Data Science

Assignment 1 – Problem 1 – House Prices

Farimah Rashidi – 99222040

1. Import library and dataset

At the first we should import dataset and libraries. I did this part in Kaggle.

2. Data Exploration

The house price dataset provides comprehensive information on residential properties, including key features such as square footage, number of bedrooms and bathrooms, location, and sale prices. This dataset is a valuable resource for real estate market analysis, helping to understand housing market trends and factors influencing property values.

Data fields:

SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.

MSSubClass: The building class

MSZoning: The general zoning classification

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access

Alley: Type of alley access

LotShape: General shape of property

LandContour: Flatness of the property

Utilities: Type of utilities available

LotConfig: Lot configuration

LandSlope: Slope of property

Neighborhood: Physical locations within Ames city limits

Condition1: Proximity to main road or railroad

Condition2: Proximity to main road or railroad (if a second is present)

BldgType: Type of dwelling

HouseStyle: Style of dwelling

OverallQual: Overall material and finish quality

OverallCond: Overall condition rating

YearBuilt: Original construction date

YearRemodAdd: Remodel date

RoofStyle: Type of roof

RoofMatl: Roof material

Exterior1st: Exterior covering on house

Exterior2nd: Exterior covering on house (if more than one material)

MasVnrType: Masonry veneer type

MasVnrArea: Masonry veneer area in square feet

ExterQual: Exterior material quality

ExterCond: Present condition of the material on the exterior

Foundation: Type of foundation

BsmtQual: Height of the basement

BsmtCond: General condition of the basement

BsmtExposure: Walkout or garden level basement walls

BsmtFinType1: Quality of basement finished area

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Quality of second finished area (if present)

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

HeatingQC: Heating quality and condition

CentralAir: Central air conditioning

Electrical: Electrical system

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Number of bedrooms above basement level

Kitchen: Number of kitchens

KitchenQual: Kitchen quality

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality rating

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

GarageType: Garage location

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

GarageCond: Garage condition

PavedDrive: Paved driveway

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Fence: Fence quality

MiscFeature: Miscellaneous feature not covered in other categories

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold

YrSold: Year Sold

SaleType: Type of sale

SaleCondition: Condition of sale

3. Data Preprocessing

Data Shape: I checked the shape of the dataset using df.shape to understand the number of rows and columns. This helps in getting an overview of the dataset's size which is (1460, 81)

Checking for Missing Data: I used df.isnull().sum() to check for missing data to find the number of missing values in each column.

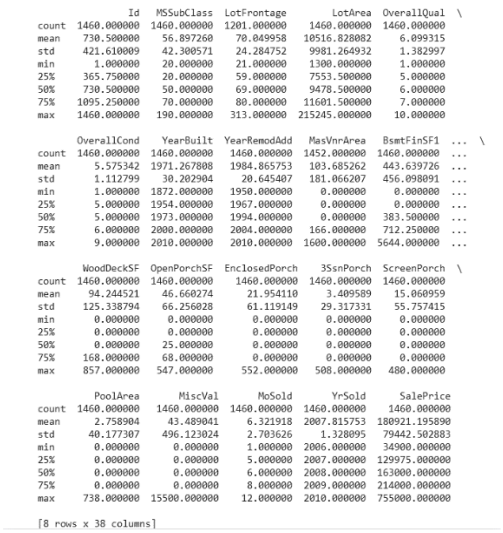
Data Types: Using df.dtypes, I verified the data types of every column. It is necessary to comprehend data kinds to manipulate and analyze data appropriately.

Resul:

```
Id                int64
MSSubClass        int64
MSZoning         object
LotFrontage     float64
LotArea           int64
                  ...
MoSold            int64
YrSold            int64
SaleType         object
SaleCondition    object
SalePrice         int64
Length: 81, dtype: object
```

Displaying the First Few Rows: I printed the first few rows of the dataset using print(df.head()) to get a glimpse of the actual data and its structure.

Summary Statistics: I used print(df.describe()) to provide summary statistics for numerical features. This offers important statistical information that might be used for initial data exploration, such as mean, standard deviation, and quartiles.



	Id	MSSubClass	LotFrontage	LotArea	OverallQual	\
count	1460.000000	1460.000000	1201.000000	1460.000000	1460.000000	
mean	730.500000	56.897260	70.049958	10516.828082	6.099315	
std	421.610009	42.300571	24.284752	9981.264932	1.382997	
min	1.000000	20.000000	21.000000	1300.000000	1.000000	
25%	365.750000	20.000000	59.000000	7553.500000	5.000000	
50%	730.500000	50.000000	69.000000	9478.500000	6.000000	
75%	1095.250000	70.000000	80.000000	11601.500000	7.000000	
max	1460.000000	190.000000	313.000000	215245.000000	10.000000	

	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	...	\
count	1460.000000	1460.000000	1460.000000	1452.000000	1460.000000	...	
mean	5.575342	1971.267808	1984.865753	103.685262	443.639726	...	
std	1.112799	30.202904	20.645407	181.066207	456.098091	...	
min	1.000000	1872.000000	1950.000000	0.000000	0.000000	...	
25%	5.000000	1954.000000	1967.000000	0.000000	0.000000	...	
50%	5.000000	1973.000000	1994.000000	0.000000	383.500000	...	
75%	6.000000	2000.000000	2004.000000	166.000000	712.250000	...	
max	9.000000	2010.000000	2010.000000	1600.000000	5644.000000	...	

