Predictive Analysis of Housing Sales Prices in Ames, Iowa: A Two-Part Study for Century 21 Ames

Authors: Antonio Debouse and Rafia Mirza

Introduction

Century 21 Ames is looking to gain insight on what variables can best predict the sales prices of a house. We will initially see what correlations can be drawn between sales prices and increased living area in three specific neighborhoods. Secondly, we will expand our analysis to all neighborhoods, choosing additional variables based on their predictive strength. We will assess each model's predictive accuracy on unseen data to recommend the most accurate model.

Data Description

The dataset we are using for this analysis contains 79 explanatory variables describing many aspects of residential homes in Ames, Iowa. The files of interest in regards to our analysis are: train.csv(460.68 kB), test.csv (451.41 kB) and data_description.txt(13.37 kB). Additional information can be found at the hosting site Kaggle.[1] For Analysis 1, the file used is train.csv and the specific variables of interest are: SalePrice , Neighborhood and GrLivArea. For Analysis 2 the files used are train.csv and test.csv, and the specific variables of interest are: OverallQual, LotArea, YearBuilt, YearRemodAdd, GrLivArea, TotalBathrooms, TotRmsAbvGrd, PoolArea, YrSold, and MoSold. A full description of the data can be found in the data_description.txt file.

Analysis Question 1

Restatement of Problem

In our analysis, we examine Century 21 Ames' real estate market in Ames, Iowa, focusing on the NAmes, Edwards, and BrkSide neighborhoods. Our objective is to establish the correlation between house sale prices and their living area (GrLivArea), while considering the influence of neighborhood location. We estimate the relationship between sale price and living area in 100 sq. ft. increments, providing detailed figures with confidence intervals. Additionally, we scrutinize model assumptions and address outliers to ensure the robustness of our findings, aiming to quantify the impact of living area on sale prices in these neighborhoods.

[1] Anna Montoya, DataCanary. (2016). House Prices - Advanced Regression Techniques. Kaggle. https://kaggle.com/competitions/house-prices-advanced-regression-techniques

Build and Fit the Model

We utilized the train.csv dataset, filtering it to focus solely on properties in the NAmes, Edwards, and BrkSide neighborhoods. After confirming no missing values in key columns such as SalePrice and GrLivArea, we converted GrLivArea into 100 sq. ft. increments. We then built a linear model (lm) to predict SalePrice, using the transformed GrLivArea and neighborhood as variables. We evaluated the model's assumptions for linear regression accuracy, finding issues with linearity, homoscedasticity, and normality in the dataset. These issues could impact the model's accuracy. Additionally, we identified 'influential' data points, specifically houses with unique features that significantly influence our analysis. We found high leverage points, which are properties that deviate from general trends due to factors like unusually large living areas or atypical pricing within their neighborhood. A few houses with exceptionally high sale prices might distort our understanding of average prices, leading to potential inaccuracies in predicting prices for more typical neighborhood houses. We also identified influential points based on Cook's D; these properties, if excluded, would notably change our analysis's results. These influential points are illustrated in Figure A.

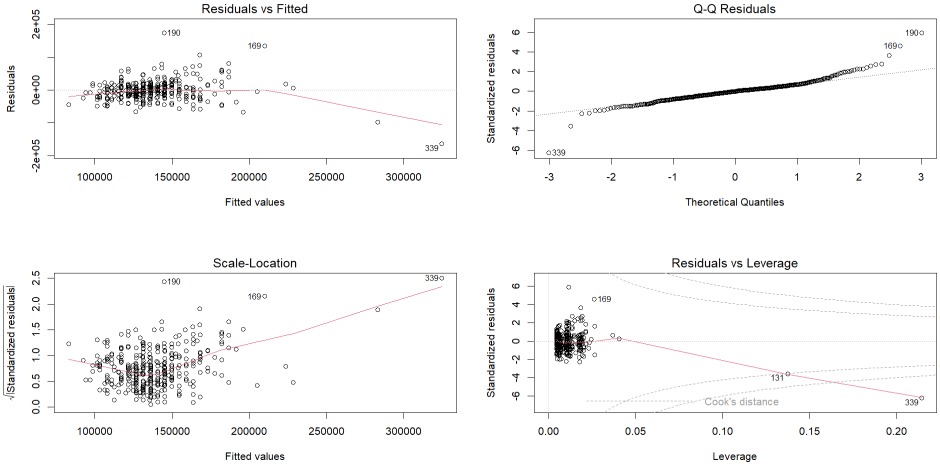


Figure A: Diagnostic Plots

We recommend conducting further analysis of the dataset to decide whether these points should be included in the analysis. This additional examination should focus on understanding the underlying reasons for their distinctiveness. For example, these outliers might be properties with unique characteristics not accounted for by the current model, like distinctive architectural features or historical significance.

Findings

The analysis indicates a significant positive correlation between house living area and sale prices, with the NAmes neighborhood exhibiting higher prices compared to the baseline BrkSide. However, in the Edwards neighborhood, increased living area does not significantly impact sale prices. This finding suggests that generally, larger houses tend to be pricier. Specifically, in the NAmes neighborhood, we are 95% certain that the true increase in sale price for every additional 100 square feet of living area ranges between $7,305.514 and $24,548.610. In the BrkSide neighborhood, this increase is estimated to be between $3,992.525 and $5,231.554. In contrast, for the Edwards neighborhood, larger living areas do not significantly affect sale prices.

