Analysis of House Prices Dataset

Sarah Hosseini Feshtami - 400222026

Shahid Beheshti University

1. Data Exploration

The dataset available at [House Prices - Advanced Regression Techniques](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data) provides valuable information for conducting data exploration and analysis. This data exploration aims to provide insights into the dataset's structure, key variables, distributions, and potential relationships among variables.

The dataset consists of multiple variables related to residential properties. It includes both numerical and categorical variables, providing a comprehensive view of the various aspects influencing house prices. The dataset is divided into two separate CSV files: `train.csv` for training purposes and `test.csv` for testing or prediction purposes.

The dataset contains the following attributes:

- 1. Id: The unique identifier for each house.
- 2. MSSubClass: The building class of the house.
- 3. MSZoning: The general zoning classification of the property.
- 4. LotFrontage: Linear feet of street connected to the property.
- 5. LotArea: Lot size in square feet.
- 6. Street: Type of road access to the property.
- 7. Alley: Type of alley access to the property.
- 8. LotShape: General shape of the lot.
- 9. LandContour: Flatness of the property.
- 10. Utilities: Type of utilities available.
- 11. LotConfig: Lot configuration.

- 12. LandSlope: Slope of the property.
- 13. Neighborhood: Physical locations within the Ames city limits.
- 14. Condition1: Proximity to various conditions (e.g., main road or railroad).
- 15. Condition2: Proximity to various conditions (if more than one is present).
- 16. BldgType: Type of dwelling.
- 17. HouseStyle: Style of dwelling.
- 18. OverallQual: Overall material and finish quality of the house.
- 19. OverallCond: Overall condition rating of the house.
- 20. YearBuilt: Original construction date.
- 21. YearRemodAdd: Remodel date (same as construction date if no remodeling or additions).
- 22. RoofStyle: Type of roof.
- 23. RoofMatl: Roof material.
- 24. Exterior1st: Exterior covering on house.
- 25. Exterior2nd: Exterior covering on house (if more than one material).
- 26. MasVnrType: Masonry veneer type.
- 27. MasVnrArea: Masonry veneer area in square feet.
- 28. ExterQual: Exterior material quality.
- 29. ExterCond: Present condition of the exterior.
- 30. Foundation: Type of foundation.
- 31. BsmtQual: Height of the basement.
- 32. BsmtCond: General condition of the basement.

- 33. BsmtExposure: Walkout or garden level basement walls.
- 34. BsmtFinType1: Quality of basement finished area.
- 35. BsmtFinSF1: Type 1 finished square feet.
- 36. BsmtFinType2: Quality of second finished area (if present).
- 37. BsmtFinSF2: Type 2 finished square feet.
- 38. BsmtUnfSF: Unfinished square feet of basement area.
- 39. TotalBsmtSF: Total square feet of basement area.
- 40. Heating: Type of heating.
- 41. HeatingQC: Heating quality and condition.
- 42. CentralAir: Central air conditioning.
- 43. Electrical: Electrical system.
- 44. 1stFlrSF: First floor square feet.
- 45. 2ndFlrSF: Second floor square feet.
- 46. LowQualFinSF: Low quality finished square feet (all floors).
- 47. GrLivArea: Above ground living area square feet.
- 48. BsmtFullBath: Basement full bathrooms.
- 49. BsmtHalfBath: Basement half bathrooms.
- 50. FullBath: Full bathrooms above grade.
- 51. HalfBath: Half baths above grade.
- 52. BedroomAbvGr: Number of bedrooms above basement level.
- 53. KitchenAbvGr: Number of kitchens above grade.
- 54. KitchenQual: Kitchen quality.

