Analysis of Delaware Real Estate Data Set

Group 2: Huyen V, Jerome J, Michelle C, Radhika S

MATH 4339 - 15445

Instructor: Dr. Poliak

Data Set Introduction

The data set used in the this project is a subset of the USA Real Estate data set posted on Kaggle by the author, Mr. Sakib, who extracted the data from realtor.com, which is the second most visited real estate website in the U.S. with over 100 million active users. The original data set included an excess of 20K listings spanning 18 states. To make it easier to perform analysis, the subset was created by focusing solely on the state of Delaware with around 2K listings and 6 columns for variables. Through performing analysis on this subset, the goal is to primarily find answers to two important questions.

- **Question 01:** Does the size of the house(sq-ft, acre_lot, # of beds, # of baths) vary significantly based on ZIP code?

- **Question 02:** Does the price of the house vary significantly based on size of the house (sq-ft, # of beds, # of baths) ?

Through the applications of multivariate data analysis techniques learned through the course of this semester such as MANOVA, PCA, and Multiple Linear Regression we hope to find answers to these questions and to expand our understanding of Multivariate Data Analysis.

Also, since ZIP code is an ubiquitous concept used daily, we wanted to have a closer look at its origin and reasons behind the popularity. Commonly consisting of 5 digits, the U.S. ZIP codes, or **Zone Improvement Code Plan code**, was first implemented in 1963, thanks to Mr. Robert Moon, who developed the system after a 20-year development process. Prior to its implementation, USPS used a less structured system that was only applicable to cities. The current system evolved out of necessity due to a rapid swell in the population, as well as a boom in commerce following the first and second World War eras.

There are three main parts of the 5-digit ZIP Code — the national area, the region or city, and the delivery area. The United States Postal Service (USPS) has segmented the country into 10 ZIP Code areas. Starting in the northeast, they are number 0-9. The addition of the 4 digits in the 1980s called the ZIP+4 system, allowed senders to indicate an even more precise location, such as particular block or apartment building and even which side of the street. So that is a quick take on ZIP codes. Now, let's bring the focus back to the data set in question.

Data set Analysis

- **Assessing and cleansing the data set**

code:

```
realtor = read.csv(file.choose())
cleanRealtor = na.omit(realtor)

state_to_filter = "Delaware"
subset_data_state <- cleanRealtor[cleanRealtor$state == state_to_filter, ]

numeric_data <- subset_data_state[, sapply(subset_data_state, is.numeric)]
```

Will be grouping the ZIP codes and analyzing the data for change of variance among the 4 response variables (“number of baths”, “number of beds”, “size of the lot”, and “house size”) using MANOVA.

- **Find the unique ZIP codes**

code:

```
unique_values <= unique(numeric_data $ zip_code)
sorted <- sort(unique_values) #used to sort in ascending order
print(sorted)
```

```
##  [1] 19701 19702 19703 19706 19707 19709 19711 19713 19720 19730 19734 19801
## [13] 19802 19803 19804 19805 19806 19807 19808 19809 19810 19901 19977
```

- **Sort the 23 ZIP codes into 3 groups (8 in first two 7 in the last one) inorder to perform MANOVA**

Group01: 19701 19702 19703 19706 19707 19709 19711 19713
Group02: 19720 19730 19734 19801 19802 19803 19804 19805
Group03: 19806 19807 19808 19809 19810 19901 19977

- **Sorting the entire data set ascending order in terms of ZIP__code**

code:

```
sorted_RE <- numeric_data[order(numeric_data $ zip_code), ]
View(sorted_RE)
```

- **Create 3 groups based on ranges**

```
sorted_RE$groups <- cut(sorted_RE$zip_code, breaks = c( 0,19720, 19806, 99999999), labels = c("Group 1"

attach(sorted_RE)
```

- **Hypothesis of the Test**

Null: The three group mean vectors are equal to each other.
Alt: At least one of the group mean vector is different.

- **Assumptions of the Test**

- Multivariate Normal - Hypothesis test

- No sub populations within the 3 species.

- Common Variance - Covariance Matrix - BoxM test

- Independent samples

Since the data set is quite large, we can assume MVN based on the Central Limit Theorem.
Next, use the BoxM test to check Homogeneity of Covariance Matrix.

code:

```
library(biotools)
boxM(sorted_RE[, c(1, 2, 3, 5)], sorted_RE[,7])
```

```
## Loading required package: MASS
```

```
## ---
## biotools version 4.2
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  sorted_RE[, c(1, 2, 3, 5)]
## Chi-Sq (approx.) = 2791.5, df = 20, p-value < 2.2e-16
```

In the output, it is shown that the p-value is 2.2e-16. This means that there is strong evidence that the variance-covariance matrix is not the same among all three groups.

- **Perform the MANOVA. Give important details**

code:

```
RE_data.1 = manova(cbind(bed, bath, acre_lot, house_size) ~ groups, data = sorted_RE)
summary(RE_data.1, test="Wilks")
```

```
##             Df   Wilks approx F num Df den Df    Pr(>F)
## groups       2 0.83621   39.787      8   3402 < 2.2e-16 ***
## Residuals 1704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is listed as 2.2e-16 in the summary, which indicates that there is strong evidence against the NULL hypothesis of the MANOVA. At least, one of the ZIP code based groups have a difference in number of beds, number of baths, acre_lot or house_size.
Test-statistic: 39.787

- **Bonferroni corrected ANOVAs to assess the significance of individual variables**

code:

```
RE_data.lm1 = lm(bed ~ groups , data = sorted_RE)
anova(RE_data.lm1)

RE_data.lm2 = lm(bath ~ groups , data = sorted_RE)
anova(RE_data.lm2)

RE_data.lm3 = lm(acre_lot ~ groups , data = sorted_RE)
anova(RE_data.lm3)

