**Final Project**

Edmund Leong

206049891

STATS 10

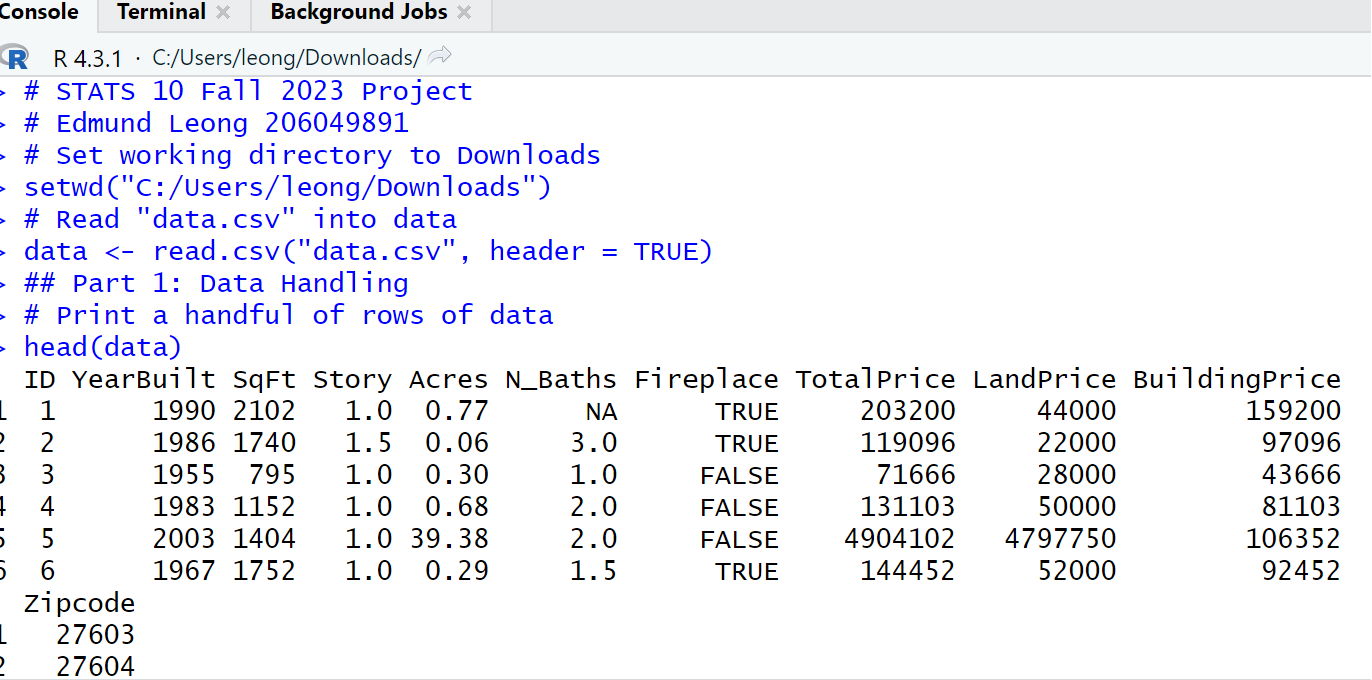
Introduction to Statistical Reasoning

**Introduction**

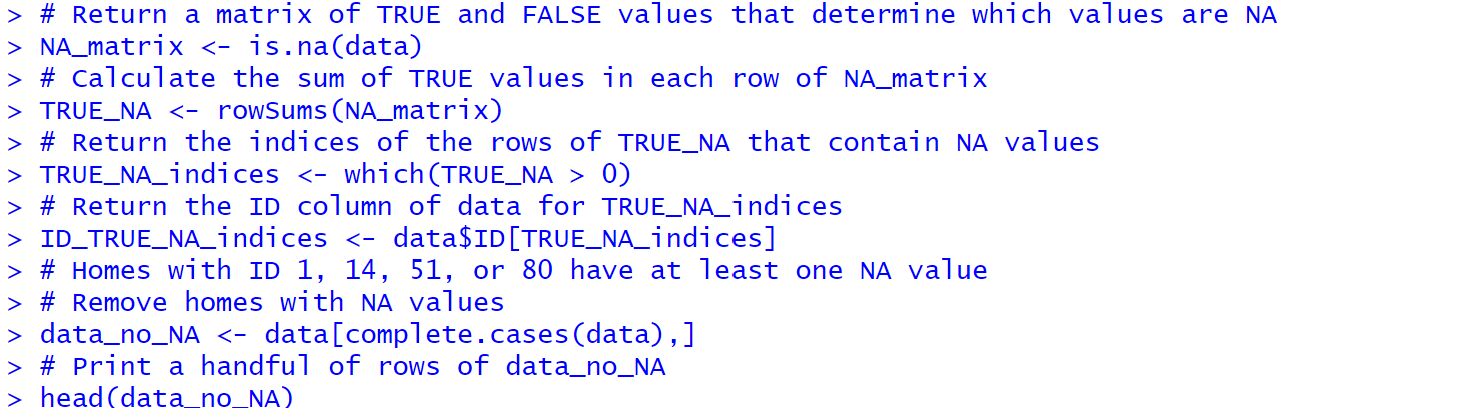
In this paper, we’ll examine a dataset containing several real estate listings and many of their characteristics, which we can categorize as either discrete or continuous variables. We consider ID, year built, story count, number of baths, and the presence of a fireplace to be discrete variables. Hence, square footage, acres occupied, total price, land price, and building price are considered to be continuous variables. Constructing boxplots and histograms allows us to visualize the distribution of each continuous variable and identify each of their potential outlier. We also calculate the correlation coefficient (*r*2) among several pairs of variables to identify any strong, linear relationship. Upon finding a high correlation coefficient, we then perform linear regression analysis to impose the best-fit regression line upon the scatterplot of both variables, including the regression line equation. Throughout this report, we explain the meaning of each block of code and visualization produced by our R code to cater toward readers who may lack experience in the R programming language.

**Detecting and Handling Missing Data**

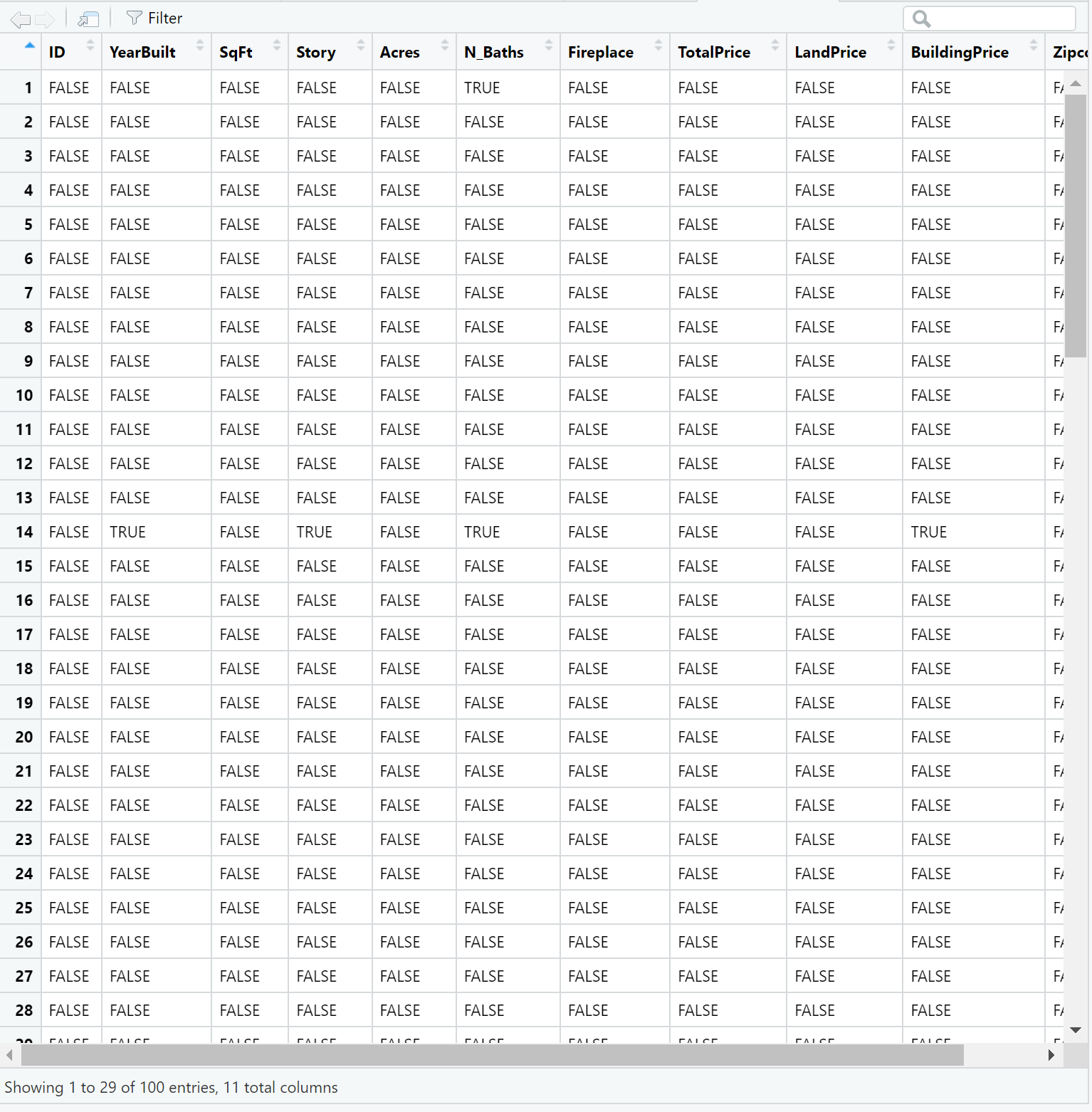
After loading the “data.csv” file into R as the object data, we first checked the observations in the dataset by performing head(data):



