Driving up House Sale Price in Iowa

Team 3 (Iowa Realtors)
(Manasi Kulkarni, Nagasoundari Balamurugan, Aditya Wakade, Mervin Sundaram)
MKTG 562

1. Overview:

When someone asks about an American dream, house ownership is probably one of the few things that come to mind. For some Americans, it is probably the largest purchase that they will ever make in their entire lifetime. Owning a real estate and raising a family is so interconnected and there is a lot of emotion involved in it. When we researched, we found that most conventional mortgages span for nearly 30 years. This accounts for nearly 40% of the entire lifetime. The decision to decide the dream house has to be sound and based on a lot of factors, not just for the buyers, but also for the builders; because they are the ones building that dream house for the customers.

From a Business and Big Data perspective, understanding the customer choice to drive up the sales price and utilizing these patterns to improve on the current and future decisions is extremely crucial and is the focus of our analysis.

2. Research Dataset:

Our dataset is titled 'House Prices: Advanced Regression Techniques' on Kaggle . It is compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset. This dataset has 79 explanatory variables and they span across every detail of the house.

3. Customer-Centric Research Question and Goal:

Based on our the dataset and the problem we were trying to address, we came up with a customer-centric research question: *'Which features of the Iowa Housing dataset influences the customer buying decision and how well can these features be used to predict future sale price of a house and attract potential house buyers'*

The goal of our research is to identify the factors that are most important to the buyers and which directly affects the sale price. Also, to create a regression model that is able to accurately estimate the price of the house given the features.

4. Data Preprocessing:

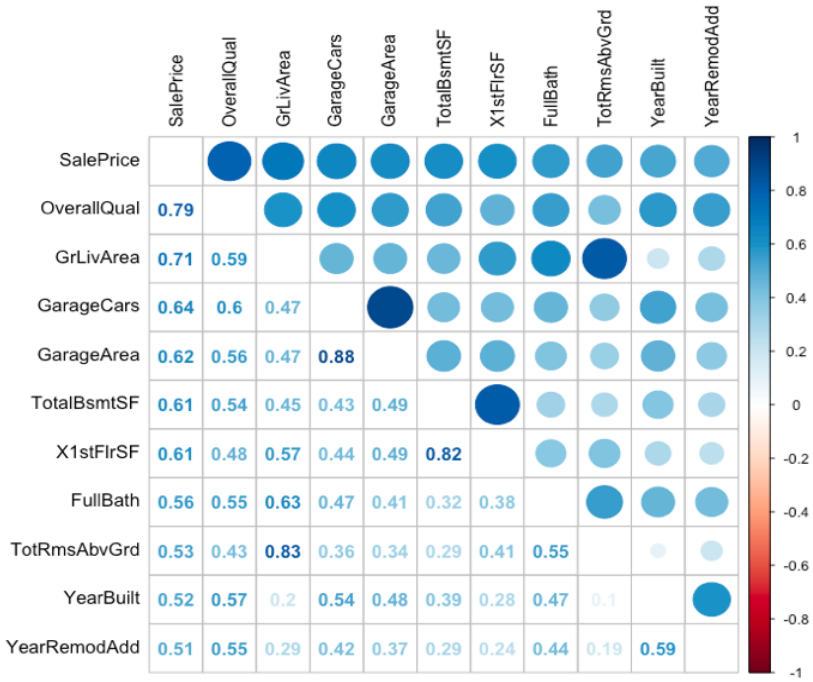
The dataset consists of features in various formats. It has numerical data such as prices and numbers of bathrooms/bedrooms/living rooms, as well as categorical features such as zone classifications for sale, which can be 'Agricultural', 'Residential High Density', 'Residential Low Density', 'Residential Low Density Park', etc. Considering 79 variables were too many for our analysis, we wanted to reduce them and only include those which had the maximum correlation with the dependent variable - Sale Price.

