MA308 Project Report

Gengshang Dong, Vincent William Hadiasali, Yuang Li, SiHan Fu

January 3, 2024

1 Introduction

The project content of our team is to analyze the housing price data of the three cities of Guangzhou, Xiamen and Shenzhen. The data set we mainly use is "Data/Guangdong Xiamen Shenzhen 2018.csv". We used linear models to fit the best models for different cities and different districts, and used random forest models to predict district variables through other variables, and achieved relatively good results. At the same time, methods such as variance analysis were used to analyze the possible impact of discrete variables on the unit price of housing prices. First, we give an overview of the data.

danjia	area	floor	hall	room	school	chaoxiang	year	subway	district	city	age	price
109243	372	中层	5	3	1	南北向	2008	0	南山	深圳	10	40638396
110000	371	中层	6	3	1	其他	2016	0	南山	深圳	2	40810000
77838	370	高层	6	3	1	南北向	2009	0	宝安	深圳	9	28800060
95890	365	高层	7	4	1	南向	2006	0	南山	深圳	12	34999850
83333	360	高层	5	3	1	南北向	2006	0	南山	深圳	12	29999880
106205	358	低层	5	3	0	南向	2008	0	宝安	深圳	10	38021390
108262	351	低层	5	3	0	南向	1998	0	南山	深圳	20	37999962
85470	351	中层	7	3	0	其他	2016	0	宝安	深圳	2	29999970
77143	350	中层	5	2	0	南北向	2016	0	宝安	深圳	2	27000050
82514	350	中层	6	3	0	南北向	2016	0	宝安	深圳	2	28879900
78571	350	中层	6	3	0	其他	2016	0	宝安	深圳	2	27499850
82671	337	低层	6	3	0	南向	2014	0	宝安	深圳	4	27860127
77612	335	中层	5	2	1	其他	2012	0	福田	深圳	6	26000020
77612	335	中层	5	2	1	其他	2013	0	福田	深圳	5	26000020
65672	335	中层	5	2	0	南北向	2013	0	宝安	深圳	5	22000120
65582	335	高层	5	3	0	南北向	2015	0	宝安	深圳	3	21969970
71284	335	高层	5	2	0	南北向	2015	0	宝安	深圳	3	23880140
74627	335	高层	5	2	0	南向	2015	0	宝安	深圳	3	25000045

(a) Data Overview

Column Name	Data Type	Min Value	Max Value
danjia	int64	6538	121554
area	int64	32	399
floor	object	N/A	N/A
hall	int64	1	9
room	int64	0	8
school	int64	0	1
chaoxiang	object	N/A	N/A
year	int64	1980	2018
subway	int64	0	1
district	object	N/A	N/A
city	object	N/A	N/A

(b) Datatype of the variables

Figure 1: Data Overview

2 Data Overview

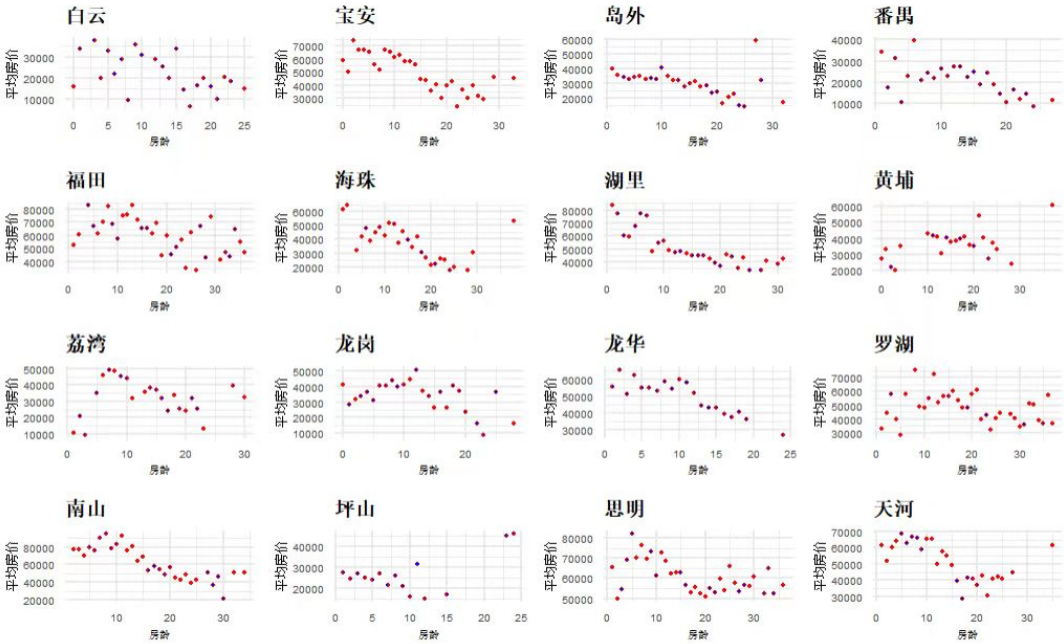


Figure 2: Price vs. Age

This data comes from a real estate agency website with a total of 8,430 samples. Data variables: unit price, area, floor, number of entrance halls, number of rooms, whether it is a school district, orientation, construction year, whether it is close to the subway station, urban area, city. Among theses variables, "school", "chaoxiang", "subway", "urban" are discrete variables. To better use the information of construction year, we use 2018 minus "year" to get the age of the house. Now the first several rows are shown in Figure 1. After generating age data, we plotted the changes in housing prices in each district with age (Figure 2). Overall, as the age of the house increases, the unit price shows a downward trend.

3 Find the best linear model

3.1 Use linear model to fit the data

First, by looking at the correlation heat map of the overall data (Figure 3), we find that hall and area have a strong correlation, so we no longer consider hall when fitting the data. By dividing the overall data into cities and regions, we obtained the optimal model corresponding to each region through stepwise regression (Figure 4). This method performs model selection by minimizing AIC.