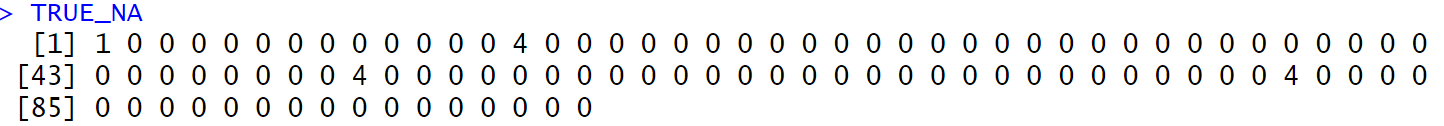
Notice the NA value in the N\_Baths column of the dataset. Since the head() function produces only a handful of observations, it is impossible to tell alone from this display how many NA values the dataset contains. To do this requires the use of logical operators:



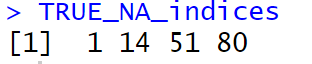
First, the is.na() function returns TRUE for values that are NA and FALSE for values that are not NA. Combining this function with an assignment operator returns a matrix (another way of saying “dataset”) for all of these TRUE and FALSE values:



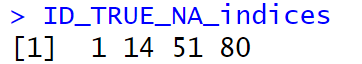
Thus, this matrix, NA\_matrix, becomes the foundation for performing additional statistical techniques, such as calculating the sum of NA values in the original dataset, which can be done by using the rowSums() function:



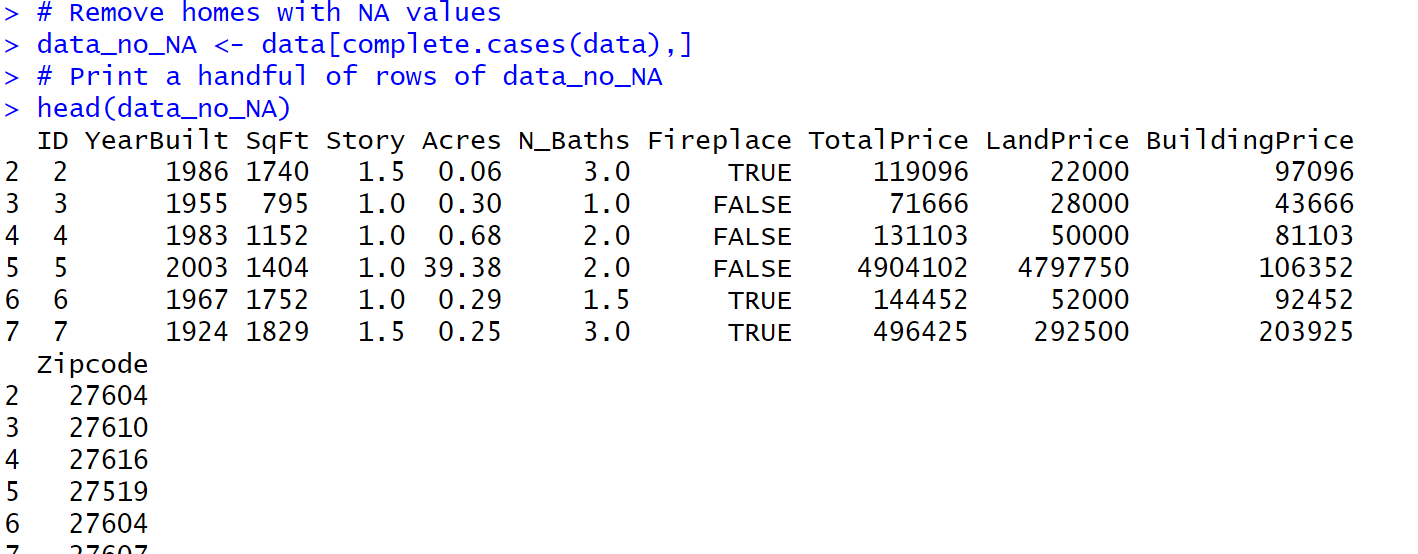
Encapsulating this function with the which() function returns all indices that contain one or more NA values:



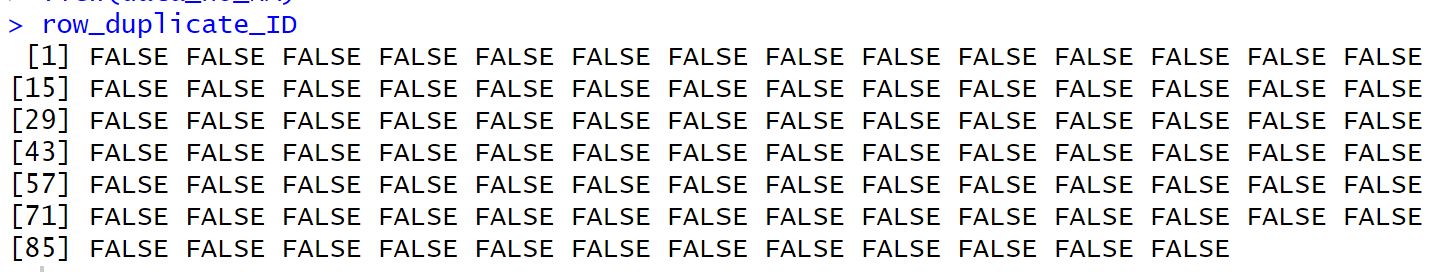
We’re not done identifying which homes contain NA values, however, because the ID column, not the actual index stored by R, is how we can identify homes in the dataset. Thus, we must subset data$ID with the indices we computed earlier to associate the right ID values with the indices corresponding to observations containing NA values:



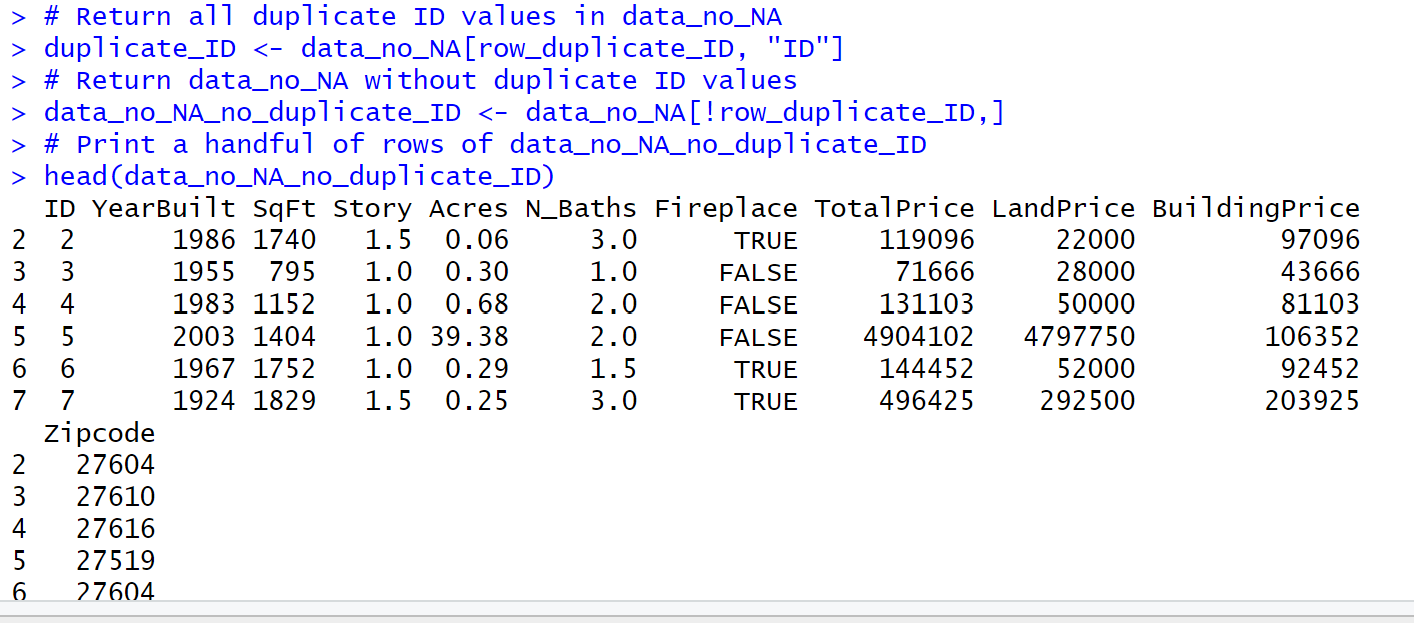
This step may appear redundant, because TRUE\_NA\_indices and ID\_TRUE\_NA\_indices are identical. However, subsetting data$ID with TRUE\_NA\_indices checks whether home ID values correspond to their predetermined indices by R. In this dataset, they are the same. After identifying that the first, fourteenth, fifty-first, and eigthieth homes contain NA values, the next step is to remove them from the dataset with the complete.cases() function, which removes either all NA values or all NA values both per row and by column from a dataset:



