STATS 503 Final Project: Iowa Housing Price

Abstract

Housing Data of the city of Ames, Iowa are addressed by the application of machine learning methods in three aspects: clustering of neighborhoods, classification of housing quality and prediction on future sale prices. Above all, a spatial structure of transactions is visualized as maps to help analyze diagnostic statistics. Then, hierarchical clustering and K-means are applied for shrinking the number of neighborhoods and strengthening the meaning behind each neighborhood. Next, a reiteration of experts evaluation for housing quality and condition can be achieved through classification. Last, various regression models are implemented to predict housing prices, and regularization regression models perform better. Our results demonstrate a flexible and competitive approach to other housing issues in general.

Contents

1 Introduction

House sales are determined based on the Standard & Poors Case-Shiller home price indices and the housing price index of the Office of Federal Housing Enterprise Oversight (OFHEO). These reflect the trends of the US housing market. In addition to these housing price indices, the development of a housing price prediction model can greatly assist in the prediction of future housing prices and the establishment of real estate policies. This study uses machine learning algorithms as a research methodology to develop a housing price prediction model. Machine learning has been used in disciplines such as business, computer engineering, industrial engineering, bioinformatics, medical, pharmaceuticals, physics, and statistics to gather knowledge and predict future events. With the recent growth in the real estate market, machine learning can play an important role to predict the price of a property. However, few researchers have experimented on the selling price for real estate properties using machine learning algorithms.

It is a well-known fact that housing price valuation is one of most important trading decisions affecting a national real estate policy. In this study, we create models using machine learning algorithms such as hierarchical clustering, K-means++, LDA, Knn, SVM, decision tree, random forest to explore the inner connection between variables. Moreover, we apply regression methods include GLM, polynomial, ridge, lasso and elastic net to predict the housing price.

The remainder of this paper is organized as follows. Section 2 introduces the data source of our study. Section 3 summarizes the data. Experimental results are presented and analyzed in Section 4, and finally, our concluding remarks are provided in Section 5.

2 Data Source and Selection

For the experiment, we used the real estate datasets from two different sources: address-report.com [1] and Kaggle.com [2]. The training data includes 1460 observations of 81 explanatory variables describing different aspects of residential homes in Ames, Iowa. The test data has 1459 observations. Table 1 and Table 2 are the lists of variables selected.

Table 1: List of physical feature variables selected.

Category	Attribute	Descriptions	Data type
Physical features (15)	Basement	Features of basement (Height, general condition, exposure, area)	nomial, numerical
	Bathroom	Features of bath (basement/above grade full/half bathroom)	numerical
	Bedroom	Number of bedrooms	numerical
	Exterior	Features of exterior (meterial, quality, condition)	nomial
	Fireplace	Features of fireplace (number, quality)	nomial, numerical
	Garage	Features of garge (location, year built, capacity, size quality)	nomial, numerical
	Heating	Features of heating (type, quality)	nomial
	Storey	number of stories	numerical
	Roof	Features of roof (type, material)	nomial
	Lot	Features of lot (distance, size, shape, configuration)	numerical, nomial
	Kitchen	Features of kitchen (number, quality)	numerical, nomial
	Pool	Features of pool (size, quality)	numerical, nomial
	Garden	Features of garden (porch, fence, wooddeck, etc.)	numerical, nomial
	Masonry veneer	Features of Masonry veneer (size, type)	numerical, nomial
	General	Total area, total rooms, sale type, land contour, etc.	numerical, nomial

Table 2: List of neighborhood feature variables selected.

Category	Attribute	Descriptions	Data type
Neighborhood (6)	County	Name of county	nomial
	Cost of living	Cost of living in 2010	numerical
	Income	Average income in 2010	numerical
	Annual property tax	Annual property tax in 2010	numerical
	School	Number of schools nearby	numerical
	Crime score	Crime score in 2010	numerical

3 Summary

3.1 Descriptive statistics

Figure 1 shows the distribution of the sale prices and the number of transactions on geometric scale. The black polygons are the administrative regions of the city of Ames. In addition, there are some data from the distant area, so some of the points locate outside the city. In terms of the Sale Prices, the highest price appears in the north of the city, which can be told by the size and the color of the point, and generally, the houses from the north of the city have a better price than the other area. The trees and the lake in the north might be a possible reason. With respect to the frequency of transactions, the higher frequency means the more active house market. Therefore, it is not surprising the large numbers of transactions appear in the downtown and the Iowa State University. However, transportation seems to influence neither the sale prices nor the frequency of transactions. The areas close to the main roads do not demonstrate special patterns. In conclusion, the influential factors of higher sale prices and the larger number of transactions seem to be opposite. Residents of the city prefer to live in the area surrounded by natural views, while the housing market is more active in the center of the city.