	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch	\
count	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	
mean	94.244521	46.660274	21.954110	3.409589	15.060959	
std	125.338794	66.256028	61.119149	29.317331	55.757415	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	25.000000	0.000000	0.000000	0.000000	
75%	168.000000	68.000000	0.000000	0.000000	0.000000	
max	857.000000	547.000000	552.000000	508.000000	480.000000	

	PoolArea	MiscVal	MoSold	YrSold	SalePrice
count	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000
mean	2.758904	43.489041	6.321918	2007.815753	180921.195890
std	40.177307	496.123024	2.703626	1.328095	79442.502883
min	0.000000	0.000000	1.000000	2006.000000	34900.000000
25%	0.000000	0.000000	5.000000	2007.000000	129975.000000
50%	0.000000	0.000000	6.000000	2008.000000	163000.000000
75%	0.000000	0.000000	8.000000	2009.000000	214000.000000
max	738.000000	15500.000000	12.000000	2010.000000	755000.000000

[8 rows x 38 columns]

Column Names: I printed the column names using print(df.columns) to have a list of all the features in the dataset.

Data Information: I got a summary of the dataset using df.info(), which included the number of non-null values and the types of data.

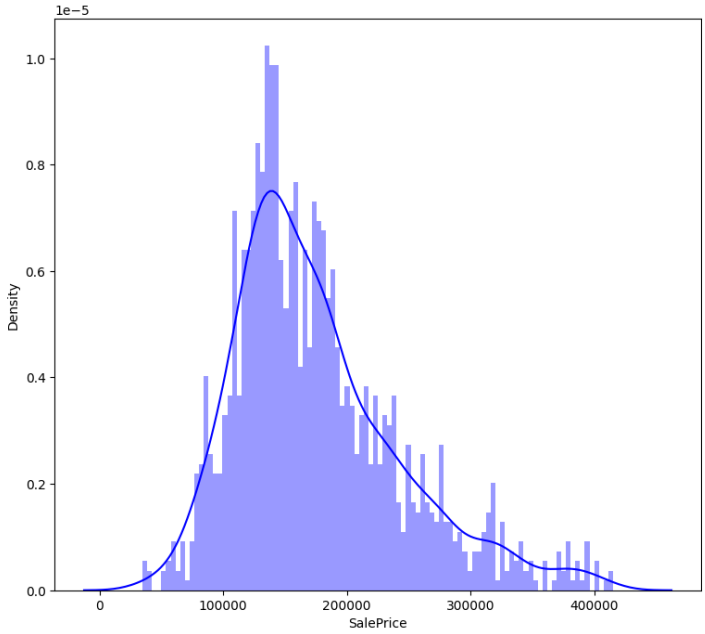
Checking for Missing Data (Revisited): I calculated missing data again and stored it in the missing_data variable. This step is a repetition of checking for missing values to verify if any were missed.

Feature Engineering: I created a new feature called 'TotalSF,' which is the overall square footage of the property, to show off feature engineering.

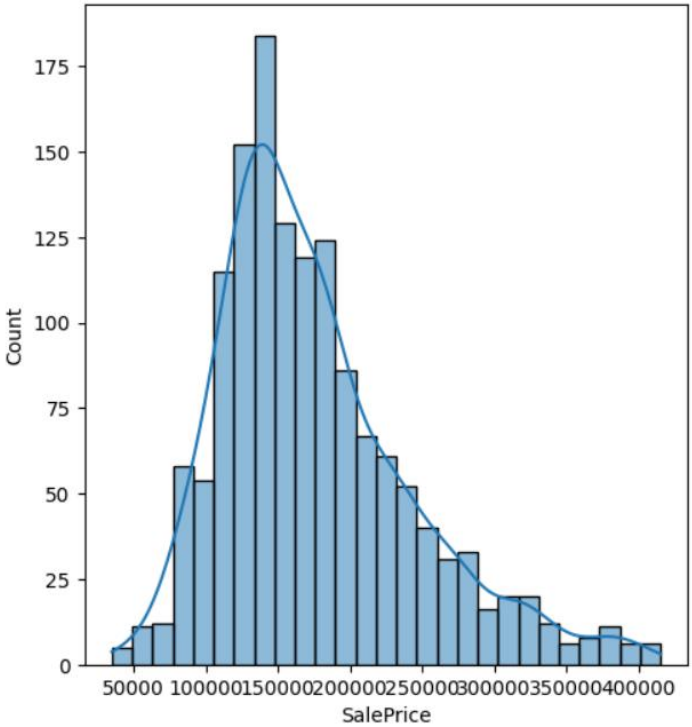
Outlier Handling: I gave an illustration of how to handle outliers with the z-score. This stage is crucial for locating and maybe eliminating extreme data points that might have an impact on the study.

4. Exploratory Visualization

Now let us look at how the house prices are distributed.

