Capstone Project – The battle of the Neighborhoods

Diego Belotto

August 2020

1. Introduction

Asunción is the capital of Paraguay and at the same time is the city with the largest quantity of venues and diversity of activities of the country, so for a young Paraguayan who wants to be independent or an adventurous tourist it seems as the ideal place to rent a house or an apartment.

The described situation generates an interesting problem to solve for the audience mentioned before, because there is a vast quantity of factors that need to be weighted to make a renting decision. Let us start describing some of these factors:

- The neighborhood
- The quantity of venues in each neighborhood
- The type of activities based on the venues
- The renting prices
- The types of renting places
- How many rooms have the houses or apartment?

Based on these premises we are going to try to help our audience to get a better understanding of the city using basic data analytics techniques and give tools to them to determine what neighborhoods would be suitable to start hunting for houses and try to predict what would be fair price to negotiate based on location data and renting data.

2. Data acquisition and cleaning

a. Data sources

- List of neighborhood from Asunción: for this purpose we'll be using the table from GeoHidroInformatica - Itaipu that contains data such as neighborhood name, district name, department name (equivalent to federal states), quantity of houses and codes for all of these fields as well a field with multipolygon data that contains delimiters for each neighbor
  http://geohidroinformatica.itaipu.gov.py/geoserver/wfs?typename=cih%3Aparaguay_2012_barrrios_localidades_wgs84&outputFormat=csv&version=1.0.0&request=GetFeature&service=WFS

- Neighborhood coordinates: open street map
  https://nominatim.openstreetmap.org
- Venues data for each neighborhood using de Foursquare API
  https://api.foursquare.com/v2/venues/
- Renting data: containing type of place, number of rooms, number of bathrooms, price, neighborhood, coordinates. We scrapped the data from a website called Infocasas.

https://www.infocasas.com.py/alquiler/inmuebles/asuncion/

b. Data Cleaning

- List of neighborhoods: from this data we picked the following fields.
  - - distrito = contains the code of the district
  - - dpto_desc = contains the name of the department
  - - dist_desc = contains the name of the district or city
  - - barlo_desc = contains the name of the neighborhood
  - - cant_viv = contains the name of the houses on each neighborhood

  Then added an additional field with the combined data from the 'barlo_desc' and 'dist_desc', to create the address field that we used to get the coordinates from each neighborhood using *ratelimiter* from *nominatim*. For this new data we created two new fields including the location data and point data (coordinates)

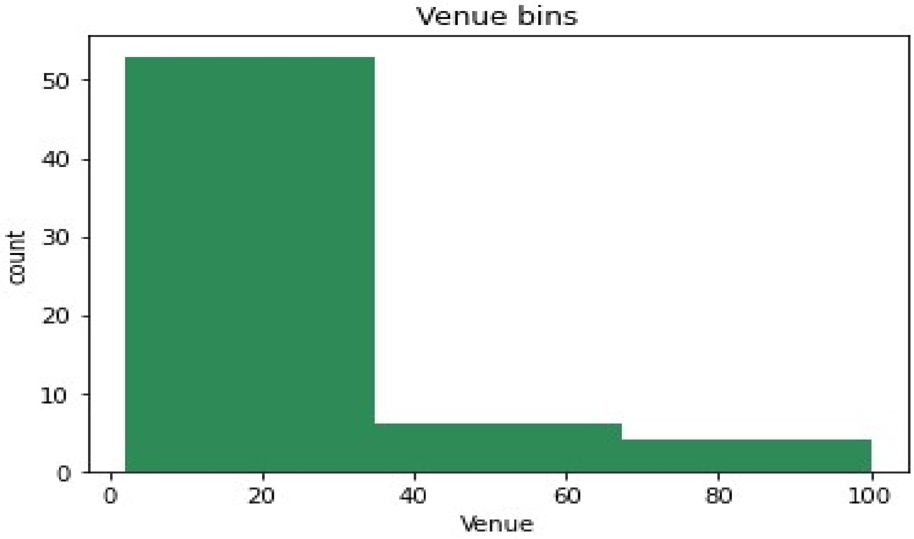
  Then, we checked if there was any row with no location data, finding 8 neighborhoods with no coordinates. We evaluated the quantity and determined that it was low, so in order to get the location data for these places, we've replaced manually the neighborhood names with the ones that appears in the google maps site and then update the dataframe. Later, we split the point data into three new fields: latitude, longitude, and altitude.

