# Documentation - exercise 1
# dataset : Housing Prices

Professor : Dr. Kheradpisheh

Teacher Assistant : MohammadReza Khanmohammadi

By: Katayoun Kobraei

## 1. Introduction to Dataset

The "House Prices" dataset is a comprehensive and detailed dataset commonly used in data science and machine learning for regression and predictive modeling tasks. This dataset provides a wide range of information about residential properties, with a primary focus on predicting the sale prices of houses. Below is a description of the key attributes and features found in the House Prices dataset:

- **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

Clearly, a number of these features are of no practical value. Consequently, we opted to eliminate the following functionalities:

‘Id’: If the dataset includes a column that serves as a unique identifier for each record (such as an "Id" column), it is typically not useful for predictive modeling and can be dropped.

‘Street’: This column typically indicates the type of road access to the property, but it might not have a strong impact on house prices in many cases; Because in manu cases we have stress’s values as ‘Pave’.

‘Alley’: Similar to the "Street" column, it represents the type of alley access and might not be a strong predictor of house prices.

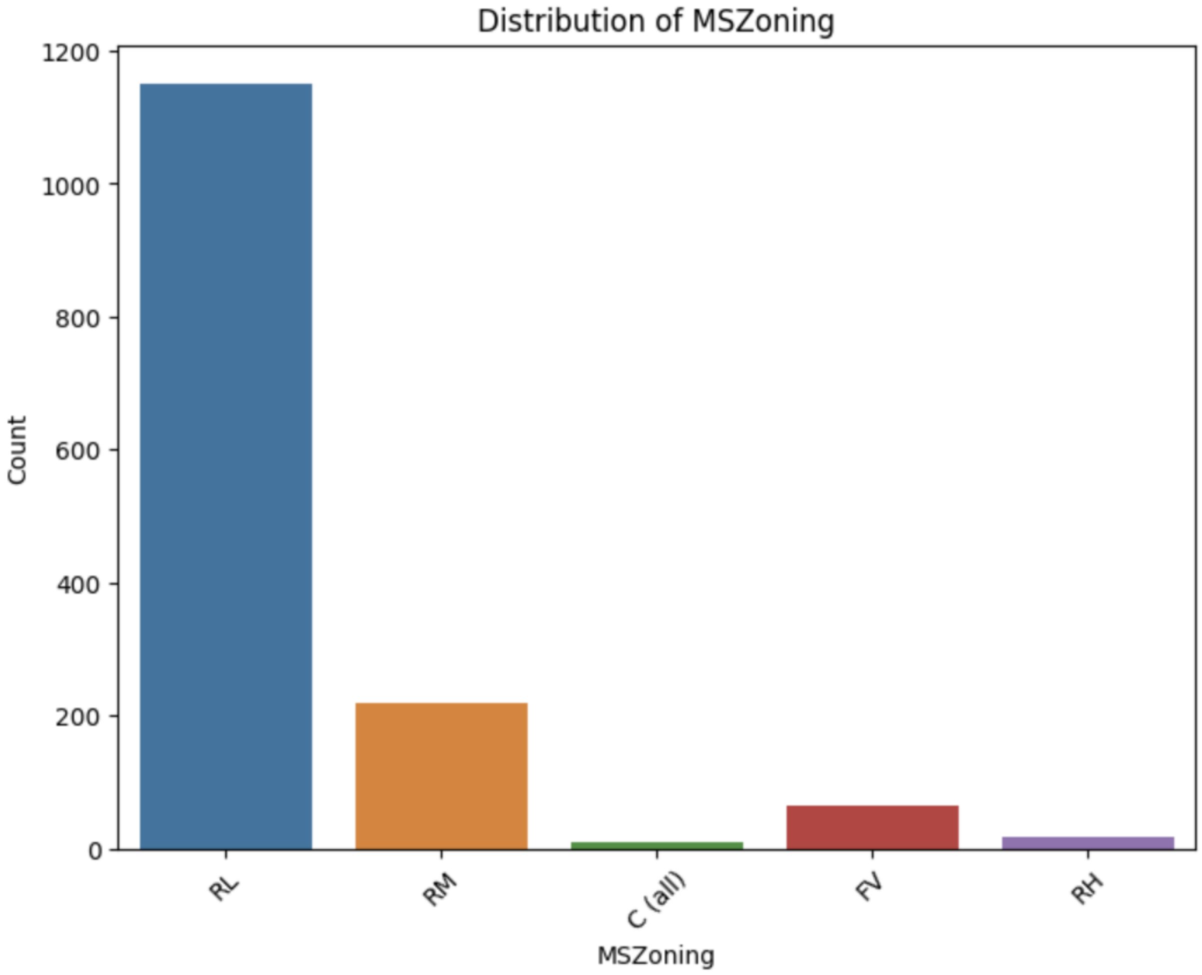
‘PoolQC’, ‘Fence’ and ‘MiscFeature’: If a significant portion of the data has missing values for this column, it may not be a strong predictor. These columns have too much null values. So it is better to drop all of their values.

## 2. Abstract

In this dataset, our goal is to prepare the data for analysis by revealing hidden patterns and relationships among its features. First, we start by identifying the most important and valuable features that we'll focus on in our analysis. Then, we tackle the issue of missing data by filling in those gaps using imputation techniques. Afterward, we take a closer look at the data, creating visualizations and exploring the features to uncover potential connections and patterns. Lastly, we formulate five hypotheses aimed at uncovering any concealed relationships among the features, and we use various statistical methods like t-tests, One-way ANOVA, Chi-squared tests, and regression analysis to help us uncover these insights.

