# Analysis of Real Estate Prices in Mexico

\*\*Bootcamp Final Project\*\*

## Overview

After analyizing several options, we decided to select this project because it involves almost all topics covered in the course, from webscrappng, ETL, working with Pandas, SQL, reading csvs, visualization and machine learning. Also, we think it could be useful for people looking into buying a new home or simply for those seeking to maximize investing in real estate in Mexico. Furthermore, this analysis could expand in determining price changes overtime by scrapping information periodically.

The purpose of this analysis is to determine a model that predicts real estate prices by taking into consideration several features that answers the following questions:

1. Where are the places in Mexico with the highest and lowest prices per square meters by asset type (house, apartment, land)?

2. According to the predicted prices, where are the best opportunities of undervalued assets?

Our analysis includes several steps to achieve our goal:

1. \*\*Webscraping\*\*. Create a python code with beautiful soup and splinter to retreieve asset data from several webpages, such as metroscubicos.com and create a raw data csv.

2. \*\*ETL\*\*. Raw data needs to be cleansed, transformed and add new information from other sources to generate a final SQL database.

3. \*\*ML model\*\*. We will use a supervised linear regression machine learning model that helps us determine our price target variable based on features such as location, type, land size, construction square meters, number of rooms, number of bathrooms, etc.

4. \*\*Dashboard\*\*. We will create visualizations and interactions to display our results in Tableau.

## Team Members and roles

|Member |Segment 1 Role |

|:------------:|:-------------------------------------------:|

|Brenda Treviño|ML model ![Triangle](/Resources/triangle.png)|

|Luis Carmona |Technology ![Red](/Resources/x.png) |

|Carlos Acosta |Database ![Green](/Resources/circle.png) |

|Jorge Castro |Repository ![Square](/Resources/square.png) |

## Resources and Techonology

- Data source: raw data scraped with from real estate websites. This raw data will be stored as csv.

- ETL process: using Pandas for reading and transforming csv to dataframes for cleaning and transformation, as well as to merge additional data. Resulting data will be stored into an SQL database.

- ML model: We'll use scikit libraries to create a linear regression model and test other models to get the best accuracy for pricing prediction.

- Dashboard: Using Tableau we will import our data and try to use geographical coordinates to show our analyses by entities.

- Software use to perform the analysis: Jupyter Notebook v6.4.5, Pandas, , SQLite, Scikit-learn v0.24.2, Tensorflow 2.0 and Tableau.

## Github

EXPLAIN BRANCHES CREATED AND COMMITS

### Communication Protocols

For the purpose of communicating among the team members, we decided to create a group in Whatsapp. The reasoning as why not to use Slack is because all team members have different activities and schedules so checking Slack throughout the day is not as common as checking our smartphones for general messages.

It seems to be working.

Also, we decided to establish zoom meetings, outside those for classes, if there is something we need to review among us.

## Database

The following image shows our two main tables in our database:

![database tables](/Resources/relationalTables.png)

Besides the data from the webscraping we will obtain information regarding location (lat, lon) for each \*\*\*municipio\*\*\* so that we can plot them on a map, and make analysis by zip codes.

Our mockup database is /Resources/mockData.csv and contains what we think we might get from the webscrapping process. The following image shows the structure of the csv loaded into a pandas DataFrame:

![database preview](/Resources/databaseDF.png)

## Machine Learning model

For this particular project out model will be done with supervised learning, using examples and labels to find patterns in our data. We will be using a linear regression model under the assumption that the reationship between the price (dependent continuous variable) and the other more explanatory house features is linear.

It is also worth mentioning that this is a regression model because we wnt to predict a numerical outcome (price). Because of this, our first approach is the Random Forest classifier, in order to learn from data and to aggregate their resulting predictions.

The following image shows our first attempt to model our data after all the necessary imports:

![ML\_model1](Resources/ML\_model1.PNG)

We are aware that this is our first approach and that we will have to go through a number of iterations and even alterations to get our model up to a high standard.