Statistical Analysis Of House Prices Data

Mohammad Rouintan

Computer Science, Shahid Beheshti University

Abstract

My work presents a comprehensive analysis of house price data using Python. The dataset consists of 1460 samples and encompasses 80 features that capture various aspects of a house, potentially influencing its sale price. The objective of this study is to perform exploratory data analysis and apply statistical tests, such as t-tests, ANOVA tests, and correlation analysis, to uncover significant relationships and insights within the data. The results of this analysis aim to provide valuable information and understanding for stakeholders in the real estate industry.

Introduction

The housing market is a complex and dynamic system influenced by numerous factors. Understanding the key determinants of house prices is crucial for various stakeholders, including homeowners, real estate agents, and policymakers. In this report, we present a comprehensive analysis of a house price dataset comprising 1460 samples and 80 features. Our primary objective is to explore the relationships between these features and the sale price, which serves as the target variable.

The dataset encompasses a wide range of features that capture various aspects of a property, including its physical characteristics, location, amenities, and quality. These features include variables such as the building class, zoning classification, lot size, street type, and overall material and finish quality. Additionally, the dataset includes information about the year of construction, remodeling, roof type and material, exterior covering, basement characteristics, heating and cooling systems, number of rooms and bathrooms, garage features, and presence of additional amenities such as pools or fences.

To gain insights into the factors influencing house prices, we will employ various statistical tests and exploratory data analysis techniques. Statistical tests, including t-tests and ANOVA, will help us assess the significance of individual features on the sale price. We will also conduct correlation tests to examine the strength and direction of relationships between pairs of variables. Exploratory data analysis techniques, such as data visualization and descriptive statistics, will provide a deeper understanding of the dataset's characteristics and reveal any underlying patterns or trends.

By conducting this analysis, we aim to identify the key determinants of house prices and provide valuable insights for homeowners, real estate professionals, and policymakers. The findings from this study will enhance our understanding of the dynamics of the housing market and have practical implications for pricing strategies, investment decisions, and urban planning.

Methodologies

Handling Missing Values and Inefficient Features

Feature ID is not useful for us, so we delete it. Some features of the dataset have a large number of missing values compared to the total number of samples, and they cannot be analyzed and checked, so we can remove them. There are other features that have missing values, but we can analyze them and there is no need to remove them.

Hypothesis Testing

For all the statistical tests used, the alpha value is equal to 0.05, which is compared with the p-value. If the p-value is greater than alpha, the null hypothesis is accepted, otherwise it is rejected.

Question1: Average of SalePrice is 175000 (one sample T-test)

$$H_0 : \mu = 175000 \tag{1}$$

$$H_1 : \mu \neq 175000 \tag{2}$$

accept null hypothesis

Question2: Average of SalePrice in Pave street and Grvl street is equal (2 sample independent t-test)

$$H_0 : \mu_{Pave} = \mu_{Grvl} \tag{3}$$

$$H_1 : \mu_{Pave} \neq \mu_{Grvl} \tag{4}$$

accept null hypothesis

Question3: Roof Style effects on SalePrice (ANOVA)

$$H_0 : \mu_{Gable} = \mu_{Flat} = \mu_{Mansard} \quad (5)$$

$$H_1 : \mu_{Gable} \neq \mu_{Flat} \neq \mu_{Mansard} \quad (6)$$

accept null hypothesis

Question4: Street as categorical feature and RoofStyle as categorical feature (chi2 Test) Retain H0,There is no relationship between 2 categorical variables

Question5: What are The building class and why it is important? We divide the MSSubClass into three ranges in this way, first we sort the unique feature values and divide them into three groups of 5, that is, the first 5, the second 5, and the third 5.

$$H_0 : \mu_{Range1} = \mu_{Range2} = \mu_{Range3} \quad (7)$$

$$H_1 : \mu_{Range1} \neq \mu_{Range2} \neq \mu_{Range3} \quad (8)$$

reject null hypothesis

Question6: How does the overall quality (OverallQual) of a house relate to its sale price? This time, we act like the method of the previous section and divide it into two ranges.

$$H_0 : \mu_{Range1} = \mu_{Range2} \quad (9)$$

$$H_1 : \mu_{Range1} \neq \mu_{Range2} \quad (10)$$

reject null hypothesis

Question7: How do the different types of heating (Heating) affect the sale prices?

