**Ames Housing Data set**

# Introduction

The aim of this assignment is to analyze the Ames Housing dataset to understand correlations, patterns, and regression models better. The dataset contains information from the Ames Assessor's Office, which was used to calculate assessed property values for individual residential properties sold in Ames, Iowa, between 2006 and 2010.

The dataset comprises 2930 observations, which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables. These variables are essential in assessing a property's value and determining its sale price.

We focused on a subset of properties within the dataset to create a predictive model for the sale price of a typical single-family residence in Ames, Iowa. We used simple and multiple regression models to identify correlations between variables and the sale price of a property. We also performed a log transformation of the response variable to enhance the accuracy of our predictions.

Our analysis aimed to identify patterns and trends in the data to gain insight into the factors that influence property values in Ames, Iowa. We considered different components of the regression model, including the coefficients, standard errors, and p-values, to ensure the reliability and robustness of our model.

Our analysis of the Ames Housing dataset provided valuable insights into the factors that influence property values in the area. Our predictive model can be useful to prospective buyers and sellers, as well as real estate professionals and analysts.

# Exploratory Data Analysis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Order | PID | MS.SubClass | MS.Zoning | Lot.Frontage | Lot.Area | Street | Lot.Shape |
| 1 | 526301100 | 20 | RL | 141 | 31770 | Pave | IR1 |
| 2 | 526350040 | 20 | RH | 80 | 11622 | Pave | Reg |
| 3 | 526351010 | 20 | RL | 81 | 14267 | Pave | IR1 |
| 4 | 526353030 | 20 | RL | 93 | 11160 | Pave | Reg |
| 5 | 527105010 | 60 | RL | 74 | 13830 | Pave | IR1 |
| 6 | 527105030 | 60 | RL | 78 | 9978 | Pave | IR1 |
| 7 | 527127150 | 120 | RL | 41 | 4920 | Pave | Reg |
| 8 | 527145080 | 120 | RL | 43 | 5005 | Pave | IR1 |
| 9 | 527146030 | 120 | RL | 39 | 5389 | Pave | IR1 |

****

The given dataset consists of 82 variables, which can be categorized into different types based on their nature:

Nominal variables: There are 23 nominal variables in the dataset. Nominal variables are categorical variables that do not have any inherent order or ranking among them.

Ordinal variables: There are 23 ordinal variables in the dataset. Ordinal variables are categorical variables that have a natural ordering or ranking among them.

Discrete variables: There are 14 discrete variables in the dataset. Discrete variables are variables that can only take integer values and are usually counts of some kind.

Continuous variables: There are 20 continuous variables in the dataset. Continuous variables are variables that can take on any numerical value within a certain range and can be measured on a scale.

Therefore, the dataset can be summarized as containing 23 nominal variables, 23 ordinal variables, 14 discrete variables, and 20 continuous variables.

The code utilizes the str() function and summary() function to obtain descriptive statistics for the provided dataset. str() function provides a summary of the structure of the dataset, including the number of observations, variables, and variable types. summary() function, on the other hand, provides descriptive statistics such as mean, median, quartiles, and minimum and maximum values for each variable. These functions are commonly used for exploratory data analysis to gain an understanding of the distribution and characteristics of the data. We can see how it is shown in R after running the code.

## summary(ameshousing)

|  |
| --- |
| 1 Order PID MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street |
| 2 Min. : 1.0 Min. :5.263e+08 Min. : 20.00 Length:2930 Min. : 21.00 Min. : 1300 Length:2930 |
| 3 1st Qu.: 733.2 1st Qu.:5.285e+08 1st Qu.: 20.00 Class :character 1st Qu.: 58.00 1st Qu.: 7440 Class :character |
| 4 Median :1465.5 Median :5.355e+08 Median : 50.00 Mode :character Median : 68.00 Median : 9436 Mode :character |
| 5 Mean :1465.5 Mean :7.145e+08 Mean : 57.39 Mean : 69.22 Mean : 10148 |