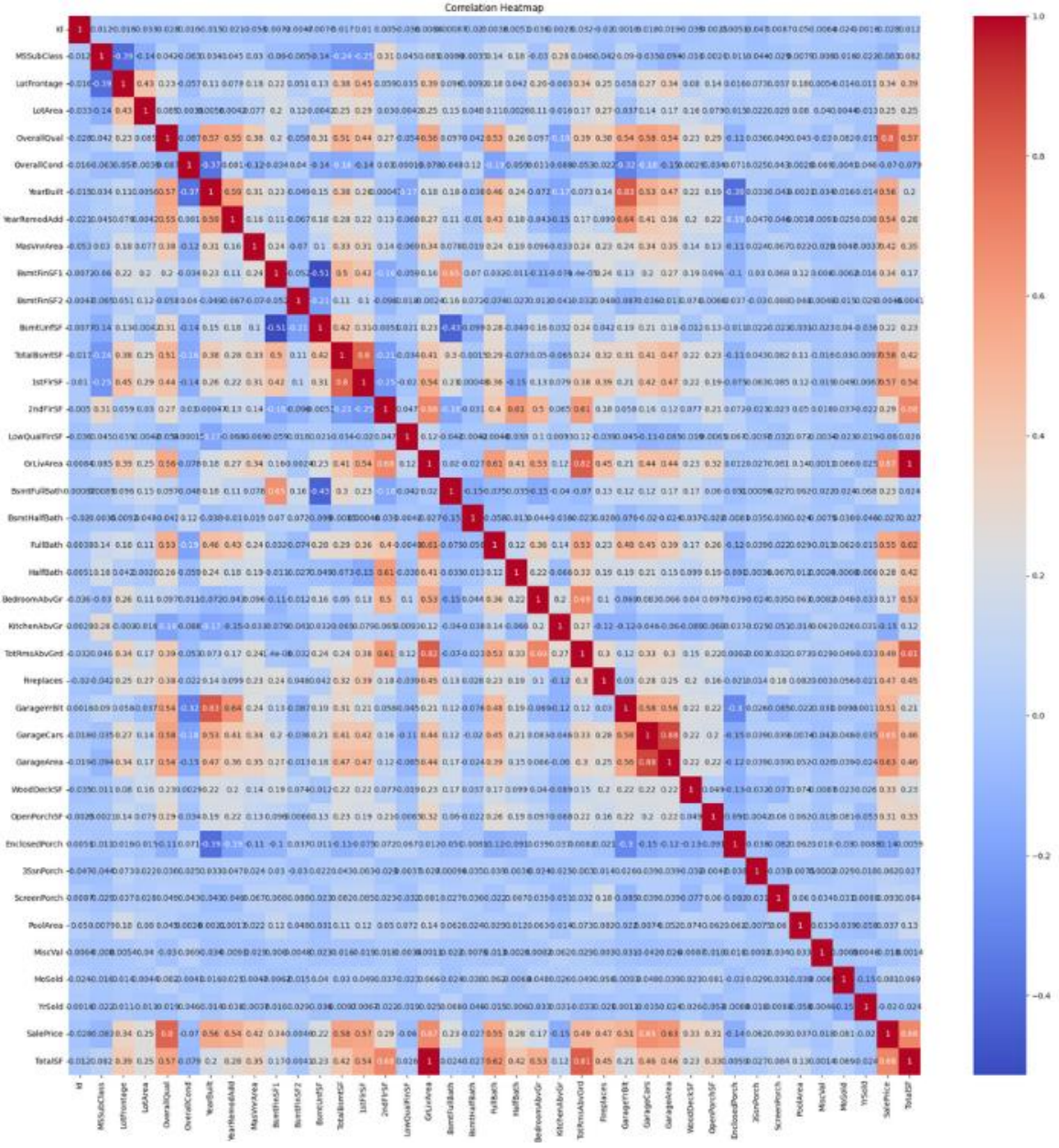


Also, this plot shows the distribution of sale prices in the dataset and provides an estimate of the probability density of sale prices at different values:

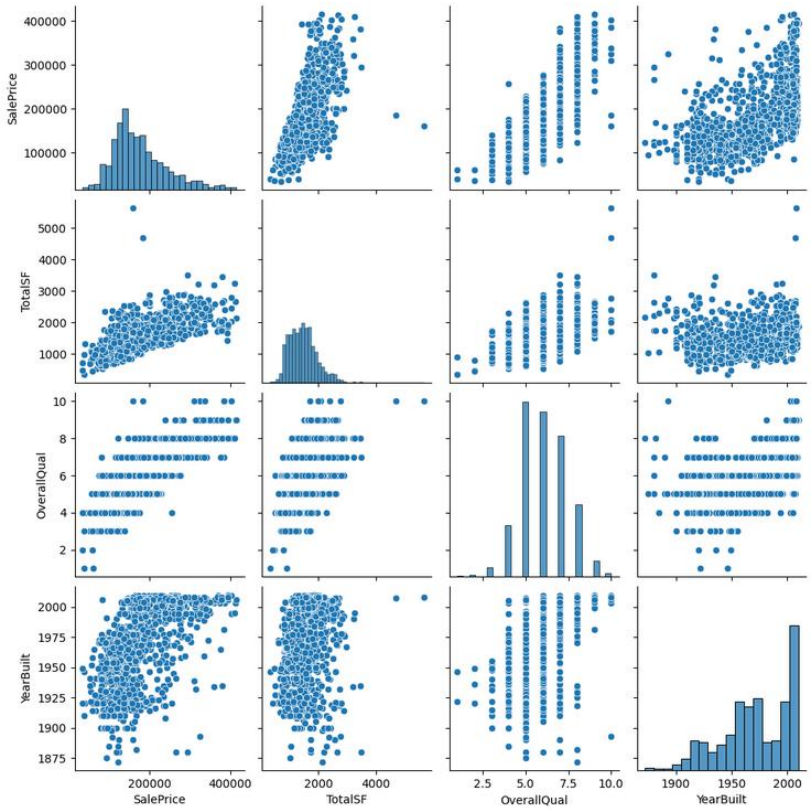


Correlation Heatmap:

The correlations between the numerical variables in our dataset are shown visually by the heatmap. The correlation between two variables is shown by each cell in the heatmap. Each cell's color represents the direction and strength of the correlation.