- 55. TotRmsAbvGrd: Total rooms above grade (excluding bathrooms).
- 56. Functional: Home functionality rating.
- 57. Fireplaces: Number of fireplaces.
- 58. FireplaceQu: Fireplace quality.
- 59. GarageType: Garage location.
- 60. GarageYrBlt: Year garage was built.
- 61. GarageFinish: Interior finish of the garage.
- 62. GarageCars: Size of garage in car capacity.
- 63. GarageArea: Size of garage in square feet.
- 64. GarageQual: Garage quality.
- 65. GarageCond: Garage condition.
- 66. PavedDrive: Paved driveway.
- 67. WoodDeckSF: Wood deck area in square feet.
- 68. OpenPorchSF: Open porch area in square feet.
- 69. EnclosedPorch: Enclosed porch area in square feet.
- 70. 3SsnPorch: Three-season porch area in square feet.
- 71. ScreenPorch: Screen porch area in square feet.
- 72. PoolArea: Pool area in square feet.
- 73. PoolQC: Pool quality.
- 74. Fence: Fence quality.
- 75. MiscFeature: Miscellaneous feature not covered in other categories.
- 76. MiscVal: $Value of miscellaneous feature.

- 77. MoSold: Month sold.
- 78. YrSold: Year sold.
- 79. SaleType: Type of sale.
- 80. SaleCondition: Condition of sale.
- 81. SalePrice: Sale price (the target variable).

These attributes provide information about various aspects of the houses, such as their location, size, quality, condition, amenities, and sale-related details.

2. Exploratory Data Analysis (EDA):

Performing exploratory data analysis is crucial to gain a deeper understanding of the dataset. We calculate summary statistics such as count, mean, standard deviation, minimum, and maximum sale prices for each overall quality category using groupby() and describe(). This gives us an overview of the distribution of sale prices based on different overall quality levels.

To visualize the relationship between overall quality and sale price, we create a box plot using sns.boxplot() from seaborn. This plot shows the median, quartiles, and outliers of the sale prices for each category of overall quality. The x-axis represents the overall quality levels, and the y-axis represents the sale prices. Some key steps and techniques to consider during the EDA process include:

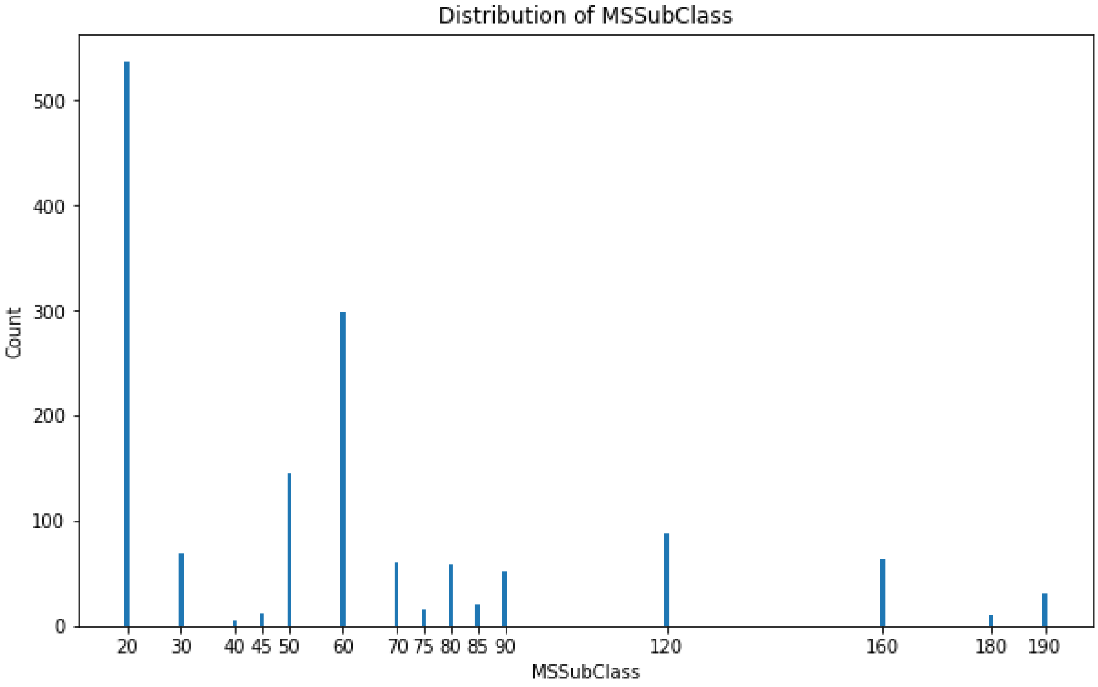
Data Cleaning: Check for missing values, outliers, and inconsistencies in the dataset. Handle missing data by imputation or elimination based on the specific context. Address outliers appropriately, considering their impact on the analysis.