- Foursquare data: using the foursquare API we collected the venues data from each neighborhood in the neighborhood's list, selecting the venues name, venues coordinate and venue category and created a dataframe with these columns and the neighborhood names and coordinates.
- Renting data: we scrapped the ***infocasas*** data and found a field on the website with json datatype that contains all the necessary information and first transformed this data into pandas dataframe. After that, we identified that the price field contains data in two currencies: United States Dollar (USD) and Paraguayan Guarani (PYG), so we added an additional calculated column with price in PYG Currency using the exchange price of the month for every row valued in USD. Following, we deleted the plus sign in the bathroom and rooms fields and converted the datatype of these fields from object to float. Then, we converted the house type to string. Finally, we dropped all the fields that we are not going to use, getting a dataframe with the following fields.
  - Id = property id
  - Zona = neighborhood
  - tipoPropiedad = house type
  - m2
  - dormitorios = rooms
  - banios = bathrooms
  - lat = latitude
  - lng = longitude
  - Moneda = Currency
  - Precio_Origen = Original Price
  - Precio_Gs = Price in PYG

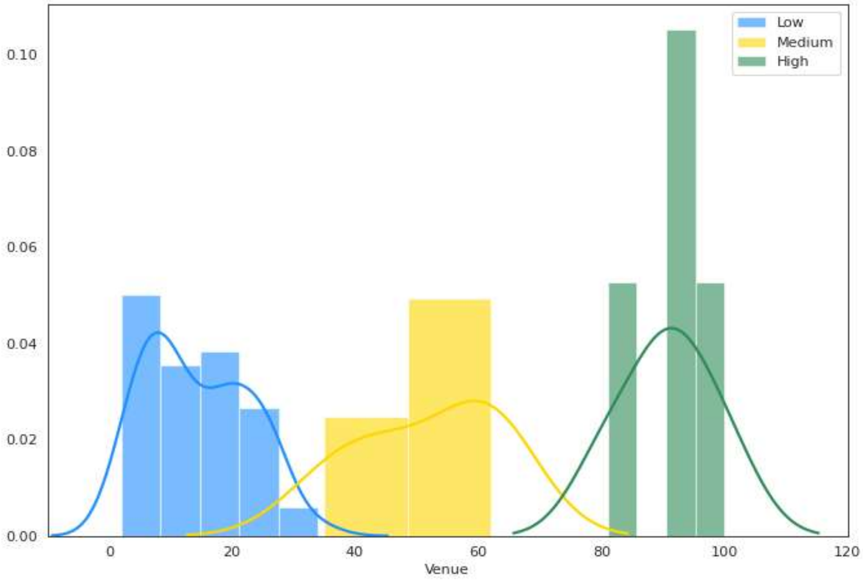
3. Exploratory Analysis

a. Distribution of venues per neighborhood and most common venue

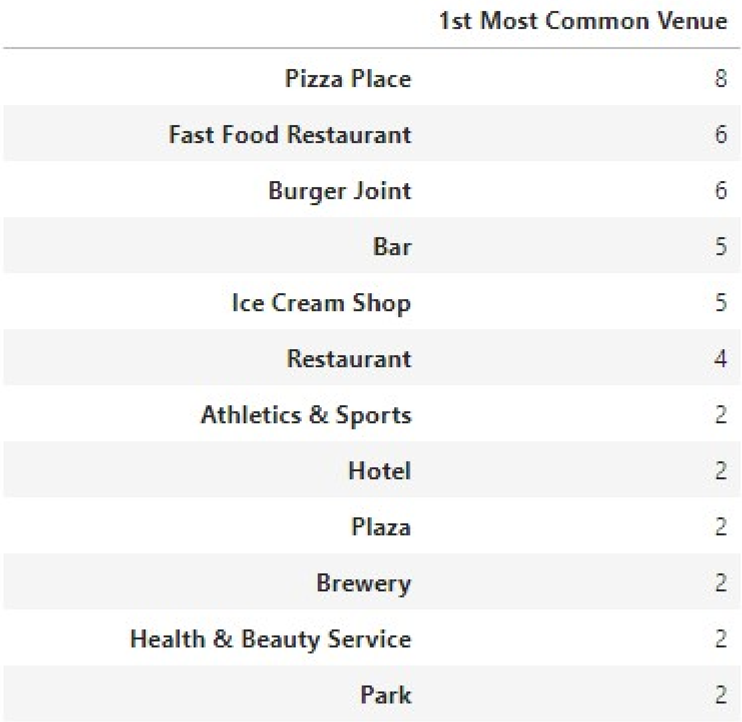
Firstly, we found that the city indeed has a wide variety of activities to look for, with 225 unique categories of venues, a number remarkably close to a bigger city as Toronto for example. However, considering the distribution of these venues by neighborhood we found that most of the neighborhoods have a low density of venues, where more than 50 neighborhoods have less than 34 venues. As a result, a person who frequently look for a lot of options to go out, has fewer neighborhoods to pick.



Then looking at the distribution by the categories of neighborhoods per venues (low, medium and high), it seems that in the low-density neighborhoods the quantity of venues is more evenly distributed than in the other two categories. Consequently, if someone chooses one with less venues, probably will not notice the difference comparing with most of the same category neighborhoods.



Now that we know the distribution of venues per neighborhood, we inspected which is the most common site that someone would find when renting a place.

