Modeling Texas Home Prices

Andrew Pfeifer

DSC 550

Summer

August 8, 2024

Mortgage companies on the occasion have customer's disputing the property value that was provided by the appraiser for their home. The mortgage company would then need to complete a review of the appraisal and do research that can take a few weeks to determine the validity of the customer's complaint. This is time that can be used in assisting others to close on their new homes or complete a refinance or to examine the home buying process to see if it can be made more streamlined. This problem should be solved through modeling to free up this time. Also, if the model is accessible on the mortgage companies' website, the homeowner could have an idea of their property value prior to an appraiser coming out.

This is precisely how I would pitch this project to stakeholders as well. We can save the company time in labor from preventing unnecessary research into property values. Also, if there is a repeat offender appraiser that is consistently undervaluing homes, that appraiser can be cut off from future transactions, saving the company money. This project is even more beneficial to mortgage companies that are tied to a real-estate business. If this were the case the realtor can use the model to track home values when a customer is selling their property to give them a better idea of what it should sell for. That said, if homeowners and potential buyers/sellers can estimate the home's value before the transaction, this can create a more determined customer. The ones that see that their home value is not as high as they thought, may end up not applying for a refinance. This may give them the incentive to pay down the home more prior to committing to a refinance or a sale.

The data set for this project can be found on the following link (https://www.kaggle.com/datasets/kanchana1990/texas-real-estate-trends-2024-500-listings/code). The data contains a mix of values and categories that can be used for the analysis. However, there are a couple of columns that may not provide insight to the model. Once the model is complete the mortgage company will be able to input characteristics of customer's homes into the model to predict what the value of the home should be. This will aid in the mortgage companies' decision to either ignore what value the appraiser has provided for the home as a work around or support what they are providing as the value. Using this to help keep records of home values will also allow the mortgage company to track the performance of the appraisers that they are employing. The main characteristics of a home will be the primary focus of the project.

While exploring the data with histograms in the images below, most homes in this data set a new or left skewed (figure 1). Since most of the homes are new, this may end up being a feature that we cannot rely on when predicting home values. The Home's age may cause some issues further in the data as it may not provide insight into the listing price since most homes are new.

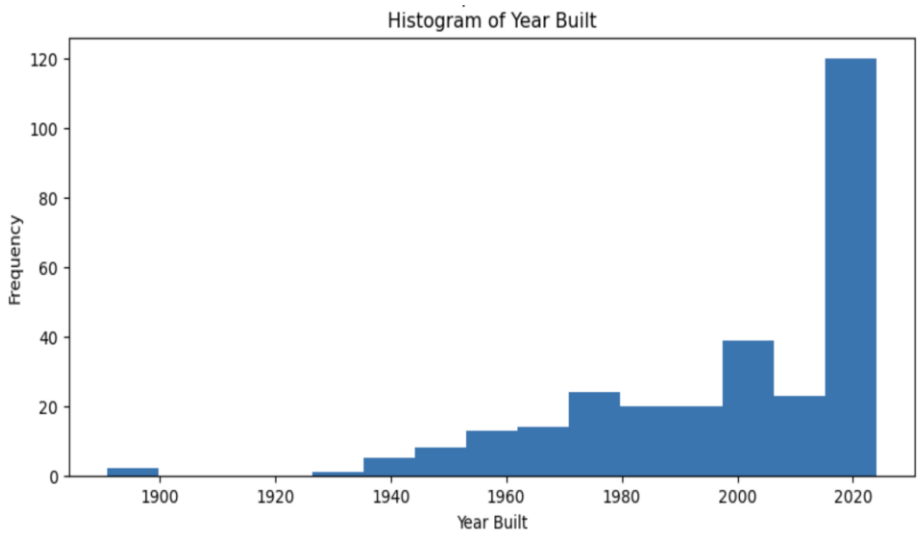

Figure 1: Histogram of Year Built

The square footage and the listing price of the home are both right skewed in figure 2. However, for the listing price, this is only due to outliers in the data. Ignoring these points one can see an almost normal distribution curve. If all the outliers are removed from the square footage image, we would still have a right skewed image so another approach would be needed to normalize this information. This could be due to the types of homes in the data set.

