2022

Descriptive & Cluster Analysis of AMESHOUSING3 Dataset

Abstract

In this research report, we will conduct descriptive analysis and cluster analysis of variables in the AMESHOUSING3 data set to investigate the determinants of home selling prices. In addition, a customer background is provided to help us assess the value of the customer needs of the house in the actual situation. Part of this analysis is based on traditional house characteristics, such as size, number of bedrooms, construction year, etc. Research shows that houses for sale are typically concentrated between 6,250 feet and 11,250 feet, most of which have three to five bedrooms and range in price from $100,000 to $150,000. In the last part, we recommend the houses that meet the requirements for customers according to the research results.

Keywords Descriptive analysis · Cluster analysis · House characteristics · Sale prices

Table of Contents

Introduction

In modern society, people's living standards are getting better and better. Everyone wants to live in a comfortable environment. With the development of the construction industry and the real estate industry and the increasing diversification of customer needs, the types and configurations of residential houses are becoming more and more abundant. These features typically include lot size, number of bedrooms, base type, garage and heating and cooling, each of which affects the sale price of the home. All types of housing are in demand, depending on the specifics of the customer (e.g., family size, salary, personal preferences, etc.).

AMESHOUSING3's dataset contains characteristic data and sales for more than 300 homes. There are a wide variety of home types, each with different configurations and different prices. For a more concrete analysis of the problem, we establish a young male worker as a virtual customer in this case. Here is some basic information about the customer and his needs: The customer is a worker who lives alone and is looking for a two-bedroom house with full bathroom, central air conditioning, heating and garage. For now, he has budgeted no more than $175,000 to buy the house and has no other requirements.

Based on the analysis of housing characteristic data and sales data, this report will discuss the sales situation of different types of houses from these 300 sales data. As a data analysis team, we will provide users with detailed analysis reports to match the most suitable home. Therefore, the first goal of this paper is to use descriptive analysis to gain a more complete understanding of the configuration of different types of houses, namely house characteristics. Cluster analysis is used to identify similar housing configurations. The second goal is to analyse the relationship between clustering and sales prices, provide the hypothetical customer with the sales situation of different types of houses in the house he needs, and recommend the most suitable listing information for him.

Methodology

All the analysis, induction and drawing of this report are carried out on SAS Viya. In this report, our first step is to sort out and summarize some variables in the data set through descriptive analysis, to understand the attributes of these variables in general. We analysed the variables of " Lot_Area", "Bedroom_AbvGr", " Full_Bathroom", " Year_Built" and "Age_Sold" in the data set respectively, and obtained the maximum / minimum values, mean, median, mode, standard deviation, and skewness of these variables. Through these, we can infer their dispersion degree, frequency distribution, concentration trend and other information. Then, through analysis, we selected the best clustering number k as 5 (dividing the data into five clusters), carried out clustering analysis on the data set twice and drew a chart.

Finally, we draw a scatter of house area and frequency and a scatter diagram of house area and selling price. Through further sorting and summarizing the data, we have reached some conclusions, which can help us make better suggestions for home buyers.

Result

The purpose of the cluster analysis in this report is to determine the impact of house characteristics on house sales prices for each of the different clusters. Therefore, the variables selected in the next section are all house characteristics, including the square footage of the house, the number of bedrooms, the original construction year, etc. The variables associated with the sales price of a home are the number of homes separated by the sales price of $175,000 and the natural logarithm of the sales price. This report plans to build two cluster analysis model with five variables, one cluster without considering customer needs and one cluster filtered according to customer needs. Next, a descriptive analysis of the variables involved will be conducted through the Appendix 1.

Descriptive analytics

Lot size in square feet: Is the size of the house that has been sold, and we need to determine whether it will affect the selling price of the house. The value of this variable ranges from 1,495 feet to 26,142 feet, and the standard deviation is 3,323.79, so the dispersion degree is large. The mean of this variable is 8,294, the median is 8,265, and the mode is 7,200. The skewness is 1.0093, so the shape of this variable is skewed to the right.

