Benjamin Martin

R Code

Nashville Housing Data Analysis

July 2025

R CODE

```
# Preprocessing

rm(list=ls())

library(readxl)

library(dplyr)

library(ggplot2)

library(car)

library(rio)


# Data Import and Initial Inspection

data <- read_excel("C:/Users/Admin/Desktop/USF/Found Bus Statistics/Final Project/Nashville Housing Raw Data.xlsx", na = c("NA", ""))

colSums(is.na(data))


# Data Cleaning

data_clean <- data %>%

  select(-contains("X"), -contains("Unnamed"), -matches("(?i)suite.*condo"), -matches("(?i)state")) %>%

  na.omit()


str(data_clean)

summary(data_clean)


# Create Primary Data Set

set.seed(6304)

primary_data <- data_clean %>% filter('Year Built' > 1949) %>% sample_n(500)


# Exploratory Data Analysis (EDA)

str(primary_data)
```

```
sale_price_ci <- t.test(primary_data$'Sale Price', conf.level = 0.95)
sale_price_ci
t.test(primary_data$'Sale Price', mu = 250000, alternative = "greater")

# Regression Part 1
model <- lm(`Sale Price` ~ `Land Use` + `Sold As Vacant` + `Multiple Parcels Involved in Sale` +
              Neighborhood + Acreage + `Building Value` + `Land Value` +
              `Finished Area` + Bedrooms + `Full Bath` + `Half Bath` +
              `Foundation Type` + `Year Built`,
            data = primary_data)

summary(model)

confint(model)

library(car)
vif(model)
par(mfrow=c(2,2))
plot(model)

new_data <- data.frame(
  `Land Use` = factor("SINGLE FAMILY", levels = unique(primary_data$`Land Use`)),
  `Sold As Vacant` = factor("No", levels = unique(primary_data$`Sold As Vacant`)),
  `Multiple Parcels Involved in Sale` = factor("No", levels = unique(primary_data$`Multiple Parcels Involved in Sale`)),
  Neighborhood = 3650,
  Acreage = 1.5,
  `Building Value` = 135000,
  `Land Value` = 106300,
```

```
  `Finished Area` = 241300,
  Bedrooms = 3,
  `Full Bath` = 2,
  `Half Bath` = 1,
  `Foundation Type` = factor("CRAWL", levels = unique(primary_data$`Foundation Type`)),
  `Year Built` = 1959,
  check.names = FALSE
)

predict(model, newdata = new_data, interval = "prediction")

# Analysis of Variance
# Bedrooms
leveneTest(`Sale Price` ~ as.factor(Bedrooms), data = primary_data)

anova1 <- aov(`Sale Price` ~ as.factor(Bedrooms), data = primary_data)
summary(anova1)

TukeyHSD(anova1)
plot(TukeyHSD(anova1), las = 1)

# Full Bath
leveneTest(`Sale Price` ~ as.factor(`Full Bath`), data = primary_data)

anova2 <- aov(`Sale Price` ~ as.factor(`Full Bath`), data = primary_data)
summary(anova2)

TukeyHSD(anova2)
plot(TukeyHSD(anova2), las = 1)
```

```
# Grade
leveneTest(`Sale Price` ~ as.factor(Grade), data = primary_data)

anova3 <- aov(`Sale Price` ~ as.factor(Grade), data = primary_data)
summary(anova3)

TukeyHSD(anova3)
plot(TukeyHSD(anova3), las = 1)
```

RESULTS WITH ANSWERS

Introduction

For this assignment, a cleaned dataset of residential property sales in Nashville was analyzed to explore what factors influence home sale prices. The focus was on a subset of 500 randomly selected homes built after 1949 and used various statistical techniques, including t-tests, regression, and ANOVA, to evaluate relationships between sale price and features like land use, square footage, number of bedrooms and bathrooms, and property grade. The goal was to uncover patterns and determine which variables significantly impact home values in the area.

Data Import and Initial Inspection

```
data <- read_excel("C:/Users/Admin/Desktop/USF/Found Bus Statist
ics/Final Project/Nashville Housing Raw Data.xlsx", na = c("NA",
""))
New names:
• `` -> `...1`
> colSums(is.na(data))
                                ...1
                                   0
                         Unnamed: 0
                                   0
                           Parcel ID
                                   0
                            Land Use
                                   0
                  Property Address
                                 159
                  Suite/ Condo   #
                             50527
                     Property City
                                 159
                         Sale Date
                                   0
                        Sale Price
                                   0
                   Legal Reference
                                   0
                    Sold As Vacant
                                   0
Multiple Parcels Involved in Sale
                                   0
                        Owner Name
                             31375
                           Address
                               30619
                              City
                               30619
                              State
                               30619
                            Acreage
                               30619
                       Tax District
                               30619
                       Neighborhood
                               30619
                              image
                               31301
                         Land Value
                               30619
                    Building Value
                               30619
                       Total Value
                               30619
                     Finished Area
```

```
                32470
      Foundation Type
                32472
          Year Built
                32471
    Exterior Wall
                32471
               Grade
                32471
            Bedrooms
                32477
           Full Bath
                32359
           Half Bath
                32490
```

```
>
```