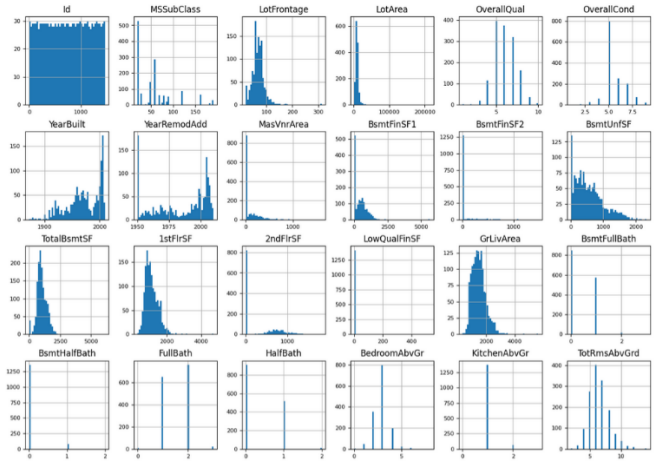
Pair plots:

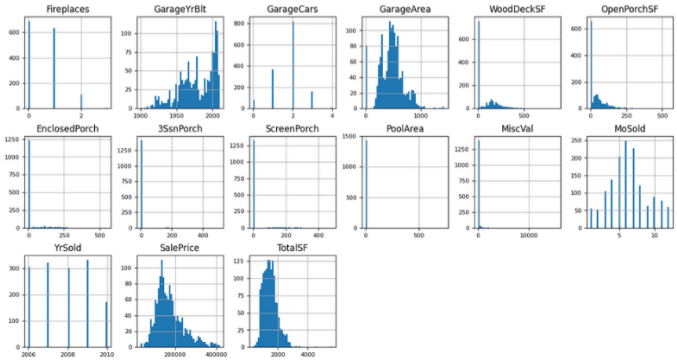


I created pair plots for a subset of numerical features in our dataset. Specifically, it plots the variables 'SalePrice,' 'TotalSF,' 'OverallQual,' and 'YearBuilt' against each other. Each scatterplot reveals how two variables are related, and the diagonal plots show the distribution of each variable.

Distribution for all the numerical features:

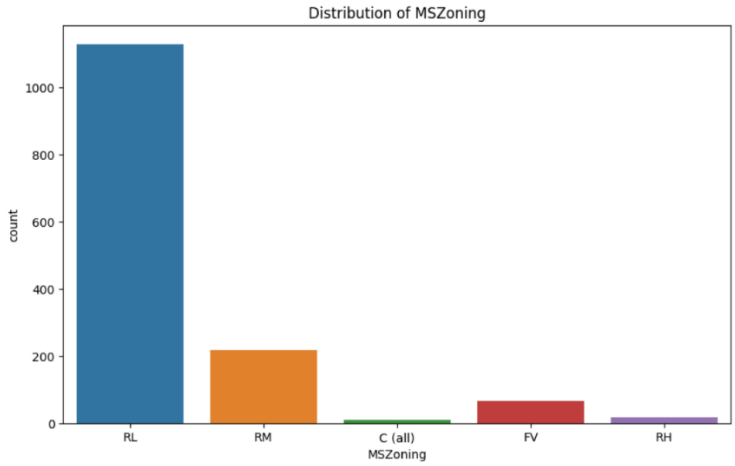
Each histogram displays the distribution of values for a specific feature.



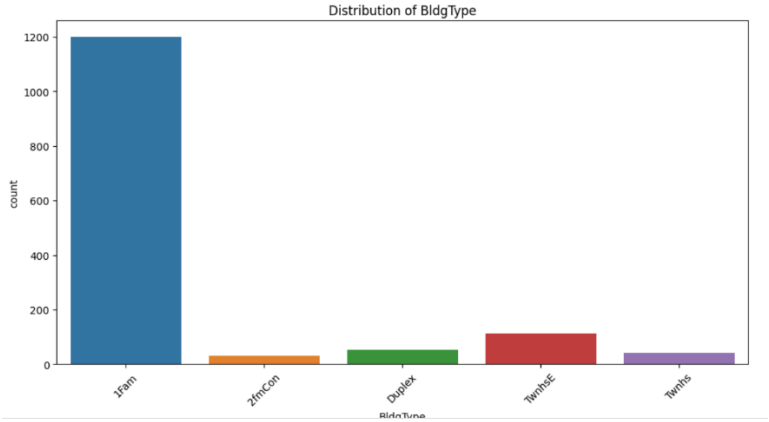


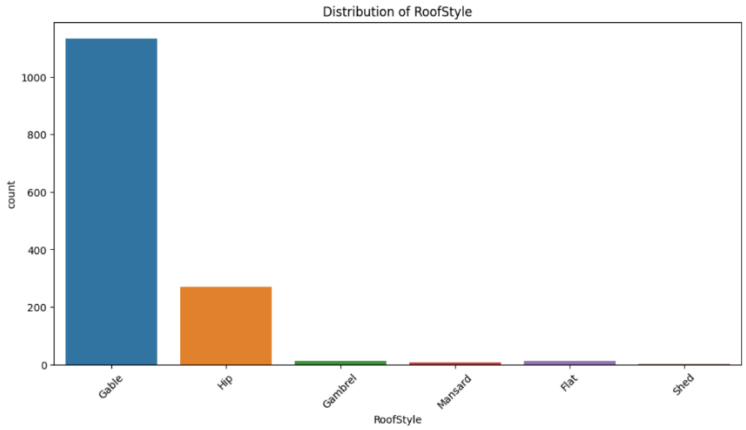
distribution of categorical variables:

This plot shows the distribution of different categories within the 'MSZoning' variable. Each bar represents a unique category or class within the 'MSZoning' variable, and the height of the bar corresponds to the frequency or count of occurrences for that category in the dataset.

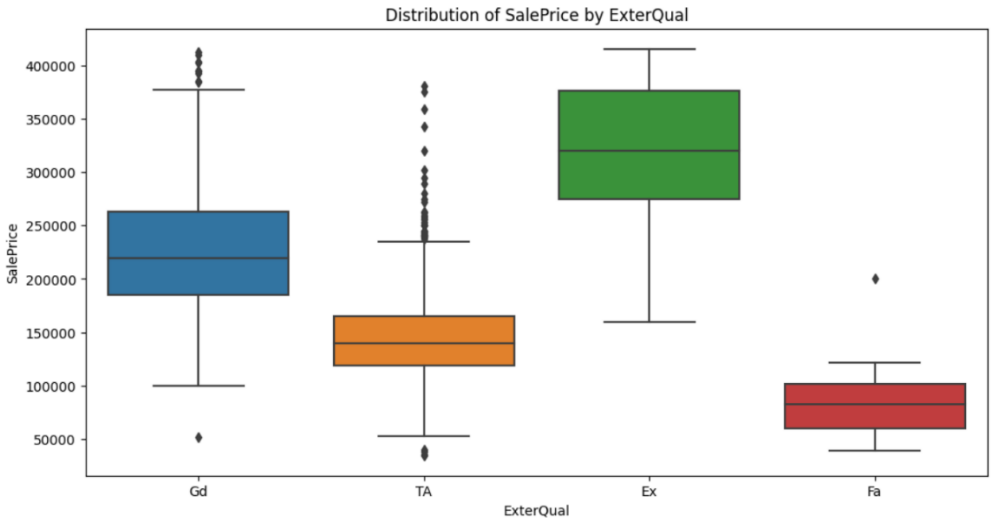


More examples:

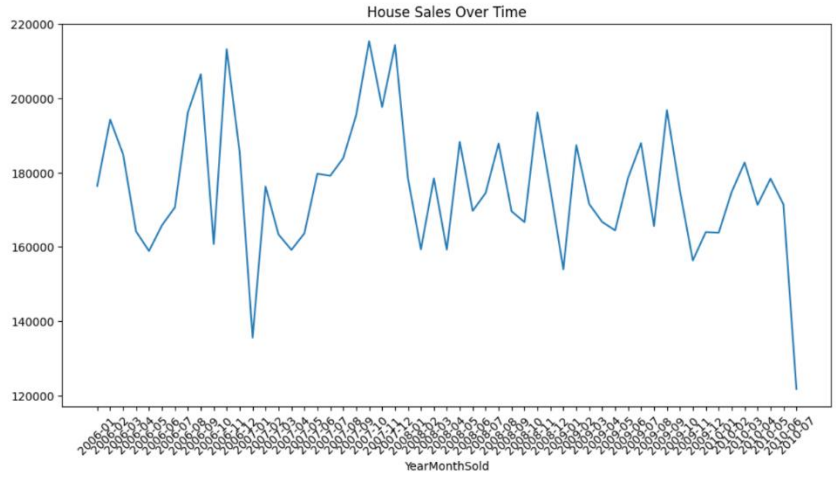




a box plot to visualize the distribution of 'SalePrice' by 'ExterQual':

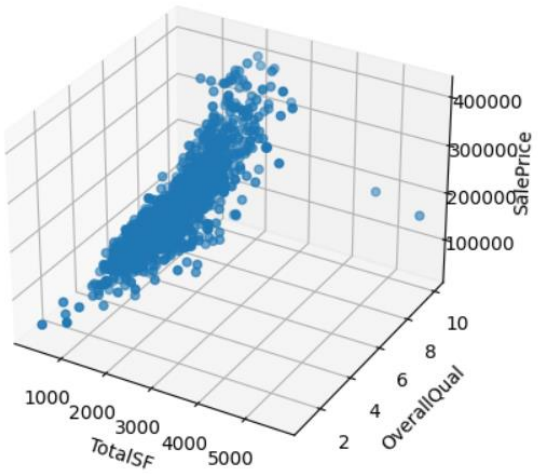


a line plot to visualize house sales over time:



a 3D scatter plot to visualize interactions between three numerical features:

Each data point is represented as a point in the 3D plot, with its position determined by the values of 'TotalSF,' 'OverallQual,' and 'SalePrice.'



5. Statistical Tests and Analysis (questions)

1) What are the building class and why it is important?

a. What a Building Class Is?