RE_data.lm4 = lm(house_size ~ groups , data = sorted_RE)
anova(RE_data.lm4)

**Create box plots

par(mfrow=c(1,2))
boxplot(bed ~ groups,data = sorted_RE)
boxplot(bath ~ groups,data = sorted_RE)
boxplot(acre_lot ~ groups,data = sorted_RE)
boxplot(house_size ~ groups,data = sorted_RE)

**Create a profile plot

library(profileR)
library(ggplot2)
group_1 <- subset(sorted_RE, groups == "Group 1")
group_2 <- subset(sorted_RE, groups == "Group 2")
group_3 <- subset(sorted_RE, groups == "Group 3")
selected_group1 = group_1[, c("bed", "bath", "acre_lot")]
selected_group2 = group_2[, c("bed", "bath", "acre_lot")]
selected_group3 = group_3[, c("bed", "bath", "acre_lot")]
m = rbind(colMeans(selected_group1), colMeans(selected_group2), colMeans(selected_group3))
rownames(m, do.NULL = F)
rownames(m) = c("Group 1", "Group 2", "Group 3")
```

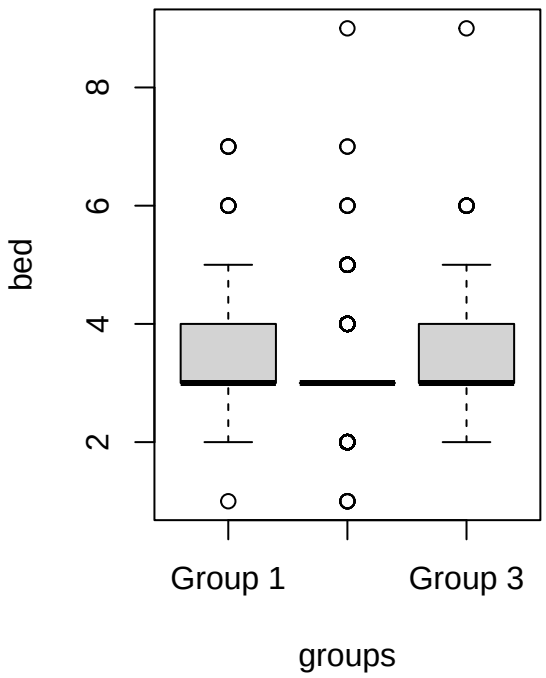
```
profileplot(m, rownames(m), standardize = F)

selected_group1 = group_1[, c("bed", "bath", "acre_lot", "house_size")]
selected_group2 = group_2[, c("bed", "bath", "acre_lot", "house_size")]
selected_group3 = group_3[, c("bed", "bath", "acre_lot", "house_size")]
n = rbind(colMeans(selected_group1), colMeans(selected_group2), colMeans(selected_group3))
profileplot(n, rownames(m), standardize = F)
```

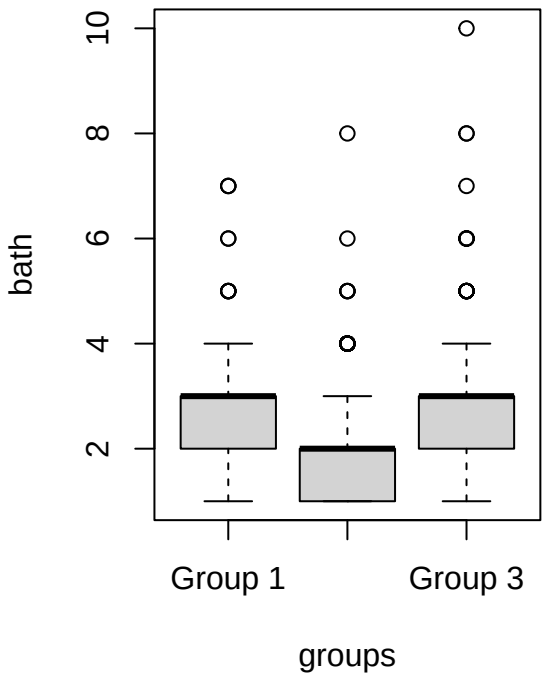
```
## Analysis of Variance Table
##
## Response: bed
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## groups       2   35.29 17.6466  25.444 1.294e-11 ***
## Residuals 1704 1181.82  0.6936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