After removing all NA values from the dataset, we must check for accidental, repeated observations with the duplicated() function:



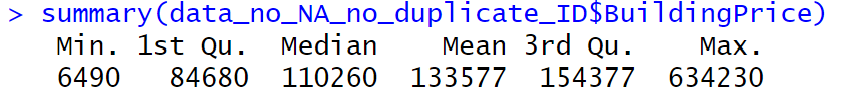
It appears from the output of row\_duplicate\_ID that no duplicates exist. Nevertheless, we perform the rest of the data cleaning process to conclude the dataset contains no duplicates:



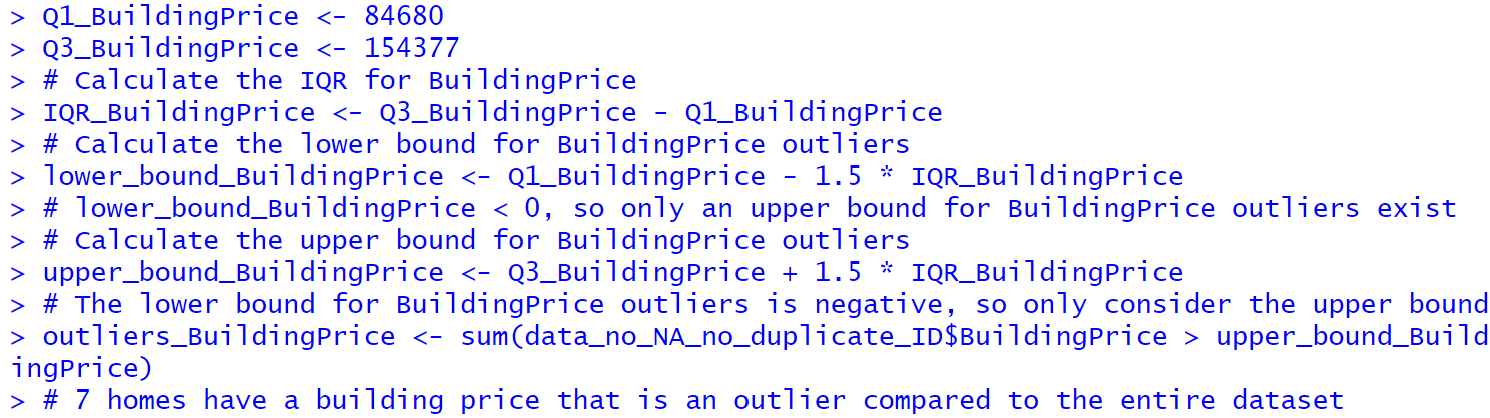
Comparing head(data\_no\_NA) and head(data\_no\_NA\_no\_duplicate\_ID) shows that both outputs are identical. For the purposes of formalizing the data cleaning process, in this project we use data\_no\_NA\_no\_duplicate\_ID as our dataset upon which we perform several statistical analyses.

**Variable Summarization**

The first variable we analyze is BuildingPrice, which by its name contains the building prices for all houses in the cleaned dataset data\_no\_NA\_no\_duplicate\_ID. Beginning with a five-number summary for BuildingPrice is optimal, because the summary() function returns both the mean and median building prices and the minimum, first quartile (Q1), third quartile (Q3), and maximum values for building prices in the dataset. Since we cannot determine whether to evaluate the distribution of BuildingPrice values by either the median and standard deviation or the first and third quartiles with their IQR, having all of these values is helpful to decide how to go about our analysis. Already, by simple observation, the maximum value suggests that at least one outlier exists in the distribution of data\_no\_NA\_no\_duplicate\_ID$BuildingPrice:



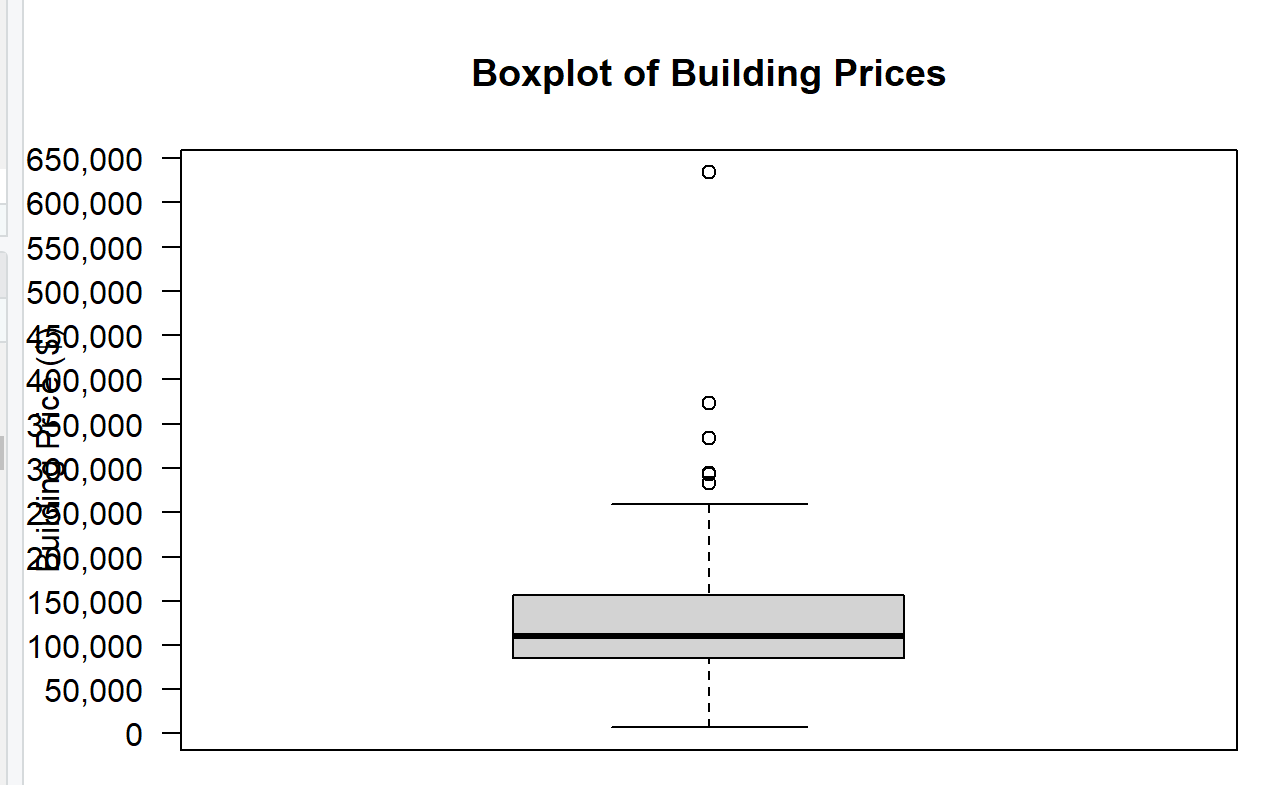
Calculating the IQR and summing up all observations that are outliers for BuildingPrice confirms this:

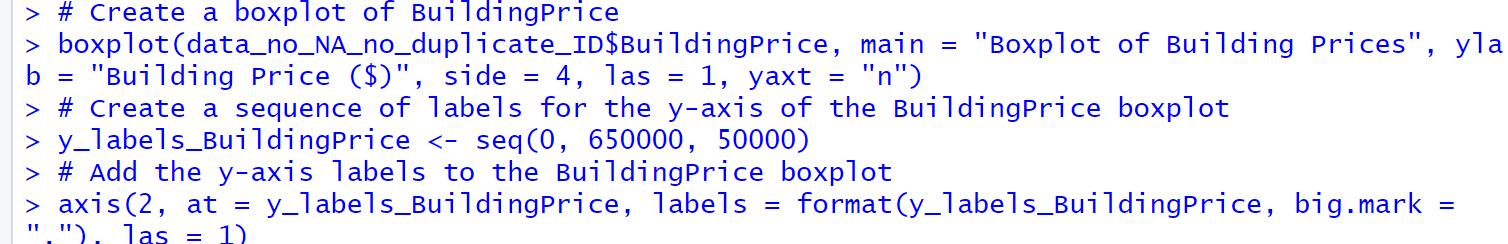


This is the value for IQR\_BuildingPrice:

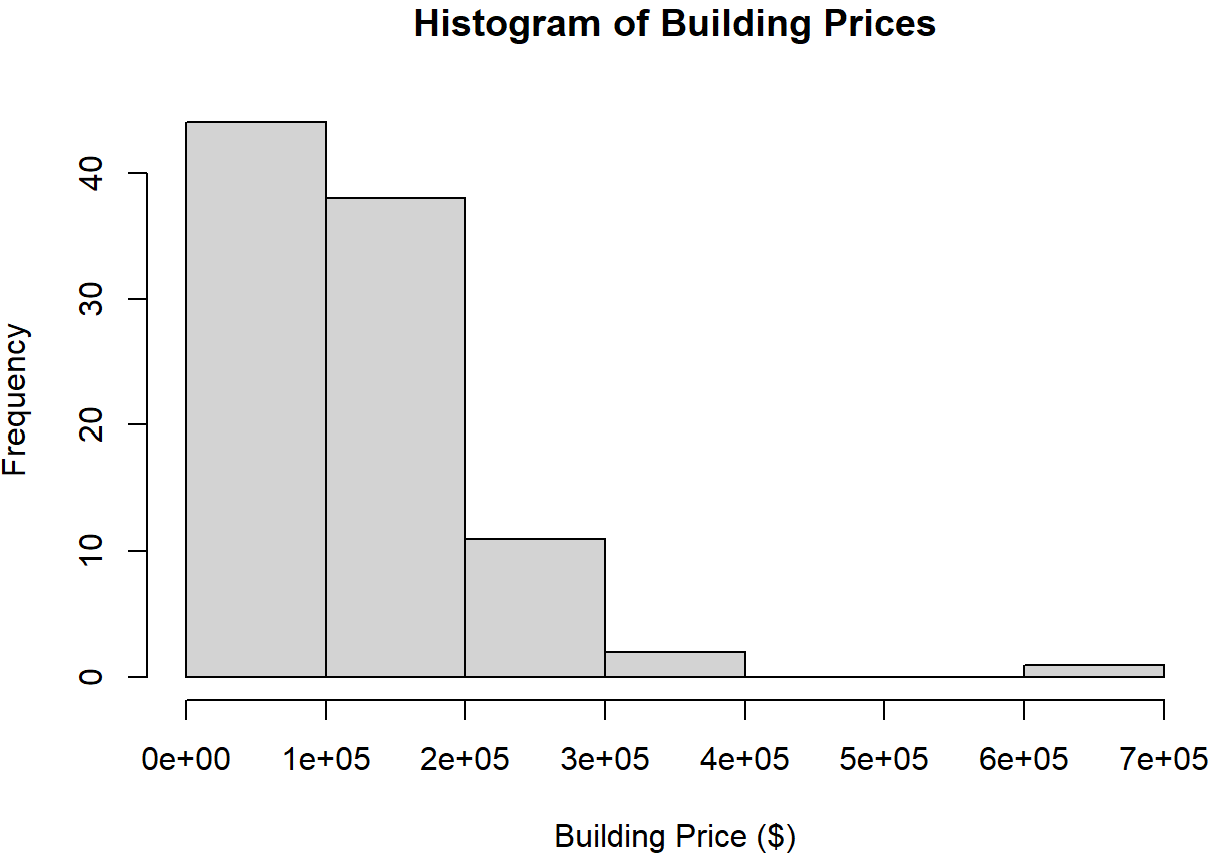


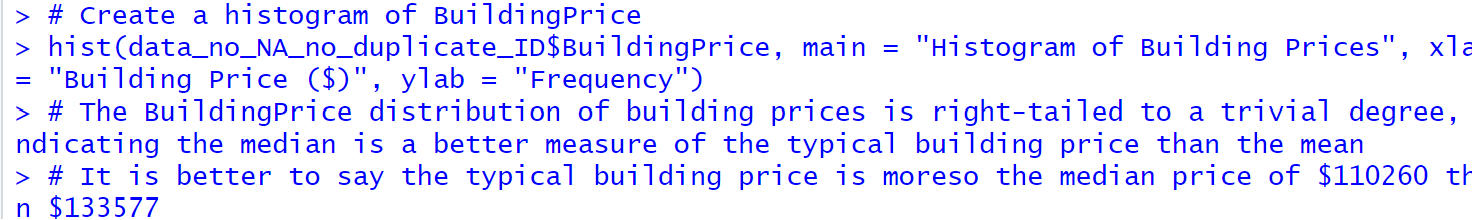
A visualization that magnifies the importance of these outliers is a boxplot, presented below along with its code:





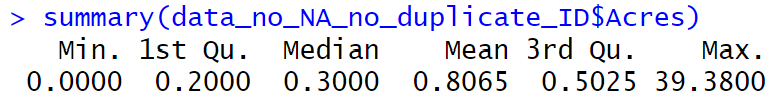
The presence of 5 outliers on a boxplot shows the distribution of building prices is right-skewed. A better way of displaying this skewness is by a histogram:



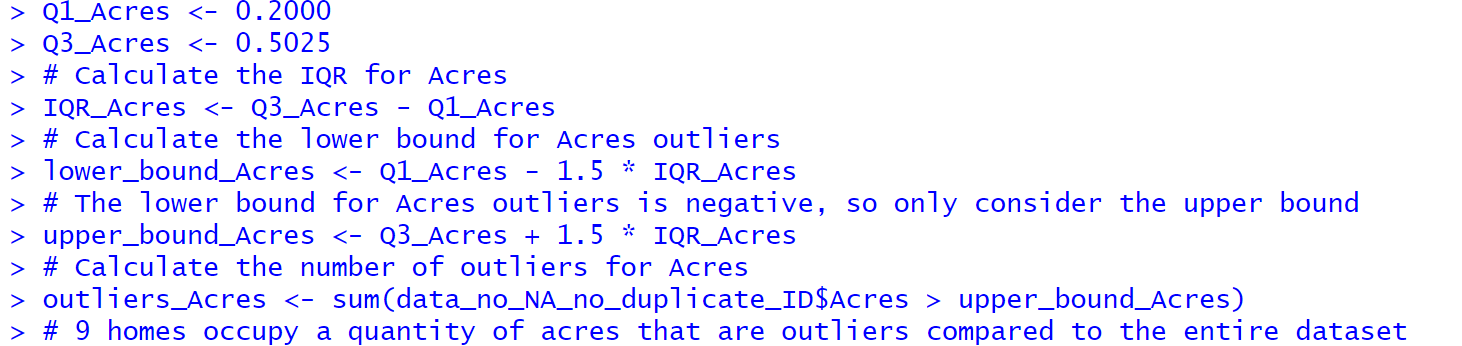


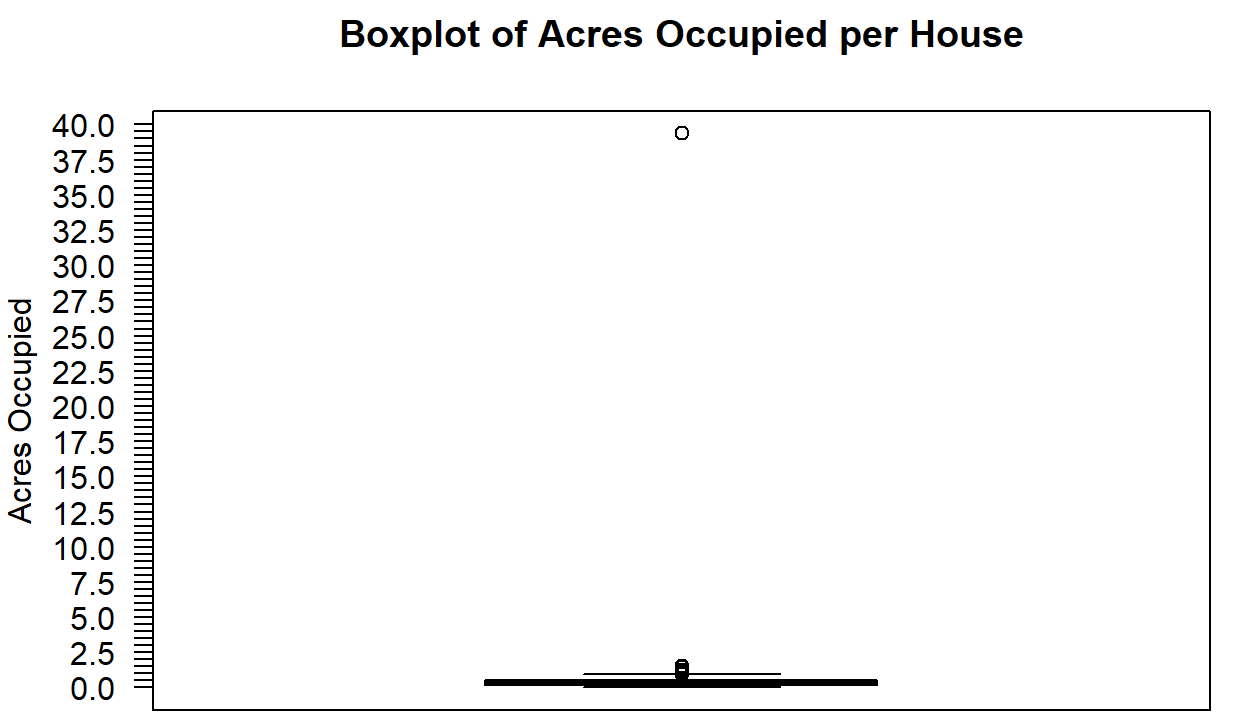
We conclude that building prices in the dataset are right-skewed due to the presence of extreme outliers above the mean, which makes the median a better measure of the typical building price among all homes sampled. In this case, the median building price is $110,260.

