House affordability in Medellin, Colombia

Ricardo Arias (ID 30550971)

April 23, 2021

1. Introduction

Medellin is the second largest city of Colombia, with 2.6 million of inhabitants. The city is divided by 22 different areas called communes where people live. In Medellín, Colombia the segregation by the income level of the people throughout the territory is clear, unlike Melbourne, there you can easily identify which is the area inhabited by people with higher income and which by people with less

With an open dataset provided by Properati (the Latin American real estate search site), an open dataset of the local government of Medellin, Colombia, and one dataset of the government of Colombia, we would like to identify what is the relationship between poverty (measured by the Multidimensional Poverty Index), communes, house pricing and house distribution.

1.1. Questions

1. How is the availability, and pricing of properties in different suburbs of Medellin, Colombia?
2. What is the distribution of poverty in different suburbs of Medellin, Colombia?
3. How is the relationship between the price of the properties in Medellin, Colombia, and the poverty?

2. Data Wrangling

2.1. Data Sources

1. **Properati Dataset:** All property listings and real estate developments that are and were published on Properati throughout Colombia from 2015. (Properati, 2021)
    a. ***URL:*** https://www.properati.com.ar/data/
    b. ***Type:*** CSV file
    c. ***Last Update:*** 07/01/2021
    d. ***Dimensions*:** 1.000.000 rows and 25 columns
    e. ***Attributes:*** Described in the proposal

2. **Cadastral Boundary of Suburbs and Neighborhoods:** It represents the boundary of the commune and suburb according to the conformation of properties in the city. (Medellin Goverment, 2021)
    a. ***URL:*** https://geomedellin-m-medellin.opendata.arcgis.com/datasets/l%C3%ADmite-catastral-de-comunas-y-corregimientos?page=2
    b. ***Type:*** ShapeFile
    c. ***Last Update:*** 05/03/2019
    d. ***Dimensions*:** 25 rows and 6 columns
    e. ***Attributes:*** Described in the proposal

3. **Multidimensional poverty measure - block-level information:** It represents the poverty in each block of the city, represented by the Multidimensional poverty measure (MPI). It was measured in the Population and Housing Census 2018. (DANE, 2021)
    a. ***URL:*** https://geoportal.dane.gov.co/visipm/
    b. ***Type:*** ShapeFile
    c. ***Last Update:*** 2018
    d. **Dimensions:** 407.851 rows and 3 columns
    e. ***Attributes:*** Described in the proposal

Dimensions	*Indicator*	*Deprived if--*	*Weights*
Education	Years of Schooling	No household member has completed five years of schooling	1/6
	Child School Attendance	Any school-aged child is not attending school up to class 8	1/6
Health	Child Mortality	Any child has died in the family	1/6
	Nutrition	Any adult or child for whom there is nutritional information is malnourished	1/6
Living Standards	Electricity	The household has no electricity.	1/18
	Improved Sanitation	The household's sanitation facility is not improved (according to MDG guidelines), or it is improved but shared with other households	1/18
	Improved Drinking Water	The household does not have access to improved drinking water (according to MDG guidelines) or safe drinking water is more than a 30-minute walk from home, roundtrip	1/18
	Flooring	The household has a dirt, sand or dung floor	1/18
	Cooking Fuel	The household cooks with dung, wood or charcoal	1/18
	Assets Ownership	The household does not own more than one radio, TV, telephone, bike, motorbike or refrigerator and does not own a car or truck	1/18

This index (MPI) measures poverty through the education, health, and standard of living indicators to determine the incidence and intensity of poverty experienced by a population. In which a larger figure indicates a higher level of poverty. (Wikipedia, 2021).
The methodology to calculate this index is shown in the following table.

2.2. Data Transformation

Using R and the library dplyr, we transformed this dataset so we could use it to reach the target of the project. The transformations made are listed below and were performed in another file upon request.

Transformations of Properati Dataset

1. ***Filter out the properties located in Medellin, Colombia***
   - With this step, we create a new dataset with dimensions 207195 rows and 25 columns

2. ***Convert the coordinates of each property to the same Coordinate Reference System (WGS84 or 4326 in R)***
   - With this step, we just changed the format of the coordinates, but the size of the dataset remains the same

3. ***Drop columns that only have the same value along all the rows, as this information is not useful***
   - Identifying the unique values of each column we found out that 4 columns (ad_type, l1, l2 and l3) have just one value in all the rows
   - With this step, we create a new dataset with dimensions 207195 rows, and 21 columns

4. ***Drop columns and rows with too many missing values***
   With Figure 2, there are some columns with too many missing values that are not going to help in the development of the project, therefore we get rid of them.
   - l6
   - surface_total
   - surface_covered
   - price_period
   - rooms
   - l5

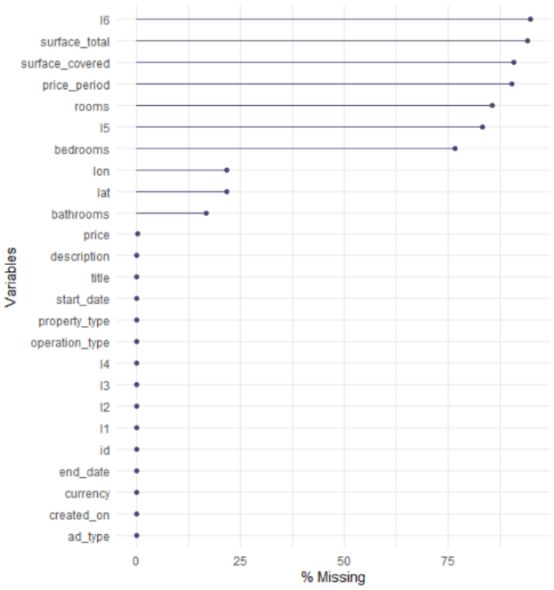


Figure 1. Missing values per Attribute

- Even though the attribute bedrooms, has too many missing values, we decided to keep it as there is a question related to it, and we can still use 30792 rows of this attribute.
- We remove the properties without coordinates as those cannot be analyzed.

5. ***Currency conversion***
   - The column price is given in two different currencies COP – Colombian Peso and USD – US Dollars, to make the project more informative we converted these prices to AUD – Australian Dollars with two different exchange rates:
     - $1 Australian Dollar is equal to $2.500 Colombian Pesos
     - $1 Australian Dollar is equal to $1.3 US Dollars

   As we already have a new column with the price of the property in Australian Dollars, we get rid of the columns price and currency

6. ***Drop useless columns***
   - start_date and end_date are useless because we have the column created_on, which gives us all the date information required for the project.
   - title and description are useless because it refers to the post of the seller, and we are not going to use text analyzer to extract more information, because Properati gives us the information complete about each property.
   - id is given but will not be used for the purpose of this project.
   - l4 is given but we will identify it better when we merge this dataset with communes.