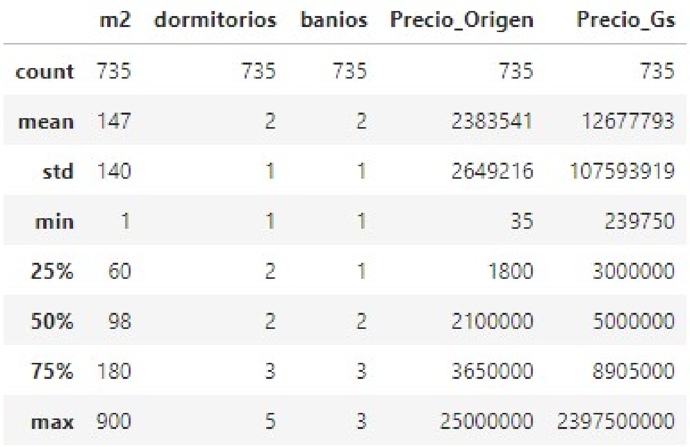


	1st Most Common Venue
Pizza Place	8
Fast Food Restaurant	6
Burger Joint	6
Bar	5
Ice Cream Shop	5
Restaurant	4
Athletics & Sports	2
Hotel	2
Plaza	2
Brewery	2
Health & Beauty Service	2
Park	2

And interestingly for a South American city, the most common location is the pizza place, therefore this shows that it is a friendly city for young people looking for fast food and this also can be a good tip for an Italian tourist and wants something familiar.

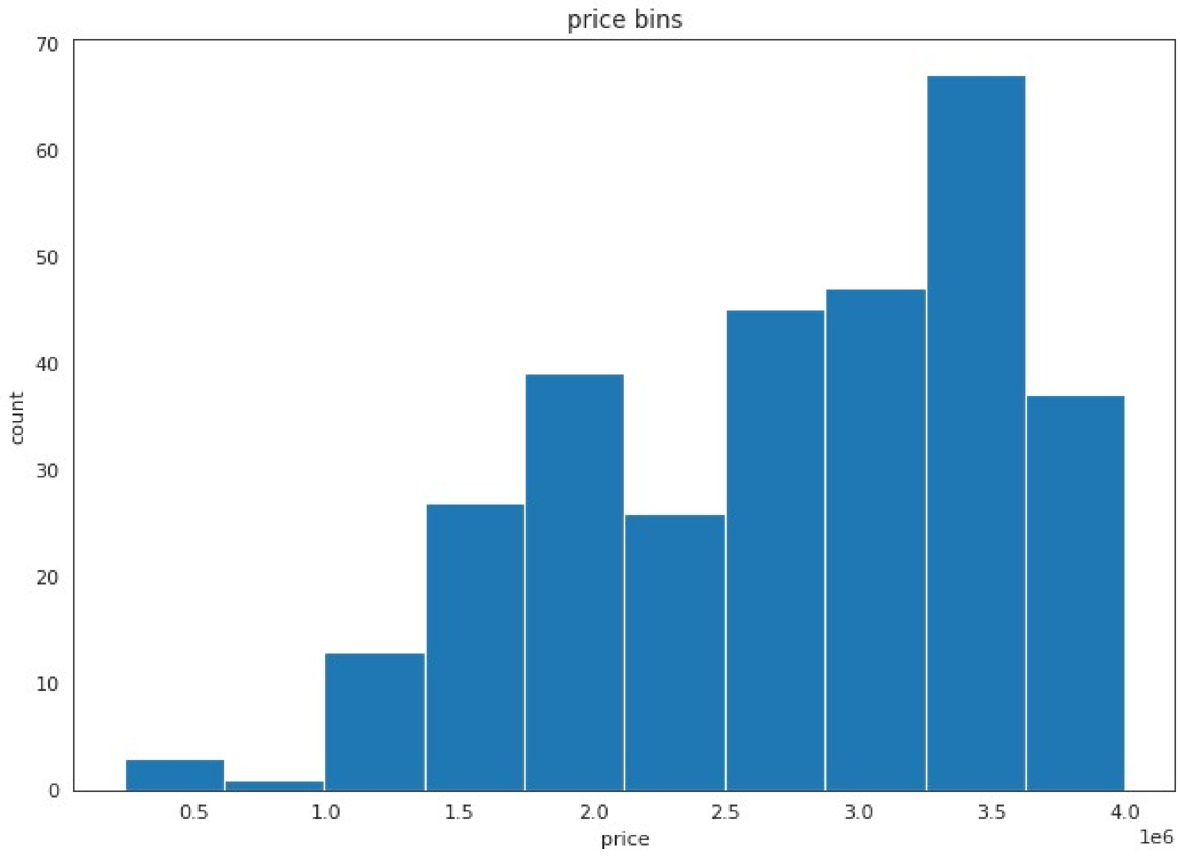
b. Renting price analysis

We have started the price analysis checking the distribution of this variable, but rightly at the start we found an extreme max value that generates a distortion in the distribution.



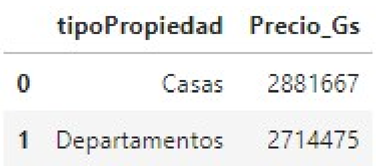
	m2	dormitorios	banios	Precio_Origen	Precio_Gs
count	735	735	735	735	735
mean	147	2	2	2383541	12677793
std	140	1	1	2649216	107593919
min	1	1	1	35	239750
25%	60	2	1	1800	3000000
50%	98	2	2	2100000	5000000
75%	180	3	3	3650000	8905000
max	900	5	3	25000000	2397500000

Therefore, in order to get a more precise analysis and moreover considering that this project was developed to solve a problem for a young person, we have decided to establish a limit in price at PYG 4.000.000

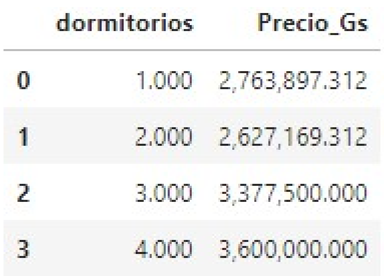


With these we can see that distribution of renting prices of our data tends to be in the highest price, starting at 2.500.000 PYG. This could demonstrate that this is the price we could expected in the city, however, could be also thanks of the population that offer places through this type of websites.

