Analysis of Ames Housing dataset

Reza Mousavi

Computer Science Faculty, Shahid Beheshti University

27 October 2023

A B S T R A C T

The Ames Housing dataset, the focus of this report, is a playground competition designed for individuals with some experience in R or Python and a basic understanding of machine learning. It challenges participants to predict the final selling prices of residential properties in Ames, Iowa, using 79 explanatory variables that cover various aspects of the homes.

The competition highlights that factors beyond the obvious, such as the number of bedrooms or the presence of a white-picket fence, play a crucial role in determining property prices. This report introduces the key concepts of creative feature engineering and advanced regression techniques, such as random forest and gradient boosting, which are instrumental in addressing this prediction task.

The dataset itself was curated by Dean De Cock and is a valuable resource for data science education. It serves as a more comprehensive and modernized alternative to the well-known Boston Housing dataset, making it an attractive choice for data scientists looking to expand their skill set. The competition provides an opportunity for data science students to apply their knowledge to a real-world problem, bridging the gap between theory and practice.

In summary, this report introduces the Ames Housing dataset competition, emphasizing its relevance for skill development in data science. By understanding the dataset and the advanced techniques involved, participants can enhance their data science proficiency and successfully participate in this competition.

1. Introduction

The Ames Housing dataset comprises around 1500 rows of data, encompassing 81 columns, each representing various aspects of residential properties. The aim of this project is to delve into this extensive dataset, employing statistical analysis to gain a deeper understanding of the underlying patterns and relationships.

Our analysis begins by assessing the extent of missing data in each column and subsequently addressing these gaps through appropriate methods. This crucial data preprocessing step ensures the completeness and integrity of the dataset, laying the foundation for our exploratory journey.

With a comprehensive grasp of the data in place, our objective is to uncover valuable insights and craft compelling narratives. This involves posing and answering a series of pertinent questions, supported by a battery of statistical tests. By rigorously examining the Ames Housing dataset, we seek to unearth the stories concealed within its columns, shedding light on the intricate dynamics of the real estate space.

2. Exploratory Data Analysis

first, for data analysis we should read description of dataset and know features of dataset in detail. This dataset contains three csv files:

train.csv: Most of the data is in this file. This file is for learning machine learning models. Our analysis and review will be based on this file.

test.csv: This file is for testing machine learning models.

$sample_submission.csv$: This file contains less data than the dataset. It is intended for understanding the dataset's structure and includes examples of data.

2.0..1. Missing data

In this section, we will address missing data within the dataset. Our general approach involves filling missing values in categorical columns with the median, while for numerical columns, we use the median as well. However, in the case of categorical data, we will only fill in the missing values if the count of missing entries is less than 100. This criterion is applied to maintain the integrity of the main data distribution. For numerical data, if the count of missing values exceeds 100, we fill the gaps with the mean.

Certain columns within the dataset contain more than 100 missing values. Since we lack knowledge of the actual data distribution and these columns represent over half of the dataset, we have decided to exclude them from our analysis.

Additionally, the 'id' column, which serves solely as a counter and does not contribute valuable information, will also be removed from the dataset.These columns are deleted with this amount of missing data:

© 0000 Tellus

[H]

