HOUSING MARKET DATA ANALYSIS

MIDTERM PROJECT

Carlyn Yeatts

DATA 1501 FALL 2023

Understand the Problem:

A real estate agency is having difficulties missing out on potential earnings, determining the price of houses on the market more accurately, and predicting what key features increase the value of a home.

Problem Statement:

We will perform an explanatory analysis to determine what factors lead to increased home value and build a model to predict the sale price based on these factors.

Assess the Data:

Table 1: Data University Data Dictionary				
Variable Name	**Descritption**	**General Type**	**Specific Type**	**Measurment Units**
House ID	Unique housing ID for observation identification	Categorical	Identifier	NA
KitchenQual	The quality of the kitchen: Ex = Excellent, Gd = Good, TA = Typical/Average, Fa = Fair, Po = Poor	Categorical	Ordinal	Ex = Excellent, Gd = Good, TA = Typical/Average, Fa = Fair, Po = Poor
Foundation	The type of foundation: Cinder Block, Poured Concrete, Slab, or Stone	Categorical	Nominal	NA
MasVnrArea	The amount of masonry veneer area	Quanitative	Continuous	Square Feet
GarageCars	The size of garage in car capacity	Quanitative	Discrete	Cars
SalesPrice	The sale price of the property in US Dollars	Quanitative	Continuous	Dollars

Diagnose the Data:

Table 2: Diagnostics Table for Housing Market Dataset				
Variable Names	Missing Values	Distribution Shape	Outliers/Errors	Preliminary Cleaning
House ID	0	NA	NA	NA
KitchenQual	9	Ex (105), Gd (594), TA (757), Fa (27), Po (4), Fair (4)	Error: Fair	Imputted missing values with the mode of TA. Corrected typo from Fair to Fa.
Foundation	5	Cblock (633),P Conc (5), PConc (688), Slab (169)	Error: P Conc	Imputted missing values with mode of Pconc. Corrected typo from P Conc to Pconc.
MasVnrArea	11	Unimodal, right skewed, 99 possible upperbound outliers.	Outliers: 99 data points	Imputted missing values with the median of 0. Imputted values greater than 425 with the max of 425.
GarageCars	7	Unimodal, symmetric, with 4 possible upperbound outliers.	Outliers: 4, 4, 4 Error: 22	Imputted missing values with the median of 2. Corrected error from 22 to 2.
SalesPrice	9	Unimodal, right skewed, 67 possible outliers.	Outliers: 67 data points	Deleted 9 rows of missing values.

Clean the Data:

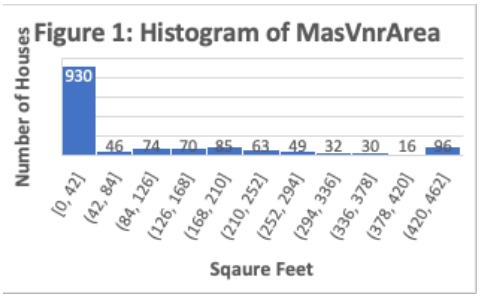
The first task for the data cleaning process will be to correct any misspelling or data entry errors. The variable names will be cleaned up first by making them more descriptive, yet concise.

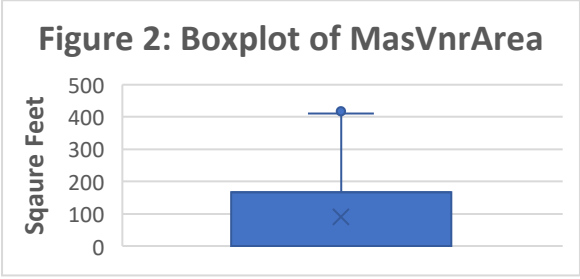
The next task will be to impute any missing values with the median for each quantitative variable and the mode or center for each categorical variable.

The last task will be to impute any outlier values that could be plausible but are highly unlikely based on the overall distribution. Outliers on the upper side of the distribution will be winsorized with a conservative maximum and outliers on the lower side of the distribution will be winsorized with a conservative minimum.