$$H_0 : \mu_{GasA} = \mu_{GasW} = \mu_{Grav} = \mu_{Wall} = \mu_{OthW} = \mu_{Floor} \quad (11)$$

$$H_1 : \mu_{GasA} \neq \mu_{GasW} \neq \mu_{Grav} \neq \mu_{Wall} \neq \mu_{OthW} \quad (12)$$

reject null hypothesis

Question8: How do the different types of utilities (Utilities) available in a property relate to sale prices?

$$H_0 : \mu_{AllPub} = \mu_{NoSeWa} \quad (13)$$

$$H_1 : \mu_{AllPub} \neq \mu_{NoSeWa} \quad (14)$$

accept null hypothesis

Exploratory Data Analysis (EDA)

In this section, we have used different charts for further analysis These charts include the charts that show the distribution of data, the count plot, the violin plot and the scatter plot, which we provide sufficient explanations below.

Show The Distribution Of Data In Some Feature:

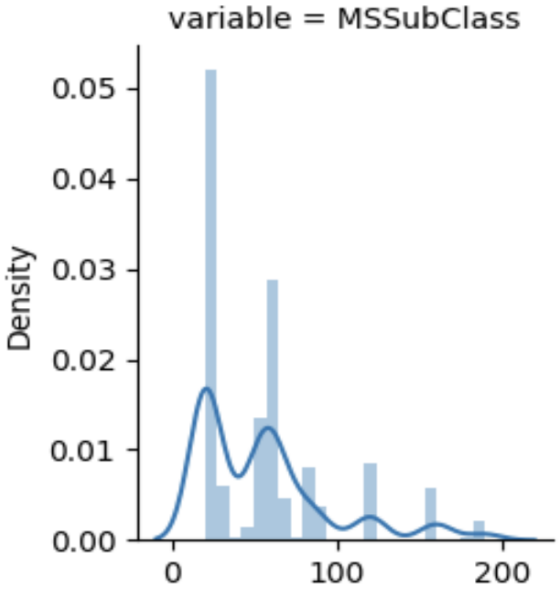

Figure 1: *The Distribution Of MSSubClass Feature : The building class. Result: Most building classes are between 20 and 90*

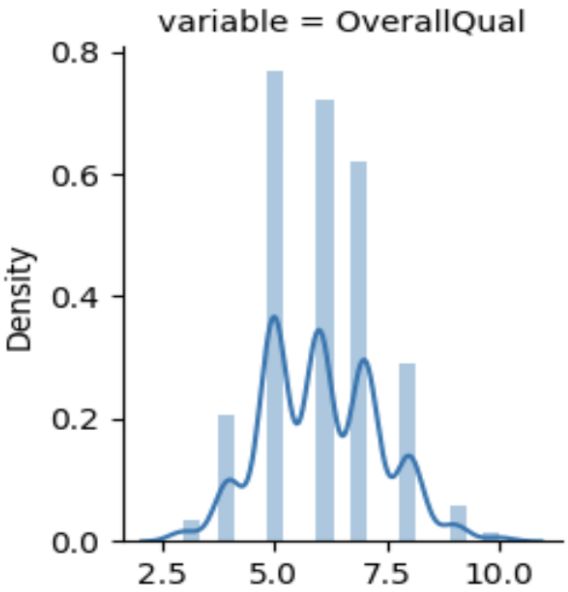

Figure 2: *The Distribution Of OverallQual Feature : Overall material and finish quality. Result: Most OverallQuals are between 5 and 7.5*

