64060 Machine Learning Final Report

Mo Zhang(810841823)

[mzhang25@kent.edu](mailto:mzhang25@kent.edu)

##### Introduction

With the continuous development of the economy, the price of housing is also undergoing great changes. As a necessity in daily life, it is of great significance to study the factors related to the price change of housing. Especially for the United States, the prosperity of the American real estate market has increased unabated. Suppose we are a house salesperson, and we want to develop a house value analysis tool. Our goal is to have a basic understanding of house prices, living area and housing conditions. and use K-means to find out the factors affecting real estate prices. What is the impact of the combination of various house attributes on the house price, what are most houses like in Washington and how to better recommend a house to customers?

##### DATA UNDERSTANDING AND DATA PREPARATION

The dataset comes from Kaggle[[1]](#footnote-1). The dataset contains 4600 samples and 18 features for each sample. The features include house selling date, house selling price, the number of bedrooms, the number of bathrooms, square footage of the home (sqft\_living), square footage of the lot (sqft\_lot), total floors (levels) in house, waterfront, view, condition, square footage of house apart from basement (sqft\_above), square footage of the basement (sqft\_basement), built Year (yr\_built), year when house was renovated (yr\_renovated), street, city, statezip and country.

Intuitively, some features are not so important for predicting the price of house and they also have no meaning in clustering. To simplify our method, we first clean our dataset by deleting these useless features. Specifically, we delete 1(date), 15(street), 17(statezip), 18(country). And the detail statistic of features are summarized in the Fig.1.

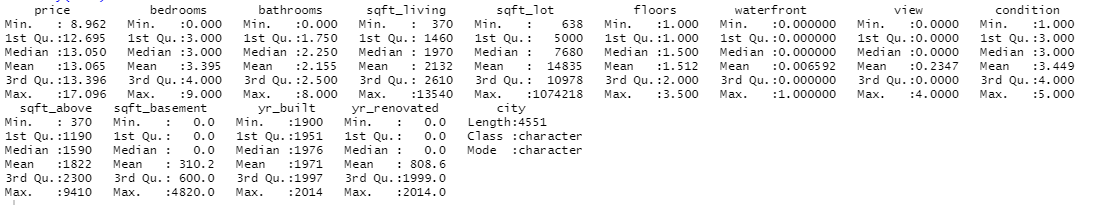


Figure 1. The detail statistic of features of all features. Specifically, the minimum value, Median value, Mean value and Maximum value of each feature is listed.

Then, we would like to analysis the relationship between location with house price. As shown in Fig.2, there are large number of high-end houses located in Medina, Clyde Hill, and Yarrow Point. Since house prices vary widely, I use "price=mean(price)/1000" to make sure the plot is reasonable. In Fig.3, we expose more statistical information about frequency of features. We can see (A) price, (B) sqft\_living, (C) bedroom and (D) bathroom have normal distribution and (E) condition and (F) yr\_built have random distribution.

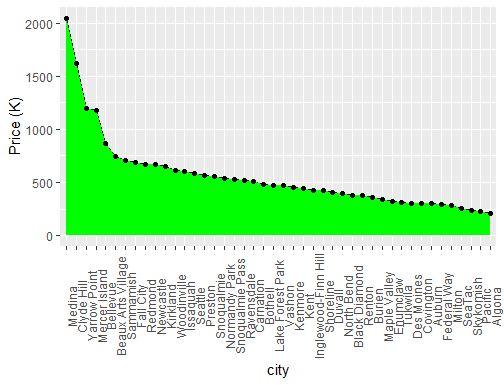


Figure 2. The relationship between city with price. The curve satisfies the long tail distribution, most house prices tend to be equal, and very few cities have high or low house prices.

图表, 瀑布图

描述已自动生成

Figure 3. We expose more statistical information of our features. Specifically, (A) shows the frequency of price; (B) shows the frequency of sqft\_living; (C) shows the frequency of bedroom; (D) shows the frequency of bathroom; (E) shows the frequency of condition; (F) shows the frequency of yr\_built.

##### Explore though different clusters

The project goal is to identify the factors affected by the price on a house and find out the type of house. We would like to adopt k-means as our main method. We choose k-means as our main method, since k-means can automatically classify samples into clusters by their features which can help us to understand the importance of each feature. Moreover, we can easily plot the figure of k-means which can help us have a better visualization of clusters.

Intuitively, people will consider more on the living area of house and condition of the house. Thus, we first divide our features into two separate classes to explore their relationship with house price. Specifically, we select the features about the area of house as (sqft\_living + sqft\_lot + sqft\_above + sqft\_basement). And we select the features about the condition of house as (condition + yr\_year + yr\_renovated). In the following parts, we would introduce the detail of our results.