7. Translation of terms

- As this is a dataset from a Latin American webpage there are many terms written in Spanish that will be translated to English for better understanding.

After this wrangling we remain with a dataset with 90.474 rows and 6 columns:

8. Remove Outliers

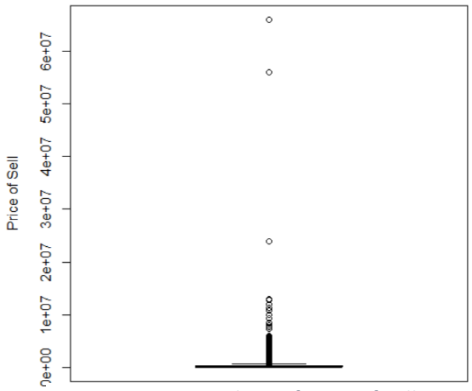


Figure 2. Outliers of Price of Sell

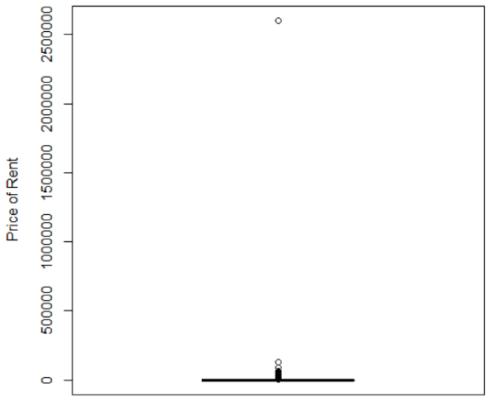


Figure 3. Boxplot of Price of Rent

By looking at the properties we found out that there are properties like building or big estates in the dataset that can affect the visualization of the properties as their prices are too high compared to most of the properties, this happens with Price of Sell and Rent as can be observed in Figure 3 and 4. That is why we proceed to remove those values with the following process.

Summary of Price of Sell

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
4000	100000	160000	260335	276000	66000000	57

Summary of Price of Rent

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
46	448	720	1300	1280	2600000	28

With the summary tables of we can calculate the IQR of each price.

IQR of Price of Sell

$$IQR = Q3 - Q1$$
$$IQR = \$276.000 - \$100.000$$
$$IQR = \$176.000$$

IQR of Price of Rent

$$IQR = Q3 - Q1$$
$$IQR = \$1.280 - \$448$$
$$IQR = \$832$$

And then calculate the parameters to define if a value is an outlier

$$median - 1.5IQR < P < median + 1.5IQR$$

Parameters of Price of Sell

$$\$160.000 - 1.5(176.000) < P < \$160.000 - 1.5(176.000)$$
$$\mathbf{-\$104.000 < P < \$424.000}$$

Parameters of Price of Rent

$$\$720 - 1.5(832) < P < \$720 - 1.5(832)$$
$$\mathbf{-\$528 < P < \$1.968}$$

And that is how we found out the parameters to accept or not a value in our dataset and remove the outliers.

Transformations of Multidimensional poverty measure - block-level information

The only transformation made to this dataset was to create a new column where we reflect the IPM in terms of categories of Incidence of Multidimensional Poverty following the advice given by the source of the data.

a. 0.1% - 20% ~ 0 < IPM < 20
b. 20.1% - 40% ~ 20 ≤ IPM percentile < 40
c. 40.1% - 60% ~ 40 ≤ IPM percentile < 60
d. 60.1% - 80% ~ 60 ≤ IPM percentile < 80
e. Greater than 80% ~ IPM percentile ≥ 80
f. 0% ~ IPM = 0 (If a zone it is not residencial, the IPM is 0 by definition)

Transformations of Cadastral Boundary of Suburbs and Neighborhoods

Drop useless columns OBJECTID and SECTOR, with the remaining columns we can perfoms the tasks

3. Data Checking

To perform the data checking we are going to build a new dataset in a tabular form, where we merge all three datasets mentioned before. Using a library **library('sf')** we are going to append 2 new columns to the Properati Dataset: Commune and IPM corresponding to the coordinate of the property. We do this with a function that looks in what shape (Commune or IPM Block) is located at one point (property coordinates).

3.1. date

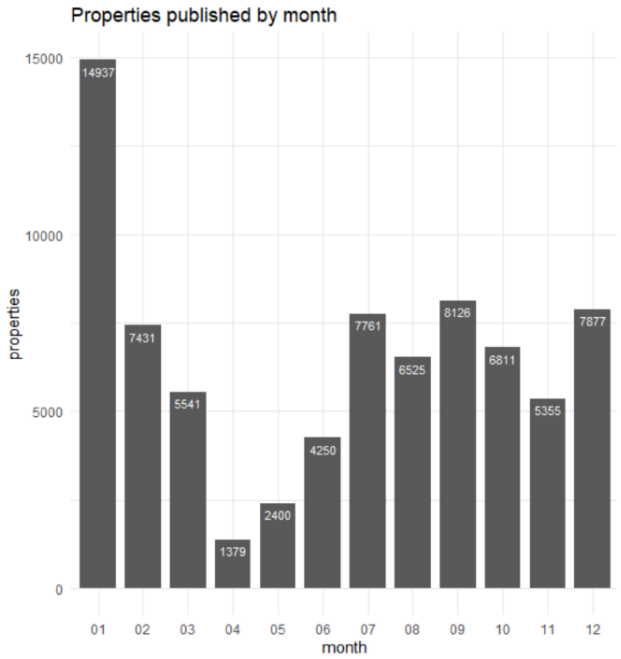


Figure 4. Properties published by month

Using a barplot we can observe the number of properties published by month in Properati. Even though the original dataset has data since 2015, when we performed all the transformations mentioned above, all the data remaining is from 2020.

From this plot, we can observe how the property selling and renting dropped drastically between February and July due to COVID-19, and we can also observe that the market has not recovered the numbers before the pandemic, as January (01) is greater than any other month by nearly the double.

The month with the least publications of properties on Properati web page was April (04) coinciding with the beginning and the worst time of COVID-19 in most countries of the world.

Even though this project is not about that we can observe the impact of this virus in this market in Colombia last year.