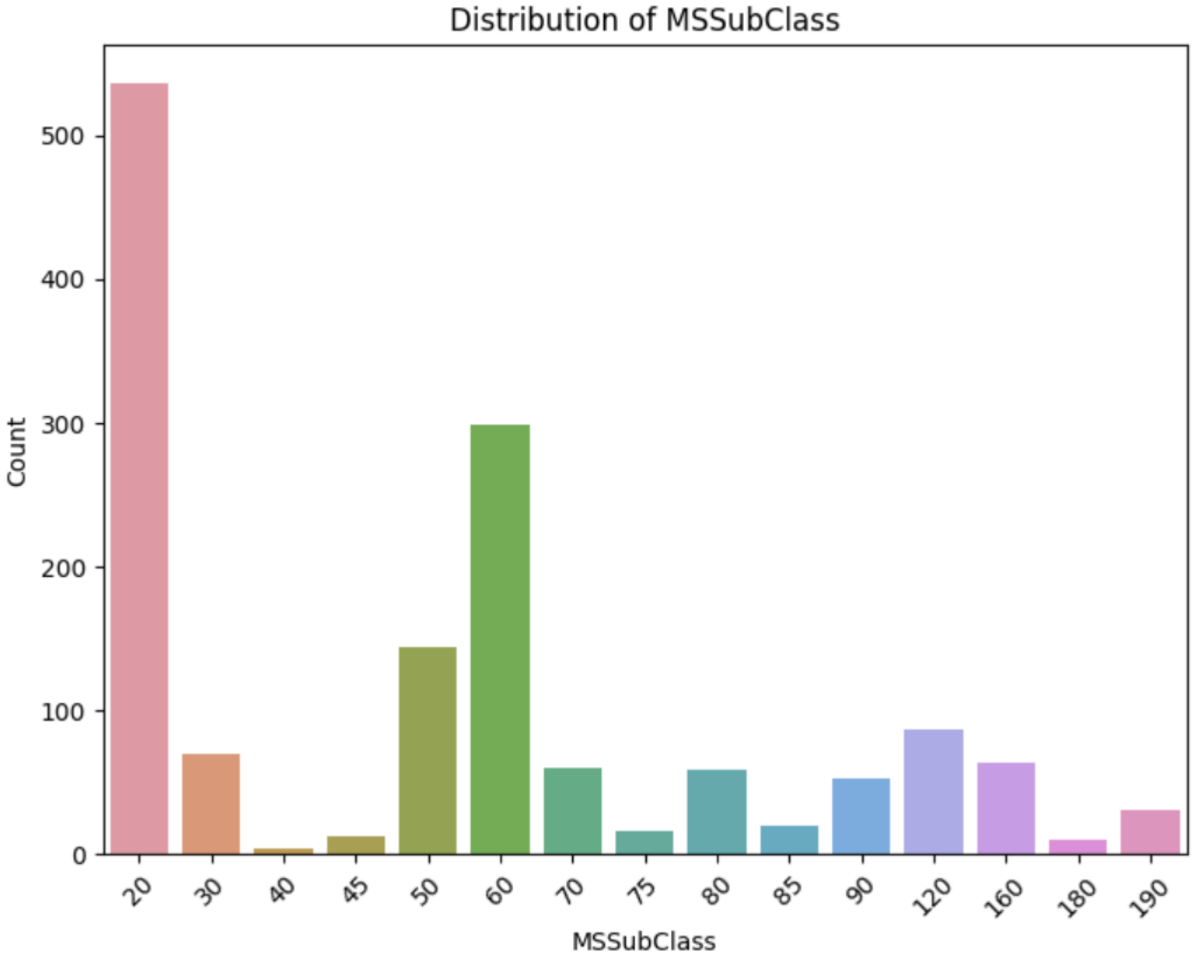
## 3. EDA (Exploratory data analysis

1. To see distribution of ‘MSZoning’ column



As you can see most of houses are in RL category of MCZoning. Residential zoning classifications are used by local governments and municipalities to regulate land use within a particular area. 'RL' zoning typically designates areas for low-density residential use, which often means that the properties in this zone are intended for single-family homes and have certain restrictions on things like lot size, building size, and land use.

2. To see distribution of ‘MSSubClass’ column (The Building Class)



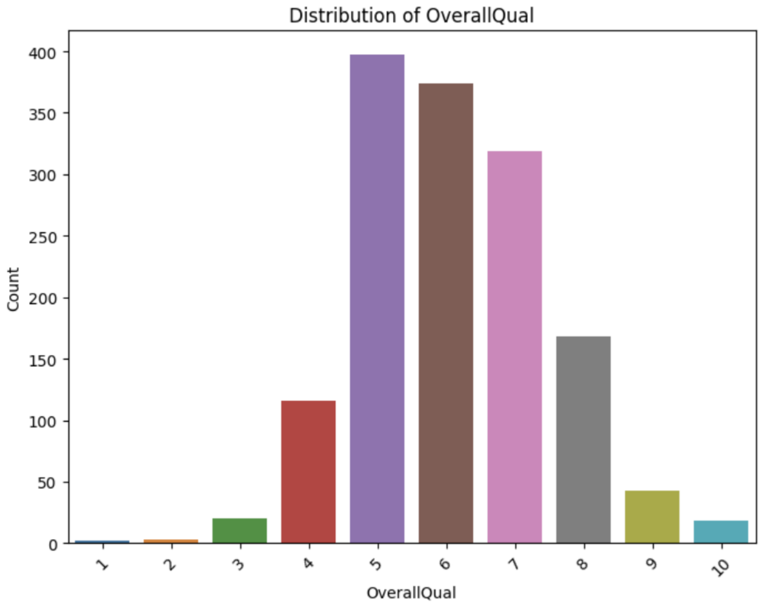
"MSSubClass" typically represents the building class or type of a property. It is a categorical feature that describes the general type of dwelling associated with each record in the dataset. common values we might find in this column include:

20: 1-STORY 1946 & NEWER ALL STYLES
30: 1-STORY 1945 & OLDER
40: 1-STORY W/FINISHED ATTIC ALL AGES 45: 1-1/2 STORY - UNFINISHED ALL AGES

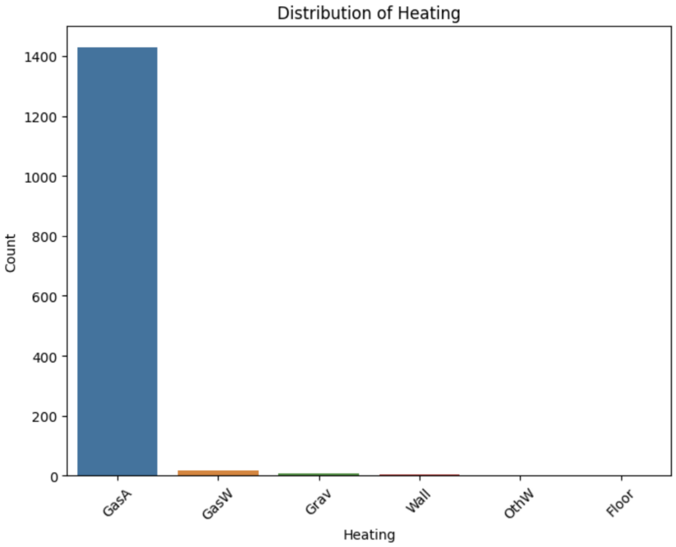
50: 1-1/2 STORY FINISHED ALL AGES
60: 2-STORY 1946 & NEWER
70: 2-STORY 1945 & OLDER
75: 2-1/2 STORY ALL AGES
80: SPLIT OR MULTI-LEVEL
85: SPLIT FOYER
90: DUPLEX - ALL STYLES AND AGES
120: 1-STORY PUD (Planned Unit Development) - 1946 &
NEWER 150: 1-1/2 STORY PUD - ALL AGES
160: 2-STORY PUD - 1946 & NEWER
180: PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190: 2 FAMILY CONVERSION - ALL STYLES AND AGES

Analyzing the distribution of "MSSubClass" alongside housing prices can help us to identify whether a property type tends to be more expensive or affordable. This column provides information about the architectural style and structure of a property. This can be essential for potential buyers or real estate professionals when evaluating the type of dwelling. Different building classes may have varying demand and market values. Analyzing the distribution of building classes in the dataset can provide insights into the preferences of buyers and the dynamics of the real estate market.

