Intro:

We are using real-estate data to analyze the relationship between housing prices and various features of houses including: the year the houses were built, their square footage, and the number of bathrooms,etc.. We analyze the data visually by using histograms and boxplots to estimate the distributions of different variables, and by using linear regression to assess the relationship between numerical variables.

Part 1:

There are 13 missing entries. The columns of the missing values are “YearBuilt” “SqFt” “Story” “Acres” “N_Baths” “Fireplace” “LandPrice” “BuildingPrice” “Zipcode”. Four rows(1,14,51,and 81) are taken out of the dataset due to the missing

```r
## detect missing values
# sum all na values
sum(is.na(RealEstate))
# sum by columns
has_na <- sapply(RealEstate, function(x) any(is.na(x)))
# make and print info of columns with missing info
columns_na <- names(RealEstate)[has_na]
print(columns_na)
# find and print ids
rows_na <- which(rowSums(is.na(RealEstate)) > 0)
print(rows_na)

## remove and make new data set
no_na <- RealEstate[rowSums(is.na(RealEstate)) == 0, ]
```

```
[1] 13
[1] "YearBuilt"     "SqFt"          "Story"         "Acres"         "N_Baths"       "Fireplace"     "LandPrice"     "BuildingPrice"
[9] "Zipcode"
[1]  1 14 51 80
```

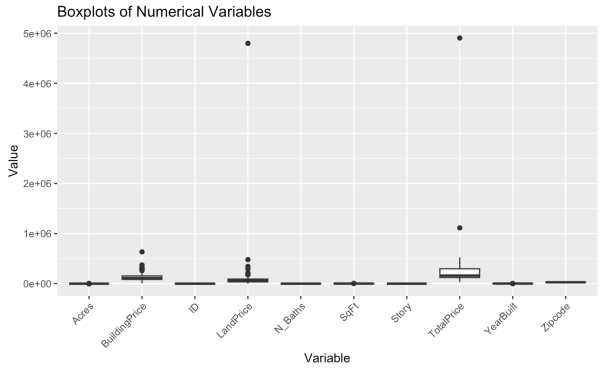
Part 2:

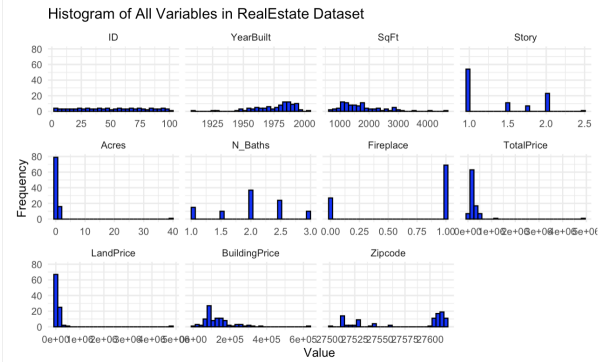
I created a list of boxplots to see all of the distributions of all the variables. Based on the boxplots, I notice that there is one extreme each in TotalPrice and LandPrice, so I decided to remove the extreme values and decided that they were the two variables I wanted to interpret. I used code to take out the outliers-- if the values are bigger than the Q3 + 1.5 * IQR or smaller than Q1 - 1.5 * IQR, they are outliers and were promptly removed. Total Price and Land Price became my two variables of interests. To summarize them I used boxplots, histograms, and the 5 number summary. I then redid the summary without the outliers to get more of an insight of the data's spread.

Step 0: The first thing I did was plot and summarize all the possible variables to see which ones interested me.

```r
#install.packages("ggplot2")
library(ggplot2)
#install.packages("tidyr")
library(tidyr)
#install.packages(reshape2)
library("reshape2")

column_classes <- sapply(no_na, class)
print(column_classes)
# Transforming data to long format
re_long <- melt(no_na, id.vars = NULL)
# Creating  plot
ggplot(re_long, aes(x = value)) +
geom_histogram(bins = 30, fill = "blue", color = "black") +
facet_wrap(~ variable, scales = "free_x") +
theme_minimal() +
labs(title = "Histogram of All Variables in RealEstate Dataset", x = "Value", y = "Frequency")

numeric_data <- no_na[, sapply(no_na, is.numeric)]
long_data <- pivot_longer(numeric_data, cols = everything(), names_to = "Variable", values_to = "Value")
ggplot(long_data, aes(x = Variable, y = Value)) +
geom_boxplot() +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) + # Rotates x-axis labels for better readability
labs(title = "Boxplots of Numerical Variables", x = "Variable", y = "Value")
```

```
          ID     YearBuilt          SqFt         Story         Acres       N_Baths     Fireplace
TotalPrice     LandPrice BuildingPrice
    "integer"     "integer"     "integer"     "numeric"     "numeric"     "numeric"     "logical"
"integer"     "integer"     "integer"
       Zipcode
     "integer"
```



