1. Project & Data Introduction

In this project, I will be conducting a comprehensive analysis of the dataset provided by the Kaggle challenge, "House Prices - Advanced Regression Techniques." This competition presents a rich dataset consisting of 1460 observations of homes in Ames, Iowa, with a detailed set of 81 features that capture various aspects of these properties. These features encompass a wide array of variables, including physical characteristics of the houses, such as their size, age, and number of rooms, as well as qualitative factors like the quality of materials used and the condition of the house. Additionally, the dataset includes contextual information, such as the geographic location of the properties, which can play a significant role in determining home values.

2.Data Loading

First of all, I will load up the data and checking the data to see if the data is containing any null value or value that I won't be use later

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	Sa
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2008	WD	
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2007	WD	
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	9	2008	WD	
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2006	WD	
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	12	2008	WD	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1455	1456	60	RL	62.0	7917	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	8	2007	WD	
1456	1457	20	RL	85.0	13175	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0	2	2010	WD	
1457	1458	70	RL	66.0	9042	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	GdPrv	Shed	2500	5	2010	WD	
1458	1459	20	RL	68.0	9717	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	4	2010	WD	
1459	1460	20	RL	75.0	9937	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	6	2008	WD	

1460 rows × 81 columns

By checking the data frame we can sees that the dataset contains 81 columns and 1460 rows, representing various features of houses in Ames, Iowa. Key features in the dataset include **MSSubClass**, **MSZoning**, **LotFrontage**, **LotArea**, and **SalePrice**, which is the target variable for predicting house prices. Some columns have missing values, such as **LotFrontage**, **Alley**, and **Fence**, which will need to be addressed during data preprocessing. The dataset includes both categorical and numerical features, requiring appropriate encoding for modeling. The variability in the **SalePrice** indicates that multiple factors influence house prices.

3.Data Cleaning

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType
0	False	False	False	False	False	False	True	False	False	False	...	False	True	True	True	False	False	False	False
1	False	False	False	False	False	False	True	False	False	False	...	False	True	True	True	False	False	False	False
2	False	False	False	False	False	False	True	False	False	False	...	False	True	True	True	False	False	False	False
3	False	False	False	False	False	False	True	False	False	False	...	False	True	True	True	False	False	False	False
4	False	False	False	False	False	False	True	False	False	False	...	False	True	True	True	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1455	False	False	False	False	False	False	True	False	False	False	...	False	True	True	True	False	False	False	False
1456	False	False	False	False	False	False	True	False	False	False	...	False	True	False	True	False	False	False	False
1457	False	False	False	False	False	False	True	False	False	False	...	False	True	False	False	False	False	False	False
1458	False	False	False	False	False	False	True	False	False	False	...	False	True	True	True	False	False	False	False
1459	False	False	False	False	False	False	True	False	False	False	...	False	True	True	True	False	False	False	False

1460 rows × 81 columns

Then, to better see the features that have null, we will use data.isna, after generating the data frame of the data set we can see there are several columns containing Null values (**Alley**, **PoolQC**, **Fence**, **MiscFeature** etc...)

```
# Drop:Alley, PoolQC, Fence, MiscFeature
data.drop(['Alley', 'PoolQC', 'Fence', 'MiscFeature'], axis=1, inplace=True)
```

Then we will perform the feature drop, since ('Alley', 'PoolQC', 'Fence', 'MiscFeature') is none fillable and it can't be use for further analysis so we can drop it and regenerate the data frame after dropped the features.

4.Topic 1 **Impact of Physical Characteristics on House Prices**

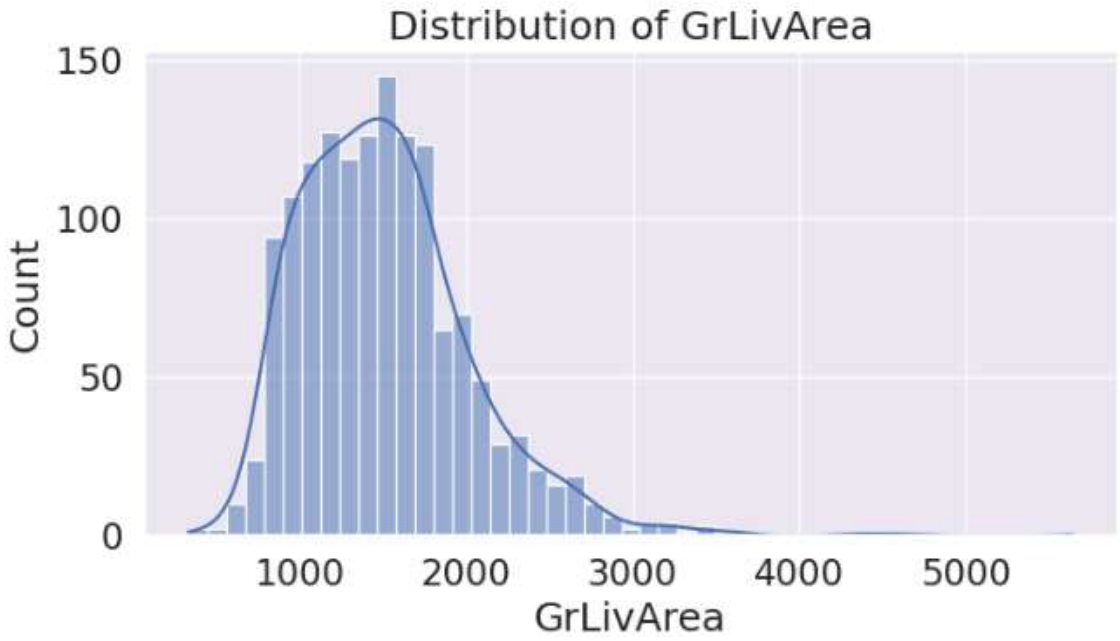
In order to find out “Impact of Physical Characteristics on House Prices” we first need to import the relative features for analysis

	LotFrontage	LotArea	OverallQual	TotalBsmtSF	1stFlrSF	2ndFlrSF	LowQualFinSF	GrLivArea	FullBath	BedroomAbvGr	Fireplaces	GarageArea	PoolArea	SalePrice
0	65.0	8450	7	856	856	854	0	1710	2	3	0	548	0	208500
1	80.0	9600	6	1262	1262	0	0	1262	2	3	1	460	0	181500
2	68.0	11250	7	920	920	866	0	1786	2	3	1	608	0	223500
3	60.0	9550	7	756	961	756	0	1717	1	3	1	642	0	140000
4	84.0	14260	8	1145	1145	1053	0	2198	2	4	1	836	0	250000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1455	62.0	7917	6	953	953	694	0	1647	2	3	1	460	0	175000
1456	85.0	13175	6	1542	2073	0	0	2073	2	3	2	500	0	210000
1457	66.0	9042	7	1152	1188	1152	0	2340	2	4	2	252	0	266500
1458	68.0	9717	5	1078	1078	0	0	1078	1	2	0	240	0	142125
1459	75.0	9937	5	1256	1256	0	0	1256	1	3	0	276	0	147500

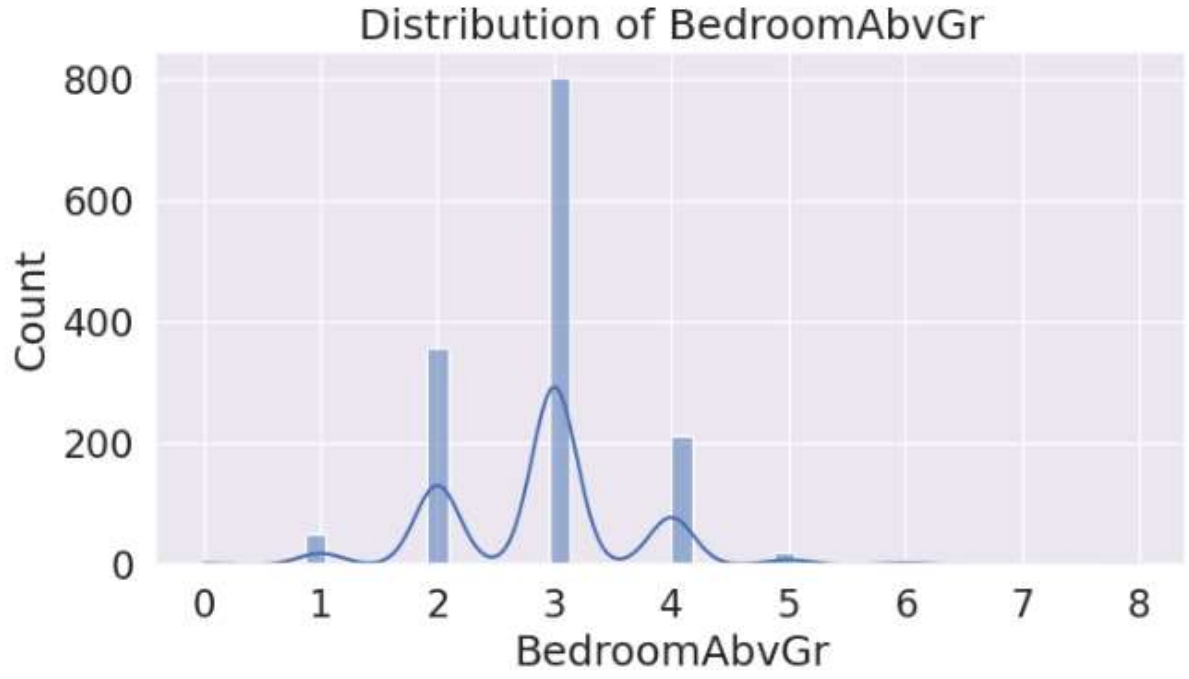
1460 rows × 14 columns

After look through the dataset description I picked 14 features for analysis in Topic 1
Then we need to analysis the distribution of each features before we build up model and draw out EDA

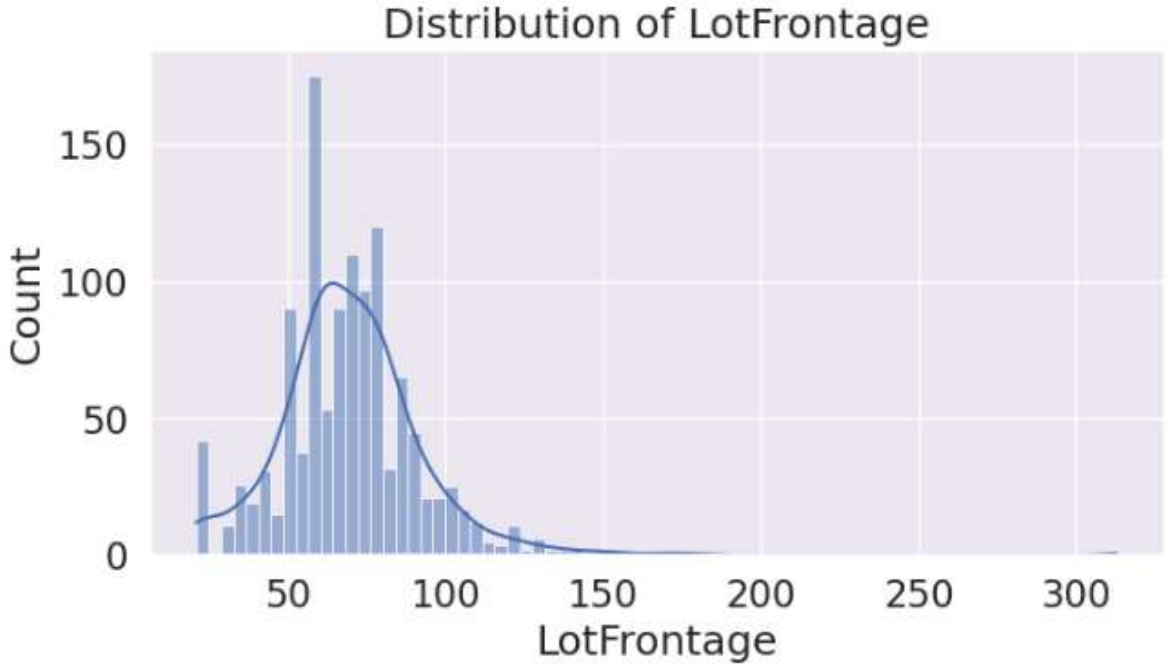
Distribution Analyze



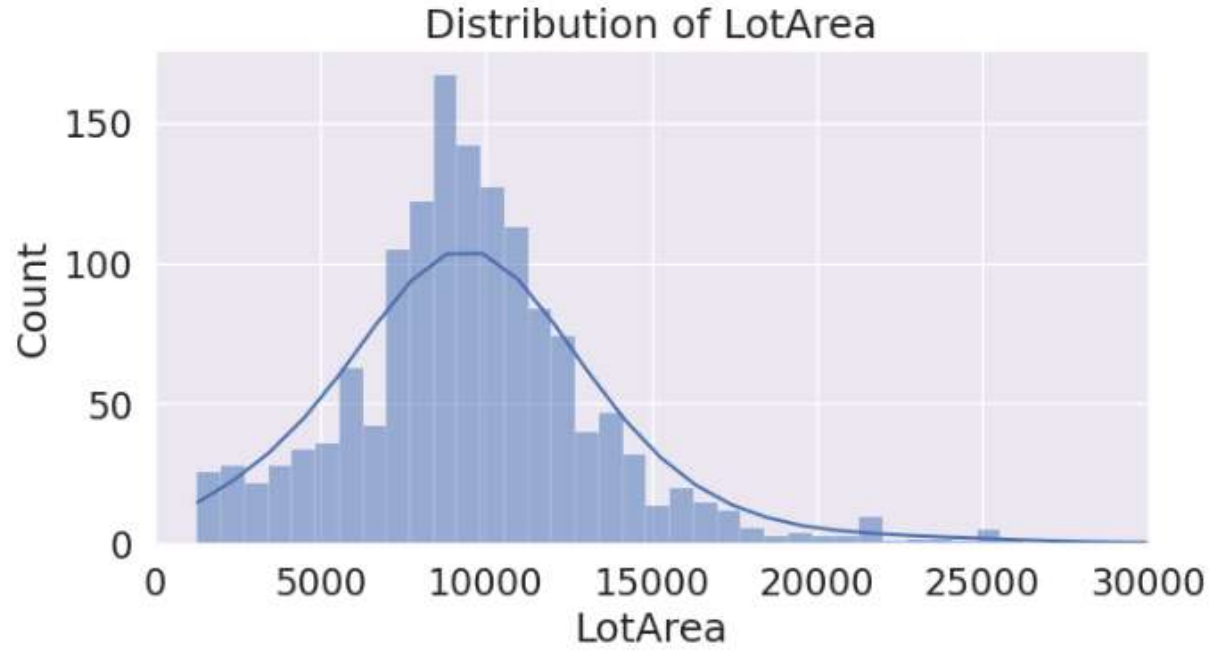
'GrLivArea': The distribution of the living area above ground is primarily concentrated between 1000 to 2000 square feet. Only a few houses have living areas above 2000 square feet, with their count typically less than 50. As the square footage increases, the number of houses decreases.