Comparision of two linear regression models

We analyzed two linear regression models predicting house sale prices in Ames, Iowa. The first, a main effects model, considers the individual impacts of living area and neighborhood on sale prices. The second, an interaction effects model, additionally explores the combined influence of these factors. Our goal was to determine which model offers more accurate predictions.

Main effects model

Looking only at main effects, the stepwise regression method based on Akaike Information Criterion (AIC), we identified significant predictors for the sale price. Initially, the model (AIC = 9170.779) improved by including GrLivArea100 and as.factor(Neighborhood), reducing the AIC to 8981.379. The final model explains about 39.5% of the variance in sale prices (based on Adjusted R-squared of 0.395). Using Leave-One-Out Cross-Validation (LOOCV), this model's predictions are, on average, about $30,732.18 away from the actual sale prices (indicated by RMSE (Root Mean Square Error) of 30,732.18, which gives more weight to larger errors). If the errors are unweighted, on average, the model's predictions are about $20,797.4 away from the actual sale prices. (indicated by MAE (Mean Absolute Error) of 20,797.4).

Interaction effects model

When we incorporate interaction terms between GrLivArea100 and neighborhood, we found the model was technically a better fit to than the main effects model (with a lower AIC of 8950.639) This model explains about 44.4% of the variance in sale prices, which is a slight improvement in explaining the variability of sale prices (shown by the adjusted R-squared of 0.444). Using Leave-One-Out Cross-Validation (LOOCV), we assessed this model's predictive accuracy on unseen data as marginally better. On average, the model's predictions are about $29,683.61 away from the actual sale prices (shown by the RMSE of 29,683.61, which gives more weight to larger errors). If the errors are unweighted, on average, the model's predictions are about $20,425.92 away from the actual sale prices. (MAE of 20,425.92). While the interaction model is slightly improved over the main effect model in regards to statistical accuracy, the practical improvement is very small.

Conclusion

Our analysis of the real estate market in Ames, Iowa, focused on the NAmes, Edwards, and BrkSide neighborhoods, examining the relationship between a house's living area and its sale price. We observed that the living area positively influences sale prices, with significant variations depending on the neighborhood. Specifically, the NAmes neighborhood demonstrated a strong positive correlation between increased living area and higher prices. Our comparison of models, using stepwise regression and Leave-One-Out Cross-Validation (LOOCV), indicated that the interaction model slightly outperforms the main effects model statistically, but the practical improvement is minimal. This suggests that while both neighborhood and living area are significant factors, their combined impact on sale prices is not markedly different from when they are considered separately.

In conclusion, our initial analysis revealed a correlation between increased living area and sale price, evident in two of the three neighborhoods. Certain data points may affect the models' accuracy. In our subsequent analysis, we plan to include additional variables beyond neighborhood and living area to develop a more precise model. This approach may confirm that these outliers are rare exceptions in the real estate market, warranting their exclusion for a more generalized understanding of market trends.

Application

You can see the relationship between sales price, greater living area and neighborhood using this R Shiny Application: [https://librarianrafia.shinyapps.io/AmesRealEstate/]

Analysis Question 2

Restatement of Problem

The goal is to create several models to predict the sales prices of all the homes in Ames, Iowa and explain which model is the best. We will be comparing a simple linear regression model with using only the total size of the lot to predict the sales price of homes with a multiple linear regression model using the above ground living area square feet and full bathrooms above ground area, and a model with explanatory variables selected from the backward variables selection technique.

Note: For our model assumptions we will assume independence for this point forward and address other assumptions in-depth going forward.

Simple Linear Regression Model

For the simple linear regression model, we will use total size of the lot (LotArea) as the explanatory variable to predict the sales price of a home. LotArea variable was selected under the assumption that the total size of the land is a strong indicator of how much a property will cost or can be sold for.