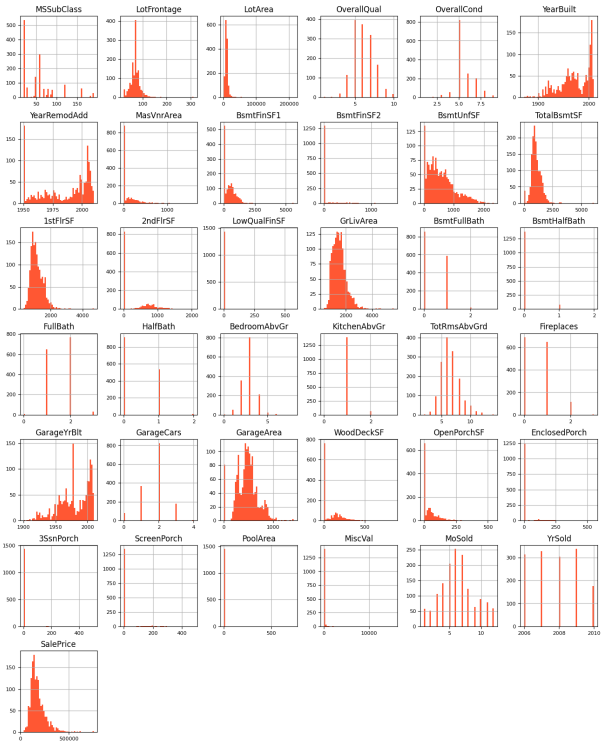

Fig. 1. Numerical data histogram plot.

Alley: 1369
MasVnrType: 872
FireplaceQu: 690
PoolQC: 1453
Fence: 1179
MiscFeature: 1406

2.0..2. Numerical columns

In this section, we will focus on the analysis of numerical columns. To initiate our exploration, we will begin by examining the following diagram. 1

Initially, it is evident that a significant surge in housing construction and renovations occurred in proximity to the year 2000. This suggests that a substantial portion of the housing stock in this area was established during that period.

What's particularly noteworthy is that despite the predominant construction around 2000, a considerable number of renovations and repairs were also concentrated in these years. This implies a high demand for housing during this period, prompting the question: Does the rise in property prices correlate with the increased construction and renovation activity observed around the same time? Is this price escalation influenced by population growth and increased housing demand in this area?"

[H]

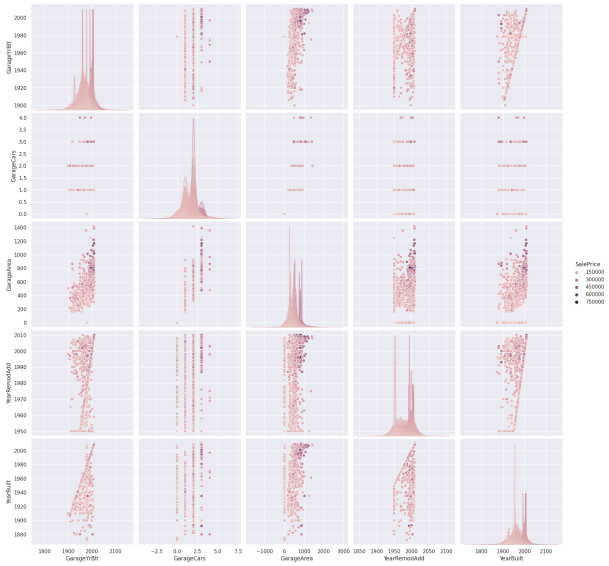

Fig. 2. Examining some columns of Numerical regarding price changes

This rephrased version maintains the flow of your ideas while enhancing the clarity of your questions and observations.

Furthermore, the plot provides valuable insights into the distribution of bedroom counts and the availability of parking spaces within the dataset. Notably, a substantial proportion of houses feature two bedrooms and two parking spaces, while various other distributions for these features are also evident.

Now let's examine these values in relation to price changes :2

Based on the chart, it's evident that houses constructed around the year 2000 and beyond tend to have larger garages and higher prices. This relationship between garage size, parking availability, and house prices is unmistakable. In general, the 20th century, spanning from 1900 to 2000, witnessed the highest rate of construction and house renovations, with the most substantial price increases occurring between 1990 and 2010.

An intriguing trend is the expansion of garage areas during renovations, which appears to contribute to the overall increase in house prices. This demonstrates the interplay between garage size and property value during remodeling and highlights the significance of garage space in the housing market.

It is evident that an increase in GrLivArea, particularly after the 1990s renovations, corresponds to higher house prices. Houses constructed from 1990 onwards exhibit a consistent trend: as GrLivArea increases, so does the house price, underscoring the significance of GrLivArea for homebuyers during this period.

