Case – Real Estate

Instructions

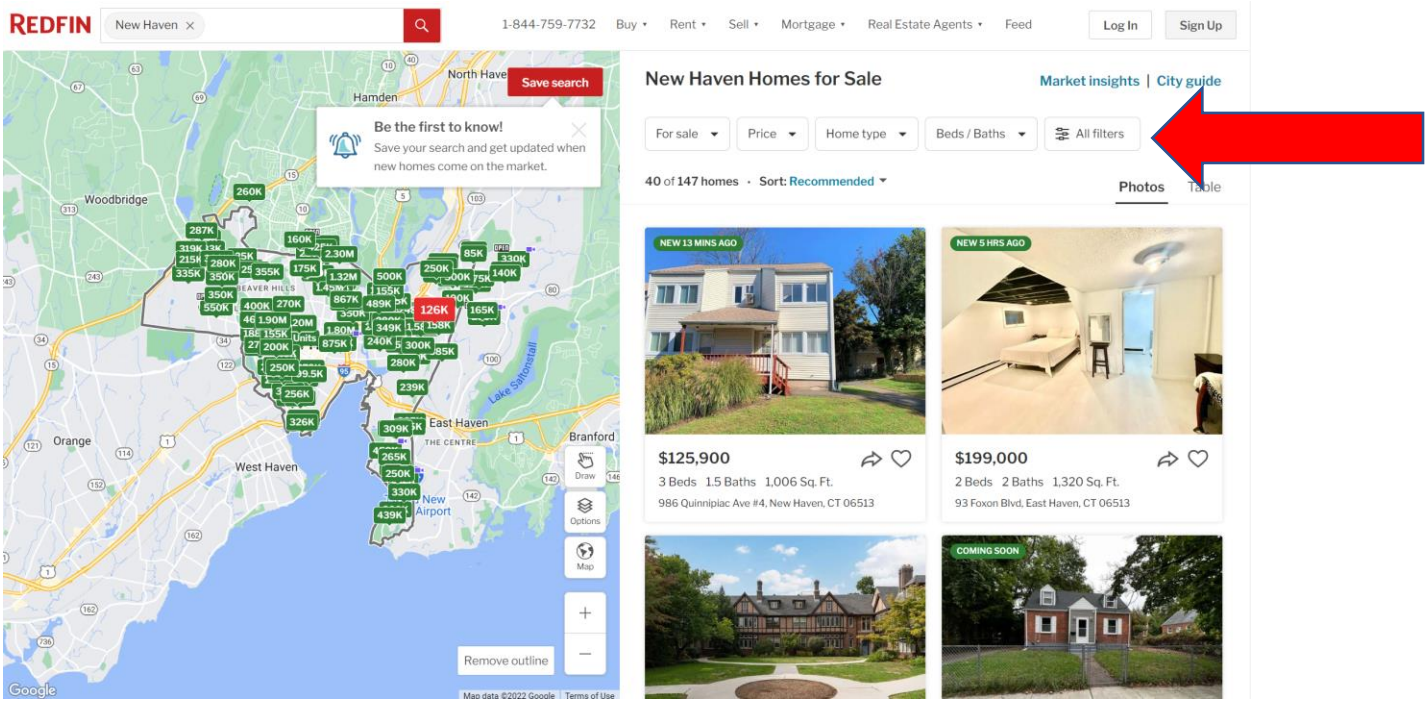
Suppose you are hired as a team of consultants by a real estate developer who wishes to better understand a specific residential real estate market to identify investment opportunities. Your goal is to **develop a regression model with multiple regressors to predict the sale price of a home in your selected market**.

Each learning team should prepare solutions to the questions below and be prepared to discuss them in class. The required deliverable is a PowerPoint presentation (≈ 10 slides) highlighting your findings to your client. One member should upload your team's slides, a video of your presentation (15mins max) and **the complete dataset and any related R code** on BruinLearn. Please provide your group number and the names of all group members. Everyone whose name appears on the write-up is assumed to have actively participated in its preparation.

