COMP SCI 7209 Big Data Analysis and Project
Project Title: Big Data: House Price Analysis and Prediction (Part B)
Ning Ni (a1869549)
Project Coordinator: Bernard Evans
1 July 2023

1 Introduction

The object is data visualization, identification of key features, clusters/patterns, and the refinement of questions.

2 Initial analysis of Ames housing

2.1 General description

The housing dataset consists 80 explanatory variables, which can be categorized into four classes:

- Structure and Layout
- Location and Geography
- Material and Quality
- Facilities

2.2 Restatement of questions

Questions in Part A primarily investigate how the structure, layout, and location of a house impact its price, as well as whether it is possible to build a reliable model for predicting house prices.

Since these features are all included in the dataset, it's adequate to address these research questions.

3 Exploratory Data Analysis (EDA)

3.1 Data Cleaning

3.1.1 Removing Nonsense Variable

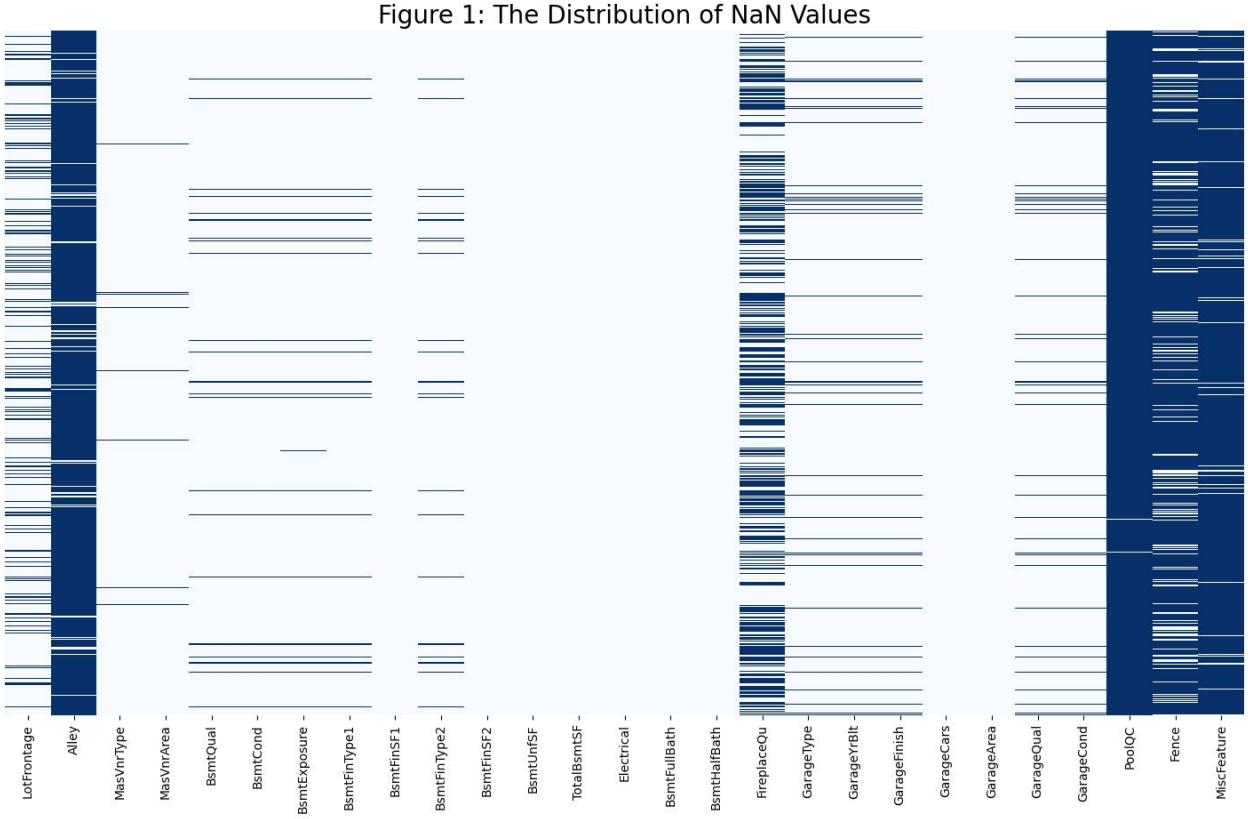
"Order" and "PID" variable represent the only label of house. "Utilities" is the same value at all sample, They can't provide any variability information,and can be safely removed.

3.1.2 Nan Value Processing

Downey et al. (2015) provided methods for imputing missing values.The missing values will be filled with the value "None" or "0". This indicates that there is no specific category

or lack of a numerical value for those instances.

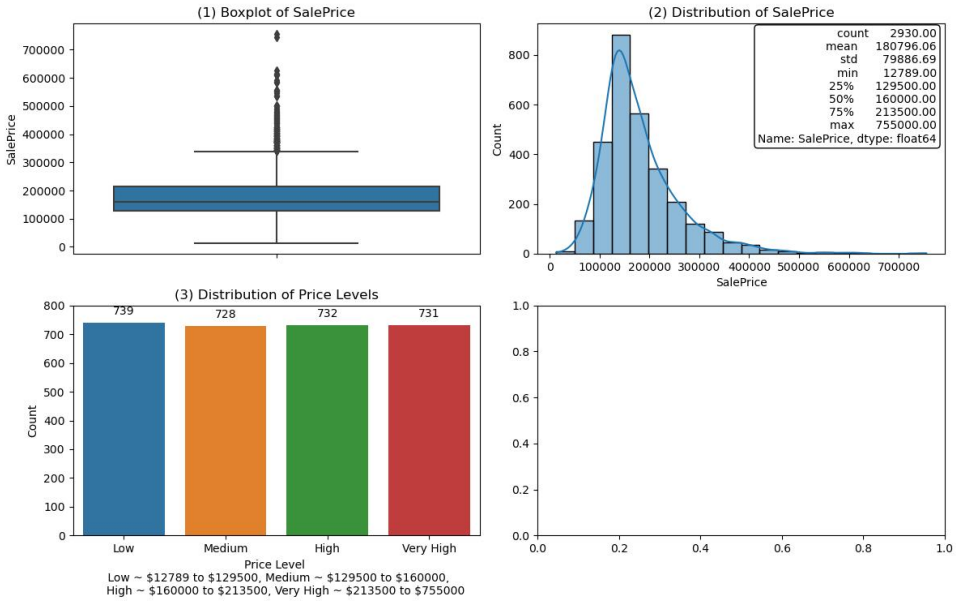
Figure 1: The Distribution of NaN Values



3.2 Target Variable

'SalePrice' is the target variable.

Figure 2: Analysis of Target Variable



'SalePrice' exhibits a slight right skewness, with a mean value of approximately $160,000. Moreover, the price was divided into four levels based on quantiles, which will be utilized for addressing classification questions in further analysis.