##### House area cluster

Firstly, I make standardization to rescale every key feature we have selected. Then, I used k-means to find if there are any natural groupings of houses. As shown in Fig.4, we get 5 clusters. The red, blue, green clusters are more compact, and pink cluster, yellow cluster are sparse. The reason is that red, blue, green clusters have a stable price: although these features within each cluster have some differences, their price is stable and change a little. As shown in Fig.7, these houses are located on Kirkland and Redmond etc. We have checked that these places have a stable price of house and they do not change much for a long time.

For pink and yellow clusters, their prices are not stable and have a big change if the house have different areas. As shown in Fig.7, these houses are located on Woodinville, Bellevue, etc. There are many areas where rich people live, and the house price is very different from the ordinary location. Specifically, the house id 1079, 123 are bought by rich person.

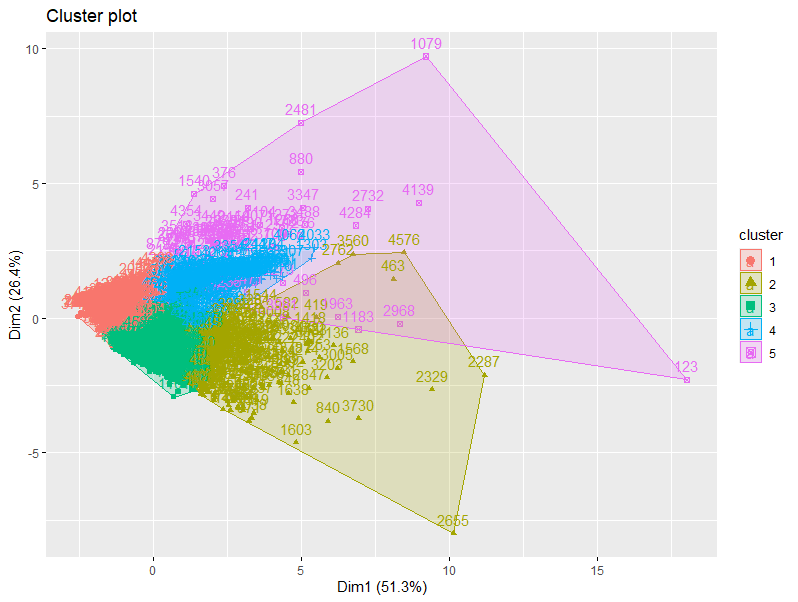


Figure 4. The visualization of five clusters. 4 features about the area of house include sqft\_living, sqft\_lot, sqft\_above, sqft\_basement are adopted by k-means.

As shown in Fig.5, the optimal number of clusters k are determined by using elbow method. We can see when the number of clusters are equal to 5, the clustering accuracy will decrease. So we choose k=5.

图表, 折线图

描述已自动生成

Figure 5. The elbow method to determine the optimal number of clusters of k-means.

For the sqft of living cluster, I use k-means models to find the steady clusters. Then I used the values of centroids to interpret the meaning of each cluster.

表格

描述已自动生成

Figure 6. We list all attributes of features on each cluster, and we give a name on each cluster.

Cluster 1 is called “single houses”. As shown in Fig.7, those located mainly in Seattle and Renton, with a small living area, small lot area, a small area above and a small basement. This single houses always have a low and stable price. As shown in Fig. 4, single houses are compact in 2-dimensional coordinate system.

I called the second cluster “luxury house”, it is interesting to see Fig.7 that most of luxury houses locate in Seattle with huge living area and huge basement. As shown in Fig 4. The cluster 2 is not compact as the cluster 1 like the house id 2287 and 2655. I analysis the reason is that some have stable price and others are in rich person area which has high price of houses.

The third cluster is “generally family house”, since those houses always have suitable living area, small lot and small garden. As shown in Fig 4, the cluster 3 is impact and it is very close to the cluster 1. We analysis the reason is the two types of houses have many similar features, such as sqft\_loft and sqft\_above. However, the two have great differences in features sqrt\_basement. Features sqrt\_basement have a great impact on the prices of the two types of houses. Families prefer to buy these types of houses because it has large sqrt\_basement.

The fourth cluster is” General house with a large garden”.As shown in Fig.4, it is similar to a general family house since they have many similar features like average sqrt\_living and average/small sqrt\_lot. But they are different in feature sqft above. Specifically, cluster 4 has a large garden and cluster 3 just has a small garden. So, garden size is a good feature to influence the prices of these two types of houses.

