Data Analysis on Houses Analysis Report

1. Data Exploration

Here is a detailed explanation of each column in the dataset:

- **Id**: A unique identifier for each house. integer
- **MSSubClass**: The building class of each house. integer
- **MSZoning**: The general zoning classification of each house. string
- **LotFrontage**: The linear feet of street connected to each property. float
- **LotArea**: The lot size of each property. integer
- **Street**: The type of road access to each property. string
- **Alley**: Type of alley access to each property. string
- **LotShape**: The general shape of each property. string
- **LandContour**: The flatness of the property. string
- **Utilities**: Type of utilities available to each property. string
- **LotConfig**: The configuration of each property. string
- **LandSlope**: The slope of the property. string
- **Neighborhood**: Physical locations within Ames city limits. string

- **Condition1**: Proximity to various conditions such as a busy street or railroad.
  Condition2: Proximity to various conditions (if more than one is present).
- **BldgType**: The type of dwelling within each house. string
- **HouseStyle**: The style of dwelling within each house. string
- **OverallQual**: Overall material and finish quality of the house. integer
- **OverallCond**: Overall condition rating of the house. integer
- **YearBuilt**: Original construction date of the house. integer
- **YearRemodAdd**: Remodel date (same as construction date if no remodeling or additions). integer
- **RoofStyle**: Type of roof on the house. string
- **RoofMatl**: Roof material on the house. string
- **Exterior1st**: Exterior covering on the house (most common). string
- **Exterior2nd**: Exterior covering on the house (if more than one material). string
- **MasVnrType**: Masonry veneer type on the house. string
- **MasVnrArea**: Masonry veneer area in square feet. float
- **ExterQual**: Exterior material quality rating for the house. string
- **ExterCond**: Present condition of the material on the exterior for the house. string
- **Foundation**: Type of foundation for the house. string

- **BsmtQual**: Height of the basement for the house (if present). string

- **BsmtCond**: General condition of the basement for the house (if present). string
- **BsmtExposure**: Walkout or garden level basement walls for the house (if present). string
- **BsmtFinType1**: Quality of basement finished area for the first type (if present). string
- **BsmtFinSF1**: Type 1 finished square feet in basement for the house (if present). float
- **BsmtFinType2**: Quality of second finished area (if present) in basement for the house (if present). string
- **BsmtFinSF2**: Type 2 finished square feet in basement for the house (if present). float
  BsmtUnfSF: Unfinished square feet of basement area for the house (if present). float
  TotalBsmtSF: Total square feet of basement area for the house (if present). float
- **Heating**: Type of heating in the house. string
- **HeatingQC**: Heating quality and condition rating in the house. string
- **CentralAir**: Central air conditioning in the house (yes or no). string
- **Electrical**: Electrical system in the house. string
- **1stFlrSF**: First Floor square feet in the house. integer
- **2ndFlrSF**: Second floor square feet in the house. integer
- **LowQualFinSF**: Low quality finished square feet in all floors except basement and attic for the house. integer
- **GrLivArea**: Above grade (ground) living area square feet in the house. integer

- **BsmtFullBath**: Basement full bathrooms in the house (if present). integer
- **BsmtHalfBath**: Basement half bathrooms in the house (if present). integer
- **FullBath**: Full bathrooms above grade in the house. integer
- **HalfBath** : Half baths above grade in the house . integer
- **BedroomAbvGr** : Number of bedrooms above basement level . integer
- **KitchenAbvGr** : Number of kitchens above basement level . integer
- **KitchenQual** : Kitchen quality rating .string
- **TotRmsAbvGrd** : Total rooms above grade (does not include bathrooms) . integer
- **Functional** : Home functionality rating .string
- **Fireplaces** : Number of fireplaces in a home . integer
- **FireplaceQu** : Fireplace quality rating .string
- **GarageType** : Garage location .string
- **GarageYrBlt** : Year garage was built .float
- **GarageFinish** : Interior finish of garage . string
- **GarageCars** : Size of garage in car capacity .integer
- **GarageArea** : Size of garage in square feet .float
- **GarageQual** : Garage quality rating .string
- **GarageCond** : Garage condition rating

- **PavedDrive**: string
- **WoodDeckSF**: integer

- **OpenPorchSF**: integer
- **EnclosedPorch**: integer
- **3SsnPorch**: integer

- **ScreenPorch**: integer PoolArea: integer
- **PoolQC**: string
- **Fence**: string
- **MiscFeature**: string
- **MiscVal**: integer
- **MoSold**: integer
- **YrSold**: integer
- **SaleType**: string
- **SaleCondition**: string
- **SalePrice**: integer

2) **Data Processing:**

- **Date Loading**

Before delving into the statistical report, it's important to become familiar with the dataset and its features. This will help us to be prepare for any issues that might arise down the road. To do this, we can use various functions, as below:

Df.columns : provides the list of the columns' names.