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	···	PoolArea	PoolQC
1	60	RL	65	8450	Pave	NA	Reg	Lvl	AllPub	···	0	NA
2	20	RL	80	9600	Pave	NA	Reg	Lvl	AllPub	···	0	NA
3	60	RL	68	11250	Pave	NA	IR1	Lvl	AllPub	···	0	NA
4	70	RL	60	9550	Pave	NA	IR1	Lvl	AllPub	···	0	NA
5	60	RL	84	14260	Pave	NA	IR1	Lvl	AllPub	···	0	NA
6	50	RL	85	14115	Pave	NA	IR1	Lvl	AllPub	···	0	NA

Figure 1: Snapshot of the training data (Train: 1460 rows, Test: 1459 rows)

As seen from Figure 1, there were columns like Alley and PoolQC that contained only NULL/NA values and they had to be removed as they were not adding any value to our analysis. We then derived the

correlation of all the variables with the Sale Price to address the problem of multicollinearity and to find out the variables that really has the highest impact on the Sale Price.



In regression, "multicollinearity" refers to features that are correlated with other features. Multicollinearity occurs when your model includes multiple factors that are correlated not just to your target variable, but also to each other.

Multicollinearity increases the standard errors of the coefficients.

Figure 2: Correlation matrix of the top variables with the Sale Price (R library corrplot)

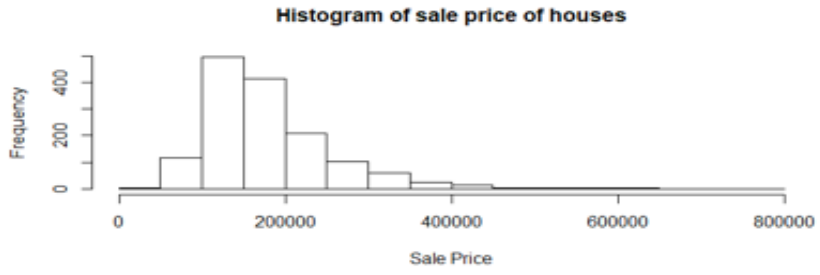
That means multicollinearity makes some variables statistically insignificant when they should be significant. So, we just considered a single feature of these multicollinear variables. Keeping the threshold of the correlation greater than 0.5, we finally zeroed down to the variables seen in the above marix. Of these, we will be exploring in depth - OverallQuality, GrLivArea, FullBath. There are some additional variables which we will be exploring as well- Kitchen, Zone, Building Type etc.

5. Exploratory Data Analysis:

We divided our variables into categories: External Factors, Aesthetics, Internal Factors.

5.1. Analysis of Dependent Variable- SalePrice

To perform an analysis of the dependent variable, it is important to know the spread and distribution of this variable. We made a histogram of the Sale Price. Most of the houses fell under the price range of 100k to 200k. This makes absolute sense as the price range is in accordance with the mid-east and mid-west regions of the United States. We found that very few houses were above 500k. These could be highly luxurious houses but are mostly outliers.



Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
$34900	$129975	$163000	$180921	$214000	$755000

Figure 3: Frequency Histogram and summary table of Sale Price

5.2. Overall Quality

The variable which had the highest correlation with the Sale Price was the Overall Quality. It had the correlation coefficient as 0.79. The trend we see here is that the houses with the highest OverallQual score has a high SalePrice. While other factors could definitely be detrimental in determining the SalePrice, this seemed to be the one that affects it the most. If this data was to be used by an Iowa based property agent who is working on a real estate website, this would be the feature he needs to highlight the most. The way this score is calculated is not known and it is a black box.