After looking at the price distribution, we have checked the average price per house type and per room's quantity. Considering that renting an apartment is in average cheaper than a house but not by that much, just PYG 170.000 that is almost USD 60. In the other hand seeing the average price we can notice that more rooms means highest price, that's quite logical, but doesn't apply to the one rooms places, that cost in average more than a place with 2 rooms, this is odd but could be because we replaced those observations without rooms with the most common, that was 2 rooms per place.

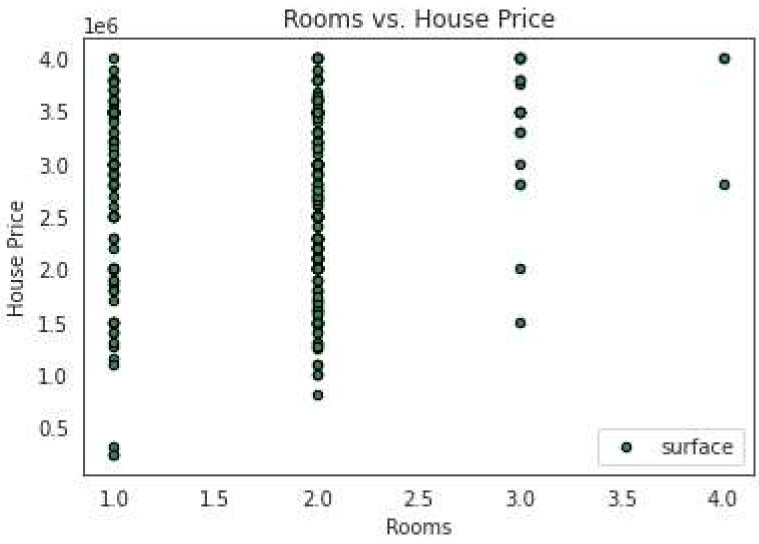


	tipoPropiedad	Precio_Gs
0	Casas	2881667
1	Departamentos	2714475

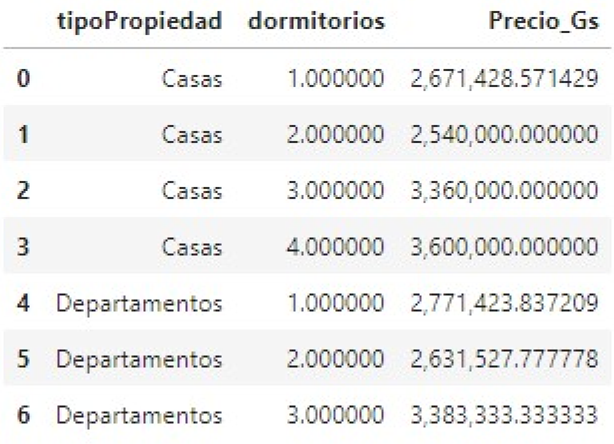


	dormitorios	Precio_Gs
0	1.000	2,763,897.312
1	2.000	2,627,169.312
2	3.000	3,377,500.000
3	4.000	3,600,000.000

Going a little deeper in the last relationship we can perceive that the average price does not reflect very well the relationship between the price and the quantity of rooms, because we can observe that there is quite a number of places with fewer rooms near the limit price and places with 3 rooms that can cost less than the average of the the places with one or two rooms. This relantionships shows that a four room place at minimun will cost more than the average of the other types.

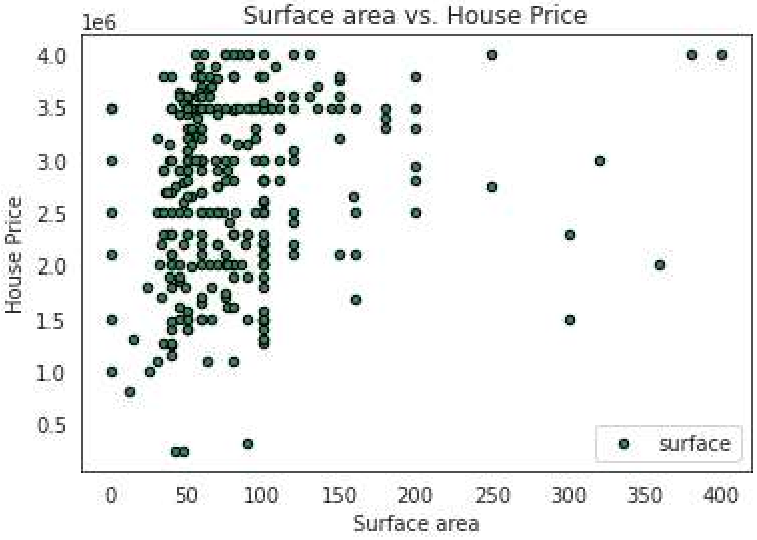


Grouping the data by types and quantity of rooms, we can see in terms of rooms it follows the same tendency we saw previously, but interestingly a house with 3 rooms is cheaper than an apartment with the same quantity of rooms. Accordingly, if someone is looking for a place with more rooms, maybe a better option could be a house instead of an apartment.