## Str(ameshousing)

|  |
| --- |
| data.frame': |
| $ Order : int 2351 |
| $ PID : int 527356020 |
| $ MS.SubClass : int 60 |
| $ MS.Zoning : chr "RL" |
| $ Lot.Frontage : int 80 |

## Input missing values

|  |
| --- |
| Exterior.2nd Mas.Vnr.Type Mas.Vnr.Area Exter.Qual Exter.Cond Foundation Bsmt.Qual Bsmt.Cond |
| 0 0 23 0 0 0 79 79 |
| Bsmt.Exposure BsmtFin.Type.1 BsmtFin.SF.1 BsmtFin.Type.2 BsmtFin.SF.2 Bsmt.Unf.SF Total.Bsmt.SF Heating |
| 79 79 1 79 1 1 1 0 |
| Heating.QC Central.Air Electrical X1st.Flr.SF X2nd.Flr.SF Gr.Liv.Area Bsmt.Full.Bath Bsmt.Half.Bath |
| 0 0 0 0 0 0 2 2 |
| Full.Bath Half.Bath Bedroom.AbvGr Kitchen.AbvGr Kitchen.Qual TotRms.AbvGrd Functional Fireplaces |
| 0 0 0 0 0 0 0 0 |
| Fireplace.Qu Garage.Type Garage.Yr.Blt Garage.Finish Garage.Cars Garage.Area Garage.Qual Garage.Cond |
| 1422 157 159 157 1 1 158 158 |

To handle the missing values in the dataset, the 'mice' package was utilized. This package offers imputation methods to replace the missing data with estimated values. In this case, the missing values were replaced with the mean value of the corresponding variable using the mice package.

# Histogram

After analyzing the data, it can be inferred that the mean selling price of the available houses is approximately $180,796. The histogram of the Sale Price distribution reveals a right-skewed pattern, implying that most of the houses are priced below the average. The histogram also indicates that the majority of the houses fall within the price range of $100,000 to $200,000, as it exhibits a peak in this range.

Chart, histogram

Description automatically generated

# Boxplot

The plot displays the pricing information for different neighborhoods. By looking at the plot, we can conclude that the top three most expensive neighborhoods are Stone Broke, Northridge Heights, and Northridge. The boxplot also gives us an idea about the range of prices for these neighborhoods, and we can see that both Northridge Heights and Stone Broke have a wide range of prices. Overall, this plot provides useful insights into the pricing trends for different neighborhoods.

Chart, waterfall chart, box and whisker chart

Description automatically generated

# Correlation Plotting

Selected numeric values using 'numeric.var', converted them to matrix using 'corr.matrix', and plotted correlation matrix with 'corrplot' and 'circle method' to estimate correlations accurately. According to the plot, the highest correlation of 0.8 is observed between Sales Price and Gr.Liv.Area (above ground living area).

Chart, bubble chart

Description automatically generated

# Scatter plots for different correlations

*We computed the correlations of all the numeric columns in the 'ameshousing' dataset with the 'SalePrice' variable using the 'cor' function. Then, we sorted the correlations in descending order using the 'sort' function.*

From the scatter plot, it is evident that there is a positive correlation between the Sale Price and Above Ground Living Area (Gr.Liv.Area) as the group has a steeper slope. This means that an increase in living space above ground leads to an increase in the Sale Price. The close fit of data points to the regression line indicates a strong association between Sale Price and Gr.Liv.Area. This is in contrast to the Scatter Plot for Enclosed.Porch with Sale Price, which has the lowest correlation.

## Sales Price vs Above Ground Area (Highest correlation)

Based on the scatter plot, it seems that there is a positive connection between the Sale Price and the Above Ground Living Area (Gr.Liv.Area). This implies that as the living area above ground increases, the Sale Price also tends to rise. The slope of the line of best fit for this group is quite steep, which strengthens this relationship. Furthermore, the data points on the scatter plot are closely clustered around the regression line, which indicates a strong correlation between the Sale Price and Gr.Liv.Area.

