**Universidad de los Andes**

**Facultad de Economía**

**Big Data y Machine Learning para Economia Aplicada**

PhD. Ignacio Sarmiento-Barbieri

**Merit Tejeda:** 202210104

**Celin Hernández:** 202210067

Repositorio Url: <https://github.com/chernan77/Taller_2>

**Problem Set 2: Making Money with ML?**

*“It’s all about location location location!!!”*

# Introduction

A new start-up dedicated to buying and selling properties just hired you and your team to develop a predictive model. Their objective is to buy the most properties in the neighbor-hood of Chapinero in Bogot´a, Colombia while spending as little as possible.

The company has a sample of individual property data on Bogota´ from <https://www.>[properati.com.co](https://www.properati.com.co/). However, information about properties in Chapinero is mostly missing. The company want’s to avoid Zillow’s fiasco.[[1]](#footnote-1) Zillow developed algorithms to buy houses. However, their models considerably overestimated the price of homes. This over- estimation meant losses of about USD 500 million for the company and an approximate reduction of 25% of their workforce. There are two expected outputs:

* 1. A .pdf document.
  2. Submissions with your team’s predictions in Kaggle. To join the competition use the following [link](https://www.kaggle.com/t/f2da8bf2ad9940169c27a709321df793).

# General Instructions

The main objective is to construct a predictive model of asking prices. From Rosen’s landmark paper ”Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition” (1974), we know that a vector of its characteristics, describes a differentiated good.

In the case of a house, these characteristics may include structural attributes (e.g., number of bedrooms), neighborhood public services (e.g., local school quality), and local amenities (e.g., crime, air quality, etc). Thus, we can write the market price of the house as:

However, Rosen’s theory doesn’t tell us much about the functional form of In this problem set, you will explore different models to yield the best prediction possible.

The document must contain the following sections:

* Introduction. The introduction briefly states the problem and if there are any antecedents. It briefly describes the data and its suitability to address the problem set question. It contains a preview of the results and main takeaways.
* Data[[2]](#footnote-2). In this problem set, you are required to add expand the variables in your data (remember to expand the training and testing data), at a minimum you have to add six extra variables:
  + At least 4 predictors coming from external sources; these can be from open street maps.
  + At least 2 predictors coming from the title or description of the properties. When writing this section up, you must:
    1. Describe the data, it’s suitability for the problem, and the sample construction process, including how the data was cleaned, combined, and how new variables were created.
    2. Include a descriptive analysis of the data. At a minimum, you should include a descriptive statistics table and two maps with its interpretation. However, I expect a deep analysis that helps the reader understand the data, its variation, and the justification for your data choices. Use your professional knowledge to add value to this section. Do not present it as a “dry” list of ingredients.
* Model and Results. This section presents the model with the best score submitted for evaluation. When writing this section up, include:
  + An explanation of the variables used to train this model, remember to use the variables you added in the previous section.
  + A detailed explanation on how it was trained, the selection of hyper-parameters, and any other relevant information.
  + A section comparing the performance of the best-scored submission to other submissions submitted to Kaggle. These submissions must include specifications trained using Linear Regression, Ridge, Lasso, Elastic Net, CART, Random Forest, and Boosting models. Please indicate in your submission file the name of the used model.

Conclusions and recommendations. In this section, you briefly state the main take- aways of your work

# Additional Guidelines

* Predictions have to be submitted on [Kaggle](https://www.kaggle.com/competitions/uniandes-bdml-202320-ps2). Check the competition website for more information.
* Turn a .pdf document in Bloque Ne´on. The document should not be longer than 8 (eight) pages and include, at most, 8 (eight) exhibits (tables and/or figures). Bibliography and exhibits don’t count towards the page limit. You are welcome to add an appendix, but the main document must be self-contained. Specifically, a reader should be able to follow the analysis in the paper and be convinced it is correct and coherent from the main text alone, without consulting the appendix.
* The document must include a link to your GitHub Repository.
  + The repository must follow the [template](https://github.com/ignaciomsarmiento/PS_Repo).
  + The README should help the reader navigate your repository. A good README helps your project stand out from other projects and is the first file a person sees when they come across your repository. Therefore, this file should be detailed enough to focus on your project and how it does it, but not so long that it loses the reader’s attention. For example, [Project Awesome](https://project-awesome.org/) has a curated list of interesting READMEs.
  + Include brief instructions to fully replicate the work.
  + The main repository branch should show at least five (5) substantial contributions from each team member.
  + The code has to be:

∗ Fully reproducible.