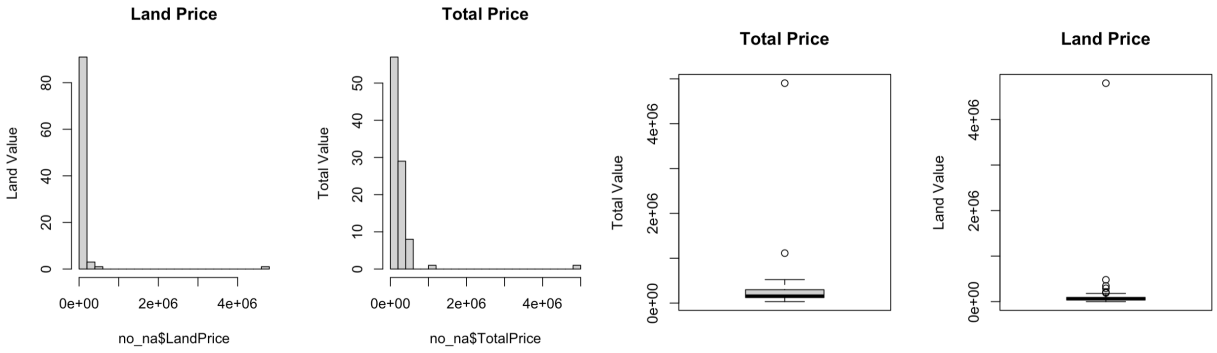



```{r}
#install.packages("ggplot2")
library(ggplot2)
#install.packages("tidyr")
library(tidyr)
#install.packages(reshape2)
library("reshape2")

par(mfrow = c(1,2))
boxplot(no_na$TotalPrice, main= "Total Price", ylab = "Total Value")
boxplot(no_na$LandPrice, main= "Land Price", ylab = "Land Value")
hist(no_na$LandPrice, breaks= 20, main= "Land Price", ylab = "Land Value")
hist(no_na$TotalPrice, breaks= 20, main= "Total Price", ylab = "Total Value")
summary(no_na$TotalPrice)
summary(no_na$LandPrice)

```

Step 1: variables at interest: land price and total price

Step 2: summarize each variable


```{r}
#install.packages("ggplot2")
library(ggplot2)
#install.packages("tidyr")
library(tidyr)
#install.packages(reshape2)
library("reshape2")

par(mfrow = c(1,2))
boxplot(no_na$TotalPrice, main= "Total Price", ylab = "Total Value")
boxplot(no_na$LandPrice, main= "Land Price", ylab = "Land Value")
hist(no_na$LandPrice, breaks= 20, main= "Land Price", ylab = "Land Value")
hist(no_na$TotalPrice, breaks= 20, main= "Total Price", ylab = "Total Value")
summary(no_na$TotalPrice)
summary(no_na$LandPrice)

