Project 2 Proposal - Web Scraping and Linear Regression
Vancouver and Seattle House Price Prediction
Flora Xinru Cheng

Introduction

Seattle and Metro Vancouver share similar climates and locations on the Pacific Coast. Both cities also offer active lifestyles and great views, and attract a lot of Millennials who are potential first-time homebuyers. For this project, I am interested in predicting the sale price of a home that meets certain filter requirements (such as size, neighbourhood, number of bedrooms, etc.).

The ultimate goal is to use the same model for datasets of both cities, and compare the predictions after applying the same set of filters, to see which is less expensive.

Potential clients: Homebuyers looking to relocate to either Vancouver or Seattle, or people looking to invest in real estate in either of the two cities

Questions we might answer for clients with data:

- How much does a certain categorical variable (HasBasement, or HasYard) increase the home price?
- Which feature has a larger impact on SalePrice, YearBuilt or YearRemodelled?
- Given a set of filters (2 bed 2 bath, 2-story townhouse, built after 2010, with view), is it cheaper to buy a home in Vancouver or Seattle?
- (if have time) For a fixed size, which neighbourhood has the cheapest homes (plot on map)?

Data

The data will be scraped from the web, from Zillow and/or Redfin (both have Canadian versions of their websites), and then stored as flat files. The recommended complexity of the dataset is 1000+ outcomes (rows) and 10+ features (columns). Some possible features to consider are:

Numerical Data:
SalePrice (y and y_predicted), Size, YearBuilt, YearRemodelled, GarageSize, Number of Stories, Number of Bedrooms, Number of Bathrooms, Number of Days Listed on Website

Categorical Data:
HomeType, Neighbourhood, HasParkingSpots, HeatingType, SaleCondition, HasBasement, HasBalcony, HasYard, HasView (City, Mountain, Park, Water), ForSaleBy(Owner or Agent)

Other:
Location Data (coordinates for mapping)

Tools
BeautifulSoup, Selenium
Scikit Learn, Pandas, NumPy
Seaborn, Matplotlib

Known Unknowns/Questions/Things to Note:

Are the home listings pages dynamic? Can we scrape them using BeautifulSoup?

Additional info (mainly categorical variables) are on separate pages for each listing, will probably need to navigate to each page using Selenium, then scrape within page using BeautifulSoup.

The .ca and .com websites have different designs so might need to use a different set of code for scraping each. In the case where there's not enough time to do both, will prioritize scraping Vancouver data.

On Zillow, one can set a price limit, but there is no currency type attached to it. Their search function treats CAD and USD the same, so it will limit to $500,000 CAD in Canada and $500,000 USD in the U.S. (currency conversion required after final prediction for direct comparison).