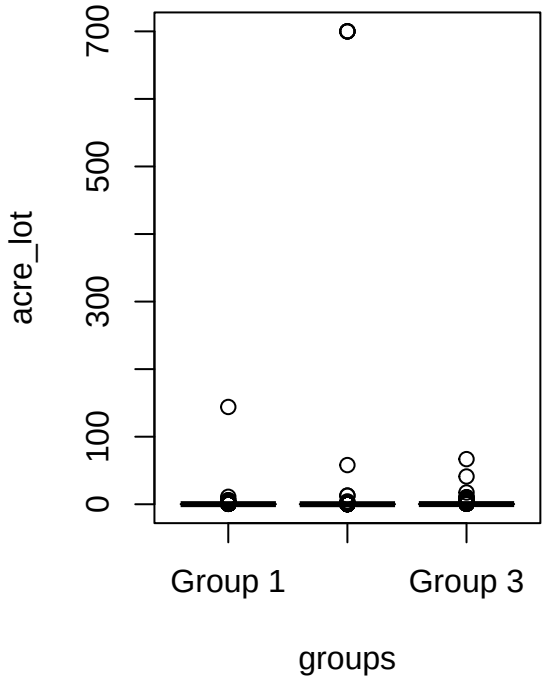
```
## Analysis of Variance Table
##
## Response: bath
##             Df Sum Sq Mean Sq F value    Pr(>F)
## groups       2  245.0 122.500  132.07 < 2.2e-16 ***
## Residuals 1704 1580.6   0.928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

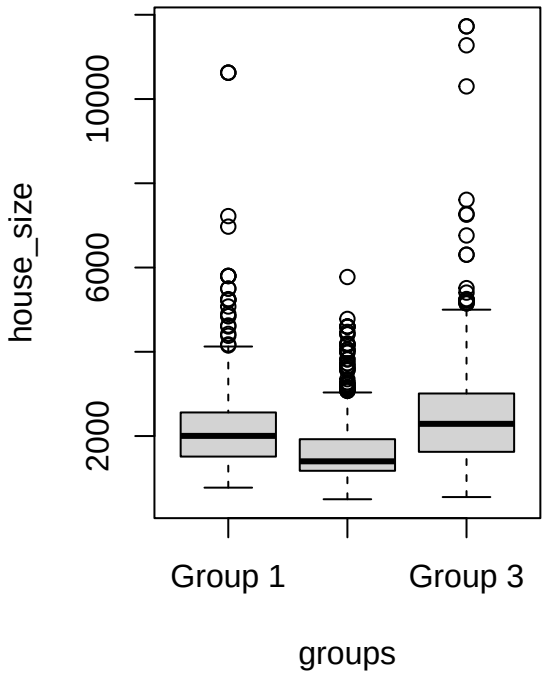
```
## Analysis of Variance Table
##
## Response: acre_lot
##             Df  Sum Sq Mean Sq F value Pr(>F)
## groups       2    3094  1547.1  1.3305 0.2646
## Residuals 1704 1981293  1162.7
```

```
## Analysis of Variance Table
##
## Response: house_size
##             Df     Sum Sq   Mean Sq F value    Pr(>F)
## groups       2  271284326 135642163  122.51 < 2.2e-16 ***
## Residuals 1704 1886664680   1107198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

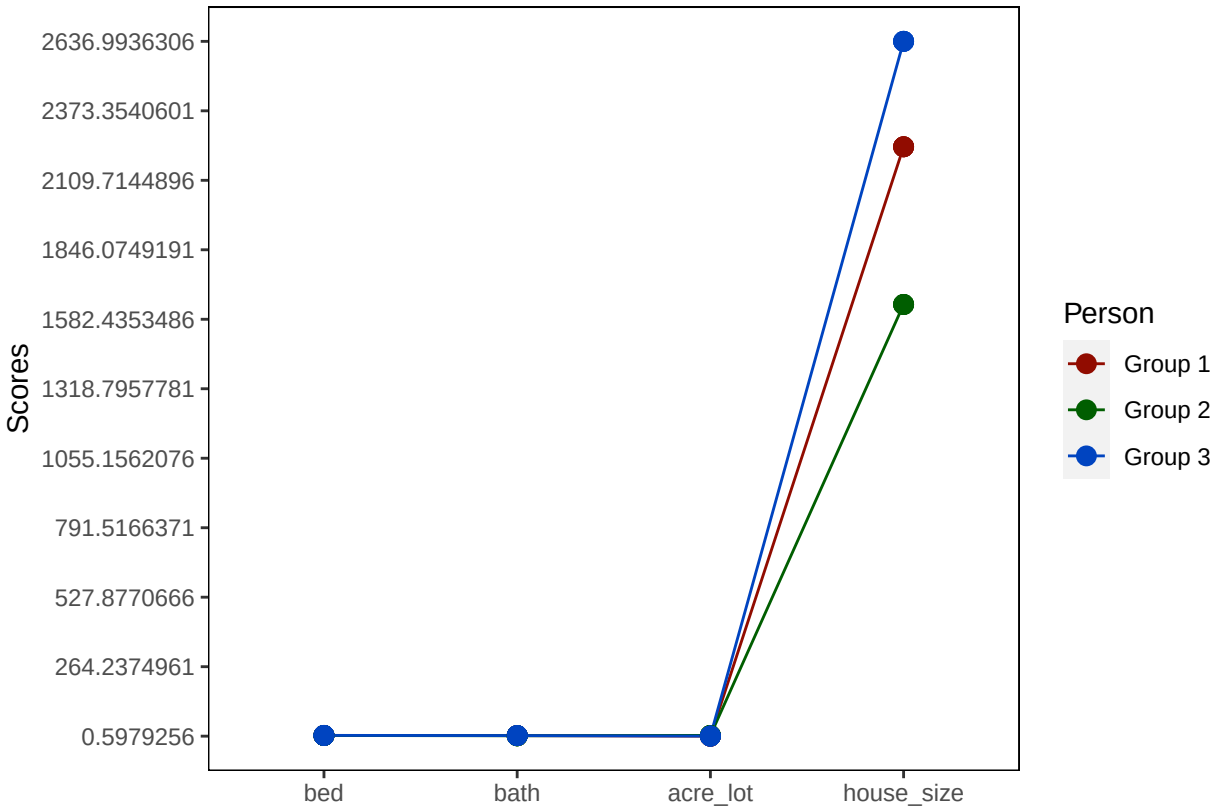

bed
2
4
6
8
Group 1
Group 3
groups


bath
2
4
6
8
10
Group 1
Group 3
groups






```
## [1] "row1" "row2" "row3"
```



After reviewing the ANOVA tables for each of the variables, acre_lot is the only one with a p-value greater than ($0.05/4 = 0.0125$). This means that there is difference in means for each variable except for acre_lot among the 3 ZIP code groups.
The size of the lot could be similar for the 3 groups, however the size of the house built on the lot is different due to income gap between different areas.

Data Set Summary for Question 1

- Based on the Box plots, Group1 and Group 3 ZIP codes appears to have consistently larger values and greater variance than Group2.
- These zip codes could possibly be inhabited by families with higher incomes. While Group2 area probably could be inhabited by lower income families. Higher income usually results in larger houses.
- Also when observing the profile plot, it is evident that the largest houses in the state are located in areas covered by Group3 ZIP codes.
- Insights like these could be used by:
- – Retail businesses to find the most suitable spot to start their next branch.
- – Banks to decide on interest rates for consumers.

