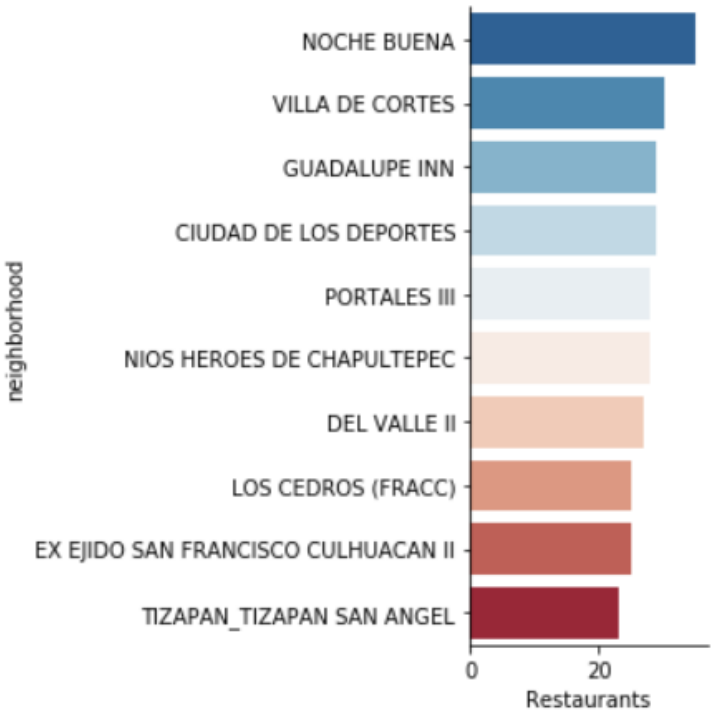
* Night Life

* Weighted Average

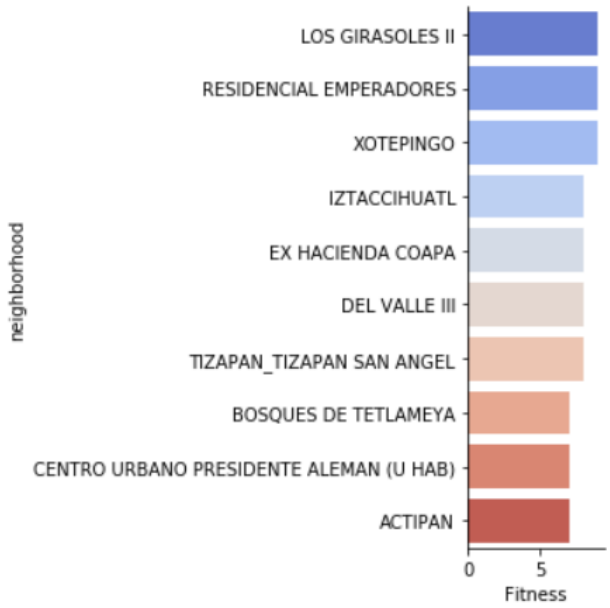
Commuting facilities



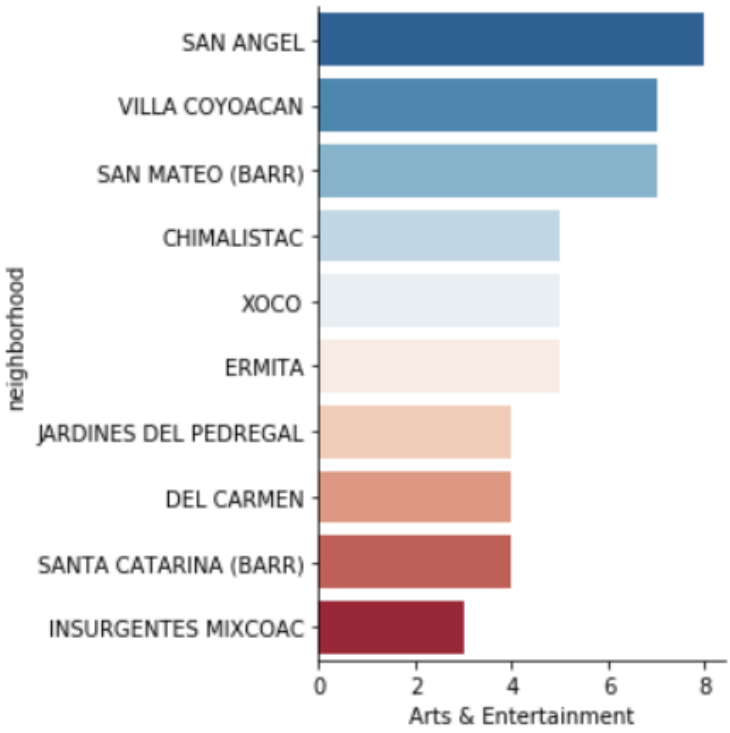
Restaurant Offer



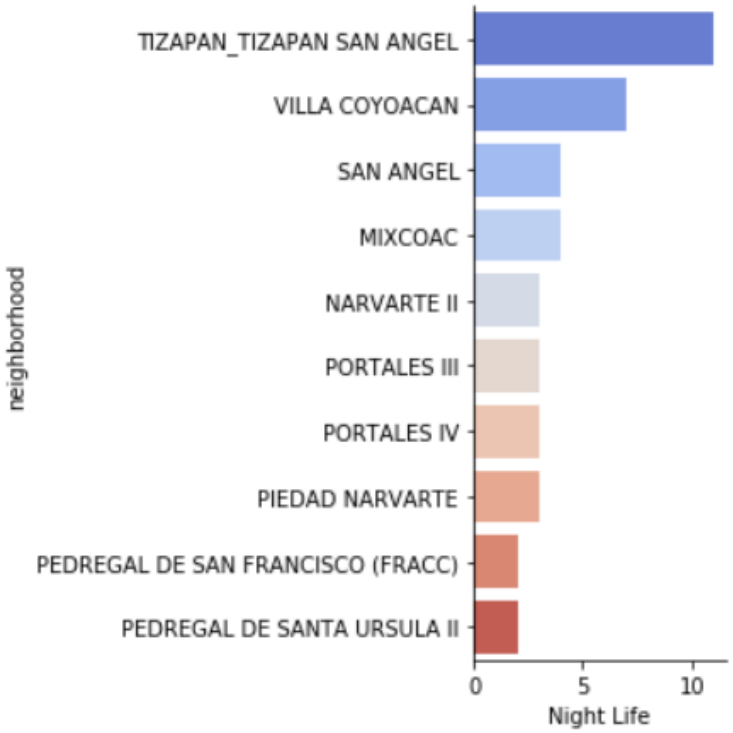
Fitness Lifestyle