Chart, scatter chart

Description automatically generated

## Sales Price vs Enclosed Porch (Lowest correlation)

According to the scatterplot, there seems to be a positive connection between Sale Price and the size of Enclosed Porch area, with the data points rising from left to right. This implies that as the Enclosed Porch area increases, the Sale Price tends to increase as well. The data points are relatively closely spaced around the regression line, which suggests that there is a significant correlation between Sale Price and Enclosed Porch area. However, the slope of the line of best fit is less steep than in the previous scatterplot, which means that the positive correlation between Sale Price and Enclosed Porch area is not as strong as the relationship between Sale Price and Above Ground Living Area.

Chart, scatter chart

Description automatically generated

## Sale Price vs First floor in sq feet (correlation closest to 0.5)

The scatter plot suggests a positive correlation between the size of the First Floor area and Sale Price, with the data points gradually increasing from left to right. This indicates that an increase in the First Floor area tends to be associated with higher Sale Prices. The slope of the line of best fit in this scatter plot is moderate compared to the previous two plots, indicating a correlation of around 0.5. Therefore, the positive correlation between First Floor area and Sale Price implies that larger First Floor areas are likely to be associated with higher Sale Prices.

Chart, scatter chart

Description automatically generated

# 8. Linear Regression Model Equation

The multiple linear regression model can be expressed as SalePrice = 11323.44 + 103.03 \* Gr.Liv.Area + 98.15 \* Wood.Deck.SF + 0.57 \* Lot.Area.

The coefficient for Gr.Liv.Area is 103.03, implying that, holding all other predictors constant, a one-unit increase in Gr.Liv.Area (in square feet) leads to a $103.03 increase in the SalePrice. Similarly, a one-unit increase in Wood.Deck.SF (in square feet) corresponds to a $98.15 increase in SalePrice, and a one-unit increase in Lot.Area (in square feet) results in a $0.57 increase in SalePrice, while other predictors remain constant.

All three predictors are positively related to the SalePrice, meaning that an increase in these predictors' values leads to a corresponding increase in SalePrice. The model's coefficient of determination (R-squared) is 0.5265, indicating that roughly 53% of the variability in SalePrice can be explained by the linear relationship between SalePrice and the three predictors. The model is significant, with a very low p-value of the F-statistic, indicating that the predictors are jointly significant in explaining the variation in the SalePrice.

# Regression Model

|  |  |
| --- | --- |
| Coefficients: | |
| Estimate Std. Error t value Pr(>|t|) | |
| (Intercept) 11323.4364 3215.9057 3.521 0.000436 \*\*\* | |
| Gr.Liv.Area 103.0315 2.1489 47.947 < 0.0000000000000002 \*\*\* | |
| Wood.Deck.SF 98.1548 8.3422 11.766 < 0.0000000000000002 \*\*\* | |
| Lot.Area 0.5671 0.1351 4.196 0.0000279 \*\*\* | |
| Residual standard error: 55000 on 2926 degrees of freedom |
| Multiple R-squared: 0.5265, |
| F-statistic: 1084 on 3 and 2926 DF, p-value: < 0.00000000000000022 |

The given equation is a linear regression model that uses Gr.Liv.Area, Wood.Deck.SF, and Lot Area to estimate the Sale Price of a property. The intercept value of 11323.4364 implies that if all predictor variables are zero, the expected Sale Price would be 11323.4364 units.

The adjusted R-squared value of 0.526 indicates that the predictor variables explain 52.6% of the variability in the Sale Price. The significance of each predictor can be assessed by their respective t-values or p-values. In this case, both the intercept and predictor variables have significantly low p-values, suggesting that there is likely a relationship between the predictor variables and the outcome variable.

Therefore, we can reject the Null Hypothesis, which states that the coefficients are zero, and proceed with the alternative hypothesis, indicating that there is a relationship between the predictor variables and the outcome variable.

## Regression Model Plots

Chart, scatter chart

Description automatically generated

1. **Residual vs Fit plot**

A linear model is deemed appropriate for a data set if the residual plot against the fitted values displays a uniform level of dispersion across the plot and the plot of fitted values against residuals exhibits no visible structure or correlation. On the other hand, if the plot displays any discernible pattern or correlation, it is a strong indication that there is nonlinearity in the data.