3.2. Bedrooms and Bathrooms

In the following figure, we plotted two different attributes, bedrooms and bathrooms. With two different plots: Scatter plot and boxplot.

From the graph, we can observe the relationship between these two variables and it can be seen that most of the properties have less than 10 rooms and 5 bathrooms, but few properties that have more than that (outliers) but we are not removing them as it could provide us information about big properties being sold or rented.

With the blue tendency line, we can identify that the relationship between them is strongly positive, with a rate of approximately 2 bedrooms per bathroom in each property.

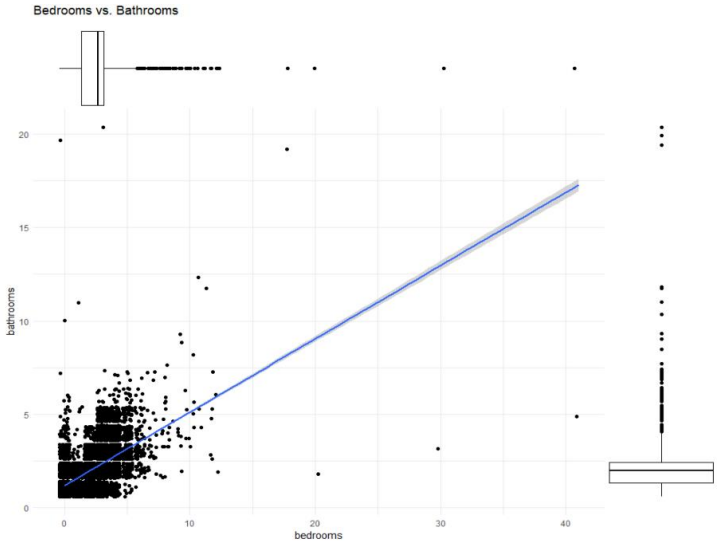


Figure 5. Distribution of properties (Bedrooms vs. Bathrooms)

3.3. Property Type and Operation Type

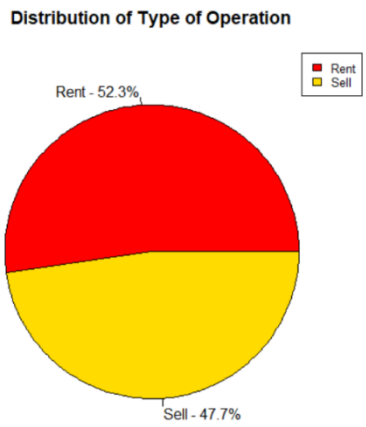


Figure 7. Distribution of Operation Type

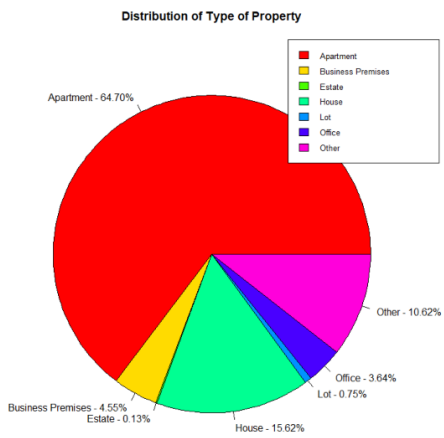


Figure 6. Distribution of Property Type

Using pie charts we can observe the proportion of properties classified by property and operation type.

From Figure 8, we identify that the properties published on Properati webpage are evenly distributed in terms of operation type. Almost half of the properties are posted to sell them and the other half to rent them. But in Figure 7, we observe the opposite, almost 2 out of 3 properties posted are apartments.

3.4. Price

After removing the outliers of price, as we showed before, we obtain the following density plots of price, wrapped by operation type, as we can not compare the price of selling with the cost of renting.

By looking at the graph, we can observe that both curves are skewed to the right, because their mode is smaller than their median. What can lead us to think that that there are many more properties with a low price and few properties with a very high price compared to the majority. It could be because of the social inequality present in Colombia, that has a Gini index of 51.3%.

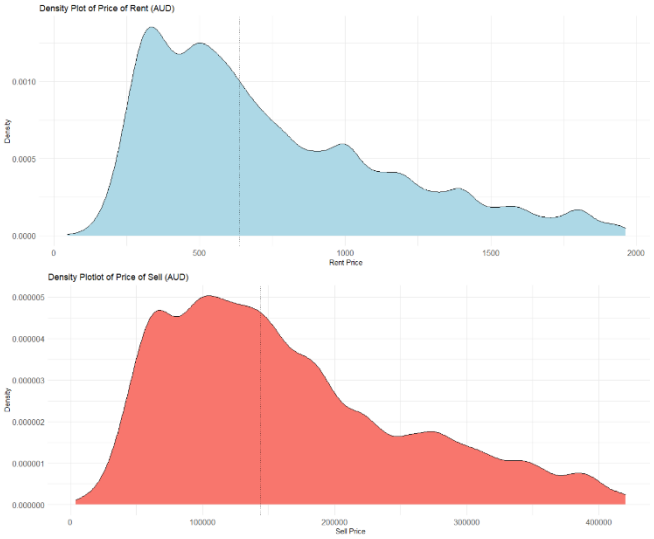


Figure 8. Density plot of price

3.5. Commune

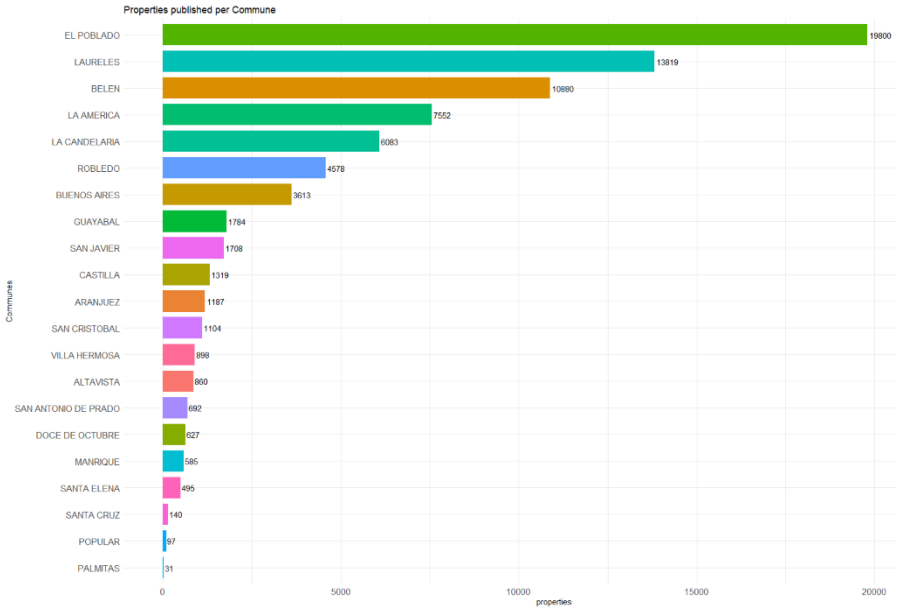


Figure 9. Distribution of properties published by Commune

Medellin is divided into 21 communes (suburbs) shown in Figure 10. From this barplot, we can observe a pretty high disbalance of properties published per commune, as the three most popular communes have approximately the 50% of all the properties of the dataset.