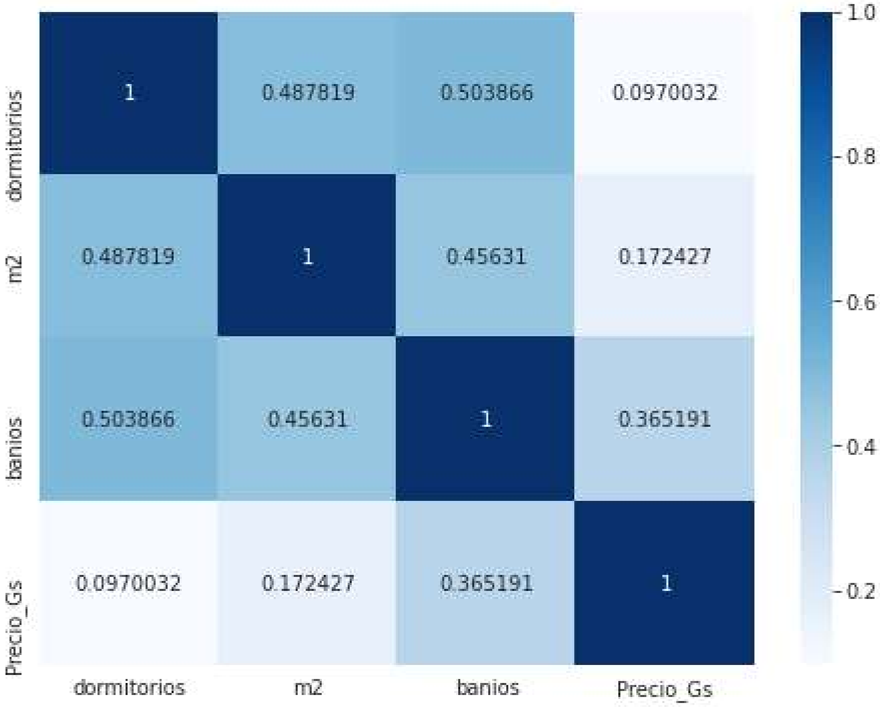


	tipoPropiedad	dormitorios	Precio_Gs
0	Casas	1.000000	2,671,428.571429
1	Casas	2.000000	2,540,000.000000
2	Casas	3.000000	3,360,000.000000
3	Casas	4.000000	3,600,000.000000
4	Departamentos	1.000000	2,771,423.837209
5	Departamentos	2.000000	2,631,527.777778
6	Departamentos	3.000000	3,383,333.333333

Then looking at the relationship between surface area (m2) and price we can see the relationship is similar to the relationship of price with number of rooms



Finally, examining at the correlation matrix, we found that there is no variable with a high correlation with price, where the quantity of bathrooms is the highest one, although, just with 0.36. This gave us a hint that we are going to need more than this renting variables to try to predict a fair price.



To add more information to this renting we are going to add the venues data analyzed first to prepare our model.

4. Modeling

As we have decided this project to help a young Paraguayan or a young tourist on how to start the rent hunting and what would be a fair price for renting, we have divided in two parts.