Data Cleaning

```
str(data_clean)
```

```
tibble [23,721 × 26] (S3: tbl_df/tbl/data.frame)
 $ ...1                          : num [1:23721] 1 2 3 4 5 7 8
9 10 11 ...
 $ Parcel ID                     : chr [1:23721] "105 11 0 080
.00" "118 03 0 130.00" "119 01 0 479.00" "119 05 0 186.00" ...
 $ Land Use                      : chr [1:23721] "SINGLE FAMIL
Y" "SINGLE FAMILY" "SINGLE FAMILY" "SINGLE FAMILY" ...
 $ Property Address              : chr [1:23721] "1802  STEWAR
T PL" "2761  ROSEDALE PL" "224  PEACHTREE ST" "316  LUTIE ST" ...
 $ Property City                 : chr [1:23721] "NASHVILLE" "
NASHVILLE" "NASHVILLE" "NASHVILLE" ...
 $ Sale Date                     : POSIXct[1:23721], format: "
2013-01-11" ...
 $ Sale Price                    : num [1:23721] 191500 202000
32000 102000 93736 ...
 $ Legal Reference               : chr [1:23721] "20130118-000
6337" "20130124-0008033" "20130128-0008863" "20130131-0009929" ..
.
 $ Sold As Vacant                : chr [1:23721] "No" "No" "No
" "No" ...
 $ Multiple Parcels Involved in Sale: chr [1:23721] "No" "No" "No
" "No" ...
 $ Owner Name                    : chr [1:23721] "STINSON, LAU
RA M." "NUNES, JARED R." "WHITFORD, KAREN" "HENDERSON, JAMES P. &
LYNN P." ...
 $ Address                       : chr [1:23721] "1802  STEWAR
T PL" "2761  ROSEDALE PL" "224  PEACHTREE ST" "316  LUTIE ST" ...
 $ City                          : chr [1:23721] "NASHVILLE" "
NASHVILLE" "NASHVILLE" "NASHVILLE" ...
```

```
 $ Acreage                          : num [1:23721] 0.17 0.11 0.1
7 0.34 0.17 0.2 0.2 0.4 0.34 0.23 ...
 $ Neighborhood                     : num [1:23721] 3127 9126 313
0 3130 3130 ...
 $ image                            : chr [1:23721] "\\114000\\91
0001.JPG" "\\131000\\191001.JPG" "\\133000\\721001.JPG" "\\134000
\\474001.JPG" ...
 $ Land Value                       : num [1:23721] 32000 34000 2
5000 25000 25000 16000 16000 25000 25000 21500 ...
 $ Building Value                   : num [1:23721] 134400 157800
243700 138100 86100 ...
 $ Total Value                      : num [1:23721] 168300 191800
268700 164800 113300 ...
 $ Finished Area                    : num [1:23721] 1149 2091 214
6 1969 1037 ...
 $ Foundation Type                  : chr [1:23721] "PT BSMT" "SL
AB" "FULL BSMT" "CRAWL" ...
 $ Year Built                       : num [1:23721] 1941 2000 194
8 1910 1945 ...
 $ Grade                            : chr [1:23721] "C" "C" "B" "
C" ...
 $ Bedrooms                         : num [1:23721] 2 3 4 2 2 2 2
2 2 3 ...
 $ Full Bath                        : num [1:23721] 1 2 2 1 1 1 1
1 1 1 ...
 $ Half Bath                        : num [1:23721] 0 1 0 0 0 0 0
0 0 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:32915] 1 7 18 19 26
27 29 30 31 32 ...
  ..- attr(*, "names")= chr [1:32915] "1" "7" "18" "19" ...
```

```
summary(data_clean)
      ...1        Parcel ID           Land Use
 Min.   :    1   Length:23721       Length:23721
 1st Qu.:13074   Class :character   Class :character
 Median :27219   Mode  :character   Mode  :character
 Mean   :27592
 3rd Qu.:42023
 Max.   :56615
 Property Address   Property City
 Length:23721       Length:23721
 Class :character   Class :character
 Mode  :character   Mode  :character


   Sale Date                     Sale Price
 Min.   :2013-01-02 00:00:00   Min.   :     100
 1st Qu.:2014-03-13 00:00:00   1st Qu.:  125000
 Median :2015-02-17 00:00:00   Median :  185000
 Mean   :2015-01-20 20:04:09   Mean   :  274912
 3rd Qu.:2015-12-15 00:00:00   3rd Qu.:  324900
```

```
Max.   :2016-10-31 00:00:00   Max.   :10750000
Legal Reference    Sold As Vacant
Length:23721       Length:23721
Class :character   Class :character
Mode  :character   Mode  :character


Multiple Parcels Involved in Sale  Owner Name
Length:23721                       Length:23721
Class :character                   Class :character
Mode  :character                   Mode  :character


  Address              City              Acreage
Length:23721       Length:23721       Min.   : 0.0400
Class :character   Class :character   1st Qu.: 0.1900
Mode  :character   Mode  :character   Median : 0.2700
                                      Mean   : 0.4558
                                      3rd Qu.: 0.4500
                                      Max.   :47.5000
 Neighborhood     image              Land Value
Min.   : 107   Length:23721       Min.   :    100
1st Qu.:3130   Class :character   1st Qu.:  22000
Median :4026   Mode  :character   Median :  29900
Mean   :4445                      Mean   :  69015
3rd Qu.:6229                      3rd Qu.:  60300
Max.   :9530                      Max.   :1869000
Building Value     Total Value      Finished Area
Min.   :   1400   Min.   :  12600   Min.   :  450
1st Qu.:  83900   1st Qu.: 109700   1st Qu.: 1242
Median : 117500   Median : 154700   Median : 1633
Mean   : 173012   Mean   : 244700   Mean   : 1919
3rd Qu.: 189200   3rd Qu.: 278100   3rd Qu.: 2214
Max.   :5824300   Max.   :6402600   Max.   :19728
Foundation Type      Year Built      Grade
Length:23721       Min.   :1799   Length:23721
Class :character   1st Qu.:1948   Class :character
Mode  :character   Median :1960   Mode  :character
                   Mean   :1964
                   3rd Qu.:1983
                   Max.   :2017
   Bedrooms        Full Bath        Half Bath
Min.   : 0.000   Min.   : 0.000   Min.   :0.0000
1st Qu.: 3.000   1st Qu.: 1.000   1st Qu.:0.0000
Median : 3.000   Median : 2.000   Median :0.0000
Mean   : 3.094   Mean   : 1.897   Mean   :0.2867
3rd Qu.: 4.000   3rd Qu.: 2.000   3rd Qu.:1.0000
Max.   :11.000   Max.   :10.000   Max.   :3.0000

>
```