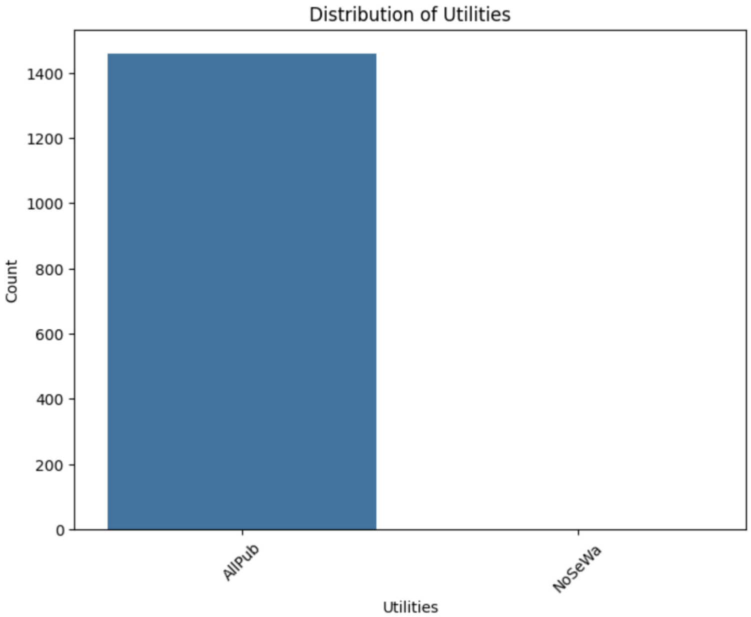
3. To see distribution of ‘OverallQual’ column



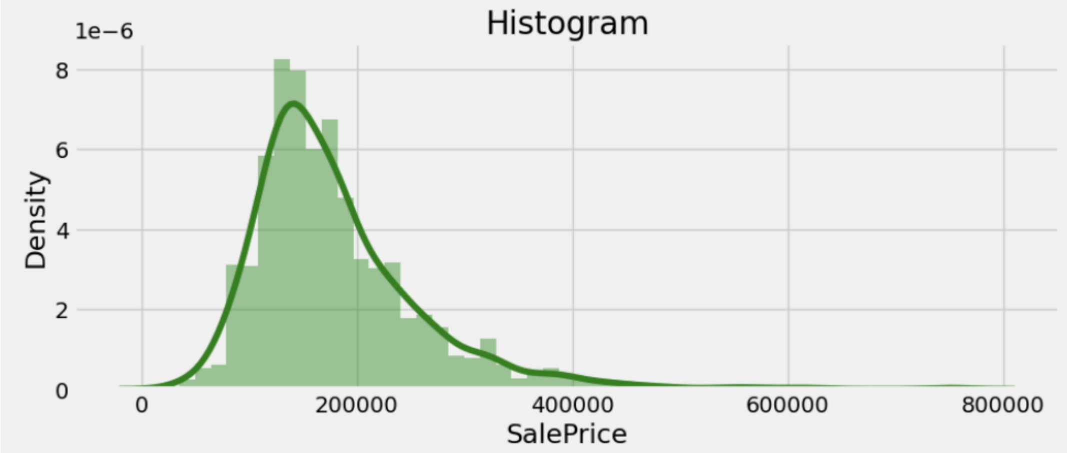
4. To see distribution of ‘Heating’ column

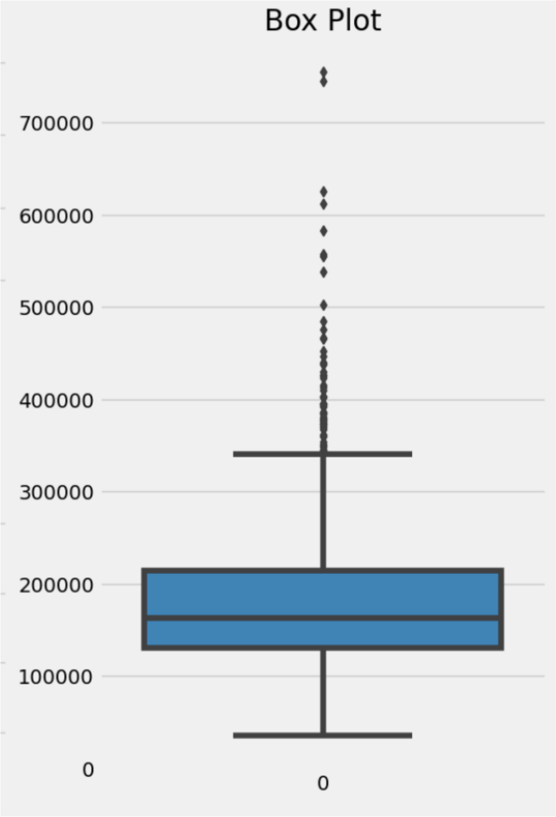


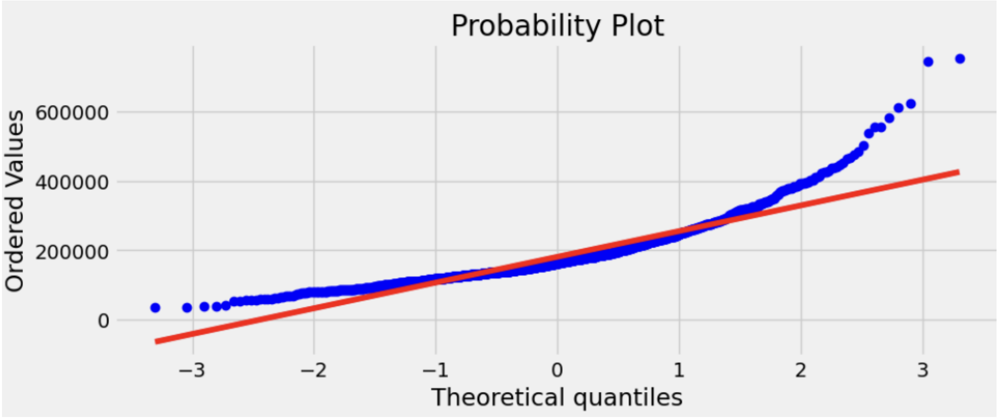
5. To see distribution of ‘Utilities’ column