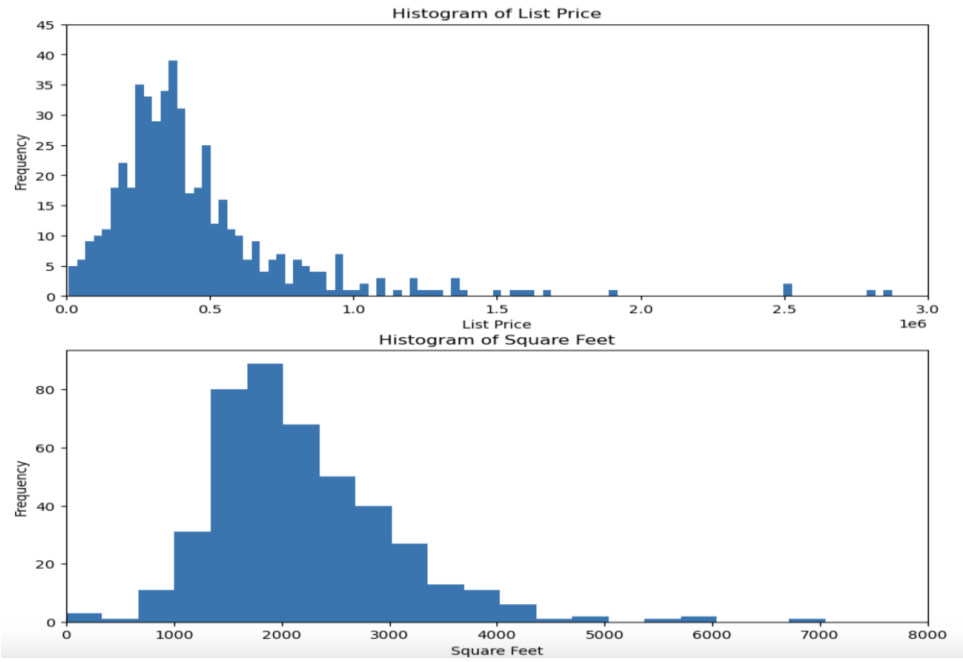

Figure 2: Histogram of List Price and square footage.

Before the data is manipulated, it is important to investigate a few key features of the data. The first feature that was chosen was the home type. The primary home type in this data is a single-family home in figure 3. This should be the primary home type for this model.

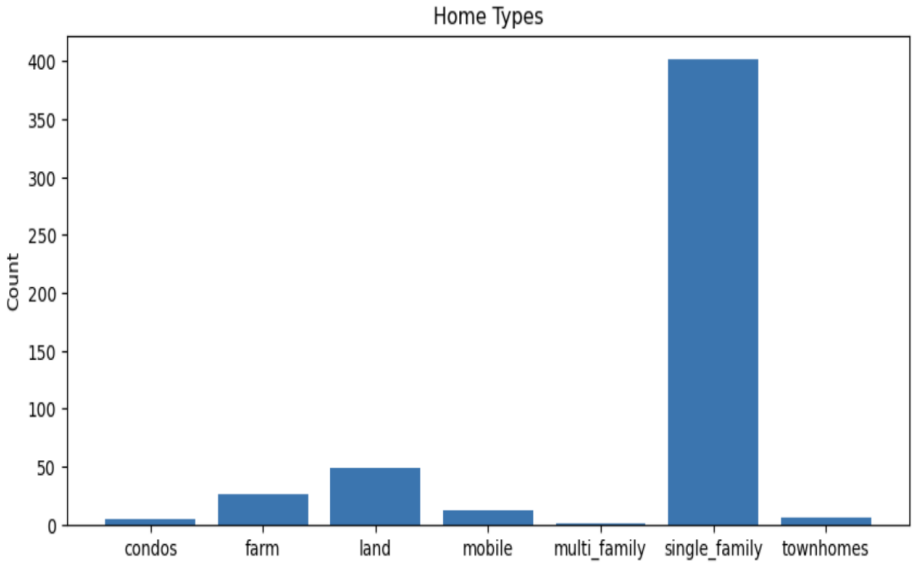

Figure 3: Bar-graphs of home type.