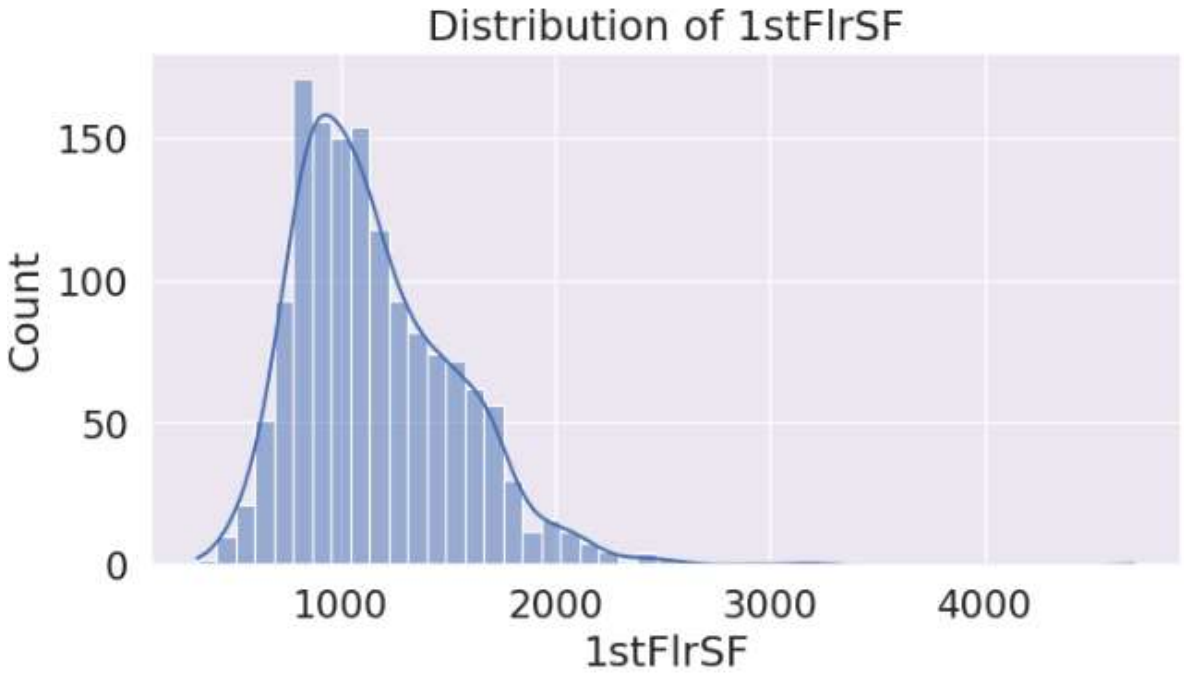
'BedroomAbvGr': According to the graph, the majority of houses have 2 to 3 bedrooms, with about a thousand counts. Houses with one or five bedrooms form a smaller group, each with fewer than 50 counts.



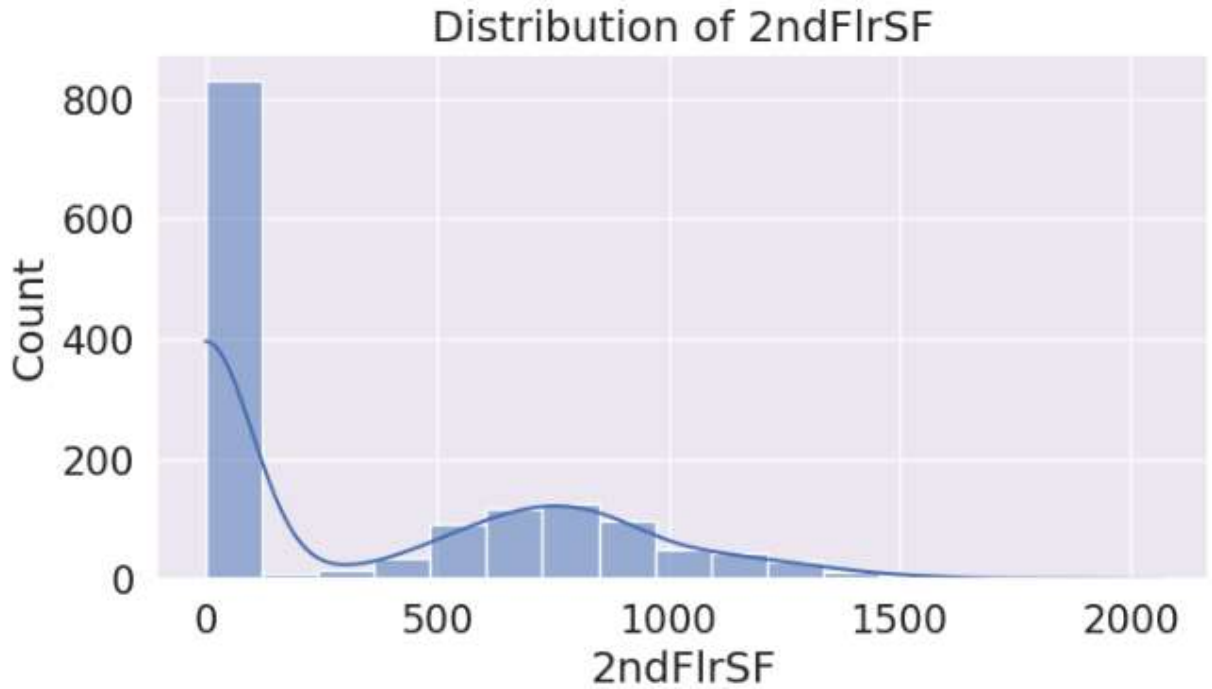
'LotFrontage': The housing frontage area is mainly between 50 to 100 square feet, which is unusual since, according to the distribution of **GrLivArea**, it is spread out evenly. The distribution of LotFrontage shouldn't show such a significant difference in central square footage.



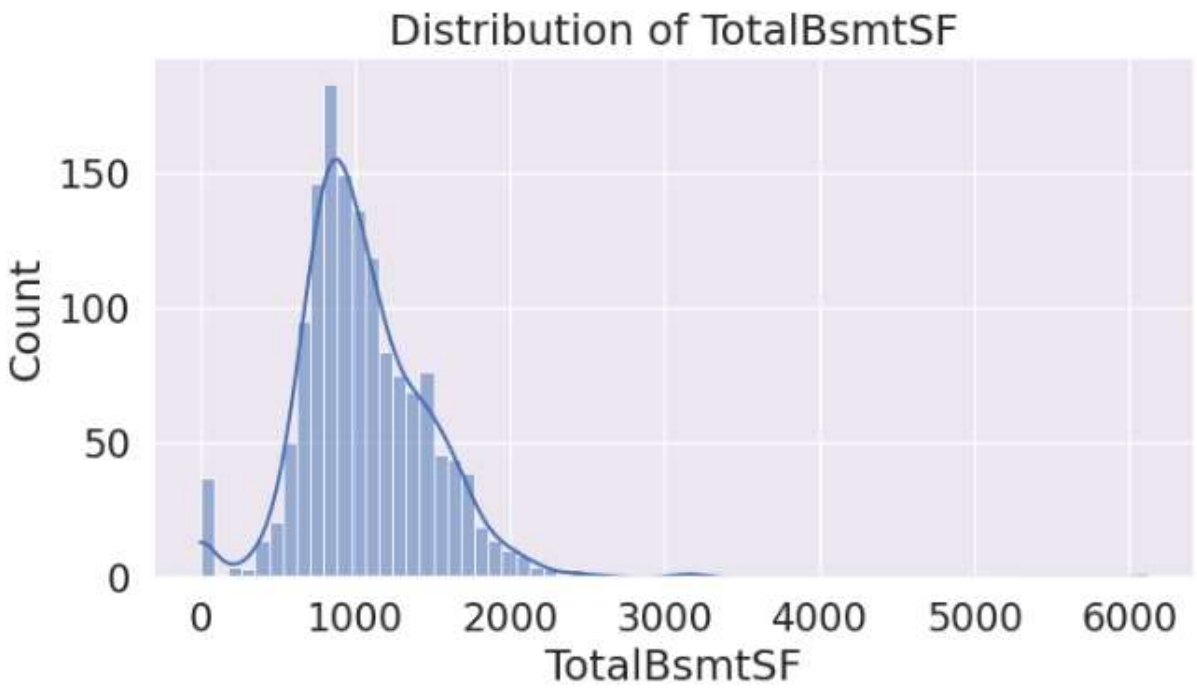
'LotArea': The total area in square feet shows a reasonable distribution close to the distribution of **GrLivArea**. This feature can be used to analyze the relationship between physical characteristics and sale price later.



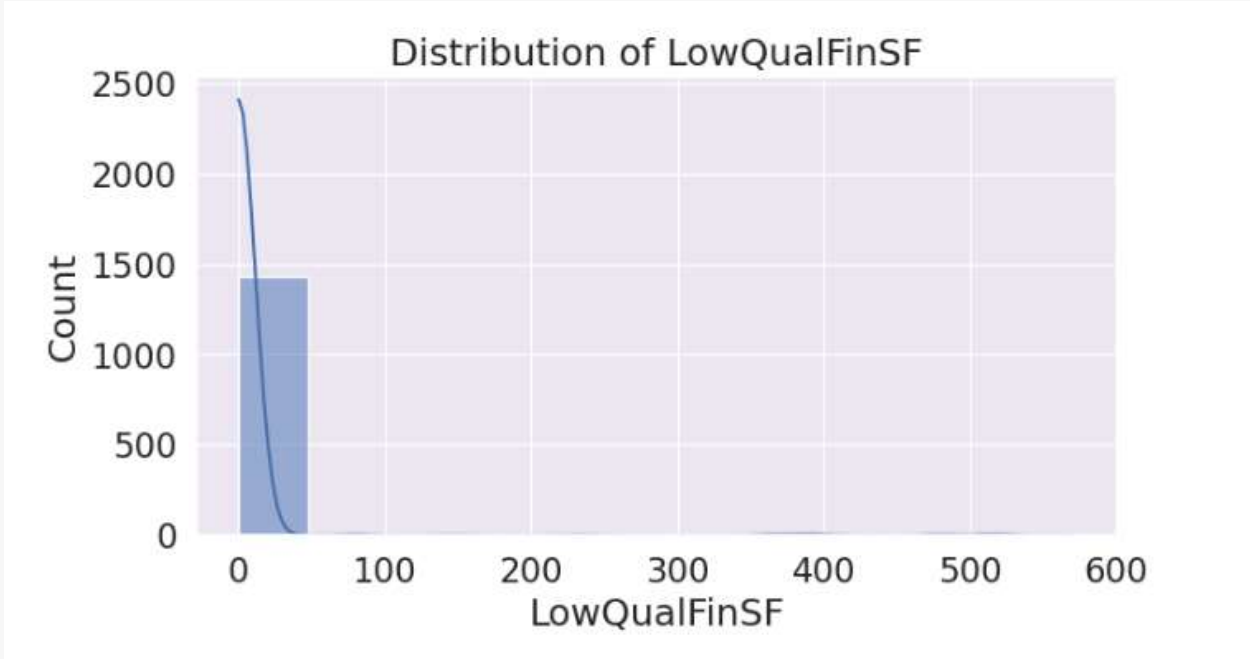
'1stFlrSF': The first floor square footage is mainly between 1000 to 1500, which is reasonable since the distribution of 1stFlSF is close to that of **GrLivArea**. Both have an even distribution.



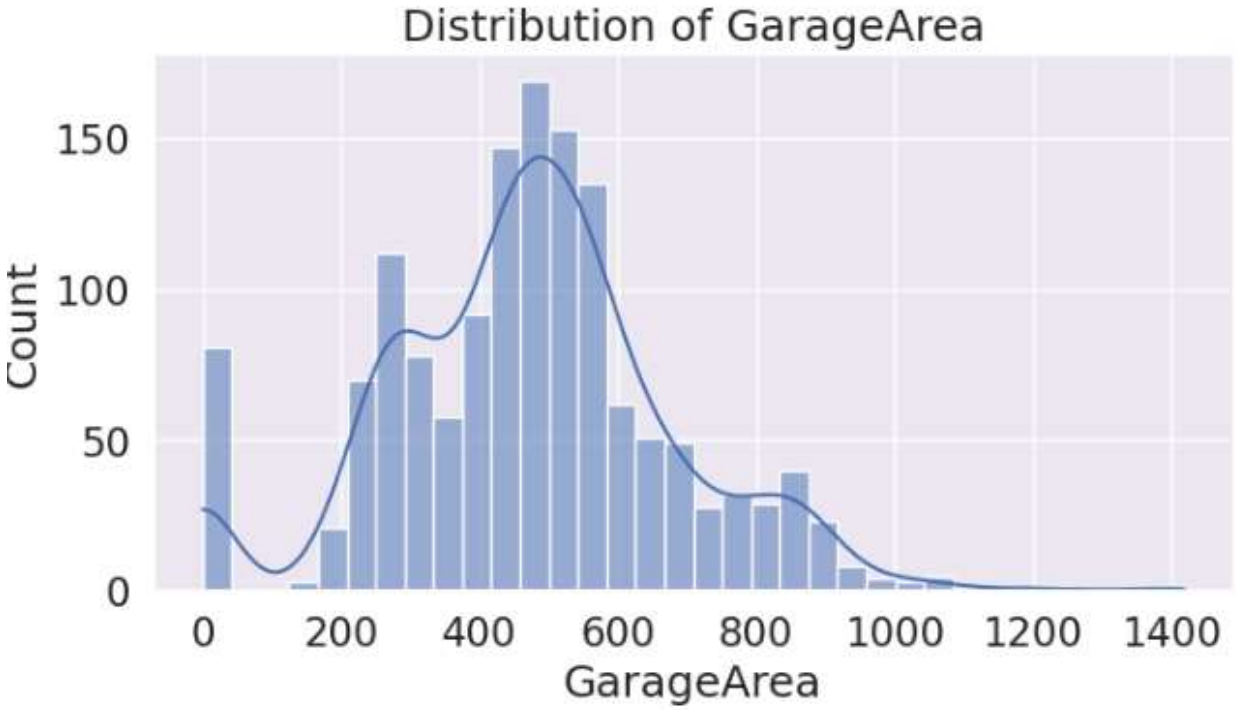
'2ndFlrSF': This column is interesting and worth deeper investigation in the future. According to the graph, fewer than 600 houses have a second floor, while the rest are single-story homes. Given that we have about 1500 houses in our dataset, only 50% have a second floor, which is a notable percentage.