3.3 Feature Analysis

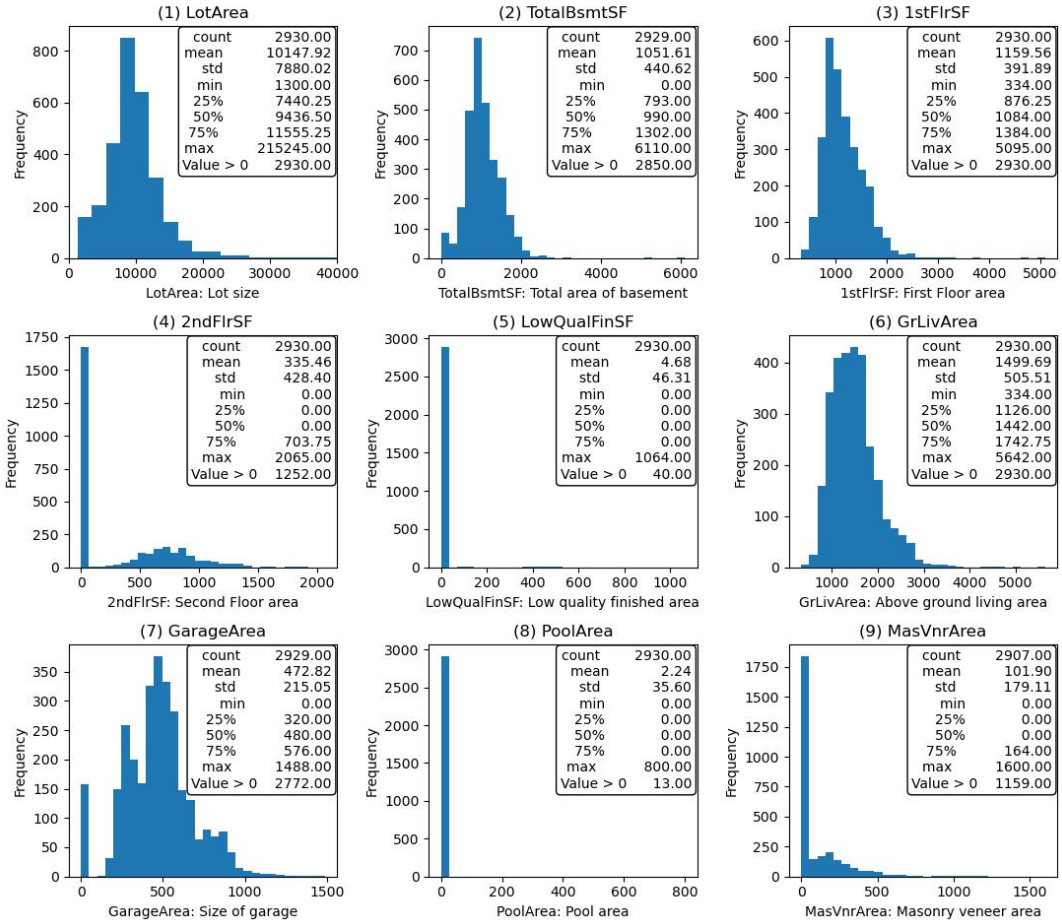
3.3.1 House Structure and Layout

3.3.1.1 Area

Figure 3 illustrates the area of subplot(1,2,3,6,7) display a relatively normal distribution, but subplot(1) exhibits a few instances of extremely large values. Approximately 40% of houses do not have a second floor or masonry veneer, and the majority of houses do not have a pool or a low-quality finished area.

Figure 4 shows as the area of the house increases, the price is showing a gradual increase. Houses with higher overall quality tend to have higher prices. Additionally, houses with larger area seem to be higher quality.

Figure 3: Histogram of Area distribution of Houses. Unit: Square Feet



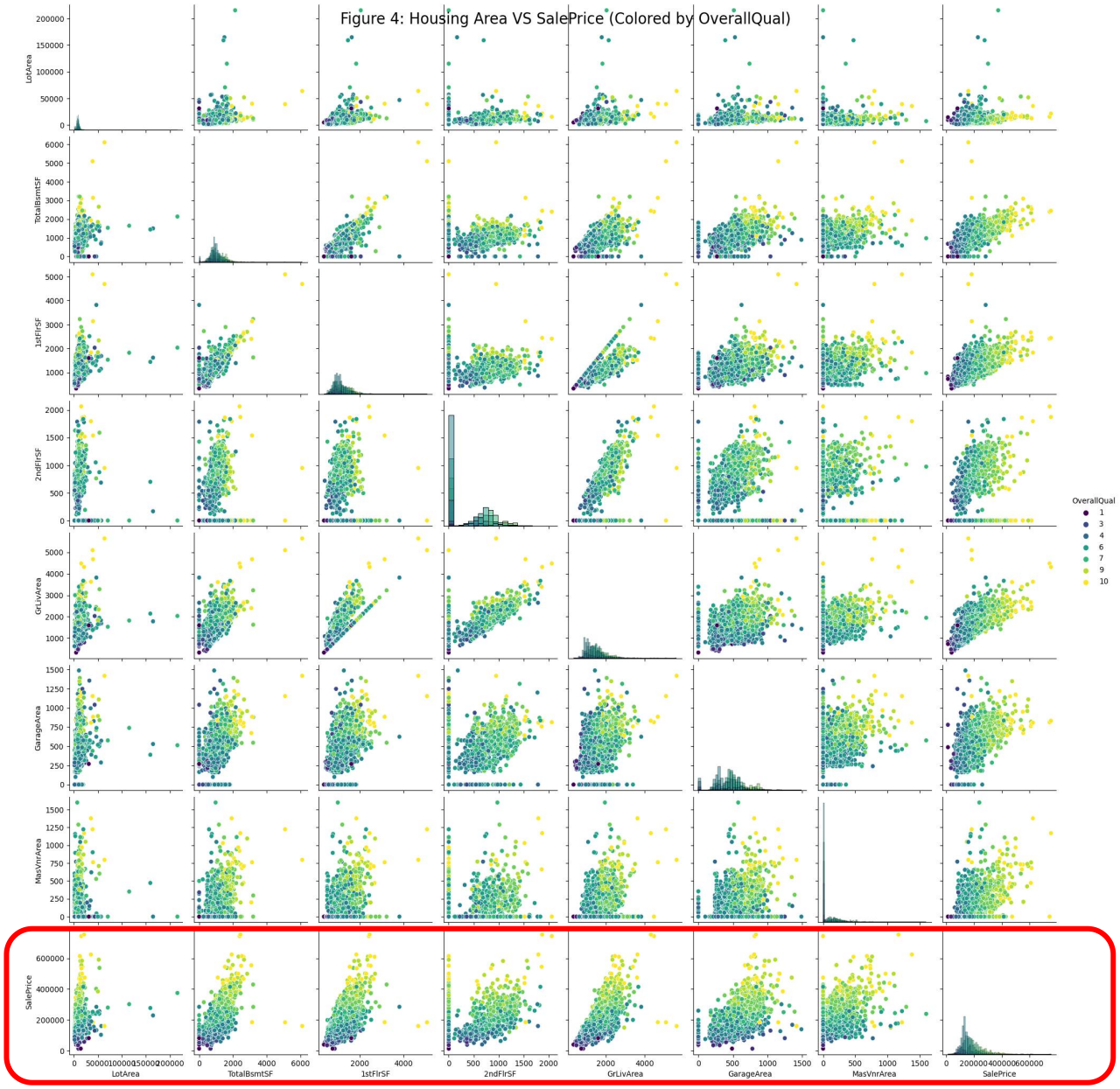


Figure 4: Housing Area VS SalePrice (Colored by OverallQual)

3.3.1.2 Room Distribution

The majority of houses have 2 to 4 bedrooms, 1 kitchen, and parking capacity for 1 to 3 cars. The total number of rooms in the houses ranges from 2 to 15, following a normal distribution pattern.

Furthermore, as the number of rooms increases, prices also tend to be higher. For houses with an equal number of rooms, those with better quality tend to have higher prices.