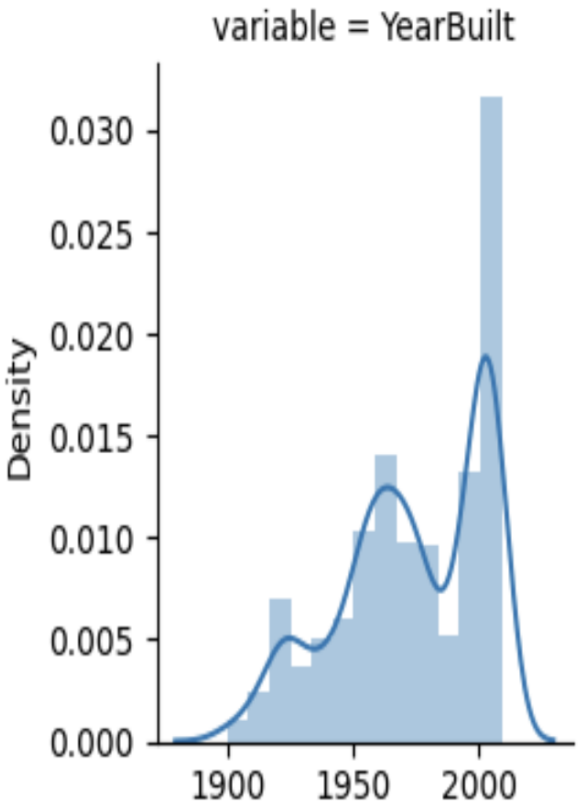

Figure 3: *The Distribution Of YearBuilt Feature : Original construction date. Result: Most YearBuilts are between 1950 and 2000*

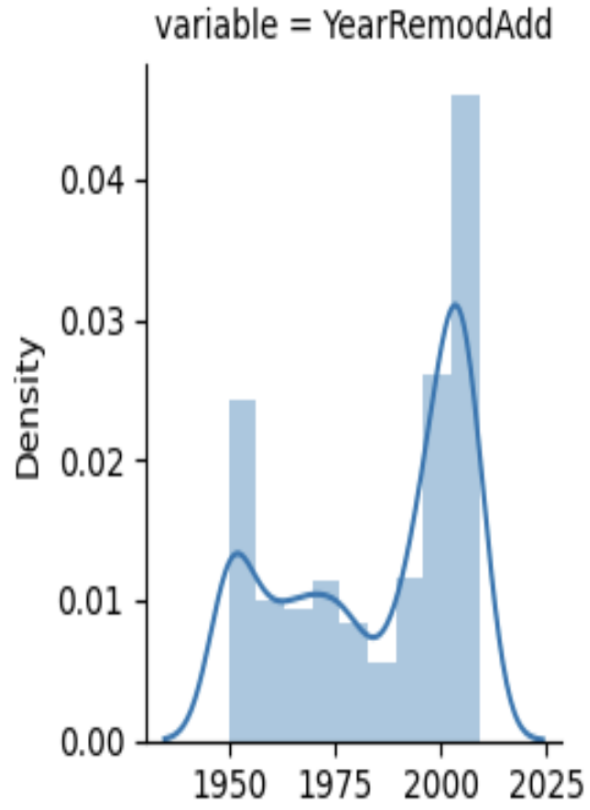

Figure 4: *The Distribution Of YearRemodAdd Feature : Remodel date. Result: Most YearRemodAdds are in 1950 and 2000*

Show The CountPlot And ViolinPlot Of Data In Some Feature:

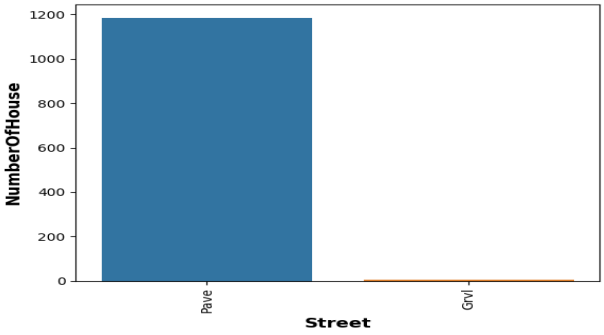

Figure 5: *The CountPlot Of Street Feature. Result: Almost all the streets of the houses are paved.*

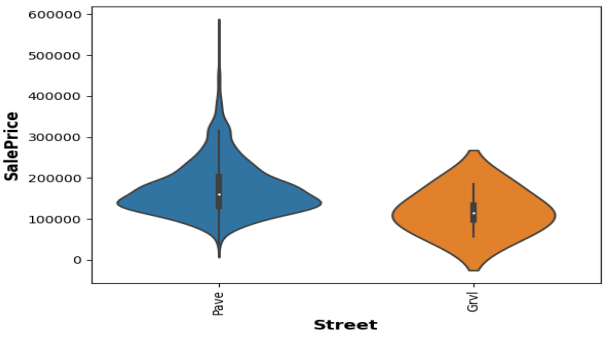

Figure 6: *The ViolinPlot Of Street Feature. Show distribution of SalePrice based on type of street*