6. Analysis target value (‘SalePrice’)

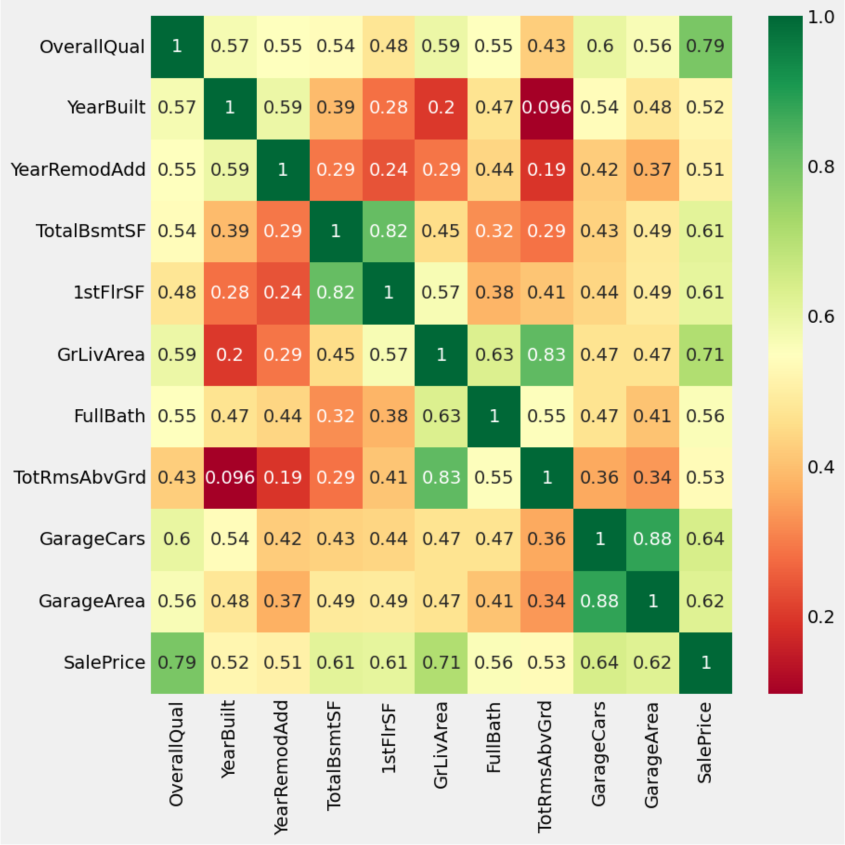






Upon inspection, it's clear that our target variable, the sale price, doesn't conform to a typical normal distribution, and we might spot a few exceptional values that warrant our attention. When we delve into the distribution of sale prices, we notice that the majority of properties tend to concentrate within a specific price range, often summarized by the mean or median sale price. This examination holds significant value because it provides us with a deeper insight into the affordability and appeal of properties within a specific market. This valuable knowledge equips real estate professionals, investors, and prospective homebuyers with the information they need to make informed and prudent decisions.

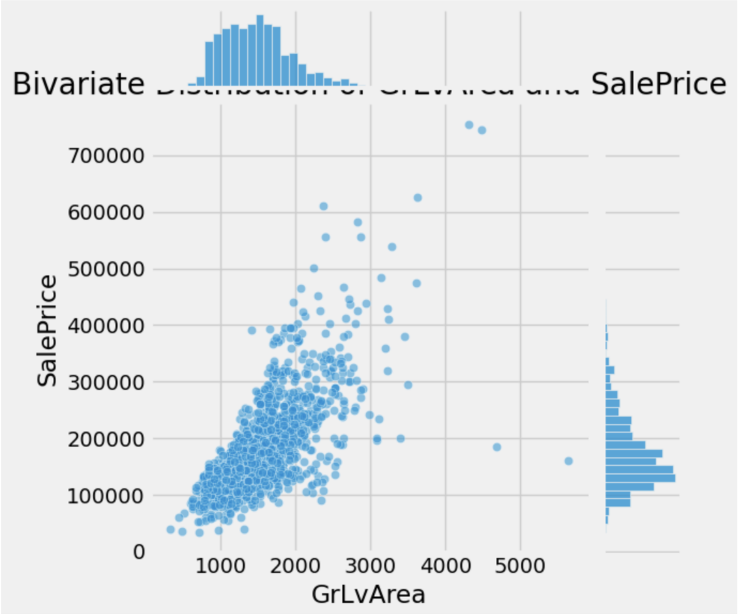
7. Correlation between inputs and target variable



| | OverallQual | YearBuilt | YearRemodAdd | TotalBsmtSF | 1stFlrSF | GrLivArea | FullBath | TotRmsAbvGrd | GarageCars | GarageArea | SalePrice |
|---|---|---|---|---|---|---|---|---|---|---|---|
| OverallQual | 1 | 0.57 | 0.55 | 0.54 | 0.48 | 0.59 | 0.55 | 0.43 | 0.6 | 0.56 | 0.79 |
| YearBuilt | 0.57 | 1 | 0.59 | 0.39 | 0.28 | 0.2 | 0.47 | 0.096 | 0.54 | 0.48 | 0.52 |
| YearRemodAdd | 0.55 | 0.59 | 1 | 0.29 | 0.24 | 0.29 | 0.44 | 0.19 | 0.42 | 0.37 | 0.51 |
| TotalBsmtSF | 0.54 | 0.39 | 0.29 | 1 | 0.82 | 0.45 | 0.32 | 0.29 | 0.43 | 0.49 | 0.61 |
| 1stFlrSF | 0.48 | 0.28 | 0.24 | 0.82 | 1 | 0.57 | 0.38 | 0.41 | 0.44 | 0.49 | 0.61 |
| GrLivArea | 0.59 | 0.2 | 0.29 | 0.45 | 0.57 | 1 | 0.63 | 0.83 | 0.47 | 0.47 | 0.71 |
| FullBath | 0.55 | 0.47 | 0.44 | 0.32 | 0.38 | 0.63 | 1 | 0.55 | 0.47 | 0.41 | 0.56 |
| TotRmsAbvGrd | 0.43 | 0.096 | 0.19 | 0.29 | 0.41 | 0.83 | 0.55 | 1 | 0.36 | 0.34 | 0.53 |
| GarageCars | 0.6 | 0.54 | 0.42 | 0.43 | 0.44 | 0.47 | 0.47 | 0.36 | 1 | 0.88 | 0.64 |
| GarageArea | 0.56 | 0.48 | 0.37 | 0.49 | 0.49 | 0.47 | 0.41 | 0.34 | 0.88 | 1 | 0.62 |
| SalePrice | 0.79 | 0.52 | 0.51 | 0.61 | 0.61 | 0.71 | 0.56 | 0.53 | 0.64 | 0.62 | 1 |