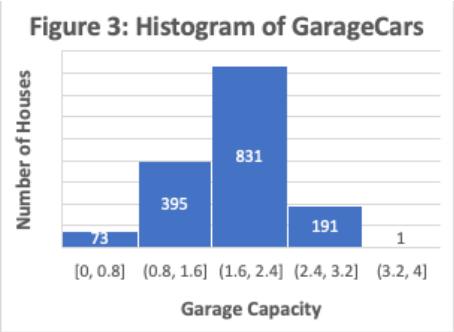
Univariate Quantitative Analysis:

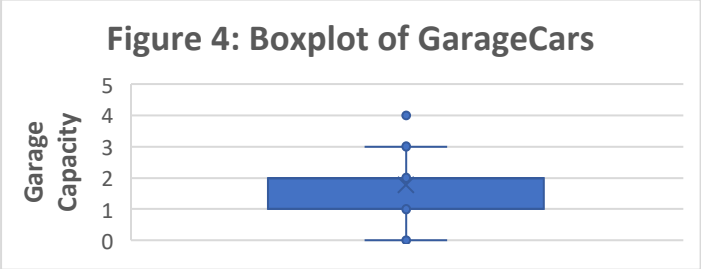
Table 6: Descriptive Statistics for Housing Market Data			
Sample Statistics	**MasVnrArea**	**GarageCars**	**SalesPrice**
Mean	90	2	179,281
Median	0	2	163,000
Mode	0	2	140,000
Standard Deviation	136	1	74,184
Range	425	4	486,900
Q1	0	1	130,000
Q3	166	2	213,192
IQR	166	1	83,192
Minimum	0	0	13,100
Maximum	425	4	500,000
Sample Size	1,491	1,491	1,491



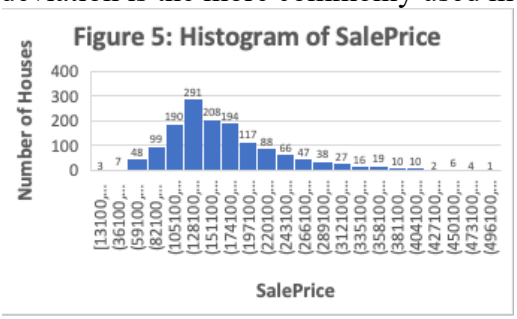


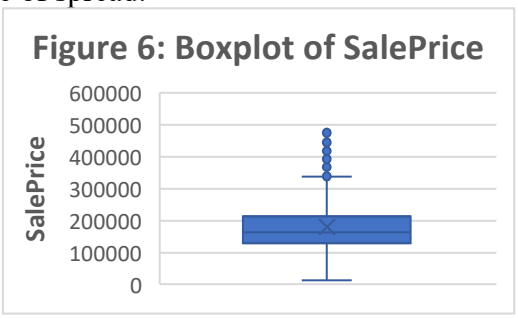
As shown in the histogram above, the variable "MasVnrArea" has a unimodal and right skewed distribution. The descriptive statistics show the mean (90 Square feet) is not equal to the median (0 square feet). The standard deviation provides a measure of spread of 136 sqaure feet. As shown in the boxplot above, the first quartile (a.k.a the 25th percentile) indicates that 25% of the houses have no MasVnrArea, while the third quartile (a.k.a the 75th percentile) indicates that 75% of the houses have under 166 sqaure feet of MasVnrArea. The interquartile range indicates that the middle 50% of houses (half of them) range from 0 to 166. Due to the unimodal nature of the variable, the median is a reliable measure of center and the IQR is the more commonly used measure of spread.





As shown in the histogram above, the variable "GarageCars" has a unimodal and approximately symmetrical distribution. The descriptive statistics show the mean (2 cars) is equal to the median (2 cars). The standard deviation provides a measure of spread of 1 car. The first quartile (a.k.a the 25th percentile) indicates that 25% of the houses have 1 car garages, while the third quartile (a.k.a the 75th percentile) indicates that 75% of the houses have 2 car garages. The interquartile range indicates that the middle 50% of houses (half of them) range from 1 to 2. Due to the unimodal nature of the variable, the median is a reliable measure of center and the standard deviation is the more commonly used measure of spread.