Bedrooms above grade: Is the number of bedrooms in a house that has been sold. The value range of this variable is 0 to 4, and the standard deviation is 0.69, so the dispersion degree is small. The mean is 2.51, the median is 3, and the mode is 3. The skewness is -0.7203, so the shape of this variable is left-biased.

Number of full bathrooms: It's the number of full bathrooms in a house. The value range of this variable is 1-4, and the standard deviation is 0.66, so the dispersion degree is small. The mean is 1.68, the median is 2, and the mode is 2. The skewness is 0.5397, so the shape of the variable is skewed to the right.

Original construction year: Is the year the house that was sold was built. The value of this variable ranges from 1875 to 2009 with a standard deviation of 27.6, so the degree of dispersion is larger than Bedrooms above grade and Number of full bathrooms, but smaller than Lot size in square feet. The mean is 1962, the median is 1963, and the mode is 2004. The skewness is -0.1821, so the shape of the variable is skewed to the left.

Age of house when sold, in years: Is the age of the houses that have been sold. The value range of this variable is 1 year to 135 years, and the standard deviation is 27.48, so the dispersion degree is close to original construction year, larger than Bedrooms above grade and Number of full bathrooms. But it is smaller than Lot size in square feet. The mean is 45.89, the median is 45, and the mode is 4. The skewness is 0.1953, so the shape of the variable is skewed to the right.

By creating the correlation matrix which shows in Appendix 2, it is obvious the correlation between age of house when sold and original construction year is greater than 0.7, which is a strong correlation. As strong correlation in same cluster could make it difficult to distinguish their influence on the dependent variable, so we decide not to put original construction year and age of house when sold together. The correlation between number of full bathrooms and original construction year is greater than 0.5, and the correlation between number of full bathrooms and age of house when sold also greater than 0.5, which are moderate correlations. The rest correlations are weak correlations. As strong correlation could be difficult for people to distinguish their influence on the dependent variable in the same cluster, so we choose not to put original construction year and age of house when sold together.

Cluster analysis

In the descriptive analysis above we mentioned the characteristics Lot size in square feet, Bedrooms above, Number of full bathrooms and Age of house when sold, in years, and set the K value to 5. We cluster these data as Appendix 3 shows and draw the following conclusions.

- **Cluster 1**: Average size around 7465.72 feet, number of bedrooms 3.07, full bath 1.2. It is defined as a mid-sized home, suitable for 3-4 people, it is the main home sold, accounting for about 30% of the overall sold homes data.
- **Cluster 2**: Average size around 10,746.42 feet, number of bedrooms 2.90, full bath 1.9. It is defined as a large home, suitable for 5-6 people, and one of the main types of homes sold, accounting for about 55% of the overall share.
- **Cluster 3**: Average size around 10,375.31 feet, number of bedrooms 2.82, full baths 3. It is defined as a large home, suitable for 5-6 people, and one of the main types of homes sold, accounting for about 55% of the overall share.
- **Cluster 4**: Average size around 7857.26 ft, number of bedrooms 1.84, full bath 1.2. It is defined as a medium sized home, suitable for 3-4 people, it is the main home sold, accounting for about 30% of the overall homes sold data.
- **Cluster 5**: average size around 4869 ft, number of bedrooms 1.77, full bath 2.1. It is defined as a small house, 15% of the overall data, suitable for 1-2 people living and basically meeting the conditions of use.

As seen in Appendix 3, the segmentation of housing characteristics in the AMESHOUSING3 dataset based on the K-Means cluster analysis method can reflect the distinct characteristics of each type of housing in a more comprehensive and detailed way, which is a feasible way to segment. Based on the housing segmentation results, we can then recommend suitable housing for the previously set virtual customer needs and meet the customer requirements more efficiently.

