**Predicting Sale Price using the given data**

**Data Analysis**

A close up of text on a black background

Description automatically generated**Data types of given data**

The data from the given data set has 27395 rows and 21 columns. From the data types you can clearly see some column values that should have a numerical data type such as ‘SALE\nPRICE’, ‘LAND SQUARE FEET’ and ‘GROSS SQUARE FEET’. Once these have been converted to a numerical data type you can then get a summary of all the data.

A screen shot of a computer keyboard

Description automatically generated**Summary of all data**

This summary is useful as it shows the distribution of data in each column when you make use of the 25%, 50% and 75% rows. You can see that for ‘SALE\nPRICE’ the majority of data points are between 0 and 1.15 million dollars whereas the mean shows a value of 1.85 million, this is because this is taking into account very large outliers in ‘SALE\nPRICE’. ‘LAND SQUARE FEET’ and ‘GROSS SQUARE FEET’ both have many 0s within the data, these are both very critical columns that affect the sale price of a house in any given area. Currently this data is very skewed with false values as you cannot have a house on 0 square feet of land. This suggests the data is incomplete, therefore making all values in these rows very difficult to analyse.

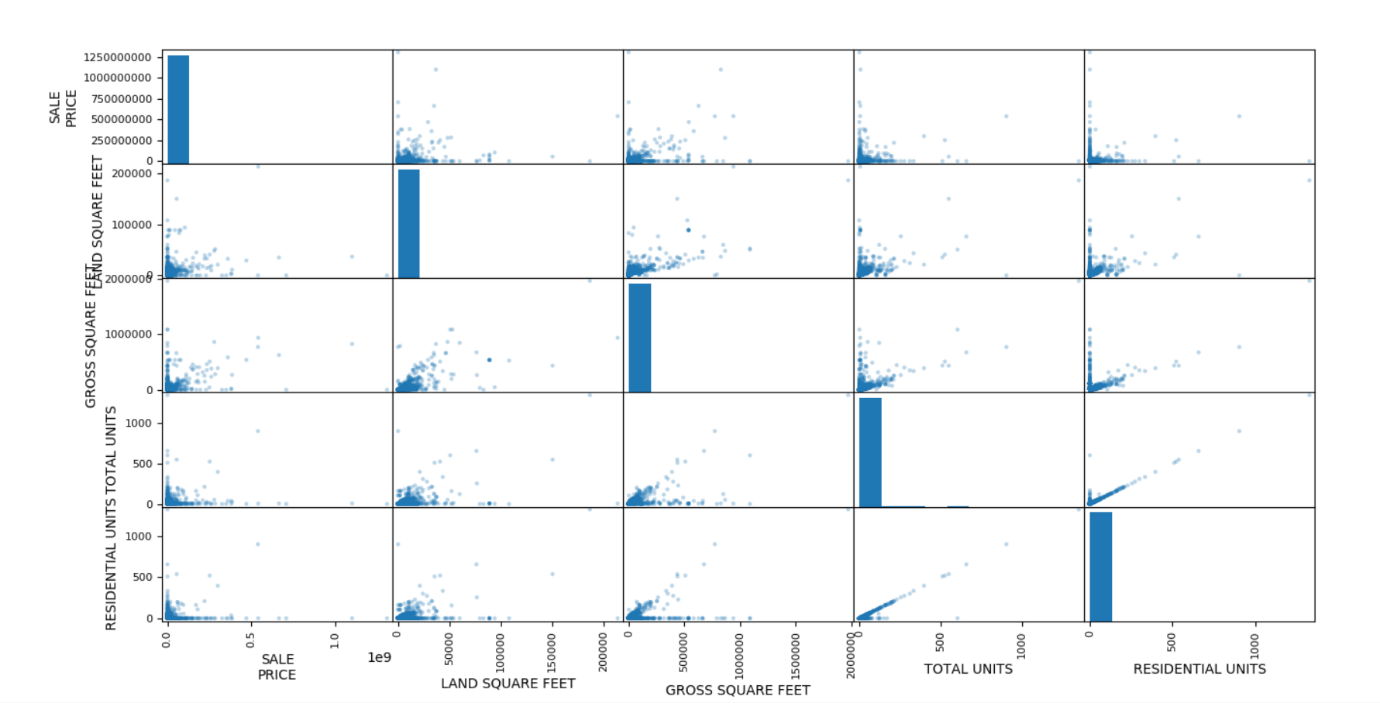
A picture containing text

Description automatically generatedA close up of text on a black background

Description automatically generated**Null values within data 0s within data**