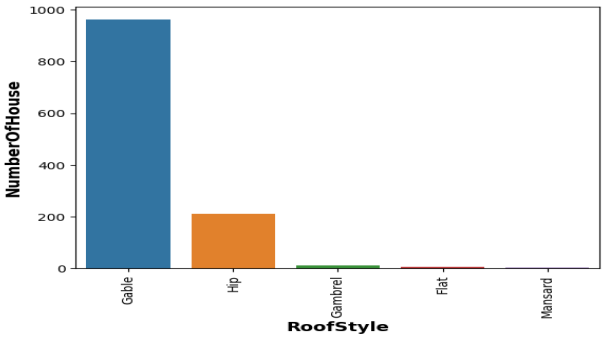

Figure 7: *The CountPlot Of RoofStyle Feature. Result: The roof style of most houses is gable and hip*

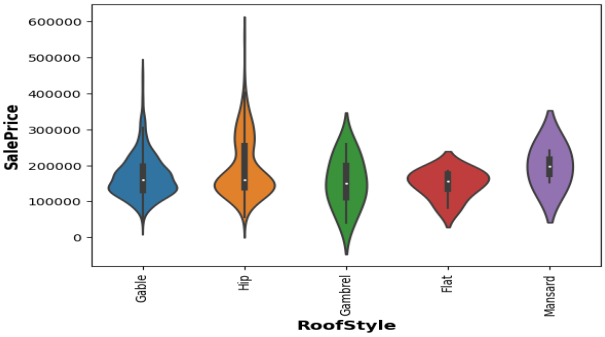

Figure 8: *The ViolinPlot Of RoofStyle Feature. Show distribution of SalePrice based on roof style*

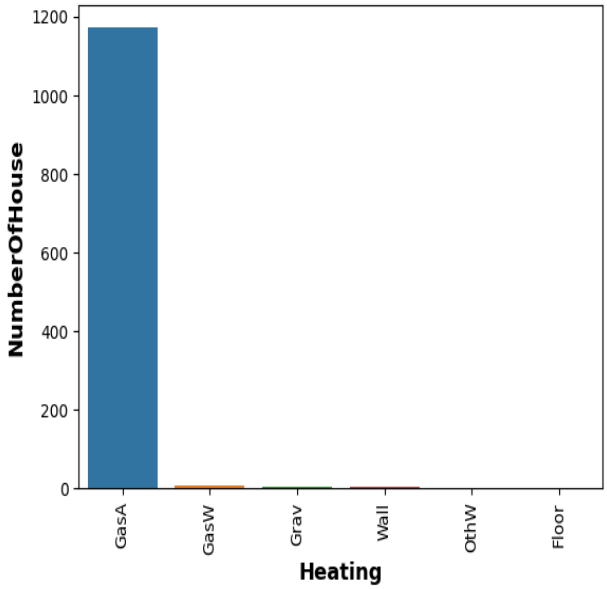

Figure 9: *The CountPlot Of Heating Feature. Result: Almost all the heating system of the houses are GasA.*

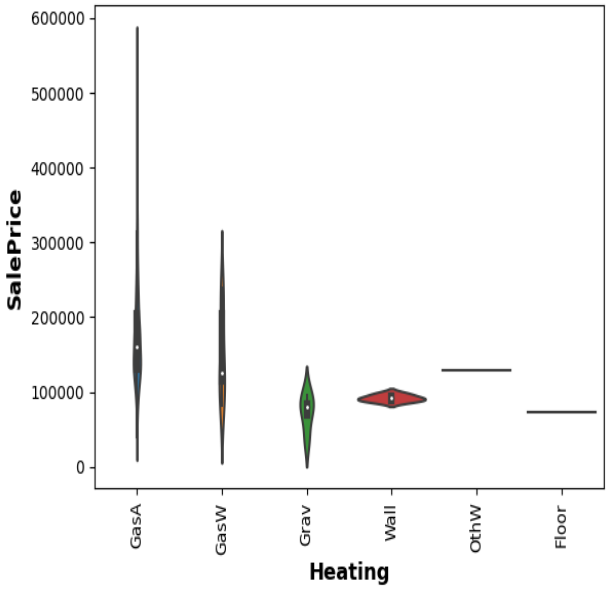

Figure 10: *The ViolinPlot Of Heating Feature. Show distribution of SalePrice based on type of heating*

