CS5610 M1: Project 1b

Exploratory Data Analysis of Real Estate Transactions in Sacramento

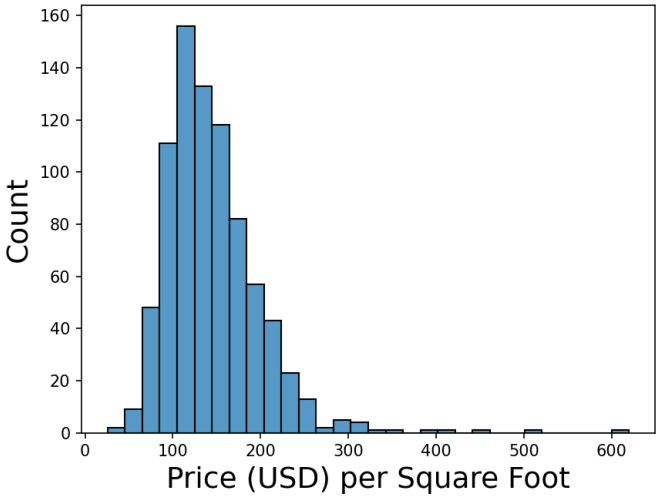
Summary of Original Data

We were given a file containing 985 real estate transactions containing the following columns:

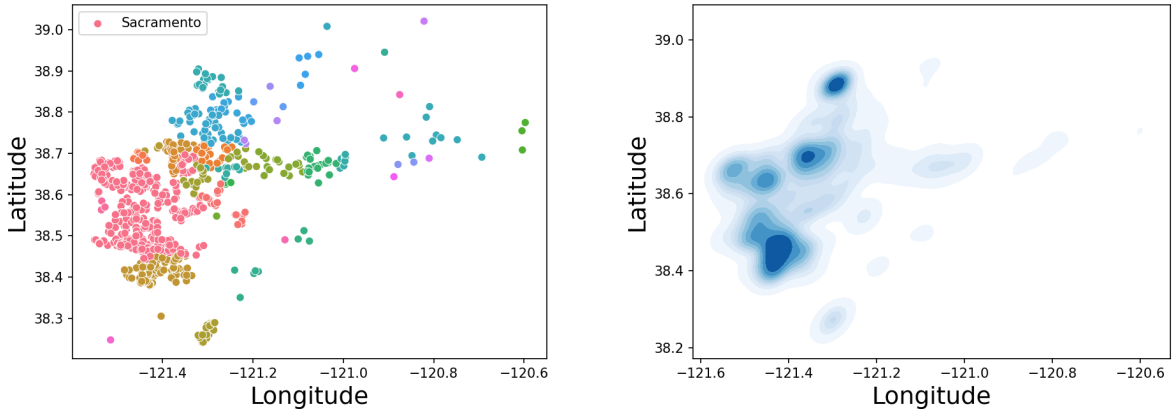
Name - Description	Value Range (mean ± std. dev.)
address - Street address (not analyzed)	981 distinct values
city - Uppercase city name	39 distinct values
zip - 5-digit zip code (does not include zip+4)	68 distinct values
state - 2 character state abbreviation	CA
beds - Number of bedrooms	0 to 8 (2.9 ± 1.3)
baths - Number of bathrooms (whole)	0 to 5 (1.8 ± 0.9)
sq__ft - Square footage of the property	0 to 5,822 ft² (1,315 ft² ± 853 ft²)
type - Property classification	917 Residential, 54 Condo 13 Multi-Family, 1 Unkown
sale_date - Date the transaction took place	May 15th to 21st, 2008
price - Sale price in whole USD($)	$1,551 to $884,790 ($234,144 ± $138,365)
latitude - Degrees north of the equator	38.2° to 39.0° (38.6° ± 0.1°)
longitude - Degrees west of the meridian	-121.5° to -120.6° (-121.4° ± 0.1°)

Patterns

The price, square footage, and price per square footage appear to demonstrate a skewed distribution. This is likely due to there being exponentially fewer large/expensive homes and would likely follow the power law distribution for wealth. The smallest values are likely due to problems with data entry ($5,000 homes don't exist, for example).



An analysis of the latitude and longitude by city shows that the clustering of properties falls within the expected range for Sacramento proper and the surrounding suburbs. The left scatter plot shows properties colored by city and the right KDE plot shows where the majority of properties were located.



Bad or Missing Data

- There was one record with the type of “Unkown” (misspelled)
- 108 records with no bedrooms listed
- 108 records with no bathrooms listed
- 171 records with a square footage of 0

Sample Size Too Small

- There were exactly 9 zip codes with only one recorded transaction.
- There were exactly 11 cities with only one recorded transaction.

Outliers

- One record with 8 bedrooms (perhaps a mansion?)
- One record with square footage above 5k (5,822 ft², next lowest was 4,400 ft²)
- 51 records with a price below $5k (next lowest was $30k)
- 172 records either below $1/ft² or over $300/ft² (mostly due to square footage of 0)

Summary of Filtered Data

We kept 794 of the original 985 records thus filtering out 191 records. The city, state, zip, and type columns were all converted to categorical data to aid in future analysis.

Based on the sales date, the data is 15 years old. So care should be taken when interpreting any results based off of the data.