Df.dtypes: provides the list of the type of each column of the table.

Df.describe(): gives us a summary like: dispersion and shape of our dataset. In other words, provides some detailed statistical information of each column, excluding NaN values:

1. **count**: The number of non-empty values.
2. **mean**: The average (mean) value.
3. **std**: The standard deviation.
4. **min**: The minimum value.
5. **25%**: The 25th percentile.
6. **50%**: The 50th percentile.
7. **75%**: The 75th percentile.
8. **max**: The maximum value.

- **Null Values**

  Now that we have a clear understanding of our dataset, our goal is to find out null values. It's not uncommon to encounter columns or features with a significant percentage of missing data in certain datasets. When this occurs, attempting to impute or fill in these null values can often prove futile. The reason for this is the lack of sufficient information or data to support effective imputation strategies. For instance, consider the "MiscFeature" column in our data frame, which exhibits a null value percentage exceeding 90%. In such instances, it is often pragmatic to remove the entire column, as it offers limited value to our analysis. The same principle applies to other columns with a high prevalence of null values.

  Dropped columns because of high percentage of null values:

  MiscFeature: 96.3 %

  Fence: 80.75%

  PoolOC: 99.52%

  Alley: 93.77%

  FirePlaceQu: 47.26%

- **Normality**

  In my statistical report, I proceeded to examine the distribution of features, aiming to determine whether they conformed to a normal distribution. I employed two distinct tests, the Anderson-Darling and Shapiro-Wilk tests, to assess the normality of these feature distributions.

- **Shapiro-Wilk test:**

  The Shapiro-Wilk test serves as a statistical tool utilized to assess whether a given data sample conforms to a normal distribution. This test operates by measuring the disparities between the actual data distribution and the anticipated normal distribution. It calculates a test statistic by summing the squared deviations between observed and anticipated values, considering the expected variance.

In common practice, the Shapiro-Wilk test is employed to scrutinize the normality of data samples. The null hypothesis of this test posits that the data sample adheres to a normal distribution. When the p-value generated by the test falls below the chosen significance level, typically set at 0.05, it leads to the rejection of the null hypothesis. Consequently, the inference drawn is that the data sample deviates from a normal distribution.

For small to moderate sample sizes, the Shapiro-Wilk test exhibits greater statistical power compared to alternative normality tests, such as the Kolmogorov-Smirnov test. However, for larger sample sizes, it may exhibit reduced statistical power when juxtaposed with other testing methods.

The Shapiro-Wilk test was conducted to assess whether the data in the 'LotArea' column follows a Gaussian (normal) distribution.

The test statistic (W) is 0.351.

The p-value associated with the test is extremely low, approximately 0.000.

Based on the Shapiro-Wilk test results, we can make the following conclusions:

Statistic: The test statistic (0.351) is less than 1, which suggests a departure from a perfectly Gaussian distribution. A value close to 1 would indicate a closer fit to normal distribution.

p-value: The p-value (0.000) is significantly smaller than the common significance levels (e.g., 0.05). Therefore, we reject the null hypothesis (H0) that the data follows a Gaussian distribution.

The data in the 'LotArea' column does not adhere to a Gaussian distribution. Instead, it exhibits significant deviations from normality. This may have implications for the choice of statistical tests and assumptions when analyzing or modeling this data.

- **Anderson-Darling test:**

  The Anderson-Darling test is a statistical tool employed for assessing the compatibility between a provided data sample and a specified probability distribution. While its primary application is in testing data samples for conformity with a normal distribution, it can also be effectively applied to investigate adherence to various other distribution types. In this procedure, the test statistic is compared against critical values, which are determined based on the chosen significance level and the sample size. If the test statistic surpasses the critical value, this leads to the rejection of the null hypothesis, indicating that the sample does not originate from the proposed distribution.