'MSSubClass' in the dataset denotes the "building class" variable, which indicates the type or classification of the residence based on many attributes. Usually, it contains details on the composition and design of the home. This variable's particular values represent various categories, each with a distinct meaning.

Common building classes, for instance, might consist of:

20: 1-STORY 1946 & NEWER ALL STYLES

30: 1-STORY 1945 & OLDER

60: 2-STORY 1946 & NEWER

b. The Importance of Building Class:

The building class holds significance in the context of real estate analysis and the dataset for multiple reasons:

Affects Property Value: The market value of a property can be greatly impacted by different building classifications. For example, a property with two stories may typically be worth less on the market than a house with one story and comparable features.

Informs Buyer Preferences: Homebuyers often have specific preferences for the style or type of dwelling they are looking for. Understanding the building class can help real estate professionals and sellers target the right audience.

guides Property Investment: The building class provides guidance to developers and real estate investors regarding the purchase, remodeling, and development of new properties. It aids in their evaluation of the possible return on investment.

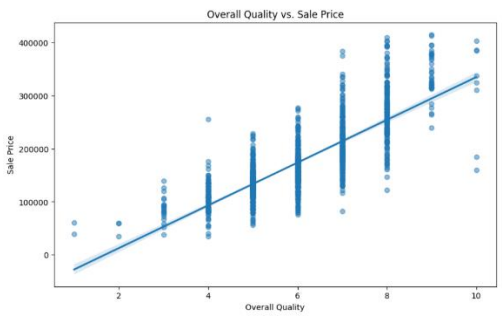
Market Analysis: Building class data is useful for understanding trends and patterns in real estate values and sales. It also helps analysts and researchers do market analysis.

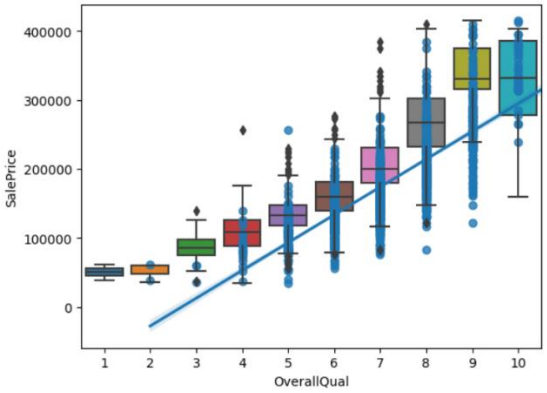
2) How does the overall quality (OverallQual) of a house relate to its sale price?

In this part I calculate the Pearson correlation coefficient. The result was:

```
Pearson Correlation Coefficient: 0.7963
```

The output (Pearson Correlation Coefficient) of 0.7963 confirms a significant and positive relationship between overall quality and sale price, meaning that higher-quality houses tend to be associated with higher sale prices.





As you can see, the line slopes upward from left to right, so it shows a positive correlation, as 'OverallQual' increases, 'SalePrice' tends to increase.

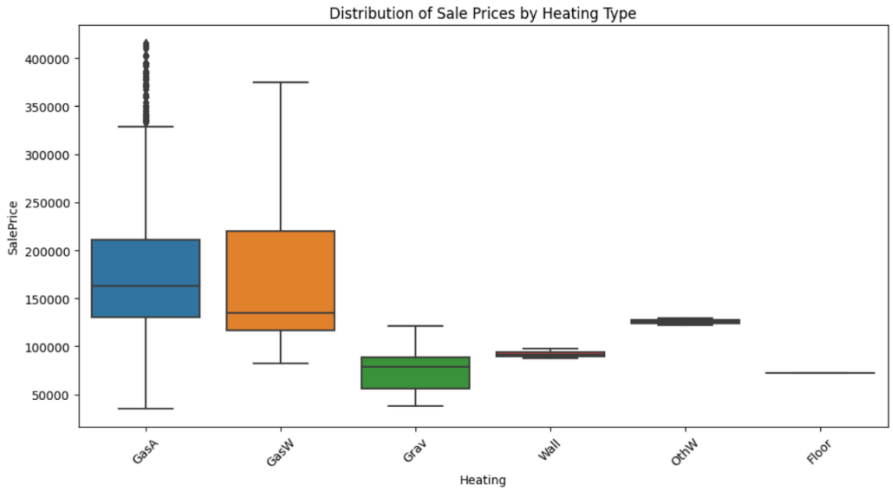
3) How do the different types of heating (Heating) affect the sale prices?

I performed ANOVA test to compare means of 'SalePrice' for different heating types.

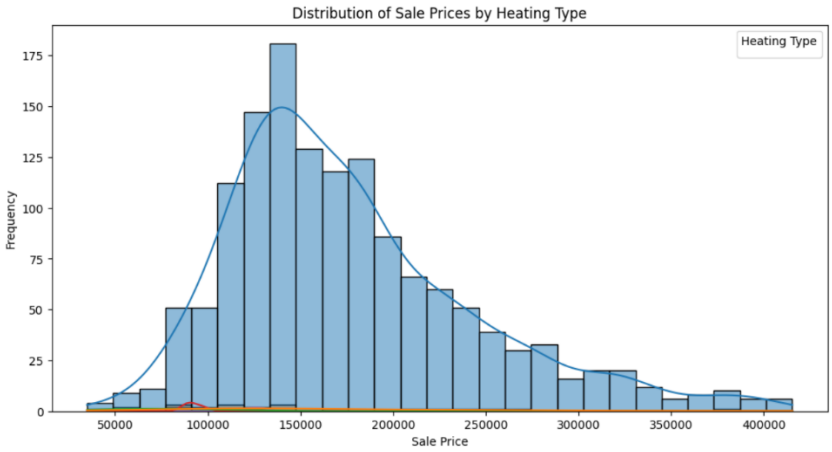
The output was:

```
ANOVA F-Statistic: 5.2373
P-Value: 0.0001
```

The ANOVA F-Statistic of 5.2373 and a p-value of 0.0001 indicate that there are statistically significant differences in sale prices among different heating types. In other words, there are differences in the average sale prices among the different heating types in the dataset.