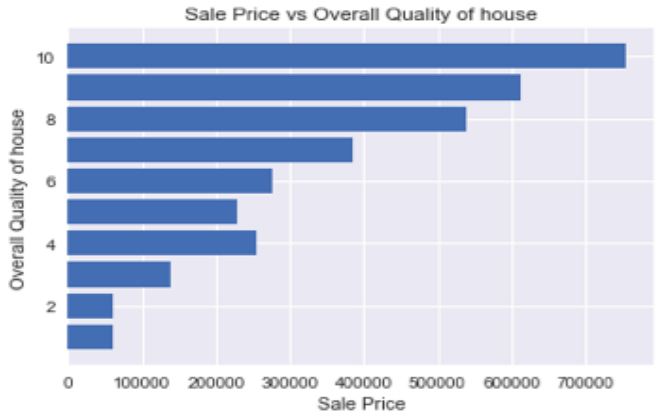


Figure 4: Bar chart of OverQual vs Sale Price

5.3. External Factor-Zoning Classification and Ground Living Area

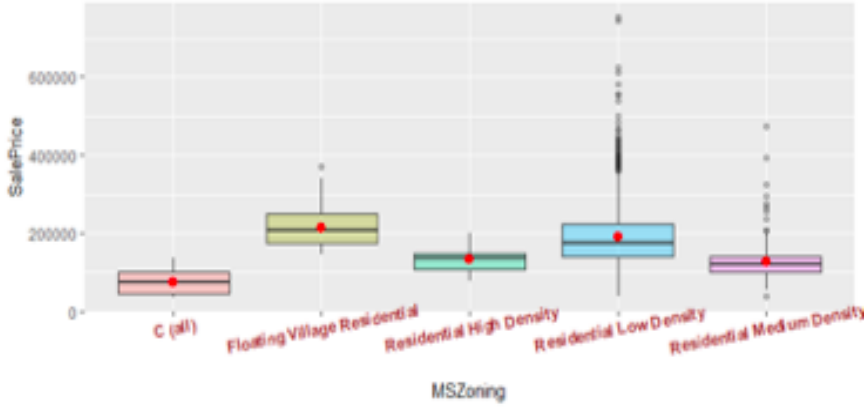


Figure 5: BoxPlot of different zones vs SalePrice

We plotted a box plot of the various available zonings vs the Sale price. As seen from the above figure, the Floating Point Village Residential account for the highest sale price followed by the Residential Low

Density and the Medium Density. Infact, the Commercial zone accounts for the lowest sale price. This was a bit absurd. Prices in commercial areas are always more considering they have the facility and convenience of a commercial neighborhood. This can happen only when people prefer house area over house neighborhood.

Zoning Type	Size(in Sq.ft)
Commercial	1191.4
Floating Village Residential	1574.5
Residential High Density	1510.1
Residential Low Density	1551.6
Residential Medium Density	1322

Figure 6: GrLivArea vs SalePrice (L), Mapping of GrLivArea to ZoneType(R)

The above figure confirms our assumption that customers tend to buy houses of a larger area and larger price than a commercial neighborhood. In other words, the price of the house does not depend on the surrounding and the external factors but the Ground Living Area of the house.

5.4 Internal Factor - Building Type

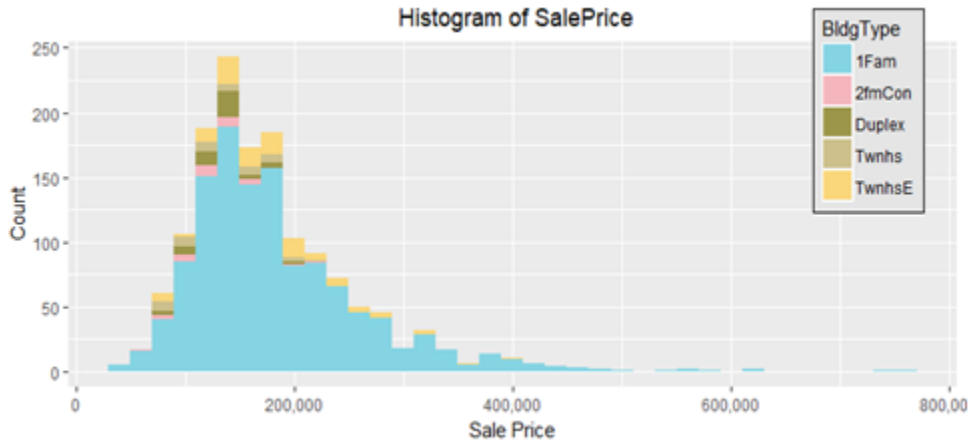


Figure 7: Histogram of different building types with its prices

We plotted a histogram of the various building types available with the Sale Price. As seen in the above figure, Single Family Homes(1Fam) were the most sold houses both in terms of frequency and in terms of the prices of the houses. The different available types of houses were categorized into four different categories, namely Single Family Homes(1Fam), Two Family Conversion homes(2FmCon) which were initially built to house a single family but then converted to fit in more than one family, Duplex, Townhouses(Twnhs) and TownHouse Edge(TwnhsE).

