4741 Project Midterm Report

Hanyu Li, Bo Zhang, Zhongyuan Ma, Liyun Wang

Data Set Description

The data set used by the team describes the sales of individual residential property in Ames, Iowa from 2006 to 2010. It has 1460 entries in total, with 79 explanatory variables involved in assessing house values as well as the prices houses are sold at. The overall quality of the dataset is great, with only a few data entries missing data. This data set was published by a researcher Dean De Cock at Truman State university, who obtained the data directly from the City Assessor's Office at Ames, Iowa. He removed variables that required special knowledge or previous calculations for their use from the original data.

The variables presented in this data are a mix of nominal, ordinal, discrete or continuous values. Attempt was made to visualize some of the variables in Julia to obtain overall understanding. Since there are 79 explanatory variables in this dataset, a few was visualized and only two are presented in this report. Figure 1 and 2 shows the frequency of sale prices and house types in training set.

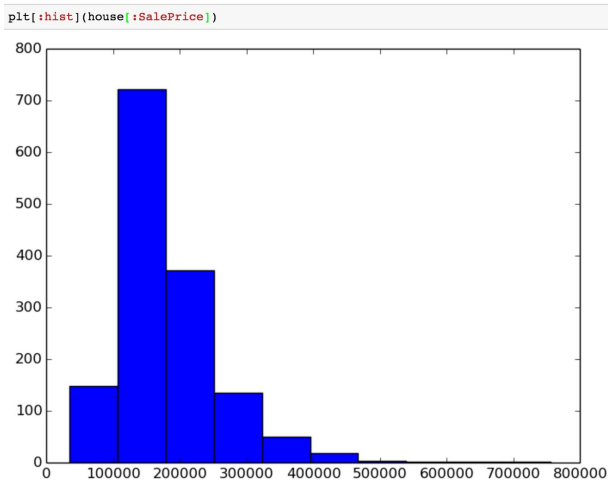

Figure 1. Histogram of Sale Prices. Most houses are sold at price from 100,000 to 300,000.

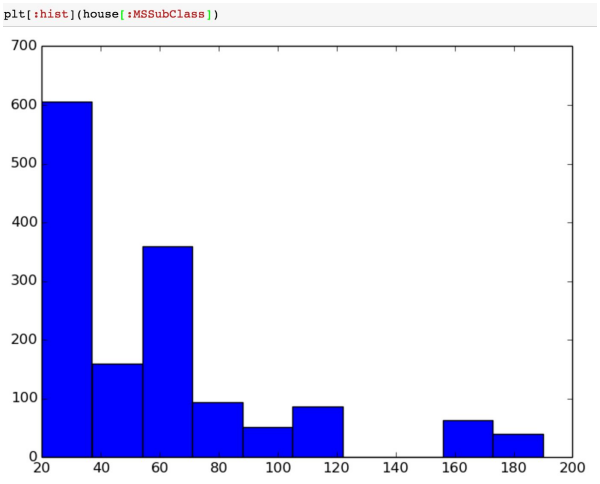

Figure 2. Histogram of Type of Dwelling. Sixteen types of dwellings are assigned numbers from the range 20-190. See details in data_description.txt.

Based on our own experience, we think some of the variables contribute to the housing price more than the others. For example, living area in square feet, neighborhood, size of garage in car capacity, size of the basement and whether it has central air system may be important indicators for the house price. Figure 3 and Figure 4 shows the car capacity relates to housing price.

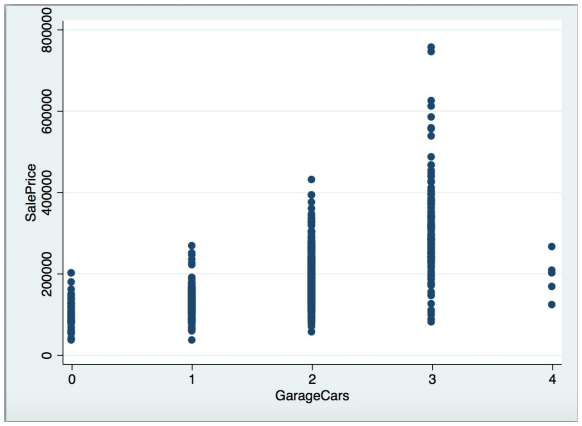

Figure 3. Scatter Plot of Housing Price vs. Garage Capacity (Measured n Number of Cars)

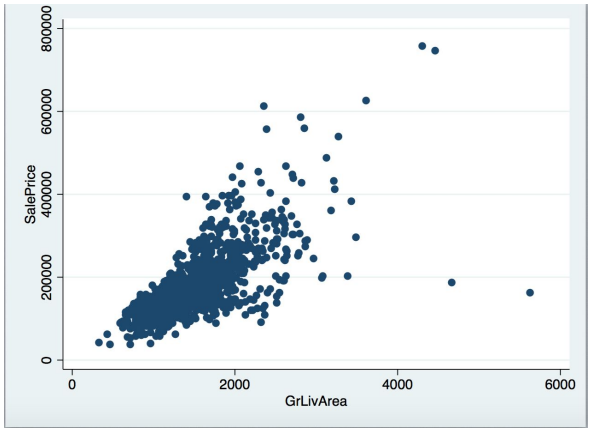

Figure 4. Scatter Plot Housing Price vs. General Living Area.