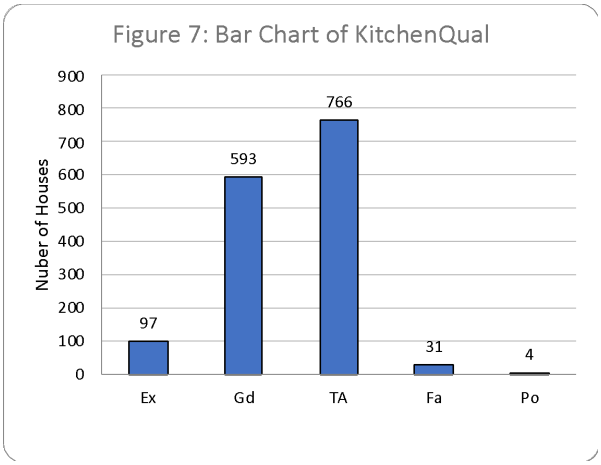


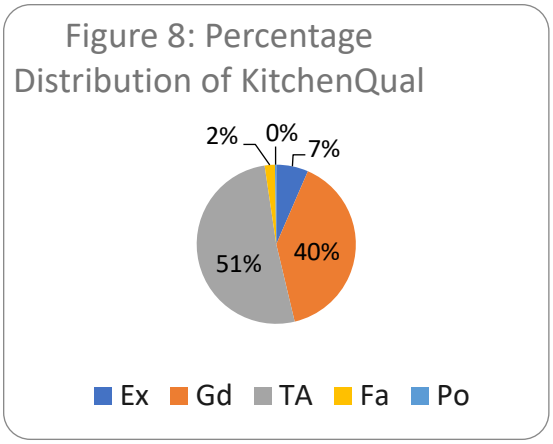
As shown in the histogram above, the variable "SalePrice" has a unimodal and right skewed distribution. The descriptive statistics show the mean ($179,281) is not equal to the median ($163,000). The standard deviation provides a measure of spread of $74,184. The first quartile (a.k.a the 25th percentile) indicates that 25% of the houses cost $130,000, while the third quartile

(a.k.a the 75th percentile) indicates that 75% of the houses cost under $213,192. The interquartile range indicates that the middle 50% of houses (half of them) range from $130,000 to $486,900. Due to the unimodal nature of the variable, the median is a reliable measure of center and the IQR is the more commonly used measure of spread.

Univariate Categorical Analysis:

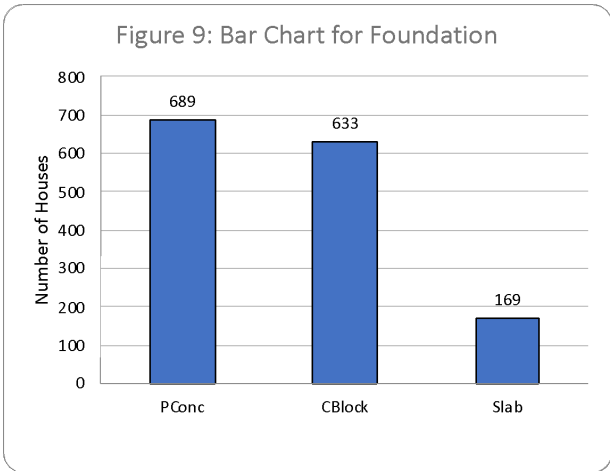
Table 4: Frequency Table for KitchenQual		
Quality	**Frequency**	**Percentage**
Ex	97	7%
Gd	593	40%
TA	766	51%
Fa	31	2%
Po	4	0%
Grand Total	1491	100%

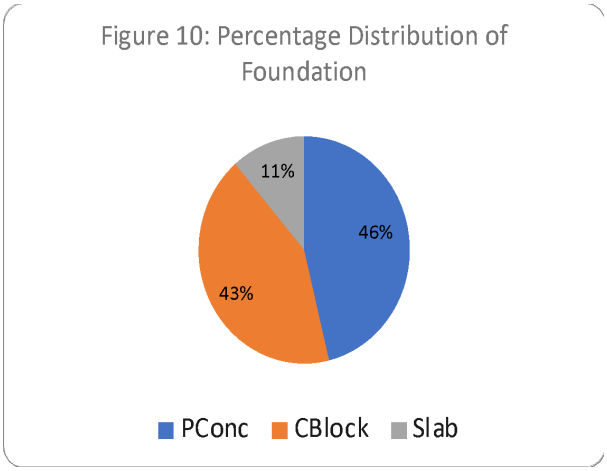




Out of the 1,491 houses sampled, 97 (or 6.5%) are Ex, 593 (or 39.8 %) are Gd, 766 (or 51.4%) are TA, 31 or (2.1%) are Fa, and 4 (or 0%) are Po.