Create Primary Data Set

```
set.seed(6304)

primary_data <- data_clean %>% filter('Year Built' > 1949) %>% sample_n(500)
```

Exploratory Data Analysis (EDA)

```
str(primary_data)
```

tibble [500 × 26] (S3: tbl_df/tbl/data.frame)

$ ...1 : num [1:500] 21836 38204 7535 9630 51216 ...

$ Parcel ID : chr [1:500] "150 05 0 225.00" "071 16 0 295.00" "146 16 0 216.00" "129 05 0 021.00" ...

$ Land Use : chr [1:500] "SINGLE FAMILY" "SINGLE FAMILY" "SINGLE FAMILY" "SINGLE FAMILY" ...

$ Property Address : chr [1:500] "401 SAFFORD VIEW DR" "728 DOUGLAS AVE" "5114 KINCANNON DR" "6608 ROLLING FORK DR" ...

$ Property City : chr [1:500] "ANTIOCH" "NASHVILLE" "NASHVILLE" "NASHVILLE" ...

$ Sale Date : POSIXct[1:500], format: "2014-09-26" ...

$ Sale Price : num [1:500] 129900 340000 224900 372000 214900 ...

$ Legal Reference : chr [1:500] "20141001-0090505" "20150918-0094903" "20130924-0100262" "20131114-0117666" ...

$ Sold As Vacant : chr [1:500] "No" "No" "No" "No" ...

$ Multiple Parcels Involved in Sale: chr [1:500] "No" "No" "No" "No" ...

$ Owner Name : chr [1:500] "SMITH, MARIO M. & KAREN R." "JORSTAD, ALISA L. & RYAN E., II" "CORRIGAN, FRANCIS W., III" "SEAMAN, JASON R. & NOLAN, ADELE F." ...

$ Address : chr [1:500] "401 SAFFORD VIEW DR" "728 DOUGLAS AVE" "5114 KINCANNON DR" "6608 ROLLING FORK DR" ...

$ City : chr [1:500] "ANTIOCH" "NASHVILLE" "NASHVILLE" "NASHVILLE" ...

$ Acreage : num [1:500] 0.28 0.17 0.5 1.85 0.38 0.51 0.26 0.27 0.21 0.28 ...