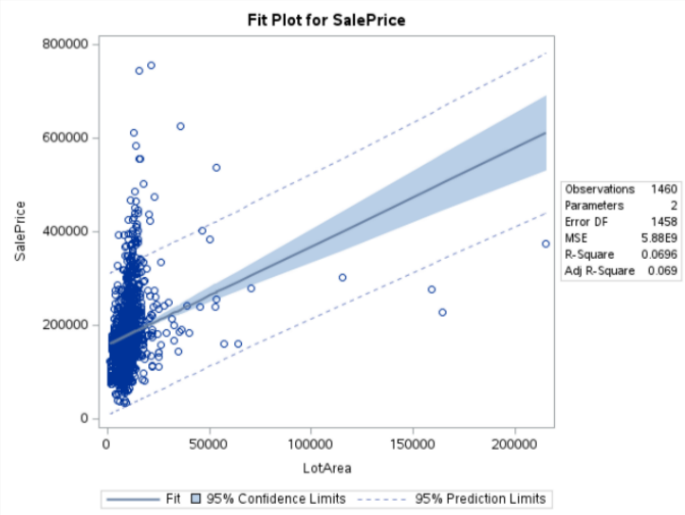


Figure 1 - Fit Plot for SalesPrice to LotArea

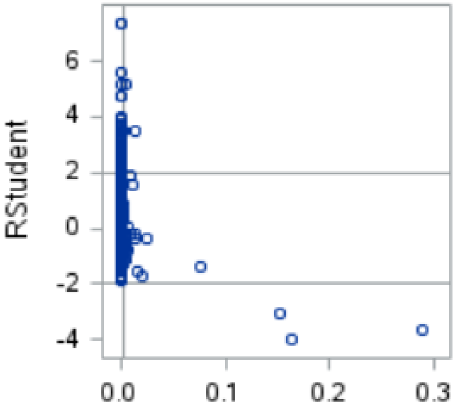


Figure 2 - Influential Residuals Plot

The SalePrice to LotArea scatterplot (Figure 1 - “Fit Plot for SalesPrice”) shows visual evidence of a linear relationship that is influenced by right-skewness in the distribution due to large outliers. Even though we have a nice cluster of properties with lot sizes around the average size of 10,516 square feet, there are several properties with lot size greater than 2 standard deviations of the mean (lots larger than 30,478 square feet). These larger lot properties appear to also cause the data to violate the equal standard deviation assumptions; as the lot size gets bigger the spread of SalePrice gets larger.

Figure – 2 Influential Residuals Plot shows some observations with major influence on the model, but they will remain in the model since we were asked to provide a model to predict all the sale price on all homes in Ames, Iowa. In the future, it is worth investigating removing these outliers noting that this will narrow the range of inference.

Looking at a Log-Log (Figure 3), the assumption for standard deviation appears to correct for unequal standard deviation.

We will continue to use a regression model although it is possible that a different model would work better considering the large cluster of observations around log_LotArea 9.

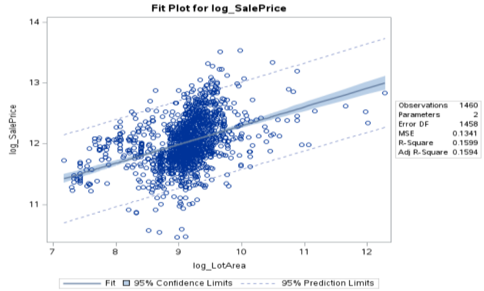


Figure 3 - Log Log Scatterplot

Multiple Linear Regression model: SalePrice ~ GrLivArea + Fullbath

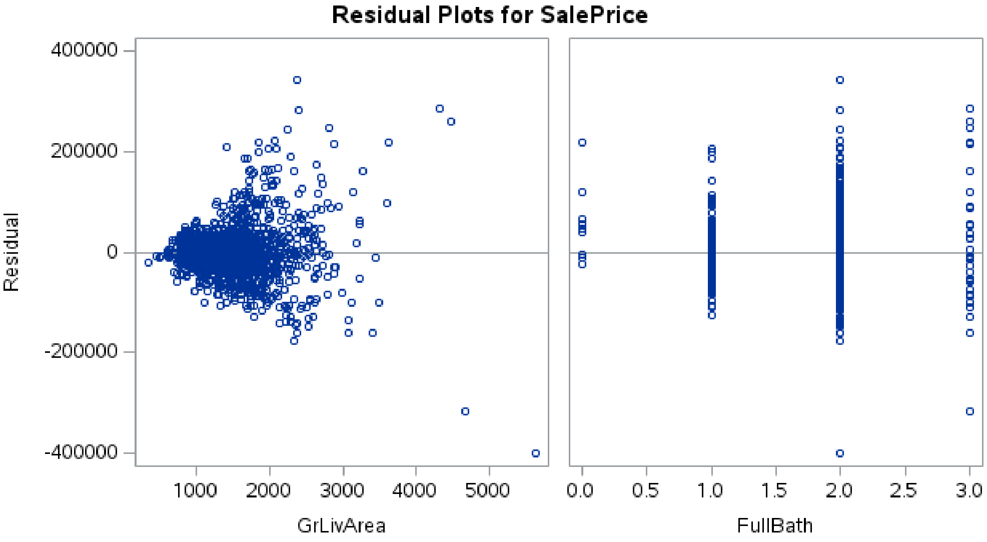


Figure 4 – Residual SalePrice Scatterplot for GrLivArea and Fullbath

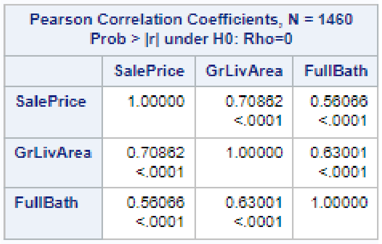
We were asked to create a multiple regression model that used GrLivArea and FullBath variables to predict SalePrice.

Checking Assumptions:

- **Normality** – There appears to be some skewness in the data from large and small outliers, but the model has 1460 observations so we can assume the Central Limit Theorem.

- **Standard Deviation** – some visual evidence against equal standard deviations between SalePrice and GrLivArea, which we have seen in our prior analysis. Simply logging the SalePrice variable does not resolve this violation and the data did not visually satisfy the standard deviation assumption until we logged both the SalePrice and GrLivArea. We needed more time to figure out how to convert predictions from a log-log model back into their original until of measure which would have been required to get a Kaggle score. Therefore, we will proceed with caution and present our solution in the *Future Analysis* section of this report.