Looking at the histogram, it is also evident that customers don't tend to buy extremely expensive homes. But it's also clear that single family homes are the more expensive houses. This could be the case of

luxurious and rich families buying such houses. Additionally, one other insight was that single family homes could not necessarily be thought of as cheap or expensive as they ranged from 100,000 to 800,000 in price ranges and this is quite a wide range of prices and this is because of the combination of various factors which combine to overall Quality Score.

5.5 Aesthetic Factors - Bathroom

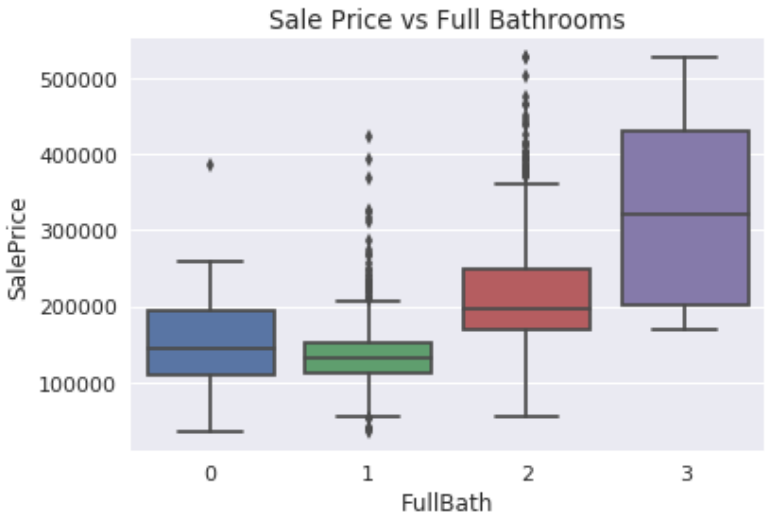


Figure 8: Boxplot of sale price vs no of full baths

We plotted a boxplot diagram to understand the relationship between the number of full bathrooms and the Sale Price. A Full Bathroom in the US is one which has a sink, a bathtub, a shower and a toilet. Any bathroom missing one of these is ¾, a bathroom missing 2 of these is ½ and so on. However, a bathtub with a shower built into it counts as both as well. As evident that the increase in the number of bathrooms implies that the house has more area in Sq Ft, the Sale price of the house was proportional to the number of bathrooms. Delving further into the data, every house which had a sale price over 200,000 and can borderline be thought of as an expensive house according to the Database had more than one bathroom. However, there could also be an inference that trying to fit in an extra bathroom within the available area without necessarily increasing the total area of the house could drive up the sale price. This was one of our final recommendations as well by suggesting that this could be done by adding an extra/guest bathroom in an existing basement if the house had any.

5.6 Aesthetic Factors – Kitchen

We then used a factor plot to understand the relationship between the presence of extra kitchens in a house and the Sale Price. Each of the data points on the plot were also of different colors based on their quality. The Qualities were Ex(Excellent), Gd(Good), Typical Average(TA) and Fa(Fair). Like in the previous relationship of bathrooms and Sale Prices, the price of the house should have gone up if an extra kitchen was added. However, that was not the case.

Figure 9: Factor plot of sale price vs no of kitchens

From the chart, we could see that one house with an excellent kitchen was worth more than a house with more number of kitchens with any quality. Although, that could be evident given that Excellent was a grade higher than Good Kitchens, our findings were more cemented when we discovered that a house with 1 good kitchen was more expensive than a house with 2 good kitchens. However, this could also be debated on the fact that how they rate kitchens on the scale is behind a black box, and for us to ascertain what range of houses have a good kitchen and if both good kitchens in the second house were borderline good, while the other kitchen in the first house could have been almost excellent but still falling into the good category. Our recommendations on the basis of this were to add more amenities and items which could improve the quality of a kitchen, rather than to have an extra kitchen.

6. Methods

To begin with we started analyzing the dependent variable sale price to build an appropriate model based on the distribution of the sale price. The distribution of the sale price looked like the below graph.