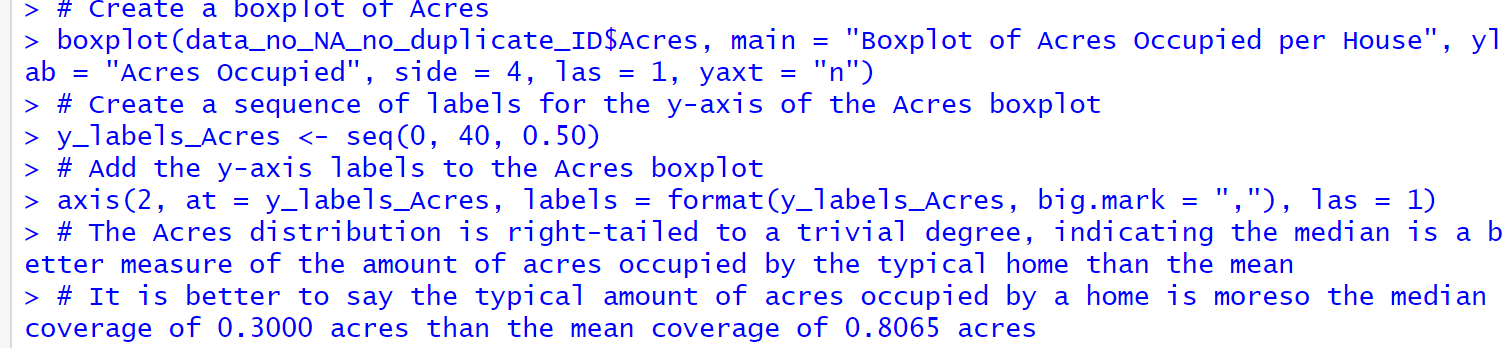
Do we observe a similar distribution in another variable, Acres, which quanitifies how many acres of land a house occupies? We present the five-number summary for this second variable of interest, once again using the summary() function:



As with BuildingPrice, the maximum Acres value is way higher than the median and mean. Calculating the IQR and the number of outliers, then visualizing this summary with a boxplot confirms these provocative findings:



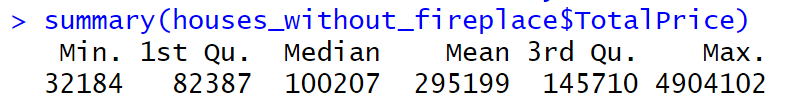


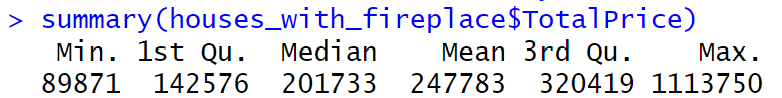


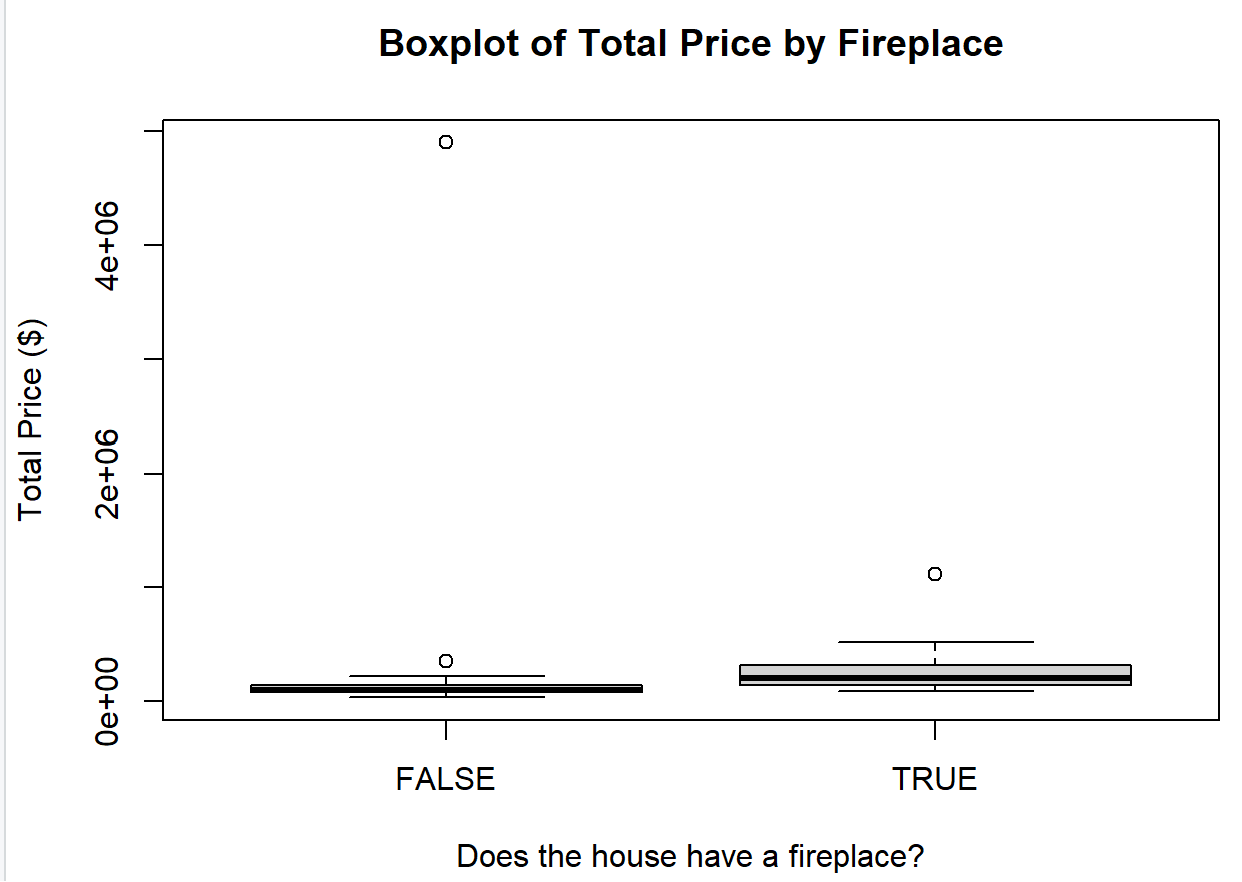
We also conclude that the distribution of acres occupied is right-skewed: some houses that occupy more acres by a significant, trivial difference compared to the mean and median cause the distribution to be non-symmetric, that is the distribution fails to be normal. As with BuildingPrice, for Acres it is better to measure the typical acres occupied per house with the median, which is 0.3000 acres.

**Price Comparison between Houses with and without Fireplaces**

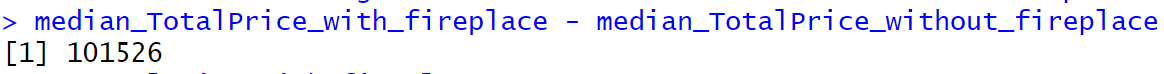
Does the presence of a fireplace increase the price of a house? If so, does the fireplace increase or decrease the total price of that house, or just either its building price or land price? We answer this question, first by testing the strength of each relationship between FirePlace, a variable whose values indicate whether a house contains a fireplace, and TotalPrice, which measures the total price of a house: we determine the value of this correlation coefficient, correlation\_FirePlace\_TotalPrice, to be approximately -0.0427717044903449. This value is almost 0, which suggests that any linear relationship between FirePlace and TotalPrice is very weak at best; other statisticians may argue that this relationship is so weak that it does not exist. This does not mean the end of our analysis; instead, we then compare the TotalPrice distributions of homes either with or without a fireplace to determine if the presence of a fireplace changes the TotalPrice value of a house. Both the summary() function and a boxplot visualization confirm our observations:







The boxplot shows the median TotalPrice value increases if a house has a fireplace, and comparing median TotalPrice values from both summary() outputs confirms this:



The summary() outputs suggest that the presence of a fireplace increases the median TotalPrice value for a house by $101,526.

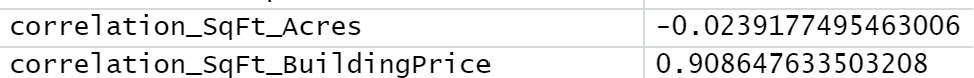
Another interesting observation about both distributinos is that the TotalPrice distribution for houses without a fireplace has two outliers, compared to one outlier for houses with a fireplace. However, this alone cannot determine whether one distribution is more skewed than the other, because there is no causal relationship between the number of outliers a distribution has and skewness. In these two particular distributions, according to the boxplot above, the houses without a fireplace have a TotalPrice distribution that is more right-skewed than the corresponding distribution for houses with a fireplace.

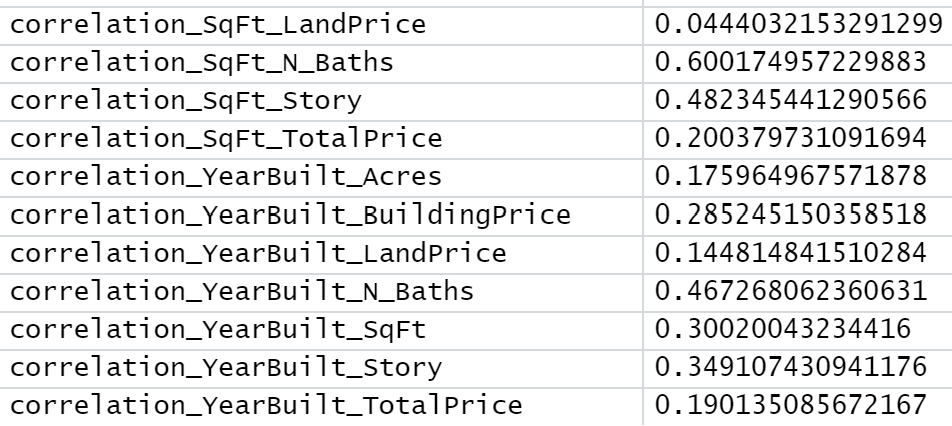
**Numerical Relationship Exploration**

The continuous variables of the dataset are SqFt, Acres, TotalPrice, LandPrice, and BuildingPrice, because they may contain decimal values and these values can exist in a continuous range with no discrete increment (or “step-size”) between them. This second criterion is why the other variables, including ID, YearBuilt, Story, and N\_Baths, are discrete and not continuous: a predetermined increment exists that separates these values. This means that N\_Baths increment by only 0.5, so a N\_Baths value of 1.75 would not be acceptable. On the other hand, incrementing a TotalPrice value by any amount is acceptable, even including increments that use up 26 decimal places.