The fifth cluster is“large house”, they have a big living area, huge sqft of the lot, big sqft\_above and big basement. Most of the large houses are in Seattle and Sammamish. We can see the prices of houses are high in this area, since rich people prefer to buy houses on this location, and they prefer huge houses.

From the analysis of the above 5 clusters, we can summary that the location can influence the area of houses like Seattle and Sammamish prefers to build large house and Renton and Kent prefer to build small houses (Fig.7). And the area features of houses can influence the price of houses since clusters may have many similar features and a kind of different feature can classify the houses into different clusters which also influences the price of these houses. In short, larger houses always have higher prices. So, It is important to know which location and the area of house they prefer, and then we can recommend the corresponding price of house to them.

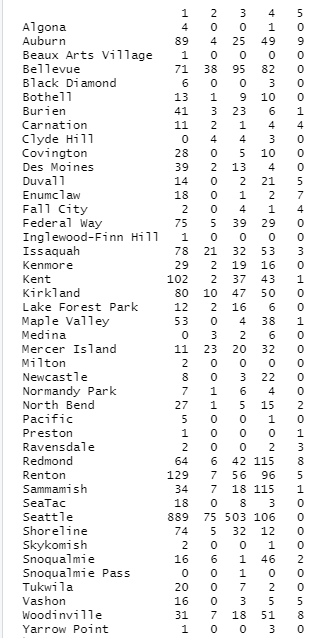


Figure 7. The number of five kinds of houses on each location. Specifically, 1 means single houses, 2 means luxury house, 3 means generally family house, 4 means General house with a large garden and 5 means large house.

##### House condition cluster

In this subsection, we select the features price, condition, yr\_year and yr\_renovated as our features to explore the relationship of these features. We also adopt k-means to detect the clusters of houses. We use elbow method to get the optimal number of clusters. As shown in Fig.8, we can get the highest clustering accuracy when k=10.

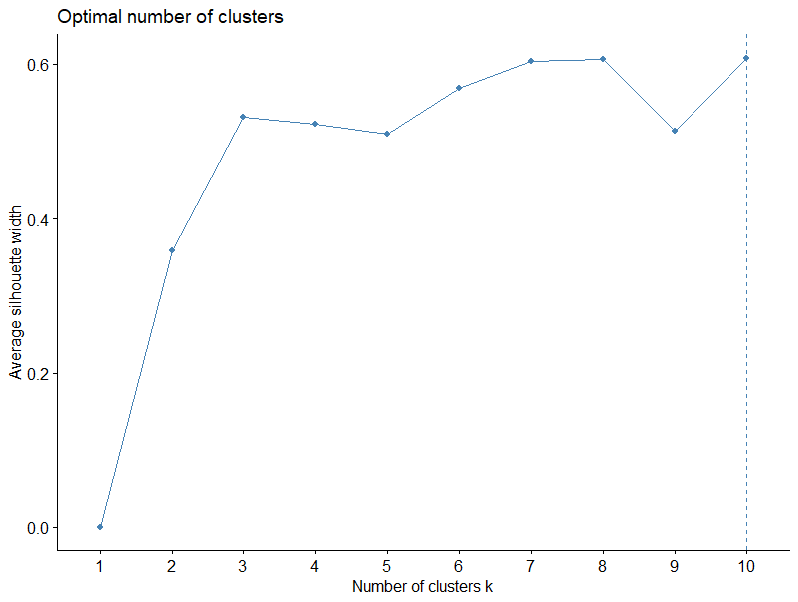


Figure 8. The elbow method to determine the optimal number of clusters of k-means. And we select k=10 as the number of clusters.

I plot the ten clusters on Fig.9. We can see all clusters are compact, and the distance between these clustering categories is very close, which shows that they have many similar features, but they also have many different features. Next, we introduce the features of each category. Finally, we analyze the relationship between housing conditions and their prices.

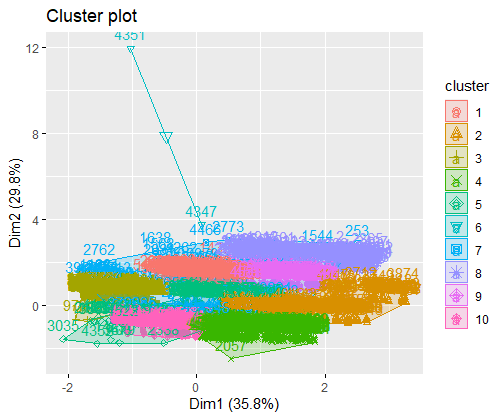


Figure 9. The visualization of five clusters. 4 features about the area of house include price, condition, yr\_year and yr\_renovated are adopted by k-means.