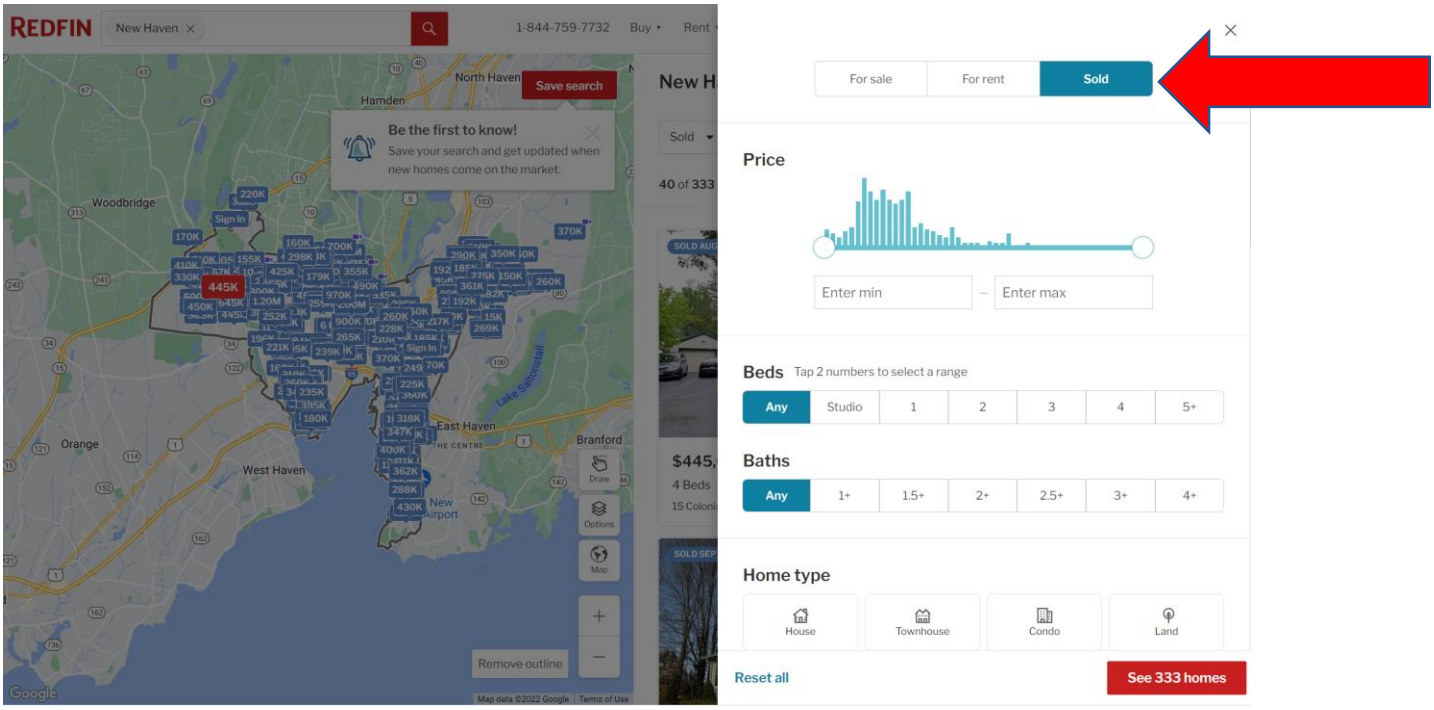
Data Summary

Each team should download a Redfin dataset of home sales for one U.S. city over the last 12 months. Please select one team member's hometown or a city you are familiar with. If you select a large city, try to limit your search to a particular neighborhood (e.g., Midtown Manhattan; East Bay, San Francisco). You may focus on just 1 type of home (e.g., single-family, condo, etc.) depending on your selected location. Aim to have **at least 300 observations** in total.