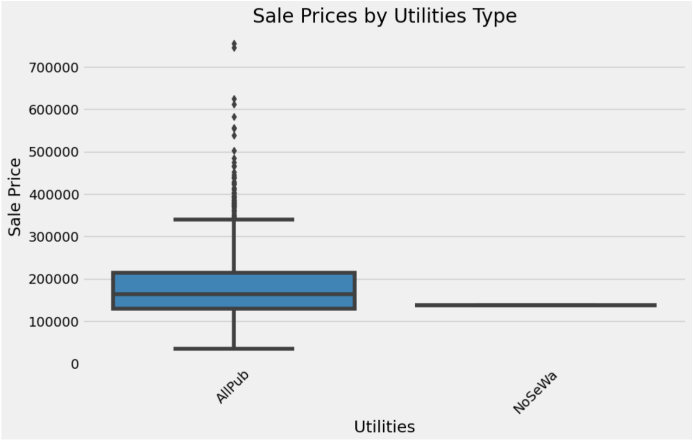
As we can see some variable are highly correlated to the target value like overallQual and GrlivArea

## 7.1. GrLivArea



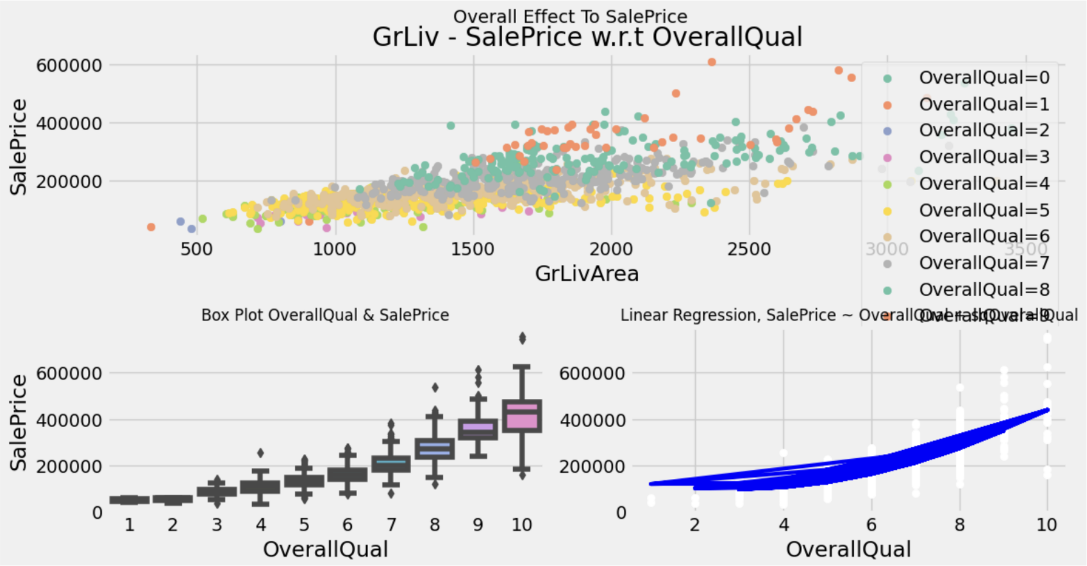
Even though GrLivArea is the most powerful variable of SalePrice, it's not enough! You can see the larger GrLivArea, the linearity was even far collapsed (In statistic, it is called 'Corn Shape Heteroscedasticity') Our mission is to let know what variables makes us to figure out inner strucutre

## 7.2. Utilities



It can be seen that Utilities is not a significant feature to predict target value.

7.3. OverallQual



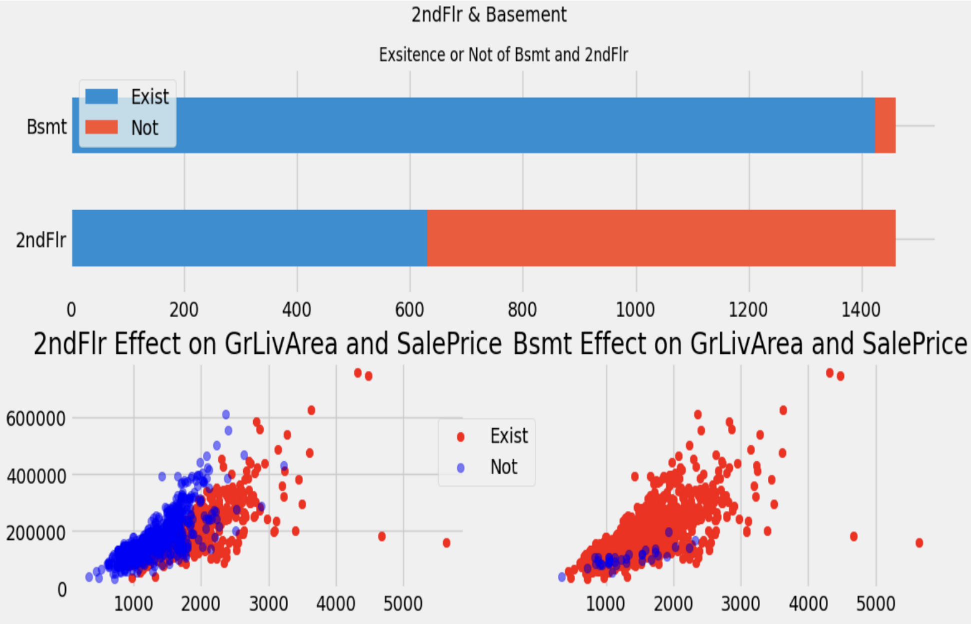
OverallQual is the best variables among Ordinal Variables regard of explaining SalePrice. I Love to see OverallQual Find

OverallQual causes different SalePrice where having same "GrLivArea". We have to put a strong attention!

OverallQual was proportional to SalePrice, and (1-2) almost identical.

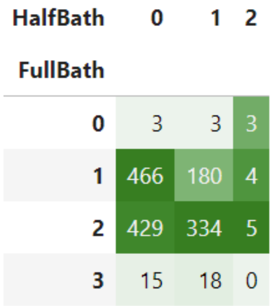
Square of OverallQual was a good variables since linear regression (SalePrice ~ O.Q + O.Q^2) has good curve shape