- – Non-profits to decide on the parts of the state that needs the most attention.
- – Law enforcement agencies to understand and mitigate crime rates within the state
- – AND SO MUCH MORE APPLICATIONS.

Correlation Analysis

```
##                    bed        bath      acre_lot     zip_code    house_size
## bed         1.000000000  0.50975367 -0.0034790978  0.048798323  0.5926197677
## bath        0.509753667  1.00000000 -0.0473446511 -0.040023288  0.7264327083
## acre_lot   -0.003479098 -0.04734465  1.0000000000  0.038195001 -0.0002100579
## zip_code    0.048798323 -0.04002329  0.0381950005  1.000000000  0.0049964836
## house_size  0.592619768  0.72643271 -0.0002100579  0.004996484  1.0000000000
## price       0.443383761  0.67515669  0.0060939315  0.027018924  0.7600489925
##                   price
## bed         0.443383761
## bath        0.675156694
## acre_lot    0.006093931
## zip_code    0.027018924
## house_size  0.760048992
## price       1.000000000
```

Price has a high correlation with house size and bath, and a moderate correlation with bed.This is expected, as larger homes often have higher prices.

House size has a high correlation with bed and bath. This indicates that,larger homes tend to have more bedrooms and bathrooms.

Acre Lot has a weak correlation with bath, bed, and zip code.

Zip code and the number of bedrooms or house size correlation is quite weak (around 0.05 and 0.005, respectively).

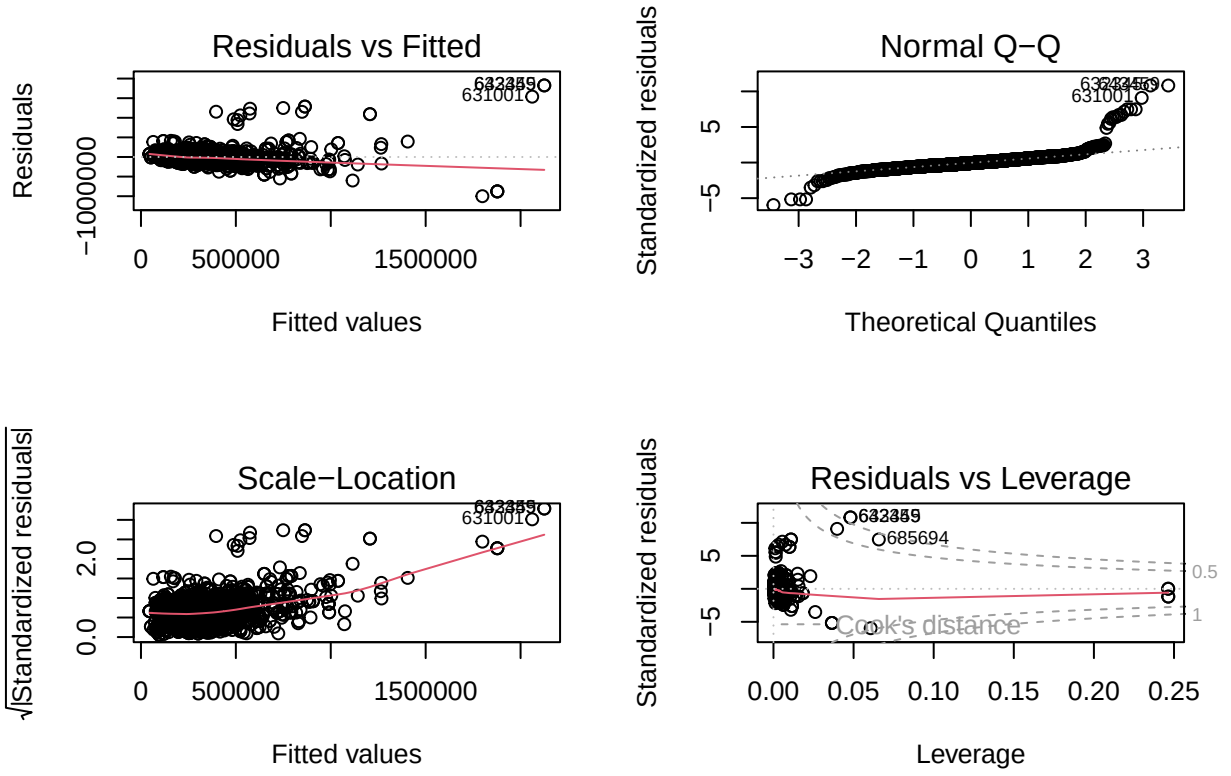
Multivariate Linear Regression

```
model.lm <- lm(price ~ acre_lot + bed + bath + house_size, data = numeric_data)
summary(model.lm)
```

```
##
## Call:
## lm(formula = price ~ acre_lot + bed + bath + house_size, data = numeric_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -999793  -75698  -11787   63675 1825172
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -71384.849  17379.889  -4.107 4.19e-05 ***
## acre_lot       152.936    123.001   1.243   0.2139
## bed          -14325.647   6215.411  -2.305   0.0213 *
## bath          72202.595   5962.227  12.110  < 2e-16 ***
## house_size      145.277      5.849  24.838  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 172900 on 1702 degrees of freedom
## Multiple R-squared:  0.6113, Adjusted R-squared:  0.6104
## F-statistic: 669.1 on 4 and 1702 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(model.lm)
```



From the linear model summary, it is evident that all the variables except acre_lot groups are significant variables in determining price. Since acre_lot is not significant, we can employ a stepwise regression procedure to determine the most optimal model.