There could be many reasons for this, one could be that these suburbs are the most residential in the city, the second one could be that those are the most popular suburbs to live in, or the people living in these communes use more digital platforms to sell their properties than other methods.

It is important to mention that in poor regions of Colombia, people use informal methods to buy and sell everything.

3.6. IPM and IPM Class

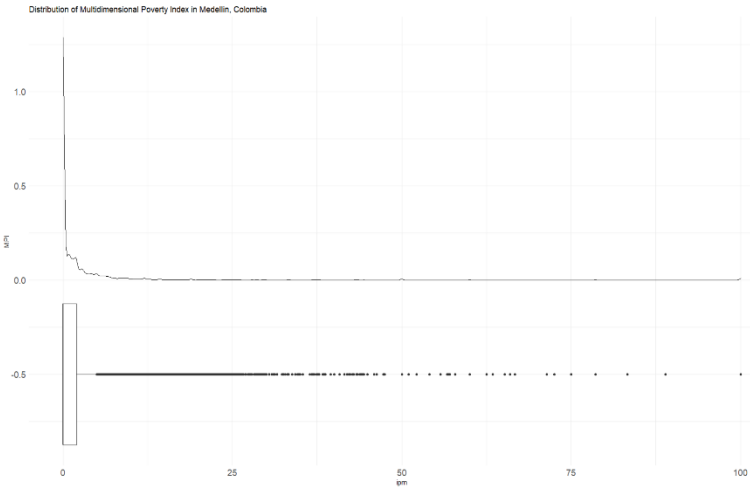


Figure 10. Distribution of IPM in Medellin, Colombia

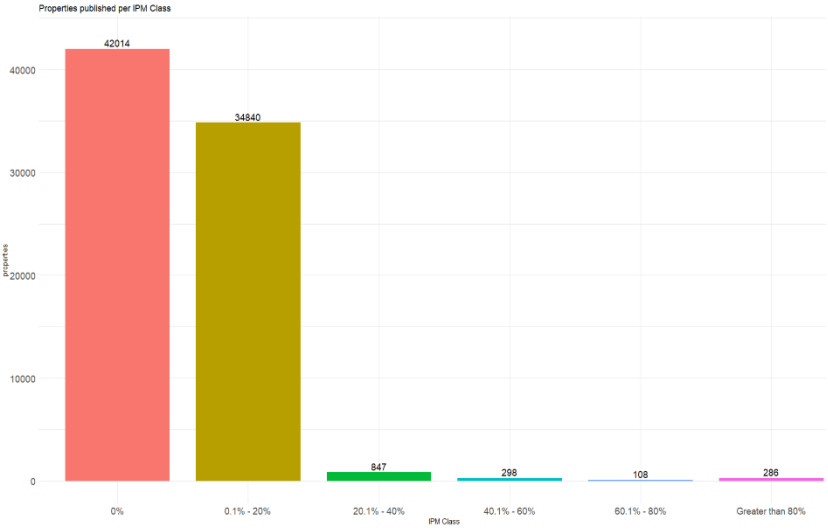


Figure 11. Distribution of properties by IPM Class

From Figures 11 and 12 we can observe the high disbalance in the IPM in Medellin, there are many blocks identified as non-residential (0%) with too many people living there. This could be because of bad urban planning, because people are living in zones destined for a different purpose, or because the method of measurement of the index is not correct, as they classified a residential zone into a non-residential one.

More than 50% of the properties are located in blocks with IPM equal to 0, and 95.5% of the remaining part of the data is classified as 0.1%-20%, which means that the properties are located in blocks where the incidence of poverty is low. Supporting the theory mentioned before in section 3.5. where we said that rich people are more likely to use digital platforms to sell or rent their properties.

4. Data Exploring

In this section we are trying to answer the questions made at the beginning of the document, by using some plots for a better understanding of it.