Target Variable Analysis: Examine the distribution of the target variable (e.g., 'SalePrice') to understand its range, central tendency, and potential skewness. Consider visualizations such as histograms, box plots, or density plots to analyze the target variable's distribution.

Variable Relationships: Explore the relationships between the target variable and other predictor variables. Utilize scatter plots, correlation matrices, or heatmaps to identify potential correlations and dependencies. Assess the strength and direction of these relationships to gain insights into the factors influencing house prices.

3. Exploratory Visualization

The following plot demonstrates the number of total incidents from each building class category (MSSubClass) in the dataset. As can be seen, the count of “20” and “60” are much higher than other types. Also, “40”, “45”, and “180” have the lowest count.



Based on the information found in ‘data_description.txt’, these numbers each represent a type of building:

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

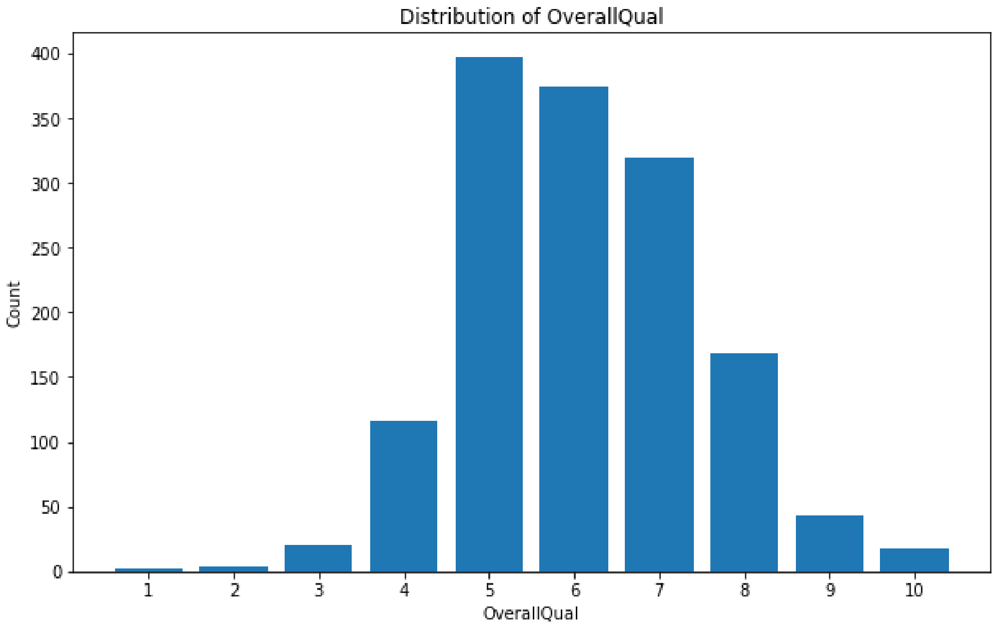
We analyze the importance of the "MSSubClass" feature by grouping the dataset by building class and calculating the median and mean sale price for each class using the groupby(), mean()

and median() methods. The resulting median and mean sale prices are sorted in descending order to identify any notable differences between the building classes.