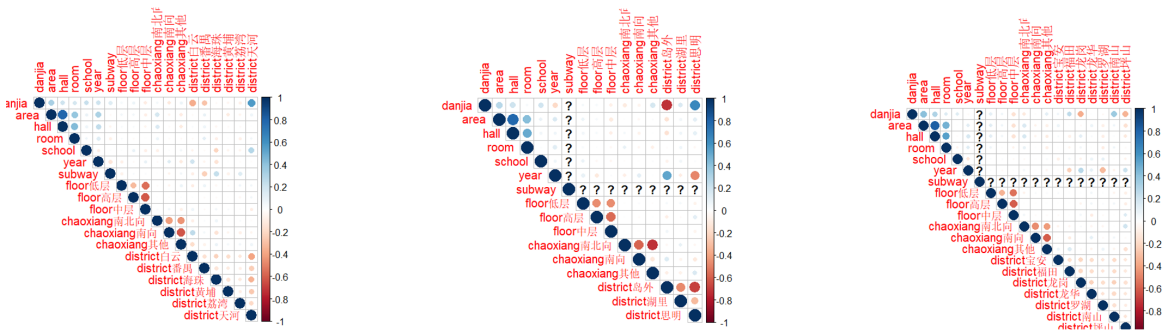


Figure 3: Variable Correlation Heat Map of different cities

Region	R-squared	Intercept	area	floor高层	floor中层	hall
深圳	0.5553165	43036.7528	102.8815	-2472.9472	-3147.2756	2809.
广州	0.5311387	18702.92067	95.35846	360.97906	-1778.05343	-1887.
厦门	0.6654445	36405.26149	47.42591	-1619.91826	-1237.91637	-703.8
白云	0.3433051	24180.31965	59.71754	-1163.89340	-3202.93428	NA
宝安	0.4126014	52296.47424	86.87547	-2803.79890	-4345.05547	2310.5
岛外	0.230701	27324.73926	12.67935	NA	NA	1160.2
番禺	0.5105763	13370.31420	69.41752	NA	NA	NA
福田	0.2480397	67057.7093	145.4451	-3196.4765	-6636.3025	9955.(
海珠	0.4376859	33210.7665	55.7610	-4205.9654	-5427.3026	2534.
湖里	0.4384147	73719.56664	69.85124	-5311.56790	-3957.25430	-1552.
黄埔	0.4377136	16776.6213	150.1089	-4377.0317	3160.3309	4113.2
荔湾	0.3404471	8428.8914	204.2611	-3718.2021	-7622.1192	-4981.
龙岗	0.3800812	16477.93464	66.14573	828.71744	-1394.97205	3483.
龙华	0.4214741	46028.91086	85.74019	-3391.80852	-2872.52797	2979.5

Figure 4: Best linear model of cities and districts

3.2 Delete outliers

We found that some districts have very low R-square, such as Futian District and Island Waiwai District. Therefore, we screened out the data of Futian District, used stapAIC() to fit Futian District alone, and then performed model diagnosis (Figure 5). This shows that some of the data are outliers. We filter out the outliers by setting the threshold of cook's diatance. By looking at the outliers, we find that the common characteristics of these outliers are that the area is very large (more than 300) or very small (less than 50). After deleting this part of outliers, use stepAIC() again to fit the unit price in Futian District. The original data of Futian District has 469 rows. After removing outliers, the data becomes 433 rows, and the R square increases from 0.24 to 0.43. At the same time, we also found that when the partition dimension is city, the fitting degree of the three cities is very good. This is because the district is a very important variable when fitting the housing prices of each city. Because the prosperity of different districts will greatly affect the housing prices in this area.

3.3 Random Forest to forecast districts

Then, we took out the data of Shenzhen city separately. By using RandomForest, 80% of the data was used for training, and a model was obtained that can predict which district a house comes from through other information. In the end, the model achieved an accuracy of 63%.

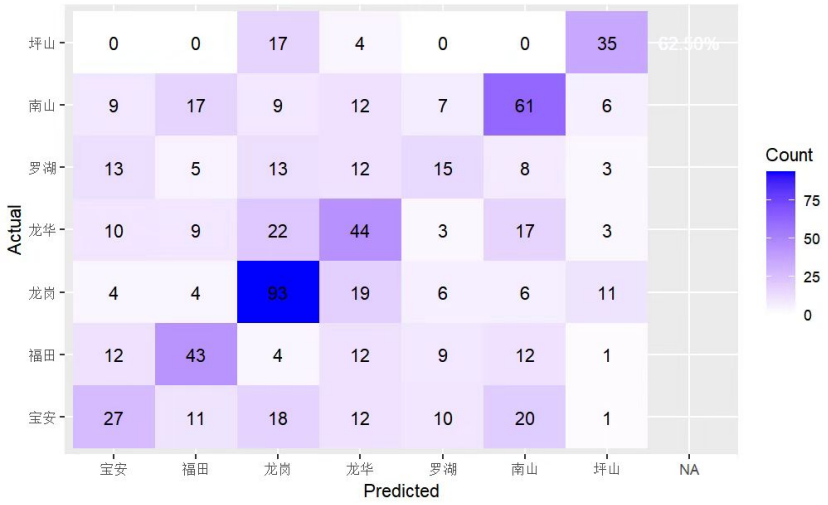


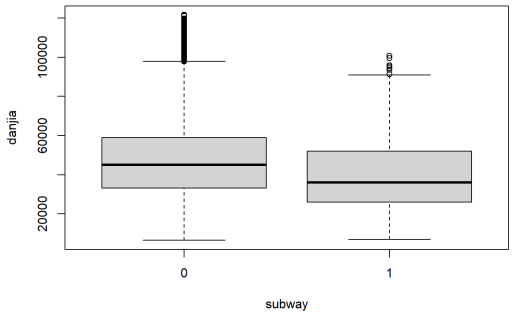
Figure 5: RandomForest to predict districts of Shenzhen

4 One-way ANOVA where price is the dependent variable

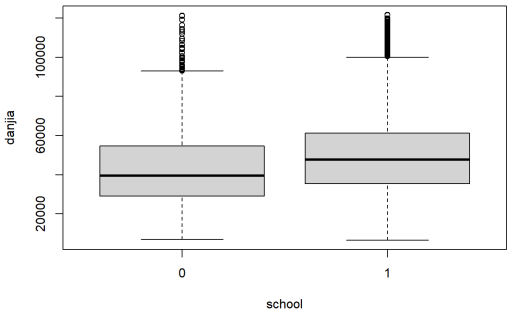
4.1 Boxplot

First, let's understand what does each feature mean in boxplot plotting

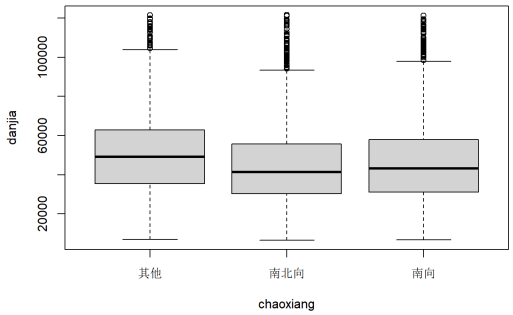
- The bottom part of whisker indicates the minimum value
- The bottom part of the box indicates the first-quantile value
- The middle line in the box indicates the median value
- The top part of the box indicates the third-quantile value
- The top part of whisker indicates the maximum value
- Any points above the maximum value are outliers