∗ Readable and include comments. In coding, like in writing, a good coding style is critical. I encourage you to follow the [tidyverse style guide](https://style.tidyverse.org/).

* Tables, figures, and writing must be as neat as possible. Label all the variables included. If you have something in your figures or tables, I expect they are addressed in the text. Tables must follow the [AER format.](https://www.aeaweb.org/journals/aer/styleguide)

1. **Introducción**

El objetivo principal de este trabajo es realizar predicciones de los precios de venta de propiedades (casas y apartamentos) de la localidad de Chapinero; de acuerdo a la Cámara de Comercio de Bogotá, Chapinero representa el 5% del área total de Bogotá y es una zona donde predominan estratos altos, el 45% de los predios corresponden a estrato 6, un 30.8% a estrato 4 y un 11.7% se ubica en estrato 5; por tanto, es la localidad de Bogotá con el mayor índice de Condiciones de Vida. [[3]](#footnote-3) En este contexto, los precios de las propiedades de Chapinero y según el [Observatorio Técnico Catastral de Bogotá (2020)](https://www.catastrobogota.gov.co/sites/default/files/Dinamica%20Inmobiliaria%20Bogota%20Region%202017%20-%202020_20210331_VersionDiseno_v1_20210520.pdf) se ubican en los más alto de Bogotá, solo superados por los precios en la localidad de Usaquen.

Para predecir los precios de casas y apartamentos en Chapinero se utilizará el planteamiento de [Rosen (1974)](https://www.journals.uchicago.edu/doi/epdf/10.1086/260169), en el cual existe una matriz de características, que incluye variables relacionadas con atributos estructurales (número de habitaciones, baños, área construida, etc.), servicios públicos del vecindario (por ejemplo, calidad de la escuela local) y servicios locales (por ejemplo, delincuencia, calidad del aire, etc.). Adicionalmente, es importante vincular dentro del modelo hedónico el análisis de datos espaciales [(Delago, et al, 2021)](https://publicaciones.eafit.edu.co/index.php/ingciencia/article/download/6772/5323/23871) para tener una visión más amplia del comportamiento de los precios en el mercado. La idea de incluir el componente espacial se refiere a la relación que tiene la fijación de precios de propiedades en función de las características de la región donde se ellas se ubican, tal el caso de los accesos o cercanías a centros educativos, transporte público y/o centros comerciales, factores que guardan una relación positiva, al menos intuitivamente, con el precio del inmueble. Esta perspectiva se ha empleado por diversos autores con el fin de cuantificar la calidad del vecindario, tal el caso de la relación de los efectos del vecindario con los precios de la vivienda [(Dubin, 1992)](https://doi.org/10.1016/0166-0462(92)90038-3).

Este trabajo busca generar predicciones de los precios de casas y apartamentos de Chapinero, a partir de la relación del precio con datos estructurales y características espaciales de las otras localidades de Bogotá; un supuesto subyacente en este análisis, es que las demás regiones de Bogotá y el modelamiento econométrico entre sus variables son capaces de inferir los precios de Chapinero.

Esto plantea un desafío importante en términos del modelamiento econométrico, dada la heterogeneidad de las localidades de Bogotá y como al combinar los datos estructurales y características espaciales de estas zonas se pueden predecir los precios de Chapinero; para tal propósito se construirán una serie de modelos econométricos que incluyen Linear Regression, Ridge, Lasso, Elastic Net, CART, Random Forest y Boosting models. El conjunto de datos a utilizar proviene de [https://www.properati.com.co](https://www.properati.com.co/) que contienen una muestra de datos de propiedades para Bogotá, Colombia.

1. **Datos**

Los datos utilizados para este análisis comprenden una base de datos de propiedades (casas y apartamentos) segmentadas para latitud y longitud en la ciudad de Bogotá, Colombia, en el período de abril 2019 a agosto 2021. Los modelos de precios hedónicos se basan en la estimación de relaciones econométricas entre el precio de las propiedades, respecto a datos estructurales y características espaciales de las regiones donde se ubican estos inmuebles.