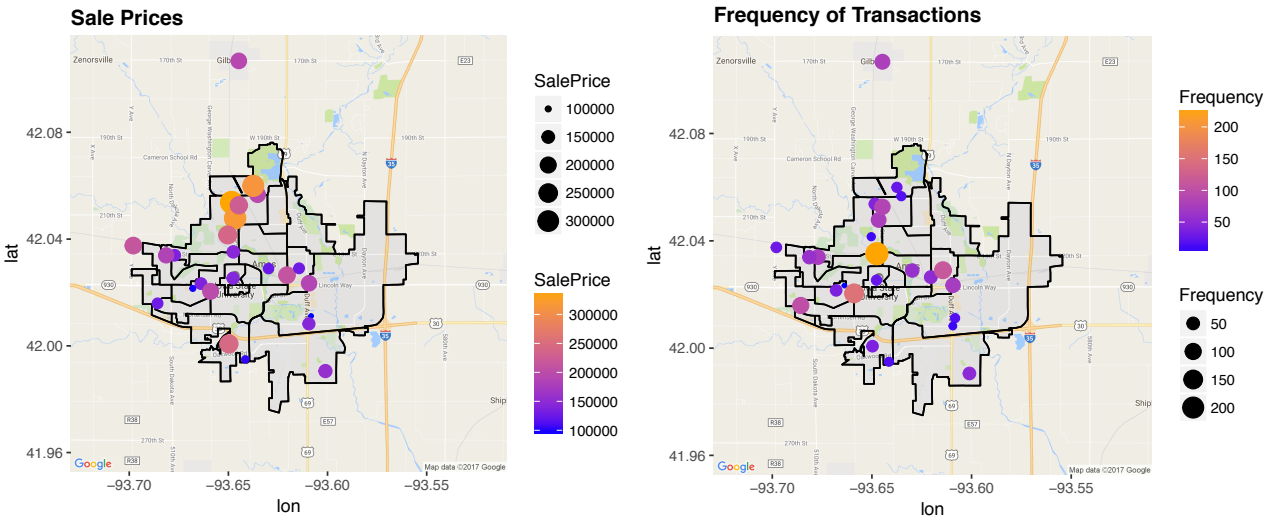

Figure 1: Maps of Ames city in geometric scale

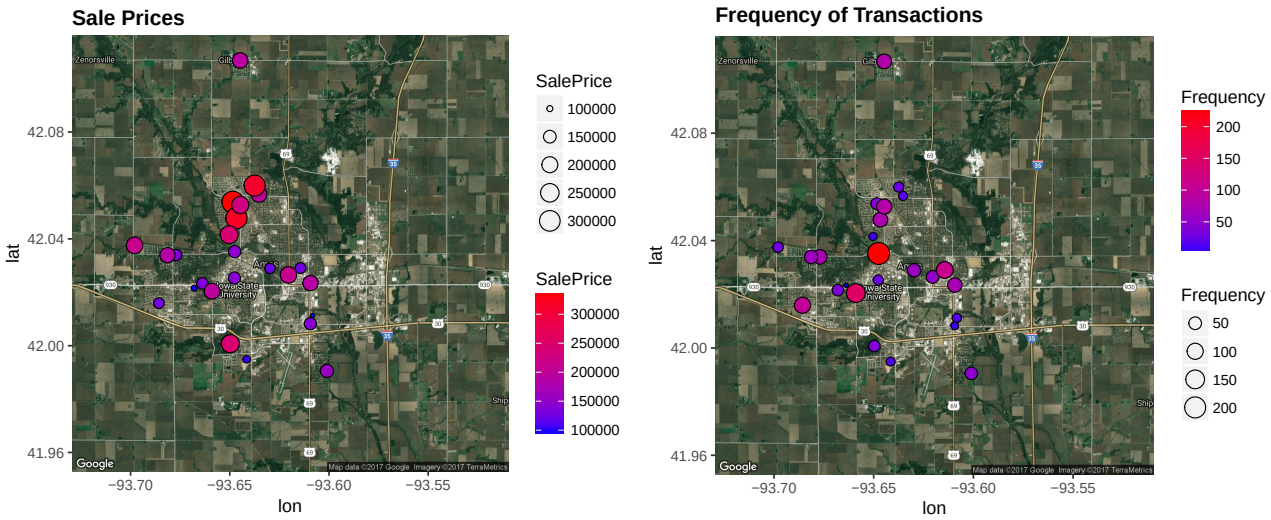

Figure 2: Real maps of Ames city

Figure 3 is a histogram of the SalePrice which is heavily right skewed. Thus, we made a log transformation to normalize the variable in order to satisfy the assumption of linear regression to make better predictions.

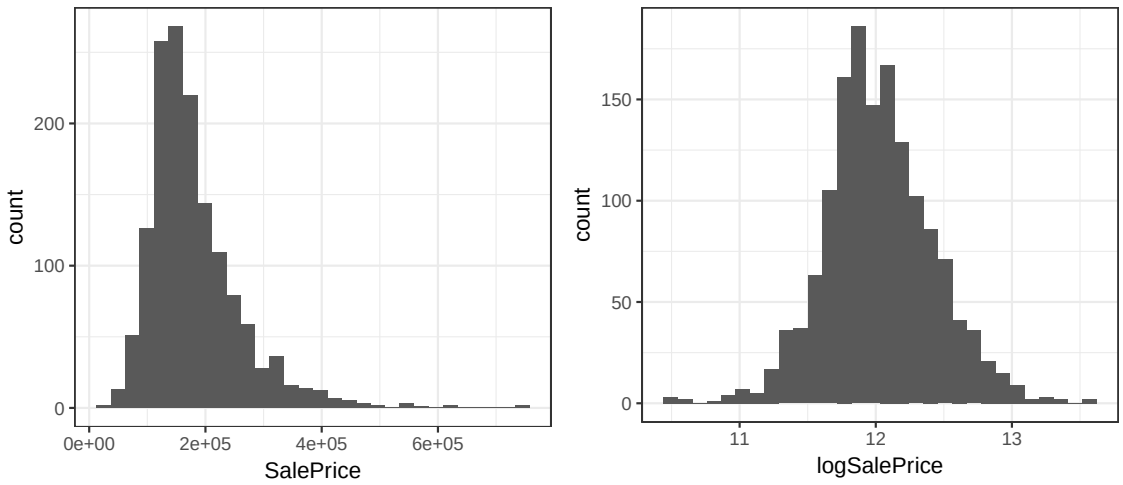

Figure 3: Histograms of original and log transformed Sale Price

Figure 4 shows eight boxplots of certain variables against SalePrice in order to confirm our understanding of what may affect a houses sale price. The following boxplots show the variable with different levels has the obviously different sale price. Houses with more full bathrooms, finished garages, central airconditioning, kitchen, fireplace, pool, basement, the material of exterior with excellent quality tend to yield higher sale prices.

