Analysis of Real Estate Pricing for Dallas

Business Understanding: Predicting sale prices for real estate properties is conditional upon various factors. Realtor agents have played a key role when determining the value of a house by looking at a plethora of factors or variables. Luckily, Big Data has helped to advance the real estate industry but providing analytical breakdowns of these specific variables and helping to create relationships. Even though there are numerous variables to consider, the most important include: location, age of house, type of property, home size, square footage, and market conditions. However, each of these variables contains other variables within. For instance, when looking at location, one will look at the community, proximity to local markets, parks, and entertainment, and school districts. Market conditions are dependent upon interest rates, job market, and interest rates when determining price. Furthermore, home maintenance and features such as tile vs. wood, stainless steel appliances vs. updated appliances, and new HVAC vs. outdated HVAC pertains to the home's value, these variables cannot predict the listing price as they are more determine as a buyer's preference.

The 3V's: In order to create a model to predict pricing of real estate properties, large volumes of data pertaining to price, location, home size, square footage, and property type could be collected. A higher rate of velocity variables to include days on the market and pricing could be collected near real-time. Finally, variety can help predict purchasing decisions and impact of predictive models based upon environmental factors such as the school district, interest rates, property taxes, and job market.

Data Understanding: After examining the data, some variables can be excluded from the model before an in-depth analysis is started. The following variables can be excluded as they have identical values for the data entries and would not impact the model: STATE, STATUS, FAVORITE, and INTERESTED. The variable SALES TYPE, SOLD DATE, NEXT OPEN HOUSE START/END TIME, URL, SOURCE, and MLS# can be excluded as these variables do not help when predicting the price of a home. Next, $/SQUARE FEET may be eliminated as this is related to PRICE. LOCATION, ZIP CODE, LATITUDE, LONGTITUDE provide detailed information pertaining to the location of the property which in result helps determine the overall value. Even though LATITUDE & LONGITUDE can help elaborate on location, these variables should be excluded as they are too specific.

Analysis: After further analysis several variables need to be reclassified, examined for missing values and outliers.

Property Type: CONDO/CO-OP, SINGLE FAMILY HOME, and TOWNHOUSES should be included as each type is comparable in size. VACANT LOT, MOBILE HOMES, and MULTI-FAMILY should be excluded as there are fewer cases and BED, BATH, and SQUARE FEET is not given.

City: 23 cities are included in the given data set, but over 87% are in Dallas. The data set is only to include Dallas, thus excluding the remaining cities as they account for only 13%.

ZIP: This was imported as continuous but changed to nominal.

BEDS: Imported as continuous and left as is. There are a few outliers with no missing values (Figure 2). The distribution is skewed to the right.

BATHS: Imported as continuous and left as is. There are few missing outliers and values (Figure 3). The distribution is skewed to the right same as BEDS.

LOCATION: This is a nominal data set with no missing values or outliers. This is due to the location being set to Dallas only.

SQUARE FEET: This is a continuous data set with several outliers, but no missing values. The distribution is skewed to the right and is an important predictor when determining the value of a property. (Figure 4).

LOT SIZE: A continuous value with over 100 missing values and a couple outliers. Thus, this may be excluded.

YEAR BUILT: This was imported as continuous and left as is. A property's value can be influenced by this variable as style changes over time. Only 29 missing values and zero outliers which means this does not have to be excluded.

DAYS ON MARKET: A continuous data set with various missing values not greater than 2%.

$/SQUARE FEET: A continuous data set with only 6 outliers and no missing values. This may still be excluded as this is a derivative of price.

HOA/MONTH: Imported as continuous and left as is. There are a handful of outliers with 57% of the data missing causing this to be excluded from the model.

PRICE: This is a continuous data set with no missing values and less than 30 outliers. The distribution is skewed to the right (Figure 4). This is an important variable to keep as this is the end value needed to create a listing.

Imputation: Used the Automated Data Imputation for missing the missing values listed in the table below. (Table 1).

Conclusion: The main variables to be included consist of PRICE, ZIP, LOCATION, BEDS, BATHS, SQUARE FEET. The outlier and missing value detection in JMP helped determine if a variable should be excluded or not. LOT SIZE, HOA, $/SQUARE FEET were among the various variables to be excluded as these variables would not be beneficial when determining the price of property.

Table 1: Summary of Variables

Variable	# of Outliers	# of Missing Values	Correct Data Modeling Type	Included?
Property Type	N/A	0	Nominal	Yes
City	N/A	0	Nominal	Yes
Zip	N/A	0	Nominal	Yes
Beds	2	0	Continuous	Yes
Bath	1	5	Continuous	Yes
Location	0	0	Nominal	Yes
Lot Size	2	107	Continuous	No
Square Feet	4	0	Continuous	Yes
Day on the Market	5	29	Continuous	No
$/Square Feet	6	0	Continuous	No
Latitude	0	0	Continuous	No
Longitude	0	0	Continuous	No
Price	25	0	Continuous	Yes

Figure 1: Histogram of BEDS

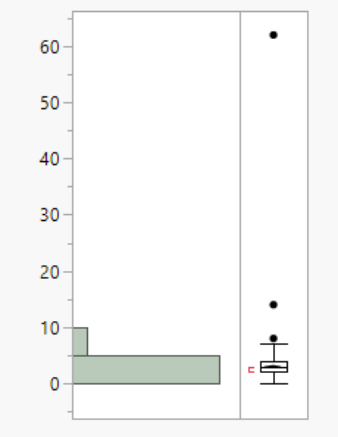

Figure 2: Histogram of BATHS

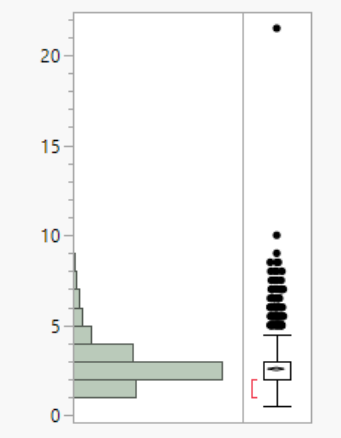

Figure 3: Histogram of Square Feet

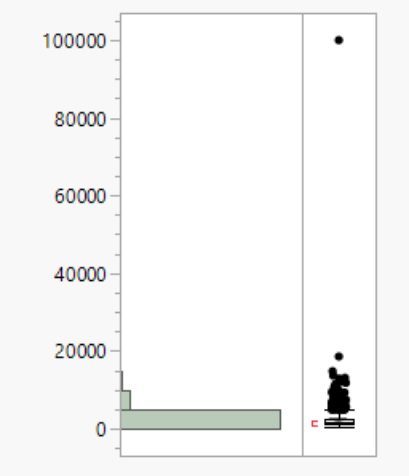

Figure 4: Histogram of Price

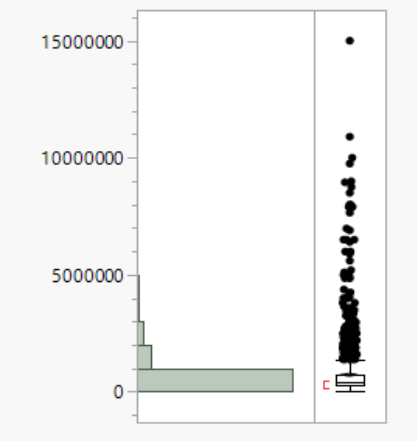

Figure 5: Map of Zip