Tyler Graham
CS5610
September 18, 2023

Project 1B Written Report

Project 1B purpose was to analyze a dataset of real estate purchases in the Sacramento, CA area. The dataset was a CSV file containing all of the purchases from May 15th-May 21st 2008. The information it contained can be broken up into three main groups for every purchase: location, information about the property and the purchase price and date. The location group contains the address, city, zip code, state, latitude and the longitude. The address, city and state are all initially stored in a python objects since they contain letters and numbers. The zip code is simply stored as an int64 since its just a integer number. The latitude and longitude are stored as float64 since they contain a decimal. The property information contains the numbers of beds and baths and the square foot of the property. All of these variables are contained in int64 since they are all integers. Finally, the sale data and price were given for each of the data sets and were store in an object and int64 respectively.

In this data set, there are initially 985 entries. Please see below a table of the numerical entries basic statistics (minimum, maximum, standard deviation and mean).

	Beds	Baths	Square Footage	Price	Latitude	Longitude
Min	0	0	0	$1,550.00	38.24	-121.55
Max	8	5	5,822	$884,790.00	39.02	-120.60
Standard Deviation	1.31	0.90	853.04	$138365.84	0.14	0.14
Mean	2.91	1.77	1,304	$234,144.26	38.61	121.36

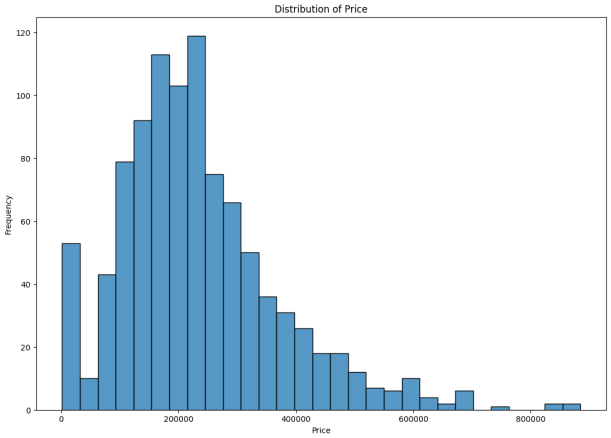
There is also 6 categorical variables. See the table below for the unique number of variables for each of these.

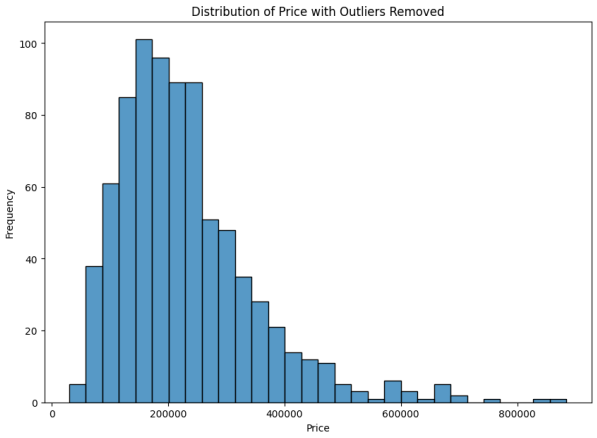
	City	Zip	State	Beds	Baths	Type
Unique Values	39	68	1	8	6	4

When analyzing the records, one of our categorical had a unique value for “Type” spelled “Unkown” and this data point was removed from the set since it was not accurate. However, the rest of the categorical did not have any data entries issues when looking at the unique values.

When plotting the numerical data using a bar chart, box chart and violin plot there were two odd relationships found in the price and the square footage data columns.

In the price category, there were 51 entries of the pricing of real estate below $5,000. The next highest after $5,000 was $30,000. The price of real estate is typically a very expensive object and it was determined to remove all of the entries below $5,000 since they were not accurate. A potential cause could be error in entry of the data since 48 of them all had the same price with zero square feet. The following table contains the statistics before and after the removal.

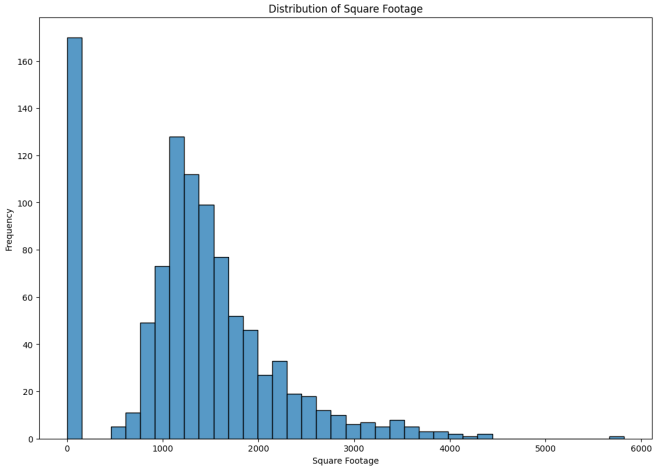


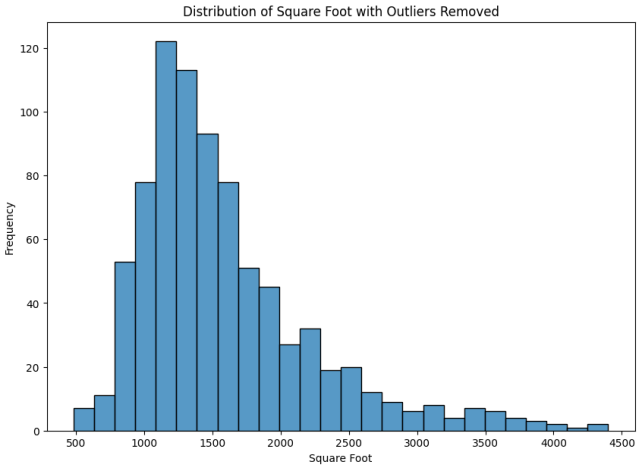


	Original Price Statistics	New Price Statistics
Min	$1,550.00	$30,000
Max	$884,790.00	$884,790.00
Standard Deviation	$138,365.84	$119,633.03
Mean	$234,144.26	$229,728.13

In addition to the price category, the square footage category contained 171 entries with 0 square feet. A piece of real estate should have above zero square foot or it is not a piece of real estate. It was determined to remove these properties. The potential reason for having these in the dataset could be it may have been land with no buildings on it or potentially just error in the data entry.

	Original Square Foot Statistics	New Square Foot Statistics
Min	0	484.00
Max	5,822.00	4,400.00
Standard Deviation	853.04	647.42
Mean	1,304.00	1585.94





Overall, with removing both of these outliers removed 172 entries from the initial data set was usable for future analysis.