$ Neighborhood : num [1:500] 6028 2026 4026 4430 6031 ...

```
 $ image                 : chr [1:500] "\\177000\\881001.JPG" "\\48000\\794001.JPG"
"\\166000\\824001.JPG" "\\141000\\312001.JPG" ...
 $ Land Value            : num [1:500] 22000 27000 47000 181100 25000 ...
 $ Building Value        : num [1:500] 97800 227900 161700 165300 116400 ...
 $ Total Value           : num [1:500] 119800 254900 208700 346400 141400 ...
 $ Finished Area         : num [1:500] 1950 2468 1205 2437 2178 ...
 $ Foundation Type       : chr [1:500] "FULL BSMT" "CRAWL" "CRAWL" "CRAWL" ...
 $ Year Built            : num [1:500] 1972 2014 1960 1958 1986 ...
 $ Grade                 : chr [1:500] "C" "C" "C" "B" ...
 $ Bedrooms              : num [1:500] 3 4 3 3 3 2 3 3 3 3 ...
 $ Full Bath             : num [1:500] 1 3 1 2 3 2 1 3 2 1 ...
 $ Half Bath             : num [1:500] 1 0 1 0 0 0 1 0 0 0 ...
 - attr(*, "na.action")= 'omit' Named int [1:32915] 1 7 18 19 26 27 29 30 31 32 ...
  ..- attr(*, "names")= chr [1:32915] "1" "7" "18" "19" ...
>
sale_price_ci

        One Sample t-test

data:  primary_data$"Sale Price"
t = 20.304, df = 499, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 255848.9 310667.1
sample estimates:
mean of x
   283258
```

Using a one-sample t-test, the 95% confidence interval for the average sale price came out to be between $255,848.90 and $310,667.10. This means we can be 95% confident that the true average sale price for homes in our sample falls somewhere in that range.

```
> t.test(primary_data$'Sale Price', mu = 250000, alternative = "greater")

        One Sample t-test

data:  primary_data$"Sale Price"
t = 2.384, df = 499, p-value = 0.00875
alternative hypothesis: true mean is greater than 250000
```

```
95 percent confidence interval:
 260268.6      Inf
sample estimates:
mean of x
   283258
```

Since the p-value is less than 0.05, we reject the null hypothesis and conclude that the mean Sale Price is significantly greater than $250,000.

Regression Part 1

```
summary(model)
```

```
Call:
lm(formula = `Sale Price` ~ `Land Use` + `Sold As Vacant` + `Multiple Parcels
Involved in Sale` +
    Neighborhood + Acreage + `Building Value` + `Land Value` +
    `Finished Area` + Bedrooms + `Full Bath` + `Half Bath` +
    `Foundation Type` + `Year Built`, data = primary_data)

Residuals:
    Min      1Q  Median      3Q     Max
-696565  -45892   -1278   44066 3719360

Coefficients:
                                     Estimate Std. Error
(Intercept)                        -1.206e+05  8.368e+05
`Land Use`QUADPLEX                 -5.786e+05  2.237e+05
`Land Use`SINGLE FAMILY            -1.438e+05  5.027e+04
`Land Use`TRIPLEX                  -3.578e+05  2.218e+05
`Land Use`VACANT RES LAND          -2.511e+05  1.404e+05
`Land Use`VACANT RESIDENTIAL LAND  -2.708e+05  1.322e+05
`Land Use`ZERO LOT LINE            -1.553e+05  7.149e+04
`Sold As Vacant`Yes                -2.565e+05  9.863e+04
`Multiple Parcels Involved in Sale`Yes  3.714e+05  6.116e+04
Neighborhood                       -1.239e+01  5.050e+00
Acreage                             3.673e+03  2.581e+04
`Building Value`                    7.169e-01  1.223e-01
`Land Value`                        1.324e+00  1.300e-01
`Finished Area`                     1.549e+01  2.661e+01
Bedrooms                            1.498e+04  1.733e+04
`Full Bath`                        -1.644e+04  1.980e+04
`Half Bath`                        -1.273e+04  2.316e+04
`Foundation Type`FULL BSMT         -2.177e+04  2.526e+04
`Foundation Type`PIERS             -4.192e+04  2.098e+05
`Foundation Type`PT BSMT            1.328e+04  3.216e+04
`Foundation Type`SLAB              -1.859e+04  4.396e+04
`Year Built`                        1.716e+02  4.319e+02
                                   t value Pr(>|t|)
(Intercept)                         -0.144  0.88550
```

```
`Land Use`QUADPLEX                        -2.586  0.01001 *
`Land Use`SINGLE FAMILY                   -2.861  0.00441 **
`Land Use`TRIPLEX                         -1.613  0.10733
`Land Use`VACANT RES LAND                 -1.789  0.07426 .
`Land Use`VACANT RESIDENTIAL LAND         -2.049  0.04104 *
`Land Use`ZERO LOT LINE                   -2.172  0.03034 *
`Sold As Vacant`Yes                       -2.601  0.00959 **
`Multiple Parcels Involved in Sale`Yes     6.074 2.55e-09 ***
Neighborhood                              -2.453  0.01452 *
Acreage                                    0.142  0.88691
`Building Value`                           5.860 8.62e-09 ***
`Land Value`                              10.184  < 2e-16 ***
`Finished Area`                            0.582  0.56085
Bedrooms                                   0.865  0.38758
`Full Bath`                               -0.830  0.40684
`Half Bath`                               -0.550  0.58287
`Foundation Type`FULL BSMT                -0.862  0.38923
`Foundation Type`PIERS                    -0.200  0.84172
`Foundation Type`PT BSMT                   0.413  0.67991
`Foundation Type`SLAB                     -0.423  0.67265
`Year Built`                               0.397  0.69141
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 208600 on 478 degrees of freedom
Multiple R-squared:  0.5718,   Adjusted R-squared:  0.5529
F-statistic: 30.39 on 21 and 478 DF,  p-value: < 2.2e-16
```