Results are obvious in visualization too.



As you can see in box plot and histogram, there are statistically significant differences in sale prices among different heating types. In other words, the choice of heating type has a meaningful impact on sale prices.

4) How do the different types of utilities (Utilities) available in a property relate to sale prices?

I performed ANOVA test to compare means of 'SalePrice' for different utility types. The result was:

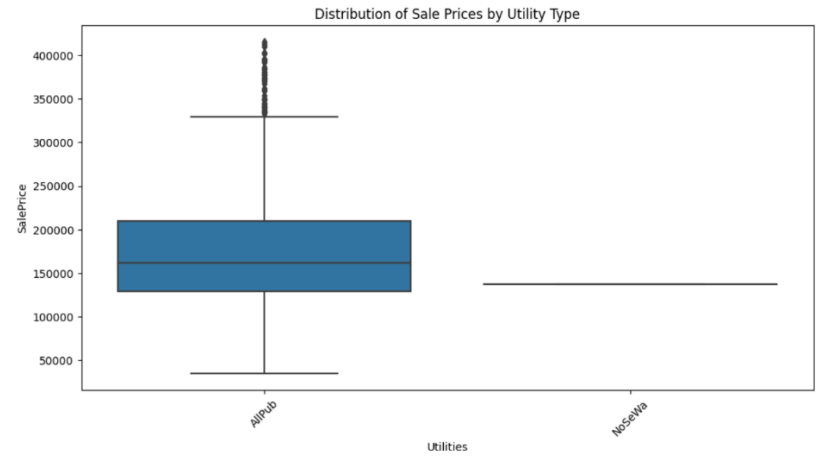
```
ANOVA F-Statistic: 0.3230
P-Value: 0.5699
```

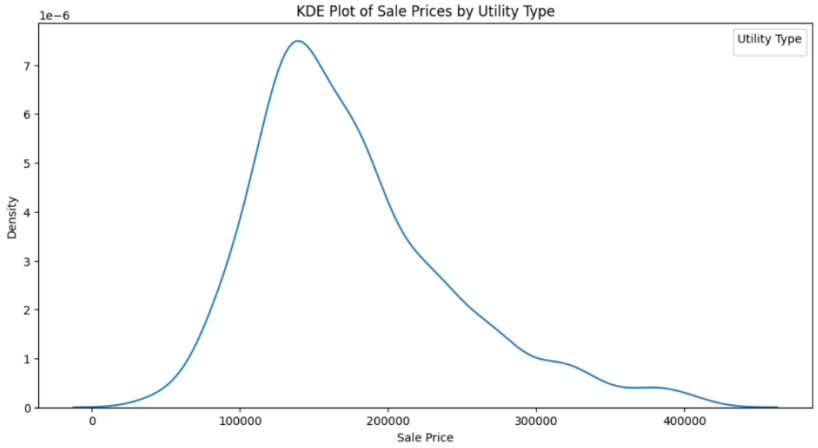
There are no statistically significant differences in sale prices between the various types of utilities that are accessible in properties, according to the ANOVA F-Statistic of 0.3230 and a p-value of 0.5699.

-The ratio of between-group variation to between-group variability is shown by the F-Statistic (F-value) of 0.3230. It is comparatively modest in this instance, indicating that differing utility kinds do not significantly affect sale prices.

-The p-value of 0.5699 is comparatively high, significantly higher than the standard significance level of 0.05. This implies that there is insufficient data to draw the conclusion that different utility kinds significantly affect sale prices.

We can infer from the outcome of this ANOVA test that there does not appear to be a significant correlation between the various utility kinds that are offered in a property and changes in sale prices. This suggests that the kind of utilities may not have a significant impact on property values in this dataset.





Results are obvious in bar plot too.

