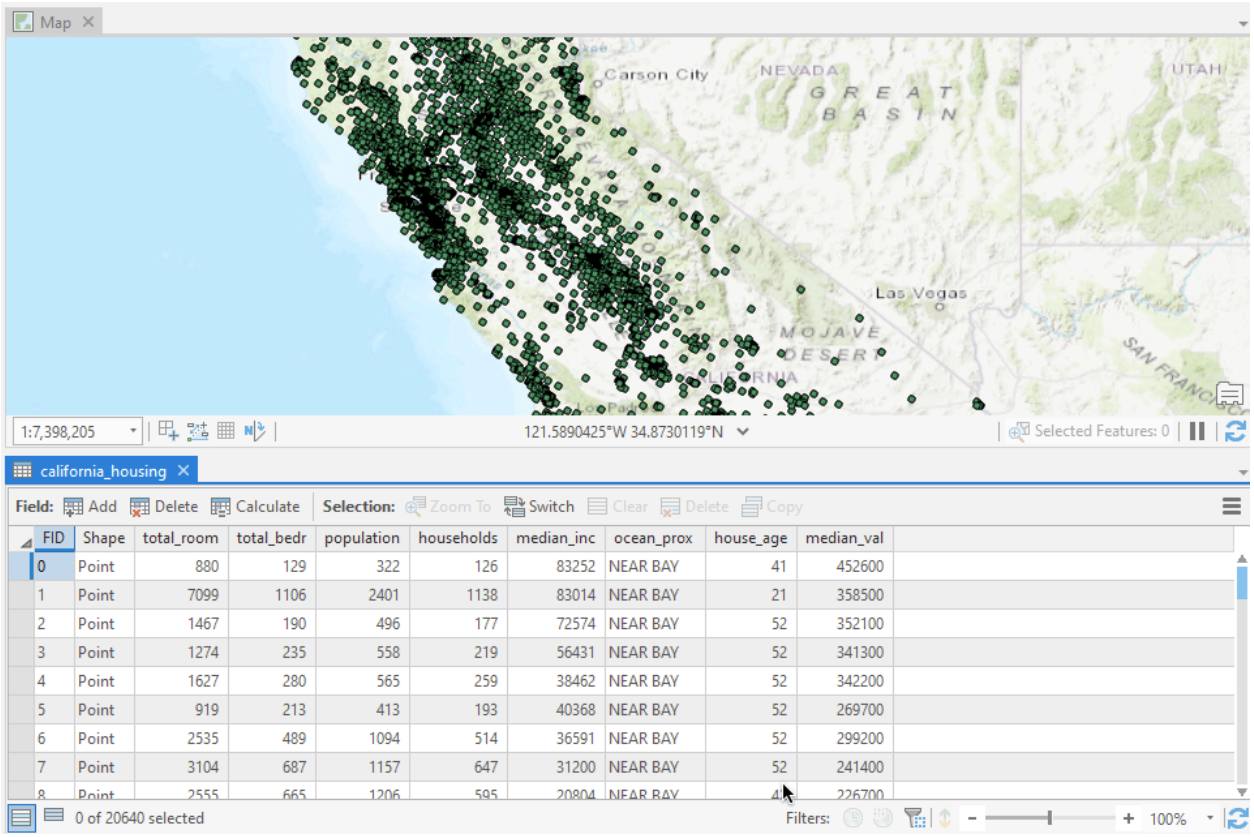
Lab 4: Making predictions using GIS, regression, and machine learning

In this lab, you will work with a shapefile dataset that contains median housing price in California at the census block level. The shapefile contains the central points of the census blocks, and it also contains some other attributes related to housing in each census block, such as the median house age and the total number of rooms. You can explore the shapefile a bit in a GIS (e.g., ArcGIS).



A bit more information about the shapefile "California_housing.shp": each row contains the data about one census block in California. There are a number of attributes (columns): geometry (i.e., central points of the census blocks), total_rooms, total_bedr (total number of bedroom), population, households, median_inc (median household income), ocean_proximity (the relation between this block and the ocean. Note that this is categorical data!), house_age, and median_val (median house value. This is the value we are trying to predict),).

Your goal is to build a linear regression model to predict "median_val". The metric you use to measure the performance of your model is RMSE:

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2}$$

Task 1: Load the shapefile data using GeoPandas, show its top 10 records, and visualize any 500 data records on an interactive map using the Folium library. (15 pts)

Task 2: Data preparation. The shapefile data is not directly ready for you to feed into a machine learning model (like most cases in the real world), and you will need to do some data preprocessing to prepare your data. (25 pts in total for the two sub tasks below)

(1) Remove artificial values (15 pts). Using the following code, you can get a quick histogram of all the numeric properties in your data. (Assuming "housing_shp" is the variable name that you use to store the shapefile data; you may need to change it to your own variable name)

```
housing_shp.hist(bins=50, figsize=(20,15))
```

From the histogram, you will notice the strange high-value bars for "house_age" and "median_val". This is because these two attributes were capped when the data were recorded. Housing median age will not exceed 52 years while median house value will not exceed $500,001. You can find out the maximum value of house age using:

```
housing_shp["house_age"].max()
```

You can then remove these artificial data records using the code below:

```
housing_shp_cleaned = housing_shp[housing_shp["house_age"] < 52]
```

In a similar way, you can remove the artificial data records associated with "median_val". After that, if you plot out the histogram again, you should no longer see the strange high-value bars for "house_age" and "median_val".

(2) Prepare dummy variables (10 pts). The "ocean_prox" attribute contains categorical values. In order to use this attribute, we will need to convert it to dummy variables. To see the unique values in "ocean_prox", you can use:

```
housing_shp_cleaned["ocean_prox"].unique()
```

To covert the attribute to dummy variables, we can do:

```
import pandas as pd

housing_shp_cleaned = pd.get_dummies(housing_shp_cleaned)
```

Task 3: Split the data into training and test and train a linear regression model to predict "median_val". (40 pts)

Tips:

- Set the random_state = 42 when you split the data into training and test
- Don't forget to standardize both your training and test data

Task 4: Evaluate the performance of your trained model on test data using RMSE. Create a scatter plot for the predictions and true values, and add a reference line for perfect prediction. (20 pts)

To submit:

- Lab4_ FirstName_LastName.ipynb