## 7.4. 2ndFlr & Basement

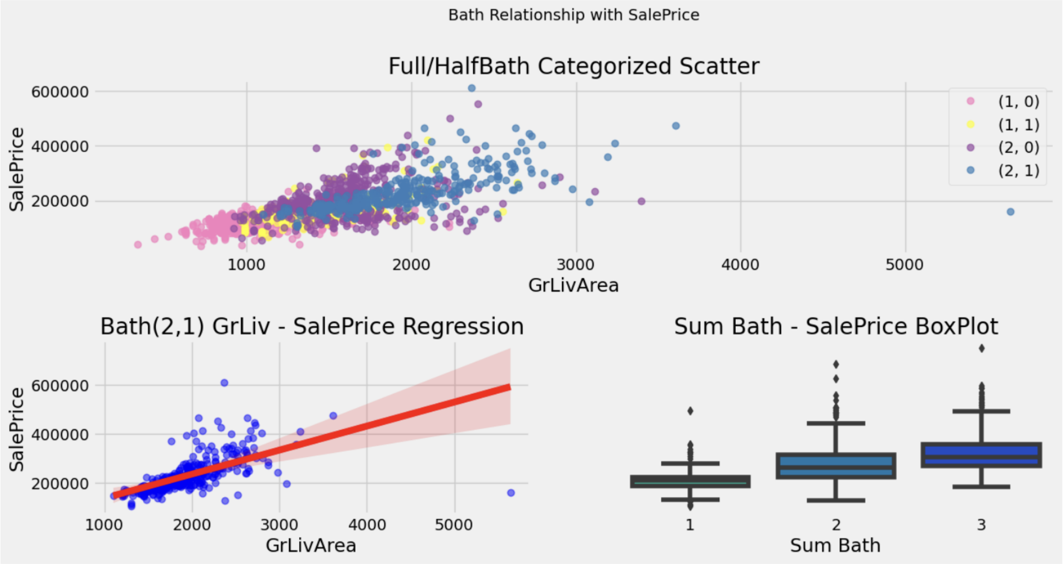


2ndFlrSF depressed the power of GrLiva toward SalePrice. Bsmt has nothing related to the price, so I won't see anything related to Bsmt.

## 7.5. Bath

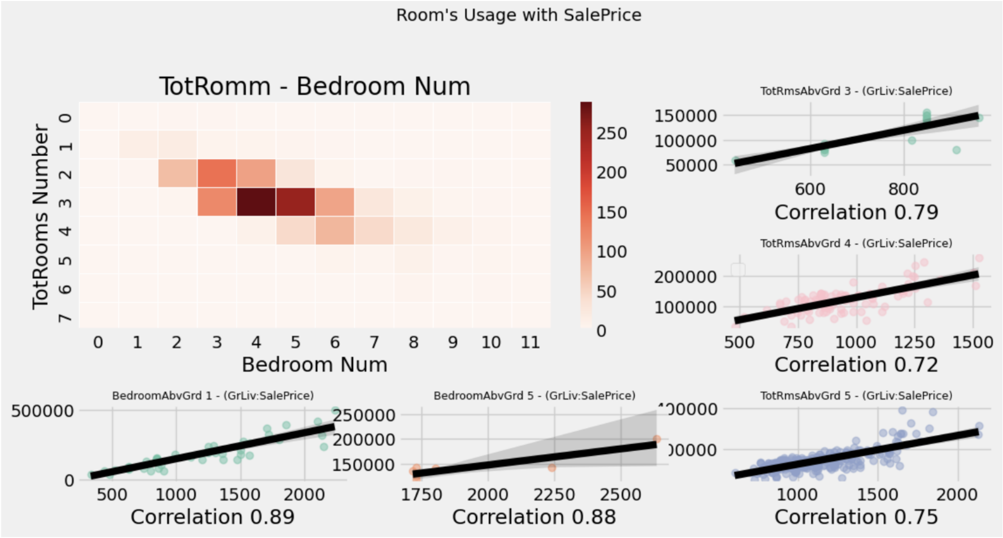


| HalfBath | 0 | 1 | 2 |
|---|---|---|---|
| FullBath | | | |
| 0 | 3 | 3 | 3 |
| 1 | 466 | 180 | 4 |
| 2 | 429 | 334 | 5 |
| 3 | 15 | 18 | 0 |



The Number of Bath usually increased the SalePrice, but especially (Full 2, Half1) improved the linearity and decreased the spreadness of SalePrice - GrLivArea.

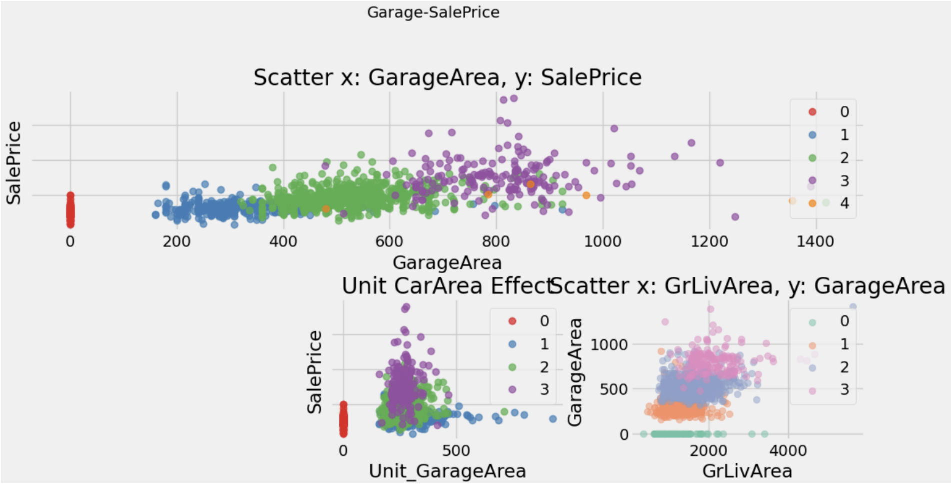
## 7.6. TotRoom



HeatMap Said "TotRomAbvGrd ~ BedroomAbvGrd ~ Bath" are proportional to each other!

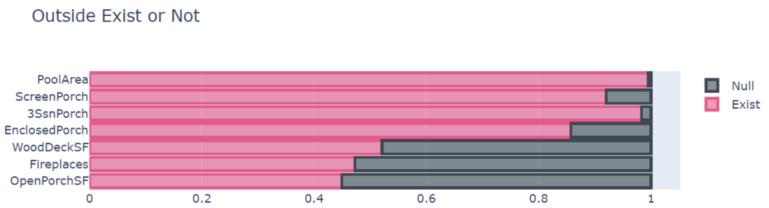
At TotRmsAbvGrd (3,4,5), the linearity w.r.t GrLiv-SalePrice was over 0.71.
At BedroomAbvGrd (0,1,2,3,4,5,6), the correlation wr.t. GrLiv-SalePrice was over 0.73

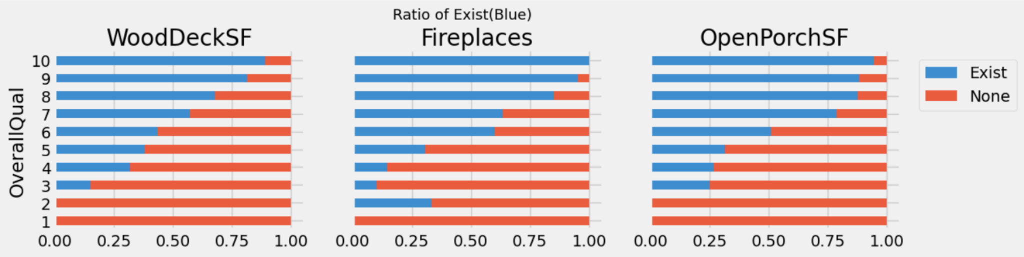
## 7.7. Garage



Most houses have two cars GarageArea makes Chunk having small lineratiy with SalePrice 0 Cars and 1 Cars has no difference in SalePrice 4 Cars are simliar with 3 Cars houses. Merge them (Update) Unt_Garage Area said that "Expensive house sustain the proper line of the area!" (Update) GrLivArea is a good variable not related to GarageArea. Those two variables enforce the prediction power.