```



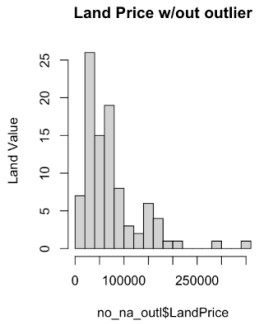

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32184  122680  161998  261119  296585 4904102
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   35500   60000  127552   91500 4797750
```

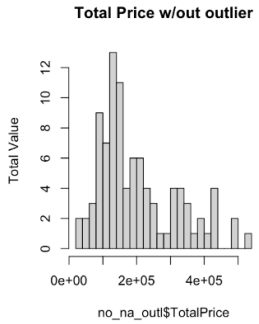
Step 3: identify unusual data(+taking out the outliers): outliers in the max of total price and land price. Identified when we used the summary function and the plots from above. Also identified through our code from when we printed out the outlier when we took it out from the code below. Outliers were– land price: 4797750, total price: 4904102

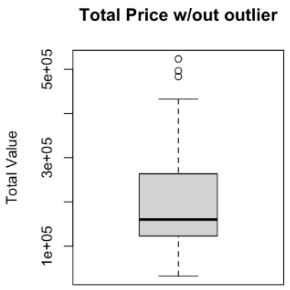
```r
# take out outliers
num_vars <- c("TotalPrice", "LandPrice", "BuildingPrice", "SqFt") # List of numeric variables you're interested in
remove_iqr_outliers <- FALSE # Set this to FALSE if you do not want to delete IQR-based outliers
rows_to_remove <- integer(0) # Initialize an empty vector to store rows to delete
# Loop through each numeric variable to identify outliers
for (var in num_vars) {
# row with the largest value for the current variable
row_with_largest_value <- which.max(no_na[[var]])
rows_to_remove <- unique(c(rows_to_remove, row_with_largest_value))
if (remove_iqr_outliers) {
# Calculate quartiles and IQR for outlier detection
Q1 <- quantile(no_na[[var]], 0.25, na.rm = TRUE)
Q3 <- quantile(no_na[[var]], 0.75, na.rm = TRUE)
IQR <- Q3 - Q1
lb <- Q1 - 1.5 * IQR
ub <- Q3 + 1.5 * IQR
# rows that are outliers for the variable
outlier_rows <- which(no_na[[var]] < lb | no_na[[var]] > ub)
rows_to_remove <- unique(c(rows_to_remove, outlier_rows))
}
}

# print IDs of outliers
if("ID" %in% names(no_na)) {
print(no_na$ID[rows_to_remove])
}

no_na_outl <- no_na[-rows_to_remove, ]

par(mfrow = c(1,2))
hist(no_na_outl$LandPrice, breaks= 20, main= "Land Price w/out outlier", ylab = "Land Value")
hist(no_na_outl$TotalPrice, breaks= 20, main= "Total Price w/out outlier", ylab = "Total Value")
boxplot(no_na_outl$TotalPrice, main= "Total Price w/out outlier", ylab = "Total Value")
boxplot(no_na_outl$LandPrice, main= "Land Price w/out outlier", ylab = "Land Value")
summary(no_na_outl$TotalPrice)
summary(no_na_outl$LandPrice)
```

Step 4: insight description of data:

After accounting for the outliers by taking them out of the data set in the previous step, I was able to better examine the data. Both graphs look like they have a positive skew. Some notable information for the total price is the 5 number summary which consists of the: min: 32184, q1:122680, median:161998, mean:261119, q3:3296585, and max:4904102. Some of the land price's notable information after the outliers were taken out are: min:0, q1:35500, median:60000, mean:127552, q3:91500, max:4797750. The previous 5 number summary(before accounting for outliers) is in the summary for part 2.





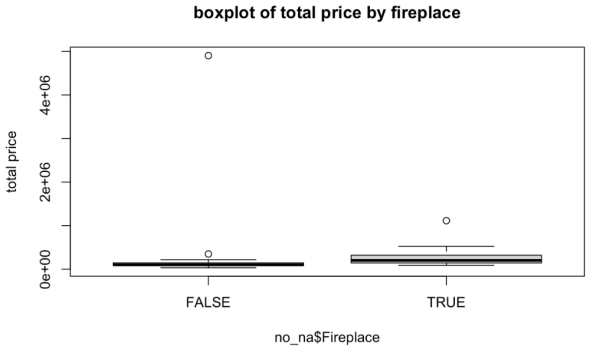


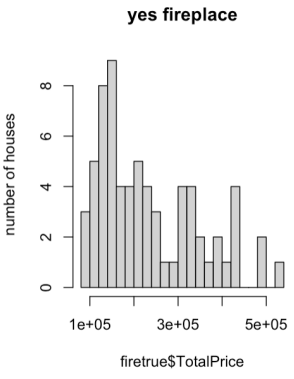
Part 3:

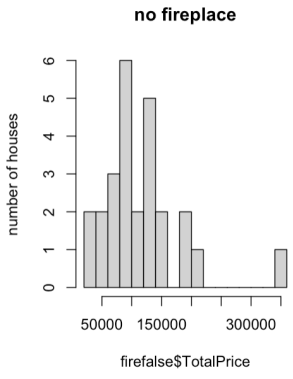
Using the visual forms of interpreting data and numerical terms, I came to the conclusion that a fireplace does play a role in the price of real estate. The five number summary shows that houses with a fireplace cost more than ones without in every category: mean, 25th, 50th, and 75th percentile, and max. In terms of the visual data-- where we used boxplots and histograms, we see that houses with fireplaces cost more than ones without confirming our numerical data.

Step 1-3: Investigate whether the presence of a fireplace is related to property price/Utilize numerical and graphical tools to compare property prices. (I am using Total Price)