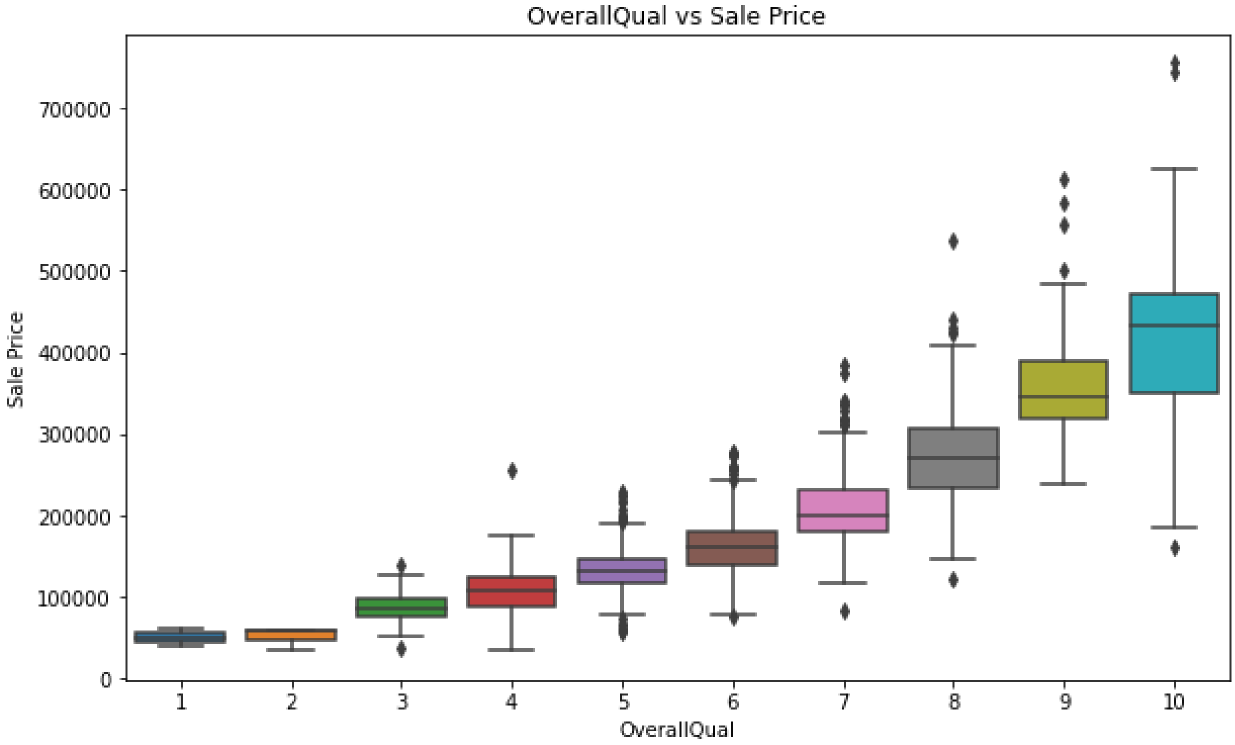
By examining the distribution, mean, and median sale prices of different building classes, we can gain insights into the importance of the "MSSubClass" feature. It helps us understand how the architectural style or type of dwelling influences the prices of houses in the Ames housing market.

We can infer that both mean and median values are highest when "MSSubClassh" equals 60 or 120 and lowest when it equals 180 or 30. This means "2-STORY 1946 & NEWER" or "1-STORY PUD (Planned Unit Development) - 1946 & NEWER" are more valuable but "PUD - MULTILEVEL - INCL SPLIT LEV/FOYER" and "1-STORY 1945 & OLDER" are less.

Another bar plot that can be insightful is the one that follows. As can be seen, the distribution of data based on overall quality looks like it can be a normal distribution. We later perform a statistical test to reject or validate this claim.



To visualize the relationship between overall quality and sale price, we create a box plot using sns.boxplot() from seaborn. This plot shows the median, quartiles, and outliers of the sale prices for each category of overall quality. The x-axis represents the overall quality levels, and the y-axis represents the sale prices.



This information can be valuable for various purposes, such as real estate market analysis, property valuation, and predicting house prices based on building class characteristics.

4. Statistical Tests and Analysis

4.1. Normality test on the distribution of overall quality

We perform an analysis of the normality of a variable using a normality test. The variable under consideration is 'OverallQual' from the given DataFrame. The normality test used in this analysis is the normaltest from the scipy.stats module.