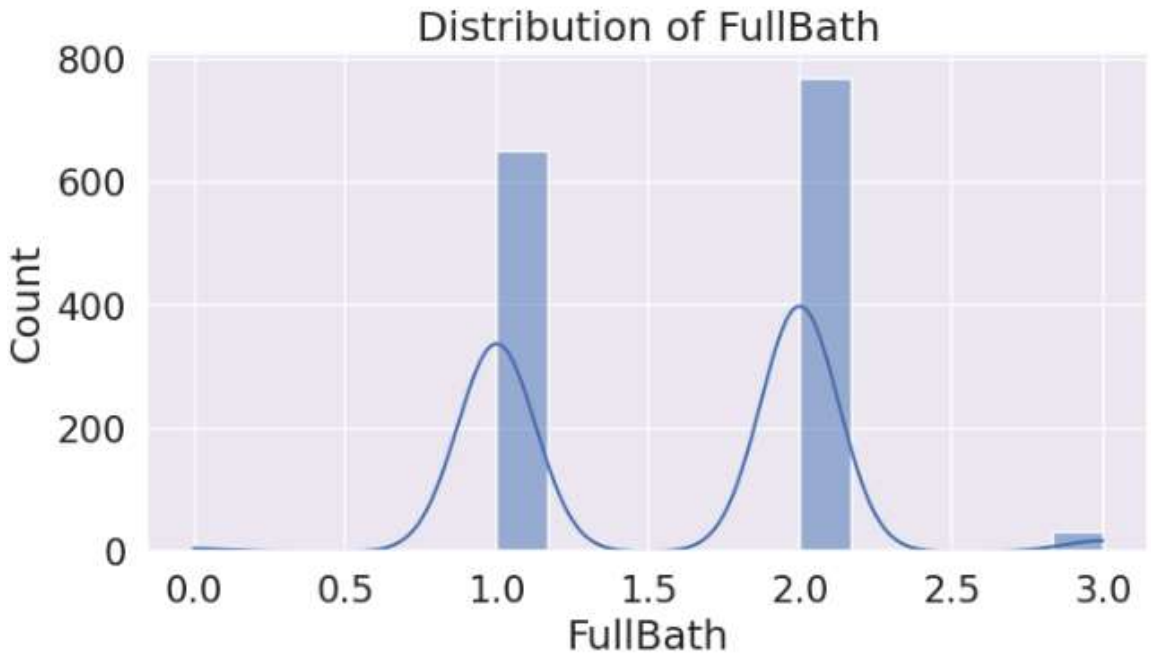
'TotalBsmtSF': The total square footage of the basement area shows a reasonable distribution, with most houses having between 1000 to 2000 square feet.



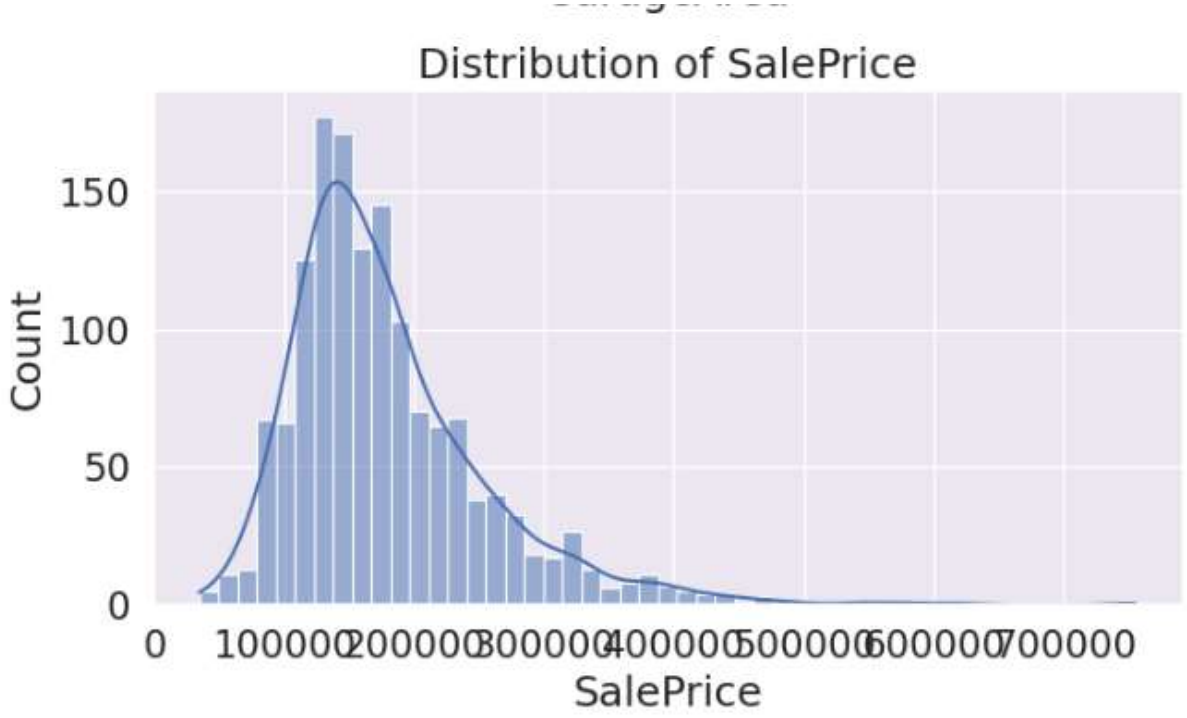
'LowQualFinSF': The distribution of low-quality finished square footage does not provide much useful information, as most houses have zero low-quality finished square footage. This feature could be influenced by multiple external factors such as company ratings, weather, or man-made damage. We may need extra data if we want to delve deeper into this feature.



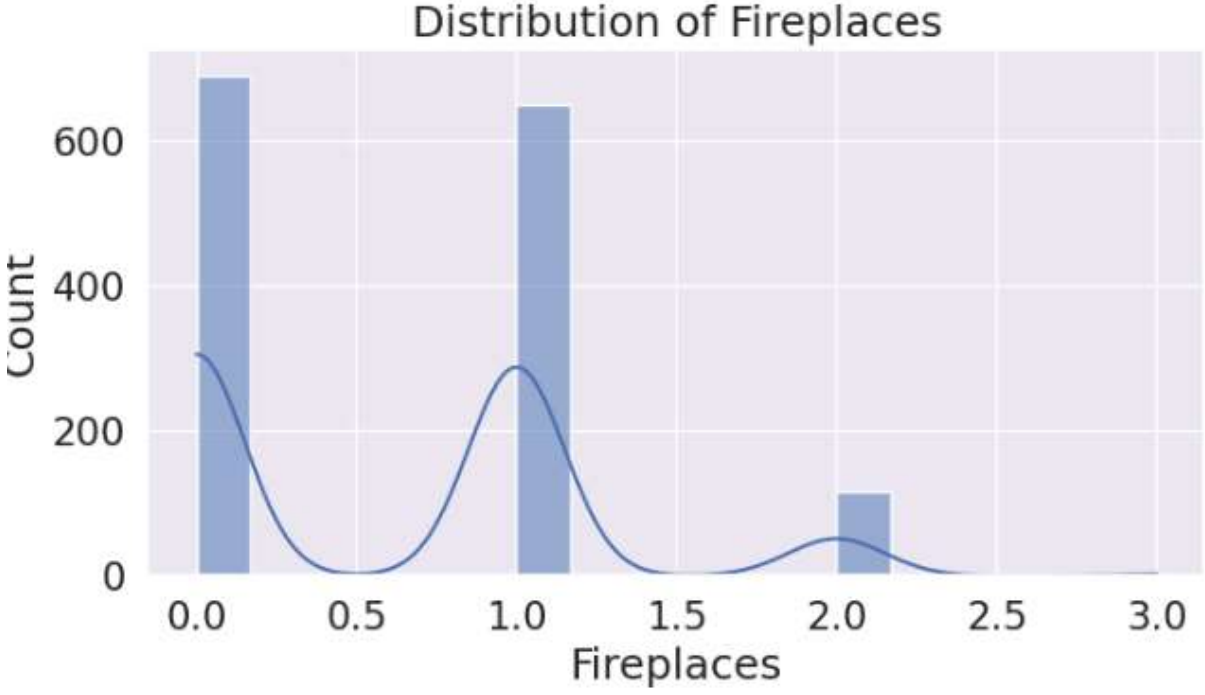
'GarageArea': The garage area has a noticeable bump in the distribution, which is unusual since most garages are designed for two cars.



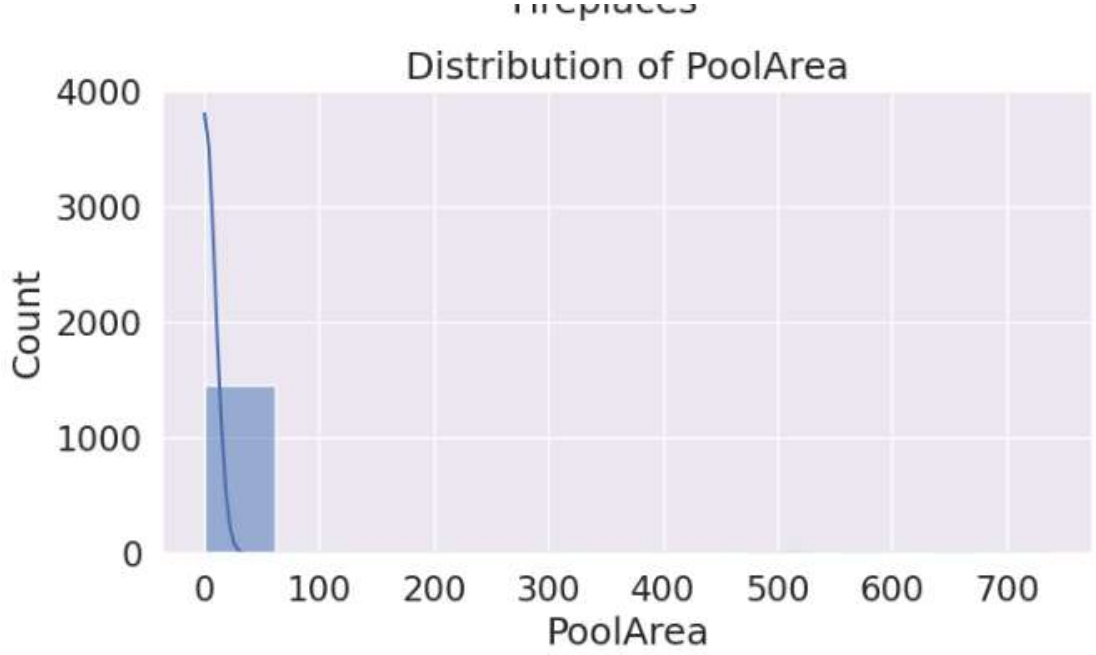
'FullBath': The number of bathrooms is dropping between 1 and 2 with less than 50 houses having 3 bathrooms.



'SalePrice': The distribution of sale prices follows a normal distribution skewed to the left. This is reasonable since most of our features also show a normal distribution skewed to the left (**1stFlrSF**, **LotArea**, **GrLivArea**).



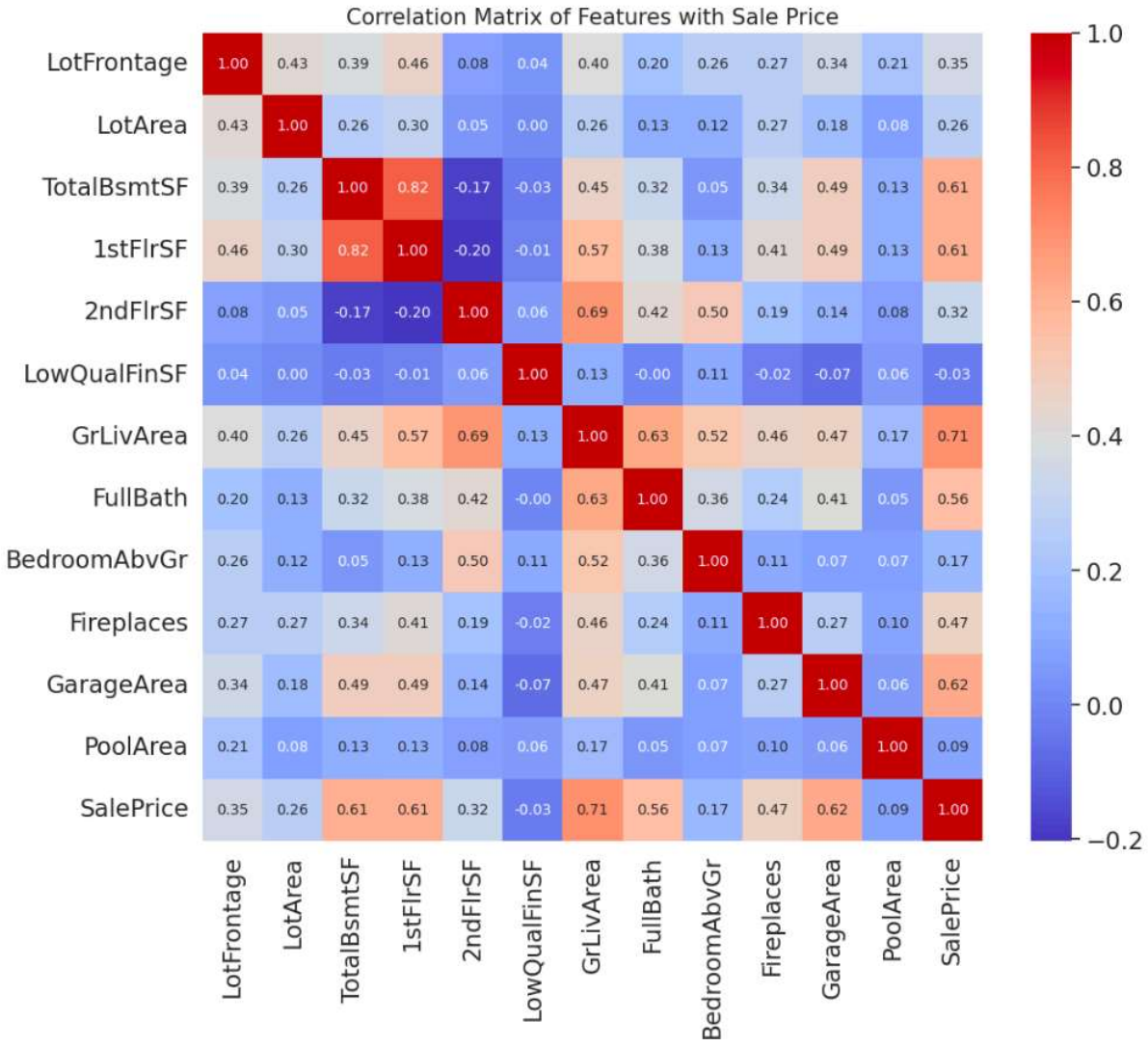
'FirePlace': The distribution of of the fireplace indicate that nearly half of the housing don't have a fireplace, which makes the research valuable, since is easier find out if fireplace will cause price change



'PoolArea': This feature seems not providing much information for the research since all the housing is having a same size pool

Correlation Analyze

After analyze the distribution of each features we now find the feature that most correlate to the sale price



When I analyze this correlation heat map, I observe that **SalePrice** is most strongly correlated with features like **GrLivArea**, **TotalBsmtSF**, and **GarageArea**, all showing fairly high positive correlations. This suggests that larger living spaces, greater basement areas, and bigger garage spaces are significant predictors of higher house prices. Additionally, there's a notable positive correlation between **1stFlrSF** and both **TotalBsmtSF** and **GarageArea**, indicating some degree of multicollinearity among these features, which I'll need to consider when building my models. Interestingly, features like **PoolArea** and **LowQualFinSF** have weak or negative correlations with **SalePrice**, implying that they have little to no impact on the overall house price, or in the case of **LowQualFinSF**, possibly even a detrimental effect.

Topic 1_Question 1How does the size of the living are(in square feet) influence house prices?

Claim: The size of the house, as measured by the above-ground living area in square feet (GrLivArea), is expected to have a significant positive association with the sale price, reflecting the influence of larger living spaces on overall property value.