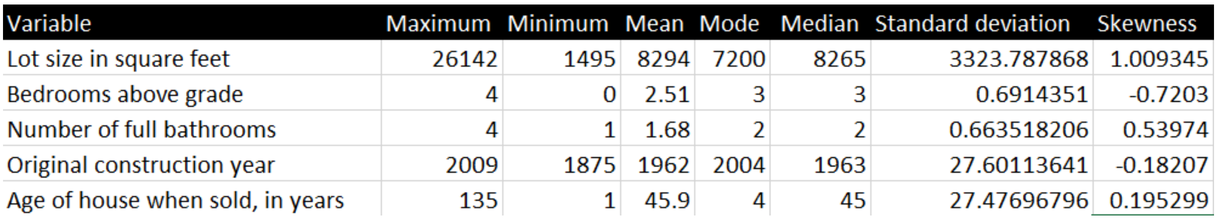
Conclusions

In the descriptive analysis part of this report, we obtained some descriptive statistical characteristics of the five data of house area, number of bedrooms, number of bathrooms, year of house construction and year of house sale through descriptive analysis. This helps us to further understand the data set and lay a foundation for the subsequent cluster analysis. According to our analysis, the sale price is influenced by the size of the home, the number of bedrooms, the number of bathrooms and the age of the home. After that, we have some insights into the distribution of this data.

To meet the needs of customers, we filter it based on the overall Cluster data, which shows in Appendix 4. We selected customer demand for 2 bedroom and full bath homes under $175,000, built after 1980, and matched 11 homes between $127,250 and $156,200, representing 3.6% of the total sales data. According to the above reasoning, consumers can buy 11 out of 300 houses that meet the requirements.