Figure 5: Barchart of Room Distribution

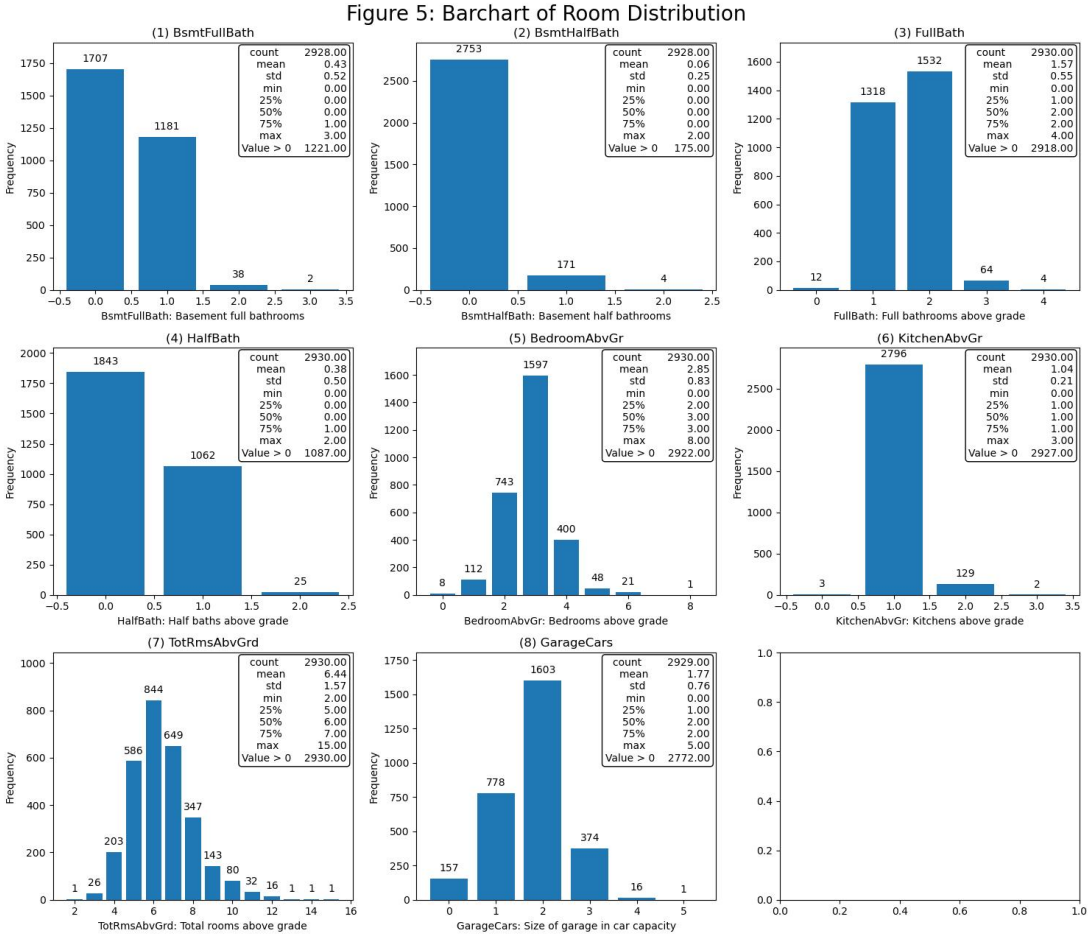


Figure 6: Boxplot of Room distribution

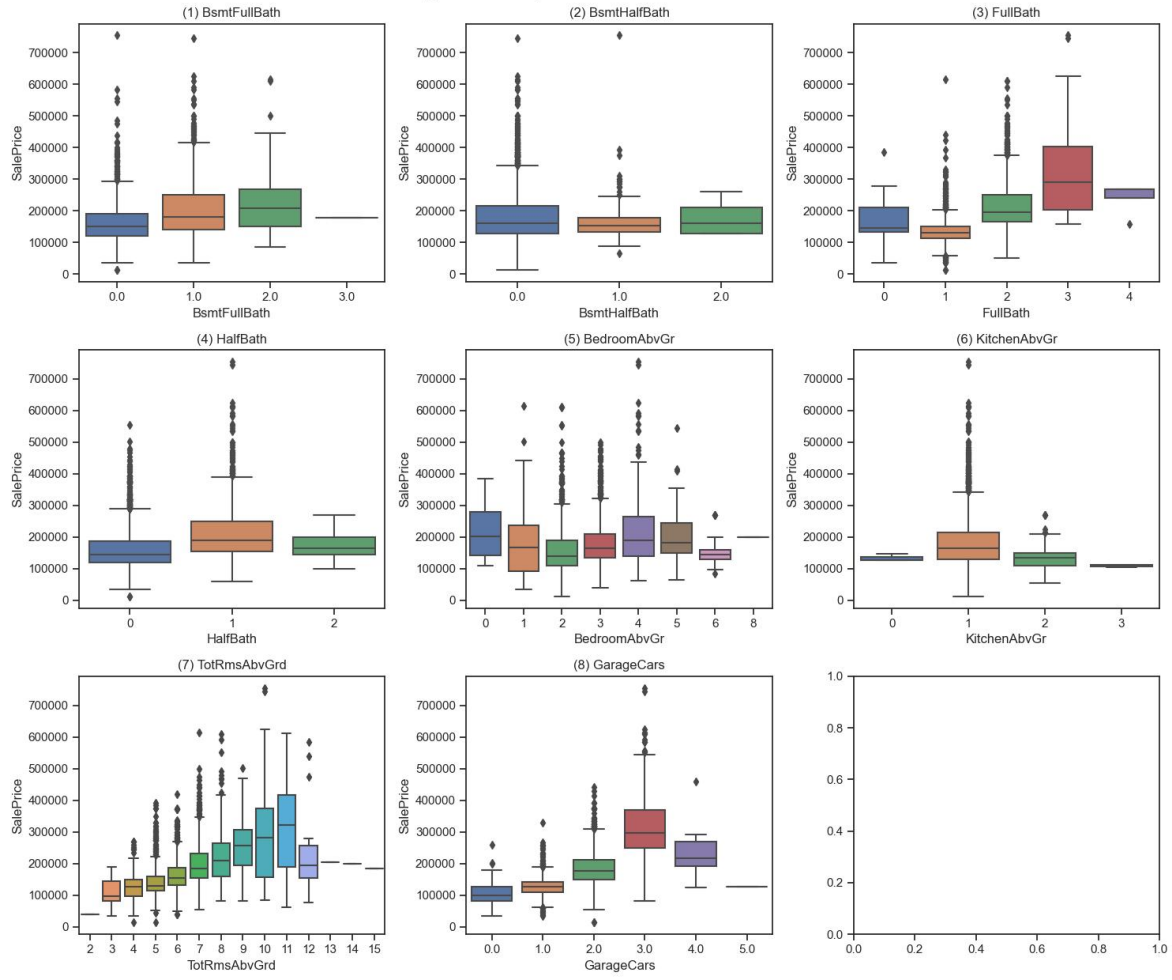
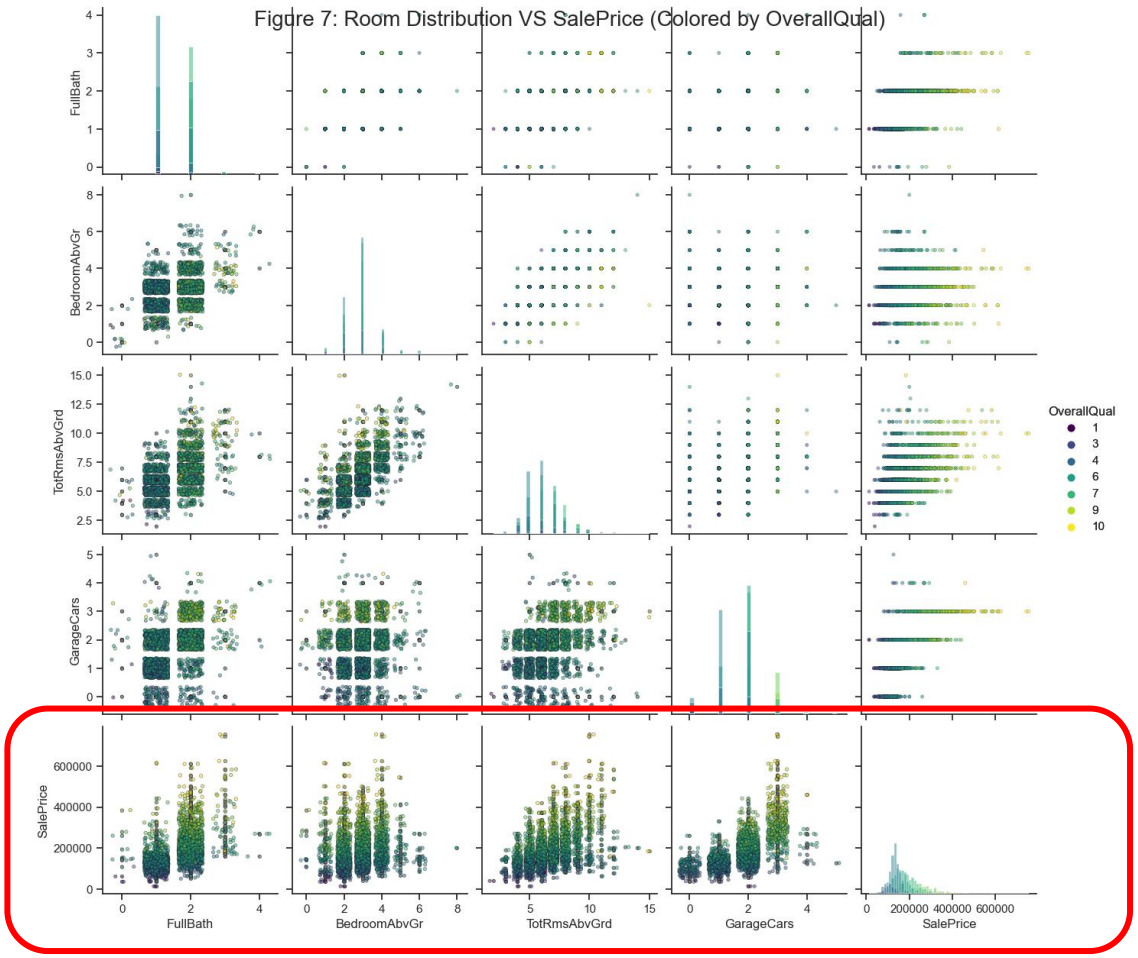