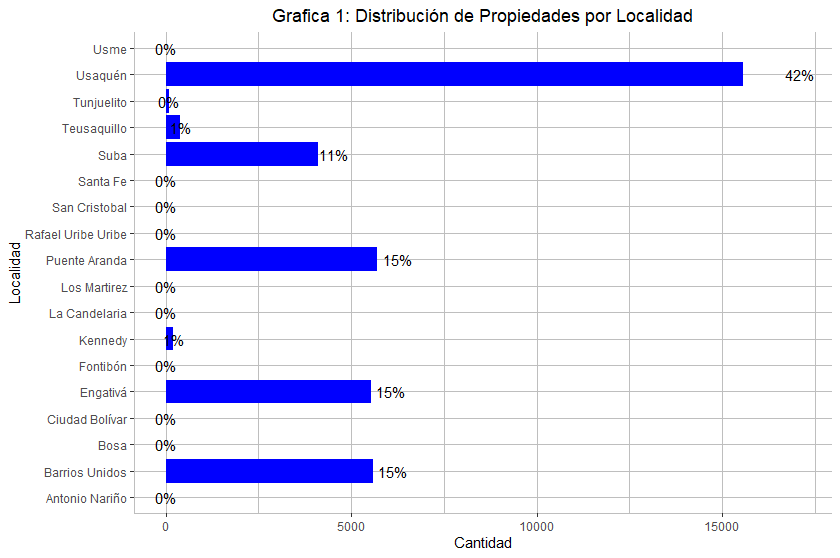
De la base de datos original se identificaron, depuraron y construyeron un conjunto de variables, que explicaran el precio de casas y apartamentos:

1. **Datos Estructurales:**

Dentro de las variables que describen datos estructurales de las propiedades y, a partir de la descripción de estos inmuebles se construyeron las siguientes variables binarias:

1. **Características Espaciales**

Las variables de distancias se obtuvieron considerando los límites geográficos de las ubicaciones de propiedades en Bogotá, de acuerdo a sus registros de latitud y longitud. Primero, se utiliza la función getbb() para obtener los límites geográficos de Bogotá, luego, con la función opq() se crea una consulta de Overpass API que busca los tipos lugares, como restaurantes, centros comerciales y educativos, etc. dentro de los límites geográficos establecidos. La función osmdata\_sf() permite cambiar el formato de los datos a un objeto sf y finalmente se genera la variables de distancias a las propiedades en metros cuadrados. Es importante mencionar que a partir de los datos de latitud y longitud se obtuvo la localidad correspondiente a cada propiedad.



La grafica 1 muestra la distribución de propiedades por localidad, según estos datos, la mayor parte de propiedades se ubican en Usaquen (42%), Puente Aranda (15%), Engativa (15%) y Barrios Unidos (15%).

1 Bajo-bajo,

2 Bajo,

3 Medio-bajo,

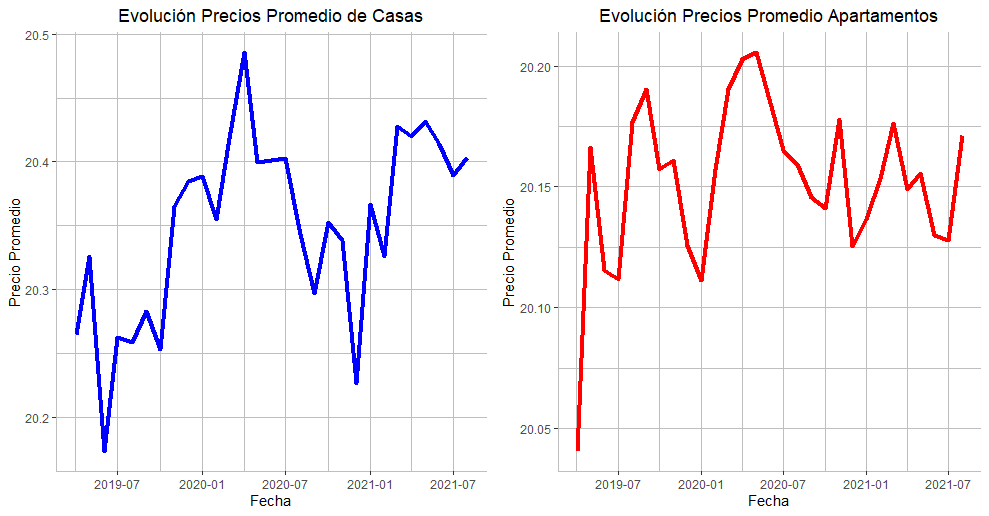
4 Medio,

5 Medio-alto,

6 Alto

Finalmente, y de acuerdo a datos de la caracterización socioeconómica de la Alcaldía Mayor de Bogotá, se identificó el estrato promedio de cada localidad.