```
library(mosaic)
#histogram(~no_na$TotalPrice | no_na$Fireplace, data = no_na, breaks= 20, layout=c(1,2))

firetrue <- no_na_outl[no_na_outl$Fireplace == TRUE,]
firefalse <-  no_na_outl[no_na_outl$Fireplace == FALSE,]

summary(firetrue$TotalPrice)
summary(firefalse$TotalPrice)
#par(mfrow = c(1,2))
#boxplot(firetrue$TotalPrice, main = "fireplace true total price")
#boxplot(firefalse$TotalPrice, main = "fireplace false total price")
boxplot(no_na$TotalPrice~no_na$Fireplace, main= "boxplot of total price by fireplace", ylab="total price")
par(mfrow = c(1,2))
hist(firetrue$TotalPrice, breaks= 20, main= "yes fireplace", ylab = "number of houses")
hist(firefalse$TotalPrice, breaks= 20, main= "no fireplace", ylab = "number of houses")
```








```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  89871  142024  201560  235048  318875  523366
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32184   78560   99082  117933  135297  349916
```

Step 4: Provide insights regarding any observed differences in property prices.
There is an observed difference in the prices of property based on whether or not a fireplace was present in the home. If we look at both the boxplot we see that places without a fireplace have lower minimums and maximums than ones with a fireplace corroborating with our findings from the 5 number summary. In addition, looking at the histogram, you can tell based on the incrementation of the histograms that houses with a fireplace generally cost more. First, the 5 number summary(fireplace top summary, and non fireplace bottom summary) shows us that in every category(after accounting for outliers) houses with a fireplace cost more.

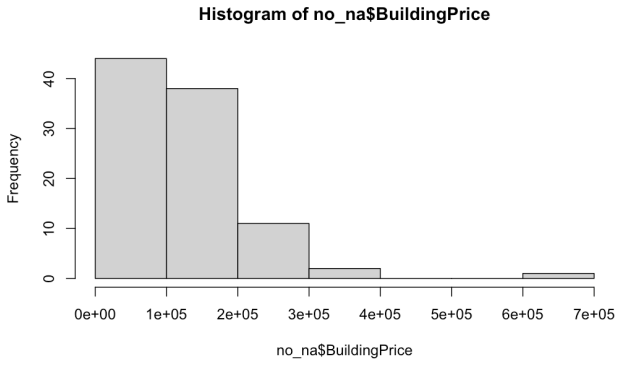
Part 4:

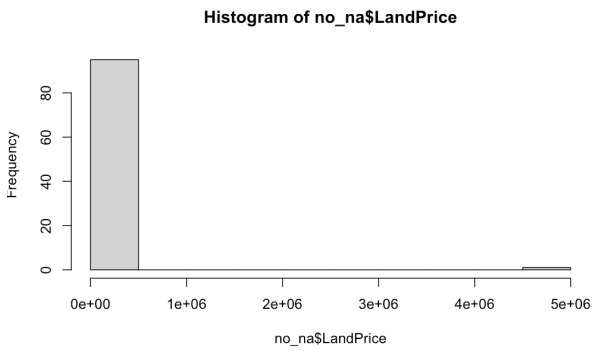
For my two variables I choose building price and sqft which have a correlation of .8839685. The graphs show a similar shape and distribution corroborating to me that they are similar. I first had to identify what my continuous variables were. To do so, I looked at numerical variables first. I wanted to find variables that were not only numerical but also decimals. Beyond that, I was looking for numbers that were also random in decimal. For example, the number of baths was not continuous because while it could have 1.5 baths, it could have 1.49 baths. This made it not continuous. Then I graphed all of my continuous variables on a histogram and choose price and square feet as my variables due to their seeming correlation.

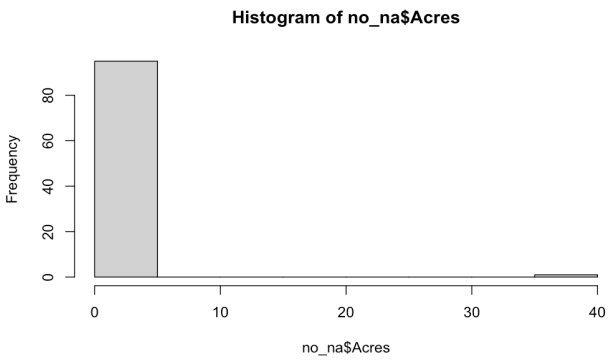
Step 1:
continuous variables: ID, YearBuilt, SqFt, Acres, TotalPrice, LandPrice, BuildingPrice
*look at code below to see what i did to rule out what was continuous

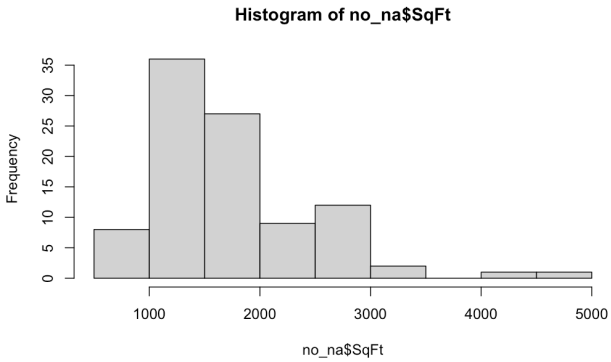
Step 2:
Explore potential relationships among different continuous variables.