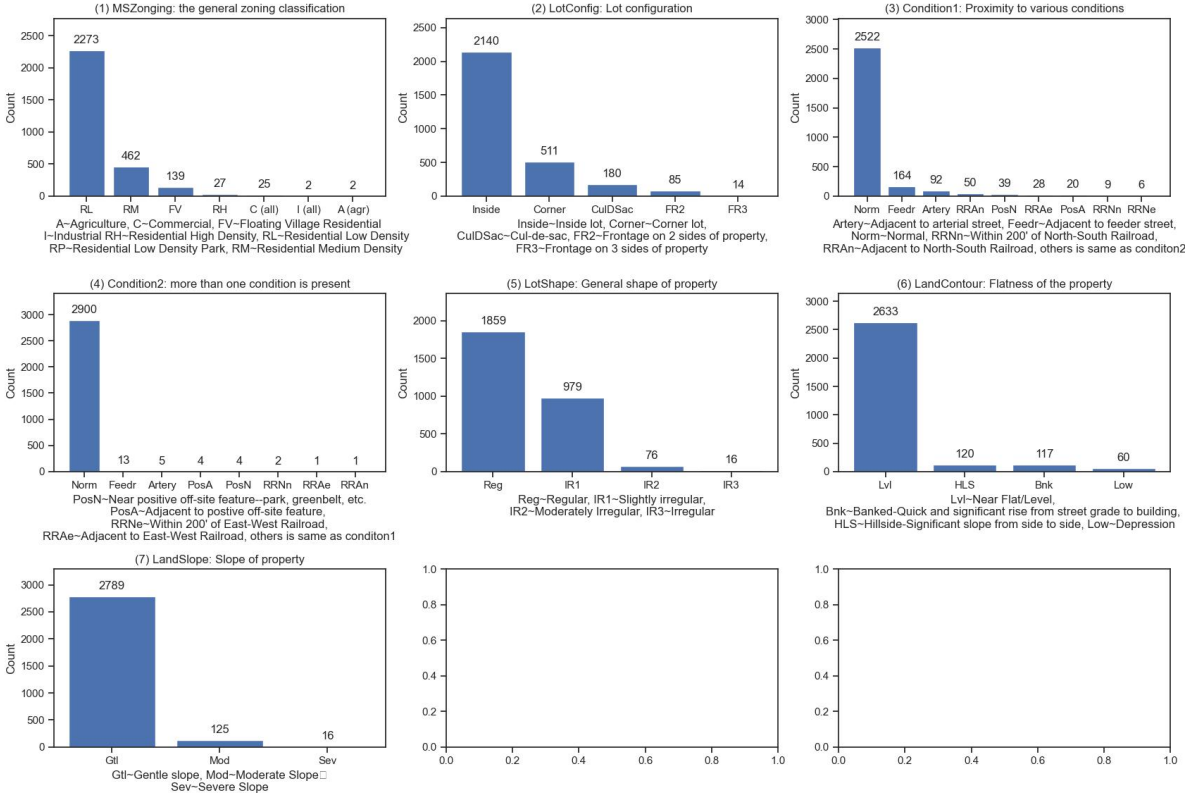
Figure 7: Room Distribution VS SalePrice (Colored by OverallQual)



3.3.2 Location and Geography

Most of the houses are located in low or medium density residential zoning. The majority of lots are situated in interior or corner locations, with flat terrain and either regular or lightly irregular shapes.

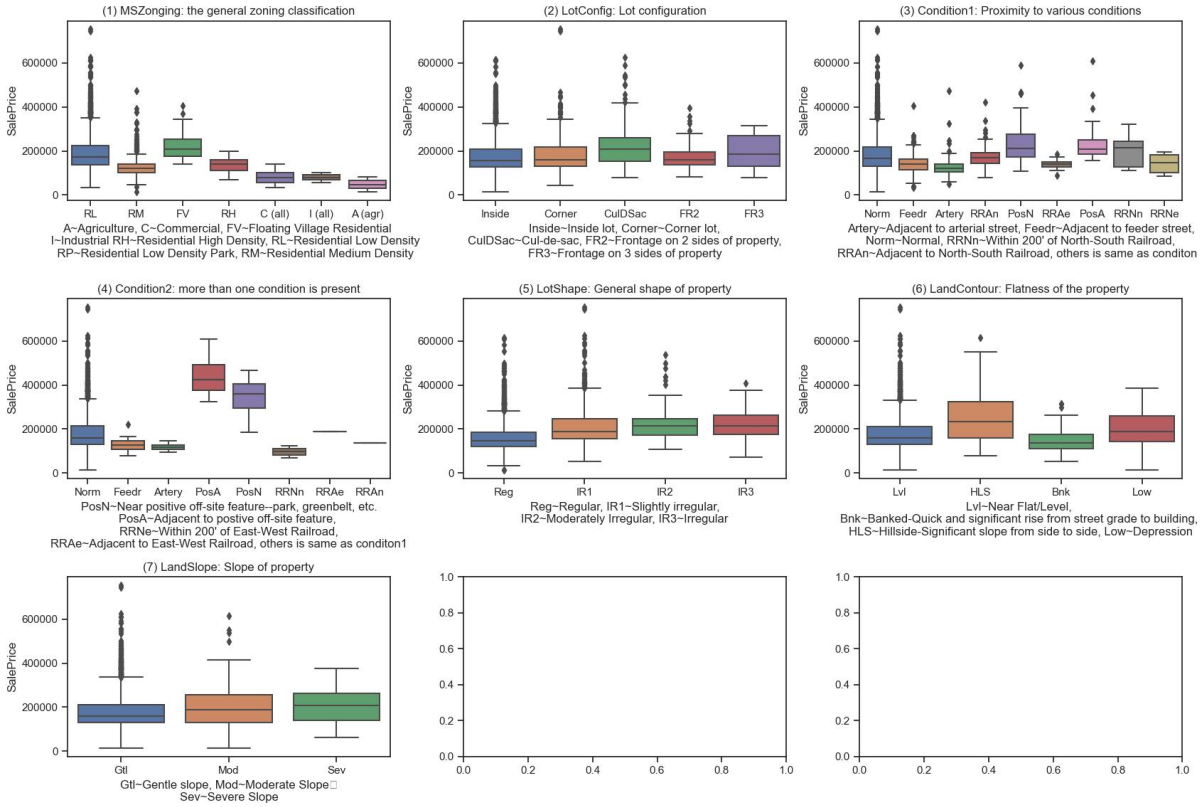
Figure 8: Barchart of House Location and Geography



Houses located close to roads tend to have lower prices while houses in proximity to parks or grasslands generally command higher prices.

An interesting observation is that some individuals may have a preference for hillside houses with a moderate slope, which are associated with higher prices.

Figure 9: Boxplots of Housing Location vs Price



3.3.3 Facilities

The majority of houses are equipped with gas air heaters and hair control, along with 1 to 2 fireplaces. Houses that lack these standard configurations tend to have lower prices.

