**Home sale price data analysis**

Annalise Bryant

DATA 1501 W03 Spring 2024

Midterm Project

**Understand the Problem:**

Real estate agents from a real estate company are having issues accurately pricing homes that would lead to their maximum profit as well as their customers.

**Problem Statement:**

With homes exhibiting various factors, we'll need to conduct an analysis by collecting data that guides us in identifying an accurate price point for each one. This information will allow realtors a better chance at maximizing their potential earnings when putting new houses on the market.

**Assess the Data:**



**Diagnose the Data:**



**Clean the Data:**

The first step in cleaning the data, we first started out by correcting any wrong abbreviations or misspelled words within our variables. For example, we change “vG” to “vGd”. After making the previous correction to each qualitative variable, we had to take those now blank categories and replace them with the mode because the mode is what appears the most often in a set.

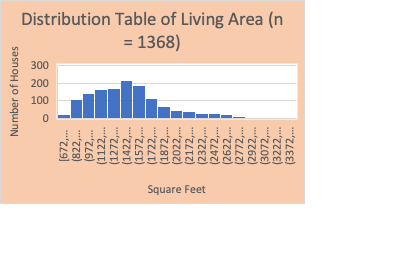
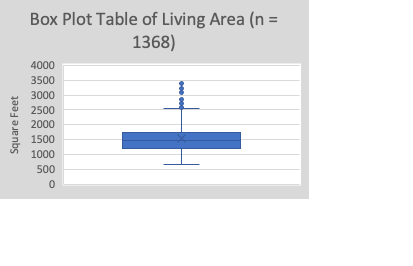
The second step we took was filtering the quantitative variables and choosing “blanks” for each one and imputing them with each variable’s median. We changed the outliers with the median as well.

After preliminary cleaning, we made the table look presentable so that it is concise and clear to the reader. We chose “all boarders”, and we edited the size of each cell based on the amount of word space that was being used. We finished it by changing font sizes, and adding some color to keep it organized.

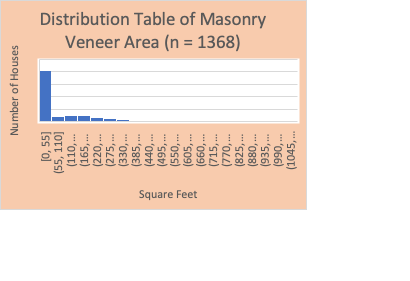
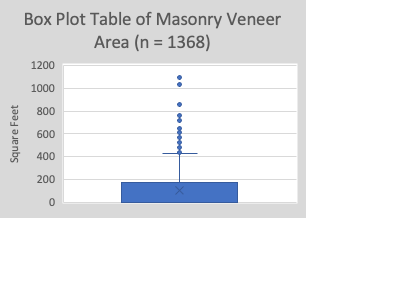
**Univariate Quantitative Analysis:**



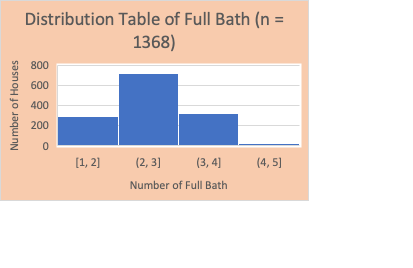
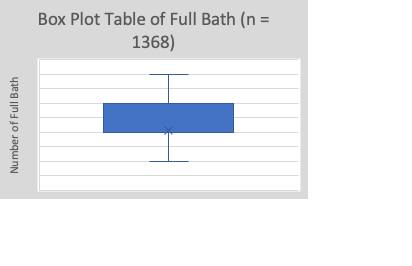
|  |
| --- |
|  |

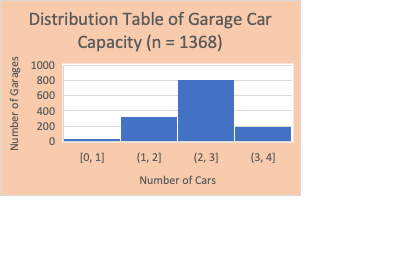
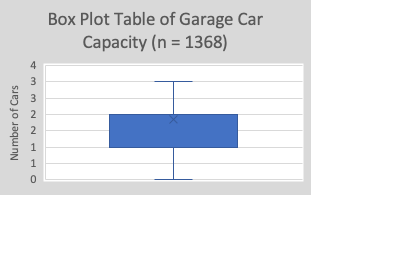
As shown in the histogram, the variable "Square Feet" for Living Area has a unimodal distribution with a skew (to the right). The descriptive statistics confirms the right skew as the mean (1532 square feet) is larger than the median (1480 square feet). The standard deviation provides a measure of spread of 446 square feet. There are 9 outliers seen in the boxplot. These outliers will not be dropped because those numbers are seen in some larger houses for the living area. The first quartile indicates that 25% of houses have less than 1200 sqft of Living Area and the third quartile indicates that 75% of Houses have less than 1743 sqft of Living Area.