The normaltest calculates a statistic and p-value to assess whether the variable follows a Gaussian (normal) distribution. In this analysis, the normaltest was conducted on the 'OverallQual' variable, resulting in a statistic value of 11.9 and a p-value of 0.003.

To interpret the results, the p-value is compared to a significance level (commonly chosen as 0.05). If the p-value is greater than the significance level, it suggests that the variable is likely to follow a Gaussian distribution. On the other hand, if the p-value is less than the significance level, it indicates that the variable is probably not Gaussian.

For the given dataset, the normaltest resulted in a p-value of 0.003. Since the p-value is greater than the significance level, we cannot conclude that the 'OverallQual' variable is normally distributed.

It is important to note that deviations from normality may affect the validity of certain statistical tests and assumptions. Thus, it is essential to consider the normality of the variable when selecting appropriate statistical techniques and interpreting the results.

4.2. Testing for the effect of different heating types on sale prices using an ANOVA

This is an analysis of the effect of different heating types on sale prices using an ANOVA (Analysis of Variance) test. To begin the analysis, a box plot was created to visualize the distribution of sale prices across different heating types. The box plot provides insights into the central tendency, variability, and potential outliers within each heating category.

Following the exploratory visualization, an ANOVA model was fitted to assess whether the heating types have a significant impact on sale prices. The ANOVA model was constructed using the `ols` function from the `statsmodels.formula.api` module. The formula 'SalePrice ~ C(Heating)' specifies that the sale price is the dependent variable, while 'Heating' is treated as a categorical independent variable.

The ANOVA test was then performed using the `anova_lm` function from the `statsmodels.api` module. This test generates an ANOVA table, which provides statistical metrics such as the sum of squares, degrees of freedom, F-statistic, and associated p-values. These metrics allow us to evaluate the significance of the differences in means among the heating types.

The sum of squares (sum_sq) for the variable "Heating" is 1.329359e+11, indicating the amount of variability explained by the different types of heating.

The degrees of freedom (df) for "Heating" is 5, which represents the number of categories within the variable.

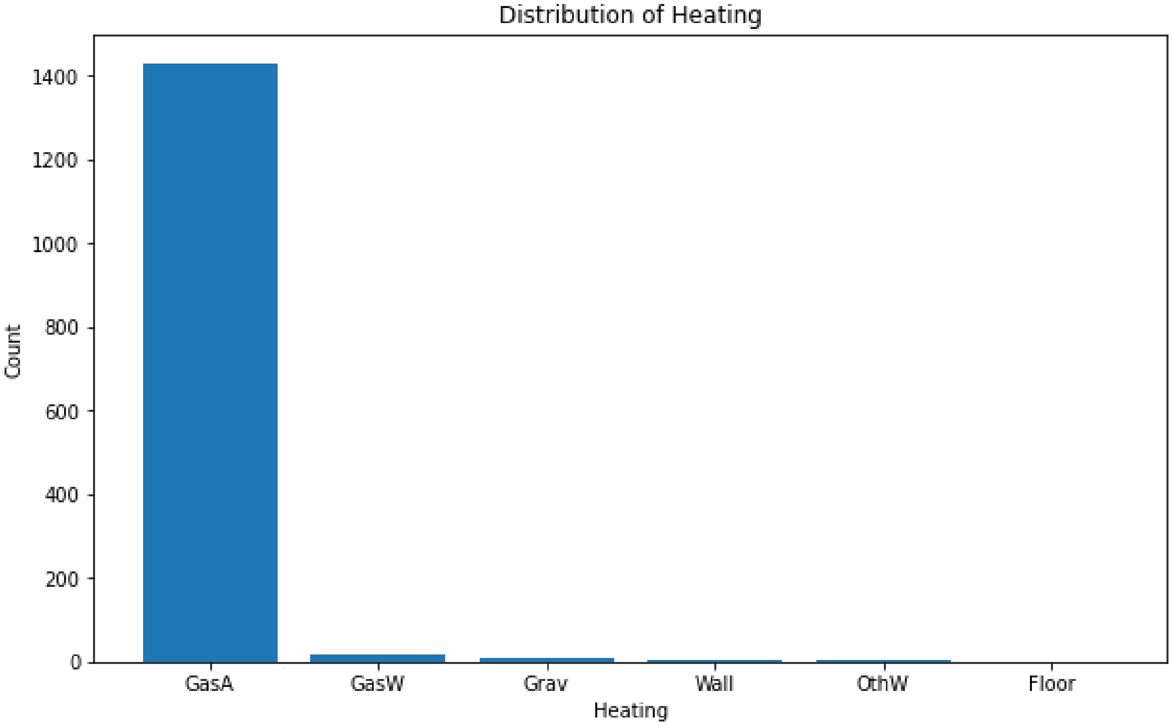
The F-statistic (F) is 4.259819, which is the ratio of between-group variability to within-group variability. It measures the significance of the differences among the means of the different heating types.

