**Raw house dataset**

* Raw house data, is the geospatial dataset that contains coordinate information like latitude and longitude which allows features to drawn on a map.
* Geospatial data can be used to create maps and analyse data in GIS (geographic information system).
* Google colab has been used for coding exercises. I started coding the data by importing the modules and mounting the drive on colab notebook.

**Data Cleaning**

* In data preparation by using data wrangling techniques which helped remove duplicates, finding out the missing values, identify the number of rows and column, validate the datatype of the columns.
* I have describe () method to understand the description of datasets like, Mean, median , std deviation.
* After finding out the missing values, I dropped the 5000 rows and16 columns to 4370 rows and 14 columns. It was imperative to drop the attributes like “kitchen feature” and “floor covering” to plot a correlation matrix.

**Methods**

* Furthermore I used data science techniques like box plot, scatter plot, pair plot, and histogram to analyse the deep information in the dataset.
* I have created a boxplot on basis of “sold price” attributes which shows that Q1-Q3 = Q2, that means data which are out of box are outliers. Removing are outliers are good so it doesn’t create a noise.
* Scatterplot can be displayed on X and Y axis respectively on the basis “square feet” and “bedrooms” , essentially 10000 square feet has 5 to 8 bedrooms.
* Histogram has represented that maximum number of bedrooms has 4 bathrooms.