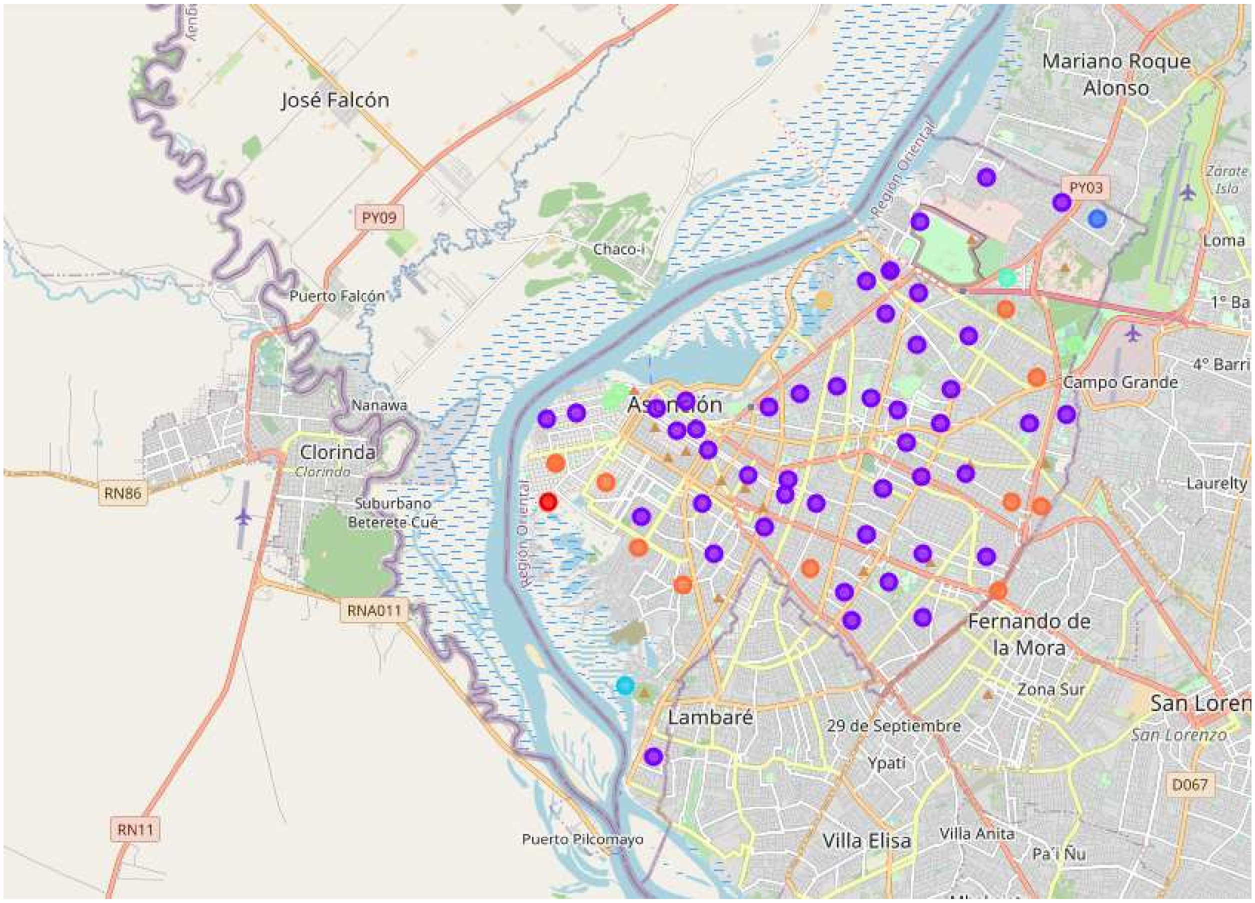
First, to give insights about the kind of neighborhoods in order to know what neighborhood would be suitable to start looking. For this part of the process we have a classification problem and we are going to use KMEANS algorithm.

For the second part, we need to predict a price and for this problem we have chosen a Regression model to solve this, testing two types of linear regression; linear and Ridge; and K Neighbors Regressor.

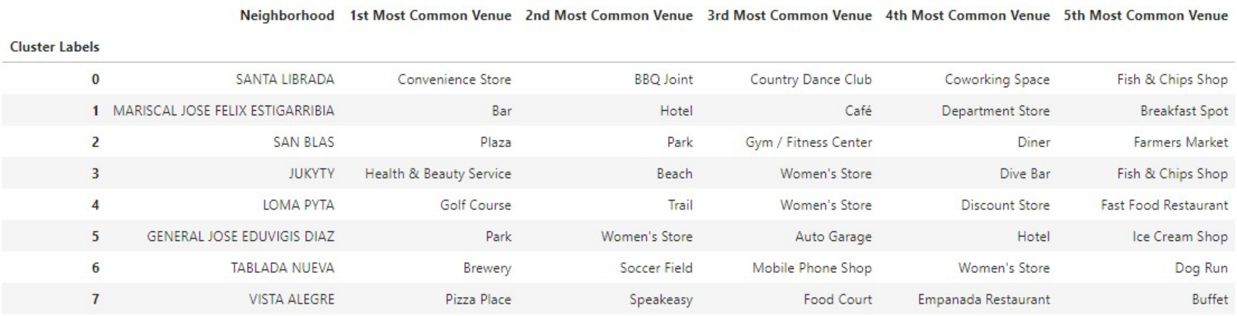
a. Classification Modeling

For the classification modeling we have used the neighborhood data combined with the foursquare venues data to generate clusters of neighborhoods based on the most common venues. Using the elbow method, we have determined that the best K for our data would be 8.

After fitting the model with our data, we have represented the cluster by colors on the map of the city. Noting that the cluster with the most neighborhoods extends across the entire city but seeing that most neighborhood near the limit of the city are in other clusters, probably more affected by the near cities and the type of venues that exist on those cities.



Then we analyzed the main characteristics of each cluster based on the most common sites.



Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	SANTA LIBRADA	Convenience Store	BBQ Joint	Country Dance Club	Coworking Space	Fish & Chips Shop
1	MARISCAL JOSE FELIX ESTIGARRIBIA	Bar	Hotel	Café	Department Store	Breakfast Spot
2	SAN BLAS	Plaza	Park	Gym / Fitness Center	Diner	Farmers Market
3	JUKYTY	Health & Beauty Service	Beach	Women's Store	Dive Bar	Fish & Chips Shop
4	LOMA PYTA	Golf Course	Trail	Women's Store	Discount Store	Fast Food Restaurant
5	GENERAL JOSE EDUVIGIS DIAZ	Park	Women's Store	Auto Garage	Hotel	Ice Cream Shop
6	TABLADA NUEVA	Brewery	Soccer Field	Mobile Phone Shop	Women's Store	Dog Run
7	VISTA ALEGRE	Pizza Place	Speakeasy	Food Court	Empanada Restaurant	Buffet

(First neighborhood of each cluster)

Cluster 0: This first cluster have just one neighborhood and have as the most common venue a convenience store and as the second most common BBQ joint, this cluster could be useful to rent because having a store such as the mentioned before is quite convenient.