```
sapply(no_na, is.numeric)
print(column_classes)

continuous_vars = c("SqFt", "Acres", "LandPrice", "BuildingPrice", "TotalPrice")
hist(no_na$SqFt)
hist(no_na$Acres)
hist(no_na$LandPrice)
hist(no_na$BuildingPrice)
```










```
          ID     YearBuilt          SqFt         Story         Acres       N_Baths     Fireplace
TotalPrice     LandPrice BuildingPrice       Zipcode
         TRUE          TRUE          TRUE          TRUE          TRUE          TRUE         FALSE
TRUE          TRUE          TRUE          TRUE
          ID     YearBuilt          SqFt         Story         Acres       N_Baths     Fireplace
TotalPrice     LandPrice BuildingPrice       Zipcode
    "integer"     "integer"     "integer"     "numeric"     "numeric"     "logical"
"integer"     "integer"     "integer"     "integer"
```

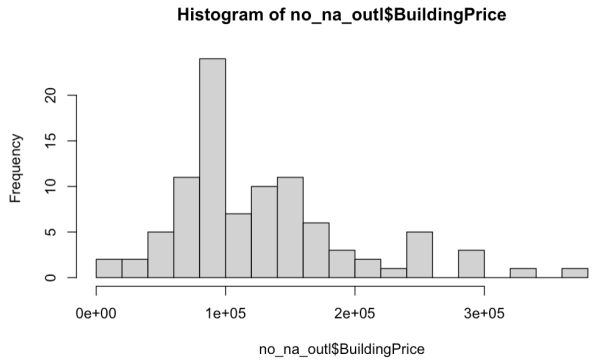
Step 3: Identify at least one pair of continuous variables:

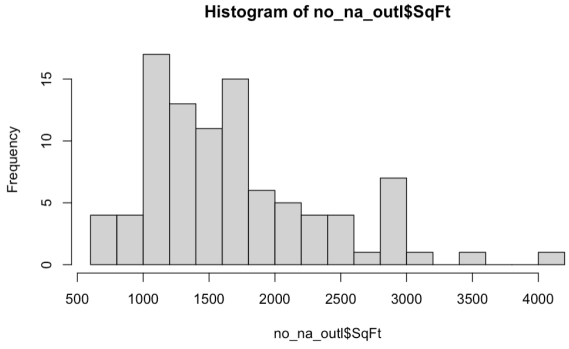
I am choosing square feet and building price. I have a feeling that since land is expensive, generally, the larger the square feet is, the more the building would cost.

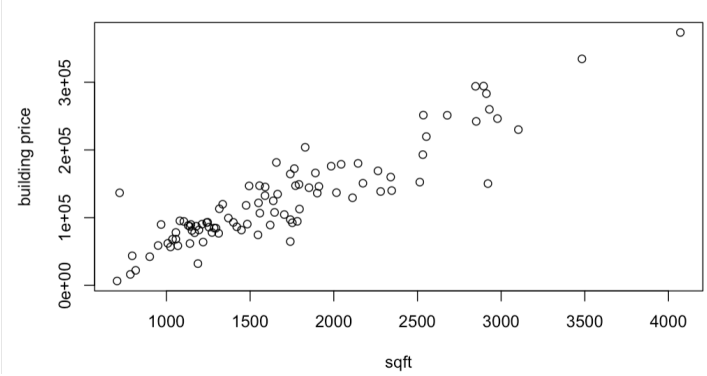
```r
cor_matrix <- cor(no_na[,continuous_vars], use = "complete.obs")
print(cor_matrix)