As you can see in the table above and reinforced with count in the summary, the ‘EASE-MENT’ column is empty as it has a count of 0 and shows 27395 null values this makes this column irrelevant to my data set. Another column that has a lot of null values is ‘APART\nMENT\nNUMBER’, this again will make little impact to predicting sale price from this data set unless I want to predict apartments from the same address.

From the ‘0s within data’ table you can now see the exact number of houses being sold that have 0 as their square feet values. This shows that there is still around 10,000 rows of data with all information within that table filled in. I believe these are the most valuable columns within the data that will help to predict housing prices.

**Scatter matrix showing different columns against SALE PRICE**

From the scatter matrix you can see a slight positive correlation to sale price against ‘GROSS SQUARE FEET’ and ‘LAND SQUARE FEET’ therefore these are the more valuable data columns in the data set and will later be used in my model.

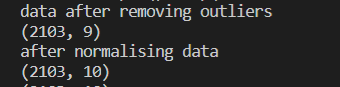
**Scatter graph of Sale price against Sale date**

A screenshot of a cell phone

Description automatically generated

The scatter matrix and this graph of sale date against sale price also show a large grouping of data points in around 0 and so to use this data effectively we must carry out data munging and cleansing. You are also able to see data points far above the rest and although this is possible within sale prices of housing, these data points change any predictions I might make using this data.

**Data Munging and Cleansing**

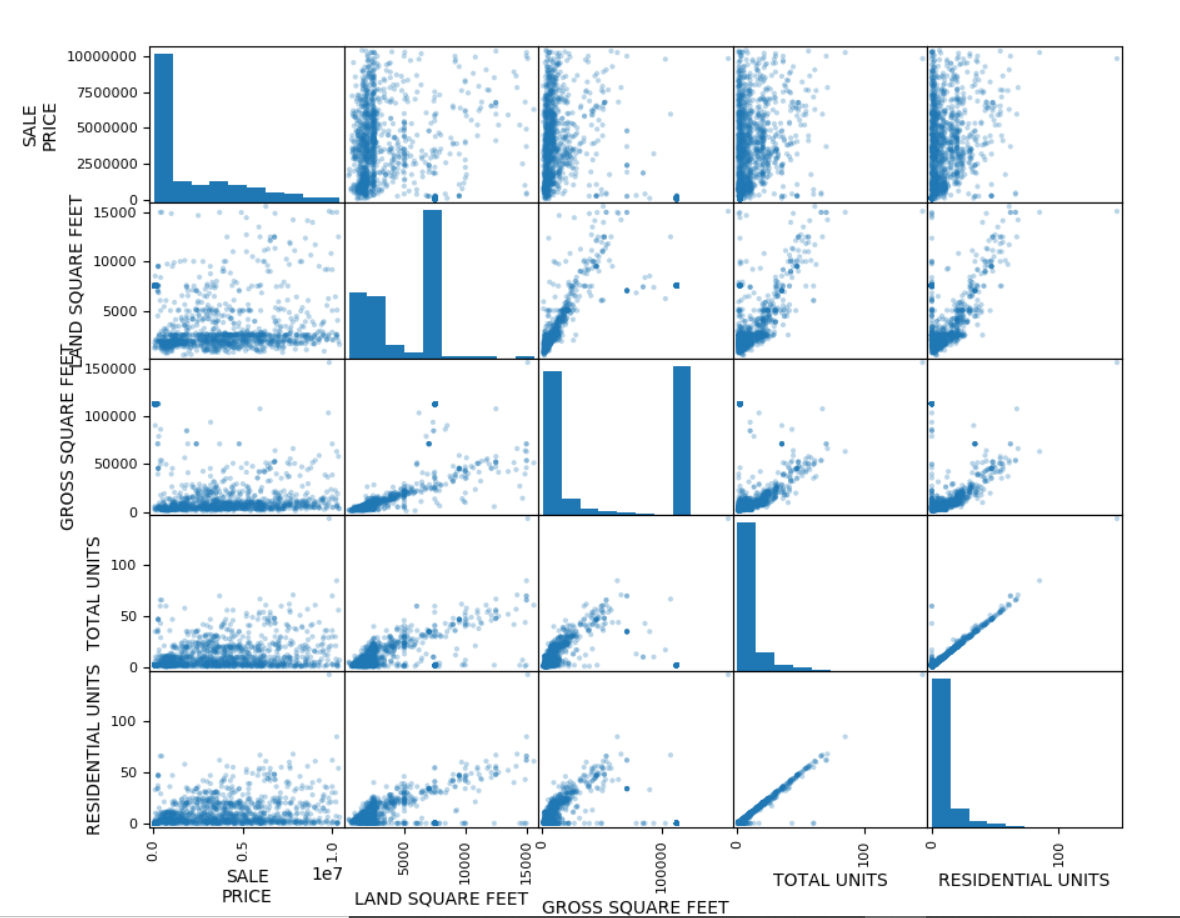
A screenshot of a social media post with text and a black background