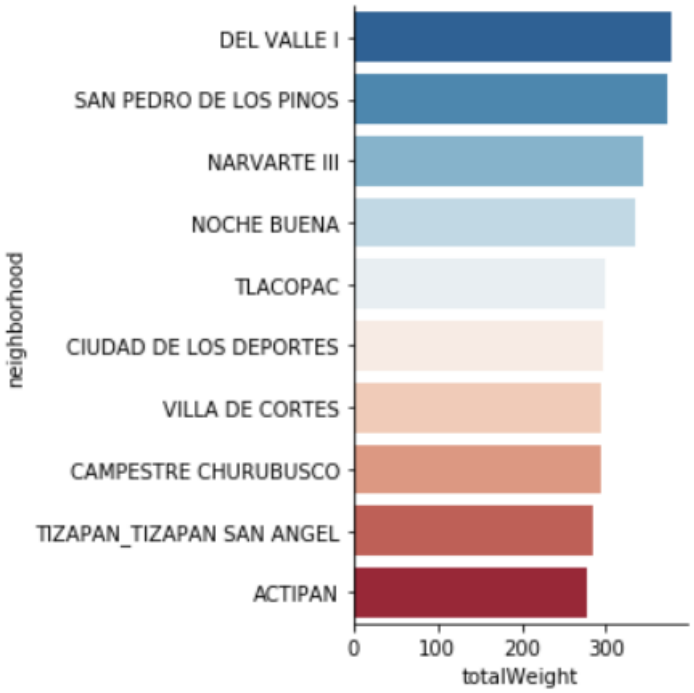
Arts, Culture and Entertainment



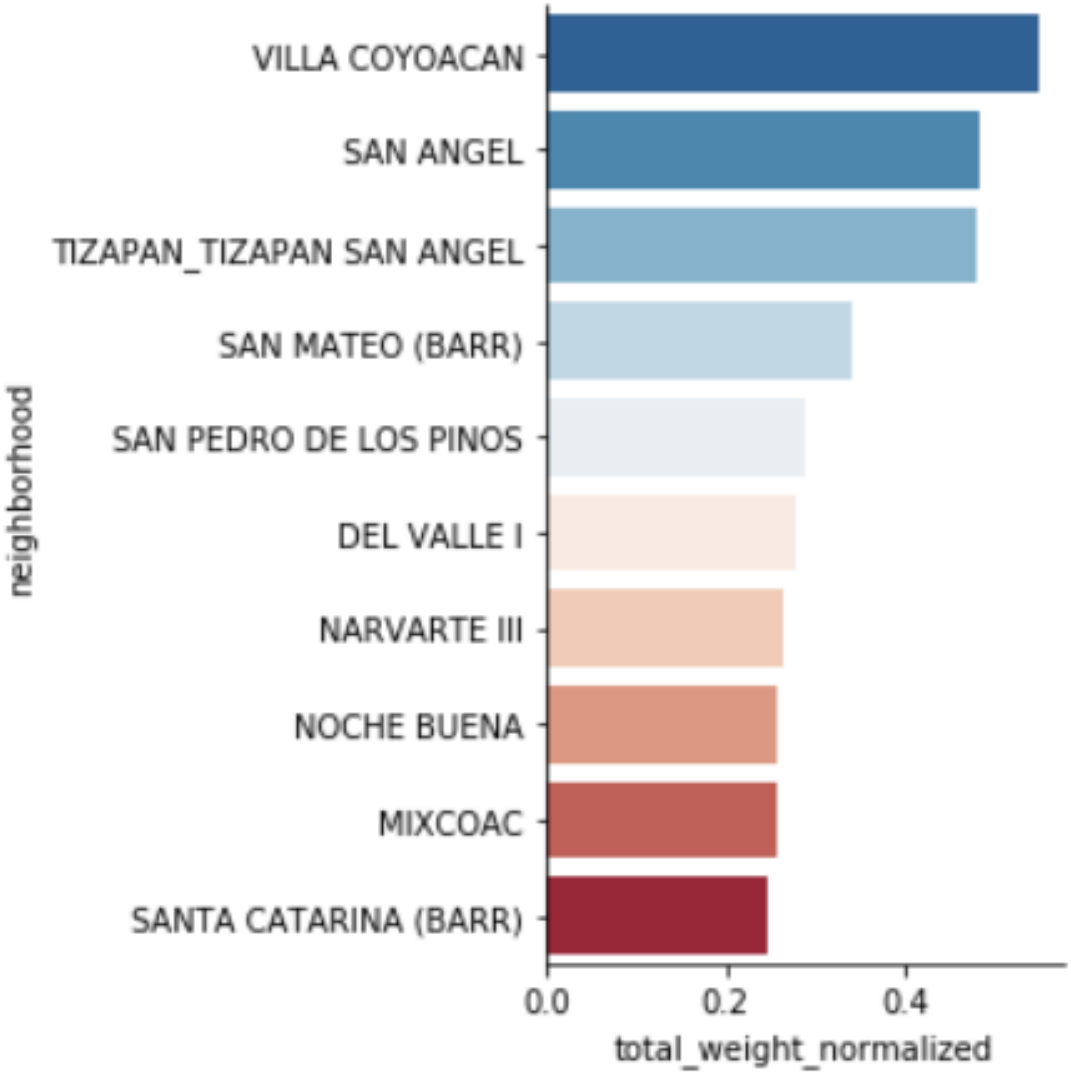
Night Life



Overall Best Neighborhoods for Millennials in Mexico City



Also, the normalized value plot is useful to consider the more well-rounded neighborhoods. This new weight considers the value of each neighborhood based on the total of venues in the surrounding boroughs. This list is what our recommender would suggest to a gen y potential customer. And, based on price range, a singular suggestion can be done.