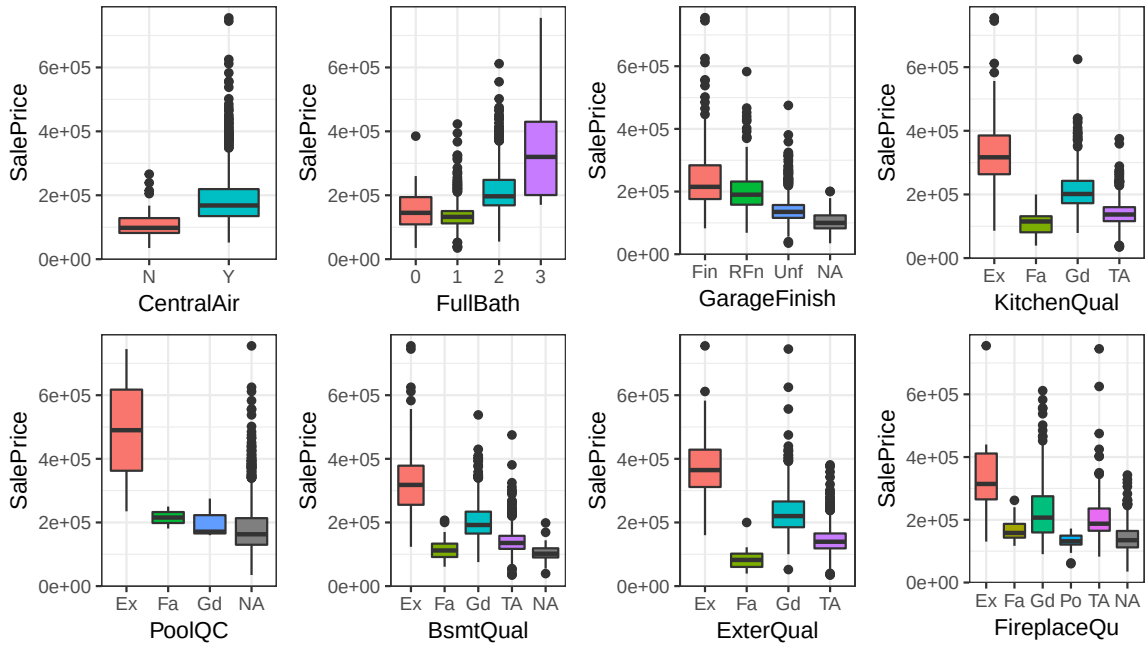

Figure 4: Boxplots of selected variables against sale price

Figure 5 shows the boxplots of OverallQuality and OverallCondition against SalePrice. Both overall quality and overall condition have 10 levels (1-10). Based on the first plot, we conclude that the higher the quality level the higher houses sale price. Based on the second plot, there are positive correlation between sale price and overall condition. However, since most housing conditions are level 5, level 5 housing price is right skewed.

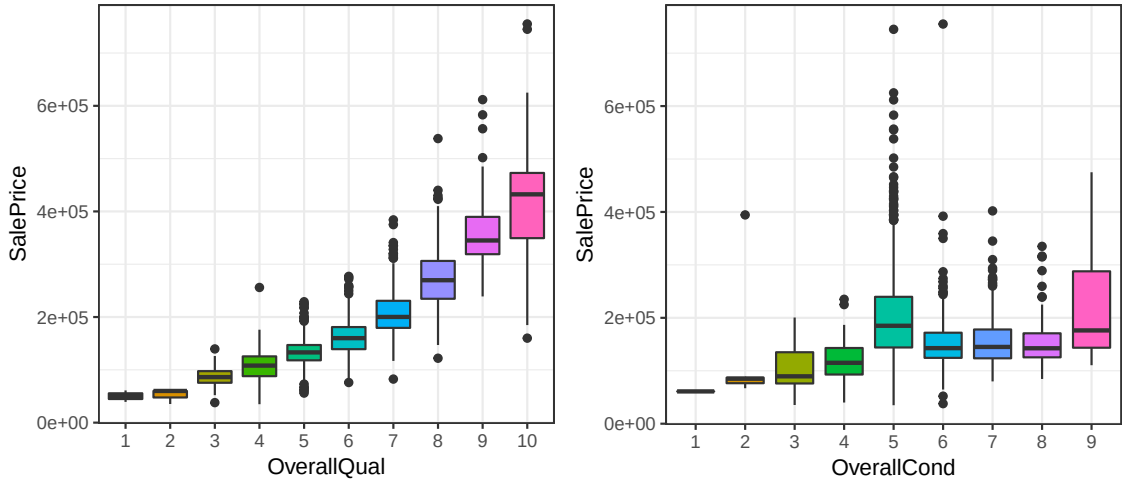

Figure 5: Boxplots of overall quality and overall condition against sale price

Figure 6 shows the correlation between numerical variables. In general, the area variables are positively correlated, while the age of house is negatively correlated with variables such as garage area.