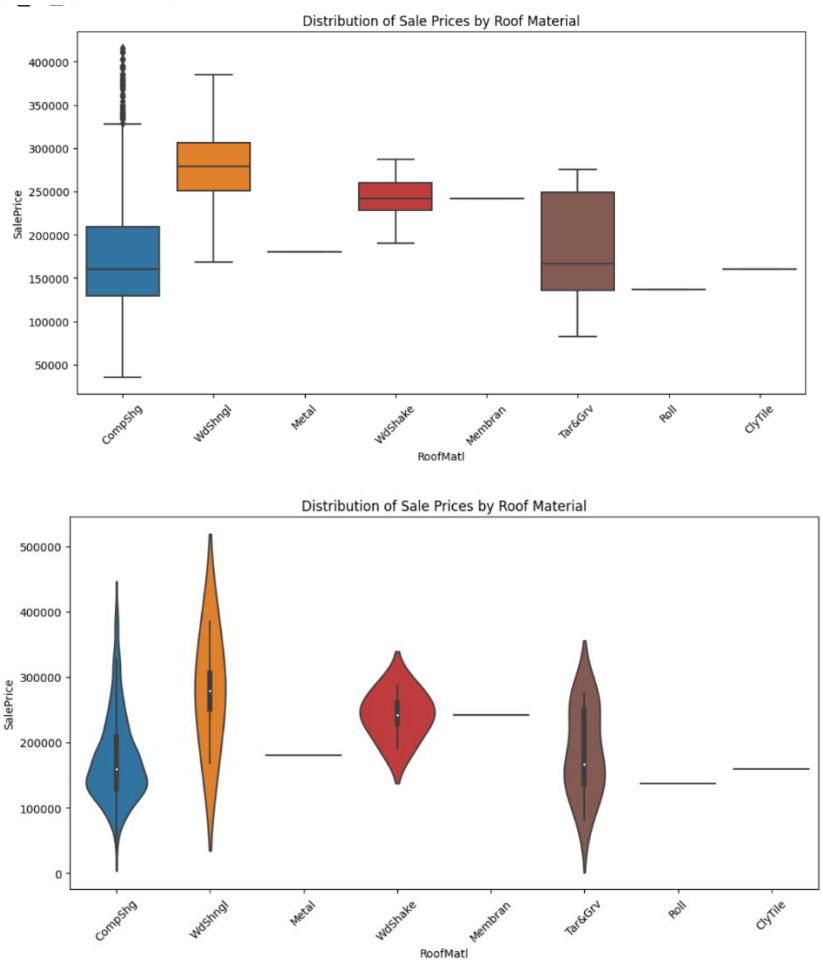
5) Are there any significant differences in property sale prices based on the type of roof material used for the houses? Do certain roof materials contribute to higher or lower sale prices, and if so, how significant are these differences?

I performed an analysis of variance (ANOVA) test to determine whether there is a significant difference in sale prices based on different roof materials. The result was:

```
P-Value: 0.029319226719694093
There is a significant difference in sale prices based on roof material.
```

The p-value in statistical hypothesis testing is a metric for the evidence that refutes a null hypothesis. Significant evidence to reject the null hypothesis is shown by a p-value less than the significance level, which is typically 0.05.

Given that the p-value (0.0293) in this instance is less than 0.05, the type of roof material utilized on homes significantly affects sale prices.



As you can see in plots, there is a significant difference in sale prices based on the type of roof material used for houses.

6) Does the age of a property (YearBuilt) have a significant impact on its sale price?

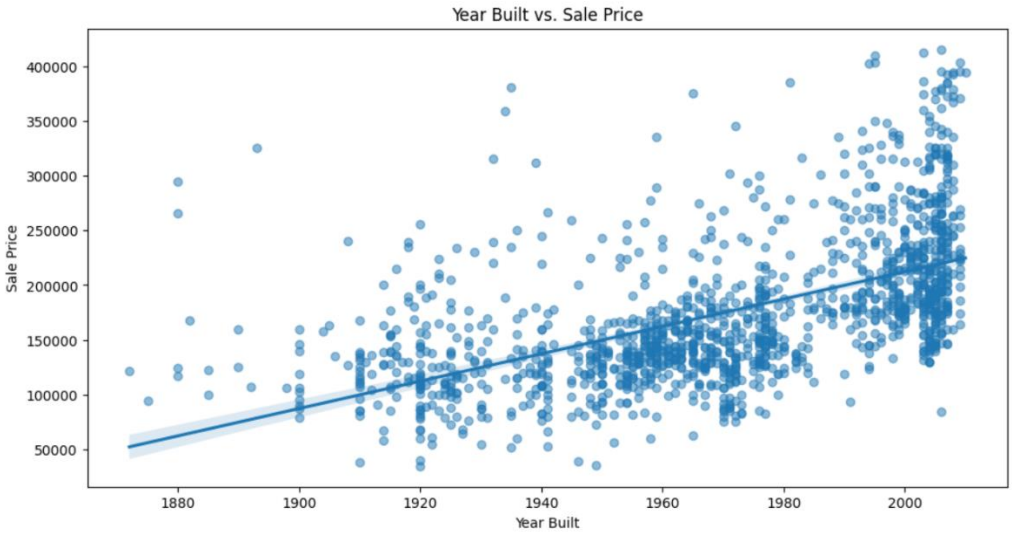
I calculated the Pearson correlation coefficient between the 'YearBuilt' (the year a property was built) and 'SalePrice' (the sale price of the property). The Pearson correlation coefficient measures the strength and direction of a linear relationship between two numerical variables.

Result:

```
Pearson Correlation Coefficient: 0.5573
```

-A positive correlation coefficient suggests that as the year in which a property was built (YearBuilt) increases, the sale price tends to increase as well.

-The value of 0.5573 indicates a moderate strength of correlation. It's not a perfect correlation, but there is a noticeable trend that newer properties tend to have higher sale prices.



There is a clear pattern visible in the scatter plot when "Year Built" and "Sale Price" are compared. There is a discernible tendency for sale prices to grow when the year a property was built (YearBuilt) increases. This pattern is highlighted by the regression line's upward slope, which shows a positive link between the property's age and sale price. Despite their scattered nature, the data points mostly match the trendline, confirming the relationship as a whole. This finding implies that newer homes typically get higher sale prices, indicating the effect of the property's age on its market worth.