Housing Features In The Real Estate Market

Sean McLean

ALY 6015

Module 1 Assignment

Introduction

The dataset studied this week looks at the features of housing that were sold over an unknown time frame. With 82 variables and over 1000 examples, it is a very detailed look at many aspects of homes and what goes into how a house is sold. An analysis of this kind of dataset can be useful for looking at certain trends in the real estate market. A thorough regression analysis with modeling and data visualizations was the focus on the dataset to provide the best opportunities for identifying patterns in the data.

Data Analysis

After loading the dataset, at first glance the variables have to deal with housing features as well the size of the yard, some qualities of the neighborhood, and some historical context like when the house was built. From the summary statistics there was evidence of missing data from some of the observations. There is a considerable range of the age of the houses as well as the sale price of the homes and potentially a correlation between the two variables. The age of the house could possibly dictate how much you would spend on a home. Several variables have almost no information in the observations, but they don't seem relevant when looking the value of a home. Some examples of this whether a home has an alley, the quality of the pool on the property, and the quality of the fireplace if the home has one.

Looking the correlation matrix between the variables, many of the strong positive linear relationships were common among housing features that would be popular to people looking to purchase a home. Some examples of this is the quality of the living room area, lot area, and the number of full bathrooms that correlate with the lot frontage. When just focusing on two variables, the price of the house and the year the house was sold, there are very strong positive correlation and the year the house sold shows more of a negative correlation among the other variables. The matrix plotting shows more positive relationships overall with the shades of blue in the plot, indicating that despite the high number of variables that go with home buying, many of these are important to the buyer. There is a small of variables that show no color or no correlation, and perhaps could be an area where they can removed from the dataset.

A regression model was created the analysis of several explantory variables and the response variable being the sale price of the house. The four independent variables being used in the analysis are lot frontage, overall quality, the year the house was built, and ms subclass. The summary of the regression model shows that the coefficient value of the overall quality of the house would vary with high standard error and t-value numbers (ChatGPT, 2024). The p-value's of each independent variable is quite low, indicating that the coefficients are not close to zero overall.

```
Call:

lm(formula = SalePrice ~ Lot.Frontage + Overall.Qual + Year.Built +
    MS.SubClass, data = ames1)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-339635  -26798   -4276   19821  392004

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -762849.54   64760.89 -11.779  < 2e-16 ***
Lot.Frontage     631.12      43.41  14.539  < 2e-16 ***
Overall.Qual   39138.79     744.82  52.548  < 2e-16 ***w
Year.Built       338.29      34.13   9.912  < 2e-16 ***
MS.SubClass      -95.54      21.29  -4.487  7.5e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44810 on 2925 degrees of freedom
Multiple R-squared:  0.6858,   Adjusted R-squared:  0.6854
F-statistic:  1596 on 4 and 2925 DF,  p-value: < 2.2e-16
```

The modeling of the regression analsyis shows a negative linear relationship between the residuals verus fitted plot with some outliers. In the Q-Q Residuals plot shows more of a normal distribution but some deviations on both ends of the line, especially as the values get bigger. The scale location plot shows a very similar negative linear relationship like the residuals versus fitted plot but with more of a spread, showing some signs of homoscedasticity. In the residuals versus leverage plot there are some outliers that are far from the other points, possibly having an impact on the regression coefficients (Chat GPT, 2024).

The model was looked at for multicollinearity by using the variance inflation factor which showed that all four independent variables were all well below five. This was revealing in that it demonstrates the strong correlation the independent variables have on the dependent variable. Because of this there was no need to make any necessary changes to strengthen the relationships. Outliers were also examined and there were only a few from the large dataset that were far from most of the other points. I would leave them in the dataset because of the small number and I do not think it would overall affect the dataset. The preferred data model overall looks pretty similar to the previous regression summary. From the analysis it looks like there doesn’t need to be too many revisions because the independent variables have fairly strong correlations to the response variable, meaning that these housing features are important to the price of the house.

```
Call:
lm(formula = SalePrice ~ Lot.Frontage + Overall.Qual + Year.Built,
    data = ames1)

Residuals:
    Min      1Q  Median      3Q     Max
-357928  -26790   -3728   19286  390300

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -767023.38   64965.59 -11.807   <2e-16 ***
Lot.Frontage     710.65      39.75  17.876   <2e-16 ***
Overall.Qual   38817.47     743.79  52.189   <2e-16 ***
Year.Built       335.83      34.24   9.809   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44960 on 2926 degrees of freedom
Multiple R-squared:  0.6836,   Adjusted R-squared:  0.6833
```

```
F-statistic:  2108 on 3 and 2926 DF,  p-value: < 2.2e-16
```