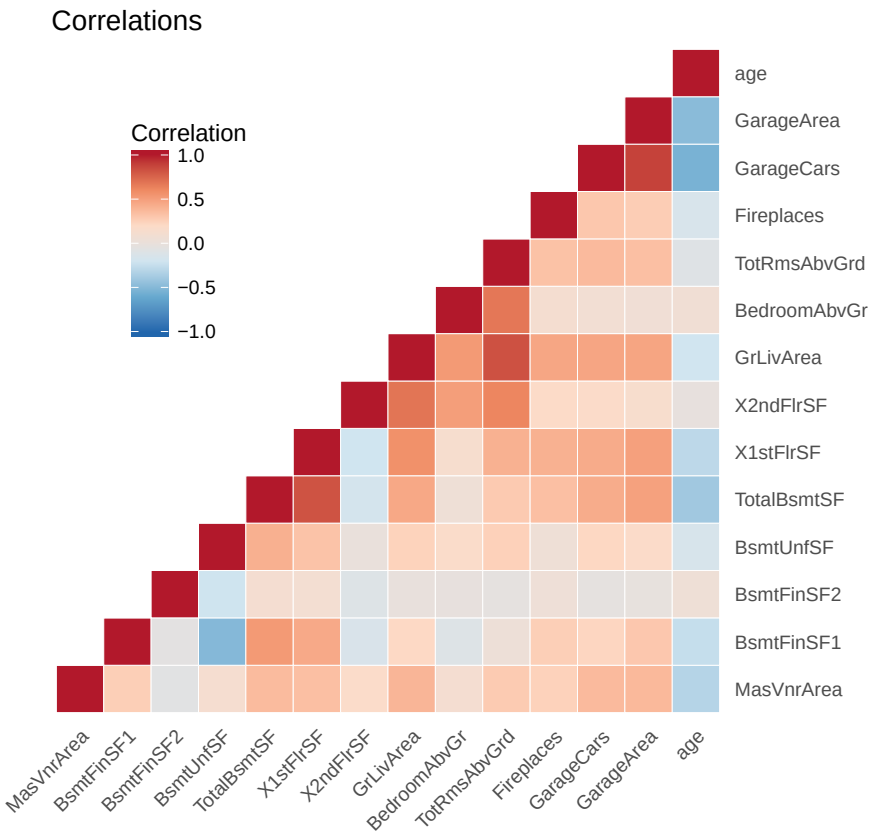

Figure 6: Heatmap of selected numerical variables

3.2 Data cleaning and integration

Since the dataset has many missing values and large amount of categorical variables, data cleaning and transformation are essential. Most missing values appear at categorical variables of equipments, so we use the area of the equipments to measure the type of them for filling in the missing values. For example, if the Fireplaces is 0, we will regard FireplaceQu to be NA(No Fireplace). After data cleaning, we use "One hot encoding" method to deal with the categorical variables. Also, for those variables with only one level significantly higher than the rest levels, we consider that there is no level for this variable. For instance, the number of the house on Pave street is 1454 and on Grvl street is 6, we change the variable to be StreetPave as a binary variable, Pave is 1 and Grvl is 0. Moreover, we do clustering on the neighbor variable to set up new neighbor level, since there are too much original levels, and thus, less levels provide more meaningful information and more suitable for the following analysis.

4 Experimental results

4.1 Clustering: Neighborhood

There are 28 communities and roads in total. Neighborhood features include Location (rough position in city: north, south, east, west and center), living cost (in dollar), Income (in dollar), owners (total number of families), annual property tax (in dollar), School (number of schools nearby), crime score and ville (indicator). Housing price features include mean, median, standard deviation and total number of houses sold. In terms of the missing values in some commuinties and roads, the modes of variables are used to fill them up.

4.1.1 Hierarchical Clustering

In data mining and statistics, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. We use gap statistic method [3] to estimate the number of clusters. The gap statistic for a given k is defined as follow:

$$Gap_n(k) = E_n^*\{log(W_k)\}$$

Where E_n^* denotes the expectation under a sample of size n from the reference distribution via bootstrapping. So The gap statistic measures the deviation of the observed W_k. Results are shown in Figure 7. We can conclude that there does not exist clear clustering between neighborhoods, so we will use all variables in regression.