- **Linearity** – there does appear to be a relationship between SalePrice and GrLivArea and SalePrice and FullBath variables (see: Figure 5 - Correlation Matrix image).

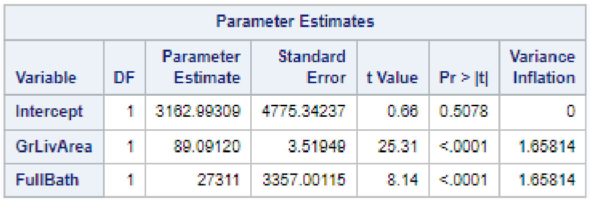


Pearson Correlation Coefficients, N = 1460
Prob > |r| under H0: Rho=0

	SalePrice	GrLivArea	FullBath
SalePrice	1.00000	0.70862 <.0001	0.56066 <.0001
GrLivArea	0.70862 <.0001	1.00000	0.63001 <.0001
FullBath	0.56066 <.0001	0.63001 <.0001	1.00000

Figure 5 - Correlation Matrix

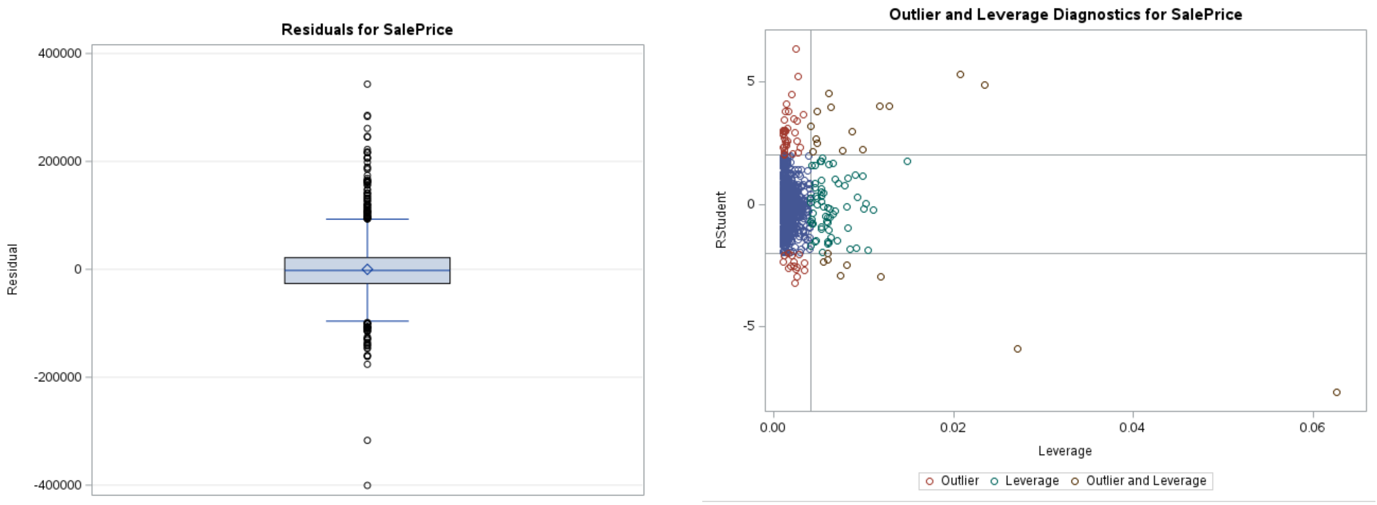
- **Independence** – there also is evidence of collinearity between GrLivArea and FullBath but the variance inflator factors are less than 2 (see MLR: SalePrice ~ GrLivArea + FullBath Parameters image). They are also significant so both explanatory variables will remain in the model.



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3162.99309	4775.34237	0.66	0.5078	0
GrLivArea	1	89.09120	3.51949	25.31	<.0001	1.65814
FullBath	1	27311	3357.00115	8.14	<.0001	1.65814

Figure 6 - MLR: SalePrice ~ GrLivArea + FullBath Parameters



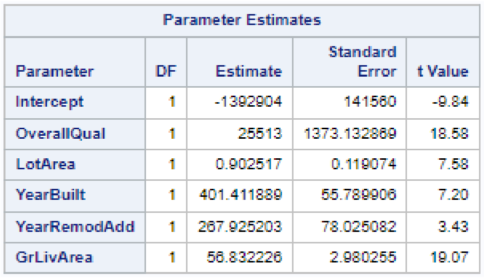
2Figure 7 - Boxplot and Influential points plot

Outliers and Influential Points: There are several properties on both the small and larger ends of the spectrum that are having a larger impact on the parameter estimates of both the MLR: SalePrice ~ GrLivArea + FullBath model and our Backward variable selection model as seen in both the boxplot and influential plot visual above. We chose to keep these observations in the models because they appear to balance each other, and it is very realistic to experience housing market that has properties that are much larger than the average and some that are much smaller (long-tails on a distribution curve).

Under the assumption that the sales price of these “outliers” also provide useful insight into the average sale price of homes in their proximity was another reason to keep these observations.

Backward Variable MLR model

The original explanatory variables that made it into the model were (OverallQual, LotArea, YearBuilt, YearRemodAdd, GrLivArea, TotRmsAbvGrd, PoolArea, YrSold, MoSold, plus the intercept). Backward selection rendered the model down to just 6 (see Backward Selection Model Results image).



Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-1392904	141560	-9.84
OverallQual	1	25513	1373.132869	18.58
LotArea	1	0.902517	0.119074	7.58
YearBuilt	1	401.411889	55.789906	7.20
YearRemodAdd	1	267.925203	78.025082	3.43
GrLivArea	1	56.832226	2.980255	19.07

Backward Selection Model Results

Comparing Competing Models

Predictive Models	**Adjusted R2**	**CV PRESS**	**Kaggle Score**
Simple Linear Regression	0.0665	7.6089E12	0.41014
Multiple Linear Regression	0.5354	3.765107E12	0.28586
Backward Selection MLR Model	0.7476	2.070991E12	0.7144

The Backward Selection MLR model perform significantly better than the SLR for SalePrice to Lot Area, and the MLR using GrLivArea and FullBath explanatory variables in all 3-performance metrics (see plot above). This could likely be due to the fact of the complexity within the housing data requiring more variables to assist the model predictions in identifying the interplay of variables that impact sales price.

This was an observational study so only association can be drawn and not causation. We also caution against using these models to make prediction on properties outside of Ames, Iowa.

Future Analysis

In future analysis we would look to investigate further if outliers should be removed, or we would segment out the properties into low, average, and high-cost housing markets. It is possible that a different method other than regression could help provide more accurate predictions on SalePrice.

Appendix

Analysis Question 1 R code

```
#load libraries
library(tidyverse)

library(caret)

library(olsrr)

#clean environment
rm(list = ls())

# Read in the data
train <- read.csv("train.csv")
head(train)

# Filter Data: from neighborhoods column, only NAmes, Edwards, and BrkSide
filter_data <- train %>%
  filter(Neighborhood %in% c("NAmes", "Edwards", "BrkSide")) %>%
  select(SalePrice, GrLivArea, Neighborhood)


head(filter_data)

##   SalePrice GrLivArea Neighborhood
## 1    118000      1077      BrkSide
## 2    157000      1253        NAmes
## 3    132000       854      BrkSide
## 4    149000      1004        NAmes
## 5    139000      1339        NAmes
## 6    134800       900        NAmes

#Check for Missing Values - SalePrice, GrLivArea columns, neighborhood data.

missing_values <- train %>%
                  select(SalePrice, GrLivArea, Neighborhood) %>%
                  summarise(across(everything(), ~sum(is.na(.))))

missing_values
```

```
##   SalePrice GrLivArea Neighborhood
## 1         0         0            0

#no missing values

#Exploratory Data Analysis: Descriptive Statistics - mean, median, standard deviati
on for SalePrice and GrLivArea by Neighborhood

descriptive_stats <- filter_data %>%
  group_by(Neighborhood) %>%
  summarise(
    Mean_SalePrice = mean(SalePrice, na.rm = TRUE),
    Median_SalePrice = median(SalePrice, na.rm = TRUE),
    SD_SalePrice = sd(SalePrice, na.rm = TRUE),
    Mean_GrLivArea = mean(GrLivArea, na.rm = TRUE),
    Median_GrLivArea = median(GrLivArea, na.rm = TRUE),
    SD_GrLivArea = sd(GrLivArea, na.rm = TRUE)
  )

#Print the descriptive statistics by neighborhood
print(descriptive_stats)

## # A tibble: 3 × 7
##   Neighborhood Mean_SalePrice Median_SalePrice SD_SalePrice Mean_GrLivArea
##   <chr>                 <dbl>            <dbl>        <dbl>          <dbl>
## 1 BrkSide             124834.           124300       40349.          1203.
## 2 Edwards             128220.           121750       43209.          1340.
## 3 NAmes               145847.           140000       33075.          1310.
## # ℹ 2 more variables: Median_GrLivArea <dbl>, SD_GrLivArea <dbl>

#Transformed the GrLivArea column into increments of 100 sq. ft
adjust_data <- filter_data %>%
 mutate(GrLivArea100 = round(GrLivArea / 100))

head(filter_data)

##   SalePrice GrLivArea Neighborhood
## 1    118000      1077      BrkSide
## 2    157000      1253        NAmes
## 3    132000       854      BrkSide
## 4    149000      1004        NAmes
## 5    139000      1339        NAmes
## 6    134800       900        NAmes

head(adjust_data)

##   SalePrice GrLivArea Neighborhood GrLivArea100
## 1    118000      1077      BrkSide           11
## 2    157000      1253        NAmes           13
## 3    132000       854      BrkSide            9
## 4    149000      1004        NAmes           10
## 5    139000      1339        NAmes           13
## 6    134800       900        NAmes            9
```