The model explains 61.13% of the variability in house price, which is reasonable considering that no transformations have been applied to the data. However, the plots reveal that the model does not meet the assumptions of linearity, normality, and homoscedasticity.

Residuals vs. Fitted Values Plot shows a clear pattern.

The normal probability plot of residuals does not show a straight line. This indicates that the residuals are not normally distributed.

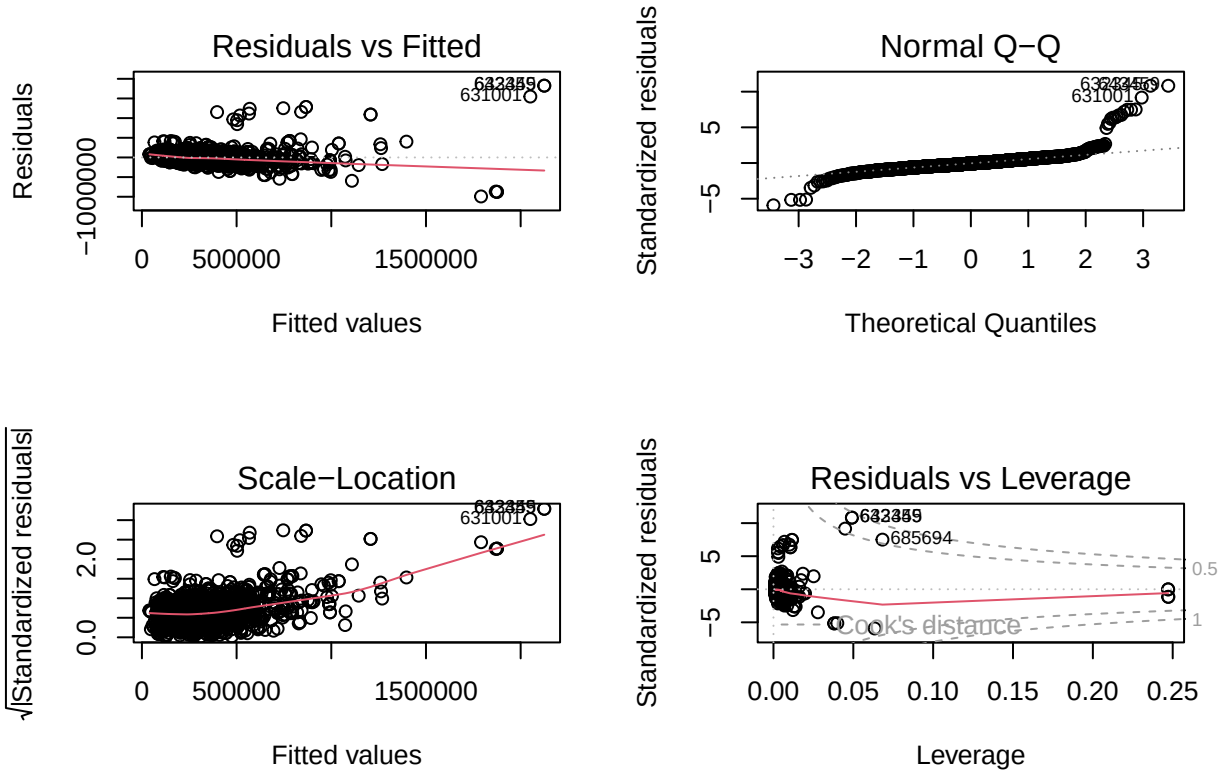
The scale-location plot does not show a horizontal line at zero. This indicates that the residuals are not homoscedastic.

The Cook's distance plot shows points that fall above the Cook's distance cutoff line and a clear pattern along the y-axis. This indicates that there are influential observations that are unduly affecting the model.

```
#adding zipcode

model.lm <- lm(price ~ acre_lot + bed + bath + house_size + sorted_RE$groups, data = numeric_data)
summary(model.lm)
```

```
##
## Call:
## lm(formula = price ~ acre_lot + bed + bath + house_size + sorted_RE$groups,
##     data = numeric_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -991814  -76090  -11419   62002 1825313
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -75606.961  18011.573  -4.198 2.84e-05 ***
## acre_lot                    153.712    123.050   1.249   0.2118
## bed                      -14455.408   6228.561  -2.321   0.0204 *
## bath                      71680.901   6010.912  11.925  < 2e-16 ***
## house_size                  145.278      5.853  24.821  < 2e-16 ***
## sorted_RE$groupsGroup 2    9008.809   9715.453   0.927   0.3539
## sorted_RE$groupsGroup 3    6565.921  12613.921   0.521   0.6028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 172900 on 1700 degrees of freedom
## Multiple R-squared:  0.6115, Adjusted R-squared:  0.6101
## F-statistic: 445.9 on 6 and 1700 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(model.lm)
```



From the linear model summary, it is evident that all the variables except acre_lot, zip code groups are significant variables of house size in determining price.

The model explains 61.17% of the variability in house price. However, the plots reveal that the model does not meet the assumptions of linearity, normality, and homoscedasticity.