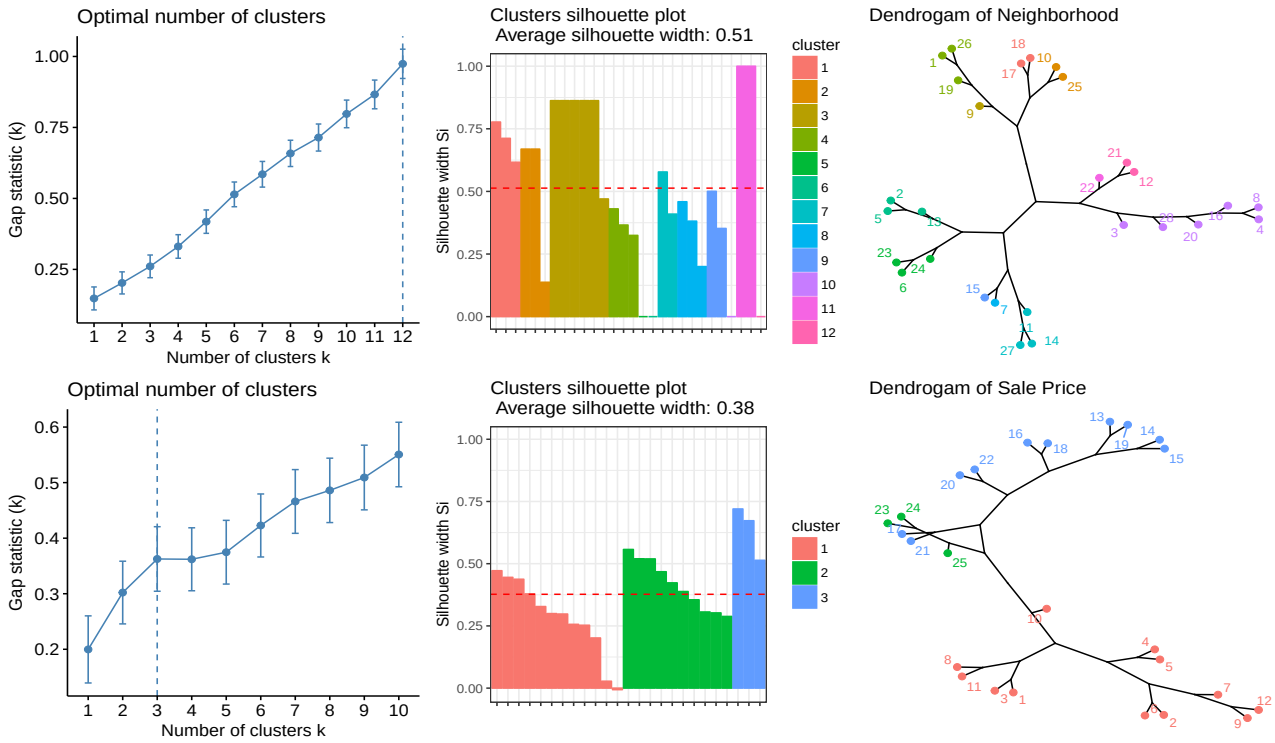

Figure 7: Hierarchical Clustering

4.1.2 K-means++

K-means++ is an algorithm for choosing the initial values (or "seeds") for the k-means clustering algorithm. The k-means method is a widely used clustering technique that seeks to minimize the average squared distance between points in the same cluster.[4] For visualization, we set the the number of clusters as 3.

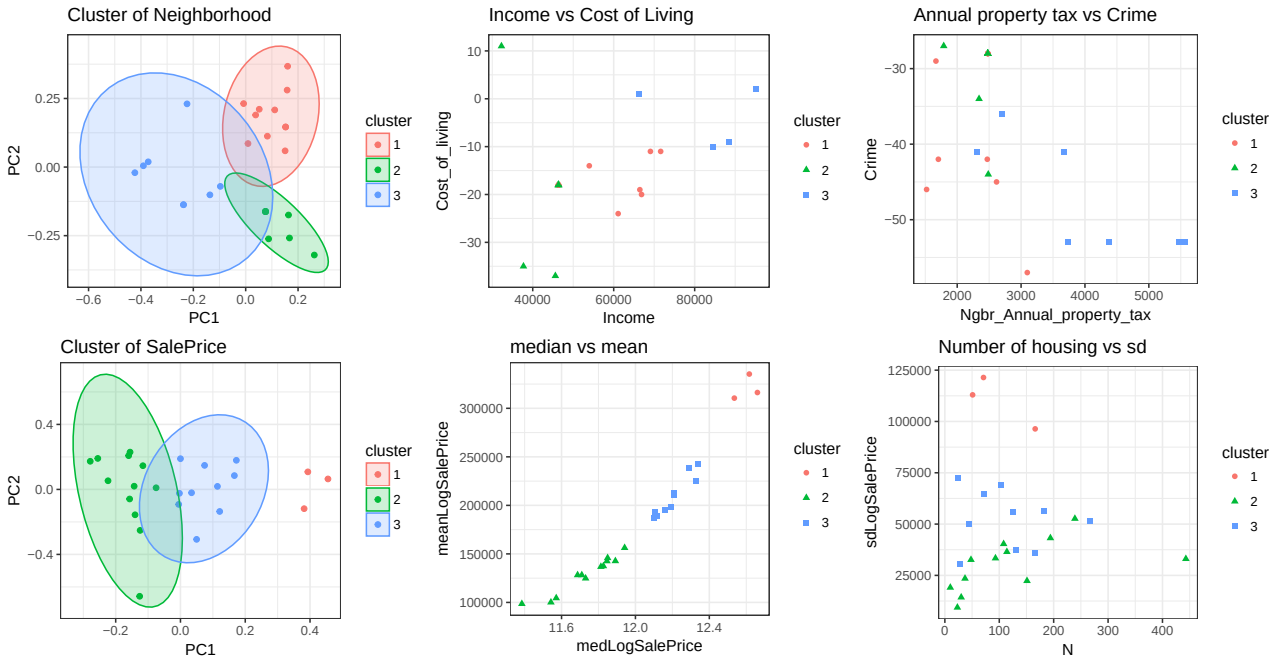

Figure 8: Cluster by K-Means++

Table 3: Locations in each neighborhood cluster.

Cluster	Location
1	Blueste, ClearCr, CollgCr, Crawfor, Landmrk, Sawyer, SawyerW
2	BrDale, BrkSide, Edwards, GrnHill, IDOTRR, MeadowV, Mitchel, NAmes, NWAmes, OldTown, SWISU, Timber, Veenker
3	Blmngtn, Gilbert, Greens, NoRidge, NPkVill, NridgHt, Somerst, StoneBr

Table 4: Locations in each sale price cluster.