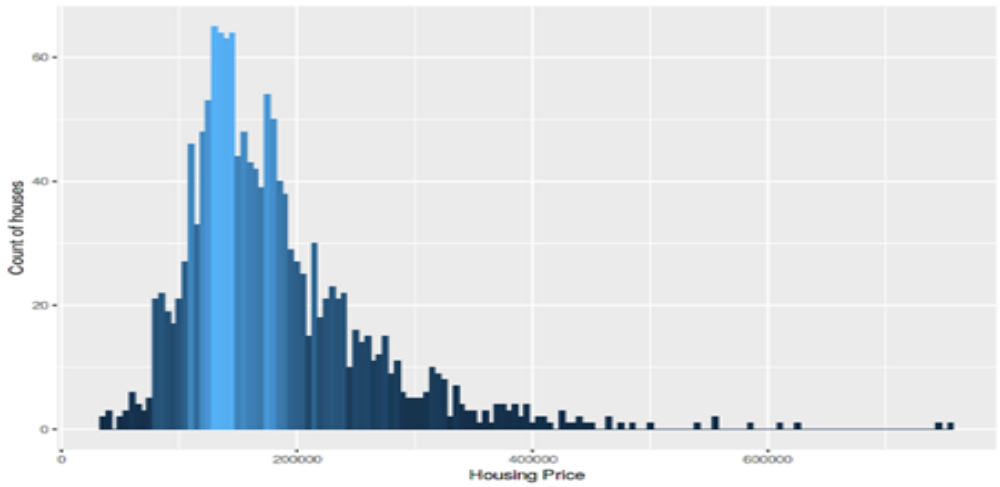


Figure 10: Histogram of sale price of houses

Though the plot look like a normal distribution, it is right skewed, Thus regression model could not be used to predict the price. Hence we applied logarithm on the sales prices and plotted it again. This changed the distribution of the plot to a normal distribution which could be seen below.

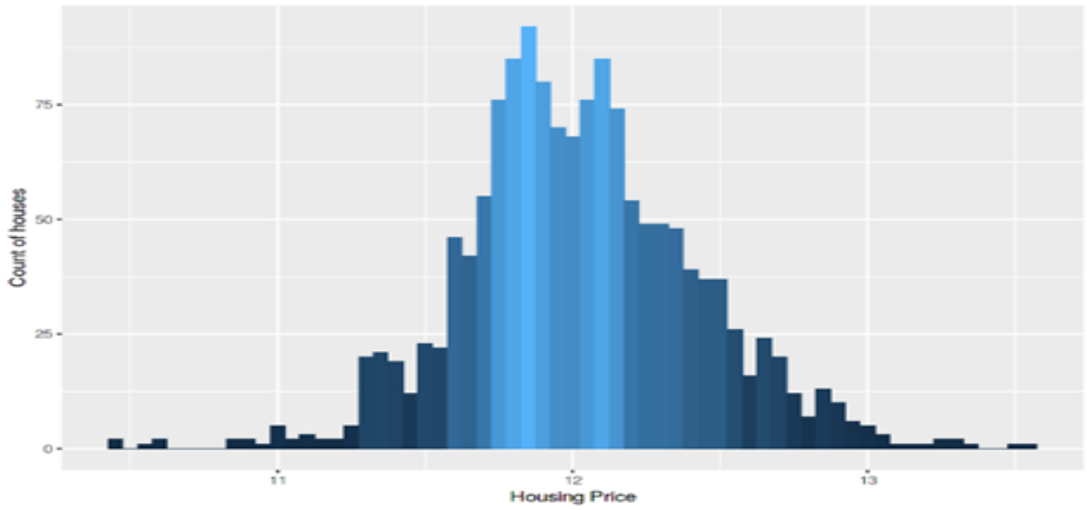


Figure 11: Histogram of logarithm of sale price of houses

Now we have a normal distribution and hence regression models could be built on it. At first, we have built a model that took all the columns as predictors and sale price as dependent factor. The model was statistically significant and had a high R-squared value of 0.8371909 and an Adjusted R-squared value of 0.8303297. But the complexity of the model is high as it has lot of variables included in it. Thus, We decided to build a regression tree to find the high weighted variables that affects the sale price of a house and use it along with the variables that are highly correlated with the sale price. The tree function was applied with Sale price as dependent variable and all other columns as independent variable. The plot of the tree model is shown below.