**Análisis descriptivo de los datos**



Gráfica 2: Evolución Mensual de los precios de Casas y Apartamentos,

Al obtener los precios promedio mensual para las propiedades (casas y apartamentos), no se observa una dinámica común en ambas series; sin embargo, ambos precios alcanzaron su punto máximo en abril 2020, lo cual podría estar relacionado con los efectos de la pandemia Covid-19 sobre el sector inmobiliario.



**Gráfica 3:** Precio por M2 de Casas y Apartamentos, según localidad

Al analizar los precios por M2 se observa que las casas con precios más elevados se encuentran en la localidad de Los Mártires; en cambio, la localidad Bosa muestra los precios más bajos, estas dos localidades no son representativas y en conjunto representan menos del 1% de la muestra de datos. En el caso de los apartamentos, los precios más altos nuevamente se ubican en Los Mártires, si bien en esta localidad predomina el estrato 3, el precio por M2 de casas y apartamentos indica la existencia de inmuebles no tan extensos, pero con un valor de mercado bastante alto.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tabla 1:** Apartamentos: Estadísticas de variables seleccionadas | | | | | |
| ================================================================================ | | | | | |
| Estadísticas | N | Mean | St. Dev. | Min | Max |
| -------------------------------------------------------------------------------- | | | | | |
| Precio | 28,093 | 620,780,098.00 | 294,379,968.00 | 300,000,000 | 1,650,000,000 |
| Precio\_M2 | 28,093 | 4,945,436.00 | 2,263,592.00 | 542,209 | 44,852,378 |
| Habitaciones | 28,093 | 3 | 0.78 | 1 | 11 |
| Baños | 28,093 | 2 | 0.79 | 1 | 11 |
| Área | 28,093 | 137.224 | 58.757 | 15.161 | 571.735 |
| M2\_por\_Habitación | 28,093 | 53.16 | 18.535 | 14 | 95 |
| Dist\_Parques | 28,093 | 162.436 | 102.752 | 0.991 | 3,344.62 |
| Dist\_Transp\_Público | 28,093 | 8,467.65 | 3,143.53 | 26.193 | 15,649.56 |
| Dist\_Establecimientos | 28,093 | 174.332 | 117.549 | 0.99 | 1,539.72 |
| Dist\_C\_Comerc | 28,093 | 541.615 | 329.281 | 1.216 | 3,803.65 |
| Dist\_Centros\_Educ | 28,093 | 268.243 | 165.331 | 0.579 | 1,357.69 |
| Dist\_Restaurantes | 28,093 | 219.491 | 150.244 | 1.171 | 1,506.62 |
| Dist\_Bancos | 28,093 | 618.33 | 443.044 | 3.498 | 3,999.40 |
| Estrato | 28,093 | 4 | 0.44 | 2 | 4 |
| ------------------------------------------------------------------------ | | | | | |

La tabla 1 presenta un conjunto de estadísticos de las variables cuantitativas seleccionadas y construidas de la base de datos de propiedades, de acuerdo con esta información, el precio promedio de los apartamentos es de COP 620,780,098.00, al obtener el coeficiente de variación:

Un C.V alto indica una mayor heterogeneidad de los precios; en tanto que menor C.V, implicaría homogeneidad en esta la variable. Dado que este coeficiente es 0.4742, lo cual se considera alto, ya que denota una fuerte variabilidad en los precios, en este sentido, el promedio no es representativo del conjunto de datos y existe una importante heterogeneidad en esta variable.

1. **Modelos and Resultados**
2. **Conclusiones y recomendaciones**

1. For more info, see the following article [here](https://insidebigdata.com/2021/12/13/the-500mm-debacle-at-zillow-offers-what-went-wrong-with-the-ai-models/). [↑](#footnote-ref-1)
2. This section is located here so the reader can understand your work, but it should probably be the last section you write. Why? Because you are going to make data choices in the estimated models. And all variables included in these models should be described here. [↑](#footnote-ref-2)
3. Cámara de Comercio de Bogotá, Perfil Económico y Empresarial, localidad de Chapinero. [↑](#footnote-ref-3)