Cluster	Location
1	StoneBr, NoRidge, NridgHt
2	MeadowV, IDOTRR, BrDale, OldTown, Edwards, BrkSide, Sawyer, Blueste, SWISU, NAmes, NPkVill, Mitchel
3	SawyerW, Gilbert, NWAmes, Blmngtn, CollgCr, ClearCr, Crawfor, Veenker, Somerst, Timber

Figure 8 indicates:

- Neighborhood

    1. Projection on first two principle components has clear clusters, however, since first two PCs can only explain 55% proportion of variance, we use all 7 variables in regression.
    2. Three clusters respectively include 7, 13 and 8 locations. The higher the income is, the higher the cost of living, however, there exists exception that low income has high cost of living. The higher the annual property tax, the lower the crime score. Locations with higher income have higher cost of living, higher tax and less crime.

- Sale Price

    1. Projection on first two principle components has three clusters, low, median and high. Since first two PCs can only explain 95% proportion of variance, we use the clusters for regression.
    2. Sale price has three clusters, low, median and high. Median price and mean price are linearly correlated. Expensive house has larger standard deviation. Popular houses has lower prices with smaller variance.

4.2 Classification

- Overall Quality
  We use variable, OverallQual as our class since the overall quality of a house is based on all the house factors. Thus, we suggest that to use all the variables according to the house condition to decide the quality. For classification, we choose 5 levels instead of 10 due to the limitation of methods. We suggest to combine two levels to one, so the five classes are, poor, below average, average, above average, and good. The size of data inspire us to reduce dimension.However, dimension reduction including PCA, FA and MDS fail to perform well. Therefore, we decide to use 160 variables to fit the dataset.

  Now, we have totally 166 variables, which is quite large, but after the transformation the dataset will be easier for the following process. Linear Discriminant Analysis(LDA), K-nearest neighbors(Knn), Support Vector Machines(SVM), Decision Tree, and Random Forest are used to fit the dataset. Table 5 shows that SVM with kernel=radial and cost=2.718282 has the lowest training error, 0.056. In comparison, LDA and SVM have almost the same testing errors, though the training errors are different. We suggest that LDA and SVM perform better than the rest models.

Table 5: Error Rate: Overall Quality

	LDA	Knn	SVM			Tree	Forest
			Linear	Radial	Polynomial		
Train Error	0.182	0.240	0.117	0.056	0.082	0.219	0.210
Test Error	0.225	0.304	0.230	0.220	0.223	0.242	0.211

Based on the information of random forest, we know the importance of variables. Thus, we decide to use those variables, whose MeanDecreaseGini is larger than 10(Table 2). We only use 19 variables, ExterQual, YearBuilt, KitchenQual, GarageYrBlt, Tota- lArea1stAnd2nd, GrLivArea, TotalBsmtSF, YearRemodAdd, GarageArea, TotalArea, LotArea, BsmtUnfSF, OverallCond, BsmtFinSF1, LotFrontag, TotRmsAbvGrd, WoodDeckSF, FoundationPCon, and OpenPorchSF, to fit the model. Table 6 is the training error and testing error of the different classification models with 19 variables. Ir shows the higher error rate than Table 5 does, no matter which method we used. However, the error rate of Table 7 is close to Table 5. Therefore, we suggest that the 19 variables contain the most information and the performance is close to that of 159 variables.

Table 6 shows the importance of variable and most of the variables in the Table 3 are about area size. Therefore, we suggest that the size of different kinds of area is the determining factor.

Table 6: Variable Importance of Overall Quality: MeanDecreaseGini

	ExterQua	YearBuilt	KitchenQual	GarageYrBlt	TotalArea1stAnd2nd
MDGini	114.690	71.472	46.639	43.768	42.352
	GrLivArea	TotalBsmtS	YearRemodAdd	GarageArea	TotalArea
MDGini	41.471	29.648	25.416	25.053	22.149
	LotArea	BsmtUnfSF	OverallCond	FoundationPConc	BsmtFinSF1
MDGini	20.074	19.003	18.134	15.916	13.780
	LotFrontage	TotRmsAbvGrd	WoodDeckSF	OpenPorchSF	
MDGini	12.987	12.508	11.921	11.856	

Table 7: Error Rate: Overall Quality with less variables

	LDA	Knn	SVM			Tree	Forest
			Linear	Radial	Polynomial		
Train Error	0.221	0.319	0.176	0.206	0.162	0.216	0.222
Test Error	0.246	0.354	0.218	0.230	0.227	0.240	0.215

- Overall Condition

  Focusing on overall condition, same methods including LDA, Knn, SVM, decision tree

and random forest for classification are applied. The error rates of different methods(Table 8) are really close to the overall quality performance(Table 5). SVM with kernal=radial and cost=2.718282 has the lowest training error, 0.045 and the testing errors of all models are around 0.2. Again, we use the same way to select and keep those variables, whose MeanDecreaseGini is larger than 10(Table 9). The results of the models with merely 19 variables are in Table 10 and the performance of these variables is almost as good as the 159 variables. In addition, Table 9 shows the importance of variable and the error rate of most of the variables in Table 9 are about area size. Therefore, we suggest that the sizes of a different kind of area are the determining factor.