Stories were chosen as well just to have an idea of the different home builds. This may not impact the model, but it was important to pull to check for outliers. Based on figure 4, it appears that anything above 2 stories might be an outlier, but this could be influenced by other property types.

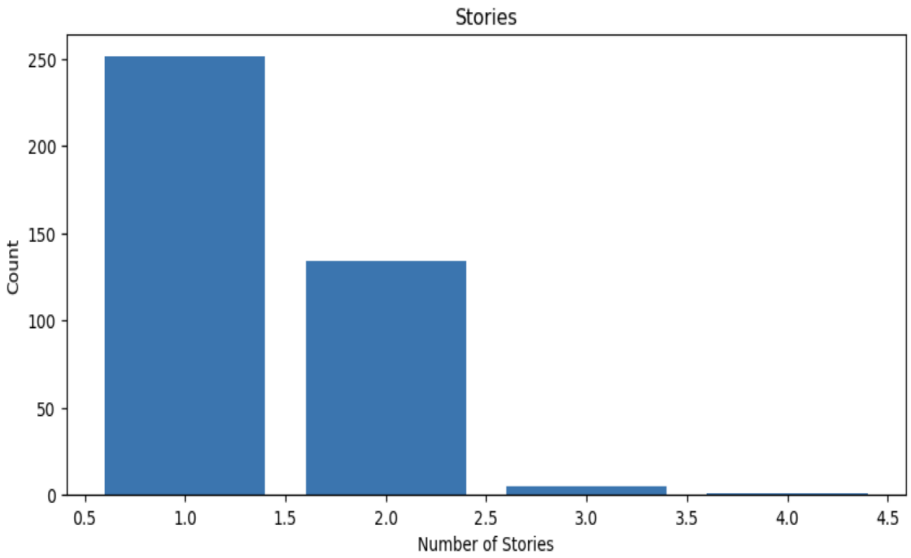

Figure 4: Bar-graphs of stories.

While reviewing the bar graphs in figure 5, the number of beds appears normally distributed but when you consider a modest house a 2 or 3 bedroom is considered normal. Luxury type homes may be influencing the higher total of beds. While considering the output for the total number of baths, there is a value that stands out and that is 0. This would not make sense for a home to have zero baths. However, the data does have lots listed in it. Ignoring the zero, we can clearly see that the majority of these homes have between 2-3 baths. The totals on the upper end of the chart could be from luxury type homes. A separate data set should be considered to exclude luxury homes.