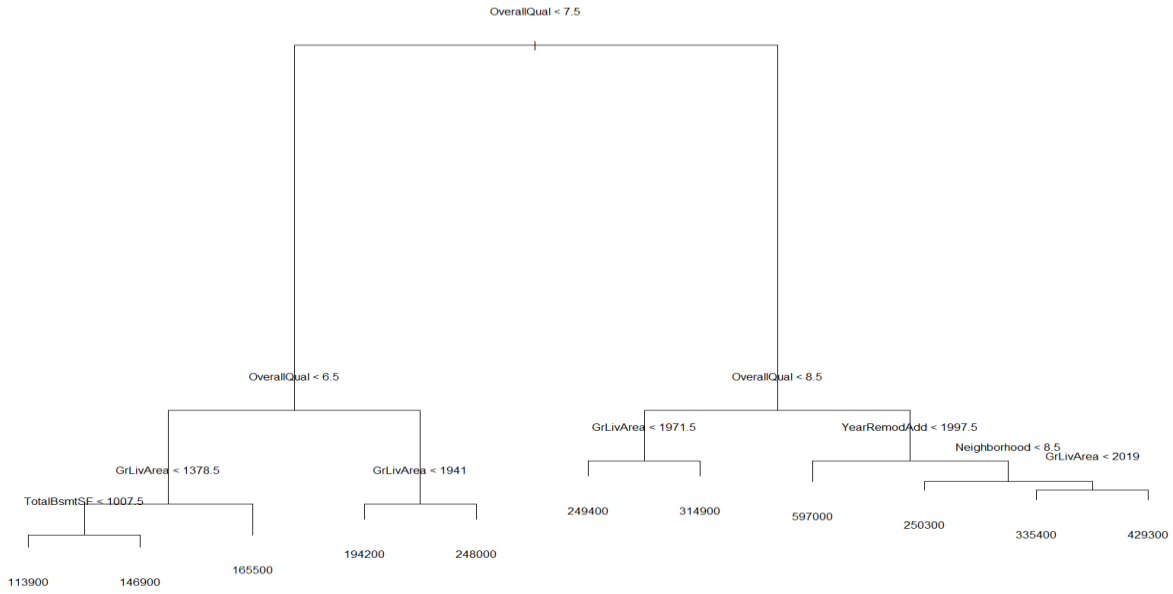


Figure 12: Plot of tree model with all variables as predictors

The variables actually used in the construction of tree were "OverallQual", "GrLivArea", "TotalBsmtSF", "YearRemodAdd" and "Neighborhood". As said earlier, in addition to these variables, we have also taken the variables that are highly correlated with the sale price to build a new regression model. The predictors included in this model are OverallQual, GarageCars, YearBuilt, Fireplaces, WoodDeckSF, X2ndFlrSF, LotArea, BsmtFullBath, RoofStyle, SaleCondition, Neighborhood, BedroomAbvGr, RoofMatl, Functional, ScreenPorch, Exterior1st, LandSlope, Street, LandContour, Condition2, YrSold, OverallCond, MSSubClass, KitchenAbvGr, KitchenQual, ExterQual. This model was also statistically signficant and had a Multiple R-squared value of 0.8322 and Adjusted R-squared value of 0.8289. Though the R-squared value is bit less(0.14) than the all-factor model, this one is quite simpler. Also the residual plot had a low variance with all data points evenly spread across 0 axis. The residual plot of this model is shown below.