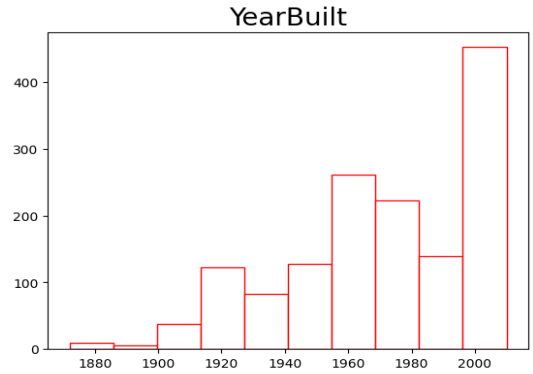
  The Anderson-Darling test i've performed on the 'OverallQual' data provides a statistic and a set of critical values to help assess whether the data follows a normal distribution.

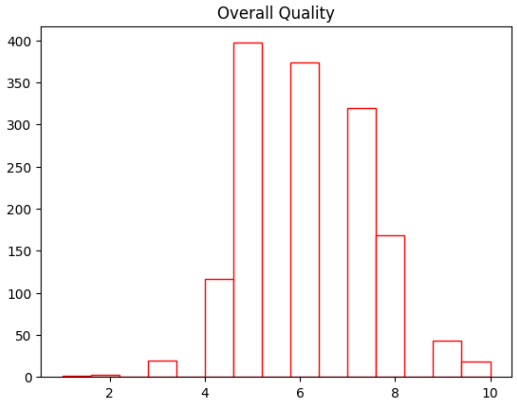
  The Anderson-Darling test was conducted on the 'OverallQual' data, resulting in a test statistic of 35.230. The critical values used for the test were [0.574, 0.654, 0.785, 0.916, 1.089]. The test statistic significantly exceeds the largest critical value, suggesting that the 'OverallQual' data does not follow a normal distribution. This indicates that the distribution of 'OverallQual' is not consistent with a normal distribution and exhibits significant deviation from normality.

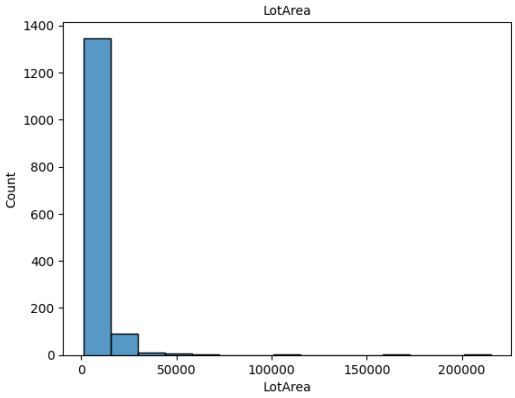
- **Visualization-check:**

  An alternative method for assessing the distribution of a feature involves visual examination, where you create a plot (such as a histogram) to represent the feature. While this approach is less precise compared to the aforementioned statistical tests, it serves as a valuable tool for gaining a visual understanding of the data's characteristics.

  To conduct the Anderson-Darling normality test, I focused on a select set of features, including Overall Quality, Bedroom above grade, and Year built. In the case of the Shapiro-Wilk test, I specifically chose to analyze Lot area and visualized its distribution through a histogram generated with seaborn. As anticipated, none of the features exhibited a normal distribution.