The model explains approximately 55.3% of the variance in Sale Price. Significant positive predictors include Building Value, Land Value, and having Multiple Parcels Involved. Pro perties sold as vacant and some land use categories (e.g., Quadplex, Triplex) were associate d with significantly lower Sale Prices.

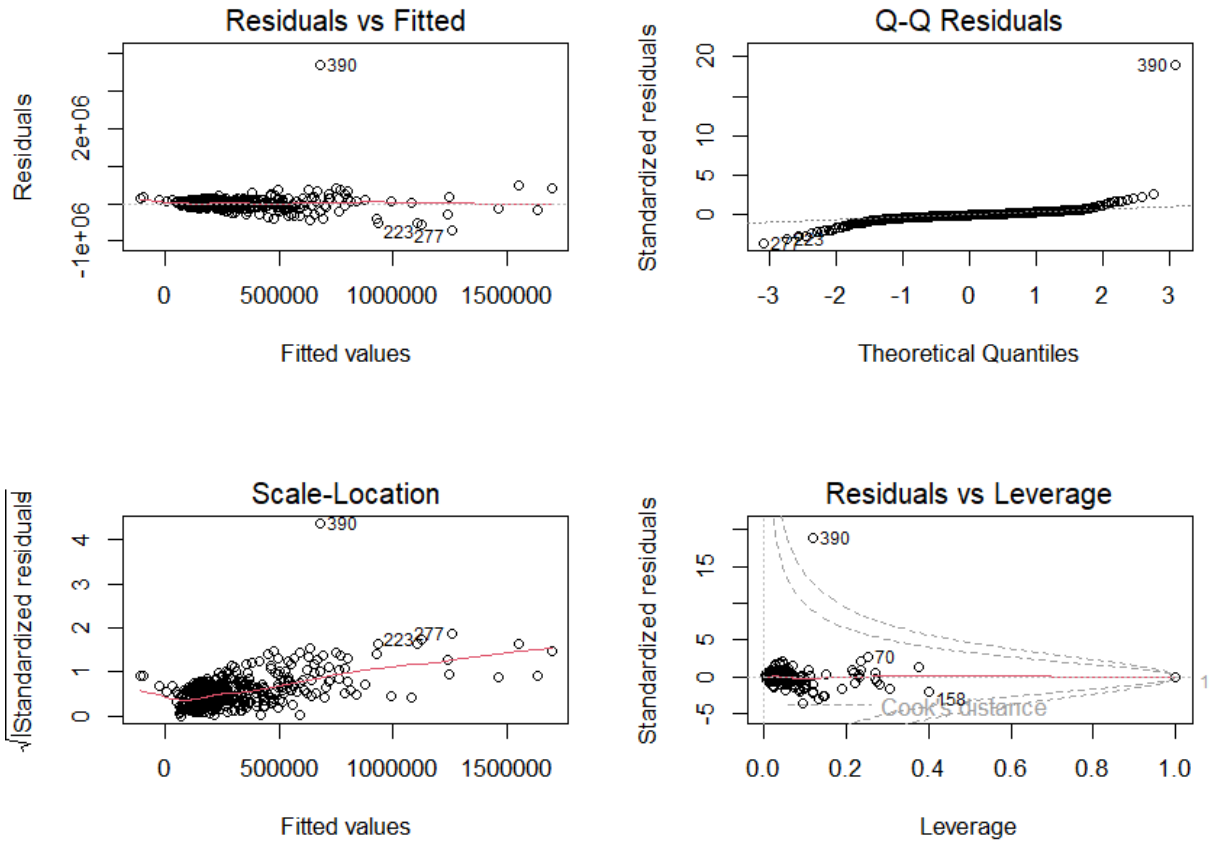
```
> confint(model)
                                              2.5 %
(Intercept)                           -1.764851e+06
`Land Use`QUADPLEX                    -1.018219e+06
`Land Use`SINGLE FAMILY               -2.425979e+05
`Land Use`TRIPLEX                     -7.936402e+05
`Land Use`VACANT RES LAND             -5.268628e+05
`Land Use`VACANT RESIDENTIAL LAND     -5.304678e+05
`Land Use`ZERO LOT LINE               -2.957480e+05
`Sold As Vacant`Yes                   -4.502911e+05
`Multiple Parcels Involved in Sale`Yes  2.512757e+05
Neighborhood                          -2.231213e+01
Acreage                               -4.704466e+04
`Building Value`                       4.764963e-01
`Land Value`                           1.068392e+00
`Finished Area`                       -3.680168e+01
Bedrooms                              -1.906139e+04
`Full Bath`                           -5.533604e+04
`Half Bath`                           -5.822392e+04
```

```
`Foundation Type`FULL BSMT             -7.139401e+04
`Foundation Type`PIERS                 -4.541472e+05
`Foundation Type`PT BSMT               -4.991460e+04
`Foundation Type`SLAB                  -1.049656e+05
`Year Built`                           -6.771382e+02
                                              97.5 %
(Intercept)                             1.523709e+06
`Land Use`QUADPLEX                     -1.389128e+05
`Land Use`SINGLE FAMILY                -4.502814e+04
`Land Use`TRIPLEX                       7.798141e+04
`Land Use`VACANT RES LAND               2.470681e+04
`Land Use`VACANT RESIDENTIAL LAND      -1.107151e+04
`Land Use`ZERO LOT LINE                -1.481045e+04
`Sold As Vacant`Yes                    -6.269272e+04
`Multiple Parcels Involved in Sale`Yes  4.916189e+05
Neighborhood                           -2.465163e+00
Acreage                                 5.438991e+04
`Building Value`                        9.572288e-01
`Land Value`                            1.579266e+00
`Finished Area`                         6.777573e+01
Bedrooms                                4.902907e+04
`Full Bath`                             2.246462e+04