The p-value (PR(>F)) for "Heating" is 0.000753, which is below the typical significance level of 0.05. This indicates that there is strong evidence to reject the null hypothesis and conclude that there is a statistically significant effect of the different types of heating on sale prices.

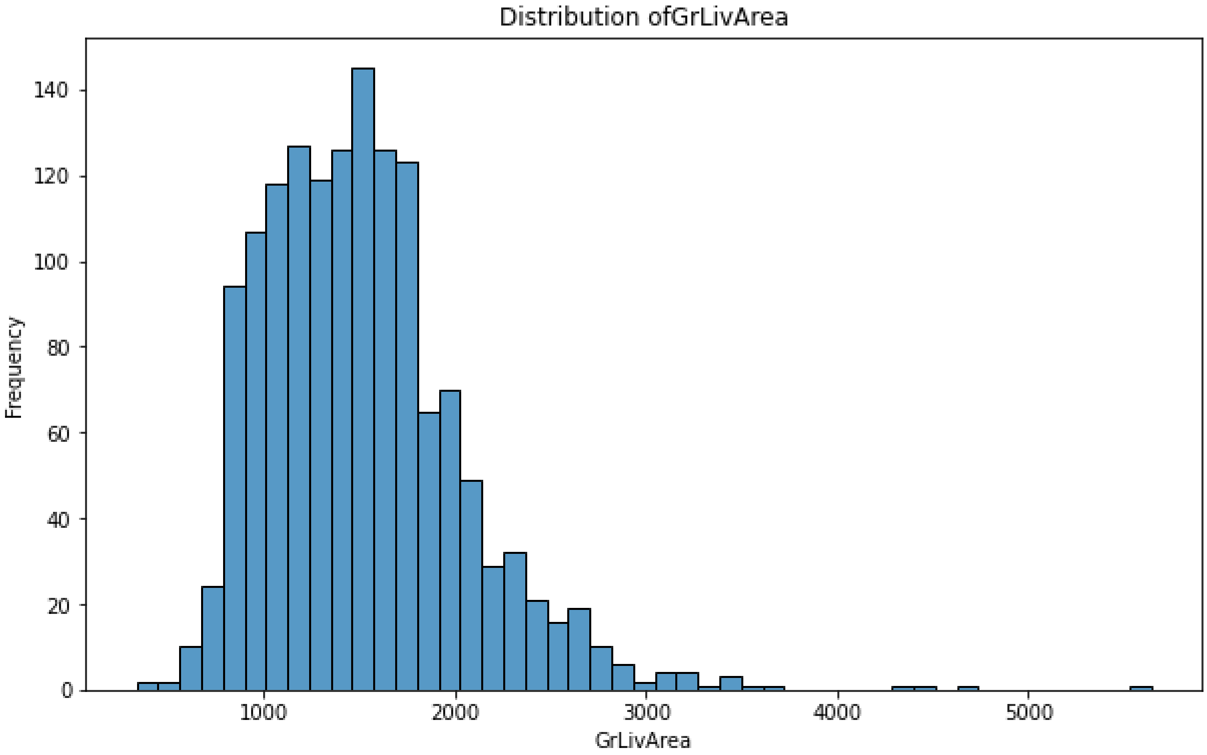
The sum of squares (sum_sq) and degrees of freedom (df) for the residual term (Residual) represent the variability that is not explained by the "Heating" variable.