Variable Selection

There are 79 variables and many of them have nominal values, bringing our total number of factors to around 200. Given only 1460 data entries, some insignificant variables need to be eliminated, or we will easily run into the problem of overfitting and lengthy run times. As a result, all variables are screened and 20 variables deemed to be significant for assessing house values are selected to construct a model. The variables selected are:

LotArea: lot size in square feet
LotShape: general shape of property (regular, slightly irregular, moderately irregular, irregular)
Neighborhood: physical locations within Ames city limits
BldgType: type of dwelling (single-family detached, two-family conversion, duplex, townhouse…)
HouseStyle: style of dwelling (one story, one and one-half story, two story, two and one-half story…)
OverallQual: rates the overall material and finish of the house (1-10)
OverallCond: rates the overall condition of the house (1-10)
ExterQual: evaluates the quality of the material on the exterior (excellent, good, average, fair, poor)
ExterCond: evaluates the present condition of the material on the exterior (excellent, good, average...)
BsmtQual: evaluates the height of the basement (excellent, good, typical, fair, poor, no basement)
TotalBsmtSF: evaluates the general condition of the basement (excellent, good, typical, fair, poor, no...)
CentralAir: central air conditioning (no, yes)
GrLivArea: above grade living area square feet
BedroomAbvGr: bedrooms above grade
KitchenAbvGr: kitchens above grade
GarageCars: size of garage in car capacity
SaleType: type of sale (warranty deed - conventional/cash/VA loan, new, contract 15% down payment…)
SaleCondition: condition of sale (normal, abnormal, adjoining land purchase, allocation)
TotalBath: total number of bathrooms (BsmtFullBath + BsmtHalfBath + FullBath + HalfBath)
YearSold: year sold (yyyy)

Preliminary Regression Analysis

For preliminary analysis, we broke the original data set into a training set (80%) and a testing set (20%). We first ran a linear regression model without regularization on the data and then ran a second model with lasso regularizer. The loss function is squared error. The risk function chosen is the root mean squared errors (Lasso_RootMSE=43251.26, Linear_RootMSE=36255.47). Although our result shows that linear regression has lower root mean square error than lasso regression, we finally chose lasso regression since it has less variation and can possibly produce a better prediction on a different data set. Another reason is that the lasso regression model tends to produce sparse solutions, thus has coefficients that are easier to be interpreted.

The lasso regression model is chosen by running cross validation on different lambda values. We decided to use the lambda that that gives a more regularized model but is within 1 standard deviation of the minimum error. In this process, we did cross validation on R and picked Lamda=3837.

The w we got from our analysis which is showed below:

x	Variable Names	Type*	w	Weight Values	Interpretation	Comment
1	LotArea	U	1	0	Insignificant	Reasonable
2	LotShape	O	2	0	Insignificant	Reasonable
3	Neighborhood	O	3	NoRidge 7623.29835	Houses in NoRidge neighborhood are significantly more expensive. Strong Impact.	Reasonable
				NridgHt 15326.20023	Houses in NridgHt neightborhood are significantly more expensive. Strong Impact.	
4	BldgType	O	4	0	Insignificant	Reasonable
5	HouseStyle	O	5	0	Insignificant	May correlate with GrLivArea
6	OverallQual	U	6	15915.2391	Overall quality has a great impact on house price.	May correlate
7	OverallCond	U	7	0	Insignificant	
8	ExterQual	U	8	13017.20117	Exterior quality has a great impact on house price.	May correlate
9	ExterCond	U	9	0	Insignificant	
10	BsmtQual	U	10	9401.77511	Basement quality has a strong impact on house price.	Need further investigation
11	TotalBsmtSF	U	11	20.13025	Total basement square feet has a little impact on house price.	Reasonable
12	CentralAir	O	12	0	Insignificant	Reasonable
13	GrLivArea	U	13	37.97014	Above ground living area has a little impact on house price.	Make no sense
14	BedroomAbvGr	U	14	0	Insignificant	May correlate with GrLivArea
15	KitchenAbvGr	U	15	0	Insignificant	May correlate with GrLivArea
16	GarageCars	U	16	11048.20174	The higher the garage capacity, the higher the house price. Strong impact.	May correlate with GrLivArea
17	SaleType	O	17	New 532.94666	House just contructed and sold are more expensive than other types. Medium impact.	Reasonable
18	SaleCondition	O	18	0	Insignificant	Reasonable
19	TotalBath	U	19	0	Insignificant	May correlate with GrLivArea
20	YearSold	U	20	5721.8512	The later the house is sold, the higher the price is.	Reasonable
*Type: U stands for numerical value; O stands for nominal value.						

Table 1. Variables x and fitted w.

Plan For Rest of the Semester

1. We will refine variable selections by eliminating some insignificant variables indicated by preliminary analysis and looking into the correlations between variables. We will also revisit the variables we dropped earlier and make sure we did not drop anything important. A way to find out is to look at their correlations with the house sale prices.
2. We will try different cross validation methods to gain a more accurate picture of how well our model generalizes to an independent data set. Also we will bootstrap to refine our model since the data is insufficient given such a big set of variables.
3. We will try different loss functions (maybe huber, quantile regression) and compare which one works better.
4. With these plans, the goal is to find a model that reduces out of sample error to a satisfactory level.