`Half Bath`                             3.277285e+04
`Foundation Type`FULL BSMT              2.786184e+04
`Foundation Type`PIERS                  3.703148e+05
`Foundation Type`PT BSMT                7.646768e+04
`Foundation Type`SLAB                   6.779433e+04
`Year Built`                            1.020239e+03
> library(car)
> vif(model)
                                    GVIF Df
`Land Use`                      3.797958  6
`Sold As Vacant`                2.191386  1
`Multiple Parcels Involved in Sale` 1.250995  1
Neighborhood                    1.309770  1
Acreage                         1.705177  1
`Building Value`                4.017888  1
`Land Value`                    2.092901  1
`Finished Area`                 7.810206  1
Bedrooms                        2.198481  1
`Full Bath`                     3.842523  1
`Half Bath`                     1.397610  1
`Foundation Type`               1.330532  4
`Year Built`                    1.527455  1
                                GVIF^(1/(2*Df))
`Land Use`                             1.117624
`Sold As Vacant`                       1.480333
`Multiple Parcels Involved in Sale`    1.118479
Neighborhood                           1.144452
Acreage                                1.305824
`Building Value`                       2.004467
`Land Value`                           1.446686
`Finished Area`                        2.794675
Bedrooms                               1.482727
```

```
`Full Bath`                    1.960235
`Half Bath`                    1.182206
`Foundation Type`              1.036342
`Year Built`                   1.235902
> par(mfrow=c(2,2))
> plot(model)
```

par(mfrow=c(2,2))

plot(model)




```
> predict(model, newdata = new_data, interval = "prediction")
      fit      lwr      upr
1 4005848 -8501503 16513199
```

At the 5% significance level, the variables that stood out as having a real effect on sale price were: Land Use (specifically Quadplex, Single Family, Vacant Residential Land, and Zero Lot Line), whether the property was sold as vacant, if multiple parcels were involved, Neighborhood, Building Value, and Land Value. These factors clearly impact sale prices. For example, more valuable buildings and land raise the price, while being vacant or certain land types like a Quadplex tend to lower it.

Other factors like Triplex, Acreage, Finished Area, number of Bedrooms or Bathrooms, Foundation Type, and Year Built didn't show a strong enough statistical link to sale price in this model.