Chart, line chart

Description automatically generated

**2. Normal QQ plot**

The presence of heavy tails in the residual plot's lower and upper regions suggests that the residuals deviate significantly from normality, violating the expected normal distribution pattern. This deviation from normality may be caused by the non-uniform variance observed in the Residuals vs. Fitted plot.

Chart, scatter chart

Description automatically generated

**3. Scale location plot**

The observation of an upward-sloping red line in the plot indicates the presence of heteroscedasticity, where the residual variance exhibits a non-constant pattern throughout the range of the fitted values. This heteroscedasticity is evident in the standard deviation of the residuals, which increases as the fitted value increases, violating the assumption of constant variance in a linear regression model.

Chart, line chart

Description automatically generated

**4. Residuals-Leverage plot**

A well-fitted linear regression model should have random and evenly distributed residuals around zero and leverage values across the predictor variables, which is seen from this graph as well, so the regression model is appropriate.

# Multi – collinearity

The variance inflation factor (VIF) is a metric used to assess the extent of multicollinearity in a multiple regression model. A VIF value exceeding 5 or 10 for any variable usually indicates the presence of multicollinearity.

However, in this scenario, the VIF values for all predictor variables are considerably lower than 5, which suggests that multicollinearity is not a major concern. Consequently, there is no requirement to eliminate any of the correlated variables or adopt dimensionality reduction techniques like principal component analysis. In summary, since the VIF values for all predictor variables are relatively low, it can be inferred that multicollinearity does not pose a significant issue in this multiple regression model.

|  |
| --- |
| Variables Tolerance VIF |
| 1 Gr.Liv.Area 0.8752363 1.142549 |
| 2 Wood.Deck.SF 0.9294137 1.075947 |
| 3 Lot.Area 0.9105858 1.098194 |

# Checking for Outliers

Chart

Description automatically generated

From the above plot, we can see there is a presence of outliers, so further functions were done to remove any outliers, and from the below box plot we can see that outliers were removed.

Chart

Description automatically generated with low confidence

# Conclusion

In our analysis, we utilized Gr.Liv.Area, Enclosed Porch, and First Floor as predictor variables in a linear regression model to estimate the Sale Price. We selected Gr.Liv.Area because of its high correlation with Sale Price and because Sale Price had a correlation closest to 0.5. The model accurately forecasted the Sale Price as the predictor variables had very significant p-values.

Despite a few outliers and influential observations in the model, the Residual versus Leverage plot indicated that removing them would not lead to significant improvements in accuracy.

To improve the model's accuracy, we explored both Stepwise Selection and Best Subset Modelling. We ultimately opted for Best Subset Modelling as it examines all possible models and ranks them based on certain criteria such as adjusted R-squared, AIC, or BIC.

It is important to note that Best Subset Modelling can be computationally expensive for large datasets, as it involves analyzing all possible predictor combinations. As such, it may not be suitable for high-dimensional datasets. Additionally, the optimal model selection process may vary based on the research question and available data, and the Best Subset Model may not always be the most optimal.

# Appendix

#Avi Milan Jani

#ALY6015 Module 1 - Regression Diagnostics with R

#Loading and installing the libraries and packages

library(dplyr)

library(ggplot2)

library(corrplot)

library(ggcorrplot)

library(RColorBrewer)

install.packages("leaps")

library(leaps)

install.packages("mice")

library(mice)

install.packages("caret")

library(caret)

#Importing the data set

ameshousing <- read.csv("C:\\Users\\avija\\Downloads\\AmesHousing.csv", header=TRUE)

#REMOVE UNWANTED COLUMNS

ameshousing <- ameshousing %>%

select(-c(Alley, Low.Qual.Fin.SF))

install.packages("openxlsx")

library(openxlsx)

write.xlsx(ameshousing, file = "ameshosuing.xlsx", sheetName = "Sheet1")

#1.Performing exploratory analysis and descriptive statistics

str(ameshousing)

str(na.omit(ameshousing))

# Store the output of str() in a variable

str\_output <- capture.output(str(ameshousing))