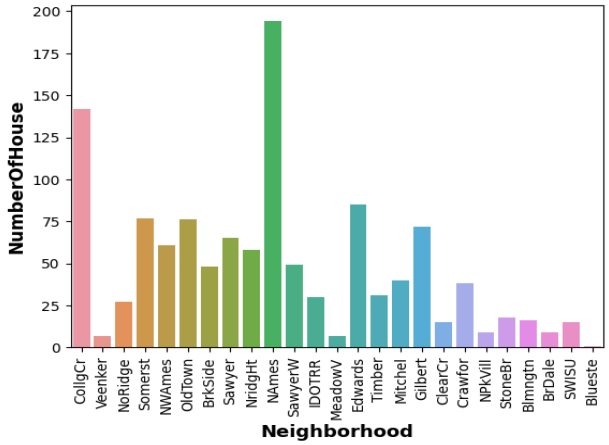

Figure 11: *The CountPlot Of Neighborhood Feature. Result: Top 3 Neighborhood is NAmes, CollgCr, Edwards*

Show The ScatterPlot Of Data In Some Feature:

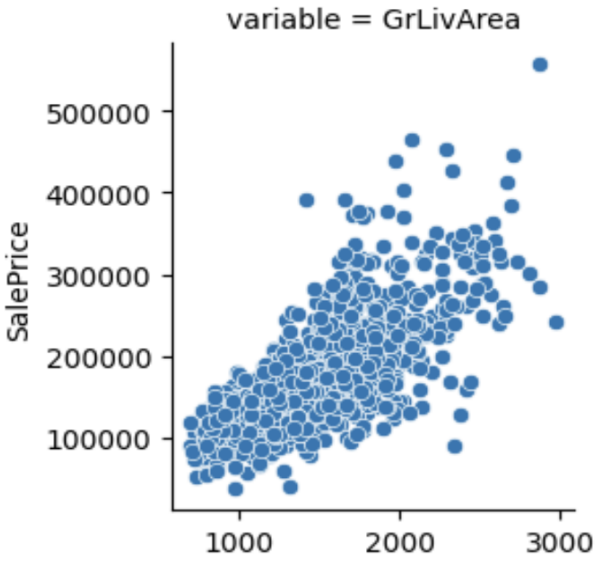

Figure 12: *The ScatterPlot Of GrLivArea Feature : Above grade (ground) living area square feet.*

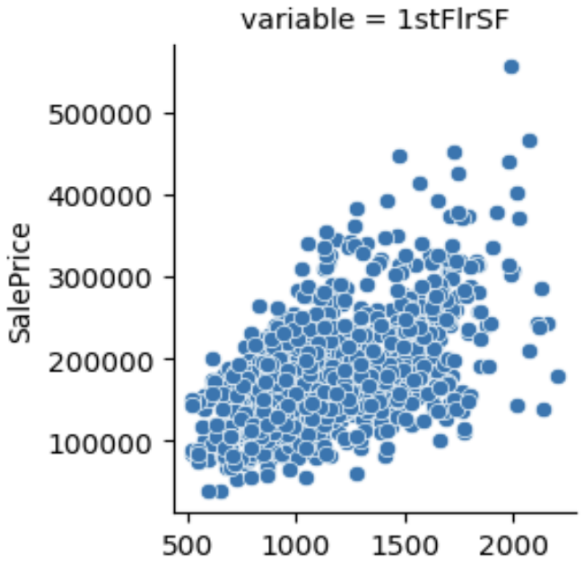

Figure 13: *The ScatterPlot Of 1stFlrSF Feature : First Floor square feet.*

Results

The main goal of this project is to learn how to use statistical tests and exploratory data analysis to analyze the problem as much as possible. Now, the dataset selected for this work is HousePrice dataset, which is about the price of houses according to the different characteristics of the houses.

But finally, by using these techniques, we can partially understand the relationships and behavior between features so that we can work as an analyst in

this field. To achieve better results using this data, we can go one step further and use machine learning models such as regression or neural network models such as MLP to estimate house prices based on features.