In summary, the ANOVA test suggests that there is a significant effect of the different types of heating on sale prices in the housing dataset we analyzed.

However, if we look at the bar plot of the distribution of data based on hating, we get the following plot. This shows that most of the data is in the 'GasA' category, making the data heavily imbalance. So the test can be unreliable.



4.3. Testing for relationship between the sale price of properties and their corresponding Above grade (ground) living area square feet



In this analysis, we investigated the potential relationship between the sale price of properties and their corresponding Above grade (ground) living area square feet.

To assess the strength and significance of the relationship, we employed the Pearson correlation coefficient and conducted a hypothesis test. The Pearson correlation coefficient measures the linear association between two variables, ranging from -1 to 1, where values close to 1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and values close to 0 suggest a weak or no linear correlation.

Our analysis revealed a correlation coefficient of 0.70 between the sale price and GrLivArea. The corresponding p-value was calculated as e−223, which is commonly used to determine the statistical significance of the relationship.

Based on our hypothesis test, we found that the p-value is less than the significance level of 0.05. Therefore, we reject the null hypothesis that the sale price and GrLivArea are independent. Consequently, we conclude that there is likely a significant relationship between the sale price and GrLivArea.

4.4. the correlation of overall condition and construction date

This is an analysis of the potential relationship between the year a property was built and its overall condition. To ensure reliable results, any rows containing missing values in either of these columns were removed from the analysis.

Based on our analysis, we obtained a correlation coefficient of -0.37 between the year built and the overall condition of the properties. The associated p-value was calculated as e-50. The p-value is commonly used to determine the statistical significance of a relationship.

Upon conducting a hypothesis test, we found that the p-value was less than the significance level of 0.05. Therefore, we reject the null hypothesis that the year built and overall condition are independent. This indicates that there is likely a significant relationship between the year a property was built and its overall condition.

4.5. testing whether the prices of houses in different neighborhoods are significantly different

This is an analysis of the differences in means among different neighborhoods based on prices. The DataFrame was grouped based on the 'Neighborhood' column, and the mean values of the SalePrice were compared among the groups using an ANOVA (Analysis of Variance) test.

The ANOVA test allows us to determine if there are significant differences in means among the groups. In this analysis, the F-statistic and corresponding p-value were calculated using the `f_oneway()` function from the `scipy.stats` module.

For the given dataset, the ANOVA test resulted in an F-statistic of 71.78 and a p-value of e-225. The F-statistic measures the ratio of between-group variability to within-group variability, while the p-value indicates the statistical significance of the observed differences in means.

Based on the obtained results, if the p-value is less than the chosen significance level (usually 0.05), it suggests that there are significant differences in means among the groups. Conversely, if the p-value is greater than the significance level, it indicates that there is insufficient evidence to conclude that the means are different.

In this analysis, the obtained p-value is less than the significance level. Therefore, we reject the null hypothesis that the means of the target variable are the same among the groups. This implies that there are likely significant differences in means among the groups.

It is important to note that the ANOVA test assumes certain assumptions, such as the normality of the data and equal variances among the groups. Violations of these assumptions may impact the validity of the results. Therefore, it is recommended to carefully assess the assumptions and consider additional analyses if necessary.

5. Results

The following list summarizes the key and valuable conclusions derived from the aforementioned processes:

- The distribution of overall quality is probably not Gaussian.
- There is likely a significant relationship between the sale price and GrLivArea. A positive coefficient means this relationship is positive.
- There is likely a significant relationship between the year a property was built and its overall condition. A negative correlation coefficient means that the relationship is inverse or negative.
- the prices of houses in different neighborhoods are significantly different.