4.1. How is the availability, and pricing of properties in different suburbs of Medellin, Colombia?

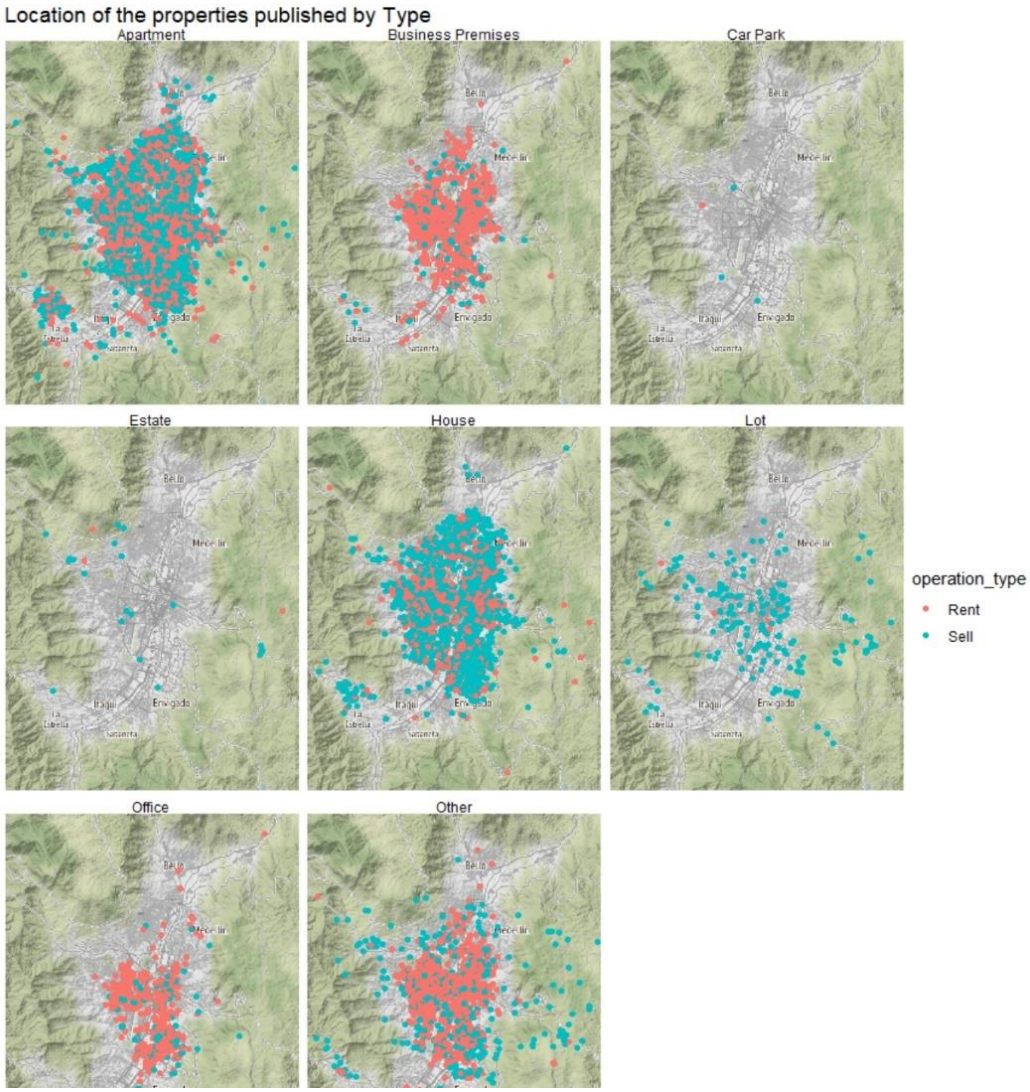


Figure 12. Localization of Properties by property_type and operation_type

With Figure 13 we can visualize the localization of each property wrapped by its type and coloured by its operation type.

With this plot, we can identify that there are not so many Car Parks and Estates available, as the maps seem to be pretty empty.

Besides, we can observe a pattern: properties like Houses and Lots are mostly to be Sold represented by the blue dots, and Offices, Others, and Bussiness Premises are mostly to be Rented represented by the red dots, and Apartments are equally distributed. With theses patterns, we can identify a little bit how the property market is in Medellin.

Finally, we can also identify that you can find Apartments, Houses, Business Premises and Others are everywhere in the city, but properties like Offices are more likely to be in the southern part of it, while the lots are just in the center of it.

With respect to the price there is a pretty clear pattern in both Rent and Sell type properties. The most expesive properties are located in the Southestern part of the city, as we can see the lighter blue dots are located there. While the cheaper one tend to be in the north. There is a transition in between. Letting us know that the more to the north you live, the cheaper are the properties

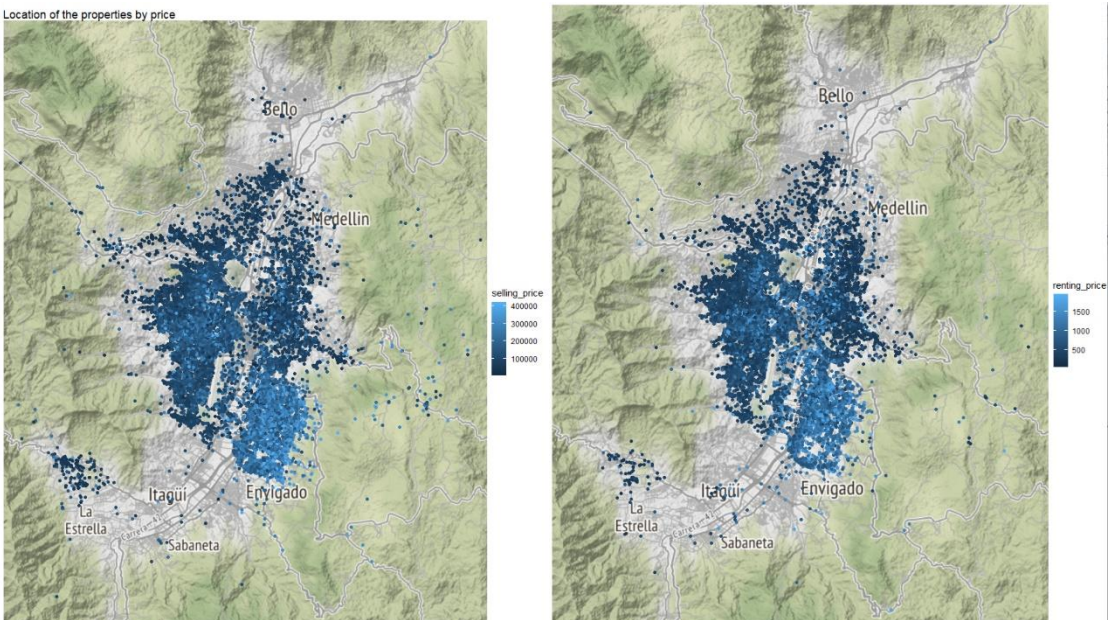


Figure 13. Localization of properties in Medellin by Price and Operation Type

In order to support Figure 14, we calculated the mean renting and selling price of the properties located in each Commune (suburb). Finding out the same pattern as we saw before, the most expensive properties are located in the south east part of the city. In suburbs like El Poblado, Guayabal, and Santa Elena. Even though there are some suburbs located in the north like Popular which has one of the highest renting price of the city, as well as Palmitas in the west that has one of the highest selling price.