Figure 10: Barchart of House Facility

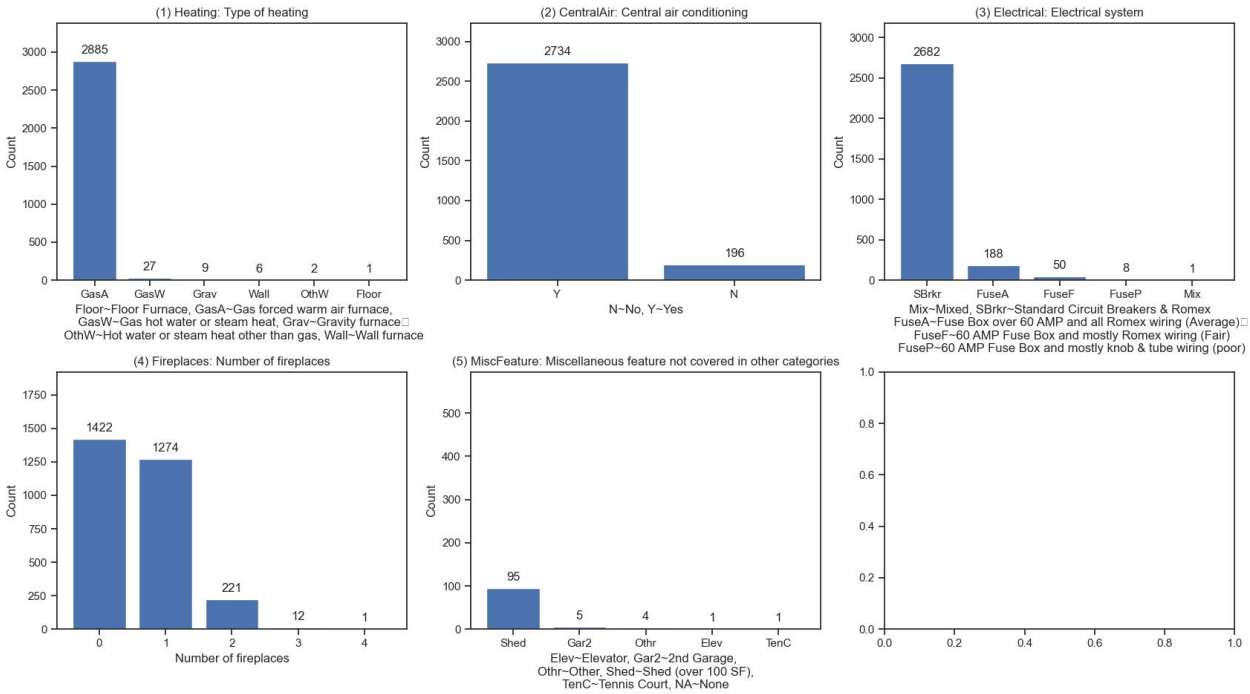
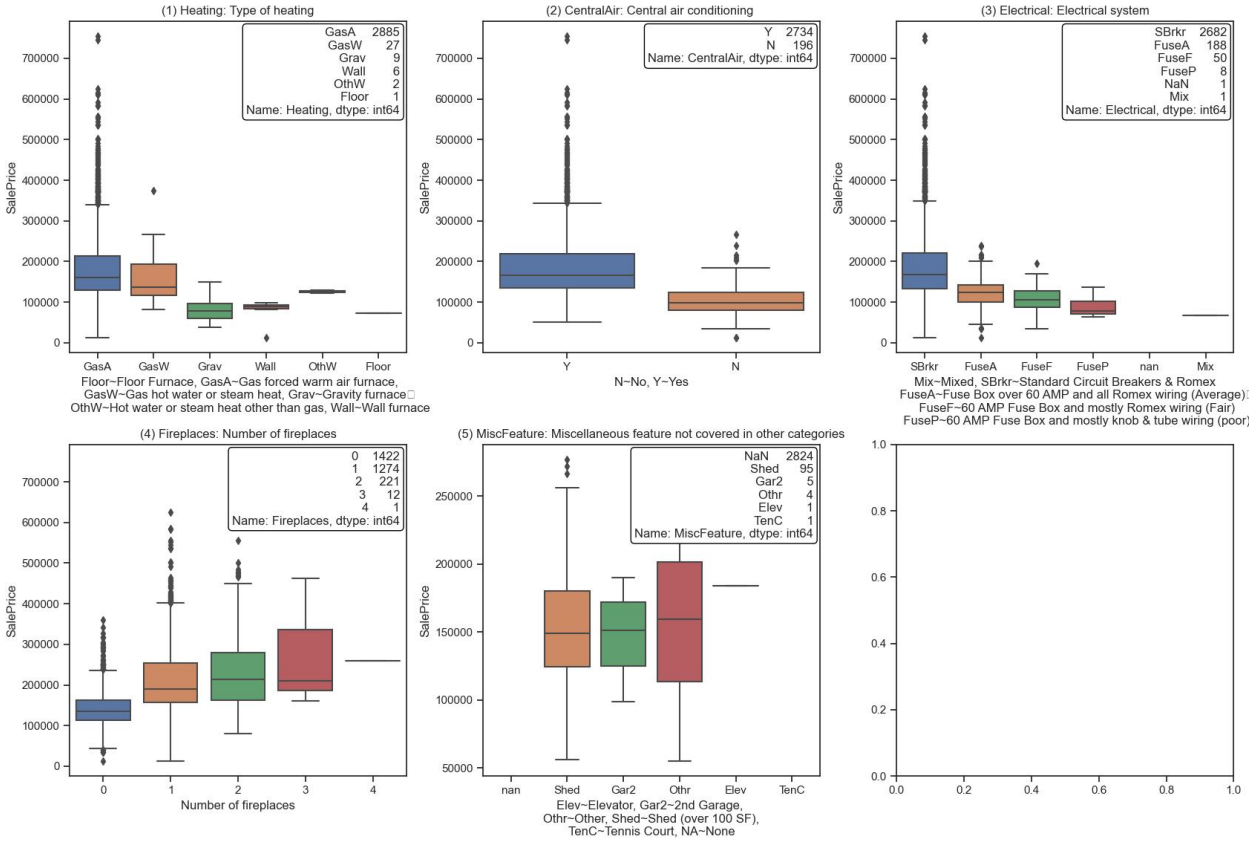


Figure 11: Boxplots of Housing Facility VS Price



3.3.4 Material and Quality

Housing quality tends to concentrate around the average level. As the housing quality improves, the prices tend to increase.