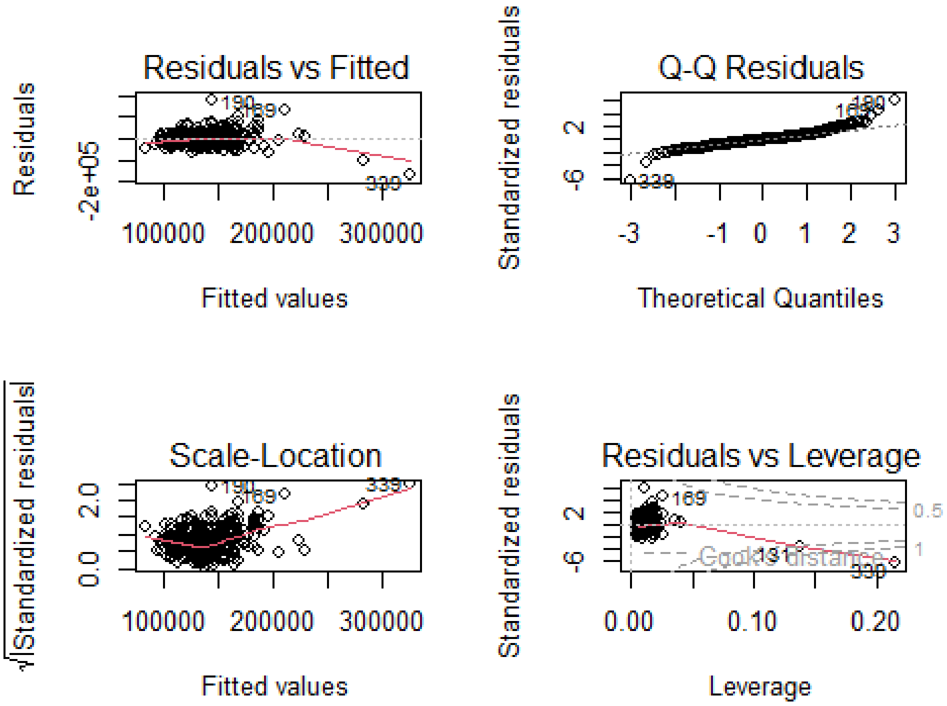
```
# Convert 'Neighborhood' to factor
adjust_data$Neighborhood <- as.factor(adjust_data$Neighborhood)

# Fit the linear model
model <- lm(SalePrice ~ GrLivArea100 + as.factor(Neighborhood), data = adjust_data)

# Summary of the model
summary(model)
```

```
## 
## Call:
## lm(formula = SalePrice ~ GrLivArea100 + as.factor(Neighborhood), 
##     data = adjust_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -164693  -16595    -294   13318  175177 
## 
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                      69410.1     5433.8  12.774  < 2e-16 ***
## GrLivArea100                      4612.0      315.1  14.638  < 2e-16 ***
## as.factor(Neighborhood)Edwards   -2991.7     4918.3  -0.608  0.54337    
## as.factor(Neighborhood)NAmes     15927.1     4384.8   3.632  0.00032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 29680 on 379 degrees of freedom
## Multiple R-squared:  0.3996, Adjusted R-squared:  0.3949 
## F-statistic: 84.09 on 3 and 379 DF,  p-value: < 2.2e-16
```

```
# confidence intervals for the model
confint(model)
```

```
##                                      2.5 %    97.5 %
## (Intercept)                      58725.865 80094.253
## GrLivArea100                      3992.525  5231.554
## as.factor(Neighborhood)Edwards  -12662.317  6678.939
## as.factor(Neighborhood)NAmes      7305.514 24548.610
```

```
# Diagnostic plots
par(mfrow = c(2, 2))
plot(model)
```



```r
#Calculate leverage values
leverage_values <- hatvalues(model)

#Calculate Cook's distance
cooks_d_values <- cooks.distance(model)

#Number of observations
n <- length(cooks_d_values)

#Number of predictors (including intercept)
k <- length(coef(model))

#Threshold for high leverage
high_leverage_threshold <- 2 * (k + 1) / n

#high leverage points
high_leverage_points <- which(leverage_values > high_leverage_threshold)

#Threshold for influential points based on Cook's Distance
cooks_d_threshold <- 4 / n

#Identify influential points based on Cook's Distance
influential_points <- which(cooks_d_values > cooks_d_threshold)

#summary
cat("High leverage points (index):", high_leverage_points, "\n")
```

```
## High leverage points (index): 53 78 80 131 136 169 339

cat("Influential points based on Cook's D (index):", influential_points, "\n")

## Influential points based on Cook's D (index): 19 48 58 64 70 80 104 131 136 140
157 167 169 180 186 190 205 227 240 262 302 322 339 370 372

#We found high Leverage points (53, 78, 80, 131, 136, 169, 339), which indicate hou
ses that don’t quite fit the general data patterns. These houses might have an unus
ually large living area or might be priced much higher or lower than other houses i
n the same neighborhood. A few houses with very high sale prices can cause us to th
ink the average house prices is higher than it is, which may lead to less accurate
predictions, especially for houses that are more typical to the neighborhoods. We a
lso found Influential Points Based on Cook’s D (19, 48, 58, 64, 70, 80, 104, 131, 1
36, 140, 157, 167, 169, 180, 186, 190, 205, 227, 240, 262, 302, 322, 339, 370, 372)
: These are houses that, if removed, would significantly change the results of our
analysis.
```

#Diagnostic plots #Residuals vs Fitted Plot (check the assumption of linearity and homoscedasticity/variance of errors). The residuals should be randomly dispersed around the horizontal line at zero, indicating that the relationship is linear and the variance of the errors is constant.This is not the case, there are indications of large residuals (observations 190, 169, and 339) Indicates potential outliers.