When we use the regression model to predict sale price for a specific home, we get about $4,005,848. But the prediction interval is extremely wide (from -$8.5 million to $16.5 million), meaning the model is highly uncertain, probably due to multicollinearity or outlier values. So while the estimate gives us a ballpark, we need to be cautious when using it to make real-world decisions.

Analysis of Variance

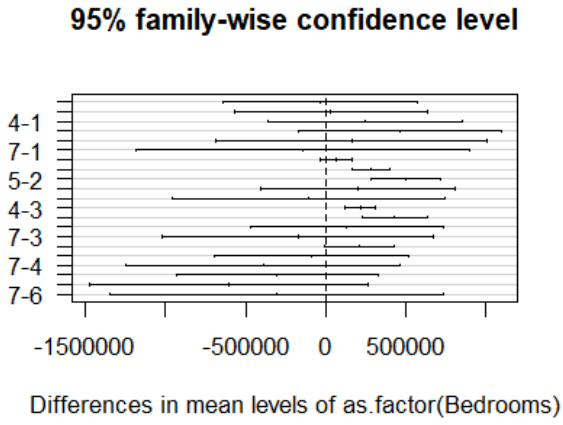
Bedrooms

```
leveneTest(`Sale Price` ~ as.factor(Bedrooms), data = primary_da
ta)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   6  7.9712 3.199e-08 ***
      493
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>
```

```
summary(anova1)
                     Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(Bedrooms)   6 7.852e+12 1.309e+12   15.85 <2e-16
Residuals           493 4.070e+13 8.257e+10

as.factor(Bedrooms) ***
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>
```

```
TukeyHSD(anova1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = `Sale Price` ~ as.factor(Bedrooms), data = primary_data)

$`as.factor(Bedrooms)`
          diff          lwr       upr     p adj
2-1  -35082.45  -642483.851  572318.9 0.9999979
3-1   32041.18  -571731.824  635814.2 0.9999987
```

```
4-1  250782.03  -356346.395  857910.5 0.8851255
5-1  466493.06  -167578.032 1100564.1 0.3093188
6-1  165100.00  -685595.632 1015795.6 0.9974867
7-1 -140400.00 -1182285.113  901485.1 0.9996858
3-2   67123.63   -31847.284  166094.5 0.4109057
4-2  285864.48   168143.092  403585.9 0.0000000
5-2  501575.51   284090.918  719060.1 0.0000000
6-2  200182.45  -407218.949  807583.9 0.9590466
7-2 -105317.55  -960173.087  749538.0 0.9998134
4-3  218740.85   121459.283  316022.4 0.0000000
5-3  434451.88   227316.838  641586.9 0.0000000
6-3  133058.82  -470714.182  736831.8 0.9949284
7-3 -172441.18 -1024722.452  679840.1 0.9968315
5-4  215711.03    -1010.008  432432.1 0.0519902
6-4  -85682.03  -692810.451  521446.4 0.9995896
7-4 -391182.03 -1245843.629  463479.6 0.8252251
6-5 -301393.06  -935464.143  332678.0 0.7978931
7-5 -606893.06 -1480899.733  267113.6 0.3809970
7-6 -305500.00 -1347385.113  736385.1 0.9770670
```

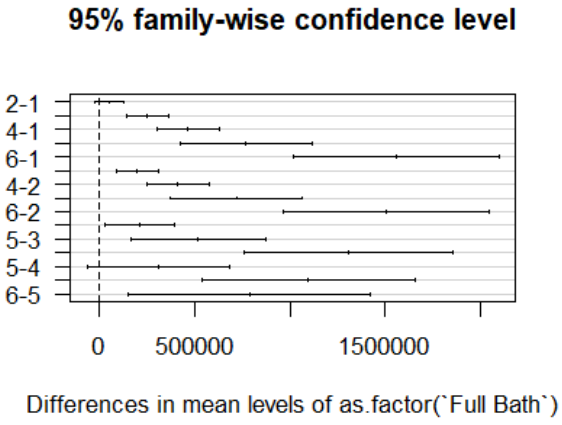


Levene's Test showed that the variation in sale prices isn't the same across bedroom groups. The ANOVA results confirmed that the number of bedrooms has a real impact on sale price. Tukey's test highlighted that homes with 4 or 5 bedrooms sold for a lot more than 2 bedroom homes, and those differences were statistically significant.