Conclusions/Intrepretations

While there are some minor corrections that could be implemented, the dataset overall shows what is valuable data to a home buyer and what isn't. Because the cost associated with the house is one of the most important factors, it serves as a good response variable when researching the correlations between the other variables. I would recommend using the sale price variable for future analysis in understanding trends in home buying and how other variables could play a role in shaping those values.

References

Bluman, A. G. (2017). Elementary Statistics: A Step-by-Step Approach. 10th edition. McGraw-Hill Education.

Chat GPT. (2023, December 10th). Default (GPT 3.5)

Appendix

```
library(dplyr)
library(ggplot2)
library(corrplot)
library(RColorBrewer)
library(car)

#Question 1
ames <- read.csv("AmesHousing.csv", header=TRUE)
ames

#Question 2
summary(ames)
View(ames)
dim(ames)
head(ames)
tail(ames)
str(ames)
table(ames)
class(ames)
range(ames$Lot.Config)

#Question 3
```

```
missing_values <- colSums(is.na(ames))
print(missing_values[missing_values > 0])
ames1 <- ames
for (col in names(ames)[missing_values > 0]) {
mean_value <- mean(ames[[col]], na.rm = TRUE)
 ames1[[col]][is.na(ames[[col]])] <- mean_value
}
head(ames1)
#Question 4
numeric_cols <- sapply(ames1, is.numeric)
cor_matrix <- cor(ames1[, numeric_cols])
print(cor_matrix)

#Question 5
corrplot(cor_matrix, method = "color", type = "upper", order = "original", tl.col = "black", tl.srt = 90)
title("Correlation Matrix")

#Question 6
set.seed(123)
correlations <- cor(cor_matrix)
highest_corr_variable <- names(which.max(correlations["SalePrice", ]))
lowest_corr_variable <- names(which.min(correlations["SalePrice", ]))
closest_corr_variable <- names(which.min(abs(correlations["SalePrice", ] - 0.5)))
par(mfrow = c(1, 3), mar = c(4, 4, 2, 1))
plot(cor_matrix ~ SalePrice, cor_matrix[[highest_corr_variable]], xlim = c(-1, 1),
    main = paste("SalePrice vs.", highest_corr_variable, "Correlation =", round(correlations["SalePrice", highest_corr_variable], 2)),
    xlab = "SalePrice", ylab = highest_corr_variable, col = "blue")
```

```
plot(cor_matrix ~ SalePrice, cor_matrix[[lowest_corr_variable]],
    main = paste("SalePrice vs.", lowest_corr_variable, "Correlation =",
round(correlations["SalePrice", lowest_corr_variable], 2)),
    xlab = "SalePrice", ylab = lowest_corr_variable, col = "red")
plot(cor_matrix ~ SalePrice, cor_matrix[[closest_corr_variable]],
    main = paste("SalePrice vs.", closest_corr_variable, "Correlation =",
round(correlations["SalePrice", closest_corr_variable], 2)),
    xlab = "SalePrice", ylab = closest_corr_variable, col = "green")


#Question 7
model <- lm(SalePrice ~ Lot.Frontage + Overall.Qual + Year.Built + MS.SubClass, data =
ames1)
summary(model)


#Question 9
plot(model)


#Question 10
library(car)
vif(model)


#Question 11
outlierTest(model = model)
hat_values <- hatvalues(model)
plot(hat_values, pch = 19, main = "Hat Plot", xlab = "Observation", ylab = "Leverage")
abline(h = 2 * nobs(model)/length(hat_values), col = "red", lty = 2)


#Question 13
install.packages("leaps")
```

```
library(leaps)

all_subsets <- regsubsets(SalePrice ~ Lot.Frontage + Overall.Qual + Year.Built, data = ames1, method = "exhaustive")

best_model <- which.min(summary(all_subsets)$adjr2)

summary(all_subsets)$which[best_model, ]

preferred_model <- lm(SalePrice ~ Lot.Frontage + Overall.Qual + Year.Built, data = ames1)

summary(preferred_model)

plot(preferred_model)
```