#Normal Q-Q Plot: If the residuals are normally distributed, the points should fall approximately along the reference line. Deviations from the line, especially in the tails, indicate departures from normality. Points at the ends (like 190, 169, and 339) deviate from the line, suggesting potential issues with normality.

#Scale-Location Plot: A horizontal line with equally (randomly) spread points along the line would suggest homoscedasticity. A funnel shape (either opening up or down) indicates heteroscedasticity.In this plot, the presence of a pattern or a non-random spread could suggest non-constant variance.

#Residuals vs Leverage Plot: (Levergae = influential cases) Points outside the dashed Cook’s distance lines (which are at a Cook’s distance of 0.5 and 1) are considered to be potentially influential. Observation 339 has high leverage and a large residual, making it particularly influential.Observations 190, 169, and 131 also stand out.

#Diagnotic plots suggest the presence of outliers and influential observations (like points 190, 169, and 339) that could potentially affect the model’s performance and should be investigated further.

#The model has several observations identified as having high leverage or high cook’s d (influential). If additional research is conducted, it important to determine whether these observations are data entry errors, outliers, or legitimate values that represent important aspects of the dataset.

```
#Perform stepwise regression based on AIC
stepwise_model <- ols_step_both_aic(model, details = TRUE)



#stepwise findings:output shows that adding GrLivArea100 and as.factor(Neighborhood
) reduces the AIC, suggesting they are important predictors.
#Step 0: AIC = 9170.779  SalePrice ~ 1
#Step 1 : AIC = 9010.12  SalePrice ~ GrLivArea100
#Step 2 : AIC = 8981.379 SalePrice ~ GrLivArea100 + as.factor(Neighborhood)
```

```r
#cross-validation using the "Leave-One-Out Cross-Validation" (LOOCV) method
train_control <- trainControl(method = "LOOCV")

#Train the model using specified predictors
model2 <- train(SalePrice ~ GrLivArea100 + as.factor(Neighborhood),
               data = adjust_data,
               trControl = train_control,
               method = "lm")

#Output performance metrics
print(model2)
```

```
## Linear Regression
##
## 383 samples
##   2 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 382, 382, 382, 382, 382, 382, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   30732.18  0.3541232  20797.4
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```r
#linear model is trained using SalePrice as the response variable and GrLivArea100
and neighborhood as predictors, applying LOOCV. The output indicates the model's pe
rformance metrics, with an RMSE of approximately 30,732, a R-squared of 0.354, and
an MAE of approximately 20,797.

# Considering interactions in the model
fit_interaction <- lm(SalePrice ~ (GrLivArea100 + as.factor(Neighborhood))^2, data
= adjust_data)
#new model is fitted considering interaction terms between GrLivArea100 and as.fact
or(Neighborhood). This model examines if the effect of living area on sale price ch
anges depending on the neighborhood.

#Perform stepwise regression based on AIC for interaction model
stepwise_interaction_model <- ols_step_both_aic(fit_interaction, details = TRUE)
```

```
## Stepwise Selection Method
## -------------------------
##

##  Step 0: AIC = 9170.779
##  SalePrice ~ 1
##
```

```
##  Step 1 : AIC = 8970.64
##  SalePrice ~ GrLivArea100:as.factor(Neighborhood)
##
##
##  Step 2 : AIC = 8950.639
##  SalePrice ~ GrLivArea100:as.factor(Neighborhood) + as.factor(Neighborhood)
##
##  Step 3 : AIC = 8950.639
##  SalePrice ~ GrLivArea100:as.factor(Neighborhood) + as.factor(Neighborhood) + Gr
LivArea100
##
-----------------------------------------------------------------------------------
----------------------------------------------

#Stepwise regression based on AIC is performed again, this time including interacti
on terms. The process results in a final model with an AIC of 8950.639, indicating
that the interaction terms are significant and improve the model compared to the on
e without interactions.

#Train the interaction model using cross-validation
model_interaction <- train(SalePrice ~ (GrLivArea100 + as.factor(Neighborhood))^2,
                           data = adjust_data,
                           trControl = train_control,
                           method = "lm")

#Output the interaction model's performance metrics
print(model_interaction)

## Linear Regression
##
## 383 samples
##   2 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 382, 382, 382, 382, 382, 382, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   29683.61  0.3986143  20425.92
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

#The interaction model is trained using LOOCV for cross-validation. The output show
s the performance metrics for this model, with an RMSE of approximately 29,683, a R
-squared of 0.398, and an MAE of approximately 20,425.

#Adjusted R-squared: (Higher values are better. They indicate that the model explai
ns more variability in the response variable and has potentially better explanatory
power.)
```

```
#Internal Cross-Validation (CV) PRESS: RMSE (Root Mean Square Error) and MAE (Mean
Absolute Error) metrics from LOOCV): Lower values are better. They indicate that th
e model has better predictive accuracy on new, unseen data.
```