Overall, there is a noticeable consistency in house sizes and dimensions during these years.3

[H]

Fig. 3. Examining the relationships of some numerical columns based on price changes

[H]

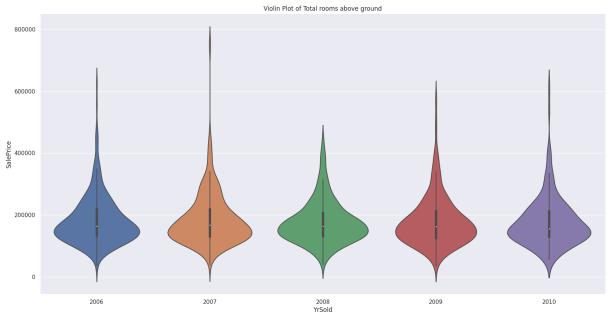

Fig. 4. price in different years

2.0..3. *Categorical columns*

Let's begin by examining house prices across different years:4

The distribution of sales quantities across different years reveals a consistent pattern with nearly equal amounts.

Now, let's assess house prices by considering different months:5

The prices show minimal variation across different months, with consistent sales volumes and distributions. However, it's worth noting that prices tend to experience substantial increases at the beginning and middle of the year.6

Observationally, houses that feature a kitchen tend to be associated with higher prices.

[H]

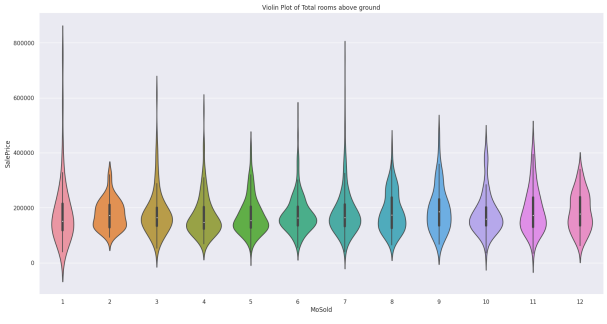

Fig. 5. price in different month

[H]

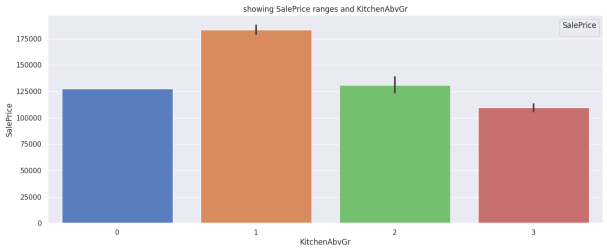

Fig. 6. Number of kitchens

Next, we will analyze the relationship between the number of rooms in a house and its price:7

Now, let's explore the relationship between the number of parking spaces and the price of a house:8

It is clear that the number of houses with 3 parking spaces is less, but they have the highest price.

Finally, we'll examine house prices based on the type of street.9

Clearly, house prices vary depending on the type of street,

[H]

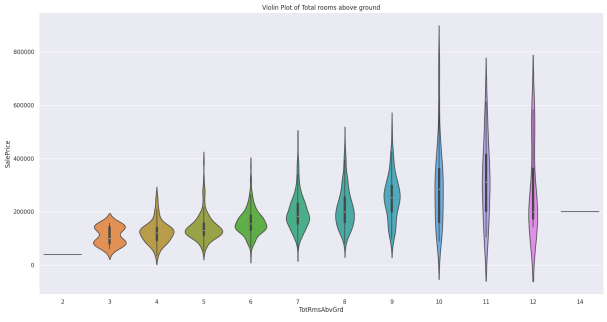

Fig. 7. Number of rooms based on price

[H]

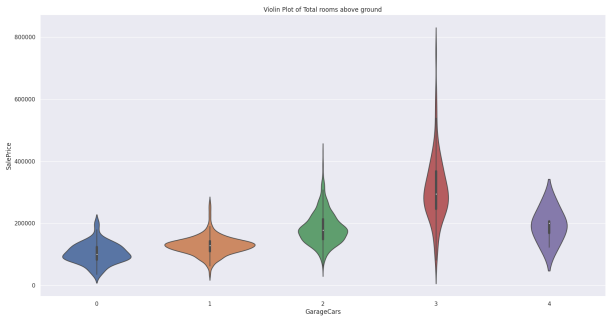

Fig. 8. The number of parking spaces for a house based on price

[H]