# Create a data frame from the str output

str\_ameshousing <- data.frame(str\_output)

# View the resulting data frame

str\_ameshousing

summary(ameshousing, na.rm = TRUE)

# Store the output of str() in a variable

summary\_output1 <- capture.output(summary(ameshousing))

# Create a data frame from the str output

summary\_ameshousing <- data.frame(summary\_output1)

# View the resulting data frame

summary\_ameshousing

output <- capture.output(str(ameshousing))

#Clean the data set

na\_count <-sapply(ameshousing,function(y) sum(length(which(is.na(y)))))

na\_count

#Input missing values by mean method

input <- mice(ameshousing, method ="mean", m =1, maxit=1)

input

# Change color values

color\_values <- c("#E87653", "#5E4FA2", "#F4A259", "#4D9AC4", "#94B447", "#DC4C46", "#0072B2", "#FFD700")

# Create a histogram of SalePrice

library(scales)

ggplot(ameshousing, aes(x = SalePrice, fill = ..count..)) +

geom\_histogram(binwidth = 50000, color = "black", alpha = 0.8) +

scale\_fill\_gradient(low = "white", high = color\_values[4]) +

xlab("Sale Price") + ylab("Count") +

ggtitle("Sale Price Distribution") +

xlim(c(0, 600000)) +

scale\_x\_continuous(labels = comma) +

theme\_minimal()

mean(ameshousing$SalePrice)

# Create a boxplot of SalePrice by Neighborhood

library(RColorBrewer)

my\_palette <- colorRampPalette(c("#f7fbff", "#deebf7", "#c6dbef", "#9ecae1", "#6baed6", "#4292c6", "#2171b5", "#084594"), space = "rgb")(28)

ggplot(ameshousing, aes(x = Neighborhood, y = SalePrice, fill = Neighborhood)) +

geom\_boxplot(outlier.shape = NA) +

scale\_fill\_manual(values = my\_palette) +

xlab("") + ylab("Sale Price") +

ggtitle("Price of Houses by Neighborhood") +

theme\_bw() +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5)) +

scale\_y\_continuous(breaks = seq(0, 800000, by = 100000),

labels = scales::dollar\_format(prefix = "$"))

# Check for missing values in numerical variables

numeric\_vars <- sapply(ameshousing, is.numeric)

summary(ameshousing[,numeric\_vars])

# Check for infinite values in numerical variables

is\_inf <- apply(ameshousing[,numeric\_vars], 2, function(x) any(!is.finite(x)))

ameshousing[,numeric\_vars][,is\_inf]

###################################################################################

# Select only the numeric columns

numeric.var <- sapply(ameshousing, is.numeric)

numeric.df <- ameshousing[, numeric.var]

# Remove rows with missing or infinite values

numeric.df <- na.omit(numeric.df)

# Create the correlation matrix

corr.matrix <- cor(numeric.df)

# Find highly correlated variables

highlyCorrelated <- findCorrelation(corr.matrix, cutoff = 0.6)

# Subset the correlation matrix to include only highly correlated variables

corr.matrix\_subset <- corr.matrix[highlyCorrelated, highlyCorrelated]

# Plot the correlation matrix

corrplot(corr.matrix\_subset,

method = "circle",

tl.col = "dark red",

tl.srt = 45,

tl.cex = 0.8,

cl.cex = 0.8,

font = 2,

addCoef.col = "transparent",

title = "\n\nCorrelation Plot for Highly Correlated Numerical Variables")

#6.