# choose 2 variables: sqft and building price
plot(x=no_na_outl$SqFt, y=no_na_outl$BuildingPrice, xlab = "sqft", ylab = "building price")
hist(no_na_outl$SqFt, breaks = 20)
hist(no_na_outl$BuildingPrice, breaks = 20)
```

Step 4: Utilize appropriate graphical tools to visualize and understand the relationship between the variables of your choice.

```
                    ID  YearBuilt        SqFt       Acres   LandPrice BuildingPrice TotalPrice
ID           1.00000000 0.06685582 -0.01672184 -0.15734003 -0.15828515    0.01896842 -0.1505548
YearBuilt    0.06685582 1.00000000  0.30020043  0.17596497  0.14481484    0.28524515  0.1901351
SqFt        -0.01672184 0.30020043  1.00000000 -0.02391775  0.04440322    0.90864763  0.2003797
Acres       -0.15734003 0.17596497 -0.02391775  1.00000000  0.98805174   -0.01678308  0.9574992
LandPrice   -0.15828515 0.14481484  0.04440322  0.98805174  1.00000000    0.07503101  0.9850017
BuildingPrice  0.01896842 0.28524515  0.90864763 -0.01678308  0.07503101    1.00000000  0.2459641
TotalPrice  -0.15055476 0.19013509  0.20037973  0.95749918  0.98500166    0.24596406  1.0000000
```





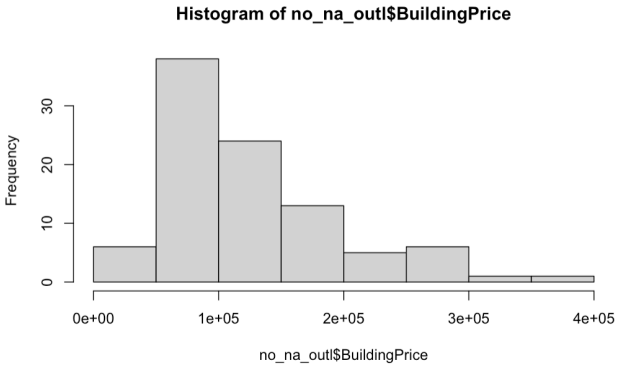


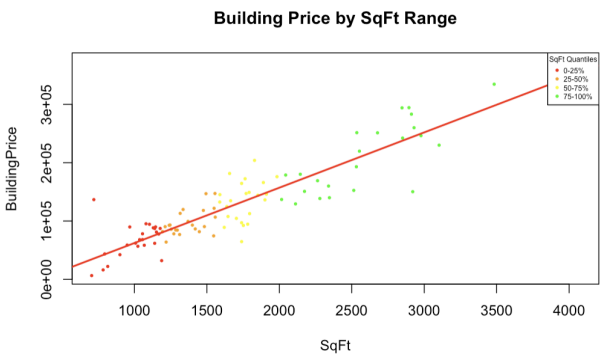
Step 5:Identify, interpret, and address any outliers that may be present:
Already accounted for since I am using my data set which has no outliers

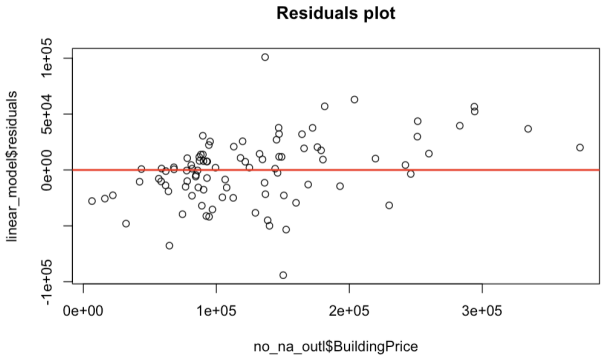
Part 5:

Step 1: Conduct a linear regression analysis on two selected variables utilizing insights from the previous step:

```
# chose SqFt/BuildingPrice
hist(no_na_outl$BuildingPrice)
hist(no_na_outl$SqFt)

# quantiles for SqFt
sqft_quantiles <- quantile(no_na_outl$SqFt, probs = c(0, 0.25, 0.5, 0.75, 1))
# Cut SqFt into ranges based on quantiles
sqft_levels <- cut(no_na_outl$SqFt, breaks = sqft_quantiles, include.lowest = TRUE)
sqft_colors <- c("red", "orange","yellow", "green")
sqft_styles <- c(10, 4, 1, 2) # Example styles: circle, triangle, etc.
# Plot BuildingPrice by SqFt with color-coded ranges
plot(no_na_outl$SqFt, no_na_outl$BuildingPrice, xlab = "SqFt", ylab = "BuildingPrice",
main = "Building Price by SqFt Range", pch = 16, cex = 0.5,
col = sqft_colors[as.numeric(sqft_levels)])
linear_model <- lm(no_na_outl$BuildingPrice~no_na_outl$SqFt)
abline(linear_model, col = "red", lwd = 2)
# Add legend
legend_labels <- c("0-25%", "25-50%", "50-75%", "75-100%")
legend("topright", legend = legend_labels, col = sqft_colors, pch = 16, cex = 0.5, title = "SqFt Quantiles")

plot(linear_model$residuals ~ no_na_outl$BuildingPrice, main = "Residuals plot")
abline(a = 0, b = 0, col = "red", lwd = 2)

summary(linear_model)
```








```
Call:
lm(formula = no_na_outl$BuildingPrice ~ no_na_outl$SqFt)

Residuals:
   Min     1Q Median     3Q    Max
-94138 -17173   1159  14563 100946

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -32519.308   8436.979  -3.854 0.000215 ***
no_na_outl$SqFt     94.817      4.622  20.515  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29950 on 92 degrees of freedom
Multiple R-squared:  0.8206,    Adjusted R-squared:  0.8187
F-statistic: 420.9 on 1 and 92 DF,  p-value: < 2.2e-16
```

Step 2:

I conducted a linear regression analysis on BuildingPrice based on the SqFt. The regression model is $BuildingPrice= 94.540 * SqFt -34119.116. For every unit of increasing SqFt, the building costs 94.540 more. The intercept is negative, so that means this model is only valid after some values of SqFt.

Step 3:

The residuals are symmetric and of equal variance, also based on the scatter plot, the model is linear, so all assumptions are met. Furthermore, the R-squared value after accounting for outliers is 0.8298, so the model is a good fit.

Conclusion:

First we cleaned and evaluated our data, plotting and summarizing them. We also removed outliers. We found out that fireplaces are correlated with more expensive building prices from part 4. In addition, we found out that building prices are strongly correlated with the amount of square feet a building has via part 5. Some other notable things which we did during the project was identifying which variables were continuous and making a linear regression model + interpreting the estimated coefficients. Finally, we evaluated the goodness of fit with our relevant metrics.