Figure 12: Barchart of House Quality

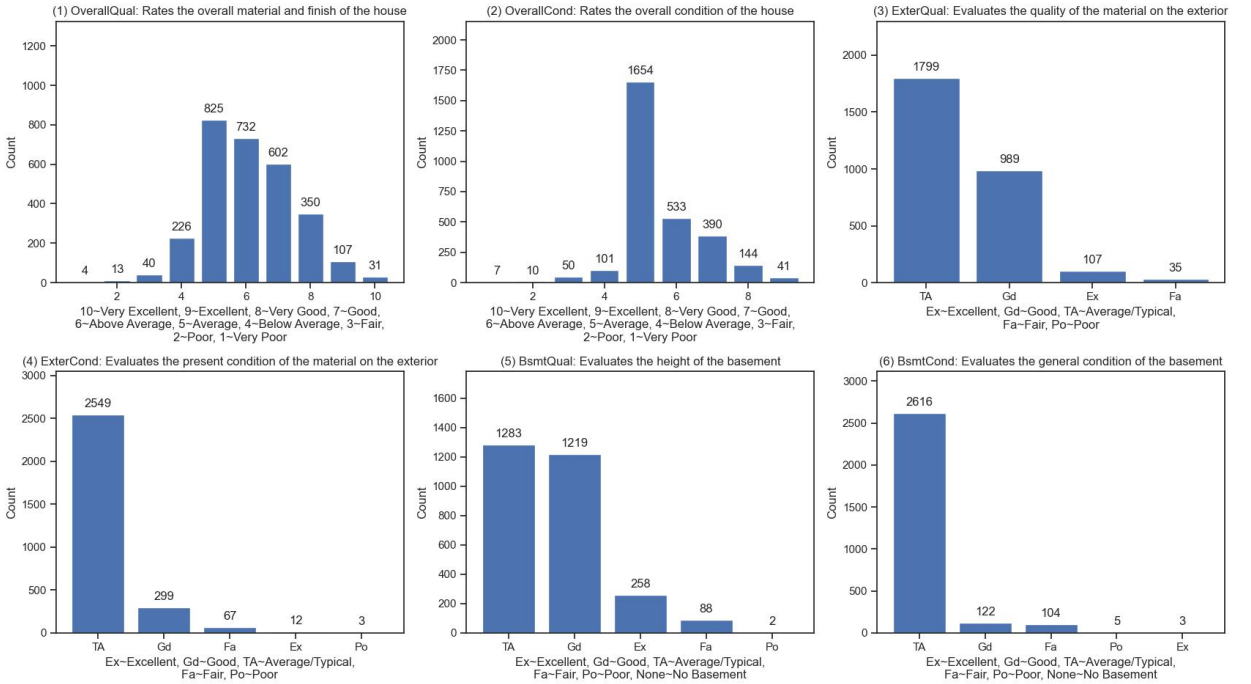


Figure 13: Boxplots of Housing Quality VS Price

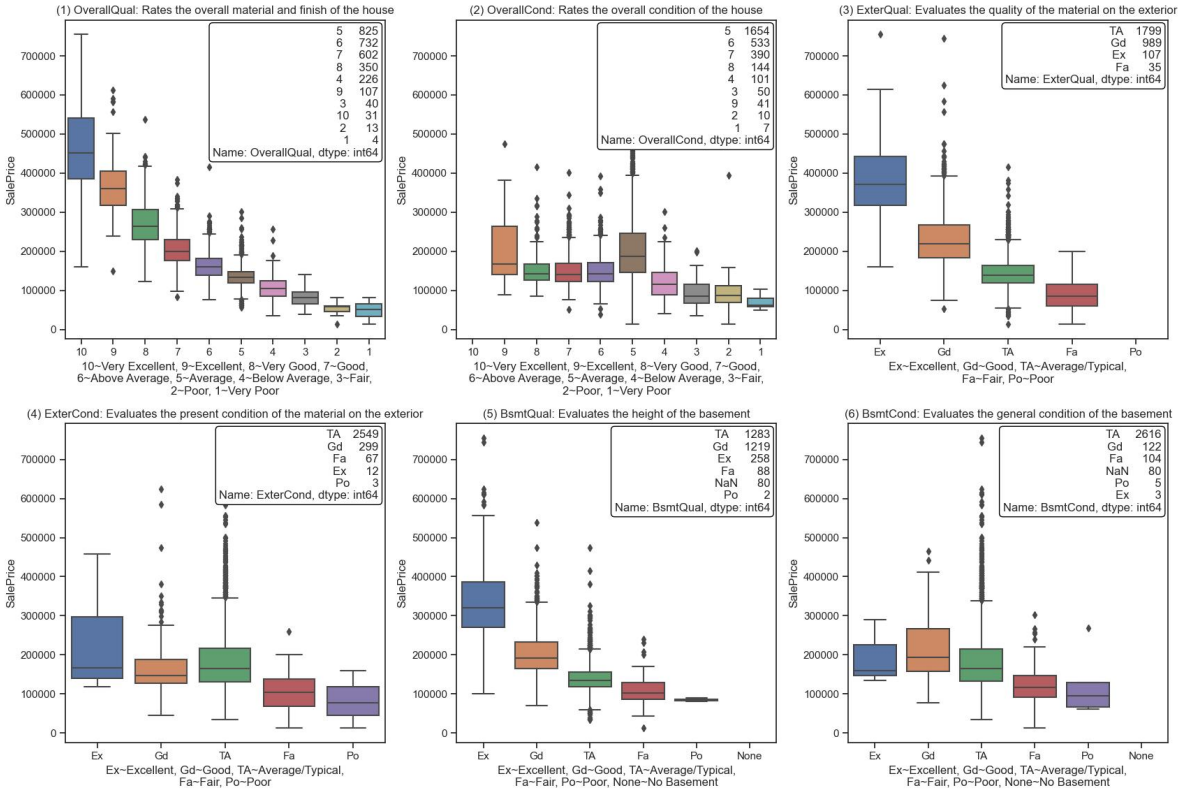
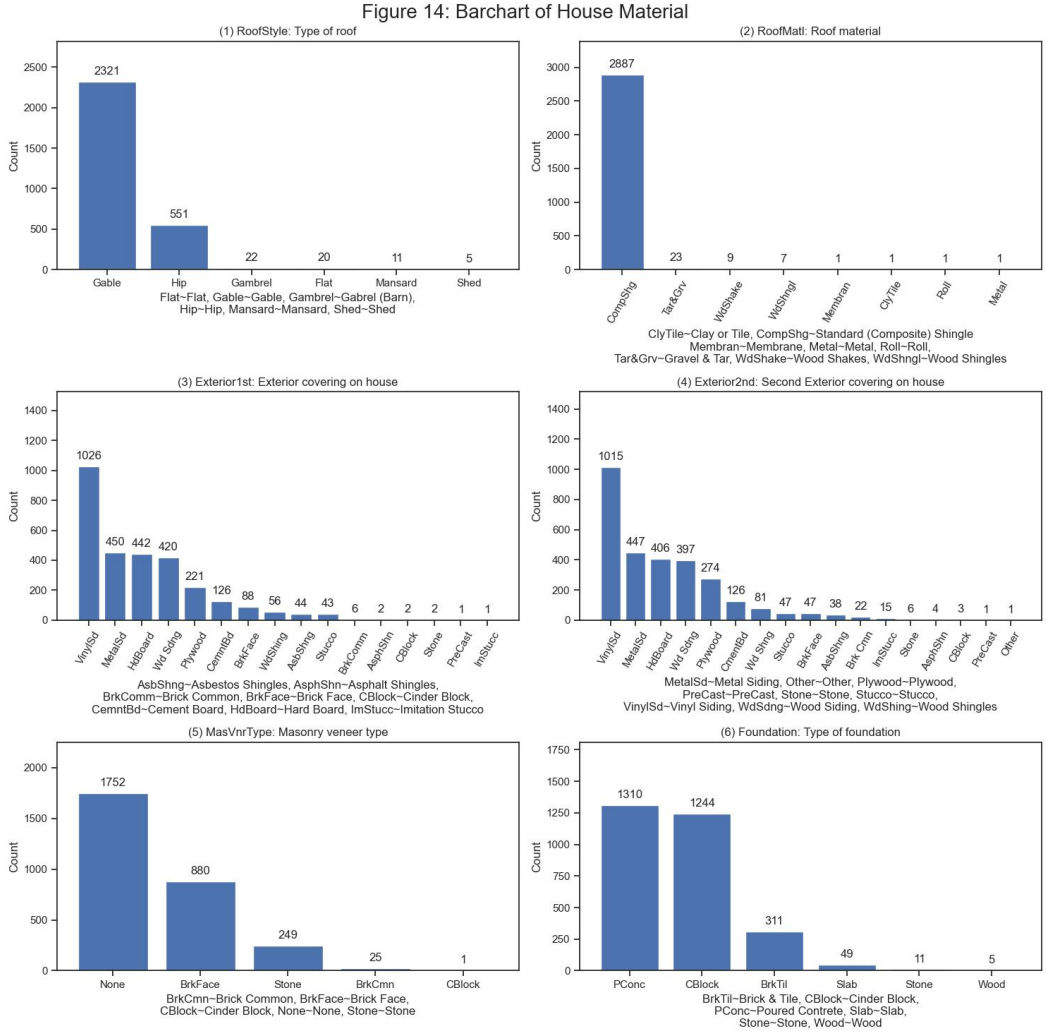


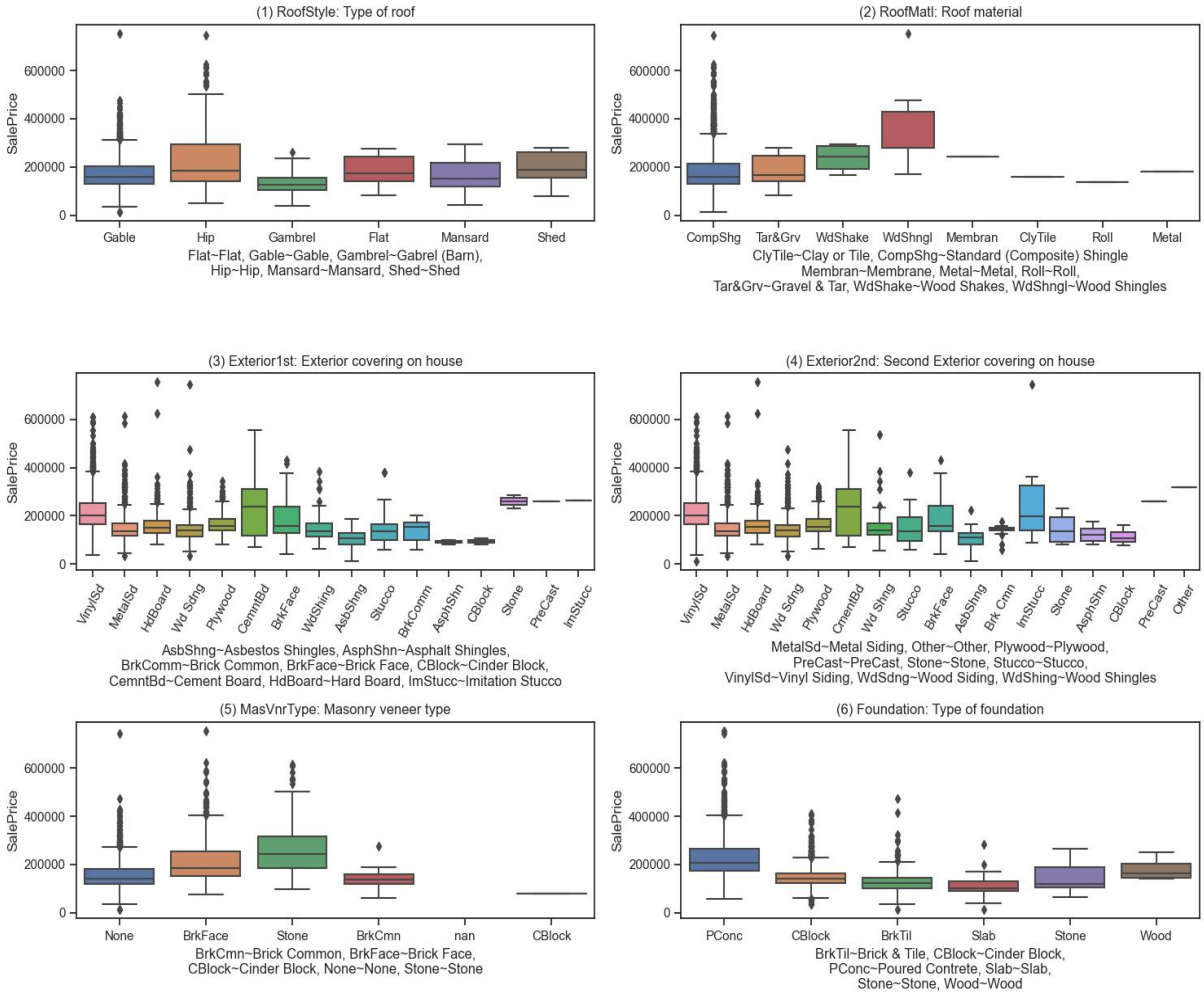
Figure 14 illustrate most houses are characterized by Gable roofs covering with Shingles. The foundation is typically made of Concrete and Cinder blocks. The external walls are usually adorned with Vinyl, metal, or wood siding.

Figure 14: Barchart of House Material



The price of houses with masonry veneer and concrete foundation is notably higher than those without. Additionally, houses with cement board siding tend to a higher mean price.