Analyze:

The heat map shows that **GrLivArea** is most correlate to the sale price with 0.71 correlation and **GarageArea**, **1stFlrSF**, **TotalBsmtSF** is also worth to be analysis with 0.61-0.62 correlation
Then we fit a linear regression model on the **GrLivArea** and SalePrice and perform a null Hypothesis to analysis the relationship

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.502
Model:	OLS	Adj. R-squared:	0.502
Method:	Least Squares	F-statistic:	1471.
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	4.52e-223
Time:	17:42:09	Log-Likelihood:	-18035.
No. Observations:	1460	AIC:	3.607e+04
Df Residuals:	1458	BIC:	3.608e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.857e+04	4480.755	4.144	0.000	9779.612	2.74e+04
GrLivArea	107.1304	2.794	38.348	0.000	101.650	112.610

Omnibus:	261.166	Durbin-Watson:	2.025
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3432.287
Skew:	0.410	Prob(JB):	0.00
Kurtosis:	10.467	Cond. No.	4.90e+03

Base on the summary:

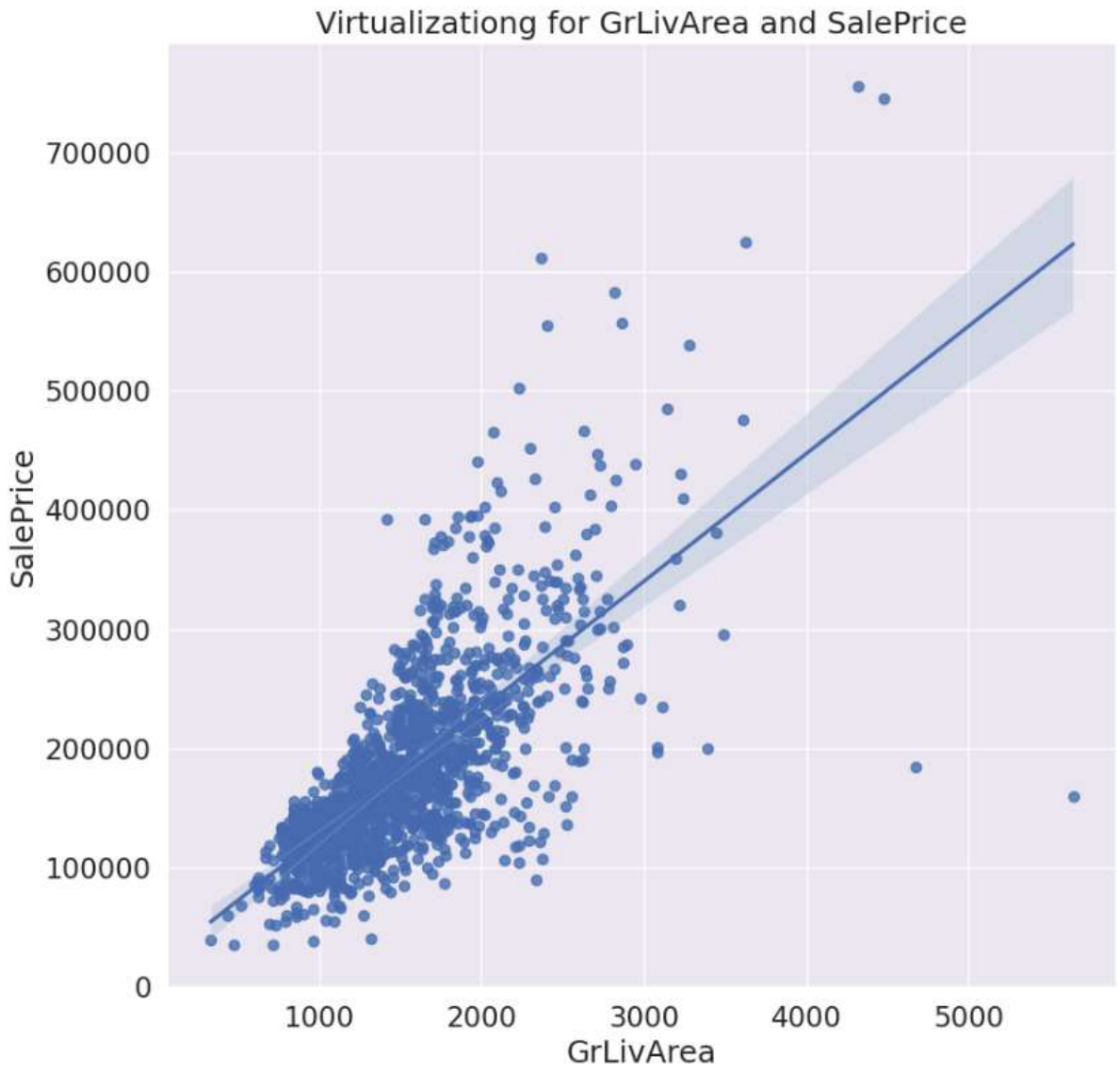
- Correlation for 'GrLivArea': The corelation is approximately 107.130 which means that for every additional square foot increase of living area the sale price will increase 107.130 dollars
- F-statis & P-Value: The F-statistic is 1471 with p-value = 4.52e-223 meaning the model is statically significant
- R^2 : The R Square Value = 0.502 indicates that about 50% Sales Price data is Explain by the model

Then we will do a NullHypothesis to see if the GrLivArea is influencing the House Sales Price

- GrLivArea

  Null Hypothesis(H0): The size of the living area(**GrLivArea**) does not significantly influence the house price (GrLivArea /= 0) Alternative Hypothesis(H1): The size of the living area significantly influence the house price (GrLivArea = 0) The p-value for "**GrLivArea**" is 4.52e-223 < 0.05 indicating that the relationship between GrLivArea and SalePrice is significant hence we reject the Null hypothesis

  Then to further understand the the association between GrLivArea and SalePrice we will virtualize the regression of the **GrLivArea**



After generated the regression graph for **GrLivArea** and SalePrice we can sees that this scatter plot with a regression line shows a positive linear relationship between **GrLivArea** (above-ground living area) and

SalesPrice, indicating that larger living areas tend to correspond with higher house prices. The majority of data points cluster around lower values of both variables, while a few outliers, particularly at the upper right, represent significantly larger and more expensive homes. The shaded area around the regression line reflects the confidence interval, suggesting a reasonably confident fit.
However, since the R^2 is only 0.502 which indicates that the single regression model is inaccurate and lacking holistic hence we should create a multiple regression model for better accuracy.

OLS Regression Results

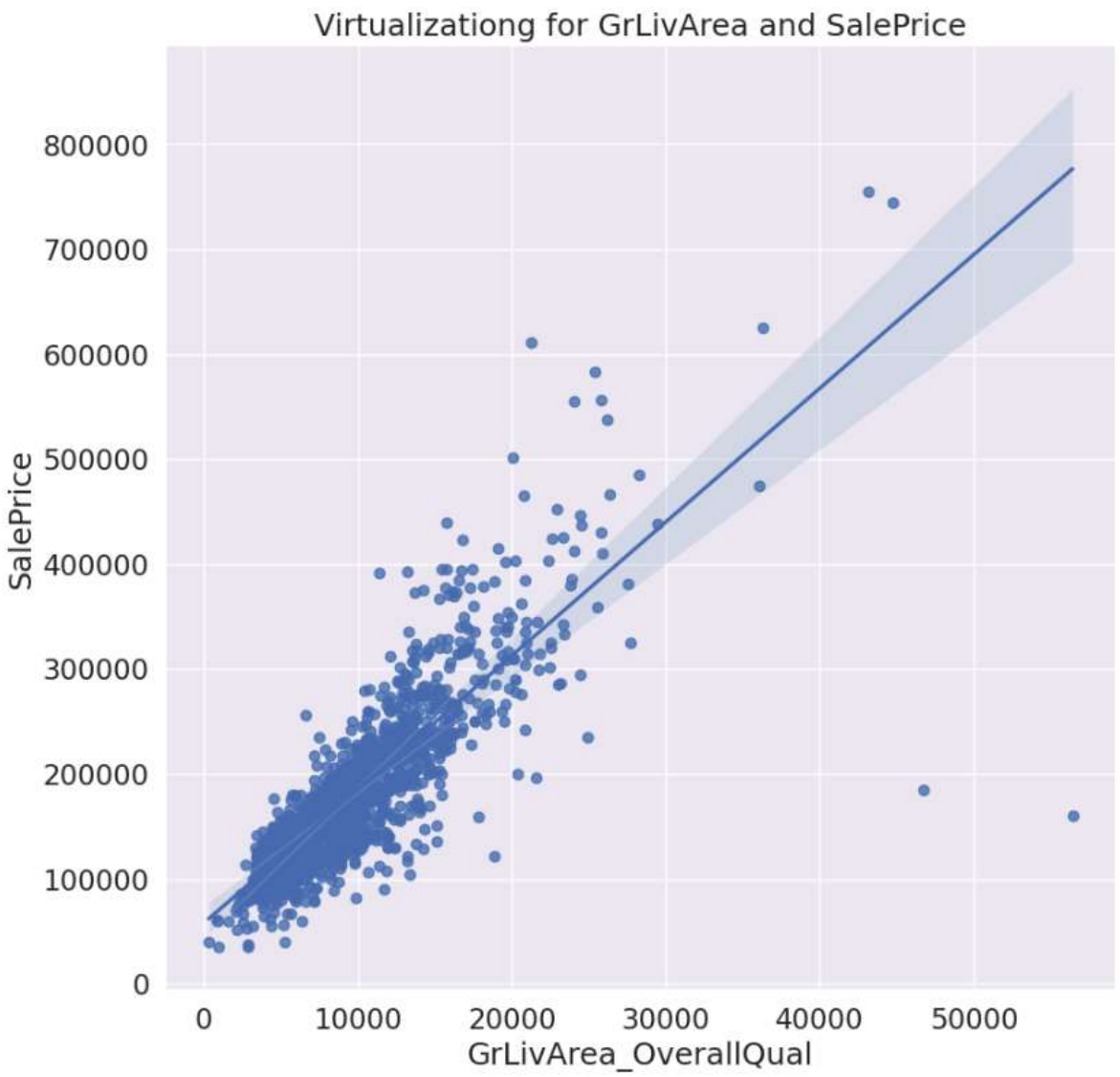
Dep. Variable:	SalePrice	R-squared:	0.771
Model:	OLS	Adj. R-squared:	0.770
Method:	Least Squares	F-statistic:	814.8
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	0.00
Time:	17:42:10	Log-Likelihood:	-17468.
No. Observations:	1460	AIC:	3.495e+04
Df Residuals:	1453	BIC:	3.499e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.032e+04	1.14e+04	-0.904	0.366	-3.27e+04	1.21e+04
GrLivArea	-21.6985	7.697	-2.819	0.005	-36.797	-6.600
OverallQual	1.163e+04	1836.448	6.332	0.000	8025.497	1.52e+04
GarageArea	55.7167	5.966	9.340	0.000	44.015	67.419
TotalBsmtSF	17.0083	4.277	3.977	0.000	8.619	25.397
1stFlrSF	13.9518	4.945	2.821	0.005	4.251	23.652
GrLivArea_OverallQual	9.5764	1.064	8.996	0.000	7.488	11.665

Omnibus:	1176.134	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	210649.081
Skew:	-2.897	Prob(JB):	0.00
Kurtosis:	61.559	Cond. No.	1.30e+05

For this multiple regression model I chose **OverallQual** as the interaction term with **GrLivArea**, according to the Correlation Heatmap the **OverallQual** is having the greatest correlation with **GrLiveArea**.

$R^2 = 0.771$ indicates that about 77.1% of the variability in **SalePrice** is explained by the model. This regression model focuses on the impact of living area size, overall quality, garage area, basement size, and the first floor square footage on **SalePrice**, also considering interactions between living area and overall quality. Intriguingly, the coefficient for **GrLivArea** is negative (-21.6985), which might seem counterintuitive as larger living areas are generally expected to increase property value. However, this could be due to the interaction term **GrLivArea_OverallQual** capturing the nuanced effects more effectively, as this term is significantly positive (9.5764), suggesting that higher quality combined with larger living areas enhances property values significantly.



The graph shows the relationship between the composite variable **GrLivArea_OverallQual** and **SalePrice**. The plot again reveals a positive linear correlation, where an increase in the combined measure of **GrLivArea** and **OverallQual** corresponds with higher **SalePrice**. The data points are more tightly clustered along the regression line compared to the first graph, indicating a stronger linear relationship with fewer deviations. The confidence interval is also narrower, suggesting a higher confidence in the fit of the regression line. When comparing the **GrLivArea_OverallQual** graph with the first, it's evident that incorporating **OverallQual** into the measure of living area results in a tighter and more consistent relationship with **SalePrice**. The first graph, which only considers **GrLivArea**, shows more dispersion and outliers, especially at higher values of living area, indicating that **GrLivArea** alone may not be as strong a predictor of **SalePrice**. By combining **GrLivArea** with **OverallQual**, the second graph provides a clearer and more reliable indication of how these factors together influence house prices, suggesting that quality significantly impacts the price in addition to size.

Topic1_Question 2 What is the effect of the number of bedrooms and bathrooms on the final sale price?

Claim: The number of bathrooms is expected to significantly increase the sale price and contribute to greater consistency in property valuations, as additional bathrooms enhance the functionality and desirability of a home.

Analyze:

To find out the effect of the bedrooms number we first pull out the features we gonna use and fit a single regression model

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.316
Model:	OLS	Adj. R-squared:	0.315
Method:	Least Squares	F-statistic:	336.2
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	8.61e-121
Time:	17:42:10	Log-Likelihood:	-18267.
No. Observations:	1460	AIC:	3.654e+04
Df Residuals:	1457	BIC:	3.656e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.244e+04	6921.659	9.021	0.000	4.89e+04	7.6e+04
FullBath	8.299e+04	3353.978	24.743	0.000	7.64e+04	8.96e+04
BedroomAbvGr	-3976.8756	2265.028	-1.756	0.079	-8419.939	466.188

Omnibus:	578.158	Durbin-Watson:	1.986
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3391.792
Skew:	1.746	Prob(JB):	0.00
Kurtosis:	9.600	Cond. No.	14.7

- Coefficient for **'FullBath'**: The coefficient for Full Bath is about 8.299e+04 indicating that with each unit of full bathroom increases in the housing the sale price will increase about 8.299e+04 dollars
- Coefficient for **'BedroomAbvGr'**: The coefficient is approximately -3976.8756 indicating that each additional bedroom above ground will increase the house sale price by -3976.8756 dollars
- P-value: the p-values for both 'FullBath' are less than 0.05, indicating that the relationship between these features and 'SalePrice' is significant. But the p -value for **'BedroomAbvGr'** is 0.079 > 0.05 which indicate that the relationship between **'BedroomAbvGr'** and 'SalePrice' is not significant

- F-statistic; The F-static is 336.2 with p-value = 8.61e-121 , indicating that the model is significant
- R_Square: The R^2 value is 0.316 indicating that only about 32% of the Sales_Price data is explained by the model

Then we will do a Null Hypothesis test to see if the feature is influencing the Sales_Pirce

- full bathrooms:

  Null Hypothesis(H0): The number of full bathrooms do not significantly influence the house price
  Alternative Hypothesis(H1): The number of full bathrooms and the number of bedrooms above ground does significantly influence house price

  Given that the 'FullBath' are significantly less than 0.05, hence we reject the null hypothesis. so the 'FullBath' does significantly influence the house price