Table 5: Frequency Table for Foundation		
Seat Location	**Frequency**	**Percentage**
PConc	689	46.2%
CBlock	633	42.5%
Slab	169	11.3%
Grand Total	1491	100%



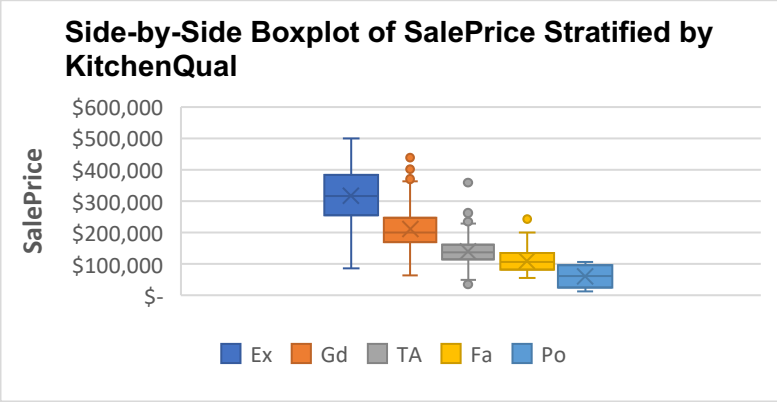


Out of the 1,491 houses sampled, 689 (or 46.2 %) are PConc, 633 (or 42.5 %) are CBlock, and 169 (or 11.3%) are Slab.

Bivariate Each Categorical and the Target Analysis:

Table 6: Descriptive Statistics for SalePrice Stratified by KitchenQual

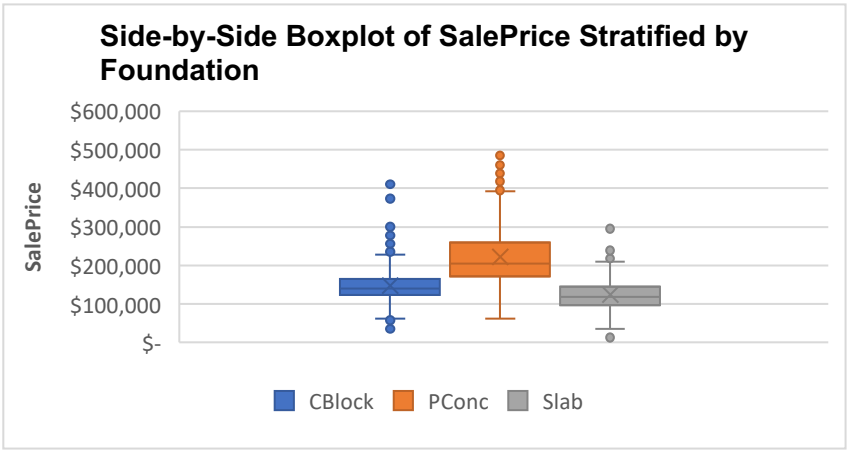
Sample Statistics	MasVnrArea	GarageCars	SalesPrice
Mean	90	2	179,281
Median	0	2	163,000
Mode	0	2	140,000
Standard Deviation	136	1	74,184
Range	425	4	486,900
Q1	0	1	130,000
Q3	166	2	213,192
IQR	166	1	83,192
Minimum	0	0	13,100
Maximum	425	4	500,000
Sample Size	1,491	1,491	1,491



There does appear to be distinct differences between the distributions of SalePrice between Ex, Fa, Gd, Po, and TA, therefore, the differences are significant. When comparing the measures of center, we can see that when the quality of the kitchen increases, so does the SalePrice. There are three standard deviations that are similar Fa, Po, and TA. Ex has the highest standard deviation of $96,639. None of the categories have similar measures of spread.