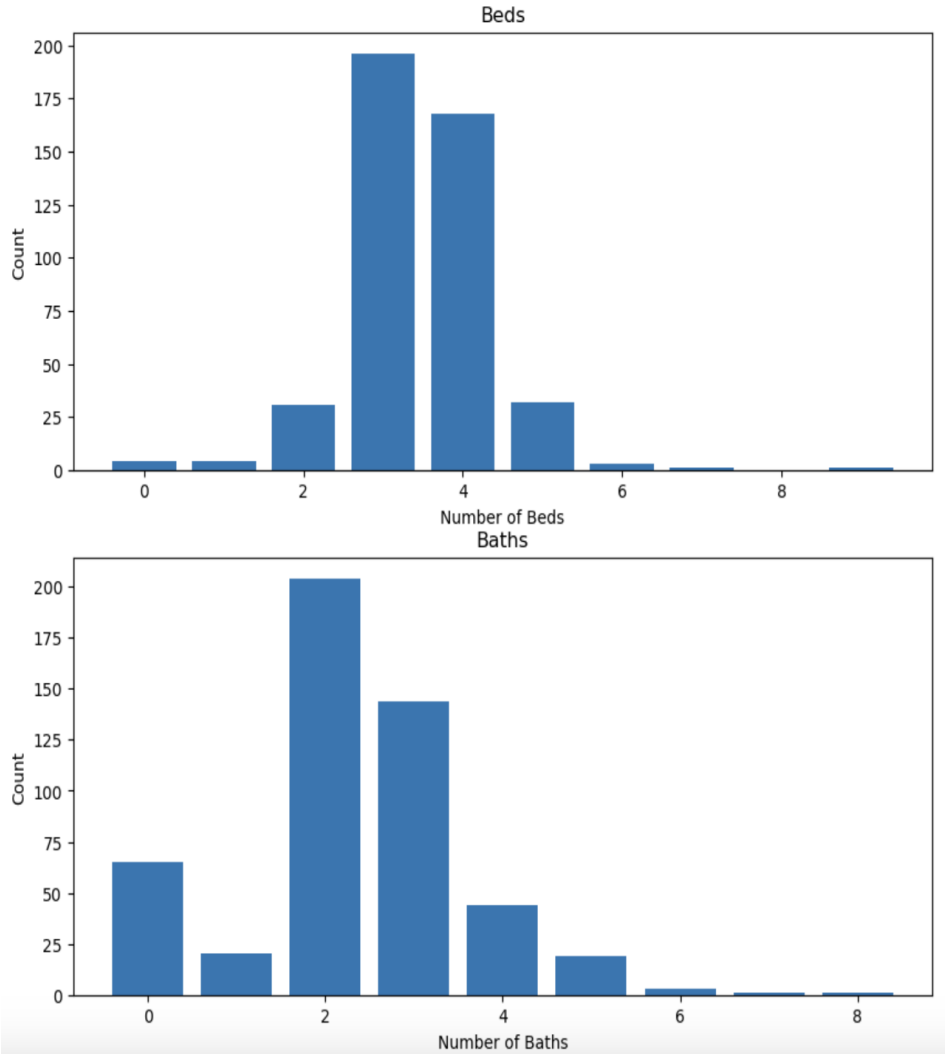

Figure 5: Bar graphs of beds and baths.

The first step in the data cleaning process would be to drop the columns that would provide no value to the model. The columns that were removed in this case were the URL, status, and ID. The ID was only an identifier for the seller, status had only one value and was the same for all homes. Lastly, the URL would not be something that would assist in the model making process. Next, we needed to handle the NA values. I chose a threshold that would eliminate any column the is more than 25% NA values. In this instance there was only one column that was removed, and this was subtype. The next step was to

remove the text column. I did not just eliminate this column, but made a second data frame containing the listing price and the text in case we can use this in the future to determine if the comments can affect a home's price.

Now that the data is split, the data frame with no text was reviewed for the NA values. For the rows that were missing the listing price, these were removed. Since we are modeling for listing price, not having this attribute would be pointless for the data. Next, any float or integer value columns in the data set had the mean of their column replace the NA values.

Now that the data is cleaned, we can now look at the outliers in the data set. At an initial glance of a box plot for the listing value in figure 6, It is shown that there are multiple outliers on the upper end of the of the listing prices. To resolve this issue the z-score of the data was taken and a threshold for the score at 0.5 was made. However, this is the lowest z-score I could use without getting an error in python. Again, in the box plot it is shown that there are still outliers within this range. To combat this, I made another dataset that would contain no luxury homes. For this consideration, a cut off point was made at one million dollars. The last step that needs to be taken is to create dummy variables for any categorical data. For the text data set I removed the NA rows since there was only two in each column and port stemmed the text data.