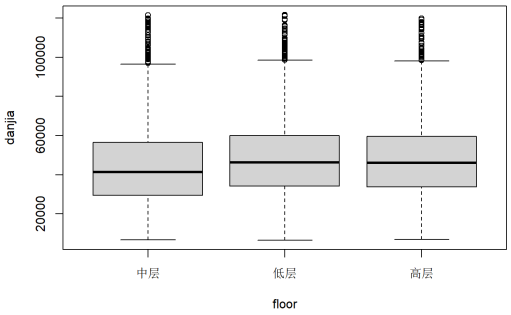
(a) danjia with subway



(b) danjia with school



(c) danjia with chaoxiang



(d) danjia with floor

Figure 6: Boxplot results

Our prediction is that housing with subway should have higher price. However, the boxplot surprisingly shows that housing without subway nearby has a higher price generally and numerous outliers. As we analyzed the data, we found out that both Shenzhen and Xiamen only has 0 as their subway value. Besides, they have a high price. The second boxplot shows that housing prices with a school nearby are higher in all aspects of boxplot which make sense. As South direction is a good fengshui, we expect that any direction involving South has a higher price, but the result is the opposite where other direction has the highest price relatively compared all direction involving South. While all aspects of pricing in boxplot do not differ much between low-level floor and high-level floor, mid-level floor has the lowest price generally.

4.2 VIF

From what we learned in the lecture, the major-thumb rule deciding whether there is a multicolinearity problem is when $\sqrt{VIF} > 2$. If a variable has its VIF root value exceeding it, deleting it is one of the solution.

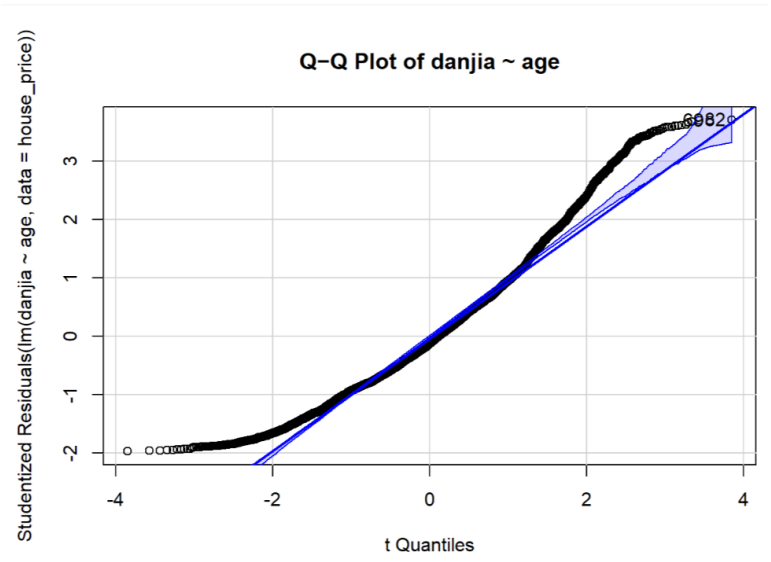
```
##          area      floor低层      floor高层          hall          room
##      2.732734       1.195996       1.195611       2.912058       1.372898
##        school chaoxiang南北向   chaoxiang南向        subway           age
##      1.030035       1.522505       1.441114       1.077787       1.068301
```

Figure 7: VIF result

According to the VIF result, there is no need to erase any variable.

4.3 One-way ANOVA

The assumption of One-way ANOVA is the same as that of linear regression where the dependent variable is assumed to be normally distributed and have equal variance in each group.




```
## [1] 69 82
```

```
bartlett.test(danjia ~ age, data = house_price)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  danjia by age
## Bartlett's K-squared = 376.75, df = 38, p-value < 2.2e-16
```

```
outlierTest(aov(danjia ~ age, data = house_price))
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 82 3.705056          0.0002127           NA
```

Figure 8: Assumption test result between danjia and age

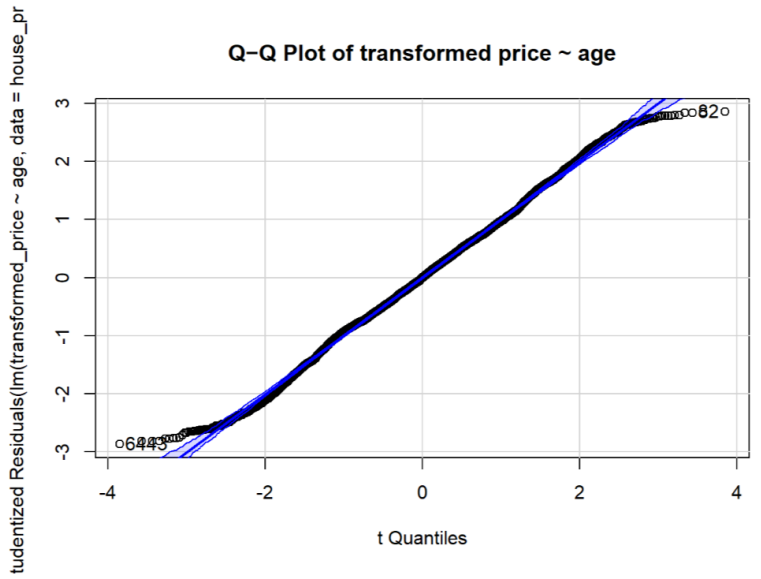
Let us take the assumption test result for danjia $\sim$ age, it violate the one-way ANOVA assumption as most of the points are outside of the blue area and the p-value is less than 0.05 in Bartlett's test which indicates that they are not normally distributed and have significantly different variances respectively. Additionaly, ANOVA methodologies can be sensitive to the presence of outliers. We can test for outliers using the outlierTest() function in the car package, as the Bonferonni p-value does not show NA, it means that there is an outlier problem. The rest of assumption test results that can be seen in Vincent's R Markdown code also have similar result.

