# Final Capstone - Predicting Real Estate Prices

IBM - Capstone

**by**

**Hector Hernandez Heimpel**

# Content

1. Introduction
2. Data Sources
3. Methodology
4. Results
5. Discussion
6. Conclusion

# Introduction

# Background

- What the right price is for a real estate listing?
- We usually rely on data regarding similar listings and build an idea from there.
  - We tend to do this without much analytical rigor.
  - Build a predictive model for real estate prices.
  - Predict the prices of real estate locations in Mexico City.

# Stakeholders

- Model can help real estate agencies in the pricing of new listings,.
- Help customers get an idea as to the price of real estate they might be searching for.
- It can flag relatively cheap real estate, etc.
- To build the model the Foursquare API can be hugely beneficial because it can identify commercial locations around the real estate and not only the usual features of a listing.

# Data Sources

# Real Estate Listings - Mexico City

- The first data-set contains real estate listings from a real estate agency's web-page. This data set was built by scraping the following url: https://www.remax.com.mx/propiedades/ciudad+de+mexico_ciudad+de+mexico/venta.

- This data set contain the features or characteristics of listed real estate properties such as number of rooms, area, neighborhood and so forth; it also contains the price of each listing.

# Foursquare Venue Data

The second database I use, which is merged with the previous data set, is foursquare venue data which contains commercial venues around a defined radius of each listing in the previous data set. To get this data I will use the foursquare API and the proper queries.

# Methodology

# Methodology

I used a linear regression model with real estate listing price as the dependent variable and the listing's features as independent or predicting variables.

- Continuous Variable - Price
- Easy interpretation of results

# Data Processing

- Neighborhoods into clusters (K-means clustering) - Choose number of clusters with highest out of sample R^2
- Polynomial fit degree = 1 (Overfitting data with more degrees.)
- Standardize Data
- Drop observations containing missing variables

# Results

# Results

Table 1: Coefficients in Real Estate Price Regression

| | |
|---|---|
| room | 0.0 |
| bathroom | -841581.94 |
| m2 | 331197.41 |
| parking | 19633.38 |
| commercial_count | 2974303.56 |
| type_Bodega | -30713.33 |
| type_Casa | -8791145.54 |
| type_Comercial | -11438220.61 |
| type_Departamento | -10165014.68 |
| N_cluster_7_1 | -6870873.17 |
| N_cluster_7_2 | -226706.43 |
| N_cluster_7_3 | -2145533.37 |
| N_cluster_7_4 | -3373129.76 |
| N_cluster_7_5 | -1672112.90 |
| N_cluster_7_6 | -2902659.77 |
| Observations | 168 |
| Out of sample $R^2$ | 0.34 |

# Discussion

# Discussion

- The best insight we got from the data is that neighborhood zero when making seven cluster is the most expensive neighborhood and that the number of commercial venues is a significant predictor of price; the more commercial venues there are the higher the price.
- The best R^2 that I managed to get in an out of sample data was 0.34. Which can be interpreted as 34% of the variance in price is explained by the features.

# Conclusion

# Conclusion

In this capstone I set about building a predictive model in order to predict the price of a real estate listing given its features. To build the data set I scraped a web-page for their real estate listings and then I used the foursquare API to get the nearby commercial venues.

Next, I searched through different linear regression models, including polynomial features and different neighborhood clusters, and found many of them to over-fit the data; I ended up using a normal linear regression with no polynomial features and with 7 clusters of neighborhood as the best predictive model.