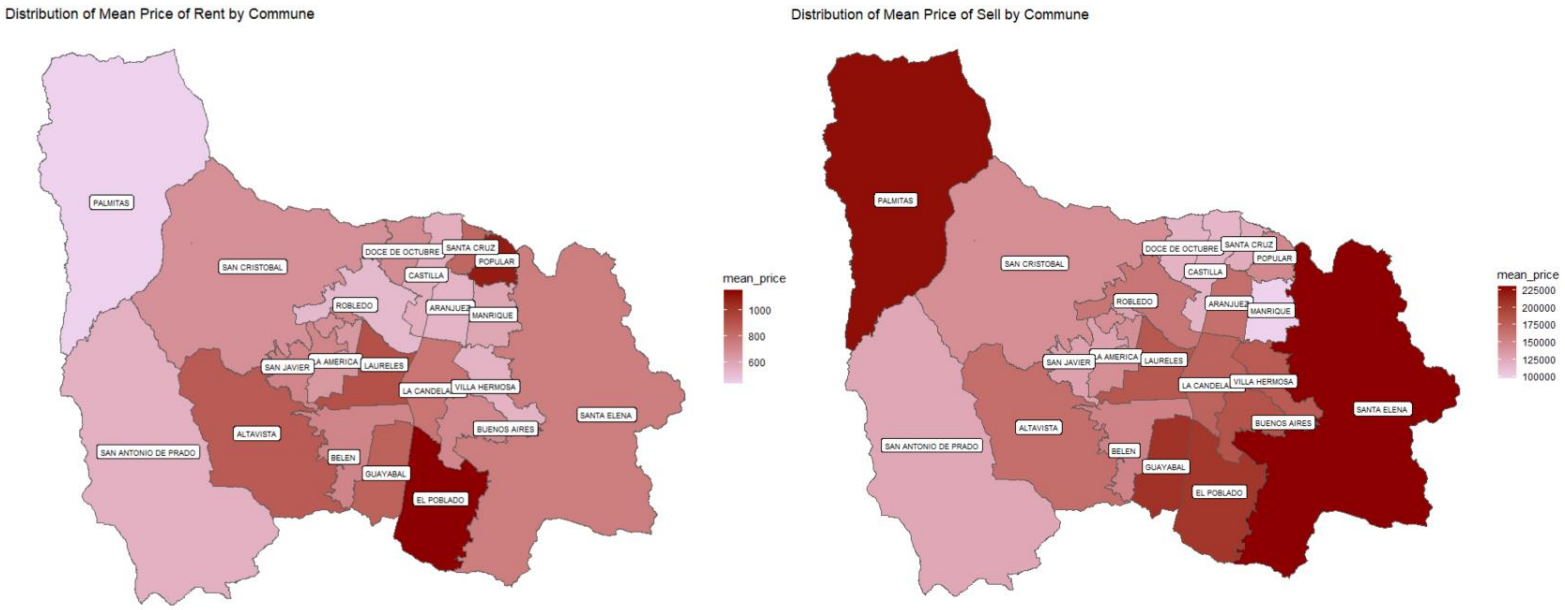


Figure 14. Distribution of Mean Price of Properties by operation_type and Commune

4.2. What is the distribution of poverty in different suburbs of Medellin, Colombia?

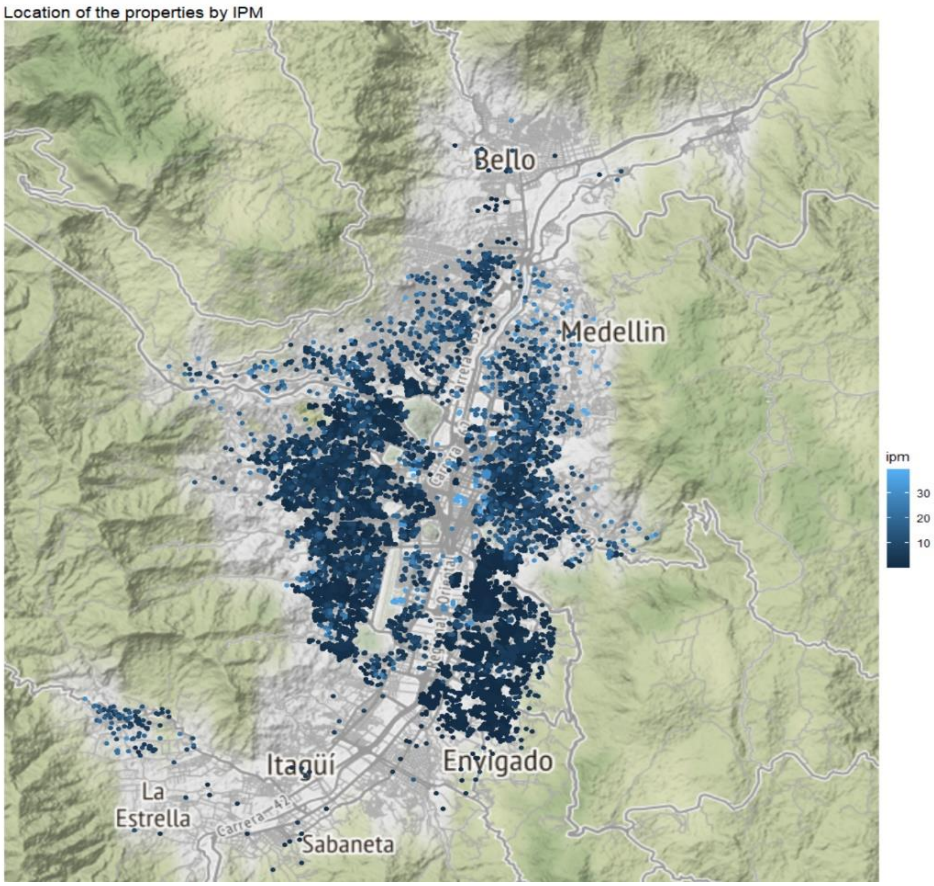


Figure 15. Localization of Properties in Medellin by its IPM

Before localizing the properties by it IPM in the map of Medellin, we filtered out some values greater than 40 and the zeros, because of the reasons given in the data checking, where we mentioned that there are no so much data and in order to make it more visible we filtered out this data.

Now that we plotted this information we can observe again that the southern part of the city has an IPM smaller compared with the IPM of the northern part.

Comparing this map with the map of prices, this finding is very meaningful, because it is relating poverty with the price of the properties, and it has too much sense that in the poorest part of the city the properties are going to be cheaper in order to be more affordable.

It is also worth it to say that even though Medellin is not a rich city, it can be observed that their inhabitants are not so poor as many properties are located in spots with an IPM low

Again, in order to support Figure 16, we created Figure 17 a map with the mean IPM of each commune.

With this map, we can identify that the poorest suburb in the city is Santa Cruz, with an IPM of 12.5%. The richest suburb is Santa Elena or Palmitas with almost 0%

Besides, we observe that the lightest suburbs are located in the south and the darkest in the north, confirming the theory that the richest people of the city live in the south and the poorest people live in the north

There is clear social segregation in the city, it can be seen not only in the mean IPM but also in the mean price of the properties.