To fix this problem, we use Box-Cox transformation with the best transformation parameter. the best transform parameter can be found by powerTransform(). However, we can apply summary function to the previous function together to get more insight.

```
## bcPower Transformation to Normality
##                    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## house_price$danjia    0.4416        0.44       0.4043       0.4789
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df       pval
## LR test, lambda = (0) 568.4741  1 < 2.22e-16
##
## Likelihood ratio test that no transformation is needed
##                             LRT df       pval
## LR test, lambda = (1) 803.1579  1 < 2.22e-16
```

Figure 9: Result of summary(powerTransform()) on danjia column

According to the result, the log transformation and no transformation will not be a normal distribution. So, we use 0.4418 as the transformation parameter. To apply the Box-Cox transformation, we use BoxCox() function from forecast package.




```
## [1]   82 6443
```

```
bartlett.test(transformed_price ~ age, data = house_price)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  transformed_price by age
## Bartlett's K-squared = 234.76, df = 38, p-value < 2.2e-16
```

```
outlierTest(aov(transformed_price ~ age, data = house_price))
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 6443 -2.86939          0.0041229           NA
```

Figure 10: The assumption test result for transformed price and age

Taking independent variable age as an example again, we can see that all plotting points in each Q-Q Plot are finally straight and inside of the blue zone meaning that they follow the normal distribution. However, the p-value of transformed price with age are less than 0.05 in Bartlett's test which indicates that the homogeneity variance does not hold. Lastly, the outlier problem has been finally resolved. While for the rest of selected independent variables for One-Way ANOVA which can be seen in Vincent's R Markdown file, all plotting points in each Q-Q Plot are finally straight and inside of the blue zone meaning that they follow the normal distribution. Also, the p-value of transformed price with every single dependent variables are more than 0.05 in Bartlett's test which indicates that it has non-significantly different variances. Finally, all of the Bonferonni p-value in outlier test show NA which means the outlier problems have been finally resolved. Hence, except for transformed price $\sim$ age, all assumption tests are finally obeyed.

(a) Untransformaed ANOVA Result

(b) Transformed ANOVA Result

Figure 11: ANOVA Result

Now, let us compared the result of ANOVA. p-value less than 0.05 implies that those variables are significantly important and the similarity that both ANOVA share is that they only have age as a non-significant variable amongst the five observed independent variables. However, the Sum of Squared dramatically decrease compared to the untransformed ANOVA observation. Hence, the observation with BoxCox transformation betters off.

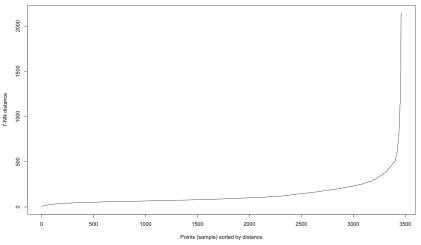
5 Clustering Analysis

We want to use different classifiers on the dataset to estimate different types of houses in different cities. We use data in shenzhen as a example and later check the data in Guangzhou too. At first we tried DBSCAN directly thinking that it is more likely to achieve good results as relationship between house price and these variables are quite complicated, however the results are quite bad. We tried K-means latter by choosing 6 variables which have low correlation to each other. And later we tried to use PCA before DBSCAN and K-means which make the result much better than before.

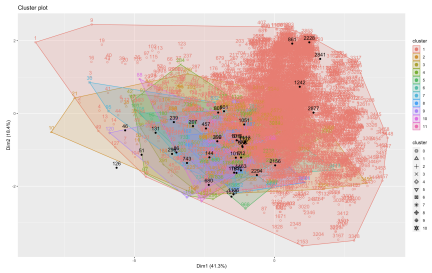
5.1 DBSCAN

We use the R-package "dbscan" to accomplish DBSCAN classification and we get the value of the size(radius) of the epsilon neighborhood by k-NN distance plot in Figure 12(a). We take eps=400 and then visualise the results of DBSCAN in two dimensions in Figure 12(b) and also the paris plot in Figure 12(c). We can see that the results have 4 clusters and one of them contains most of the data and we can see the clusters boundaries can not classify the data very well. And we make the box plot

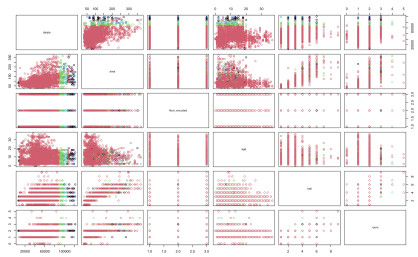
of the price of each cluster and the three other clusters which contain small number of the observations have slight difference. We believe this is not a good classification so we try K-means later.



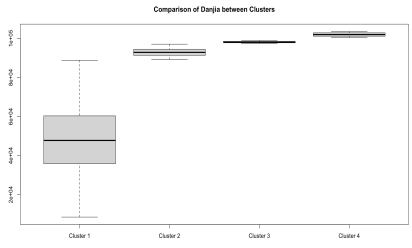
(a) Sum of squares against number of clusters



(b) Visualization in two dimensional space



(c) Pairs plot

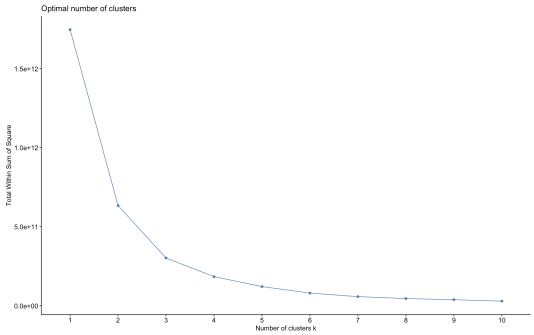


(d) Box plot

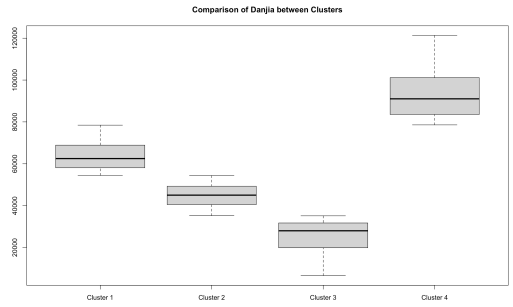
Figure 12: Results of the DBSCAN

5.2 PCA and K-means

At first we choose six variables that has the least Correlation hoping it will make the K-means result good. After using the one-hot method to deal with the categorical variables we choose "prices, area, floor, school, hall, age " as the variables in K-means. We use function in package factoextra for determining and visualizing the optimal number of clusters. We choose 4 clusters at last and get the box plot of price for each cluster. We can see that the price has big difference in each cluster in Figure 13(b) so we want to check other variables.