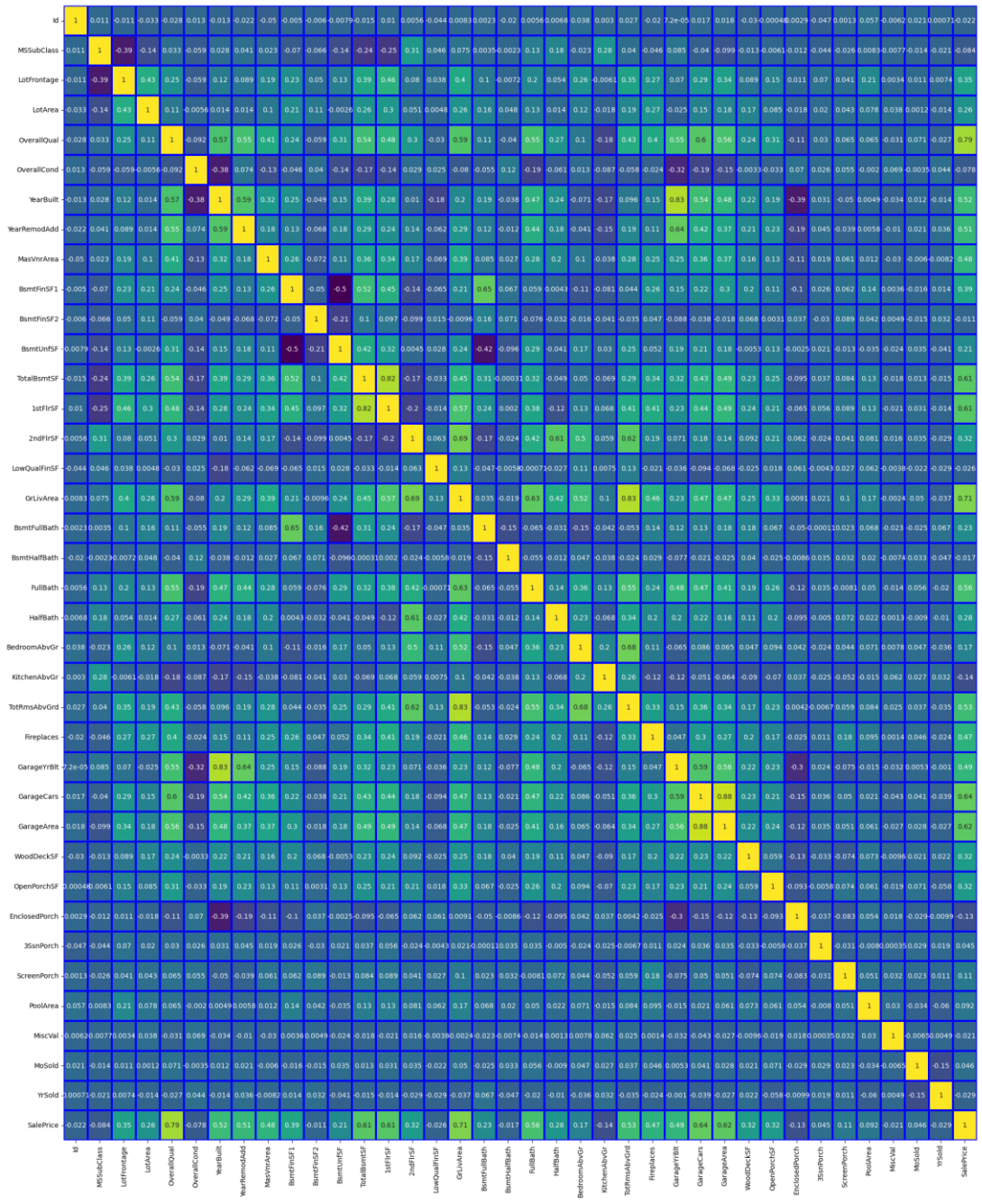


3.Correlation

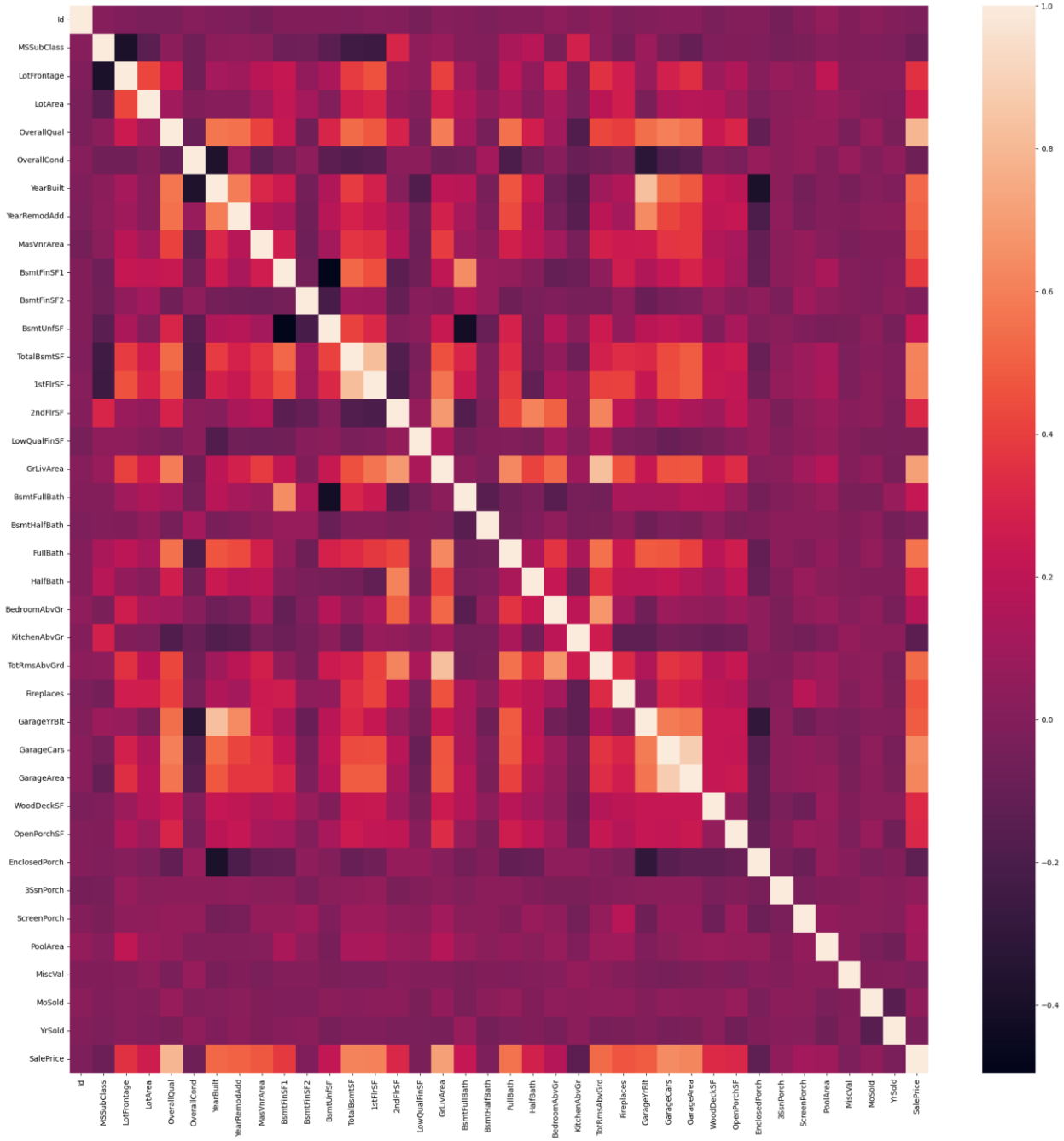
In this section, we employ correlation tests to ascertain the presence of any correlations among the features.

- **Correlation heatmap:**

  I generated a correlation heatmap specifically for numerical columns in the dataset. This heatmap is a visual representation of the relationships between multiple variables. It employs a color-coded matrix format to illustrate the degree of association between different variables. In this matrix, each variable is depicted by both a row and a column, while the cells display the correlation values between them. The intensity of the colors in these cells corresponds to the strength of the correlations, with darker hues representing stronger relationships.


Id
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
MiscVal
MoSold
YrSold
SalePrice
Correlation_matrix
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

And then another for all of the table's columns:



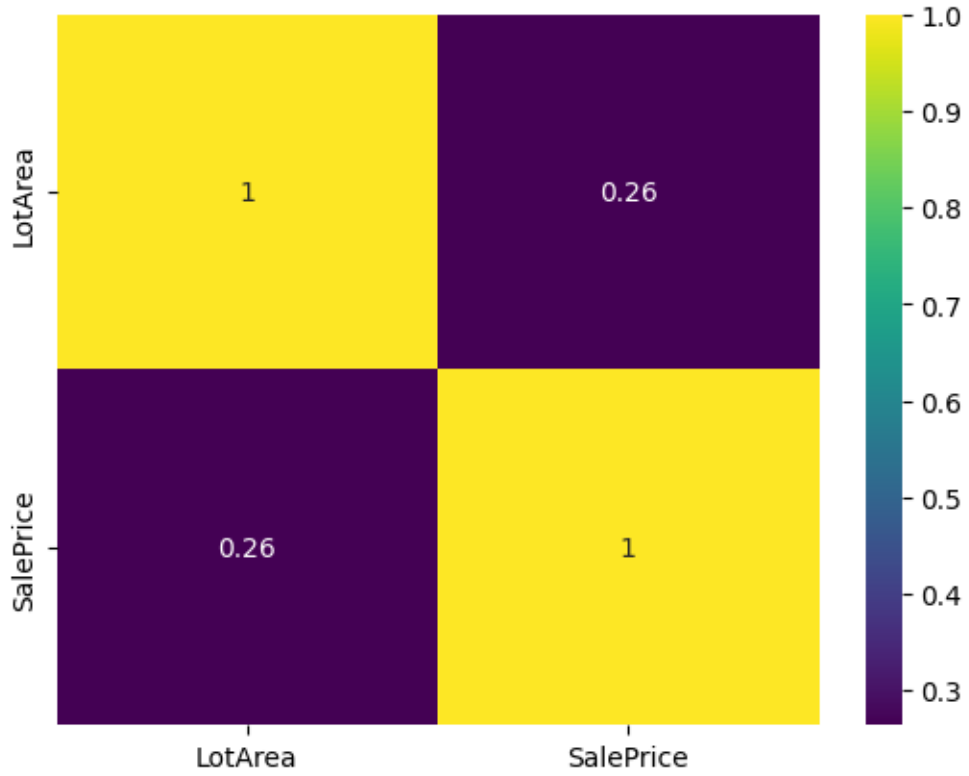
- **Correlation tests:**
  Correlation tests are employed in statistical analysis to assess the relationship between two variables and quantify the strength of that relationship. These tests can reveal whether the variables exhibit a positive or negative correlation or if they are entirely uncorrelated.

- **Pearson test:**

  In my statistical report, I utilized the Pearson correlation test to assess the extent and nature of the linear association between two continuous variables. This test quantifies the level of association between the variables and yields a correlation coefficient within the range of -1 to 1.

  Upon applying the Pearson correlation test to MSSubClass and Sale Price, the analysis revealed a relatively weak correlation between the two. Similarly, when extended to Lot Area and Sale Price, the results also displayed a modest level of correlation.

  Here's the heatmap for Sale Price and Lot Area:



- **Anova test:**

  The ANOVA test, which stands for Analysis of Variance, is employed in statistical analysis to ascertain the presence of statistically significant distinctions among the means of two or more groups. It accomplishes this by comparing the variance between the groups to the variance within the groups, enabling the determination of the significance of differences between the groups.

  In this study, a one-way ANOVA test was employed to investigate the degree of dissimilarity between Neighborhood and Sale Price. The results revealed a statistically significant difference, indicating that this disparity cannot be attributed to chance alone and, consequently, suggesting a correlation between these two variables.

  The purpose of this analysis is to determine whether there is a statistically significant difference in the mean sale prices based on different neighborhood values. We used the Analysis of Variance (ANOVA) test to achieve this goal

  The ANOVA test was conducted using the Python scipy library. The results of the test are as follows:

  F-statistic (f): 71.785

  p-value (p): 0.000

  Based on the results of the ANOVA test, we reject the null hypothesis (H0) as the p-value is extremely low ($p<0.05$). This suggests that there is a statistically significant difference in the mean sale prices between different neighborhoods. In other words, the neighborhood has a significant impact on sale prices.

  - The objective of this analysis is to determine whether there is a statistically significant difference in the mean sale prices of different Sale Conditions in our dataset.

  After conducting the ANOVA test, we obtained the following results:

  F-Value: 45.578

  p-Value: 0.000

Based on the results of the ANOVA test, we reject the null hypothesis (H0) as the p-Value is much less than the typical significance level (e.g., 0.05). This indicates that there is a statistically significant difference in the mean sale prices for different Sale Conditions.

- **T-test**

  A t-test serves as a statistical tool for comparing the means of two data groups, thereby assessing the significance of any differences between them. This test is commonly employed in hypothesis testing to ascertain whether a particular process or treatment has a measurable impact on the target population or whether two groups exhibit distinguishable characteristics.

  In the following section, I conducted a t-test analysis on Sale Price and 2ndFlrSF, which subsequently led to the determination of a significant relationship between these two variables.