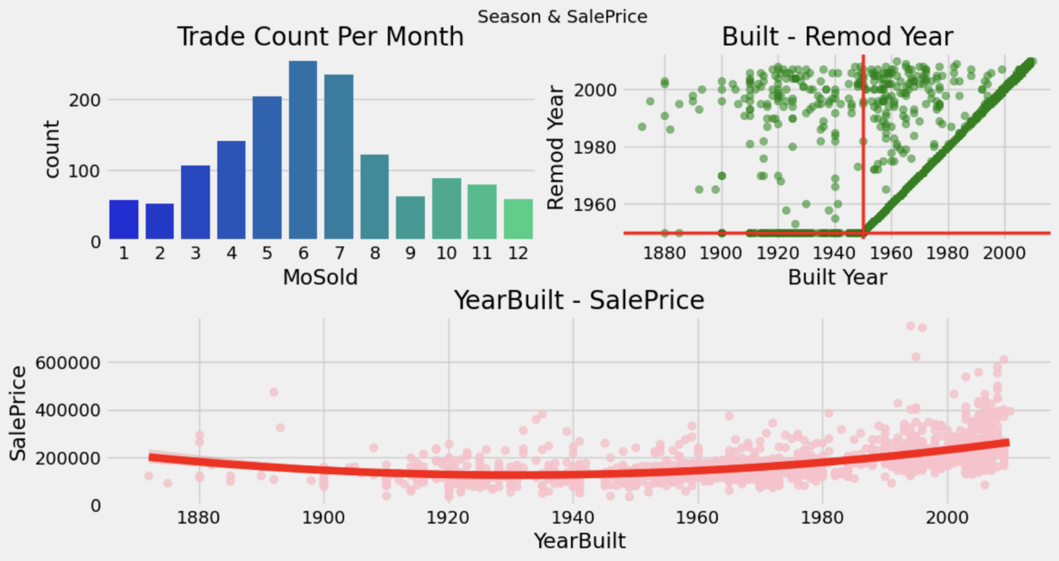
## 7.8. Outsides





Good Quality House has more outside instrumental places.
PoolArea, ScreenPorch, 3SsnPorch was almost not appeared

## 7.9. Season



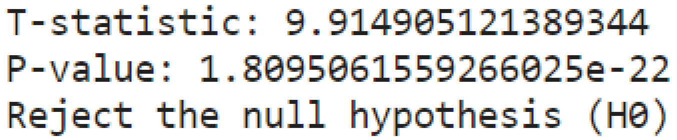
The amount of trade was increased by rising temperatures. Most old houses were remodeled in 1950. The part of house, built after 1950, was not remodeled yet YearBuilt^2 is proper if the variables is used to predict

## 8. Hypothesis tests

### 8.1. Hypothesis Test for the Difference in Sale Prices of Houses with and without Central Air Conditioning

**Null Hypothesis (H0)**: The presence or absence of central air conditioning does not significantly affect the sale price of a house.

**Null Hypothesis (H1)**: The presence or absence of central air conditioning does not significantly affect the sale price of a house.



```
T-statistic: 9.914905121389344
P-value: 1.8095061559266025e-22
Reject the null hypothesis (H0)
```

Null Hypothesis (H0): This is like our baseline assumption. It says that having central air conditioning in a house doesn't really change the sale price much when compared to houses without central air conditioning. It suggests that any differences in sale prices we see could just be due to chance.

Alternative Hypothesis (H1): On the flip side, the alternative hypothesis argues that central air conditioning does indeed affect the sale price of a house. It proposes that there's a real and noticeable difference in sale prices between houses with central air and those without. It implies that these differences aren't random but are influenced by the presence or absence of central air conditioning.

getting to the results:
T-statistic: Think of this as a measure of how much the average sale prices differ between the two groups (with and without central air conditioning). In this case, the T-statistic is pretty high, which means there's a significant difference in the average sale prices.

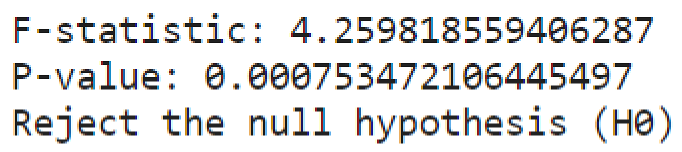
P-value: This little number tells us how likely it is to get results as extreme as the ones we've observed, assuming the null hypothesis is true. When your P-value is super low (like the one you have, close to zero), it's a strong hint that the differences in sale prices aren't happening by chance. Instead, it suggests that the presence or absence of central air conditioning is making a real impact on sale prices.

In a nutshell:
With such a tiny P-value and a high T-statistic, you've got good reason to toss out the null hypothesis. This means there's solid statistical evidence to support the alternative hypothesis: central air conditioning really does affect the sale price of a house. To put it plainly, houses with central air conditioning tend to have different (usually higher) sale prices compared to those without. It's not just luck; it's a real influence.

## 8.2. One-way ANOVA test to assess how different types of heating affect sale prices

**Null Hypothesis (H0)**: Different types of heating do not significantly affect sale prices.

**Null Hypothesis (H1)**: Different types of heating significantly affect sale prices.



```
F-statistic: 4.259818559406287
P-value: 0.000753472106445497
Reject the null hypothesis (H0)
```

F-statistic: This statistic measures how much the differences between groups (in this case, the types of heating in houses) compare to the differences within those groups. It helps us figure out if there are big differences in sale prices between the different heating types. The F-statistic you have is 4.26.

P-value: This tells us the likelihood of getting results like the ones we see if the null hypothesis (which says heating types don't really affect sale prices) were true. Your P-value is quite small, about 0.00075.

Interpretation:
Null Hypothesis (H0): This is the idea we're testing. It suggests that the different heating types don't really change sale prices much. Any differences we see in sale prices between heating types could just be random.

Alternative Hypothesis (H1): This is the opposite idea. It says that the different heating types do have a real effect on sale prices, meaning there are actual and important differences in sale prices between heating types.