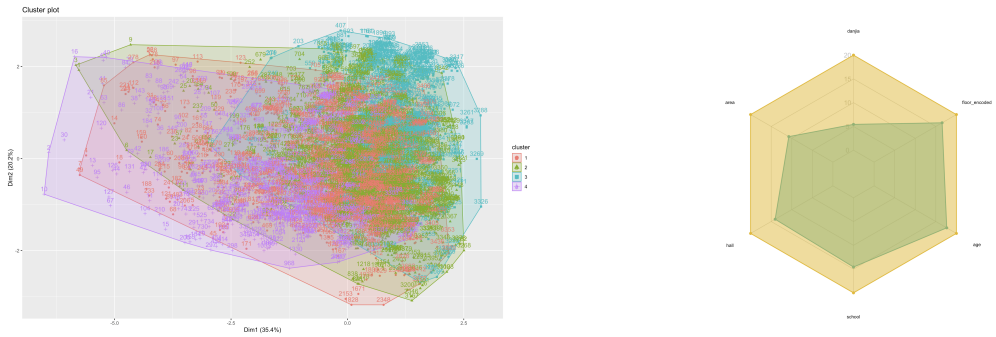
(a) K-NN distance



(b) Visualize Clustering Results

Figure 13: Results of the K-means

We first visualise the data in two dimensional space and we met the same problems as the DBSCAN and we try the radar chart plot for the cluster3 and cluster4 which have the larggest difference in price however the radar chart plot also looks bad as the value of other variables of higher price cluster's are higher too, which indiates the correlation of the variables still have a big influence on the result

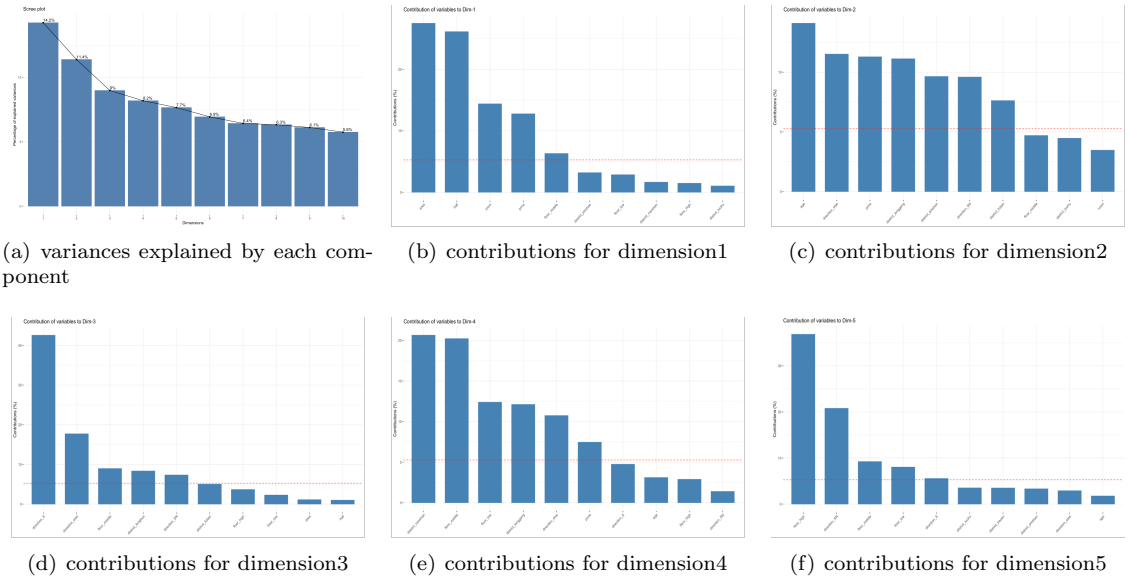


(a) K-NN distance

(b) Visualize Clustering Results

Figure 14: Visualization for the K-means

To eliminate the effects of the correlation between each variables and the selection of variables we use Principle Component Analysis PCA. To decide how many component we are going to use we first plot the variances against the number of dimensions in figure. We can see that the result is not very good as the first component only explain 14.2 percents of the variance in Figure 15(a), however we think this is because the dataset is not a very good so we continue. We take the first five components as the new variables we are going to use in K-means. And we need to define the meaning of each component, so we plot the contributions of each variables to the principle components in Figure 14. Take the principle component 1 as an example we can see the area and hall and room and price have take most of the contribution so we can label this as "Gig house". As for the principle component 2 we can see that age, and direction that is not to south and district-Longgang and district-Pinshan have the most contribution so we label this as "houses that is very old and built in very poor area". At last we label these five principle components as "PC1: "big house"PC2: house that is very old and in places that is not developed PC3: house "fengshui" and comfort PC4: mansion, houses in rich area PC5: high floor"