Full Bath

```
leveneTest(`Sale Price` ~ as.factor(`Full Bath`), data = primary_data)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   5  7.7741 4.689e-07 ***
      494
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova2 <- aov(`Sale Price` ~ as.factor(`Full Bath`), data = primary_data)
```

```
> summary(anova2)
                       Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(`Full Bath`)   5 1.355e+13 2.710e+12   38.24 <2e-16 ***
Residuals              494 3.501e+13 7.086e+10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(anova2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = `Sale Price` ~ as.factor(`Full Bath`), data = primary_data
)

$`as.factor(`Full Bath`)`
          diff        lwr       upr     p adj
2-1   52529.83  -23398.37  128458.0 0.3557060
3-1  251308.42  142781.26  359835.6 0.0000000
4-1  463265.92  298470.04  628061.8 0.0000000
5-1  769942.58  424985.36 1114899.8 0.0000000
6-1 1558542.58 1017244.18 2099841.0 0.0000000
3-2  198778.59   91244.59  306312.6 0.0000028
4-2  410736.09  246592.56  574879.6 0.0000000
5-2  717412.76  372766.70 1062058.8 0.0000001
6-2 1506012.76  964912.60 2047112.9 0.0000000
4-3  211957.50   30419.21  393495.8 0.0115462
5-3  518634.17  165372.29  871896.0 0.0004518
6-3 1307234.17  760605.93 1853862.4 0.0000000
5-4  306676.67  -67720.84  681074.2 0.1788228
6-4 1095276.67  534757.29 1655796.0 0.0000006
6-5  788600.00  151403.19 1425796.8 0.0058038
```

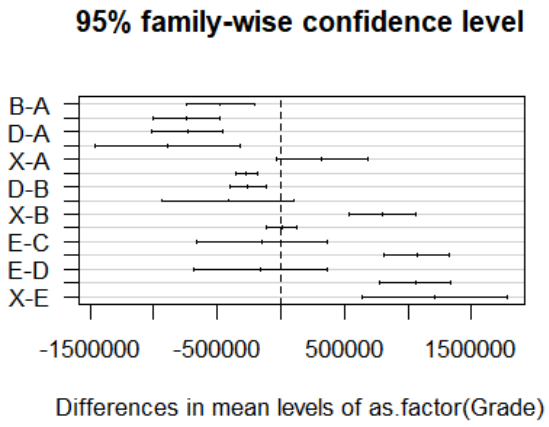


Levene’s Test showed that the spread of sale prices is not consistent across the different full bath groups. The ANOVA confirmed that the number of full baths makes a big difference in sale price. Tukey’s test showed that houses with 3 or more full baths usually sell for a lot

more, and homes with 6 full baths had the highest prices, standing out the most from all the other groups.

Grade

```
leveneTest(`Sale Price` ~ as.factor(Grade), data = primary_data)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   5  5.9136 2.534e-05 ***
      494
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova3 <- aov(`Sale Price` ~ as.factor(Grade), data = primary_data)
> summary(anova3)
                  Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(Grade)   5 1.691e+13 3.382e+12   52.79 <2e-16 ***
Residuals        494 3.165e+13 6.406e+10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(anova3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = `Sale Price` ~ as.factor(Grade), data = primary_data)

$`as.factor(Grade)`
          diff         lwr       upr     p adj
B-A -475638.31  -743431.25 -207845.4 0.0000079
C-A -743836.77 -1002705.41 -484968.1 0.0000000
D-A -733341.25 -1013792.74 -452889.8 0.0000000
E-A -888250.00 -1460719.20 -315780.8 0.0001616
X-A  324138.75   -37922.56  686200.1 0.1089564
C-B -268198.46  -355592.07 -180804.9 0.0000000
D-B -257702.94  -396547.12 -118858.8 0.0000025
E-B -412611.69  -930632.60  105409.2 0.2047074
X-B  799777.06   531984.11 1067570.0 0.0000000
D-C   10495.52  -110242.29  131233.3 0.9998703
E-C -144413.23  -657877.51  369051.0 0.9666251
X-C 1067975.52   809106.87 1326844.2 0.0000000
E-D -154908.75  -679585.44  369767.9 0.9589271
X-D 1057480.00   777028.51 1337931.5 0.0000000
X-E 1212388.75   639919.55 1784858.0 0.0000000
```



Levene's Test showed that the variation in sale prices wasn't equal between the different home grades. The ANOVA results confirmed that there's a real difference in average sale prices depending on the grade of the house. Tukey's test showed that homes with a grade of "X" sold for a lot more than the others, while grades C and D had similar prices and were both lower than grades B and A.

Conclusion

After going through the analysis, factors such as building value, land value, and how the property is used are shown to have a big impact on sale price. Some other features, like the number of bedrooms or bathrooms, were less consistent. The regression model helped estimate sale prices, but the wide prediction range showed that results should be used carefully. Overall, this assignment helped in understanding how to apply statistical tools to real housing data and revealed how important it is to clean and explore the data before drawing conclusions.