Cluster 1: This is the cluster with the most neighborhoods, so probably will end up looking places in one of them and is the one was the most common venues are Fast Food Restaurant, Burger Joint and Bars. Therefore, this cluster can be very appreciated by someone who is searching for ease cooking problems

Cluster 2: This cluster seems an adequate option for someone who prioritize a fitness life, since it considers this factor into account, however, it leaves out other important aspects

Cluster 3: The cluster seems a suitable one for self-care as it contains the beach and most common health and beauty services, yet it doesn't seem a place to rent for young population.

Cluster 4: The fifth cluster contains a great place to go shopping and take advantage of sales and have fast food as a common activity, nonetheless this doesn´t include a place to do the daily common shops

Cluster 5: This cluster have neighborhoods to pick and they seem to be interesting places for a tourist because they have parks, hotels, stores, ice cream shop, stadium as most common places.

Cluster 6: this cluster just have one neighborhood and the five most common venues do not meet the criteria for a renting place for a young adult

Cluster 7: This last cluster could be quite suitable for a person that enjoys eating, since there are many options that also includes Italian pizzas.

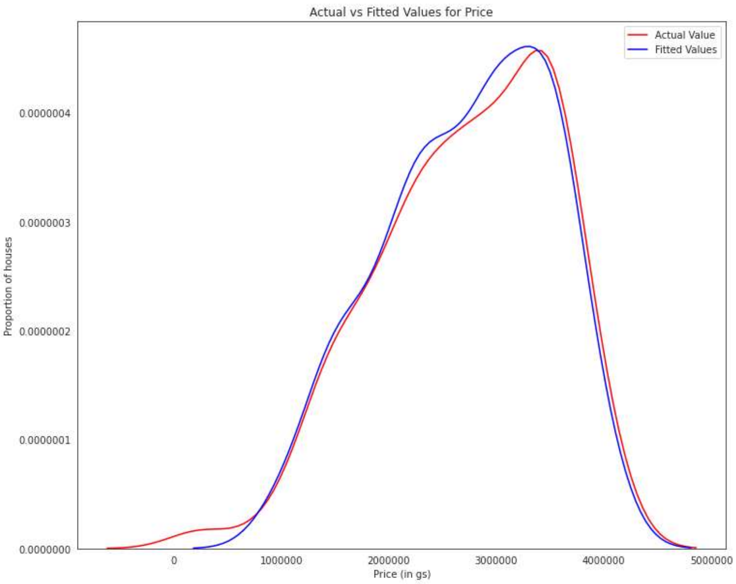
b. Regression Modeling

For the price prediction part of the project we tried 3 models to predict a renting price based on the renting data combined with the most common venues from foursquare data.

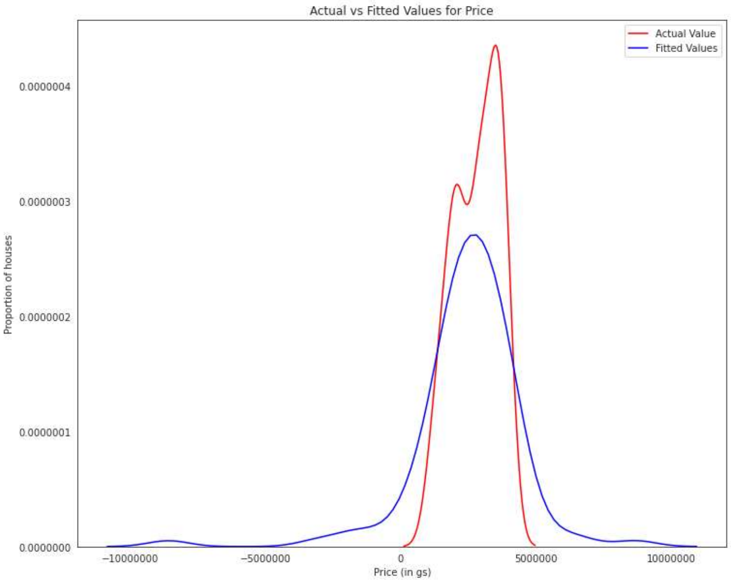
For the model selection we used the R2 score to decide which one has the best performance for our data.

The first one was the linear regression model with an R2 for the train data of 0.889, but with an R2 score of -5.60. This model performed well with the training data but fail to predict with the test data.

