Tyler Graham
CS 5610 Project 2
October 14, 2023

Project 2

In the first assignment, we were given a dataset containing all of the real estate purchases in the Sacramento California area during a week period. The dataset contains the following information: address, city, zip, state, beds, baths, square feet, sale date, price, latitude and longitude. During the first assignment, we processed the data and eliminated erroneous or outliers to have a proper dataset to develop models and interpret the relationship between various house sale information.

Starting with price, we analyzed all of the continuous variables to determine which were correlated with price. Table 1 contains if there is an association based on the plots and an association based on the r-value and p-value. In summary, price was strongly correlated with square feet, weakly correlated with longitude and had no association with latitude for the continuous variables.

Table 1: Correlation between Price & Variables

	Association based on Plots	Association based on Tests
Square Feet	Price increased almost linearly with square foot	High r-Value (.7388) and ~0 P-Value suggests price is dependent on square foot and they are correlated together
Latitude	No association based on plots	Low r-value (.05) and .104 P-value suggest latitude and price are no dependent and not correlated together
Longitude	Price increased somewhat linearly with longitude	Mild r-value (.28) and ~0 P-value suggest longitude and price are somewhat dependent and correlated together

Next we analyzed all of the variables to see which were correlated with resident type. Table 2 contains if there is an association based on the plots and if there's an association based on the tests either a Kruskal-Willis for continuous variables or a Chi-squared independence test for categorical to determine the P-Value. In summary, square feet, beds and baths had a clear association with the property type whereas latitude, longitude, city and state did not.

Table 3: Association of Variables from Plots & Tests

Variable	Association based on Plots	Association based on Tests
Square Feet	Condo was the smallest, followed by residential with multi-family being the highest square foot, very clear association.	P-Value was between (1.74E-15 & 3.47E-03 which are all very close to 0 so a strong association between property types and square feet.

Variable	Association based on Plots	Association based on Tests
Lattitude	No clear association from the plots.	P-Value was between .04 and .8. There was a very weak association between latitude and property type.
Longitude	No clear association from the plots.	P-Value was between .19 and .75. There is a very weak association between longitude and property type.
Beds	Condo's has fewer bedrooms whereas multi-family and residential had much higher amounts	Chi-squared was: 354.85 P-Value was 1.38E-68 meaning strong association
Baths	Condo's has fewer bathrooms whereas multi-family and residential had much higher amounts	Chi-squared was: 223.7 P-Value was 6.22E-44 meaning strong association
City	Too much data to make a clear association	Chi-squared was: 47.9 P-Value was .96 meaning strong no association
State	N/A - There is only one datapoint (CA)	N/A there is only one data point (CA)

Following the graphing and statistical analysis, a linear regression was created with price as an output and square feet as an input as the initial model. Each individual variable was added excluding date, zip, state and city. While adding in the data, the following condition number were calculated. Based on the errors from the regression model: address, beds, baths latitude, longitude were all removed at being too large. The table below shows the specific condition number when adding them to the model.

Table 3: Condition Number of Inputs

Input	Condition Number
Address	4.54E+28
Type	21.29
Beds	5.56E+15
Baths	2.10E+15
Latitude	11873
Longitude	124372

With the inclusion of type into the model, the calculated root means squared error (RMSE) of the final model versus it predictions was 81363.01 which translate to being an error of $81,363 when estimating the house prices. The MAPE of the new model is 30.10 which translate to an error of 30.10% when estimate the house prices. The model should not be used to accurately predict prices of the house sales, a 30% error is far to significant.

For the final model, a distribution of the residuals were plotted and they do fall into normal distribution and the mean of residuals was calculated to be -1.30e-10 which is approximately 0. When plotting the residuals vs predicted values, there does not appear to be an dependence of the residuals on the predicted values they are all scatter about. This would hold all of the assumptions for residuals to true.

Therefore, the model only included square foot and house type for inputs. The same conclusion was made earlier with the Kruskal testing showing price and type were correlated and p-value with r-value that price and square foot were correlated that the final model used. The three categories of type ("Condo", "Multi-Family", "Residential") are shown with the following coefficients, this shows the average price change from each other while keeping square feet constant. It shows generally with all things equal a residential home is more valuable than condo and multi-family real estates. Square feet shows on average through the data that per 1000 square feet it will provide 134,700 holding the type constant.

Table 4: Coefficient of Inputs

	Coefficient	Units
Condo	2521.557	Price/Type
Multi-Family	-51,180	Price/Type
Residential	18,010	Price/Type
Scaled Square Foot	134,700	Price/1000 Square Foot

The data from Sacramento area was unable to provide an accurate linear regression model. However, it did provide some insights that could be generally used to increase sale price by increase the square feet with the inclusion of addition baths and beds. Additional beds were shown to increase the price the most, but baths could continue to raise the price. Square feet was the only high r-value (.728) data to price which would explain why there is so much inaccuracy.

If the reader would like to continue to create a more accurate model some ideas would be: recent 2-year renovations or maintenance costs to see which properties were recently updated, school district score since school funding is directly related to property taxes, and total lot size as opposed to square foot to help distinguish pricing differences between the lot size and home size.

In conclusion, the linear regression model we developed, which included square footage and property type as inputs, exhibited limited accuracy in predicting house prices in the Sacramento area. While our analysis shed light on the importance of square footage and property type in explaining price variations, the model's relatively high RMSE and MAPE indicate its inadequacy for precise price predictions. For future modeling efforts, we recommend exploring additional factors such as recent renovations, school district scores, and total lot size to enhance the accuracy and reliability of housing price predictions.