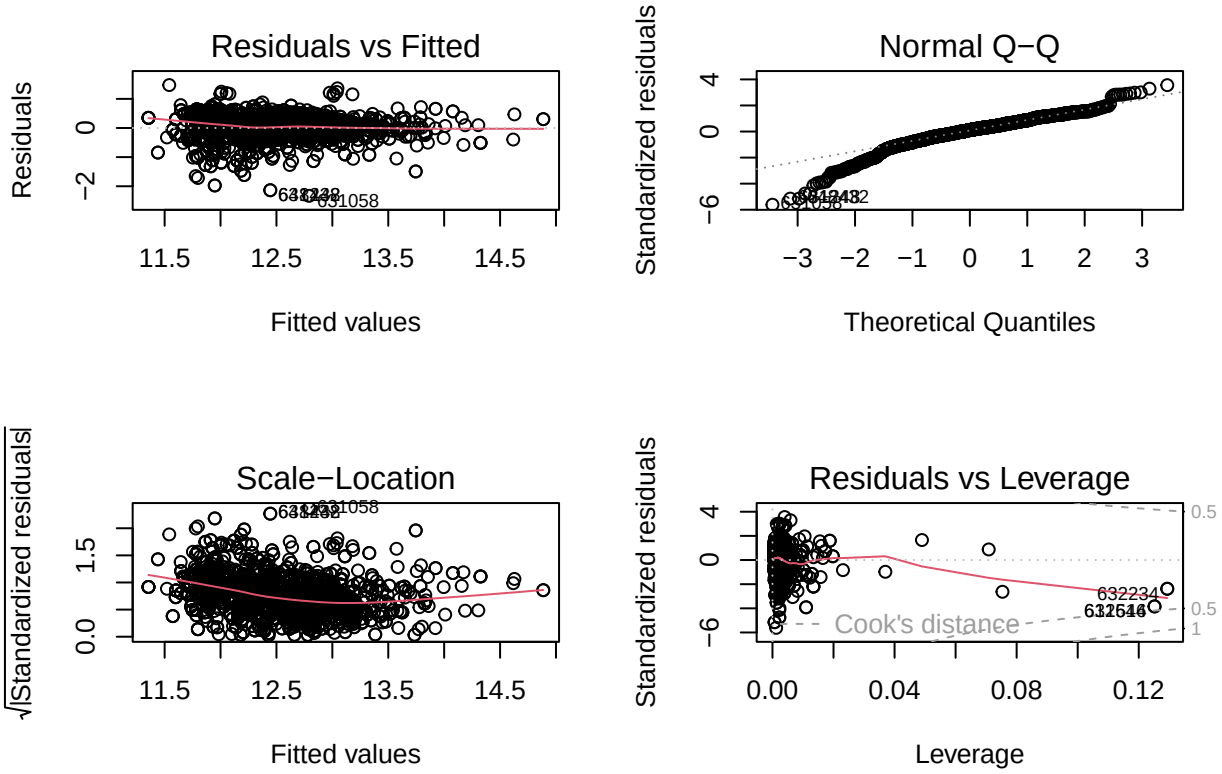
Overall, there isn't significant improvement with this model. Therefore, we will attempt to improve the model by transforming the data and performing stepwise regression.

Transforming Data

```
#Transformed bed, bath, acre_lot, house_size, price variables
transformed_data <- numeric_data
numeric_vars <- transformed_data[, c('bed', 'bath', 'acre_lot', 'house_size', 'price')]

#Logarithmic transformation
log_transformed_data <- log1p(numeric_vars)

# Replace the transformed variables in the original data frame
transformed_data[, c('bed', 'bath', 'acre_lot', 'house_size', 'price')] <- log_transformed_data


model.lm2 <- lm(price ~ acre_lot + bed + bath + house_size, data = transformed_data)
summary(model.lm2)
```

```
##
## Call:
## lm(formula = price ~ acre_lot + bed + bath + house_size, data = transformed_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32905 -0.18848  0.04679  0.25808  1.47346
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.68525    0.21486  31.114   <2e-16 ***
## acre_lot     0.21582    0.02309   9.348   <2e-16 ***
## bed         -0.16241    0.06584  -2.467   0.0137 *
## bath         0.67117    0.04851  13.834   <2e-16 ***
## house_size   0.70180    0.03706  18.937   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.415 on 1702 degrees of freedom
## Multiple R-squared:  0.596,  Adjusted R-squared:  0.595
## F-statistic: 627.7 on 4 and 1702 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(model.lm2)
```



We attempted to transform bed, bath, acre_lot, house_size, and price data to enhance the normality of the

model. While the significance of acre_lot improved, and the plots suggested slightly enhanced linearity and equal variance, the R-squared value decreased to 59.6%, and normality did not improve.

Residuals vs. Fitted Values Plot shows a clear pattern.

The normal probability plot of residuals does not show a straight line. This indicates that the residuals are not normally distributed.

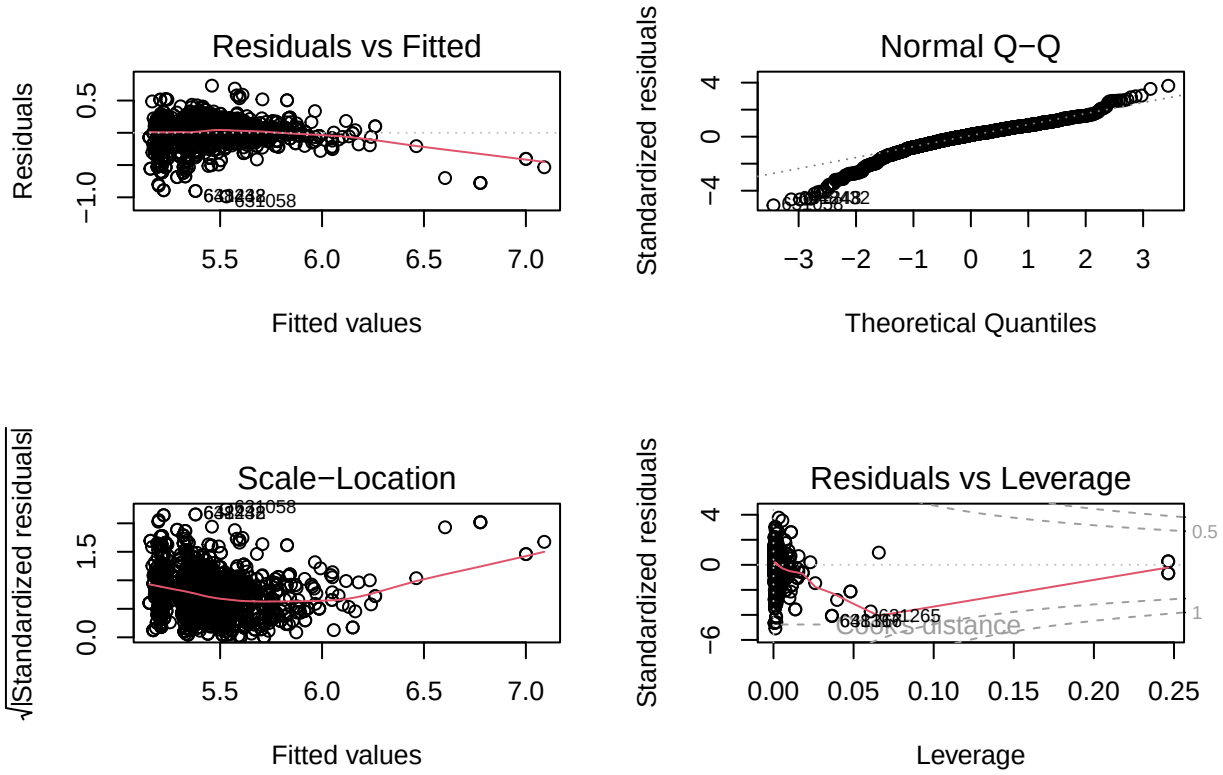
The scale-location plot does not show a horizontal line at zero. This indicates that the residuals are not homoscedastic.