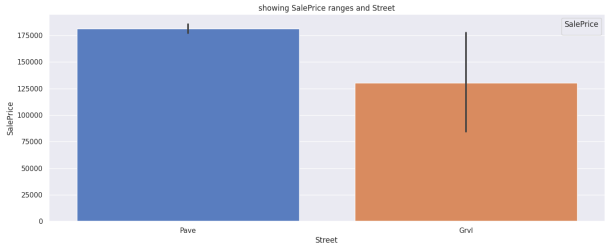

Fig. 9. House price based on street type

especially between the two street types.In this section, we conducted a general overview of the various column types and their correlation with house prices. Having familiarized ourselves with the dataset, we will now proceed to address specific questions through hypothesis testing.

2.1. Hypothesis Test

The confidence interval for all tests is equal to 95

2.1..1. Does increasing the year of manufacture increase the price?

Poisson Distribution:

H0 : There is no significant relationship between the year of manufacture and the price of houses, following a Poisson distribution. In other words, the year of manufacture does not affect the price.

H1 : There is a significant relationship between the year of manufacture and the price of houses, following a Poisson distribution. In other words, the year of manufacture has an impact on the price.

Pearson Correlation Coefficient: 0.522897332879497
P-value: 2.990229099012696e-103

Reject the null hypothesis: There is a significant correlation between price and year of manufacture.

2.1..2. Does increasing the year of manufacture increase the price?

Regression Test

Null Hypothesis (H0): The renovation date has no significant impact on the price of houses in the regression model. In other words, increasing the renovation date does not reduce the price.
Alternative Hypothesis (H1): The renovation date has a significant impact on the price of houses in the regression model. In other words, increasing the renovation date reduces the price.

Based on the results of the Ordinary Least Squares (OLS) regression analysis, we can draw important insights regarding the impact of the year of renovation ('YearRemodAdd') on the price of houses. The coefficient for 'YearRemodAdd' is approximately 1951.2994. This means that for each additional year in the renovation date, the house price is estimated to increase by around 1951.30 units.

The associated p-value for 'YearRemodAdd' is remarkably low, close to zero (0.000), indicating a high level of statistical significance. This means that 'YearRemodAdd' is a strong predictor of 'SalePrice.' In other words, as the year of renovation increases, there is a statistically significant positive impact on the house's price.

These results provide robust evidence that increasing the year of renovation corresponds to a substantial increase in the price of houses. As data scientists, understanding the relationships between various features and house prices is crucial for informed decision-making and predictive modeling

2.1..3. the value of LotFrontage in the global average equal to 80?

T-Distribution

$$H_0 : \mu_{LotFrontage} = 80$$
$$H_1 : \mu_{LotFrontage} \neq 80$$

T-statistic: -17.26253928733087
P-value: 6.63536554942959e-61
Reject the null hypothesis: The mean LotFrontage in your data is not equal to the global mean of 80.

2.1..4. Is there a significant association between the number of parking spaces and the number of rooms in a house?

Chi2-Distribution

Null Hypothesis (H0): There is no significant relationship between the number of parking spaces and the number of rooms in a house. In other words, the two variables are independent, and the distribution of the number of rooms is the same regardless of the number of parking spaces.

Alternative Hypothesis (H1): There is a significant relationship between the number of parking spaces and the number of rooms in a house. In other words, the two variables are not independent, and the distribution of the number of rooms varies based on the number of parking spaces.

Chi-squared statistic: 394.53794251977536
P-value: 7.305634237180683e-58
There is a significant association between the number of parking spaces and the number of rooms.

2.1..5. Does the type of street affect the price?

ANOVA Test

Null Hypothesis (H0): The type of street has no significant effect on the price of houses. In other words, the means of house prices are equal across different types of streets.

Alternative Hypothesis (H1): The type of street has a significant effect on the price of houses. In other words, the means of house prices are not equal across different types of streets.

F-statistic: 2.4592895583691994
P-value: 0.11704860406782483
There is no significant relationship between StreetType and the average Price of houses.

References

1. Kaggle
2. matplotlib