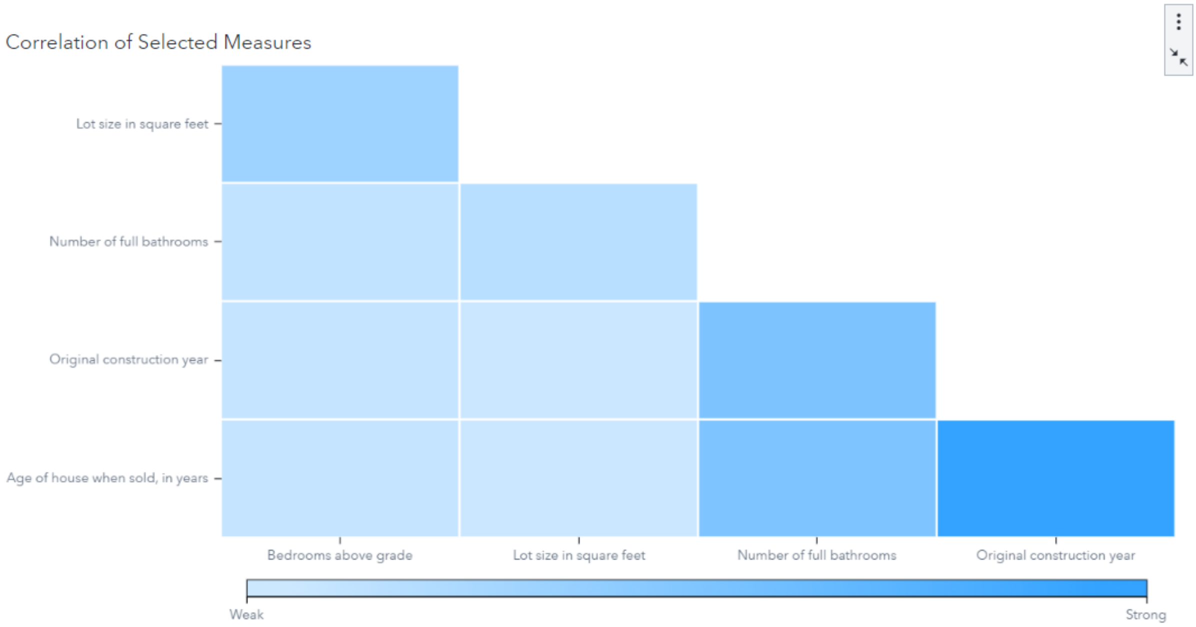
Appendix

Appendix 1: Descriptive statistics

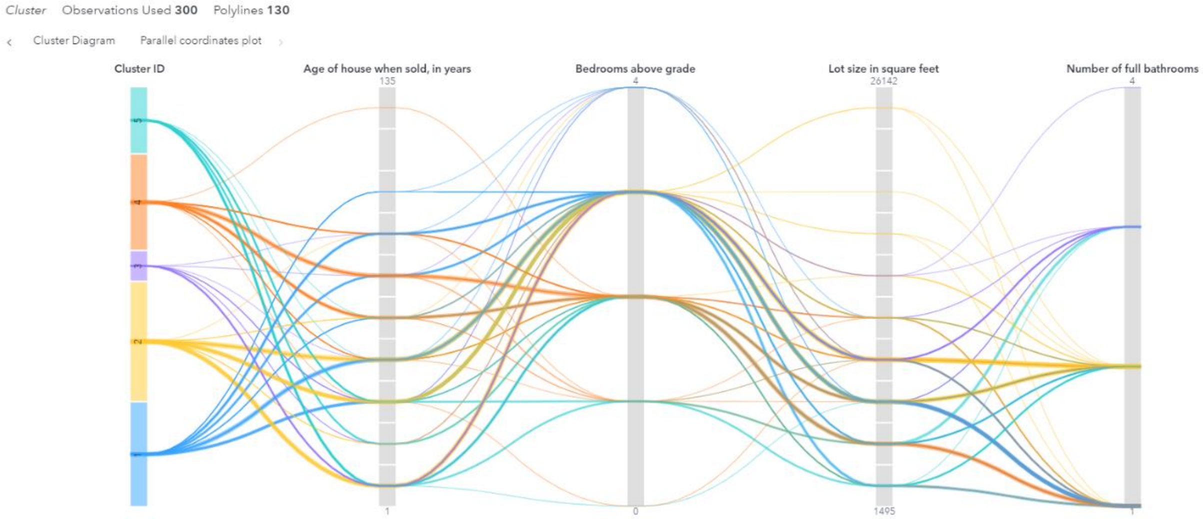


Variable	Maximum	Minimum	Mean	Mode	Median	Standard deviation	Skewness
Lot size in square feet	26142	1495	8294	7200	8265	3323.787868	1.009345
Bedrooms above grade	4	0	2.51	3	3	0.6914351	-0.7203
Number of full bathrooms	4	1	1.68	2	2	0.663518206	0.53974
Original construction year	2009	1875	1962	2004	1963	27.60113641	-0.18207
Age of house when sold, in years	135	1	45.9	4	45	27.47696796	0.195299

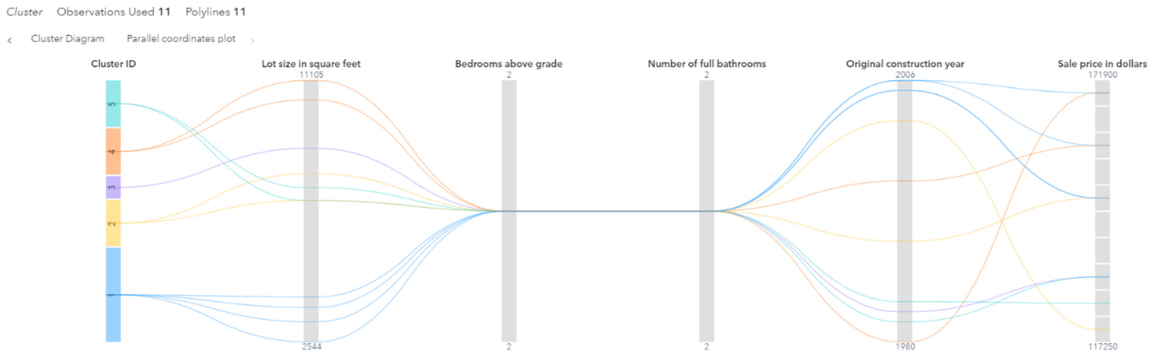
Appendix 2: Correlation Matrix



Appendix 3: Cluster 1



Appendix 4: Cluster 2 (after filter)



Cluster Observations Used **11** Polylines **11**