Santa Cruz Housing Analysis

Ayush Arora, Kyle Oda, Albert Garcia, Prajna Ta

A. Questions to be explored
   a. Is there a relationship between location & features of a house and its price?

B. Description of data
   a. We created two CSV files by web scraping Trulia. We then cleaned, wrangled, and combined the two datasets into one CSV to be used in our clustering algorithm. After running the KMeans algorithm, each home was now labeled with a cluster number.

C. Code References
   a. All of our code was created ourselves.

D. Viz References
   a. All visualizations were done ourselves.

E. Answers Provided or Final results obtained
   a. Due to Folium rendering constraints, we were not able to obtain a solid conclusion. However, in cluster 0, we can see that the apartments we properly clustered. We can conclude this because many of the homes had the same number of bedrooms, baths, and square footage. We may be able to obtain better results with a different library.

F. Libraries used:
   a. Pandas
   b. Sklearn
   c. Matplotlib
   d. Seaborn
   e. Numpy
   f. Bs4
   g. Folium

G. List of tasks accomplished
   a. Web Scraping
      i. Scraping data from Trulia. One of the challenges we encountered as that the data was very deeply nested.
   b. Data Wrangling
      i. The data we obtained from web scraping needed to be parsed and cleaned. All of the features were originally in one column, so we created a parser to separate them into different columns. We also removed rows that contained null values and removed unneeded columns

c. Visualization
   i. We used seaborn and matplotlib to create some visualizations to gain a better understanding of the dataset.
   ii. We also used Folium to visualize the data after we ran the KMeans clustering algorithm.