Since we have 10 clusters and it is hard to directly analysis the features in each cluster, we extract their clustering center as a representation of each cluster. Actually, this kind of methods are widely applied in real-world applications.

For cluster 5, cluster 8 and cluster 9, they have low price and high condition and low yr\_built and low yr\_renovated. The price is low since the build time of these house are long time ago, and the house renovated year is also too late. So, the houses are old and cannot have a high price. The location of these houses is in Bellevue, Kent and these places have many old houses.

As shown in Fig.10, the same situation occurs in cluster 2 and cluster 4. The house prices between these clusters are low since these houses are too old. The differences of their prices are influenced by the condition and yr\_renovated. In short, if the house condition is good, built and renovated closer to now, the price of the house will be higher. For example, the price of cluster 2 is higher than that of cluster 4 since cluster 2 has a better condition, and the price of cluster 4 is higher than that of cluster 5 since it has been renovated recently.

Cluster 6 has high prices since it is new and house condition is best. This kind of houses are in Issaquah and Redmond.

Cluster 1, cluster 3 and cluster 7 are built recently (higher yr\_built), so their prices are higher than classer 2. While cluster 1 has better condition than cluster 3. So, cluster 1 has a higher price than cluster 3. And cluster 7 has been renovated recently, so it has higher price than cluster 1 and cluster 3. We can get the summary that: a higher condition and higher renovate date will lead to a higher price.

Cluster 10 is a special case, since it has a high yr\_built and yr\_renovate, but it has a poor condition which significantly decreases its price.

In summary, we can see the examples from the above analysis that a better condition will lead to a higher price of houses (Cluster 2), a recent build year will lead to a higher price of house (Cluster 6) , and a recent renovated will lead to a higher price of house (Cluster4).

文本, 表格

中度可信度描述已自动生成

Figure 10. The clustering center of each cluster.

表格

描述已自动生成

Figure 11. The condition of each cluster.

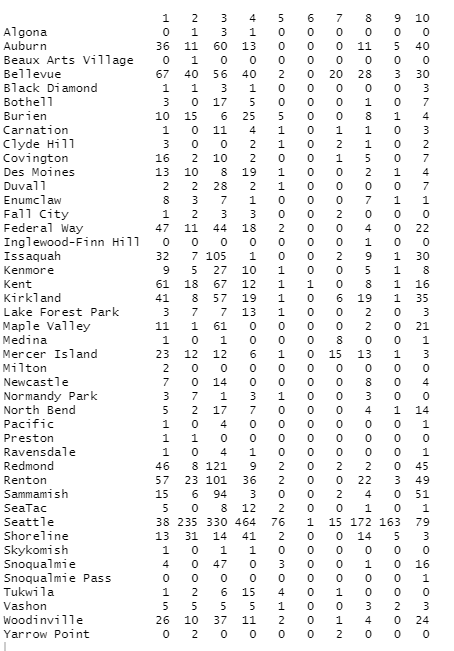


Figure 12. The number of ten kinds of houses on each location.

##### linear model cluster

After doing these two clusters, I want to import linear model into k-means. Through select the important variables by linear model and then assign a certain supervised performance to the unsupervised model. When we apply K-means to cluster, we always follow the following:

* Practical significance (variables that cannot be explained are eliminated):
* Correlation between variables (variables with too high a correlation will be removed to

ensure that the variables that remain span as many dimensions as possible).

But for this project goal, I want to find out what really influences the price. So, linear model can be a good way to filter important variables. A linear regression model was fitted to the data and l selection of the significant features with high scores. First, I would like to introduce the linear model and explain why the correlation coefficient between variables is important.

Firstly, we use y to represent the house price and to represent different attribute characteristics. According to the linear model, we can fit the value of y through the product between a set of coefficients and variables. The specific formula can be written as

where are parameters and b are a constant. As shown in Fig13, the model can be optimized by the least square method with gradient descent algorithm. However, in the process of fitting, many variables are redundant, and the weight coefficient represents the importance of these variables in the model prediction stage. The larger the coefficient, the greater the change of the variable will significantly affect the final house price. As shown in Fig.13, we can see the coefficients of variables in a linear model and Estimate represents the coefficients of linear model.

图表, 散点图

描述已自动生成

Figure 13. The linear model to predict the value of price of houses.

表格

描述已自动生成

Figure 14. The coefficients of variables in a linear model. The higher coefficient of a variable means the variable is more important.