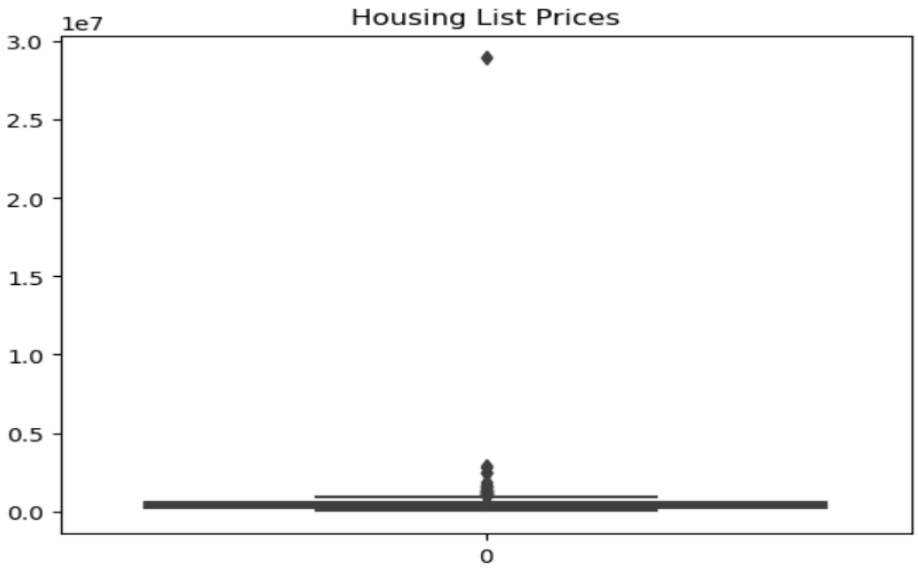

Figure 6: Box plot of listing prices.

Initially when starting the modeling process there were the two data sets for the listing prices. One included all the properties and the other had the luxury homes removed from it. These were both split into testing and training sets for modeling. Since we need to make future predictions with the data, a linear regression model would be the best choice. I made a model for the data that included all the properties first. Since accuracy score cannot be used on linear regression models, I used the R2 and RMSE of the model to gauge accuracy. The R2 was about 0.13 which is too low, we want this closer to 1. The RMSE was 288495.83 which is too high. Since this is the case, the model did not fit the data well.

The next model was for no luxury homes, in which the R2 was 0.347 which is a slight improvement. However, the RMSE was still too high at 150981 which means the model did not fit the data well. Since neither model fitted well, I considered that there may be more outliers in the data. I took the no luxury homes data set and removed all property types that

were not single-family homes. And made a model with this. The R2 ended up being .406 and the RMSE was 145131.39. Again, the model did not fit the data well enough. I decided to check the fit of the model. The fit came in at 20.27% so the model was underfitting the data.

A thought came to mind that we are choosing the wrong model for this data. I then used a model selection process to compare three linear models. The linear regression model, SVM, and linear random forest with using the power transformer scaler on the data. Power transformer was chosen because the data did not seem to have consistent variance. The best model for this data ended up being a random forest. However, the R2 came in at 0.51 and the RMSE was131598.53 which was only a minor improvement. Also, with the best model we still have underfitting since it only maps 51% of the data points.

Currently, the analysis of the model tells me that we are still underfitting. This could be due to additional outliers within the data. Also, I could have filled in the NA values improperly. Since there were categorical columns, the mode should have been used on these. The model is not currently deployable. I believe a better data set should be used to generate a model. We would want to see more features like crime and school statistics. It is even better if we had data based on zip code as this may yield better results. Lower crime and better school districts would seem to have a higher value for homes. I believe if we could split the data by zip code, we could have had better results. The current data set may have been Texas wide which could account for outliers that were not seen as well. Since this was the result, I decided not to move forward with the text data since this could result in underfitting too. However, once a properly maintained set of data is obtained, I

believe we can get better results with the regression model. Moving forward we would just need to make one model that can take in data for different zip codes. One model might be needed but the different housing data sets would need to be stored separately. This may provide additional challenges in storing the data, but his could lead to better results,which would be more important.