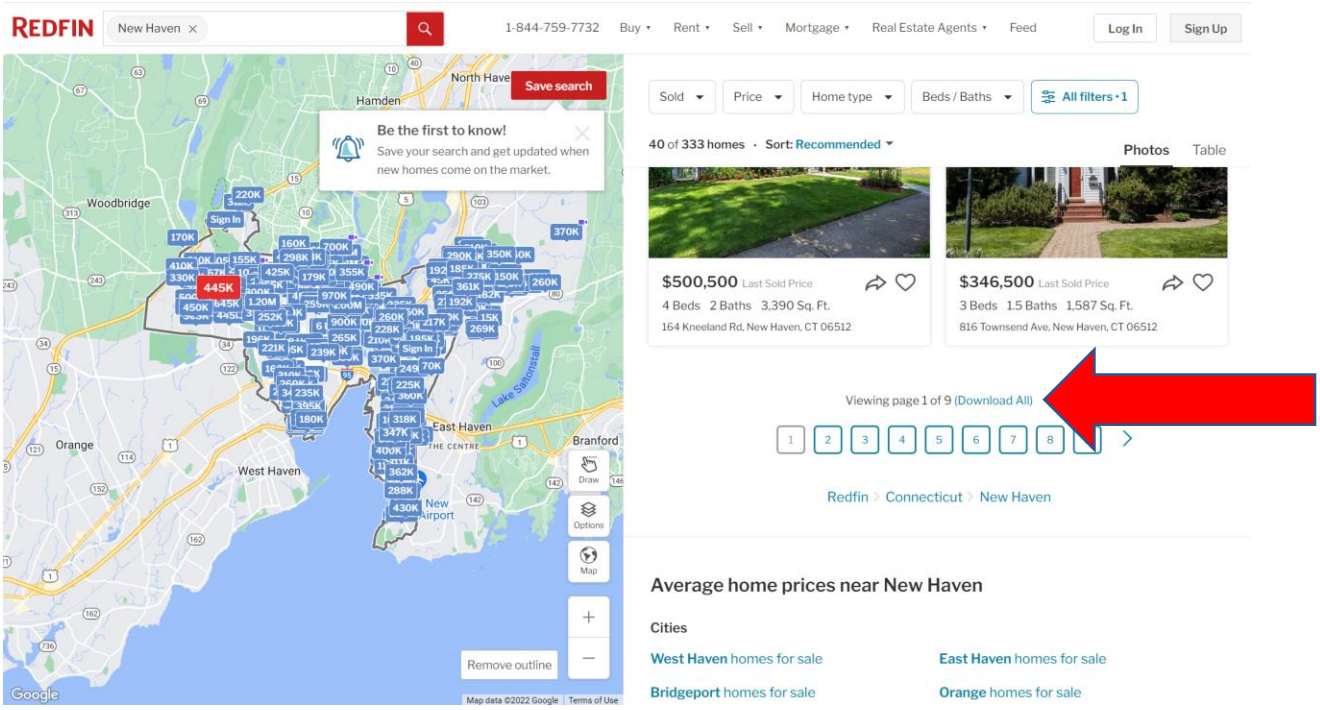
To download data: Enter the location at Redfin.com. For example, below is for New Haven, Connecticut



Click on “All Filters”, and then click on the “Sold” button to restrict your view to Sold listings.



On the right panel, below the photos is a “(Download All)” link to download the data. Redfin limits the download size to 350 observations. If your city has more than 350 sales in the past year, you can download in batches (e.g., all 1 bedroom, 2 bedrooms, etc).



When you load in your CSV file: you should use the following more detailed read_csv() command:

CHOOSE_A_NAME_df <- read_csv('YOUR_DOWNLOADED_FILE', name_repair='universal')

The additional argument just converts the column names in the csv that have spaces to column names with a period (e.g., “SOLD DATE” to “SOLD.DATE”). This just makes it simpler to refer to the variable in R (e.g., R would normally look at “SOLD DATE” and think it is referring to two variables).

A summary of key variables you may wish to explore is given below.

Variable	**Definition**
SOLD.DATE	Calendar date of property sale
PROPERTY.TYPE	Category of home: single family, condo, or townhouse
ADDRESS	Street address of property
CITY	Location of property
STATE.OR.PROVINCE	Location of property
PRICE	Property sale price ($)
BEDS	Number of bedrooms

BATHS	Number of bathrooms
LOCATION	Neighborhood
SQUARE.FEET	Indoor property area (square feet)
LOT.SIZE	Outdoor property area (square feet)
YEAR.BUILT	Year of construction

Questions to Guide Your Presentation

1. For each house, calculate the price per square foot. Report the average price per square foot in your selected city?

2. Run a simple linear regression taking the following form:
$$SALE.PRICE = \beta_0 + \beta_1 SQUARE.FEET + \epsilon$$
Interpret β_1 and its statistical significance.

3. Run a regression model with the additional explanatory variables of BEDS and BATHS (in addition to SQUARE.FEET). Write out the regression equation. What do you notice about the additional coefficients and how existing coefficients change?

4. If you are analyzing homes other than single family homes (e.g., condos, townhouses), what do you notice when you include LOT.SIZE in the regression?

5. What fraction of the variation in home prices is explained by your selected variables? How does your model change with the addition of other variables?

6. Use your model to predict a typical home price. For example, what is the predicted sale price of a 2-bedroom, 2-bath home with 1,500 square feet (or for a typical home in your chosen area)? What is the 95% confidence interval for your prediction?

7. How might you use your model to identify investment opportunities in the current housing market in your city?

8. Based on your knowledge of the city, what other variable(s) would you recommend including to better predict residential home prices? Pick one or two of these variables and collect the data for a small sample of the homes listed. Interpret your additional variable. Alternatively, consider different models or model specifications. Argued for the importance of your additional analysis for identifying investment opportunities. Be creative!

Grading

You will be graded on:

- Correctly producing and interpreting your regression outputs (70% weight)
- Business insightfulness, creativity and effective communication (30% weight)