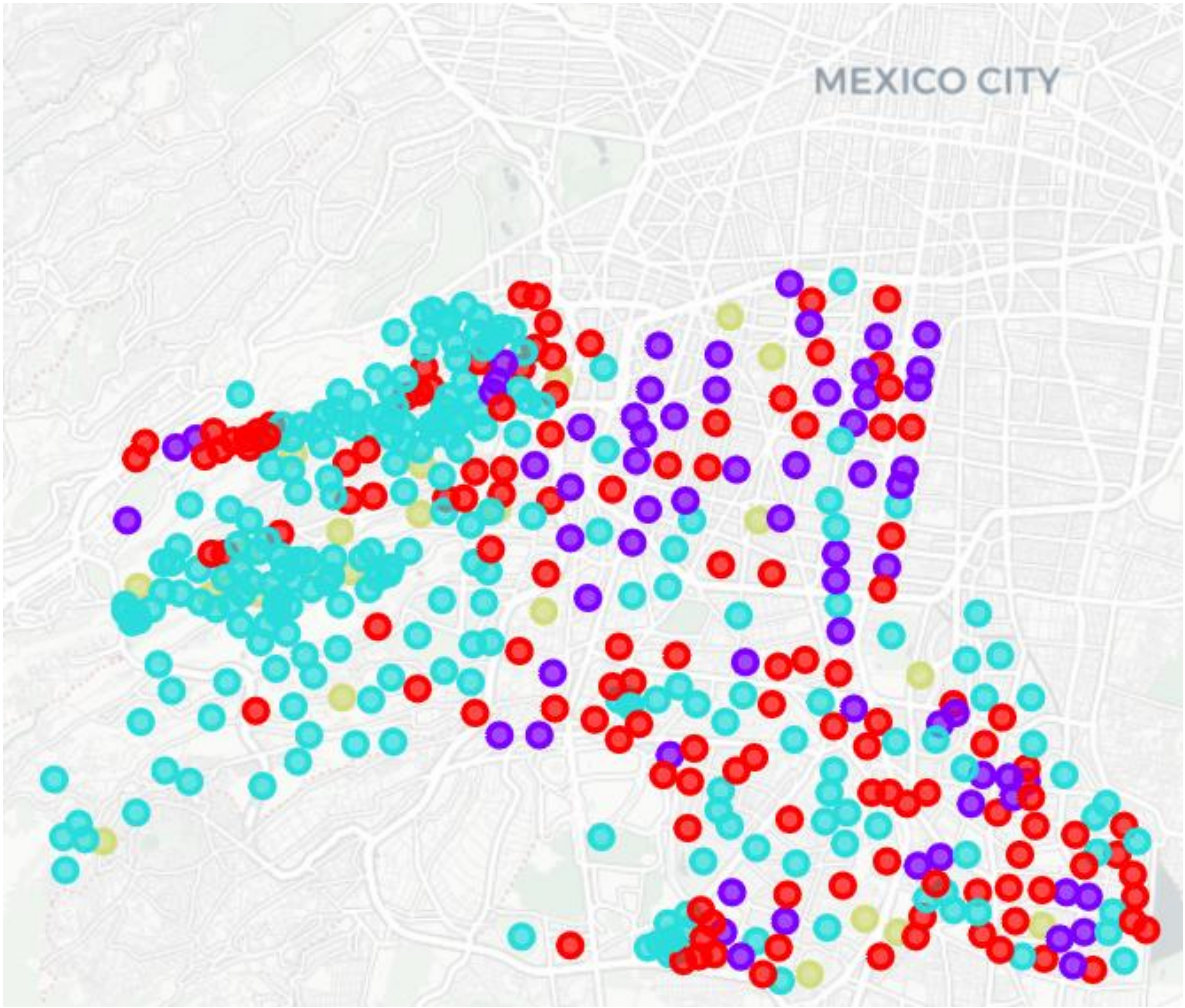
5. Discussion

The statistical description of the results is very interesting. Every category contains a lot of outliers, indicating that there are either very specialized neighbors based on category, or that there are several neighborhoods that have a lot of potential to the generation y sector.

Based on the bar plots in section 4, we can conclude that *Narvarte*, *Del Valle, Pedregal, Portales* and *San Ange*l are clusters of neighborhoods that suit any kind of preference, since they appear recurrently. But besides those neighborhoods, the vast majority of neighborhoods that come on top of each category, can be considered as “specialized neighborhoods”.

6. Conclusion

I believe the analysis of these neighborhoods can be highly beneficial to any company that sells or rents in Mexico City. The project was finished based on initial goals, and with the information obtained, new analysis can be made to the city. The following cluster is the proof of it. The distribution of the colors could lead to new insights and business opportunities, and a better understanding of the city I live in.



Transport cluster: yellow

Restaurant cluster: purple,

Fitness: red

Arts & Entertainment: light blue

References

1. ICASAS. Los 5 estados donde viven los millenials. Available at: https://www.icasas.mx/noticias/donde-viven-los-millenials/.
2. NAR. 2017 National Community and Transportation Preference survey. (2017). Available at: https://www.nar.realtor/reports/nar-community-and-transportation-preferences-surveys.
3. CDMX GOB. No TiBase de datos de las Colonias de la Ciudad de Méxicotle. (2020). Available at: https://datos.cdmx.gob.mx/explore/dataset/coloniascdmx/table/.
4. CDMX GOB. Carpetas investigacion CDMX. (2020). Available at: https://datos.cdmx.gob.mx/explore/dataset/carpetas-de-investigacion-pgj-de-la-ciudad-de-mexico/.
5. CDMX GOB. Paradas y Terminales del Sistema de Transporte Publico. (2020). Available at: https://datos.cdmx.gob.mx/explore/dataset/estaciones-paradas-y-terminales-del-sistema-de-transporte-unificado/table/.