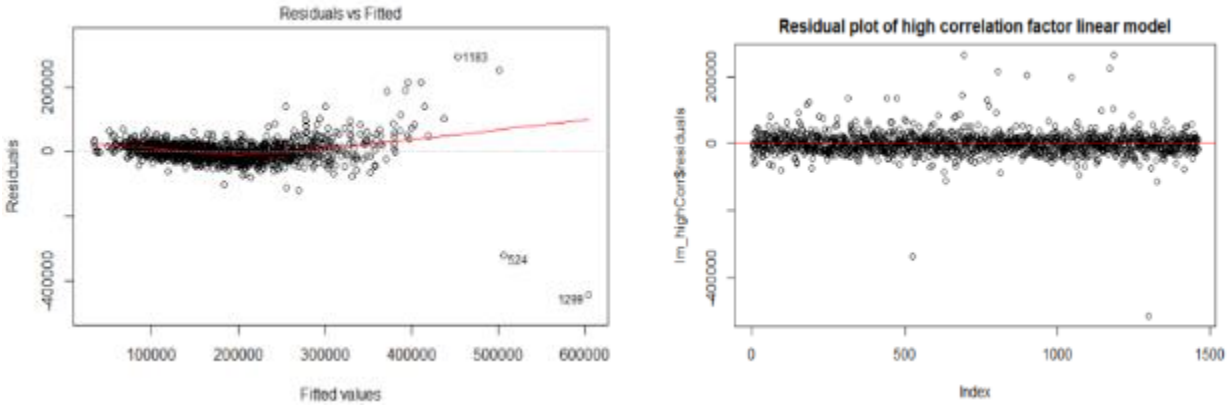


Figure 13: Residual plots of best fitted regression model

In order to check the accuracy of the model, the predicted house sale price in the test set was compared with the actual sale price and the graph is shown below.

Figure 14: Plot of predicted vs log of actual sale price in test set

From the above graph, we can see that model is pretty good in predicting the price and the accuracy was found to be 86.8%. Thus we decided to go with this model and analyzed on the values of coefficients to find various ways to increase the sale price of a house in Iowa. The top 10 predictors with high coefficient values are listed in the table below.

Predictors	Coefficient
Overall quality	12953.67
Type of road access to street	37973.33
Full bath in basement	13120.28
Garage size	10683.02
Land slope	9561.84
Home functionality	4703.95
Overall Condition	4399.53
No. of fireplaces	3980.096
No. of bedrooms	3685.47
Roof Material	3360.49

Figure 15: List of top 10 predictors with its coefficients

The numbers in the above table signifies the importance of each variable. For example, an increase in quality rating from 5 to 4, 1 being the top quality, the sale price of the house increases by $12,953. Thus concentrating on these features of houses could increase the sale price of the houses. Builders could construct houses with these features that are popular to easily sell them. Eventually, amount of investment to earnings ratio would be significantly high.

7. Recommendations:

Based on the insights from the above table, we have few recommendations for the house builders and real estate owners in Iowa region to sell a house easily and for a higher price. We have split the recommendations into two categories: New construction and Existing houses.

7.1 New Construction:

The above mentioned features are easy to implement in new construction as it is an empty canvas compared to the existing ones. Few of the recommendations to increase the sale price are mentioned below.

1. Lesser the slope of the land, higher the price. Thus, a flat land could be chosen to build a house or a sloped land could be flattened before building.
2. Larger the living area, higher the price. We have also seen that single family houses are the most popular building type in Iowa. Hence while building a community, bigger single family houses could be constructed.
3. Increasing the number of bedrooms and bathrooms also increases the house price. Having a full bath in basement/first floor increases the price of the house.
4. Higher the number of garages, higher the price. The house could be designed with more than 2 garages as most of the household would have at least two cars.
5. Adding extra fireplaces also increases the price of the house.
6. An interesting factor found was having a paved path access to street. There were two type of access to street: gravel and paved, among which a paved path access increases the price of the house by 37973 dollars approximately.
7. Using a better roofing material increases the sale price as the lifetime of the roof would be high.
8. Overall, the quality of the house is the most important factor which the buyer looks for. Hence the construction of house should use high quality materials and have a strong structural model.

7.2 Existing Construction:

There could not be much changes made to an existing construction. Also the cost to be invested would be high for the existing ones compared to the new ones. Thus features that are easy to implement could be done which are listed below.

1. The quality of the house could be improved by fixing any repairs, upgrading the kitchen equipment, counter tops of bathrooms, upgrading pipelines etc.,
2. Extra fireplaces could be added to the house as they are available readymade.
3. Roofing material could be changes as the lifetime of the roof would be extended.
4. The road access could be changed from gravel to paved paths to increase the price.
5. All the other factors that are easy to add from the coefficient table could be done.

8. Conclusion:

Working with the housing data set in this project was an eye opening experience. This work played both the parts of reaffirming some of our initial perceptions/beliefs and completely contradicting some of our initial thoughts. This was a great learning experience that proved perception are not always right and it is important to look for statistical evidence. This work also helped us identify suitable houses to buy in the future and our first step in entering a real estate business.