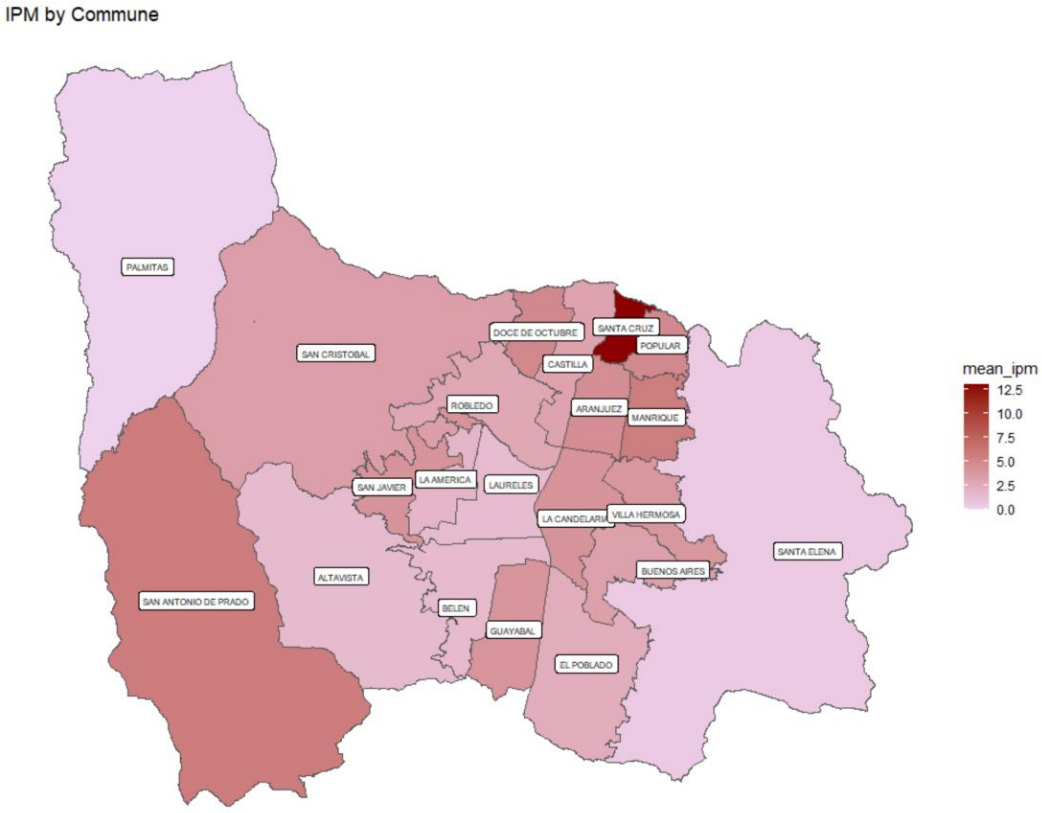
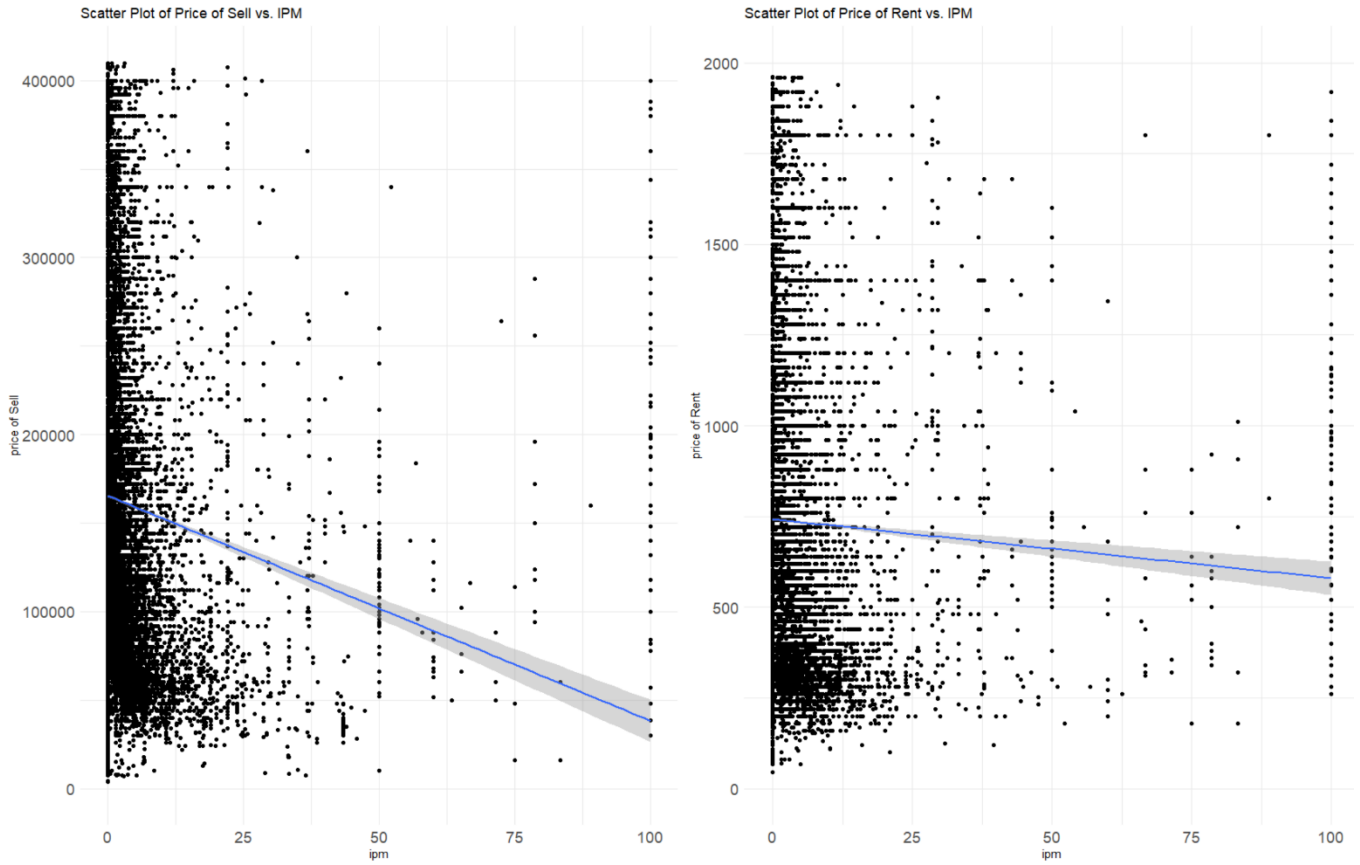


Figure 16. IPM by Commune

4.3. How is the relationship between the price of the properties and the poverty?



Additional to the relationships mentioned before between Figure 14 and 16 (Localization of properties by price vs. by IPM), we also have this graph, Figure 18, where we can observe that there is a moderate negative relationship between these two variables, which lead us to think that if the IPM decreases (richer people) the price of the properties increases no matter its operation type, and that if the IPM increases (poorer people) the price of the properties decreases.