##########################################

# Get correlations of numeric columns with SalePrice

correlations <- cor(ameshousing[, sapply(ameshousing, is.numeric)], ameshousing$SalePrice)

# Sort correlations in descending order

sorted\_corr <- sort(correlations, decreasing = FALSE)

View(correlations)

# Print correlations

print(correlations)

################################################################################################

install.packages("car")

library(car)

# Get correlations of numeric columns with SalePrice

correlations <- cor(ameshousing[, sapply(ameshousing, is.numeric)], ameshousing$SalePrice)

# Sort correlations in descending order

sort(correlations, decreasing = TRUE)

## scatterplot to identify highest correlation(1)

cor(ameshousing$SalePrice, ameshousing$Gr.Liv.Area)

scatterplot(ameshousing$SalePrice ~ ameshousing$Gr.Liv.Area, data = ameshousing,

main = "Highest correlation between Above ground living area and Sale Price",

xlab = "Above Ground Area",

ylab = "Sale Price",

col = "steelblue3")

options(scipen = 999)

### scatterplot to identify lowest correlation

cor(ameshousing$SalePrice, ameshousing$Enclosed.Porch)

scatterplot(ameshousing$SalePrice ~ ameshousing$Enclosed.Porch, data = ameshousing,

main = "Lowest correlation between Enclosed Porch square area and Sale Price",

xlab = "Enclosed Porch Square Area ",

ylab = "Sale Price",

col = "darkgreen")

options(scipen = 999)

## scatterplot closest to 0.5 correlation

cor(ameshousing$SalePrice, ameshousing$X1st.Flr.SF)

scatterplot(ameshousing$SalePrice ~ ameshousing$X1st.Flr.SF, data = ameshousing,

main = "Sale price with First Floor in square feet",

xlab = "First Floor in square feet",

ylab = "Sale Price",

col = "steelblue4")

options(scipen = 999)

#7. Fit a multiple linear regression model with three continuous predictors

regression\_model <- lm(SalePrice ~ Gr.Liv.Area + Wood.Deck.SF + Lot.Area, data = ameshousing)

# Print the summary of the model

summary(regression\_model)

plot(regression\_model)

#-------------------------------------------------------------------------------------------------------------------------------

#8.

library(broom)

# Fit the model

regression\_model <- lm(SalePrice ~ Gr.Liv.Area + Wood.Deck.SF + Lot.Area, data = ameshousing)

# Extract the coefficients from the model

intercept <- regression\_model$coefficients[1]

coef1 <- regression\_model$coefficients[2]

coef2 <- regression\_model$coefficients[3]

coef3 <- regression\_model$coefficients[4]

# Print the equation of the model

cat("SalePrice = ", intercept, " + ", coef1, " \* Gr.Liv.Area + ", coef2, " \* Wood.Deck.SF + ", coef3, " \* Lot.Area")

# Fit a multiple linear regression model with three continuous predictors

regression\_model <- lm(SalePrice ~ Gr.Liv.Area + Wood.Deck.SF + Lot.Area, data = ameshousing)

#10. Check for multicollinearity using VIF

library(car)

install.packages("olsrr")

library(olsrr)

vif(regression\_model)

ols\_vif\_tol(regression\_model)

# If VIF is greater than 5 or 10 for any variable, it indicates the presence of multicollinearity.

# To correct for multicollinearity, one can consider dropping one of the correlated variables or using dimensionality reduction techniques such as principal component analysis.

## checking for outliers

ggplot(ameshousing,

aes( x=Gr.Liv.Area+Wood.Deck.SF+Lot.Area,

y=SalePrice,color=SalePrice)) +

geom\_boxplot(fill ="red"

)

## removing outliers

p1<-boxplot(ameshousing$SalePrice)$out

p2<-boxplot(ameshousing$Gr.Liv.Area)$out

p3<-boxplot(ameshousing$Wood.Deck.SF)$out

p4<-boxplot(ameshousing$Lot.Area)$out

install.packages("leaps")

library(leaps)

ameshousing$SalePrice<-as.factor(ameshousing$SalePrice)

model <- regsubsets(SalePrice ~ ., data = ameshousing)

# Fit all possible models

regfits <- regsubsets(SalePrice ~ ., data = ameshousing)

# Find the best model using adjusted R-squared

summary(regfits)$adjr2

# Get the coefficients of the best model

best\_model <- regfits$which[which.max(summary(regfits)$adjr2), ]

coef(regfits, best\_model)

# run all subsets regression

regfit.full <- regsubsets(SalePrice ~ ., data = ameshousing)

# summary of results

summary(regfit.full)

# the best model with lowest Cp statistic

which.min(regfit.full$cp)