(a) variances explained by each component

(b) contributions for dimension1

(c) contributions for dimension2

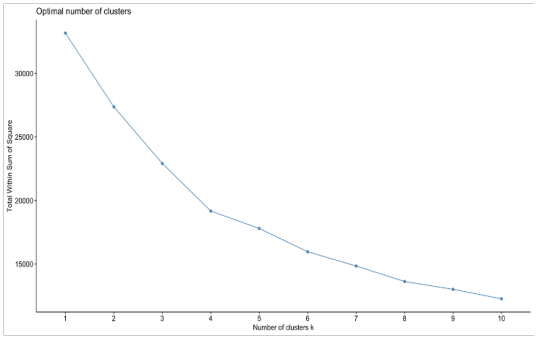
(d) contributions for dimension3

(e) contributions for dimension4

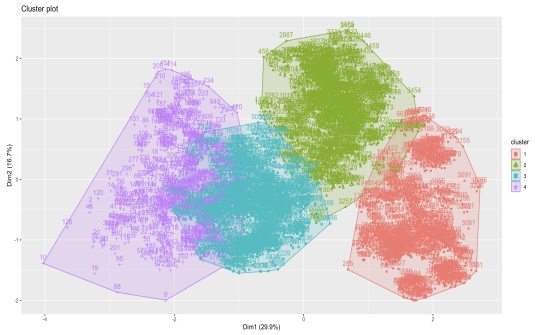
(f) contributions for dimension5

Figure 15: Visualization of PCA

After PCA we can use K-means with the results of PCA and same as before we use function in package factoextra for determining and visualizing the optimal number of clusters in Figure 16(a), we choose 4 clusters this time and we also plot the radar chart. We also visualise the result in 2 dimensional space as before, which is much better compared to the result before thus we believe PCA successfully eliminate the effects due to the correlation.



(a) Sum of squares against number of clusters



(b) Visualize Clustering Results

Figure 16: Visualization for the K-means after PCA

We can see that 4 clusters looks all different in Figure and this time there is no radar plot that have high or low value in all dimensions which prove that PCA works well, we can have a look at the cluster 1 and cluster 4, they have difference in PC1 and PC2 which are labeled as "Big houses" and "house that is very old and in places that is not developed" we can see that cluster 1 has higher value in PC2 and lower value in PC1 which make sense as there are less big houses in these poor districts that has more urban cities.

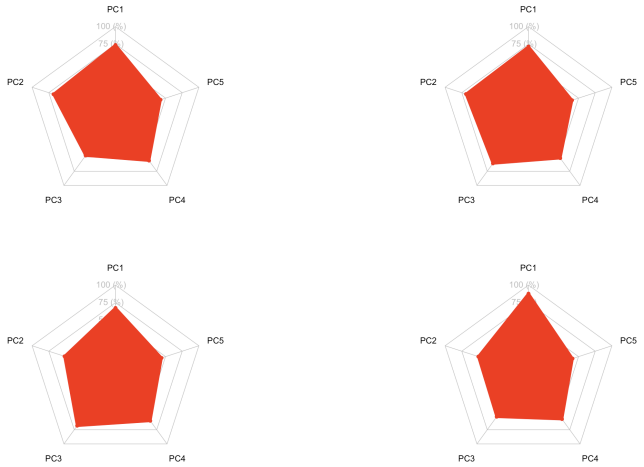
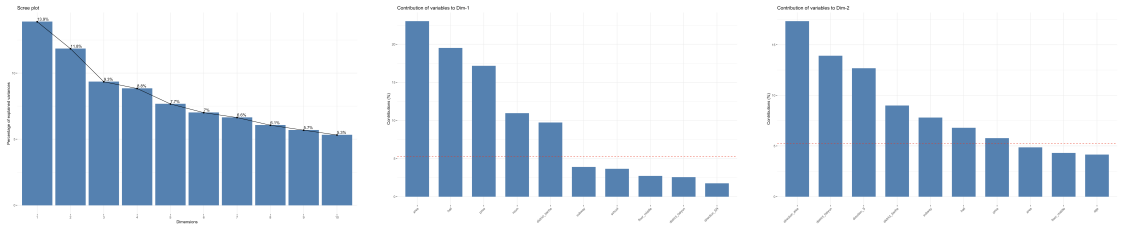
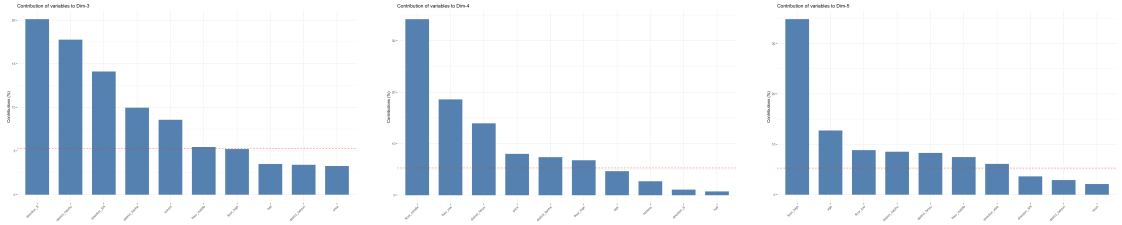


Figure 17: Radar chart

To make sure this is not an coincidence we do the PCA on data in Guangzhou too, in the same way we first dicide to ues first five components too and try to label each principle component by checking the contributions. And we label the principles components as"PC1 high quality apartment, PC2 houses in towns and villages, PC3 Urban villages, PC4 houses with low or middle floor, PC5 houses with high floor "



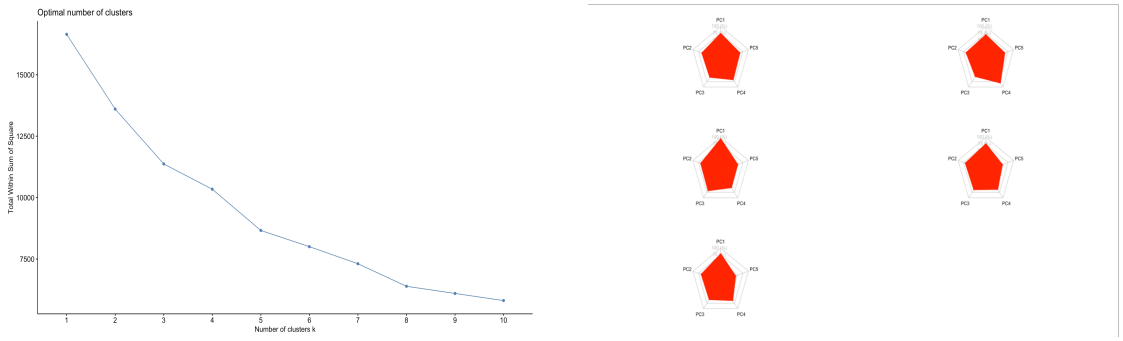
(a) variances explained by each component (b) contributions for dimension1 (c) contributions for dimension2



(d) contributions for dimension3 (e) contributions for dimension4 (f) contributions for dimension5

Figure 18: Visualization of PCA[Guangzhou]

After PCA we can use K-means with the results of PCA and same as before we use function in package factoextra for determining and visualizing the optimal number of clusters in Figure(a), we choose 5 clusters this time and we also plot the radar chart in Figure(b). After analysis we believe the results are very food and even has better explanation than data in Shenzhen, we think this is because Guangzhou has longer history.