The Cook's distance plot shows points that fall above the Cook's distance cutoff line and a clear pattern along the y-axis. This indicates that there are influential observations that are unduly affecting the model.

```
# Transformed price variable
transformed_data <- numeric_data
numeric_vars <- transformed_data[, 'price']

#Logarithmic transformation
log_transformed_data <- log10(numeric_vars)

# Replace the transformed variables in the original data frame
transformed_data[, 'price'] <- log_transformed_data

model.lm2 <- lm(price ~ acre_lot + bed + bath + house_size, data = transformed_data)
summary(model.lm2)
```

```
##
## Call:
## lm(formula = price ~ acre_lot + bed + bath + house_size, data = transformed_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98647 -0.09062  0.02599  0.12238  0.73056
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.001e+00  1.953e-02 256.082   <2e-16 ***
## acre_lot     1.772e-04  1.382e-04   1.282    0.200
## bed         -4.245e-03  6.984e-03  -0.608    0.543
## bath         1.134e-01  6.699e-03  16.920   <2e-16 ***
## house_size   9.526e-05  6.572e-06  14.495   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1942 on 1702 degrees of freedom
## Multiple R-squared:  0.5308, Adjusted R-squared:  0.5297
## F-statistic: 481.4 on 4 and 1702 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(model.lm2)
```



We attempted to transform the price data to enhance the normality of the model. In this model, acre_lot and bed are insignificant, and the plots did not enhanced linearity and equal variance, the R-squared value decreased to 53.08%, and normality did not improve. Consequently, we will continue using the untransformed dataset for our subsequent analysis.

Residuals vs. Fitted Values Plot shows a clear pattern.

The normal probability plot of residuals does not show a straight line. This indicates that the residuals are not normally distributed.

The scale-location plot does not show a horizontal line at zero. This indicates that the residuals are not homoscedastic.

The Cook's distance plot shows points that fall above the Cook's distance cutoff line and a clear pattern along the y-axis. This indicates that there are influential observations that are unduly affecting the model.

Step Function

```
step(model.lm)
```

```
## Start:  AIC=41181.72
## price ~ acre_lot + bed + bath + house_size + sorted_RE$groups
##
##                     Df  Sum of Sq        RSS   AIC
## - sorted_RE$groups  2 2.5828e+10 5.0855e+13 41179
```

```
## - acre_lot          1 4.6657e+10 5.0876e+13 41181
## <none>                           5.0829e+13 41182
## - bed               1 1.6105e+11 5.0990e+13 41185
## - bath              1 4.2520e+12 5.5081e+13 41317
## - house_size        1 1.8420e+13 6.9249e+13 41708
##
## Step:  AIC=41178.58
## price ~ acre_lot + bed + bath + house_size
##
##              Df  Sum of Sq        RSS   AIC
## - acre_lot    1 4.6193e+10 5.0901e+13 41178
## <none>                     5.0855e+13 41179
## - bed         1 1.5873e+11 5.1014e+13 41182
## - bath        1 4.3819e+12 5.5237e+13 41318
## - house_size  1 1.8433e+13 6.9288e+13 41705
##
## Step:  AIC=41178.13
## price ~ bed + bath + house_size
##
##              Df  Sum of Sq        RSS   AIC
## <none>                     5.0901e+13 41178
## - bed         1 1.5776e+11 5.1059e+13 41181
## - bath        1 4.3408e+12 5.5242e+13 41316
## - house_size  1 1.8550e+13 6.9452e+13 41707
```