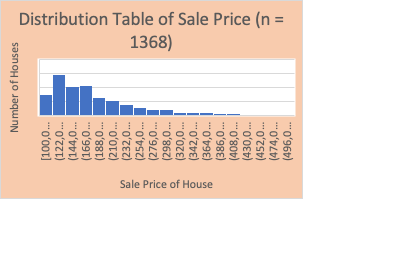
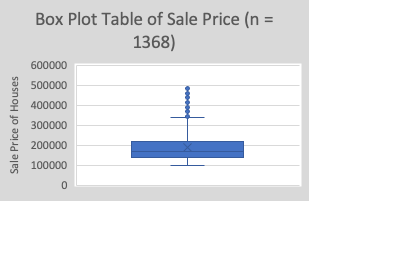
As shown in the histogram, the variable, "Square Feet" for Masonry Veneer Area has a Bimodal distribution with a slight skew (to the right). The descriptive statistics confirms the right skew as the mean (106 square feet) is larger than the median (0 square feet). The standard deviation provides a measure of spread of 172 square feet. There are 18 outliers seen in the boxplot. These outliers will be dropped because those numbers are seen at larger quantities. The first quartile indicates that 25% of Houses do not have a Masonry Veneer Area and the third quartile indicates that 75% of houses have less than 174 sqft of Masonry Veneer Area.

As shown in the histogram, the variable, "Number of Full Bath" for Full Bath has a Unimodal and symmetric distribution. The descriptive statistic also confirms this symmetry as the mean (2 full bath) is equal to the median (2 full bath). The standard deviation provides a measure of spread of (0.7) which if rounded would be 1. There are no outliers. The first quartile indicates that 25% of homes have more than 2 full bath and third quartile indicates that 75% of houses have less than 3 full baths.

As shown in the histogram, the variable, "Number of Cars" for Garage Car Capacity has a Unimodal and symmetric distribution. The descriptive statistic also confirms this symmetry as the mean (2 Cars) is approximately equal to (1.8 cars) rounding to 2. The standard deviation provides a measure of spread of (0.7) which if rounded would be 1. There are no outliers. The first quartile indicates that 25% of garages have more than a 1 car capacity garage and third quartile indicates that 75% of houses have less than a 2 car capacity garage.

As shown in the histogram, the variable, "Sale Price of Homes" for Sale Price has a Unimodal distribution and a skew (to the right). The descriptive statistic also confirms the right skew because the mean ($188,238) is larger than the median ($170,440). The standard deviation provides a a measure of spread of ($70,827). There are 11 outliers in the boxplot. These outliers will be dropped because there are many houses that are in those high price ranges. The first quartile indicates that 25% of homes are less than $137,500 and the third quartile indicates that 75% of homes are more than $219,745.

**Univariate Categorical Analysis:**



Most of the houses fall in the categories from 5 - Typical/Average to 7 - Good, with 27.7% (379) houses in the Typical/Average category, 26.7% (365) houses in the Above Average category, and 22.5% (308) houses in the Good category.

There are 2 underrepresented categories that we will need to address later. 3 - Fair only has 2 houses and 10 - Very Excellent only has 7 houses.



Most of the kitchens fall in the categories Typical Average - Good, with 48.4% (661) kitchens in the Typical Average Category, and 43.2% (591) in the Good category.

There is 1 underrepresented category that we need to address later. Poor has only 1 house.



Most of the foundation types fall in the categories Cinder Block - P Concrete with 41.7% (570) foundation types in the Cinder Block category, and 49.4% (675) foundation types in the P Concrete category.

There is 1 category represented less but still holds a normal count. Slab with 8.9% (122) foundation is a normal amount.

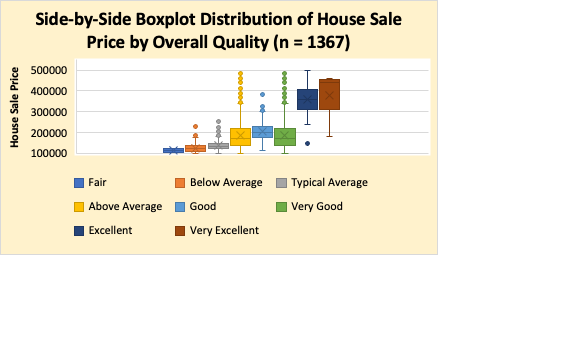


Most of the house styles fall in the categories 1 Story, 2 Story, and Laminated Veneer Lumber with 51.4% (702) in the 1 Story category, 31.4% (429) in the 2 Story category, and 17% (233) in the Split Level category.

There is 1 category represented less but it is normal because the average home does not contain 2.5 Stories.