(a) Sum of squares against number of clusters (b) Radar chart

Figure 19: Visualization for the K-means after PCA

6 XGboost Regression

We observed that the initial linear regression model, with housing prices as the target variable, did not perform well. We speculated that this could be due to a weak linear relationship between the variables and the target variable, and the presence of too many categorical variables. Therefore, we decided to use the non-parametric method, XGBoost regression, to see if the fit could be improved.

The improvement in R^2 was significant. For instance, in the case of the Xiamen dataset, without any optimization, R^2 increased from 0.66 to 0.71. Of course, we were not satisfied with this, so we tried several methods to further improve our R^2. Throughout this process, we consistently divided the dataset into a training set and a test set at a ratio of 7:3.

The optimization of the R^2 statistic was attempted through various methods:

- Parameter Tuning: The parameters of the XGBoost model were fine-tuned to achieve the best possible performance. This involved adjusting various hyperparameters such as the learning rate,

max depth of the trees, and the number of estimators, among others.

- Feature Selection: Certain features were selected based on their relevance and contribution to the model's predictive power. This process helped in reducing the dimensionality of the dataset and improving the model's performance.
- Feature Engineering: New features were engineered from the existing ones, and irrelevant features were removed. This process helped in capturing more complex relationships within the data and improving the model's predictive accuracy.

After making these adjustments, R^2 increased from 0.71 to 0.81, and the adjusted R^2 also improved from 0.71 to 0.80. The magnitude of the improvement is quite significant, and the current R^2 is already very impressive.

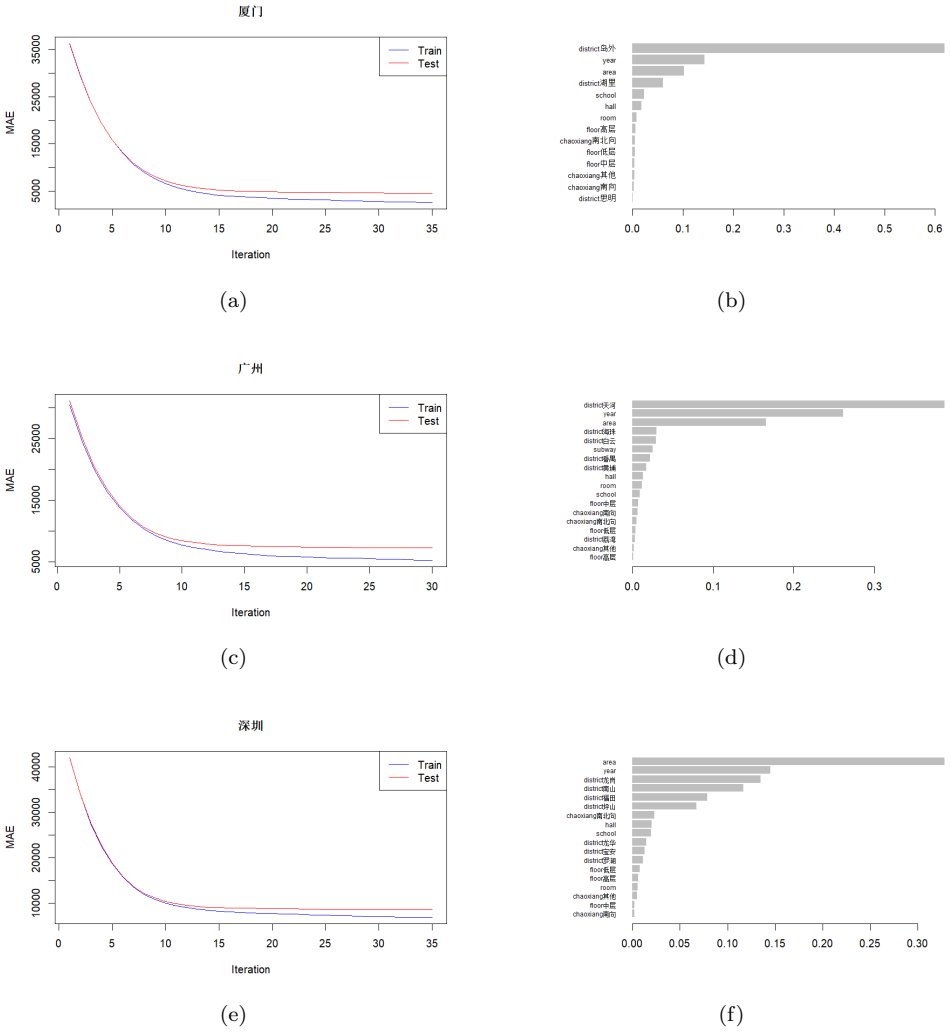


Figure 20: Final results of XGboost Regression

The above image shows the final optimized results of the XGBoost regression in the three cities. We used the Mean Absolute Error (MAE) as the test error.

7 Classification

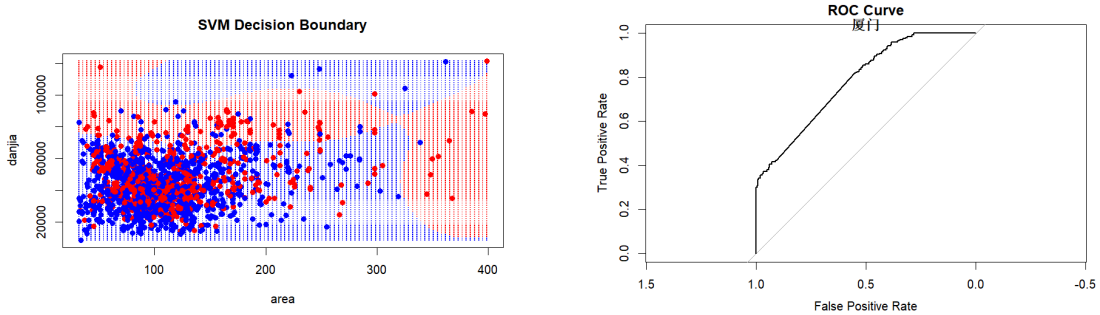
Also,we attempted a classification task, with education as the target variableThis is an important feature as houses in school districts often command higher prices due to the perceived value of nearby

educational facilities. The classification model was trained and tested on the dataset, providing valuable insights into the factors that influence a house's classification as a school district house.