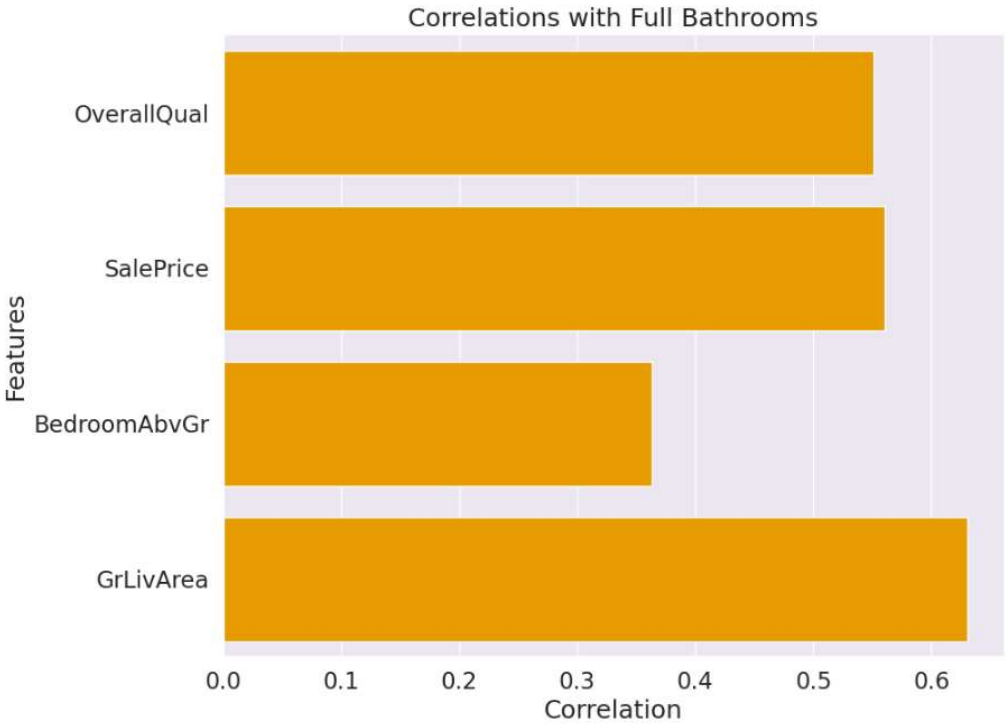
- bedrooms above ground level:

  Null Hypothesis(H0): The number of bedrooms do not significantly influence the house price

  Alternative Hypothesis(H1): The number of full bedroom above ground does significantly influence house price

  Given that the **'BedroomAbvGr'** p-value is 0.079 > 0.05, hence we fail to reject
  the null hypothesis. so the Bedroom does not significantly influence the house price. which corresponding to the heatmap

Since R^2 indicate that only 32% of the Sales_Price data is explained by the model so we need to include more variable and interaction term to increase the model accuracy
By checking the previous correlation heat map.I plan to add overall quality and groundlive are as relevent variable but before that we should check if they are having a great correlation



The analysis of the provided correlation bar plots reveals distinct relationships between house features and the number of full bathrooms and bedrooms above grade.

OverallQual shows strong correlations with **FullBath**, suggesting higher quality homes often feature more bathrooms. The correlation between **SalesPrice** and full bathrooms is notably stronger than with

bedrooms, indicating that bathrooms may have a greater impact on property values. Both **FullBath** and **BedroomAbvGr** are logically interlinked, displaying strong mutual correlations which reflect practical housing designs where more bedrooms necessitate more bathrooms. Lastly, **GrLivArea** exhibits the strongest correlation with bedrooms, underscoring that larger homes typically contain more bedrooms. This analysis highlights key aspects that real estate stakeholders might consider, especially when evaluating property values and planning constructions or renovations.

after the correlation analysis the **BedrromAbvFGr** seems won't be providing strong background to predict the sales price hence we drop **BedrromAbvFGr** from the model

OLS Regression Results

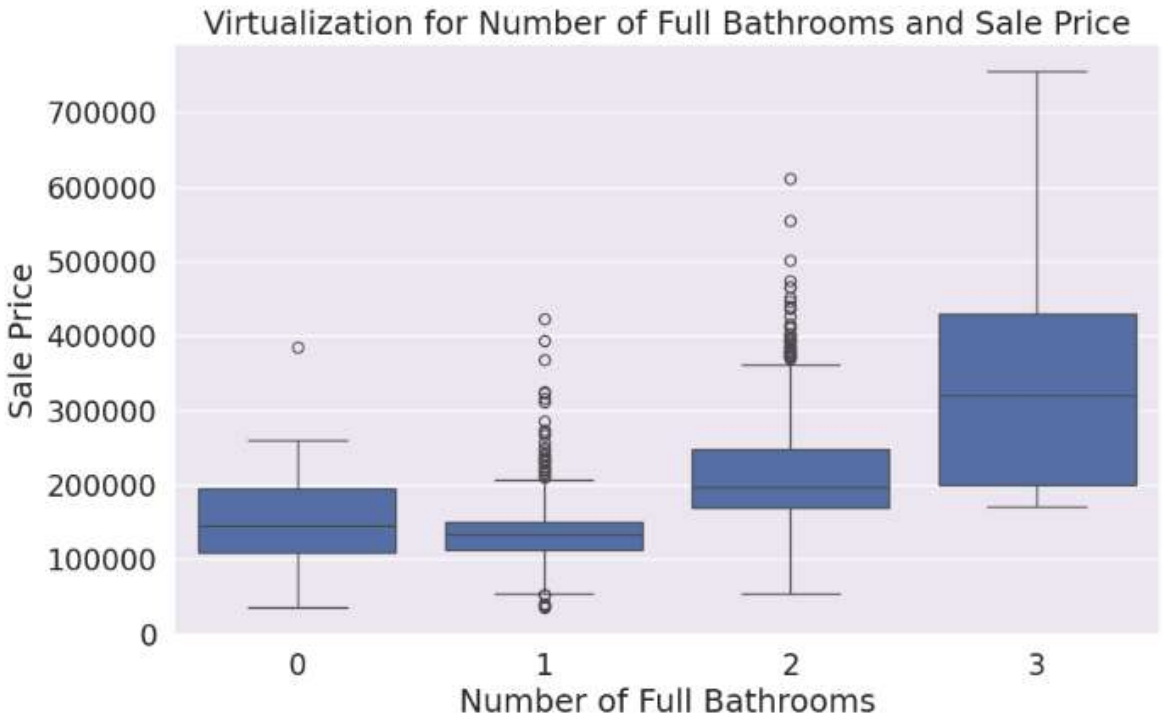
Dep. Variable:	SalePrice	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.741
Method:	Least Squares	F-statistic:	697.2
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	0.00
Time:	21:58:33	Log-Likelihood:	-17554.
No. Observations:	1460	AIC:	3.512e+04
Df Residuals:	1453	BIC:	3.516e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.497e+04	1.71e+04	1.459	0.145	-8605.574	5.86e+04
FullBath	-5.063e+04	1.73e+04	-2.923	0.004	-8.46e+04	-1.66e+04
GrLivArea	46.4736	8.698	5.343	0.000	29.411	63.536
OverallQual	1.099e+04	3219.462	3.414	0.001	4674.702	1.73e+04
FullBath_OverallQuality_LivArea	2.2251	0.764	2.912	0.004	0.726	3.724
FullBath_GrLivArea	-13.8214	8.266	-1.672	0.095	-30.036	2.393
FullBath_OverallQual	9567.2810	2513.135	3.807	0.000	4637.521	1.45e+04

Omnibus:	473.827	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23305.139
Skew:	-0.725	Prob(JB):	0.00
Kurtosis:	22.519	Cond. No.	4.99e+05

R^2 = 0.742 indicating that about 74% of the variables in sales price is explained by the model. This regression model results focusing on the impact of "**FullBath**" on "**SalePrice**" and considering interactions with "**GrLivArea**" and "OverallQual," I found some intriguing insights. The coefficient for "**FullBath**" itself is negative (-5.063e+04), which was initially surprising as more bathrooms typically add value. However, this could be due to the interactions between "**FullBath**" and other variables capturing the positive impact more effectively. Specifically, the interaction term "**FullBath_OverallQual**" is positive and significant, indicating that higher quality homes with more bathrooms are valued more. Moreover, "**FullBath_GrLivArea**," which represents the interaction between the number of bathrooms and the living area, is also significant and positive, suggesting that larger homes with more bathrooms hold higher value. Due to the lack of significant correlation with "**SalePrice**" and to streamline the model, I decided to drop the "**BedroomAbvGr**" feature from this analysis. This allowed for a clearer focus on how bathroom-related features interact with quality and size to influence house prices.

Then we Virtualize the Feature and Sale Price for better understanding



The box plot reveals that the median sale price increases with the number of full bathrooms in a house, indicating a positive association between the number of full bathrooms and sale price. Homes with three full bathrooms show the highest median sale price and the widest range of prices,But also there are some outlier indicating that some homes in the groups sold for higher or lower prices.

Topic1_Question 3 Does the additional(eg.fireplaces,pools) significantly increase the house price?

Claim: Having a fireplace/Pool is expected to contribute to more consistent sale prices, as it adds a desirable feature that enhances the overall appeal and perceived value of the home.

Analyze:

According to the correlation heat map, the 'fireplace' is having a higher correlation(0.47) with the saleprice and 'poolarea' is only have 0.09 coorelation to the sale price. Then To find out if fireplaces or pools significantly increase the house price we first need to pull out the fireplace and pools data

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.220
Model:	OLS	Adj. R-squared:	0.219
Method:	Least Squares	F-statistic:	205.9
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	1.78e-79
Time:	17:42:12	Log-Likelihood:	-18362.
No. Observations:	1460	AIC:	3.673e+04
Df Residuals:	1457	BIC:	3.675e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.457e+05	2535.755	57.471	0.000	1.41e+05	1.51e+05
Fireplaces	5.697e+04	2863.585	19.895	0.000	5.14e+04	6.26e+04
PoolArea	95.7979	45.948	2.085	0.037	5.667	185.929

Omnibus:	531.551	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2733.518
Skew:	1.630	Prob(JB):	0.00
Kurtosis:	8.857	Cond. No.	77.1

- Coefficient for 'Fireplaces ': The coefficient for Fireplaces is about 5.697e+04 indicating that with each unit of 5.697e+04 increases in the housing the sale price will increase about 5.697e+04 dollars
- Coefficient for 'PoolArea': The coefficient is approximately PoolArea indicating that each additional square feets of the pool will increase the house sale price by 95.7979 dollars
- P-value: the p-values fo both Fireplaces and PoolArea are less than 0.05, indicating that the relationship between these features and **'SalePrice'** is significant.
- F-statistic; The F-static is 205.9 with p-value = 1.78e-79 , indicating that the model is significant
- R_Square: The R^2 is 0.22 indicating that only about 22% of the Sale_Price Data is explain by the model so we will need to perform multivariate analysis for further analysis

Then we will do a Null Hypothesis test to see if the feature is influencing the Sales_Pirce

- Number of Fireplaces :

Null Hypothesis(H0): The number of Fireplaces do not significantly influence the house price

Alternative Hypothesis(H1): The number of Fireplaces Does significantly influence house price

Given that the P-value of 'Fireplaces' are 0.00 < 0.05, hence we reject the null hypothesis. so the 'Fireplaces ' does significantly influence the house price

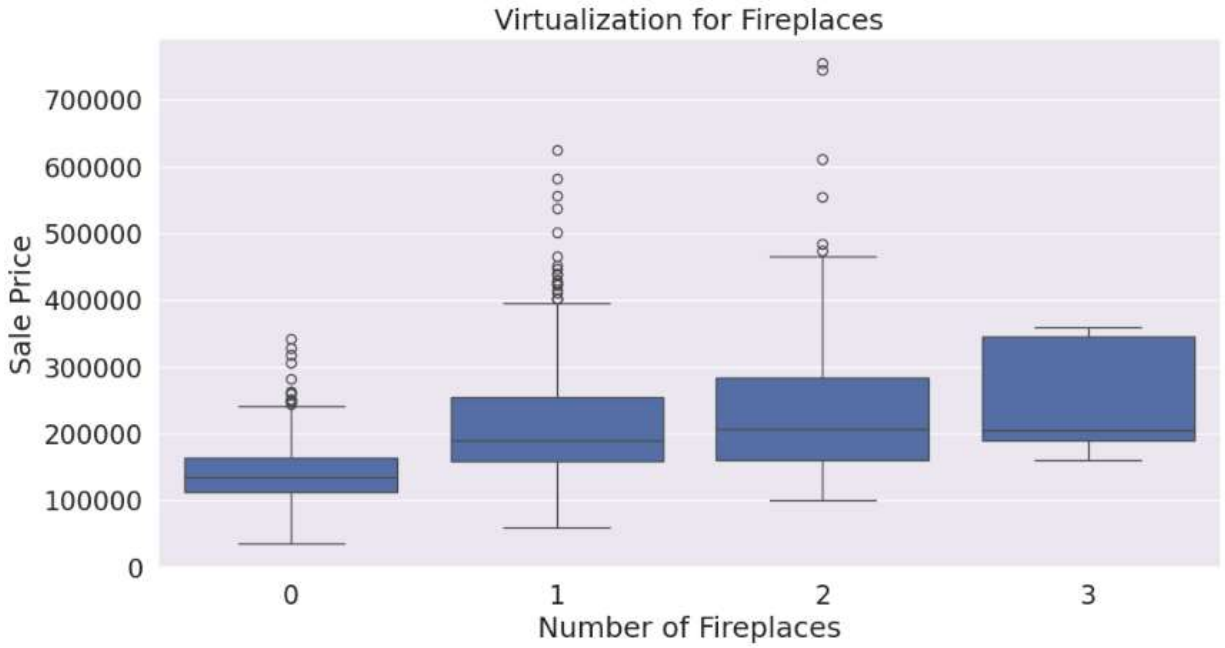
- PoolArea :

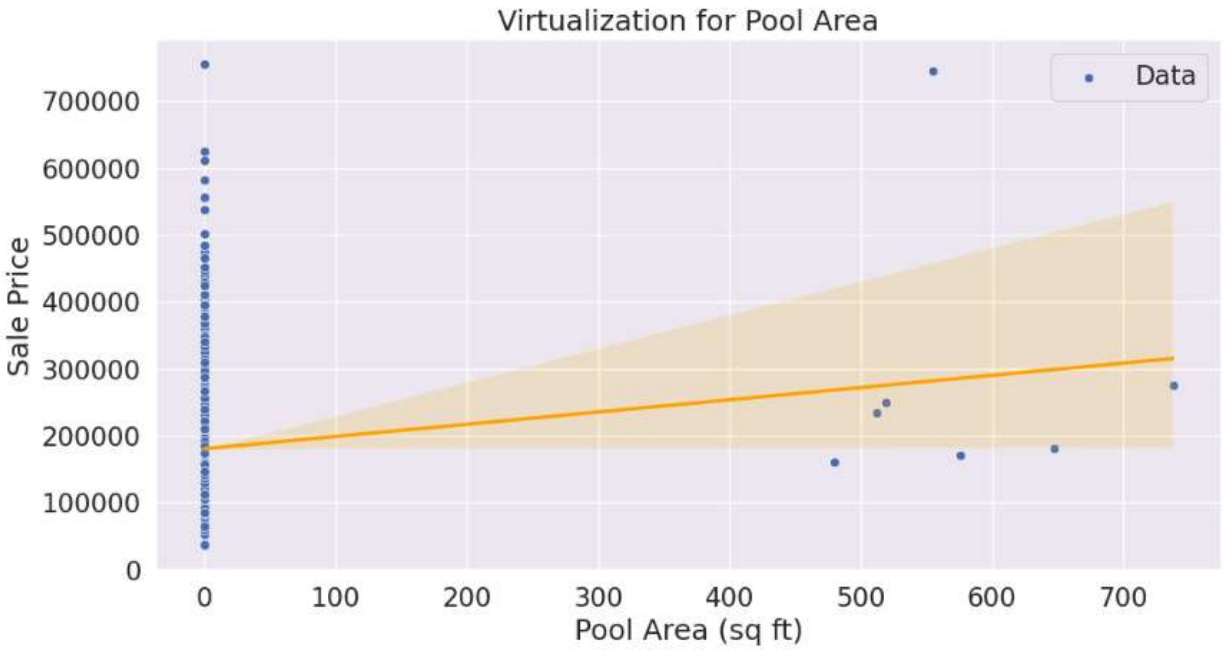
Null Hypothesis(H0): The Pool Size (in square feets) do not significantly influence the house price

Alternative Hypothesis(H1): The Pool Size (in square feets) does significantly influence house price

Given that the **'PoolArea'** p-value is 0.037 < 0.05, hence we reject the null hypothesis. so the **'PoolArea'** does significantly influence the house price

How ever by checking the R^2 we can sees that only about 22% of the data is explain by the model, which is too low so we need to adjust the model to see if we can increase the accuracy
But before that , in order to further understand the relationship between Fireplaces and Sales Price we will need to create box plot, since the data from FirePlace couldn't support linear graph and other graph type





- Fireplaces

The box plot shows a clear positive relationship between the number of fireplaces and sale price. Houses with more fireplaces tend to have higher median sale prices. The minimum sale price also tends to increase as the number of fireplaces in a house increases. The lower bound sale price for homes with no fireplaces is around 100 thousand dollars, whereas for houses with three fireplaces, the lower bound is around 200 thousand dollars. This indicates that both the central tendency and the lower quartile sale prices are higher in homes with more fireplaces.