Description automatically generated

As you can see from the above screenshots, I have printed the shape of the data at each stage I did in the data munging and cleansing process. Firstly, I removed any none numerical columns in the data set, this is because they cannot easily be used in graphs or models, only to group data. This left 9 columns and 27395 rows of data. I then went on to remove all the rows that had 0s in sale price as this is what I want to predict with my model making the data in these rows unusable. This left 19802 rows of data. After doing this I used the method of forward filling empty cells in both ‘LAND SQUARE FEET’ and ‘GROSS SQUARE FEET’ as without these being filled the majority of the data is empty and all housing being sold have to have an area greater than 0 square feet. After removing all 0s in both columns the row size becomes 2759, this is still enough to predict housing prices although this does still contain outliers.

In the image above I have printed out the quartiles of the data set for ‘SALE\nPRICE’. This data has a very large interquartile range (IQR) and so when I used the method of 1.5\*IQR below quartile 1 and 1.5\*IQR above quartile 3 I get a negative lower limit for outliers which isn’t possible with housing sale price as well as an upper limit of 12.8 million which can be the cost of some houses just very expensive houses. Because of these limits when removing outliers, I made the decision that a lower limit of $50,000 is sufficient and will increase accuracy in my model once outliers were removed.

**After outliers Data visualisation**



After removing outliers this is the scatter matrix of the graphs within my data. They now show a better correlation as there are no outliers so we have a more distributed range of values easily visible within the graph. This is better as we can now make better predictions using values within this new range of accurate data. It is only a slight positive correlation in all of the graphs though as by themselves they do not cause a big enough change in the sale price to show a strong correlation, meaning other factors can still come in to play to change the value.

**Model Evaluation**

The model I have made is a multiple linear regression model. This is a good fit for my data as I am mainly focussing around numerical data that has a linear positive correlation with many different factors that all contribute to the prediction.

My model formula is: logPrice = 0.72 + 0.25landSquareFeet + 14.05residentialUnits + 5.93commercialUnits -13.26Total units -1.09grossSqaureFeet

My model uses the columns in the dataset: ‘LAND SQUARE FEET’, ‘RESIDENTIAL UNITS’, ‘COMMERCIAL UNITS’, ‘GROSS SQUARE FEET’ and ‘TOTAL UNITS’. This is because these were the largest factors when trying to predict housing prices. The square footage plays a big part as the majority of the time the area of housing becomes larger the sale price gets higher (landSquareFeet weight coefficient is 0.25), this can change though as it varies between residential units and commercial units. Residential units plays the biggest part in my model as the weight coefficient is 14.05, this shows that as you are selling more residential housing at once the more it will be worth, which makes sense. This is counteracted by Total Units however as this takes into account the number of residential units but also commercial and other types of housing unit that all change the prediction of the housing.

**Predicted price Vs Actual Price**

A close up of a map

Description automatically generatedMy model has an R squared value of 0.87 which means it is 87% accurate. The model can predict 87% of the variation in the output by using the variation in the model. This shows a strong positive correlation as the use of the model has reduced the variability in predicting the output by 87%. Only 13% of the sample cannot be explained by my model or factors used by my model.