7.1 SVM

First, we attempted classification using SVM, again using Xiamen as an example. Without any optimization, the accuracy was 0.7305

Similarly, we tried the three optimization methods mentioned in the XGBoost regression section above. However, the accuracy only increased to 0.732, indicating that the improvement was almost negligible.



(a) SVM decision boundary in danjia area in xm

(b) ROC curve of xm

Figure 21: Final results of SVM about xm

The above image shows the SVM decision boundary drawn on the plane of 'danjia area', along with the ROC curve. The results are far from ideal. In the left image, where the density is high, the classification effect is poor, and only in areas of low density is the classification somewhat successful. The ROC curve also does not yield very good results. We speculate that this is because there are too many categorical variables, which is why SVM does not perform well.

7.2 XGboost Classification

Due to the unsatisfactory performance of SVM, we revisited our previous approach and applied XGBoost for classification on the Xiamen data. We found that the accuracy indeed improved significantly. When using objective = "binary:hinge", it reached approximately 0.8.

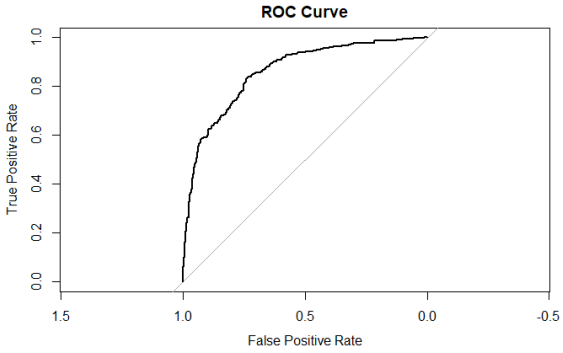
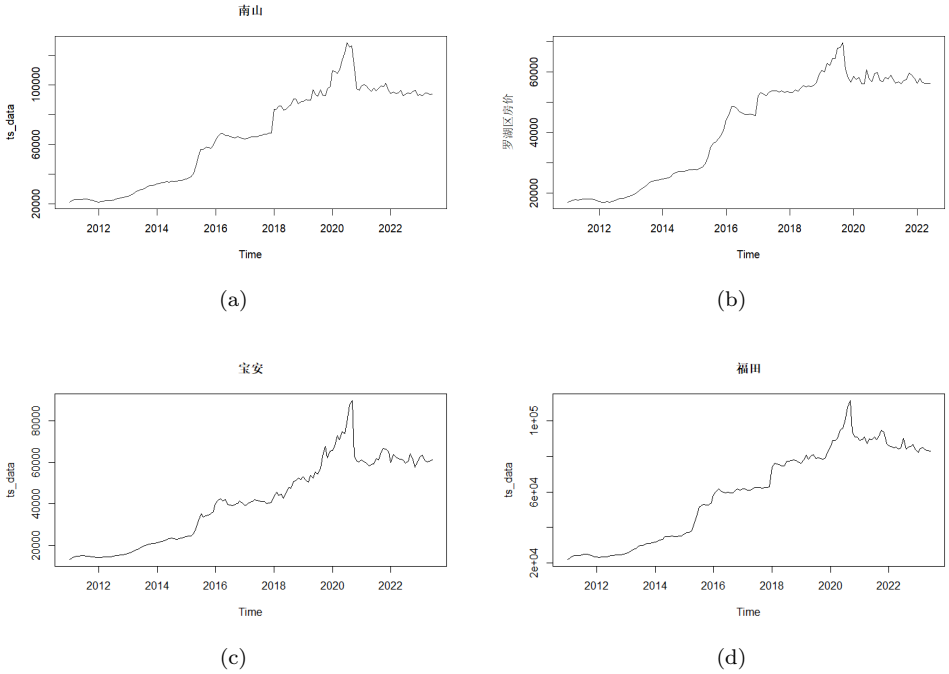


Figure 22: XGboost Roc curve of xm

This is the ROC curve of XGBoost using objective = "binary:logistic". As you can see, it performs better than SVM.

8 Time series Analysis

In addition to the provided dataset, we also sought out an additional dataset that includes second-hand housing prices in various districts of Shenzhen from 2011 to 2023. We conducted some time series analysis on this dataset.

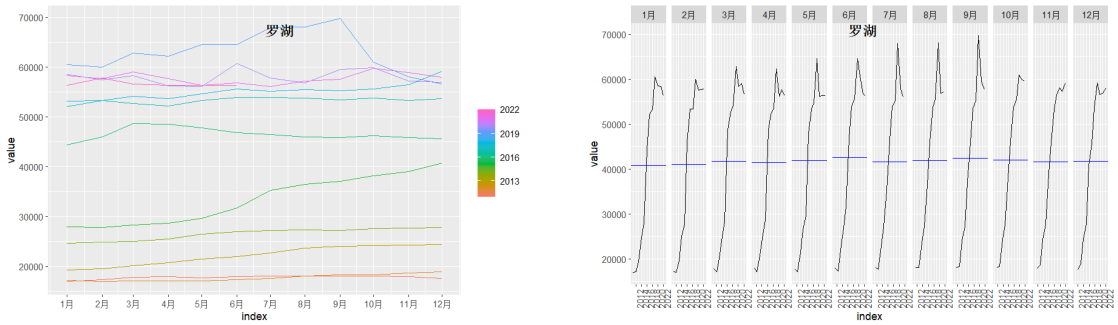


(a) (b)

(c) (d)

Figure 23: Time series

This is a graph showing the change in second-hand housing prices over time in four districts. As can be seen, their trends are almost identical - they first rise, then fall, and are now in a relatively flat stage.



(a) (b)

Figure 24: Seasonal decomposition plot

The two images above are seasonal decomposition plots of the time series of lh. In the left image, lines of different colors represent different years, and we can see that these lines generally do not intersect. The right image is a seasonal subplot, where we can see that the pattern that emerges for each month is the same. Based on these two images, we can determine that this time series has strong seasonality with a cycle of one year. Furthermore, we can observe that the trend within a year is generally to rise first and then fall, with two peak values.