Figure 15: Boxplots of Housing Matertial VS Price



3.4 Identification of Clusters and Patterns

3.4.1 SalePrice VS Categorical Variables

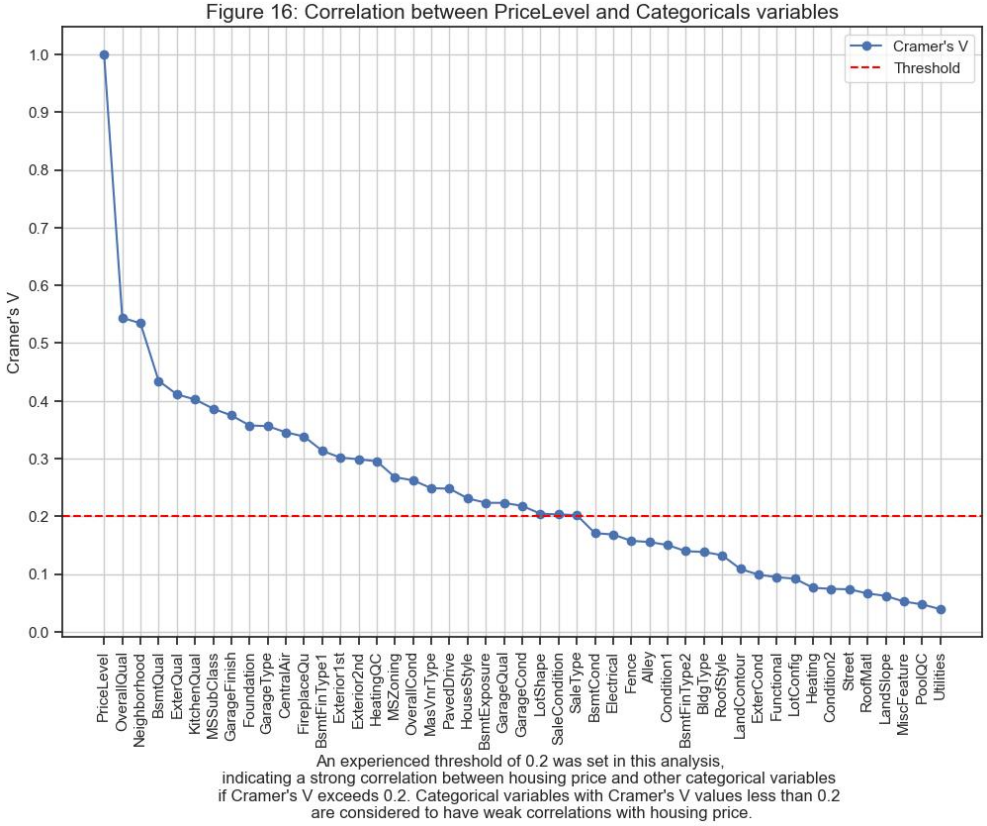
Adinyira et al. (2013) stated that Kruskal-Wallis ANOVA is particularly useful when the sample size is small and to avoid assumptions violations associated with ANOVA. Using this method, Table 3.2 reveals each categorical variable includes at least one category that exhibits a significant correlation with SalePrice,since all p_values are less than 0.05.

Table 3.2 ANOVA

	cat	h_value	p_value
3	SaleCondition	22,699.75	0
4	SaleType	22,335.92	0
5	MiscFeature	21,985.04	0
6	Fence	21,973.81	0
7	PoolQC	21,822.07	0
8	PavedDrive	21,809.68	0
9	GarageCond	21,470.58	0
⋮			
44	LotShape	2,118.29	0
45	Alley	1,750.80	0
46	Street	1,650.01	0
47	MSZoning	1,636.45	0
48	MSSubClass	1,066.59	0.00

Prematunga (2012, p.197) mentioned that Cramer's V is suitable for analyzing the correlation between nominal variables. Figures 16 illustrate the correlation between categorical variables and house prices. The overall housing quality exhibits the strongest association with price, along with 15 other variables that display significant correlations.

Figure 16: Correlation between PriceLevel and Categoricals variables



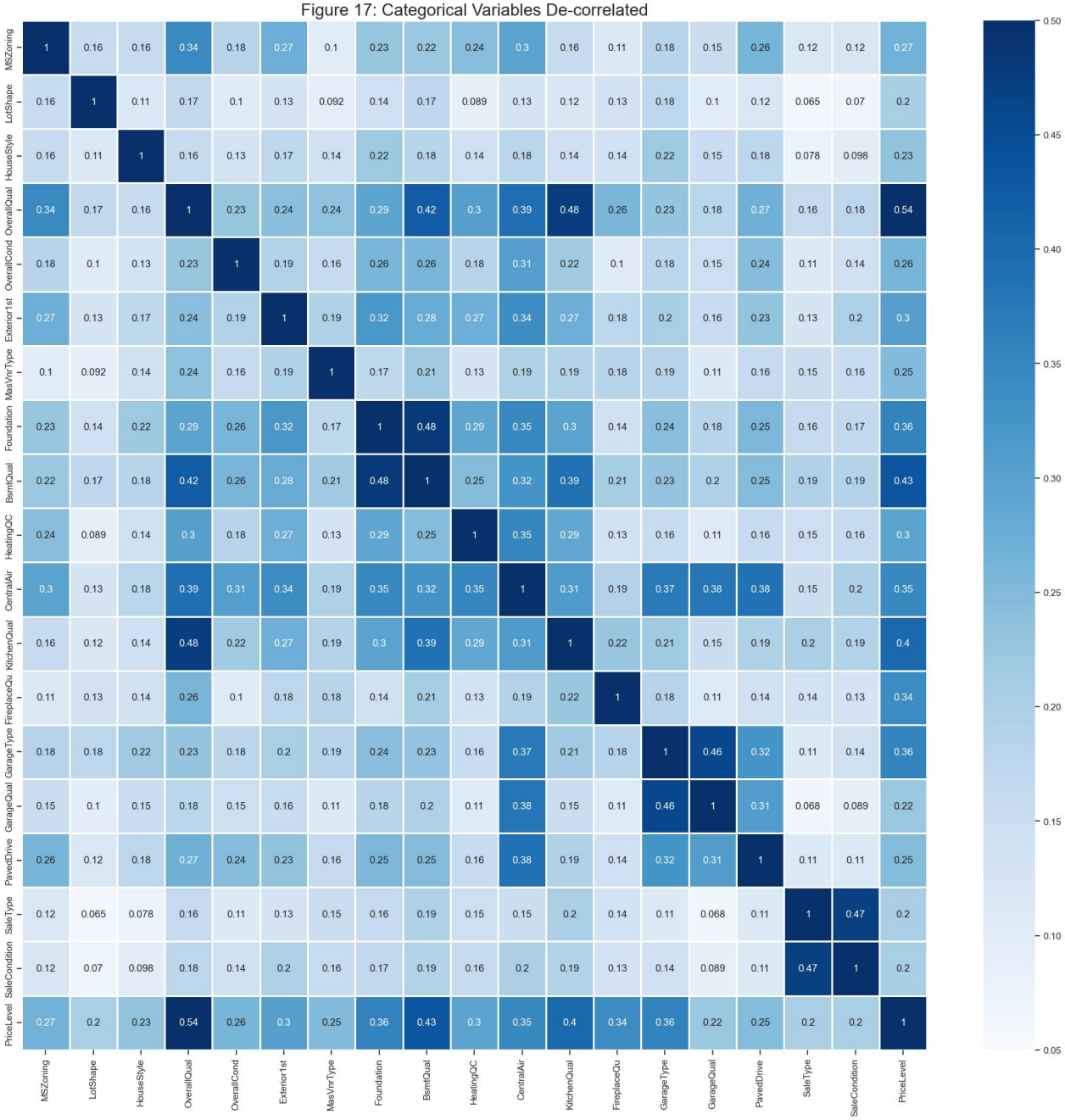
An experienced threshold of 0.2 was set in this analysis,
indicating a strong correlation between housing price and other categorical variables
if Cramer's V exceeds 0.2. Categorical variables with Cramer's V values less than 0.2
are considered to have weak correlations with housing price.

Furthermore, Cramer's V values helps eliminate high correlation features. Feature pairs with Cramer's V values exceeding 0.5 can be considered to eliminate one of the features.