Table 7: Descriptive Statistics of Saleprice Stratified by Foundation

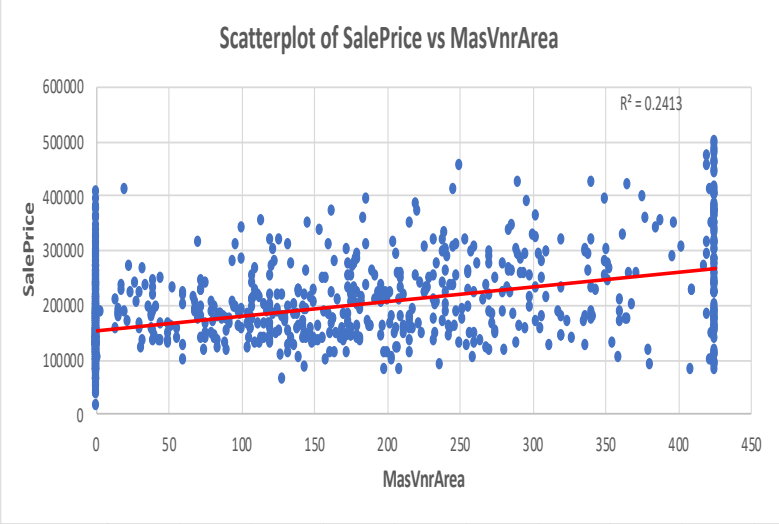
Sample Statistics	*CBlock SalesPrice*	*PConc SalesPrice*	*Slab SalesPrice*
Mean	$ 147,678	$ 221,817	$ 124,236
Median	$ 140,000	$ 205,000	$ 118,400
Mode	$ 140,000	$ 173,000	$ 110,000
Standard Deviation	$ 46,255	$ 77,608	$ 43,165
Range	$ 374,689	$ 438,500	$ 281,900
Q1	$ 123,000	$ 171,750	$ 96,900
Q3	$ 165,500	$ 260,000	$ 144,000
IQR	$ 42,500	$ 88,250	$ 47,100
Minimum	$ 35,311	$ 61,500	$ 13,100
Maximum	$ 410,000	$ 500,000	$ 295,000
Sample Size	633	689	169



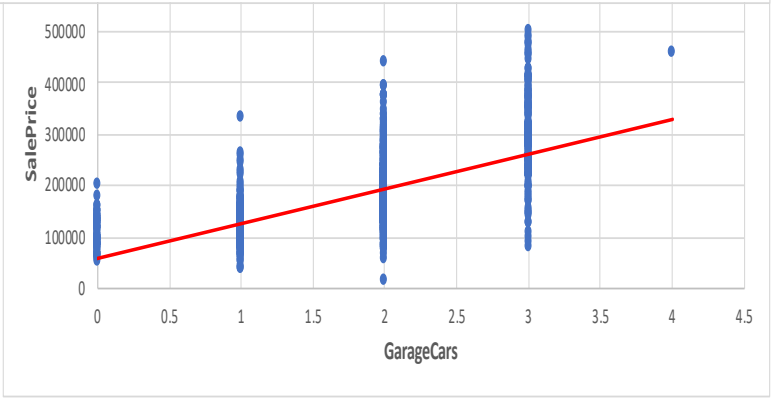
There does appear to be slight differences between the distributions of SalesPrice among CBlock, PConc, and Slab, however it is not clear if the differences are significant or meaningful. When comparing the measures of center, we can see that PConc has a mean of $221,817 which is higher than CBlock and Slab. The average SalePrice for PConc is $221,817, while the average SalePrice for CBlock is $147,678 and the average SalePrice for Slab is $124,236. Two categories (CBlock and Slab) have similar measures of spread, when comparing the standard deviation and interquartile range (CBlock Std. Dev = $46,255 and Slab Std. Dev = $43,165). There does appear to be a couple outliers for females that are creating a right (or positive skew), however skewness is not too drastic.

Bivariate Each Quantitative and the Target Analysis:

Correlation Matrix for Housing Market Dataset			
	MasVnrArea	*GarageCars*	*SalePrice*
MasVnrArea	1		
GarageCars	0.357632758	1	
SalePrice	0.491239663	0.666209598	1



The association between SalePrice and MasVnrArea is linear, positive, and moderately strong (r = 0.49). The R-sqrd value of 0.2413 indicates that 24.1% of the variation in MasVnrArea is explained by the linear relationship with SalePrice.



The association between SalePrice and GarageCar is linear, positive, and strong (r = 0.67). The R-sqrd value of 0.4438 indicates that 44.4% of the variation in GarageCars is explained by the linear relationship with SalePrice.