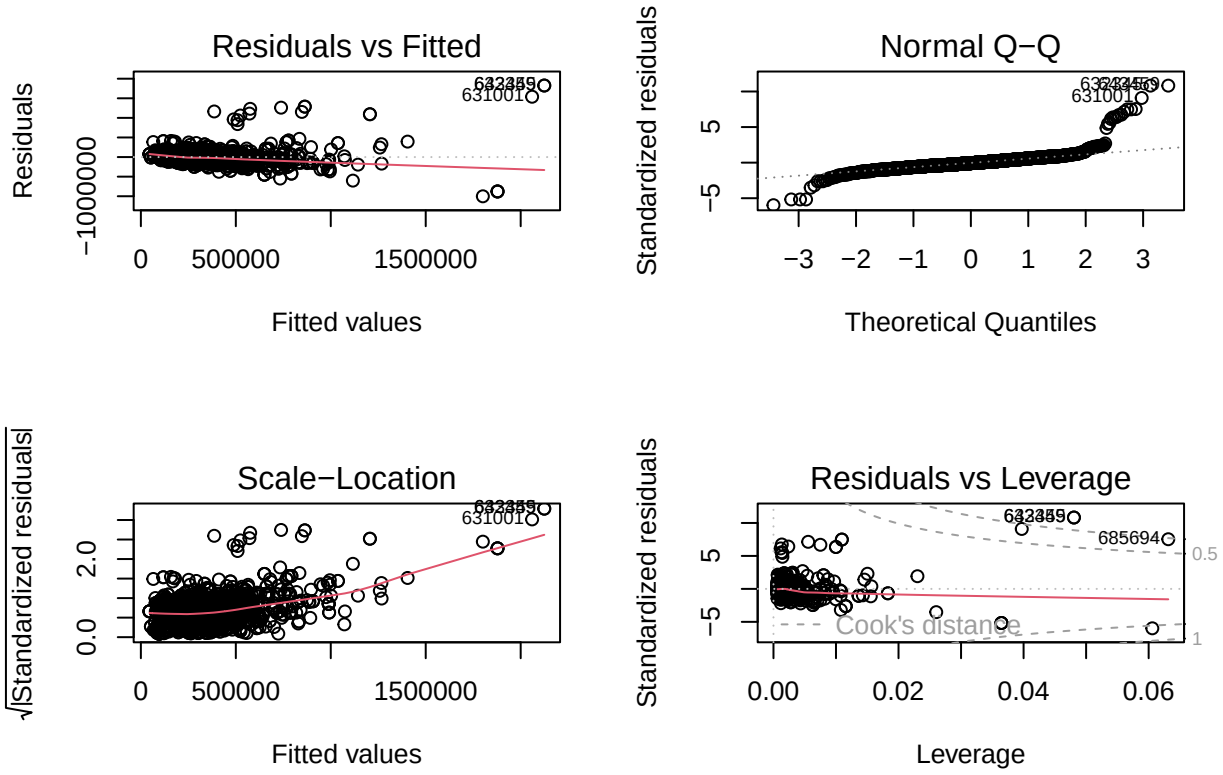
```
##
## Call:
## lm(formula = price ~ bed + bath + house_size, data = numeric_data)
##
## Coefficients:
## (Intercept)          bed         bath   house_size
##    -70654.4     -14281.3      71692.7        145.6
```

```
step.model.lm <- lm(formula = price ~ bed + bath + house_size, data = numeric_data)
summary(step.model.lm)
```

```
##
## Call:
## lm(formula = price ~ bed + bath + house_size, data = numeric_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1001879   -75554   -12373    63680  1825681
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -70654.436  17372.742  -4.067 4.98e-05 ***
## bed         -14281.291   6216.305  -2.297   0.0217 *
## bath         71692.693   5949.060  12.051  < 2e-16 ***
## house_size     145.597      5.844  24.913  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 172900 on 1703 degrees of freedom
## Multiple R-squared:  0.6109, Adjusted R-squared:  0.6102
## F-statistic: 891.4 on 3 and 1703 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(step.model.lm)
```



The linear regression model suggests that the acreage variable is not significant and can be removed. The linear model without the acre_lot & zip code: price ~ bed + bath + house_size with an AIC value of 41178.13, which is lower than the original model's AIC value of 41180.56.

Since all three variables (bed, bath, and house_size) have p-values less than 0.05, they are statistically significant predictors of price. The model also explains a substantial portion of the variability in price, with an R-squared value of 0.61.

However, the model's fit is not perfect. The diagnostic plots for the new model indicate that it does not meet the assumptions of linearity, normality, and homoscedasticity.This suggests that there may be other factors, beyond house size, that influence price.

Summary for Question 2

To answer the question, "Does the price of the house vary significantly based on the size of the house (sq-ft, # of beds, # of baths)?", we first explore the data by examining the correlation between each variable. The data reveal a high correlation between price, house size, beds, and baths. Acre Lot and Zip Code show weak correlations with all variables.

We create the initial linear model with the formula price ~ acre_lot + bed + bath + house_size. This model explains 61.13% of the house price variability, but diagnostic plots reveal that assumptions for linearity, normality, and homoscedasticity are not met. From the linear model, all the variables except acre_lot are significant predictors of house size in determining price.

In an attempt to enhance the model, we add the zip code variable. However, this model also falls short of meeting assumptions, with an explanatory power of 61.17%. All the variables except acre_lot and zip code groups are significant predictors of house size in determining price.

In subsequent attempts, we apply log transformation to bed, bath, acre_lot, house_size, and price data. However, the results do not show significant improvement, and diagnostic plots reveal continued violations of model assumptions, with the R-squared value decreasing to 59.6%.

In another attempt, we use log transformation with only the price data. Again, the results do not show significant improvement, and diagnostic plots reveal continued violations of model assumptions, with the R-squared value decreasing to 53.08%.

Consequently, we decide to continue using the first linear model for our step function analysis. The new model, discovered without the acre lot variable, demonstrates the statistical significance of bed, bath, and house_size, explaining 61% of price variability. However, diagnostic plots reveal continued violations of model assumptions.

In conclusion, while the size of the house significantly influences the price, the analysis acknowledges limitations, including the impact of unaccounted factors. Despite attempts to enhance the model through transformation and variable selection, the perfect fit is not achieved. Further exploration and identification of additional influencing factors are recommended for a more comprehensive understanding of housing price determinants.

References

- Data set - kaggle.com
- Zip code info - loqate.com , smarty.com