- PoolArea

The scatter plot with a regression line shows the relationship between pool area and sale price. The plot indicates a positive correlation between pool area and sale price, suggesting that larger pool areas are generally associated with higher sale prices. However, most houses have a pool area of zero, with a few having larger pool areas ranging up to approximately 600 square feet. The regression line, while positively sloped, indicates that the impact of pool area on sale price is relatively moderate, as reflected by the gentle incline. Additionally, the wide confidence interval at higher pool areas suggests greater variability and less certainty in the relationship for homes with larger pools, which may be due to the lack of data for homes with larger pool areas.

After understanding the the feature we will now perform the multiple regression to get a more accuracy model and gain a deeper understanding between the feature and Sales_Price
in order to do this we need to find the feature that can be use in the multi regression model according to the correlation heat map the average correlation of other features to PollArea is only about 0.15 which is low correlated also the correlation between pool area and Sale Price is only about 0.09 hence I decided to

drop the pool area due to low correlation and null hypothesis instead I will do a multi regression model on fireplace.

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.793
Model:	OLS	Adj. R-squared:	0.792
Method:	Least Squares	F-statistic:	696.4
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	0.00
Time:	23:19:33	Log-Likelihood:	-17393.
No. Observations:	1460	AIC:	3.480e+04
Df Residuals:	1451	BIC:	3.485e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.9e+04	6189.157	-7.918	0.000	-6.11e+04	-3.69e+04
Fireplaces	-1.203e+05	8519.456	-14.117	0.000	-1.37e+05	-1.04e+05
GrLivArea	35.2369	3.588	9.821	0.000	28.199	42.275
OverallQual	2.037e+04	1287.938	15.816	0.000	1.78e+04	2.29e+04
TotalBsmtSF	39.2633	3.986	9.850	0.000	31.444	47.083
GrLivArea_Fireplaces	28.5894	3.801	7.522	0.000	21.133	36.045
OverallQual_Fireplaces	1.197e+04	1385.932	8.637	0.000	9252.076	1.47e+04
TotalBsmtSF_Fireplaces	22.4093	4.905	4.569	0.000	12.788	32.031
Freatures_Fireplaces	-0.0012	8.97e-05	-13.223	0.000	-0.001	-0.001

Omnibus:	351.583	Durbin-Watson:	1.969
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5185.450
Skew:	0.700	Prob(JB):	0.00
Kurtosis:	12.126	Cond. No.	3.36e+08

After analyzing the summary I found compelling patterns. The coefficient for **Fireplaces** is significant and negative, which initially appears counterintuitive since fireplaces are often seen as a desirable feature. However, the interaction terms provide a clearer picture: **GrLivArea_Fireplaces**, **OverallQual_Fireplaces**, and **TotalBsmtSF_Fireplaces** all show significant and positive coefficients. This indicates that the value added by fireplaces is more pronounced in larger homes, those with higher overall quality, and homes with larger basement areas. The complex interaction term **Features_Fireplaces** is nearly zero, suggesting that the multiplicative effect of all these features combined with fireplaces does not significantly impact the price beyond their individual interactions.

The model's reliability, with an R-squared of 0.793, suggests that it explains a significant portion of the variability in **SalePrice**, reflecting a strong model fit.

Given the limited correlation of **PoolArea** with **SalePrice** and to simplify our analysis while focusing on more impactful variables, I decided to drop the **PoolArea** feature from this model. This approach helps to concentrate on features that significantly affect housing prices and ensures the model remains robust and interpretable.

5.Topic 2 Influence of non-physical conditions to the house prices?

To analysis this topic we will first pull out all the features we will use as well

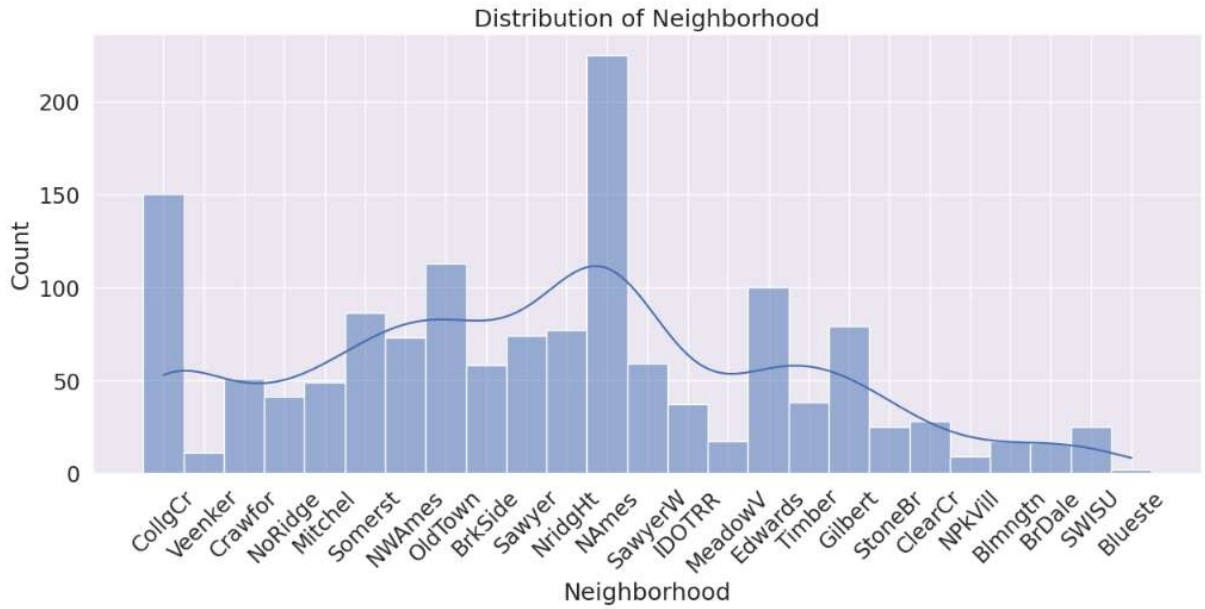
	Neighborhood	YearBuilt	YearRemodAdd	MoSold	YrSold	SalePrice
0	CollgCr	2003	2003	2	2008	208500
1	Veenker	1976	1976	5	2007	181500
2	CollgCr	2001	2002	9	2008	223500
3	Crawfor	1915	1970	2	2006	140000
4	NoRidge	2000	2000	12	2008	250000
...	...	...	...	...	...	...
1455	Gilbert	1999	2000	8	2007	175000
1456	NWAmes	1978	1988	2	2010	210000
1457	Crawfor	1941	2006	5	2010	266500
1458	NAmes	1950	1996	4	2010	142125
1459	Edwards	1965	1965	6	2008	147500

1460 rows × 6 columns

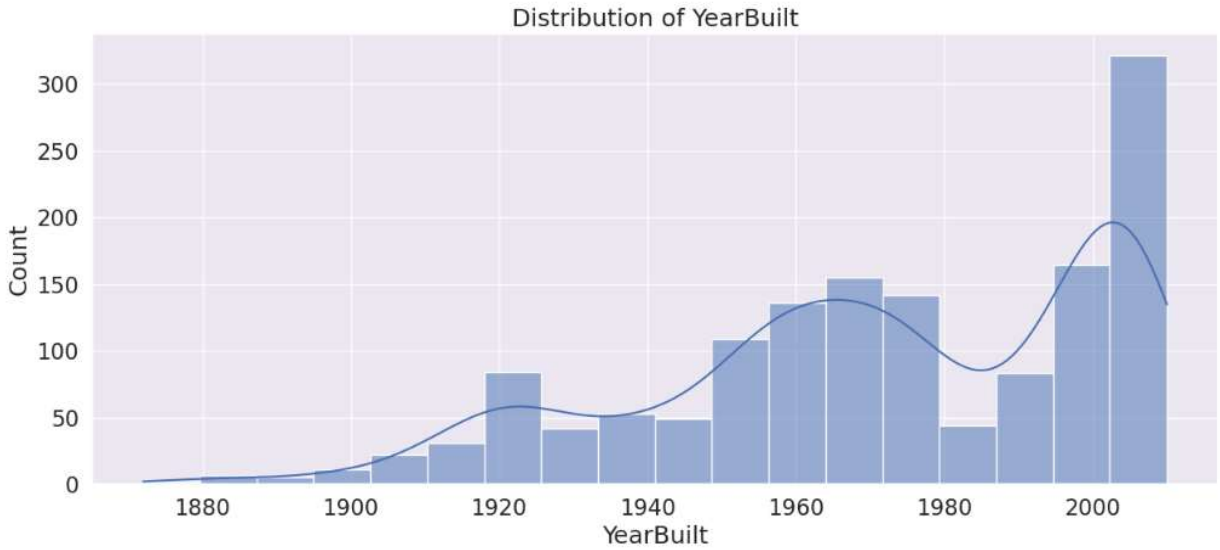
This dataset subset includes 1460 rows and six columns: **Neighborhood**, **YearBuilt**, **YearRemodAdd**, **MoSold**, **YrSold**, and **SalePrice**. It highlights key temporal factors such as the year of construction and the most recent renovation, which likely impact the house prices. The **Neighborhood** column captures the location, an important determinant of **SalePrice**, while the **MoSold** and **YrSold** columns may reveal potential seasonal trends in the real estate market.

Distribution Analyze

Then we plot out the distribution for each of the feature and analysis it in that way we can have a better understanding on the features

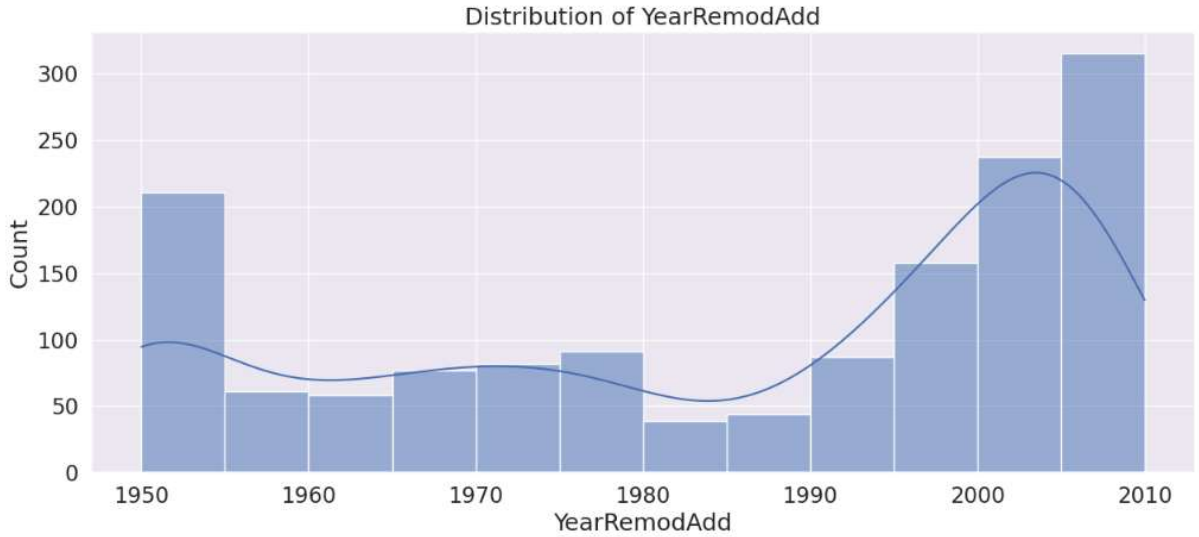


The graph displays the distribution of properties across various neighborhoods in Ames, Iowa. Notably, the 'NAmes' neighborhood has the highest number of properties, indicated by a significant peak, suggesting it's a popular area. Other neighborhoods like 'CollgCr' and 'OldTown' also show relatively higher counts, implying these areas are also well-populated or preferred. The distribution shows a long tail towards the right, with some neighborhoods like 'Blueste' having very few properties, which could indicate newer, less developed, or less popular areas. The smooth line overlay suggests a general trend, indicating that a few neighborhoods have the majority of houses, while most have fewer properties, highlighting possible preferences or economic disparities among the areas.

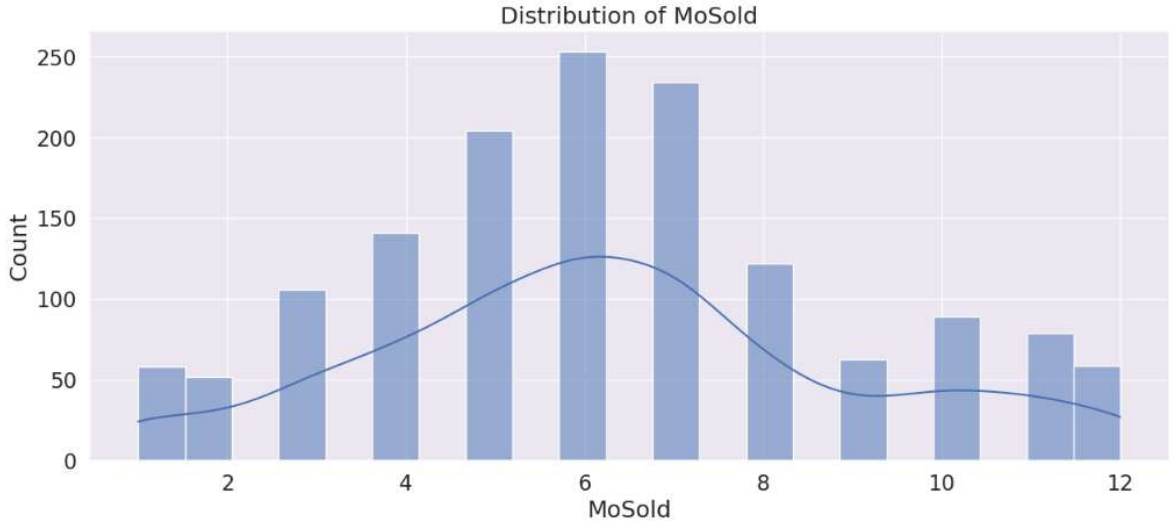


In the graph distribution of **YearBuilt**, we observe a timeline that highlights significant construction trends in Ames, Iowa. Starting from the early 1900s, there is a noticeable peak around the 1920s, which suggests a period of robust building activity. This is followed by a decline in construction during the Great Depression and World War II. A subsequent increase in the mid-20th century, particularly around the 1960s, likely reflects economic recovery and population growth. The early 2000s show another peak,

indicating a modern resurgence in housing development, possibly driven by new housing demands or urban expansion. This pattern provides valuable insights into the economic and demographic shifts influencing housing construction over the past century.