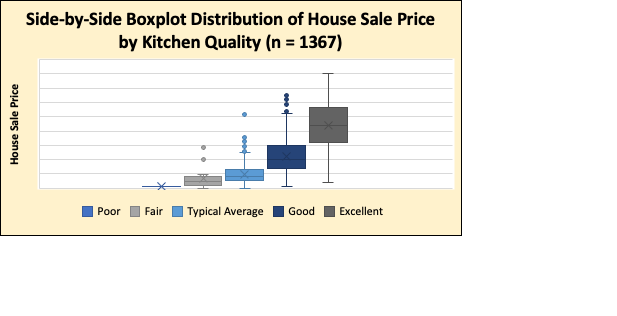
**Bivariate Each Categorical and the Target Analysis:**

****



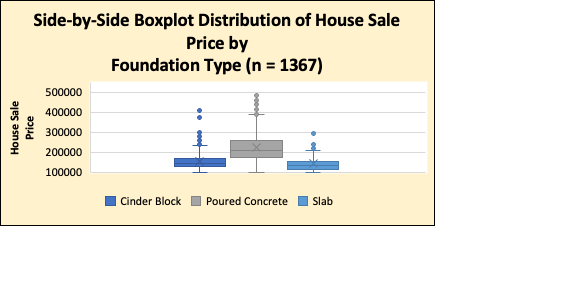
There seems to be visible differences in the distributions of the sale price and the multiple levels of overall quality. There seems to be a discernable trend where if the overall quality is higher, there is a higher sales price. There is, however, a small decrease in sale price for homes ranked with very good to those ranked with good. The average (mean) sale price for the category "good" is $207613 while the average (mean) sale price for the category "Very Good" is $188339. The measure of spread (standard deviation) is confirmed through a right skew.





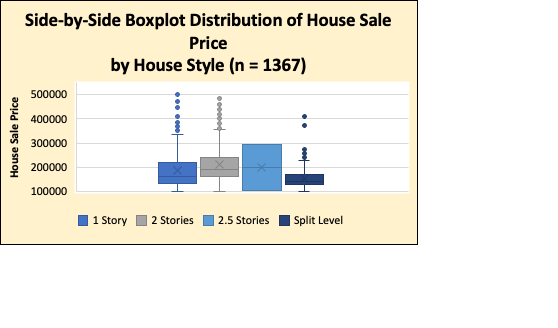
There seems to be visible differences in the distributions of the sale price and the multiple levels of kitchen quality. There seems to be a discernable trend where if the overall quality is higher, there is a higher sales price with a strong right skew. The measure of spread (standard deviation) triples from the categories Fair and Typical Average to Excellent with a difference of $58,192 in sale price. The measures of center (mean) gradually increase but stays in a $30,000 range. Makes sense when the quality increases, the prices increase.





There seems to be a difference in distributions of poured concrete compared to cinder block and slab and a small difference between cinder block and slab. There seems to be a positive skew in the categories cinder block and poured concrete. These could be significant in which we would need to text the differences between categories. The measure of spread (standard deviation) seems to increase $33,036 from the categories Cinder Block to Poured Concrete and then there is a decrease of $38,661 from the categories Poured Concrete to Slab.





There seems to be some obvious differences between each category, however, some of these differences might not have any significance or meaning. There seems to be a positive skew between the categories 1 story and 2 stories which is confirmed by the difference is their means which is $25,498 The measure of spread (standard deviation) a slight increase followed by a decrease. The measure of spread is relatively consistent.

**Bivariate Each Quantitative and the Target Analysis:**



There appears to be a moderate relationship between the living area (sqft) and the sale price. The positive relationship indicates that the more living area (sqft) there is, the higher the sale price will be. The resulting coefficient of determination is 0.4824, which indicates 48.24% of the variation in sale price is explained by the amount of living area (sqft)

There appears to be a weaker relationship between the Masonry Veneer Area (sqft) and the sale price. The positive relationship indicates that the more masonry veneer area (sqft) there is, the higher the sale will be. The resulting coefficient of determination is 0.3329, which indicates 33.29% of the variation in sale price is explained by the amount of masonry veneer area (sqft)

There appears to be a weaker relationship between the full bath per house and the sale price that is discernable. The moderately, positive relationship indicates that the more full bath per house there is, the higher the sale will be. The resulting coefficient of determination is 0.3329, which indicates 33.29% of the variation in sale price is explained by the number of full baths per house on average.

There appears to be a moderate relationship between the garage car capacity and the sale price. The positive relationship indicates that the more garage car capacity there is, the higher the sales price will be. The resulting coefficient of determination is 0.4148, which indicates 41.48% of the variation in sale price is explained by the garage car capacity.

**Bivariate Relationships:**

The variables Living Area and Sale price have the strongest positive relationship with a correlation of 0.695. The variable Garage Car Capacity and Sale price also have a strong positive relationship with a correlation of 0.644. This confirms that when the living area in sqft and garage car capacity increases, so does the sale price which we see in the scatterplot below. From observing the results of the other scatterplots, there seems to be a pattern where our dependent variable (sale price) increases along with each independent variable we analyze. This makes sense because most priced items increase when there are increased features or a better quality.

**Variable (Feature) Creation:**

For the quantitative variables that do not appear to be significant, we will create new categories by editing the values. An example would be instead of “Full Bath” we would change it to “Half Bath”, pertaining to bathrooms without a bath. For “Garage Car Capacity”, we could change it to “Driveway Car capacity” for people without garages.