#stepwise:The stepwise regression model, selected based on the lowest AIC, found that both `GrLivArea100` and `as.factor(Neighborhood)` are significant predictors of `SalePrice`. The inclusion of these variables resulted in an AIC reduction from 9170.779 to 8981.379. The model with just these predictors yielded an adjusted R-squared of 0.395, suggesting that approximately 39.5% of the variance in sale prices is explained by the model. #LOOV: The internal cross-validation using Leave-One-Out Cross-Validation (LOOCV) indicated an RMSE of 30,732.18 and an MAE of 20,797.4. These results show the model’s prediction errors and provide an insight into its predictive accuracy on new data. #Further analysis:consider interaction effects between `GrLivArea100` and neighborhoods. The interaction model reported an AIC of 8950.639, suggesting a better fit than the main effects model. This model achieved an adjusted R-squared of 0.444, indicating a slight improvement in the model’s explanatory power. Cross-validation results showed an RMSE of 29,683.61 and an MAE of 20,425.92, both marginally lower than the main effects model, pointing to a slightly better prediction accuracy.#conclusion: the model with interaction terms shows a slight increase in both explanatory power and predictive accuracy. The adjusted R-squared value is higher, and the internal cross-validation metrics (RMSE and MAE) are lower for the interaction model compared to the main effects model. This suggests that the interaction model may be the more appropriate choice for predicting house sale prices based on the given predictors. However, the differences are relatively small, indicating that the interaction terms, while statistically significant, may not lead to substantial practical improvements in prediction.

Analysis Question 2 SAS code

```
/* Generated Code (IMPORT) */
/* Source File: train.csv */
/* Source Path: /home/adebouse0 */
/* Code generated on: 12/1/23, 5:27 AM */
%web_drop_table(TrainingSet);
FILENAME REFFILE '/home/adebouse0/train.csv';
PROC IMPORT DATAFILE=REFFILE
DBMS=CSV
OUT=TrainingSet;
GETNAMES=YES;
RUN;
PROC CONTENTS DATA=TrainingSet; RUN;
%web_open_table(TrainingSet);
/*View Trainingset dataset*/
proc print data= work.trainingset;
run;
/*View descriptive information for columns in dataset*/
proc contents data=work.trainingset;
run;
/*Frequency Analysis*/
proc freq data=work.trainingset;
tables _all_ / missing;
run;
/*Summary Statistics on NUM data type columns*/
proc means data=work.trainingset n mean sum std min max;
var LotArea GrLivArea;
run;
/*Removing columns with null values greater that 50%*/
data work.trainingset;
set work.trainingset;
drop PoolQC MiscFeature Alley Fence;
run;
data work.trainingset;
set work.trainingset;
LotArea100sqft = LotArea/100;
MasVnrArea100sqft = MasVnrArea/100;
GrLivArea100sqft = GrLivArea/100;
```

```
GarageArea100sqft = GarageArea/100;
run;
/*Simple Linear Regression model*/
proc reg data=work.trainingset plots=all;
model SalePrice = GrLivArea;
run;
/*removing outliers that are +/- 2sd*/
data work.trainingset2;
set work.trainingset;
if GrLivArea > 2566.42 or GrLivArea < 464.5 then delete;
run;
/*rerun SLR*/
proc reg data=work.trainingset2 plots=all;
model SalePrice = GrLivArea;
run;
/*logging SalePrice to address unequal standard deviations*/
data work.trainingset2;
set work.trainingset2;
log_SalePrice = log(SalePrice);
run;
```

```
/*running SLR with log_SalePrice variable*/
proc reg data=work.trainingset2 plots=all;
model log_SalePrice = GrLivArea;
run;
/*running SLR with SalePrice ~ LotArea*/
proc reg data=work.trainingset plots=all;
model log_SalePrice = LotArea;
run;
/*filtering out properties with lot area greater than 115000 square feet*/
data work.trainingset2;
set work.trainingset;
if LotArea > 115000 then delete;
run;
/*rerunning SLR with SalePrice ~ LotArea*/
proc reg data=work.trainingset2 plots=all;
model SalePrice = MSSubClass;
run;
/*Replace missing values in LotFrontage with 0*/
data work.trainingset;
set work.trainingset2;
if missing(LotFrontage) or LotFrontage='N/A' then LotFrontage = 0;
run;
/*Total Bathrooms on entire property*/
data work.trainingset2;
set work.trainingset2;
TotalBathrooms = BsmtFullBath + BsmtHalfBath + FullBath + HalfBath;
run;
/*REMEMBER GOING FORWARD OUR ASSUMPTIONS ARE ONLY FOR PROPERTIES BETWEEN GrLivArea 464.5<"PROPERTY"<2566.42 AND
Lot Area less
/*MLR SalePrice ~ GrLivArea + Fullbath*/
proc reg data=work.trainingset2 plots=all;
model SalePrice = GrLivArea Fullbath;
run;
/*Backward MLR*/
proc glmselect data=work.trainingset2 plots=all seed=1565493;
partition fraction(test=.2);
model SalePrice = OverallQual LotArea YearBuilt YearRemodAdd GrLivArea TotalBathrooms TotRmsAbvGrd PoolArea YrSold
MoSold / se
run;
/*Forward MLR*/
proc glmselect data=work.trainingset2 plots=all seed=1565493;
partition fraction(test=.2);
model SalePrice = OverallQual LotArea YearBuilt YearRemodAdd GrLivArea TotalBathrooms TotRmsAbvGrd PoolArea YrSold
MoSold / se
run;
/*Stepwise MLR*/
proc glmselect data=work.trainingset2 plots=all seed=1565493;
```