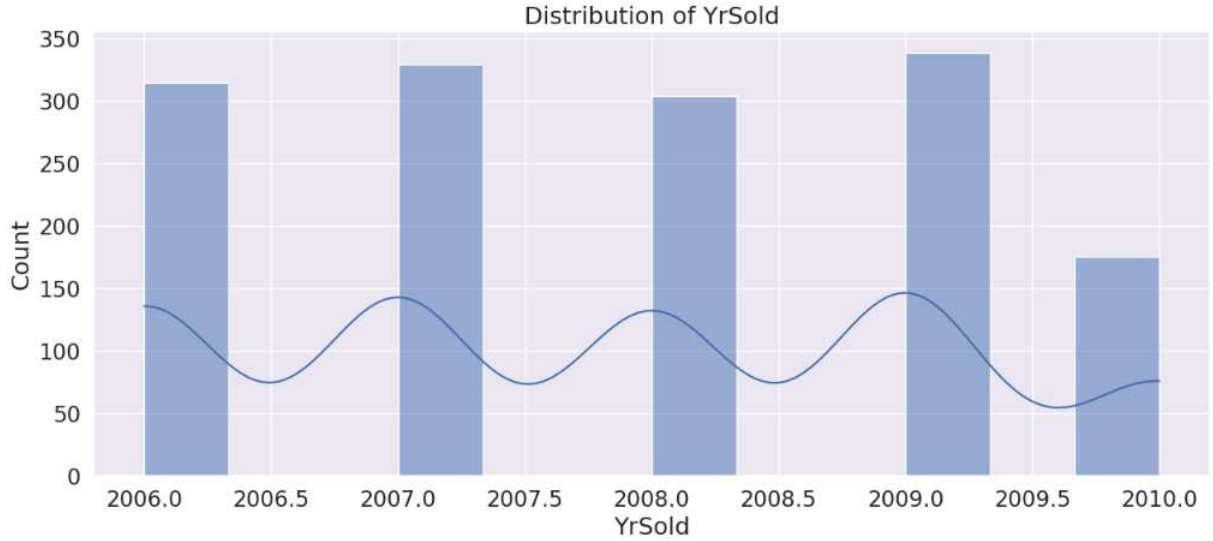


In the graph distribution of YearRemodAdd, we see the frequency of house remodels or additions by year in Ames, Iowa. The data starts with a substantial number of remodels in the early 1950s, indicating a post-war renovation wave. It then shows a general decline in remodeling activity through the 1960s and 1970s, perhaps reflecting economic downturns or shifts in housing policy. The late 1990s and early 2000s witness a dramatic increase in remodeling activities, peaking just before 2010. This resurgence could be related to economic recovery, increased home equity, or changes in homeowner preferences towards modernizing older properties.



In the graph distribution of **MoSold**, which represents the month in which houses were sold, we observe a clear seasonal pattern in Ames, Iowa. Sales activity increases starting in early spring, with the peak months for home sales occurring in June and July. This suggests that summer is the most popular time for moving, likely due to favorable weather conditions and the convenience of relocating families while school is out. Sales then decline in the late summer and fall, reaching a low in the winter months, which

could be due to the less appealing weather for moving and the start of the school year, making families less likely to relocate.



In the graph distribution of **YrSold**, which represents the years in which houses were sold in Ames, Iowa, we see fluctuations in the annual count of sales from 2006 to 2010. Sales peaked in 2007 and again in 2009, suggesting these were years of high activity in the housing market. The dip in 2008 corresponds with the global financial crisis, which likely impacted local real estate markets, causing fewer transactions. The rebound in 2009 may indicate a recovery or an adjustment in market conditions leading to an increase in sales. The drop in 2010 might suggest a stabilization of the market or a return to more typical levels of sales activity following the post-crisis recovery phase.

In the graph distribution of **SalePrice**, the data exhibits a right-skewed distribution, with most homes selling at prices around 150,000 to 200,000, as indicated by the peak of the histogram. The long tail extending towards higher values suggests that a smaller number of homes sell at much higher prices, up to $700,000. This distribution is typical in real estate markets where a majority of transactions involve moderately priced homes, while luxury properties, which are fewer, appear as outliers. The shape of the

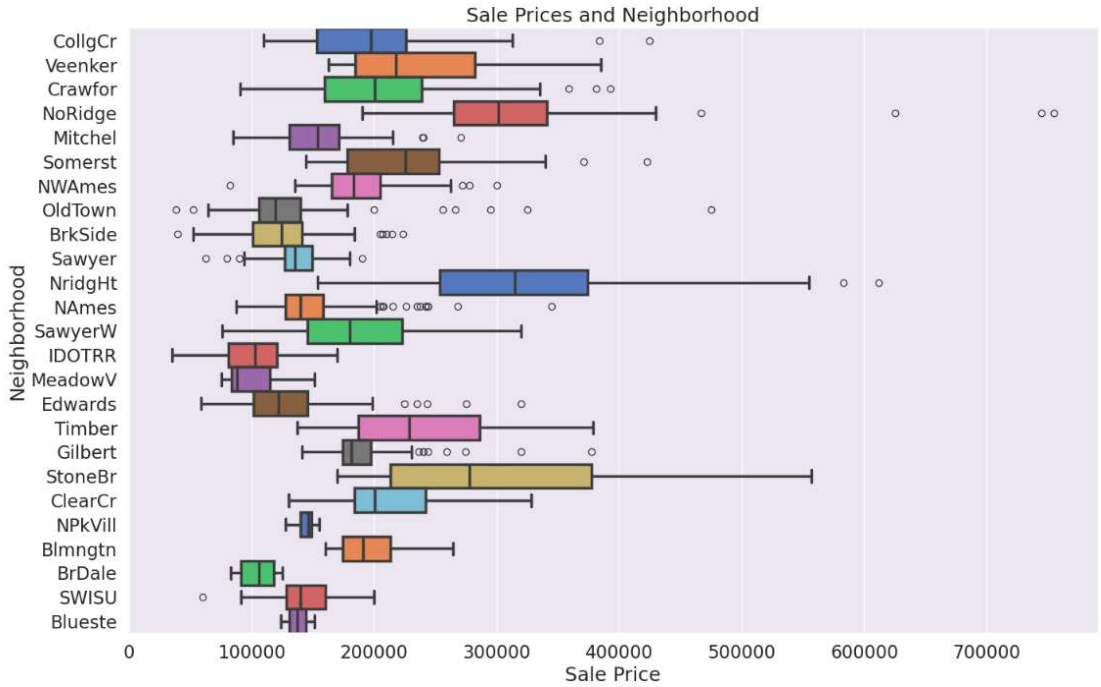
distribution, with its peak and long tail, also hints at the presence of a few highly valued properties which could significantly differ from the median housing stock in terms of features and amenities.

Topic 2_Question 1 How do different neighborhoods compare in terms of average house prices?

Claim: The type of neighborhood is expected to cause significant differences in house sale prices, as location influences factors such as desirability, accessibility, and overall market demand.

Analyze:

To find out the answer for this question I used box_plot but with color, which can let us identify the difference between each neighborhood easily



The box plot shows significant differences in median sale prices across various neighborhoods. Neighborhoods like 'NoRidge', 'StoneBr', and 'NridgHt' have higher median prices ranging from 300 thousand dollars to 400 thousand dollars, compared to other neighborhoods. In contrast, neighborhoods like 'IDOTRR' and 'MeadowV' have much lower median prices, around 100 thousand dollars. The spread of the data within each neighborhood indicates varying levels of price variability. For example, 'Sawyer' and 'StoneBr' have a wide range of sale prices, suggesting diverse housing options and unstable housing prices, while 'Blueste' and 'NPkVill' show more consistent and lower price ranges. Outliers are present in

several neighborhoods, notably in 'NridgHt' and 'NoRidge', with some outliers being significantly above the median sale price.

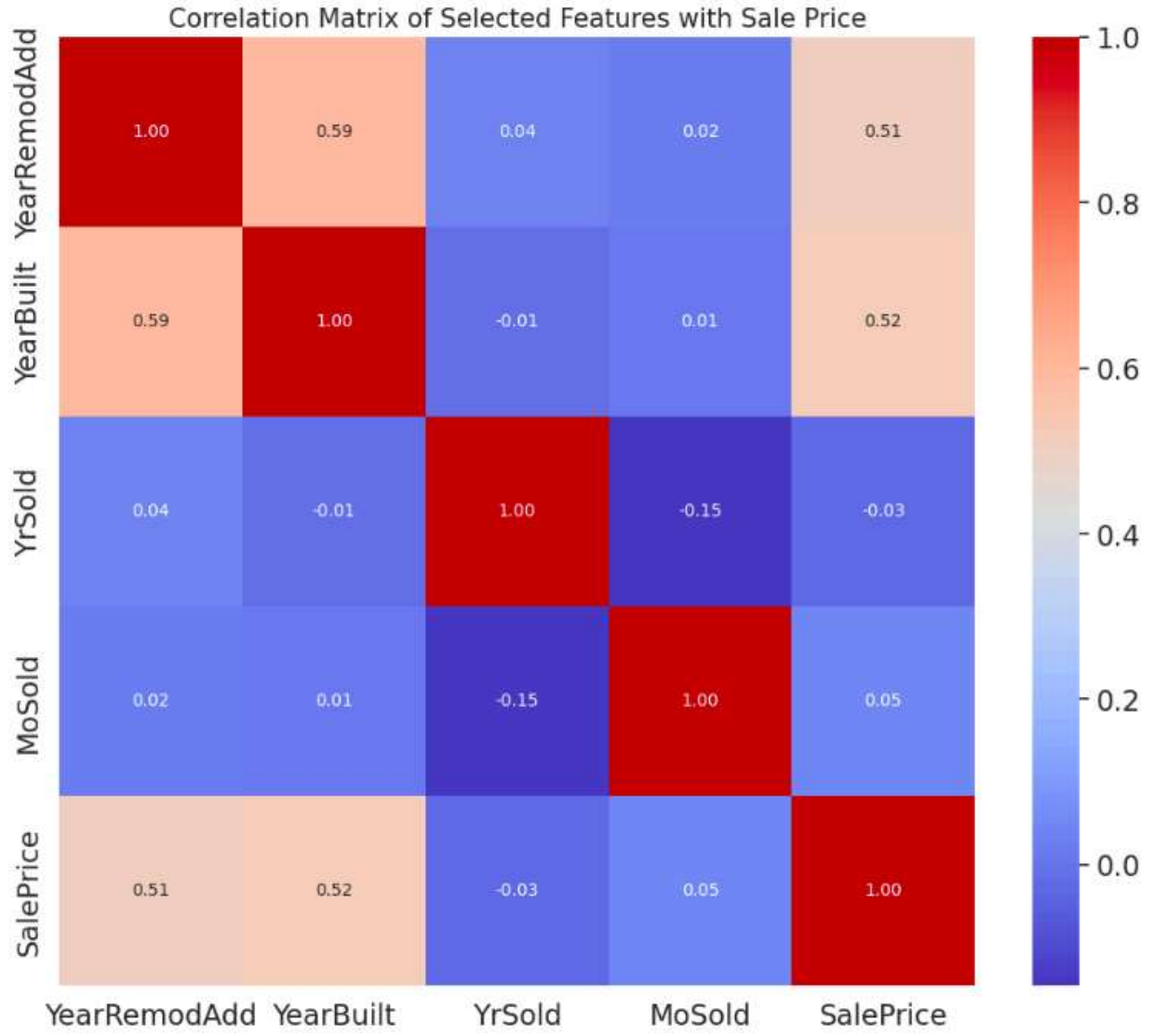
Overall, the analysis suggests that the neighborhood does influence the sale price. However, due to complex external inducing causes, we may need additional datasets to conduct deeper research into the factors affecting sale prices more comprehensively.

Topic 2_Question 2 What's the best time to sold the house

Claim: During specific times of the year, people are more likely to purchase a house, reflecting seasonal trends in the real estate market.

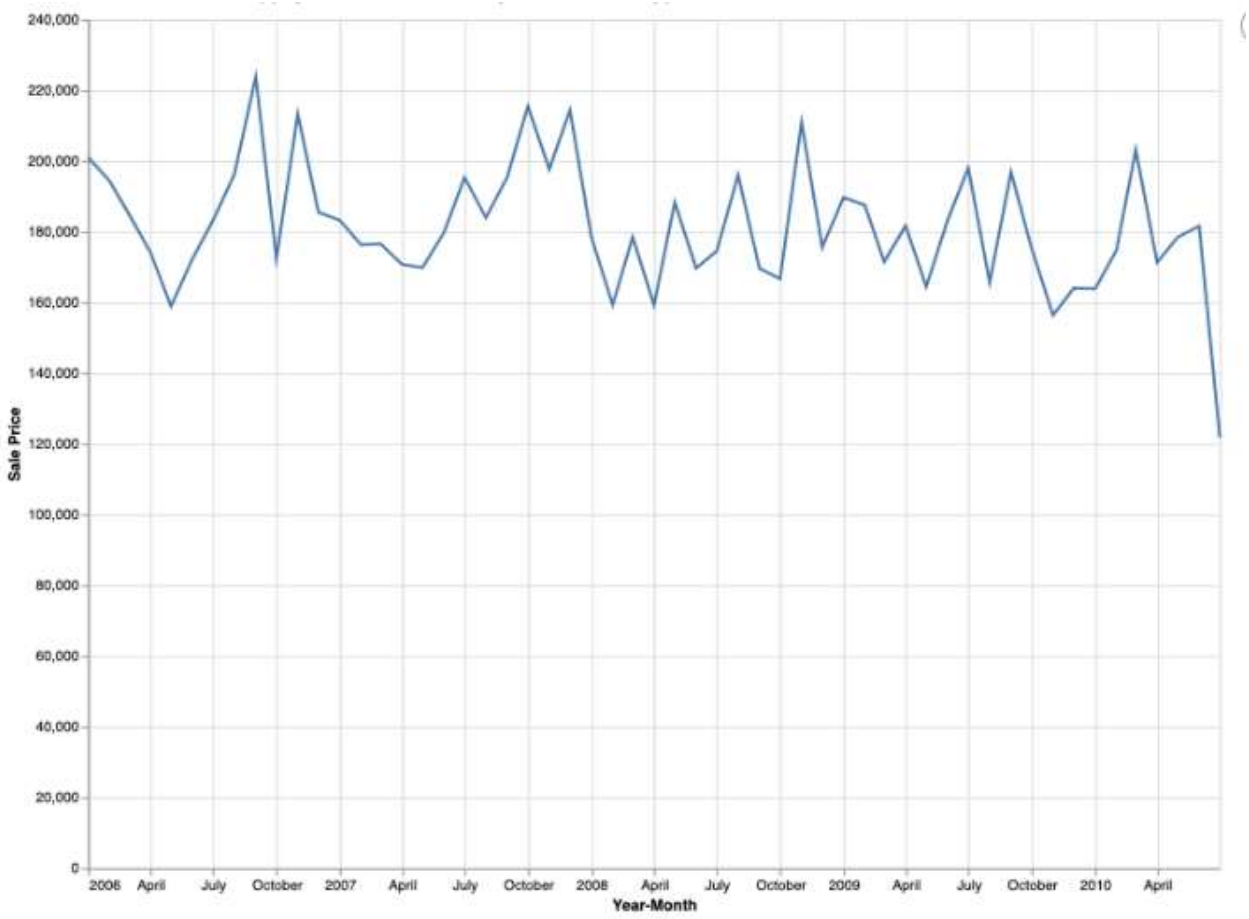
Analyze:

To find out the answer for this question we first need to create a heatmap to show the correlation



After analyzing this correlation heat map, I notice that both **YearBuilt** and **YearRemodAdd** have moderate positive correlations with **SalePrice**, indicating that newer homes and recently remodeled ones tend to have higher sale prices. Interestingly, **MoSold** and **YrSold** show very weak correlations with **SalePrice**, suggesting that the month and year of sale do not significantly impact house prices in this dataset. There is also a moderate correlation between **YearBuilt** and **YearRemodAdd**, which makes sense since newer homes are less likely to have undergone significant remodeling.