Table 3.1 Chi-square test diagram

Cat1	Cat2	chisq	p	Cramers V	dof
MSSubClass	BldgType	9,296.40	0	0.89	5
MSSubClass	HouseStyle	14,188.45	0	0.83	8
Exterior1st	Exterior2nd	24,232.11	0	0.74	16
GarageType	GarageFinish	4,134.27	0	0.69	4
GarageQual	GarageCond	5,774.22	0	0.63	6
OverallQual	ExterQual	3,332.51	0	0.62	4
GarageFinish	GarageQual	3,094.07	0	0.59	4
GarageFinish	GarageCond	3,050.72	0	0.59	4
ExterQual	KitchenQual	2,611.60	0	0.55	4
MSZoning	Neighborhood	5,200.28	0	0.54	7
BsmtQual	BsmtExposure	3,290.23	0	0.53	5
BsmtExposure	BsmtFinType1	3,237.47	0	0.53	5
BsmtQual	BsmtFinType1	4,024.66	0	0.52	6
Neighborhood	ExterQual	2,251.98	0	0.51	4

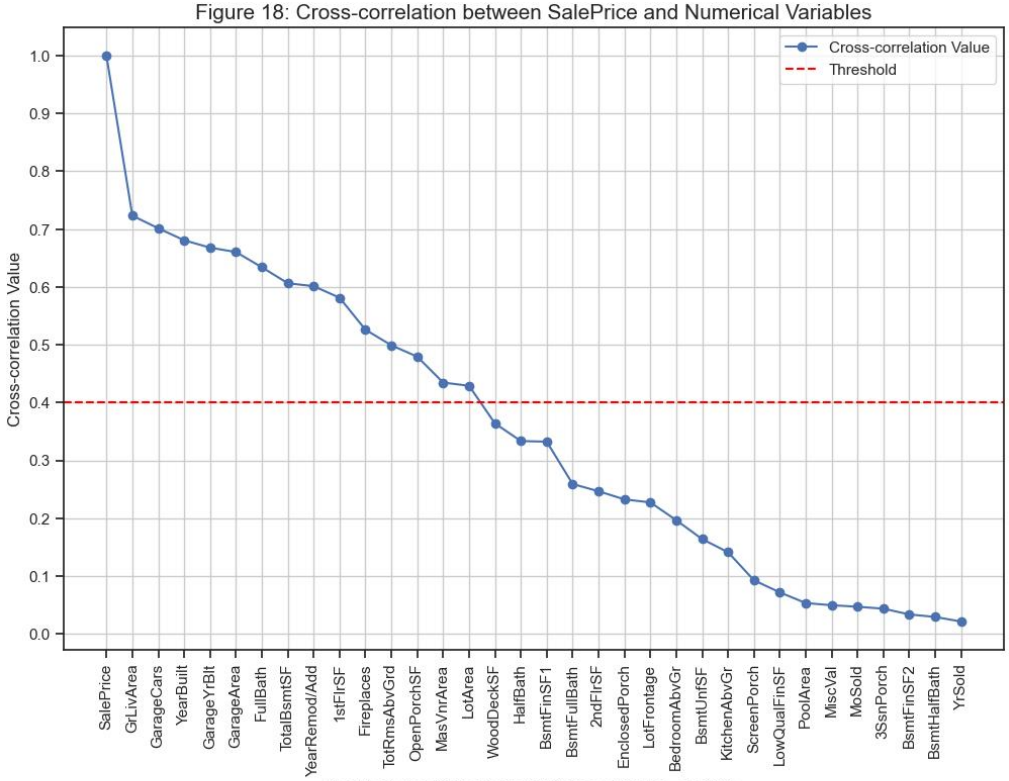
Figure 17: Categorical Variables De-correlated



3.4.2 Numerical Variables VS SalePrice

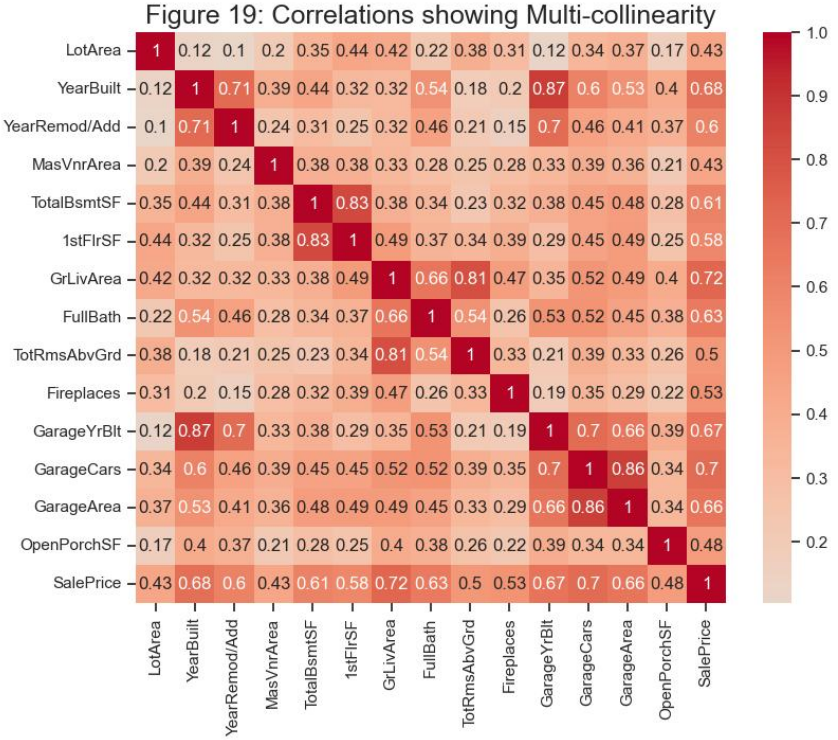
In Figure 18, the ground living area shows the strongest correlation with SalePrice. The presence of potential multi-collinearity among these variables is depicted in Figure 19. To address this issue, 7 new variables were created to mitigate the collinearity effects, as depicted in Table 3.3.

Figure 18: Cross-correlation between SalePrice and Numerical Variables



An experienced threshold of 0.4 was set in this analysis, indicating a strong correlation between housing price and other numberical variables if value exceeds 0.4. numerical variables with cross-correlation values less than 0.4 are considered to have weak correlations with housing price.

Figure 19: Correlations showing Multi-collinearity



3.4.3 Result of Identification

Based on EDA, the project excluded 57 original variables. In line with Galli's (2022, p.217) approach, mathematical functions and reference variables were employed to combine features, as depicted in Table 3.3.

Table 3.3 Result of Clusters and Patterns

Items	Feature Name	Count
Deleted numeric features	'LotFrontage','BsmtFinSF1','BsmtFinSF2','BsmtUnfSF','2ndFlrSF','LowQualFinSF','BsmtFullBath', 'BsmtHalfBath','HalfBath','BedroomAbvGr','KitchenAbvGr','WoodDeckSF','EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'MiscVal', 'MoSold', 'YrSold', 'PID', 'Order' 'GarageYrBlt','GarageArea','GarageCars','GrLivArea','FullBath','TotRmsAbvGrd','TotalBsmtSF', '1stFlrSF','OpenPorchSF',"YearRemod/Add"	29
Deleted categorical features	'BsmtCond','Electrical','Fence','Alley','Condition1','BsmtFinType2','BldgType','RoofStyle', 'LandContour','ExterCond','Functional','LotConfig','Heating','Condition2','Street','RoofMatl', 'LandSlope','MiscFeature','PoolQC','Utilities','MSSubClass','Exterior2nd','GarageFinish', 'ExterQual','Neighborhood','BsmtExposure','BsmtFinType1','GarageCond'	28
Reserved numerical features	'LotArea','YearBuilt','MasVnrArea','Fireplaces'	4
New built numerical features	GarAreaPerCar': the garage area per car, 'TotalHouseSF': the total floor area of a house, GrLivAreaPerRoom':The average area per room on the ground floor, 'TotalFullBath': the total number of full bathroom, 'TotalPorchSF': the total area of porch, 'IsRemodGar': If the garage is built after the house is constructed, 'IsRemod': Whether the house has been renovated.	7
Reserved categorical features	'MSZoning','LotShape','HouseStyle','OverallQual','OverallCond','Exterior1st','MasVnrType', 'Foundation','BsmtQual','HeatingQC','CentralAir','KitchenQual','FireplaceQu','GarageType', 'GarageQual','PavedDrive','SaleType','SaleCondition'	18

4 Refine Questions

RQ1: Describe in detail the impact of the type, size, quality and zoning of the house on the price of the house.

RQ2: Evaluate the importance of built year on the predicted house price model.

RQ3: Utilize models to quantify the impact on prices by adding additional bedrooms, while keeping other features in the same condition.

RQ4: Can these key features effectively classify different price levels for houses?

Reference

Downey, A.B., Loukides, M.K., Blanchette, M., Kersey, A., Montgomery, K. and Demarest, R., 2015. *Think stats : exploratory data analysis.* Sebastopol, California: O'Reilly.

Galli, S., 2022. *Python feature engineering cookbook, Second edition.* Birmingham, England: Packt Publishing, Limited.

Adinyira, E., Ahadzie, D.K. and Kwofie, T.E., 2013. *Determining the unique features of mass housing projects (MHPs)*. In Proceedings of the 5th West Africa Built Environment Research (WABER) Conference, Accra, Ghana.

Prematunga, R.K., 2005. Correlational analysis. *Australian Critical Care : Official Journal of the Confederation of Australian Critical Care Nurses*, 25(3), pp.195–199.