Even though we found out this relationship, it is not pretty strong and we could not find any pattern that could help ud to model the poverty as a function property price or the property price as a function IPM. As we can see in Figure 19, where we plotted the Density Polt by IPM Class, there is no visible pattern relating these two variables directly. That is why we would need more data if we wanted to model one variable as funtions of the other one. But it is pretty clear too that the relationship exists.

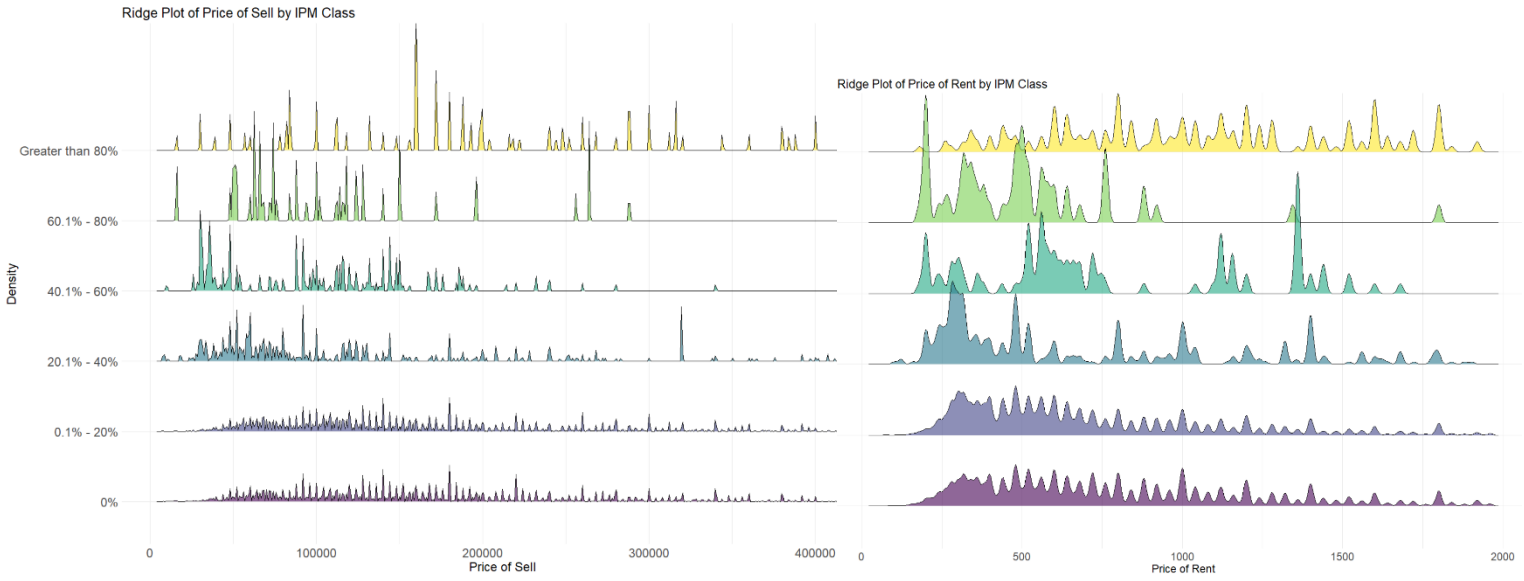


Figure 17. Ridge Plot of Price by IPM Class

5. Conclusion

As mentioned at the beginning of the document, Medellin is a city with social and economic inequality and located in a country (Colombia) that is the second most unequal country in Latin America, only behind Haiti with a Gini index of 0.538. It can be observed in figures 14, 15, 16, and 17 that the segregation of prices and IPM in the city is evident.

Even though we found out those things, the project is not conclusive as to the relationship between poverty and the price of properties is not very strong. We also discovered that this could be due to the lack of relevant data: Property area. The reason why this information is very relevant is that with it we could standardize the cost of each property in AUD/m2, and make a more fair comparison between them. In this project, we are comparing properties of small areas in the richest area of Medellin with properties of huge areas, in the poorest area, and it is obvious that even though it is not located in a better area it will have a higher cost because of its price.

Even with this limitation, the project was able to show that the cost of the properties is linked to the poverty index and its location within the city.

6. Reflection

1. From this project we learned, the importance of visualizations. All the insights that we discovered while doing it, could not be possible without them. Most of the insights were patterns of the locations of the properties along with the city, this could not be seen by just looking at the tabular data.

2. Not only it is important to plot the information but do it correctly, by using colors, size, position and also wrap the data with other information, to have another view of the same data.

3. It was really important the usage of different types of visualizations, as each of them gives a different point of view of the same information. That is why I tried to use many different visualizations, to obtain more and different insights.

4. R Studio is a very good tool to plot information, we tried to use Matplotlib from Python, and Tableau. But the library ggplot gives you too many tools to personalize the information in the way you think is better to present.

5. If I would have done something different, it would be to find a way to standardize the prices of the properties to find a better comparison between the properties. At this stage could be very difficult to find the area of each property of interest, so I would have to find a different way to do it.

7. References

DANE. (2021, 04 16). *Geoportal Dane*. Retrieved from https://geoportal.dane.gov.co/visipm/ : https://geoportal.dane.gov.co/visipm/

Medellin Goverment. (2021, 04 16). *GeoMedellin*. Retrieved from https://geomedellin-m-medellin.opendata.arcgis.com/: https://geomedellin-m-medellin.opendata.arcgis.com/datasets/l%C3%ADmite-catastral-de-comunas-y-corregimientos?page=2

Properati. (2021, 04 16). *Properati*. Retrieved from https://www.properati.com.ar/data/: https://www.properati.com.ar/data/

Santos, M. E., & Alkire, S. (2011). Training Material for Producing National Human Development Reports. *Oxford Poverty & Human Development Initiative*, 35.

Wikipedia. (2021, 04 06). *www.wikipedia.com*. Retrieved from Multidimensional Poverty Index: https://en.wikipedia.org/wiki/Multidimensional_Poverty_Index