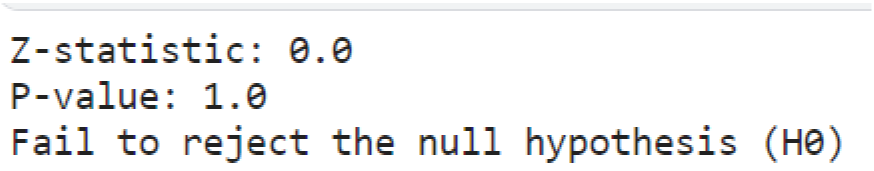
looking at the F-statistic and P-value:
The F-statistic of 4.26 tells us there are some differences in sale prices between the heating types.
The very small P-value (0.00075) suggests that the chance of seeing these differences in sale prices between heating types just by random luck is very, very low.

Conclusion:
With such a small P-value, you've got a good reason to throw out the null hypothesis (H0). This means there's strong statistical evidence supporting the alternative hypothesis (H1). In simpler terms, the type of heating in a house really does make a difference in the sale price. So, the choice of heating isn't random; it's connected to real and meaningful differences in sale prices.

## 8.3. Z-test to check how The overall quality of a house affect sale prices

**Null Hypothesis (H0)**: The overall quality of a house does not significantly affect its sale price.

**Null Hypothesis (H1)**: The overall quality of a house significantly affects its sale price.



```
Z-statistic: 0.0
P-value: 1.0
Fail to reject the null hypothesis (H0)
```

Z-statistic: This tells us how much the average quality of houses in our sample differs from what we'd expect for all houses. Here, the Z-statistic is 0.0, meaning our sample's average quality matches the expected overall average.

P-value: This measures the chance of seeing results like we have if the null hypothesis (which says quality doesn't really affect sale prices) were true. Your P-value is high, 1.0, meaning there's a big chance this result is just random.

Interpretation:

Null Hypothesis (H0): This is the idea we're testing. It says that a house's quality doesn't have a big impact on its sale price. Any connection between quality and price could be due to chance.

Alternative Hypothesis (H1): This is the opposing idea. It says the quality of a house does affect the sale price. There's a real connection between quality and price.

looking at the Z-statistic and P-value:

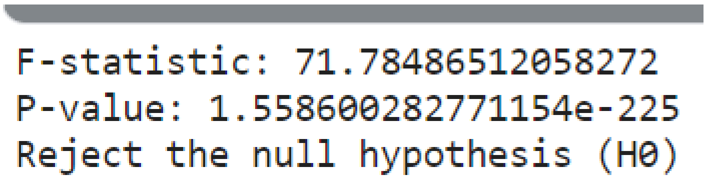
The Z-statistic being 0.0 means our sample's average quality matches what we'd expect. There's no real difference.
The high P-value (1.0) means it's very likely we're seeing these results by random chance, meaning quality probably isn't significantly tied to sale prices.

Conclusion:
With such a high P-value and a Z-statistic of 0.0, you can't reject the null hypothesis (H0). So, there's no strong statistical evidence to support the alternative hypothesis (H1). In simpler terms, this means we don't have enough data to say that a house's overall quality significantly impacts its sale price based on the Z-test we performed.

## 8.4. ANOVA Test for the Effect of Neighborhood on Sale Price

**Null Hypothesis (H0)**:Neighborhood does not significantly affect sale price.

**Null Hypothesis (H1)**: Neighborhood significantly affects sale price.



```
F-statistic: 71.78486512058272
P-value: 1.558600282771154e-225
Reject the null hypothesis (H0)
```

Null Hypothesis (H0): This is the idea we're testing. It says that where a house is located (the neighborhood) doesn't really make a big difference in the sale price. Any differences in sale prices we see among neighborhoods could just be due to chance.

Alternative Hypothesis (H1): This is the opposite idea. It says that the neighborhood does have a big impact on the sale price of a house. There are real and significant differences in sale prices among neighborhoods.

looking at the F-statistic and P-value:

The F-statistic is quite big (71.78), telling us that there are major differences in sale prices among neighborhoods.

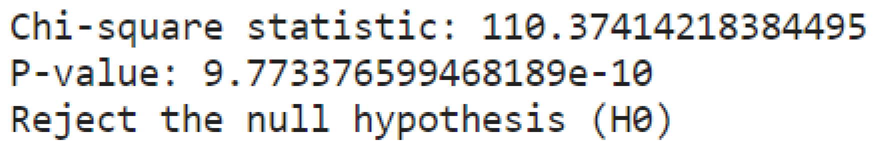
The P-value is extremely small (almost zero), meaning the chance of seeing these differences in sale prices among neighborhoods just by luck is practically impossible. It strongly suggests that the neighborhood does indeed affect sale prices.

Conclusion:
With such a tiny P-value and a high F-statistic, you've got a solid reason to toss out the null hypothesis (H0). This means there's strong statistical evidence supporting the alternative hypothesis (H1). In plain terms, the neighborhood does significantly impact house sale prices. So, your choice of neighborhood is linked to real and meaningful differences in sale prices.

## 8.5. Hypothesis Test for the Association between Roof Style and House Style (Chi-Square Test)

**Null Hypothesis (H0)**: The roof style and house style are independent of each other.

**Null Hypothesis (H1)**: The roof style and house style are associated with each other.



```
Chi-square statistic: 110.37414218384495
P-value: 9.773376599468189e-10
Reject the null hypothesis (H0)
```

Chi-square statistic: This statistic helps us see how much the actual numbers in a table (like how different roof and house styles are related) differ from what we'd expect if they were totally unrelated. In this case, the chi-square statistic is around 110.37.

P-value: This tells us how likely it is to get results as unusual as the ones we've seen if the null hypothesis (which says there's no connection between roof style and house style) were true. Your P-value is incredibly small, almost zero, which means there's very little chance these results are just random.

Interpretation:
Null Hypothesis (H0): This is the idea we're testing. It suggests that the type of roof style and house style are totally unrelated. There's no connection or relationship between them.

Alternative Hypothesis (H1): This is the opposing idea. It says the roof style and house style are connected, which means there's a real relationship or dependence between them.

looking at the chi-square statistic and P-value:
The chi-square statistic being around 110.37 tells us that there's a significant difference between what we'd expect and what we're seeing. This suggests a strong connection between roof style and house style.

The extremely small P-value (almost zero) means there's almost no chance we'd see these results by random luck. It strongly points to the idea that roof style and house style are related.

Conclusion:
With such a tiny P-value and a high chi-square statistic, you've got a good reason to toss out the null hypothesis (H0). This means there's strong statistical evidence supporting the alternative hypothesis (H1). So, in practical terms, it suggests that the type of roof style and house style are indeed related, meaning there's a significant connection between these two categorical variables.