Table 8: Error Rate: Overall Condition

	LDA	Knn	SVM			Tree	Forest
			Linear	Radial	Polynomial		
Train Error	0.171	0.258	0.116	0.045	0.251	0.207	0.212
Test Error	0.221	0.247	0.232	0.213	0.245	0.210	0.190

Table 9: Variable Importance of Overall Condition: MeanDecreaseGini

	FoundationPConc	BsmtQual	YearRemodAddBefore50	BsmtFinSF1	LotFrontage
MDGini	10.424	10.863	11.631	12.347	12.389
	KitchenQual	OverallQual	ExterCond	GarageArea	LotArea
MDGini	12.606	13.071	14.761	15.343	16.265
	GrLivArea	TotalArea	TotalArea1stAnd2nd	BsmtUnfSF	TotalBsmtSF
MDGini	16.525	16.634	16.663	17.103	17.608
	isRemod	GarageYrBlt	YearRemodAdd	YearBuilt	
MDGini	22.708	26.777	39.507	45.406	

Table 10: Error Rate: Overall Condition with less variables

	LDA	Knn	SVM			Tree	Forest
			Linear	Radial	Polynomial		
Train Error	0.214	0.266	0.208	0.138	0.064	0.207	0.201
Test Error	0.212	0.249	0.192	0.206	0.224	0.210	0.204

4.3 Regression

4.3.1 Model Introduction

To predict the sale prices of houses, several approaches of regression are considered: GLM, stepwise regression, polynomial regression, ridge regression, lasso regression and elastic net regression. The most important issues in the dataset in terms of predicting sale prices

are collinearity and high dimensions. With respect to the two problems, regularized linear regression methods like ridge, lasso and elastic net regression seem to be more appropriate because these models add a penalty term on overfitting. On the other hand, aiming for prediction instead of interpretation, the issue of collinearity is endurable if using a GLM model. Also, the sale prices transformed by logarithm follow a Normal distribution so that a GLM model with Gaussian error distribution might fit well. In addition, it is reasonable to find a model only containing a few influential variables due to a large number of variables in total. Thus, using the stepwise approach, a model is selected by comparing the AICs of models with different variables, and there are only 84 out of 192 variables are included. Some of the variables are excluded because of the collinearity. Apparently, it is a simpler model, which can help check the issue of overfitting. When it comes to the elastic net regression, use cross-validation to choose the shrinkage parameter λ and elastic net mixing parameter α. The optimal λ and α are 0.02154435 and 0.1578947. And the results of cross validation with RMSE against regularization parameter are plotted.

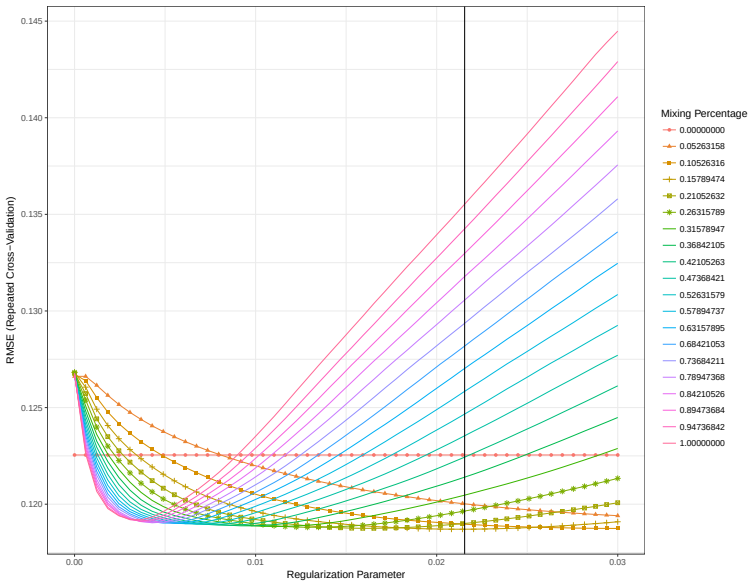

Figure 9: Cross-validation of predictors in elastic net model

4.3.2 Training via Cross-Validation

For choosing the optimal regression model, the train dataset is randomly separated into two parts; that is, 85% observations of the dataset is regarded as "train" data and the rest 15% as "test" data. Then, compare the RMSE of the sale prices transformed by logarithm in each model. From the Table 11, the average RMSEs of different models are similar except for the polynomial regression. Obviously, the model overfits the data because the RMSE of the train is low, while the RMSE of the test is high. In addition, based on the RMSEs of the train, the performance of stepwise regression is comparably worse than the others. Since less variables are included in this model, it is not surprising that the RMSE is higher. As a

result, except for the polynomial regression method, the others seem to be sufficient for the train dataset, so we fit the test dataset by using all of the five models for further analyses.

Table 11: RMSE performance under different models

RMSE	GLM	Stepwise	Polynomial	Ridge	Lasso	Elastic Net
Train	0.1002	0.1019	0.0872	0.1002	0.1017	0.1068
Test	0.2126	0.2132	0.5687	0.2172	0.2176	0.2123

4.3.3 Prediction

Although the RMSE of the ordinary test dataset should not be a standard of picking an optimal model, it is an idea to regard the ordinary test dataset as another train dataset, so the result of how these models perform under another dataset might accommodate a suggestion of picking one from the models. The following analyses are based on the ordinary test dataset. It is rather clear that Lasso regression is the best. There are two ideas to obtain fitted values: one is to use the models directly, and the other is to improve the model via generalized boosted regression model (GBM)[5] to get fitted values, take the two lists of fitted values as new variables, and use random forest to regress on the new test dataset.

Algorithm 1 Iterative Prediction Algorithm

1: Use one of the 5 models and GBM to obtain 2 lists of fitted values.
2: Take the two lists of fitted values as new variables in the test dataset.
3: Apply random forest regression to predict.
4: Compute RMSE.