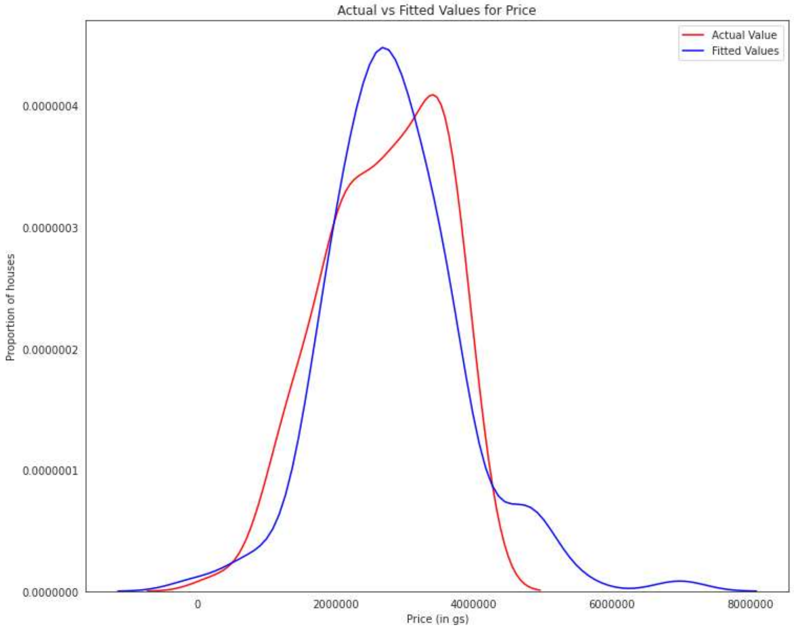
Prediction training data



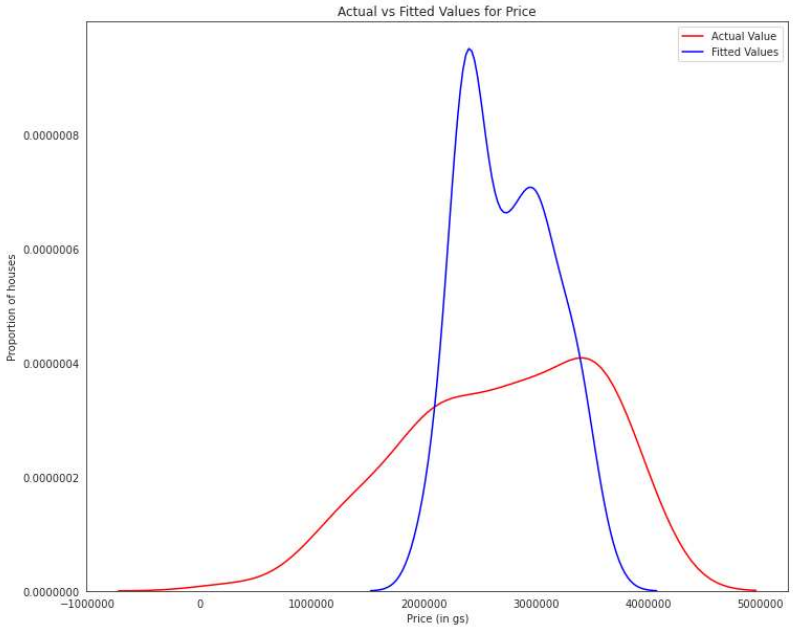
Prediction Test Data



The second model we have tried was the Ridge Regression, transforming our data to polynomial of second grade. Where we've used the gridsearch to find the best parameters for the model based on our data. The R2 of this model for the training data was 0.892, similar to the linear regression, but this model performed a bit better with the trainning data, getting a R2 score of -0.455.



The third model tried was de K Neighbors Regressor, which perfromed the worse with the trainning data, getting a R2 of 0.40 meanwhile having a 0.03 R2 score with the test data.



5. Results and discussion

In this project we determined two objectives to solve, the first one was to get insights to the kind of neighborhoods were in Asuncion, using KMeans algorithm classified each neighborhood in 8 clusters, getting that most of these neighborhoods were grouped in one cluster, giving the impression that the city is homogenous. Also, we have noticed that it seems that the border neighborhoods were part of same cluster with exemption of clusters with just one neighborhood. With these results a person gets one big cluster to start looking, a cluster of border neighborhoods influenced by the near cities, or 6 quite specific neighborhoods. One good thing of this model was that there was a group of ideal neighborhoods for a tourist who does not want to rent but wants to stay in a hotel and seize most of the city sightseeing.

For the second part of the project the regression models were used, the performance of the three were very poor at predicting a price using the test data, but with high scoring results with the test data. All of the models had high bias, probably as a result of the greater quantity of features compared with the number of observations, specially in the test data, in this part probably the solution would have been put a greater limit in the price and use more observations of our data. Also these results lets us think that there other variables with more relevance than the venues near each neighborhood or even the size or the quantity of rooms of a property, for example the model could have been improved if there was set some place of interest in the city and distance between each property, or the access to transport to each neighborhood, the level of security near each place. A positive aspect of two linear models, was that around 89% of the price variance could be explained by the features we have selected for the modeling.

6. Future Directions

Based on the findings of this project, we can conclude that there are more aspects to consider when choosing which a suitable neighborhood is or to predict with high precision the fair value to rent, not only the activities and the sightseeing around them or the rental data. Considering this point, future research could be improved adding data such as the transportation access, security level, ratings of the venues (having quantity does not mean quality), points of interest, in addition to extend the universe of observation data to prevent bias when training the model.

Finally, the use of more complex algorithms such as random forest or xgboost could have get better results in predicting price.