Finally, we use all features in K-means to get clusters. As shown in Fig.15, we can get three clusters. We analysis the features on each cluster: for cluster1 houses have high price, lots of bedroom and bathrooms, large sqft of living and lot, more floor and good view, good condition, large sqft\_above, samll sqft\_basement and old build. And for cluster2 have not too expensive price, large bedrooms and bathrooms, larger sqft of living, sqft\_above and lot, more floor, bad view and condition, small sqft\_above and new house. Cluster3, houses have low price, small number of bedrooms and bathrooms, small sqft of living and lot, small amount floor, good view, good condition, small sqft\_above, big sqft\_basement and old build. And according to the coefficients, we select sqft\_living to check if it affects the price of house. As shown in Fig.16, we can see a higher sqft\_living always has a higher price. This is easy to understand from the aspect of market that a larger square footprint of the home requires more construction costs, so the selling price will be more expensive.

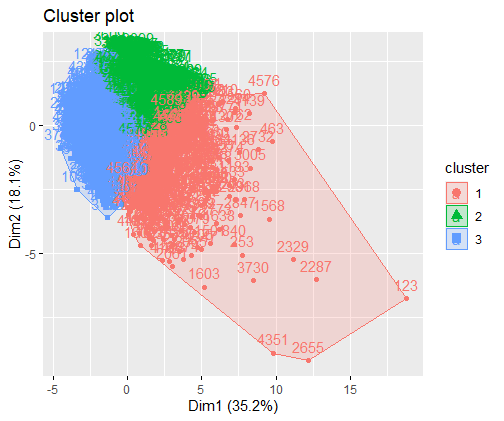


Figure 15. The visualization of all clusters.

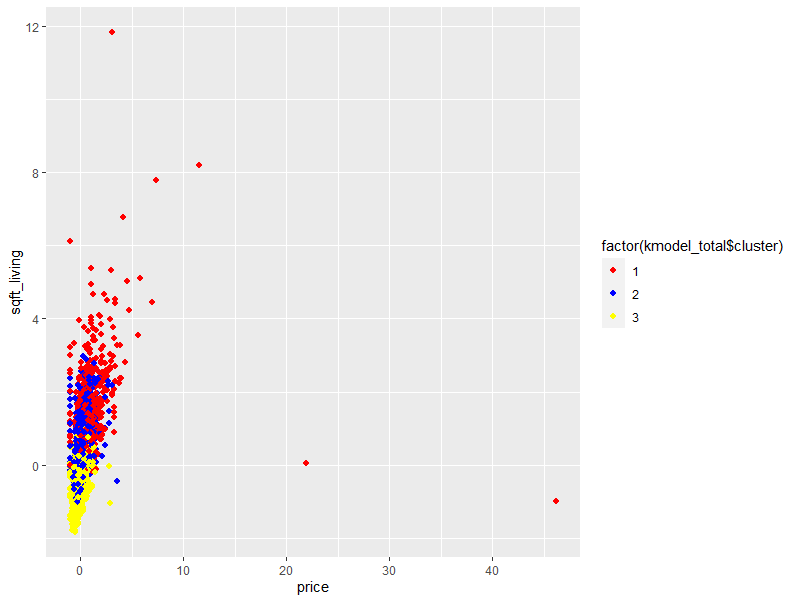


Figure 16. The relationship between sqft\_living and price.

##### Conclusion

1.The area of the house has a great impact on the price. In short, the larger the area of the house, the higher the price of the house. When the indoor area of the house is the same, the larger the area of the yard, the higher the price of the house. Of course, the price of houses in the same level will also be affected by different locations. For example, the house price in Bellevue is higher than that in Kent.

2.The conditions of the house also influence the price of the house. The longer the construction time of the house, the more expensive the price of the new house, the longer the renovation time of the house, the higher the price of the new house, the better the conditions of the house, and the higher the price of the house. Of course, the same housing conditions, different regions will also have an impact on the price of the house.

3.Through the weight of the coefficients of the linear model, it can be found that the characteristics that have the greatest impact on the house price are bedrooms, bathrooms and sqft\_ living,sqft\_lot, floors, view ,yr\_built and condition.

##### Future Deployment

1. Classifying houses, offering different houses for different kinds of guests.
2. Knowing the distribution of houses in different cities, help customers with different needs.

3. Understand the main factors affecting house prices and get a reasonable explanation from the perspective of the market and economics.

4. Analyze customer preferences and what characteristics of the house type will be more popular with customers.

1. https://www.kaggle.com/shree1992/housedata [↑](#footnote-ref-1)