Generally speaking, the predicted and the given sale prices transformed by logarithm are so close that the low RMSEs seem to indicate that either model calculated by either idea is good enough, though the lasso model is the most accurate. Surprisingly, the stepwise regression performs slightly better than the full model, which means more variables mislead the prediction on the contrary. The consequence is valuable because the stepwise model has addressed the collinearity, so the model is able to be readily interpreted, when the full model is hardly explained. Comparing the two ideas of fitting observations in test dataset, the first one is the ordinary way to fit, and the process is straightforward; the second one is at risk because it strengthens the features already figured out, so if the model is improper, the results will be worse, and in this case, the accuracy of prediction is outstandingly improved. In terms of predicting the sale prices transformed by logarithm, the RMSE of the first idea is nearly twice as much as the RMSE of the second idea, and the table below also shows the RMSE of sale prices. Basically, good models perform well by using either idea, but there is a difference, which is that the Elastic Net regression model is the second best model if using the complicated predicting method, while it is the worst if using the ordinary predicting way.

Table 12: RMSE performance under different ideas

RMSE	GLM	Stepwise	Ridge	Lasso	Elastic Net
1st idea	0.0778	0.0776	0.0754	0.0712	0.0778
2nd idea-log(SalePrice)	0.0319	0.0316	0.031	0.0297	0.03
2nd idea-SalePrice	6772.747	6675.019	6540.817	6022.359	6349.188

To sum up, lasso regression model is the best among all of the models in both fitting the train dataset and predicting the sale prices of the test dataset. At the beginning, since the collinearity among variables and the large number of variables are identified, regularized linear regression methods are recalled immediately, and thus not surprisingly, ridge, lasso and elastic net regression have desirable conclusions at last.

5 Conclusion and Discussion

To our curiosity, the paper emphasizes the contemporary environment of housing. For excavating the association among covariates and the information behind examination of the original dataset leads to discussion and improvement of the data processing. Reduced clusters of neighborhoods diminish the misleading information and magnify the important features, various classification methods attempt to interpret the overall housing quality and condition with limited covariates, and regression models construct clear numeric future sale prices by all the properties of houses.

To begin with, clustering can be regarded as a part of data organizing in order to enhance the accuracy of future sale prices predicted by regression models. Before clustering, some detailed information is cleaned, and some related information are updated. In the ordinary dataset, the neighborhoods are basically separated by geographic regions. The results are straightforward yet trivial. On the other hand, the locations of houses are valuable but rough observations. Hence, they are replaced by covariates including crime rates, costs of living, the number of schools, transportation and so on. Besides, sale prices are considered when clustering. At last, new clusters of neighborhoods are built up, and they contribute a huge partition of prediction.

As we mentioned, classification results might have connection with the actual evaluation of houses by experts. Therefore, we use different classification models to respectively fit the two covariates, Overall Quality and Overall Condition. Aside from sale prices, they tend to provide key features since they are scored by a number of factors: room sizes and ownership of a pool or a basement. However, the results of classification do not achieve the expectation. Although the training errors can be small, the testing errors hardly decrease. The testing errors are mostly around 0.2, which is nice but not satisfactory. Sorts of reasons might be causes, but one of them might be the dilemma of the method for data processing. While the prediction of sale prices in regression models improves, some of information sacrificed

in the previous process might cause to imprecise classification consequences with respect to Overall Quality and Overall Condition.

As combining every puzzles into a picture, we exploit information obtained so far by fitting regression models and producing prediction of sale prices. There are six disparate regression models, including GLM, stepwise, polynomial, ridge, lasso and elastic net, and evaluated by RMSE. To our surprise, exercising cross validation on the training dataset does not choose an obviously optimal model; five out of six models have small RMSE excluding the polynomial model. Thus, we compare the predicting performance by computing the RMSE between the true sale prices and the fitted values. There are two ways of predicting the testing dataset: one is to directly apply the models, and the other is to optimize the previous models with GBM and random forest. Finally, the lasso regression model slightly outweighs the others.

As a result, satisfactory and meaningful conclusions are generated through the techniques, i.e. clustering, classification and regression. Clustering deeply dug in the original dataset, classification evaluates the overall performance of houses with ordinary responses, and regression transforms information into precise numeric outputs. In addition, the process of analyses is forthright that it is reasonable and adjustable to address similar datasets. Therefore, not only for the dataset but also for some other open housing datasets, both the conclusion and the process provide potentials of further development.

Our study has limitations which future research could examine further, since we only focuses on a specific region, city of Ames, and on a specific type of residential properties, townhouses. In future works, this study can be extended in several ways. To begin with, it could be desirable to investigate other problem domains (real estate market prediction, economic growth rate forecasting and stock price index forecasting) to generalize the results of this study. In addition, the housing market can be influenced by macroeconomic variables. Future research should consider macroeconomic and environmental amenities variables for housing price prediction model inputs. For this purpose, more data sources are needed.

Reference

[1] "Addressreport." `https://www.addressreport.com`.

[2] "Kaggle." `https://www.kaggle.com/c/house-prices-advanced-regression-techniques`.

[3] G. W. Tibshirani, Robert and T. Hastie., "Estimating the number of clusters in a data set via the gap statistic.," *ournal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63.2, pp. 411–423, 2001.

[4] D. Arthur and S. Vassilvitskii., "k-means++: The advantages of careful seeding.," *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics*, 2007.

[5] G. Ridgeway, "Generalized boosted models: A guide to the gbm package.," 2007.