Then I Created a Time series line graph for virtualization



The line graph shows the average sale price of houses from 2006 to 2010. The data reveals seasonal fluctuations, with notable peaks and troughs in sale prices each year. There is a recurring trend where average sale prices tend to be higher in the latter half of the year, especially after July, suggesting that this period might be more favorable for purchasing houses. In contrast, the first half of the year, particularly the months leading up to July, generally sees lower average sale prices. This pattern indicates that people are more likely to purchase their houses after July, with high demand potentially being the reason causing the higher market prices. The sharp decline in average sale prices towards the end of the observed period in 2010 might suggest market volatility or external economic factors affecting the housing market. To determine the exact reasons for this decline, additional data and historical context would be necessary. Overall, this analysis highlights July and the subsequent months as the optimal time for selling houses to achieve higher sale prices.

Topic 3_Question 3 What is the impact of the year the house was built and the year it was remodeled on house prices?

Claim: The more recently a house has been remodeled or built, the higher its market price is likely to be, reflecting buyer preference for newer or updated properties.

Analyze:

Inorder to find out the impact of the year to the house we can fit a single regression model

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.333
Model:	OLS	Adj. R-squared:	0.332
Method:	Least Squares	F-statistic:	364.2
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	5.27e-129
Time:	17:42:18	Log-Likelihood:	-18248.
No. Observations:	1460	AIC:	3.650e+04
Df Residuals:	1457	BIC:	3.652e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.917e+06	1.64e+05	-23.840	0.000	-4.24e+06	-3.59e+06
YearRemodAdd	1169.4666	102.210	11.442	0.000	968.972	1369.962
YearBuilt	901.4469	69.867	12.902	0.000	764.397	1038.497

Omnibus:	770.140	Durbin-Watson:	1.955
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7817.918
Skew:	2.248	Prob(JB):	0.00
Kurtosis:	13.407	Cond. No.	2.71e+05

- Correlation for 'YearRemodAdd': The data indicates that for each year closer to the present that a house is remodeled, the sale price increases by approximately $1,169.47. This suggests that more recent renovations significantly increase the market value of a house.
- Correlation for 'YearBuilt': The data indicates that for each year closer to the present that a house is built, the sale price increases by approximately $901.45. This suggests that the newer the house, the higher the market price will be.
- F-statis: The F-statistic is 364.2 with p-value = 5.27e-129 meaning the model is statically significant
- R_Square: The R^2 is 0.333 indicating that only about 33% of the Sale_Price Data is explain by the model so we will need to perform multivariate analysis for further analysis

Then we do a Null-Hypothesis test too see if the feature is influence the Sales_Price

- YearRemodAdd:

Null Hypothesis(H0): The YearRemod of the house do not significantly influence the house price
Alternative Hypothesis(H1): The YearRemod of the house will significantly infulence house price

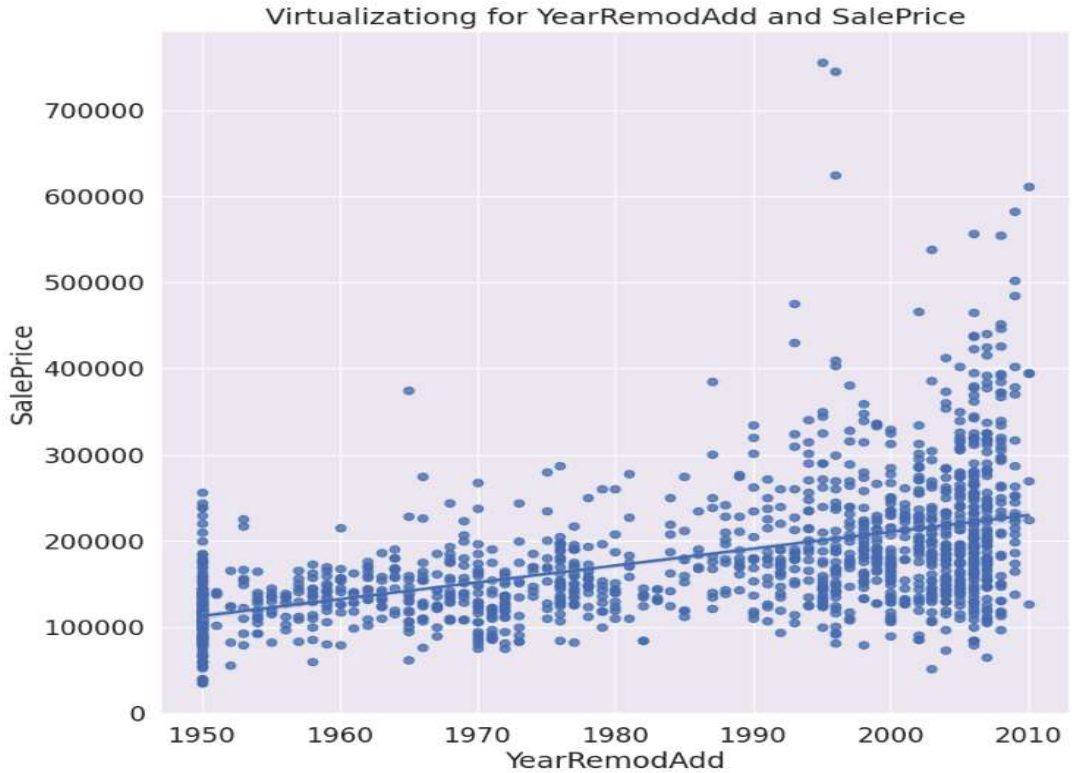
Given that the **'YearRemodAdd'** are significantly less thatn 0.00 < 0.05, hence we reject the null hypothesis. so the **'YearRemodAdd'** will significantly influience the house price

- YearBuild

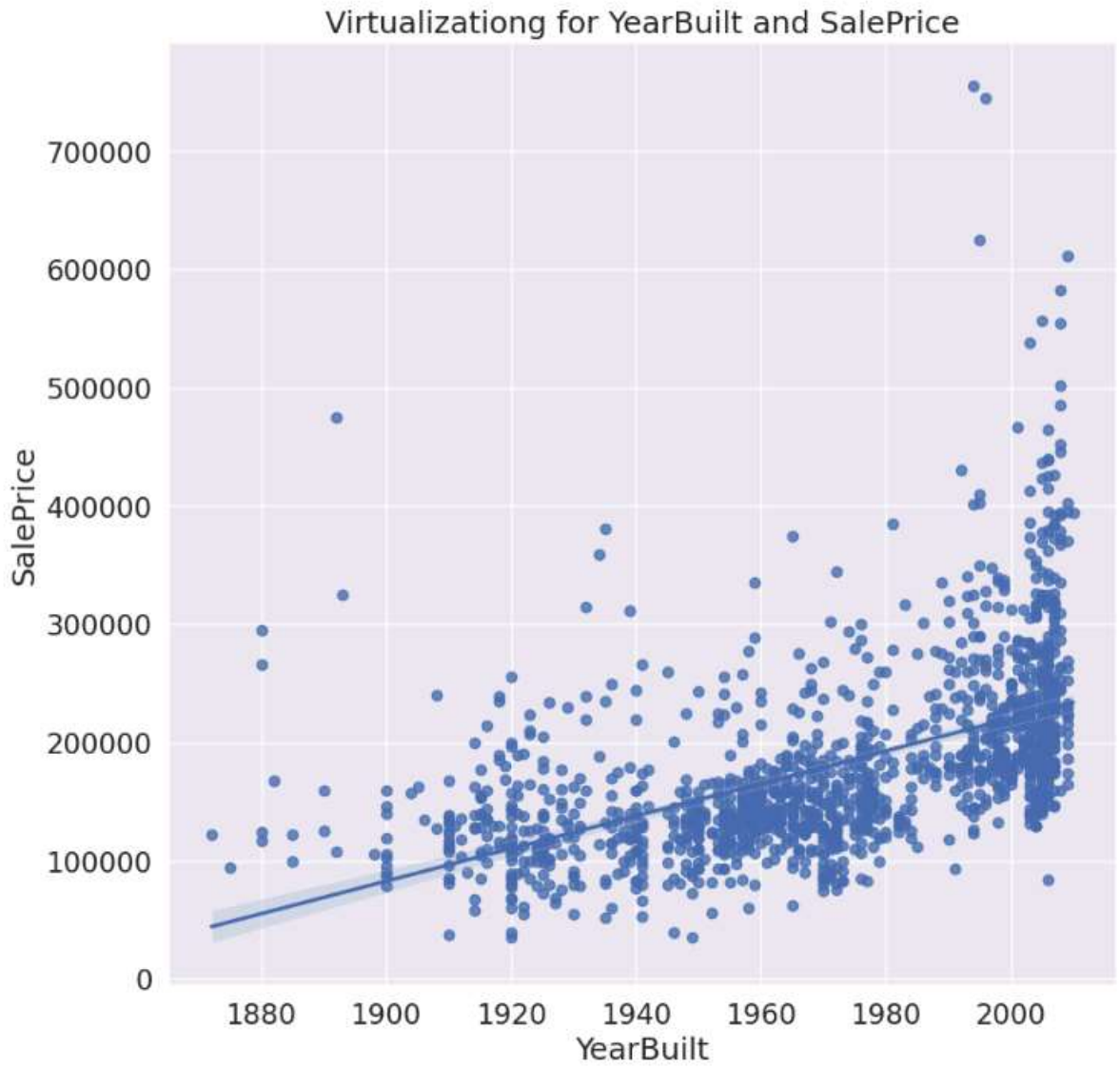
Null Hypotheis(H0): The year closer to the presnt that a house built do not significantly influence the house price
Altenative Hypothesis(H1): The year closer to the presnt that a house built will significantly influence the house price

Given that the 'YearBuild' p-value is 0.00 < 0.05, hence we reject
the null hypothesis. so the 'YearBuild' will significantly influence the house price. which corresponding to the heatmap

However as we pervious knows the R^2 for the single regression model is only explaining 33% of the data in Sales_Price hence we will need to perform a multilinear regression model for better accuracy. But before that we also need to perform the virtualization to indicate the performance of the feature



The plot shows a positive trend between **YearRemodle** and **SalePrice**, indicating that houses remodeled more recently tend to have higher sale prices. Houses remodeled around 2000 and later often achieve sale prices above 300 thousands dollars with some outliner exceeding to 700 thousands dollars. In contrast, houses remodeled in the earlier decades, such as the 1950s and 1960s, generally have lower and more consistent sale prices, mostly below 200 thousands dollars. with some outlier mostly exceeding to 250 thousands dollars which is about 65% less than house that remodeled in 2000 The increasing density of data points and the upward slope of the trend line reinforce the correlation, suggesting that more recent renovations significantly increase the market value of a house.



The plot shows a clear positive trend, indicating that newer houses tend to have higher sale prices. Houses built after 2000 frequently achieve sale prices above 250 thousands to 400 thousands, with some even exceeding 600 thousands and 700 thousands. In contrast, houses built before 1940 typically have lower sale prices, often below 200 thousands. This positive correlation suggests that for each decade closer to

the present, there is an approximate 20-30% increase in sale price, reflecting the higher market value of newer constructions.

Lastly, let's perform a multilinear regression analysis where I include **YearRemodAdd**, **YearBuilt**, **YrSold**, and **MoSold** as predictors, along with interaction terms between these variables.

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.339
Model:	OLS	Adj. R-squared:	0.335
Method:	Least Squares	F-statistic:	92.84
Date:	Mon, 26 Aug 2024	Prob (F-statistic):	1.44e-124
Time:	05:16:59	Log-Likelihood:	-18242.
No. Observations:	1460	AIC:	3.650e+04
Df Residuals:	1451	BIC:	3.655e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.287e+08	2.49e+08	1.319	0.187	-1.6e+08	8.17e+08
YearRemodAdd	5.775e+04	1.64e+05	0.352	0.725	-2.64e+05	3.79e+05
YearBuilt	-2.229e+05	1.14e+05	-1.960	0.050	-4.46e+05	186.967
YrSold	-1.654e+05	1.24e+05	-1.334	0.183	-4.09e+05	7.79e+04
MoSold	-7.698e+04	6.06e+04	-1.271	0.204	-1.96e+05	4.18e+04
YearRemodAdd_YrSold	-28.2953	81.586	-0.347	0.729	-188.335	131.745
YearRemodAdd_MoSold	40.9635	40.375	1.015	0.310	-38.236	120.163
YearBuilt_YrSold	111.4661	56.632	1.968	0.049	0.377	222.555
YearBuilt_MoSold	-1.6950	27.825	-0.061	0.951	-56.277	52.887

Omnibus:	783.122	Durbin-Watson:	1.961
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8251.921
Skew:	2.283	Prob(JB):	0.00
Kurtosis:	13.714	Cond. No.	8.25e+11

In comparing the two regression summaries, the earlier model, which includes only **YearRemodAdd** and **YearBuilt**, has an R-squared value of 0.333, indicating that 33.3% of the variance in **SalePrice** is explained by the model. The coefficients for both **YearRemodAdd** and **YearBuilt** are significant, suggesting a strong relationship with **SalePrice**. The expanded model, which introduces **YrSold**, **MoSold**, and their interaction terms, shows a slightly higher R-squared value of 0.339, but this improvement is marginal, and many of the interaction terms are not statistically significant. This suggests that the interactions between the timing of the sale and the year of remodeling or construction may not have a strong influence on sale prices within this dataset, possibly due to limited relevant interaction terms. As the R-squared value remains relatively low, it indicates that a large portion of the variance in **SalePrice** is still unexplained, and incorporating external data or additional variables might be necessary to improve the model's explanatory power.