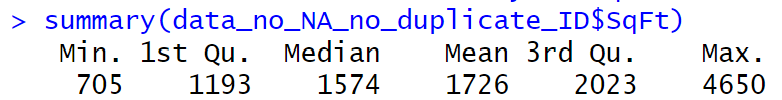
Just by the head() function, we can determine FirePlace is a discrete variable, because it can hold only TRUE or FALSE values.

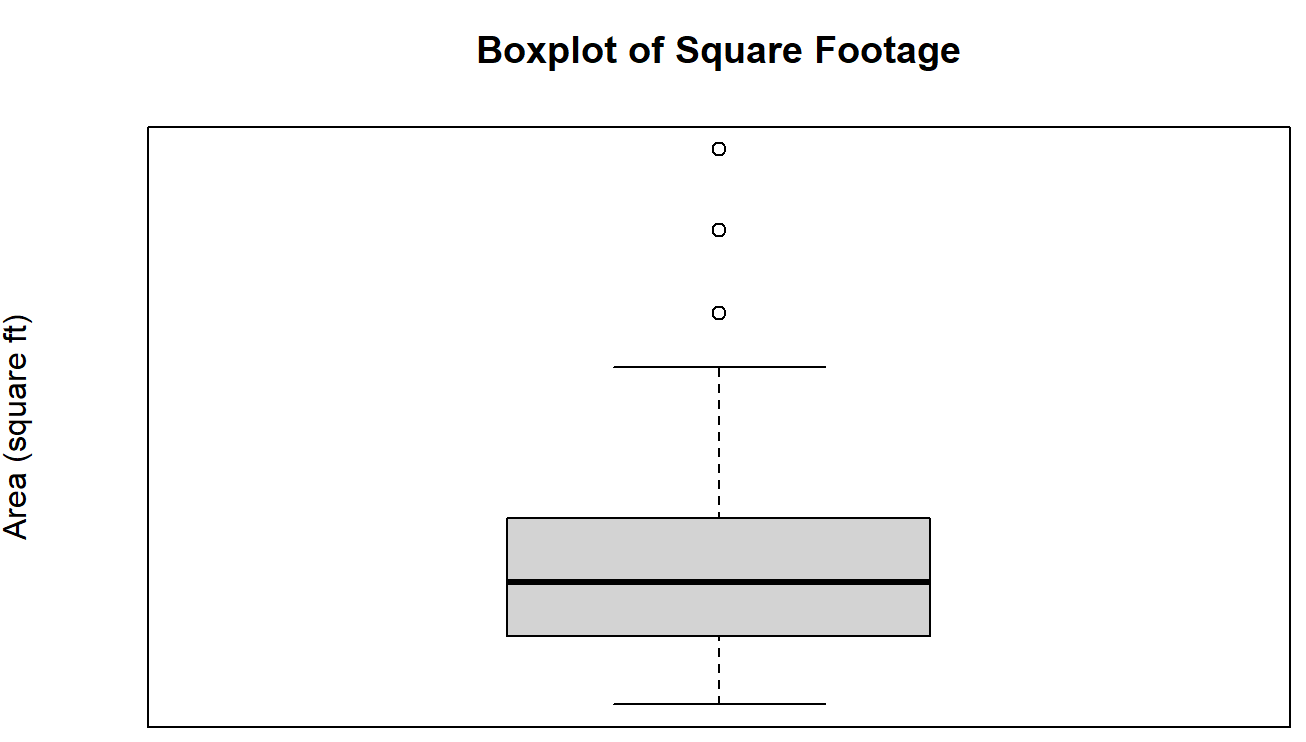
After identifying which variables in the dataset are continuous, we compute several correlation coefficient values. By intuition, it makes sense for the correlation coefficient between SqFt and housing prices to be high (its value approaches *r*2 = 1). Nevertheless, we also present several computations of other correlation coefficients for other variables, many of which seem to have no strong linear relationship, in case the dataset contains any surprises:



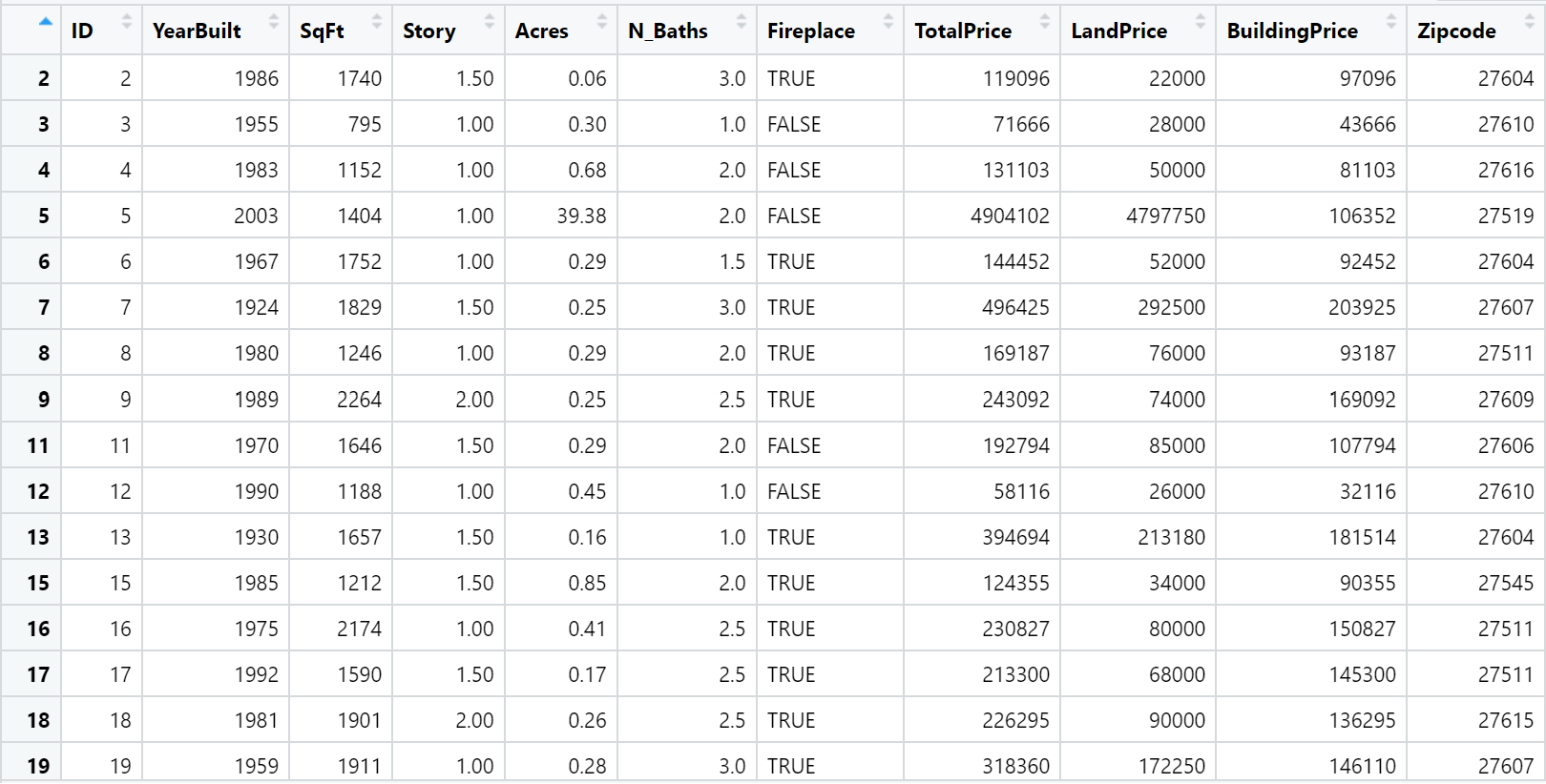


Out of all these correlation coefficients, the strongest one is between SqFt and BuildingPrice. This validates our prediction earlier, and it makes sense that a larger house would increase its BuildingPrice value. However, we cannot perform a linear regression analysis on this relationship until we verify that both SqFt and BuildingPrice follow a normal distribution. Like in previous sections of this report, we perform the summary() function and create a boxplot to see if any SqFt values are outliers:

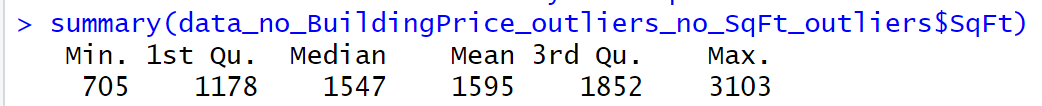


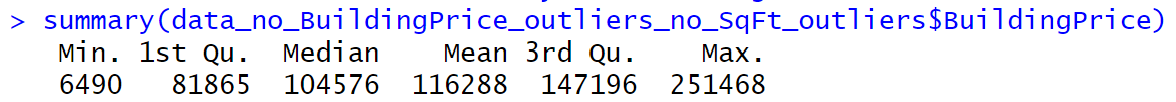


Indeed, both the summary() output and boxplot indicate SqFt contains outliers. Now, what happens to the correlation coefficient if we removed outliers from both SqFt and BuildingPrice? To test this, we first need to remove all BuildingPrice outliers from data\_no\_NA\_no\_duplicate\_ID, which we discovered in the **Variable Summarization** part of this report. This is a preview of the new dataset that excludes the outliers of both variables:

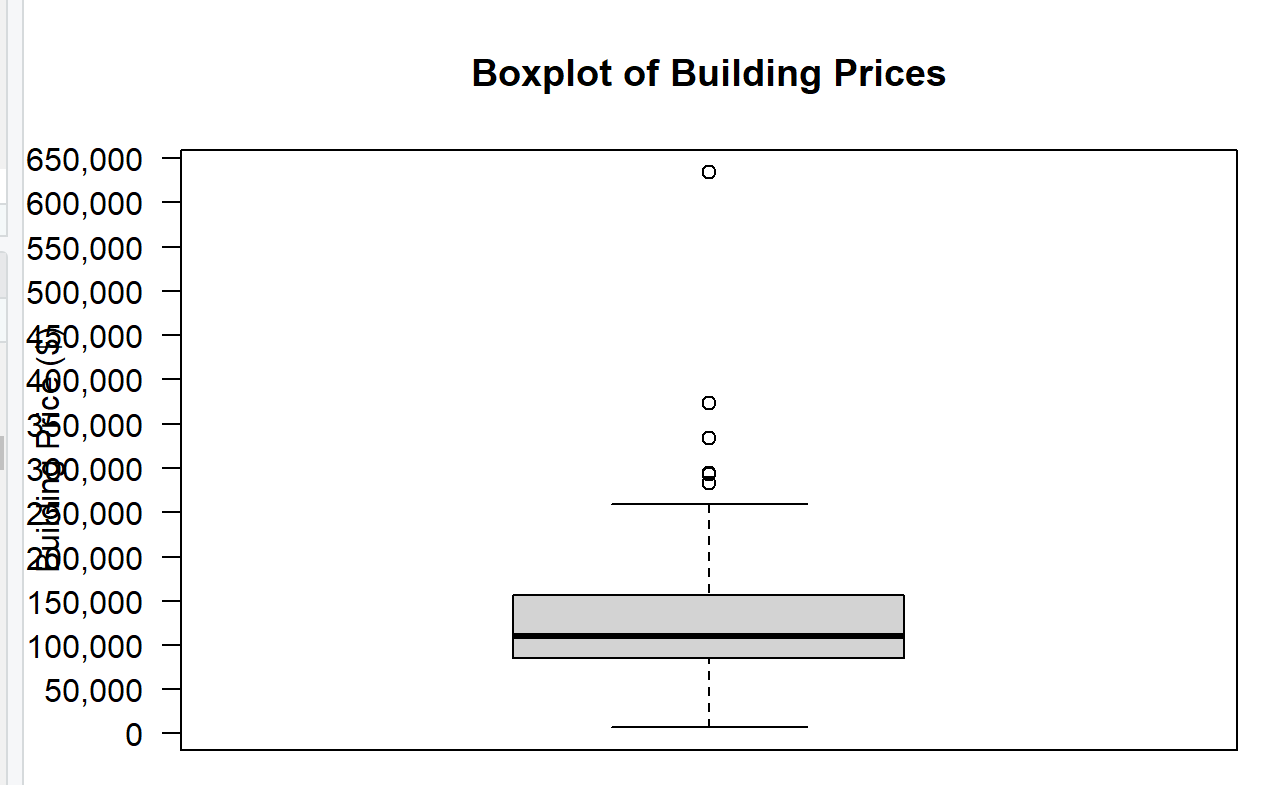


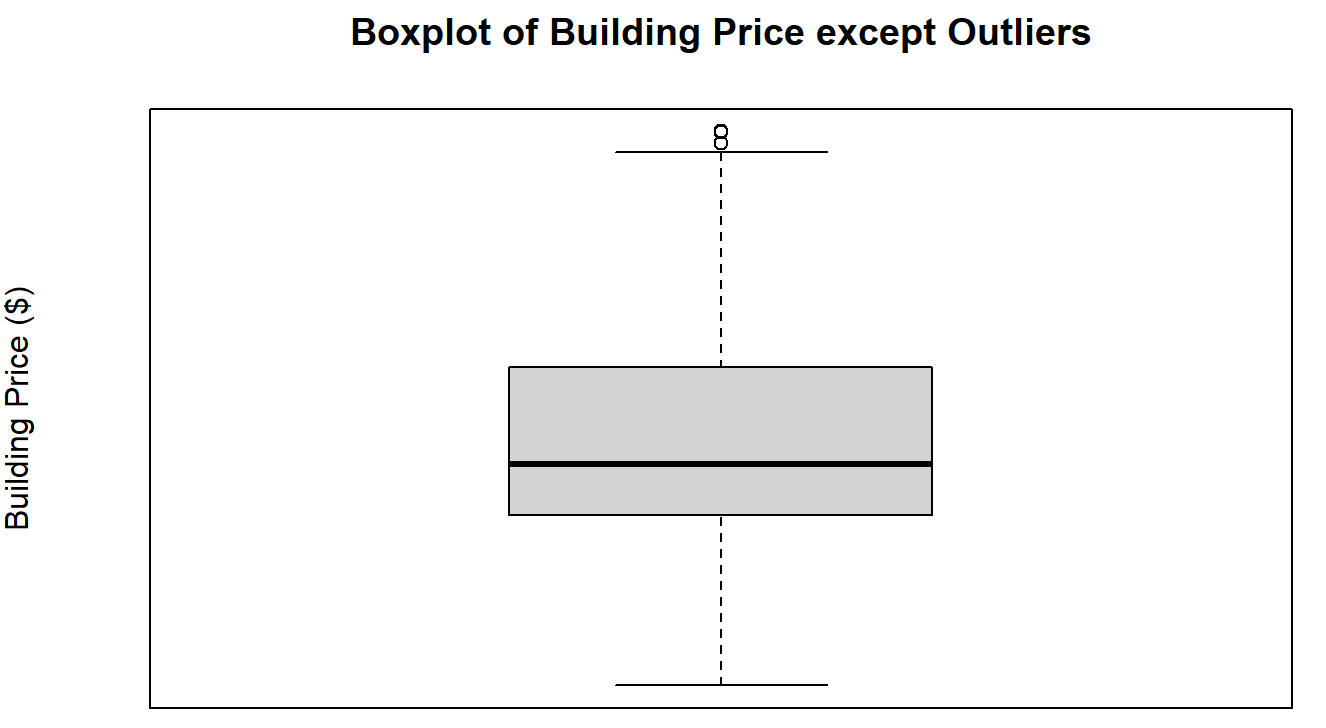
Then can we remove all SqFt outliers from the same dataset and produce boxplots for both SqFt and BuildingPrice distributions without their outliers. Here are the summary() outputs upon removing SqFt and BuildingPrice outliers for both variables:



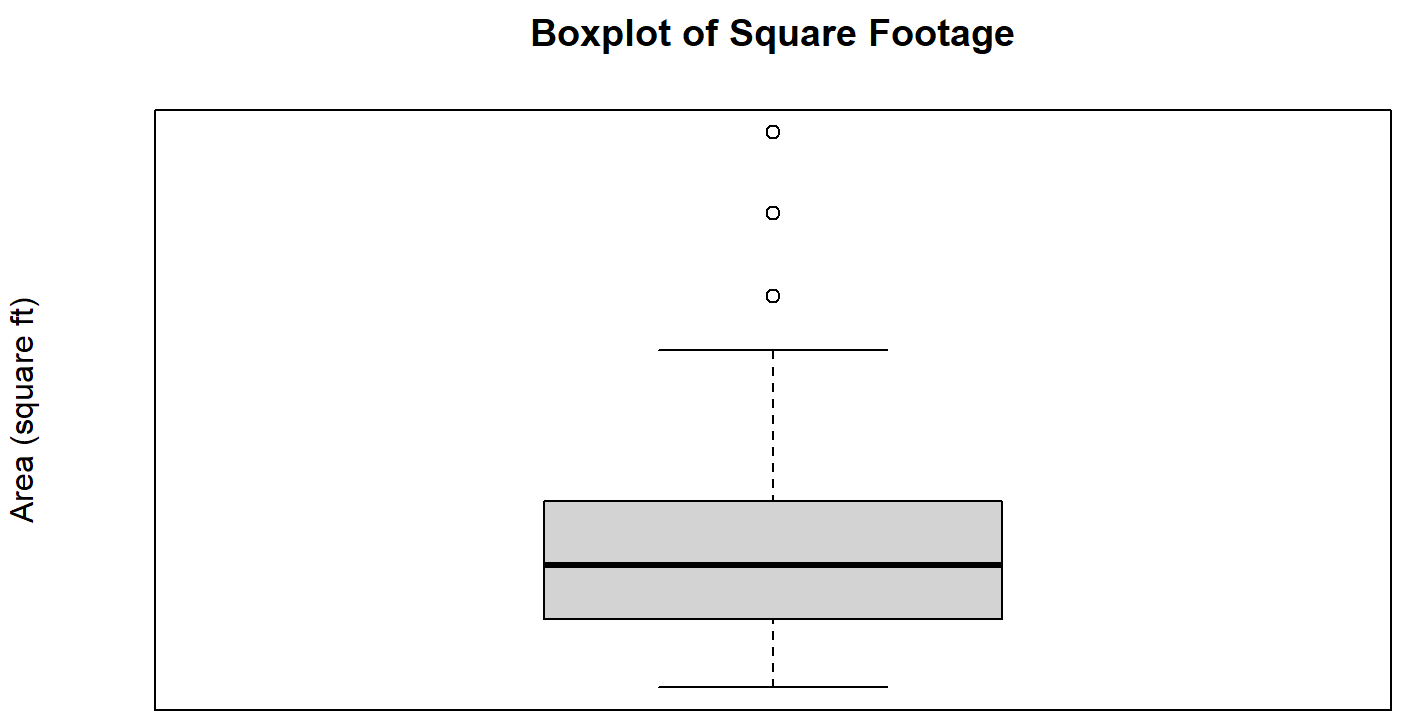


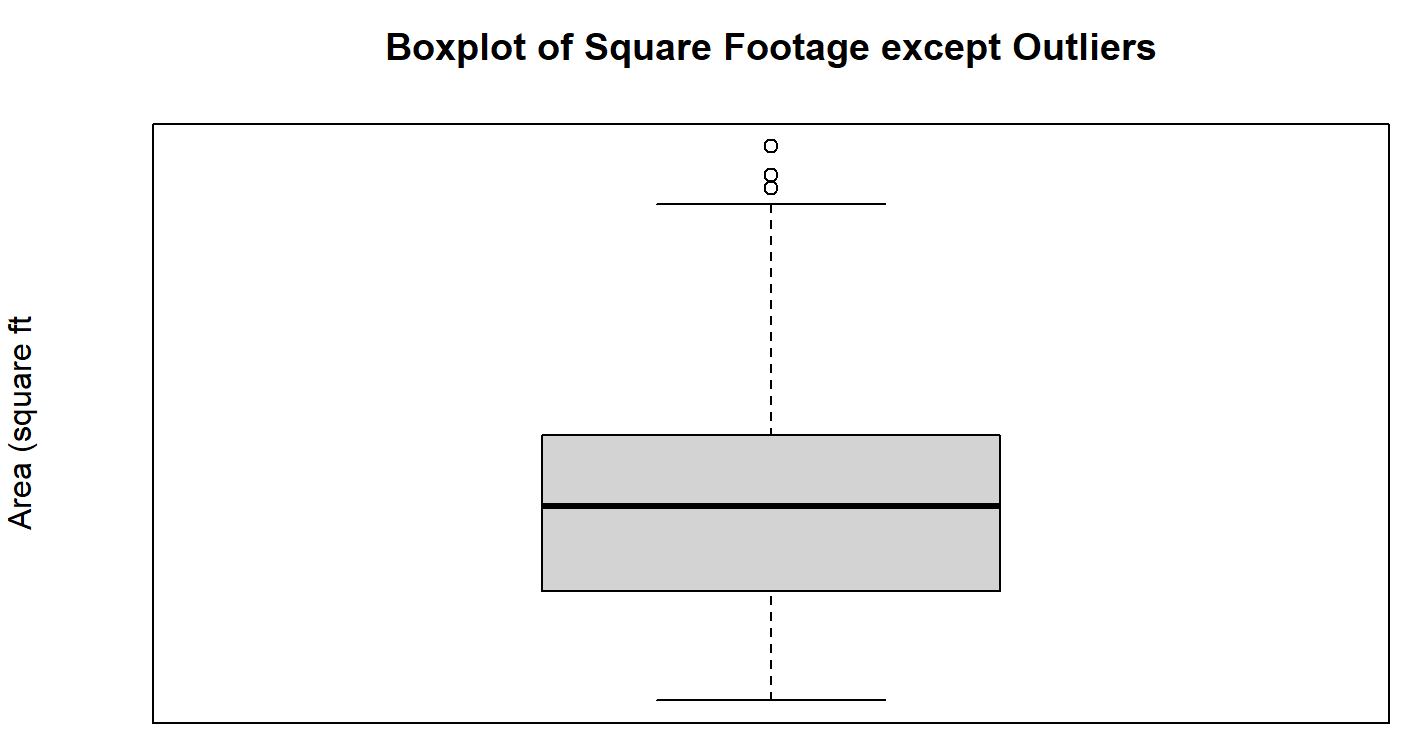
For comparison of both SqFt and BuildingPrice boxplots after removal of their outliers, we also include their original boxplots side-by-side:





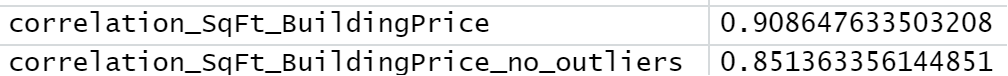
The inclusion of the cluster of values right above the top whisker in the BuildingPrice boxplot indicates that houses with a BuildingPrice value above $300,000 are outliers we removed from our new dataset.



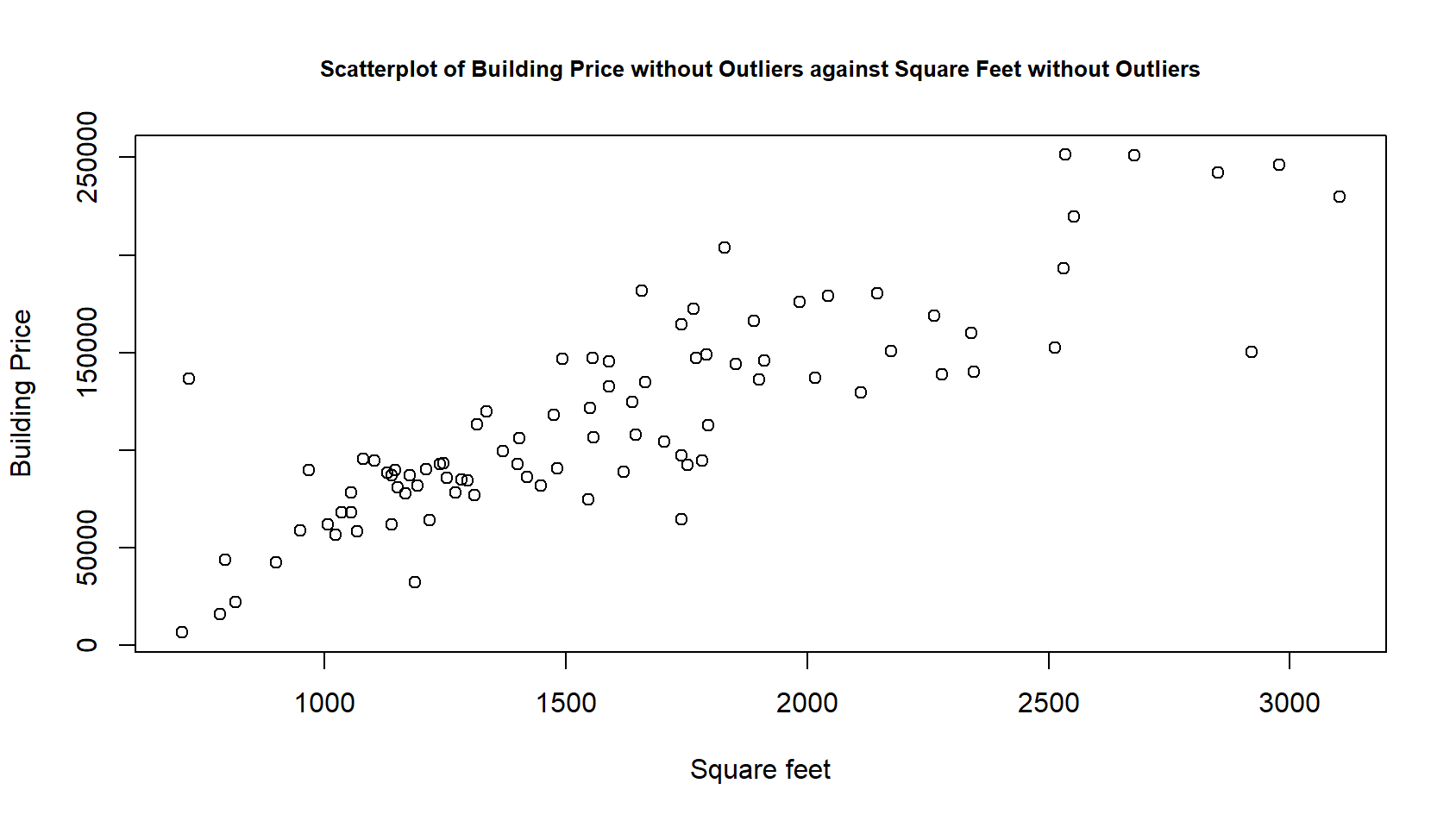


A similar pattern appears in the new SqFt boxplot. One more noticeable feature is the median SqFt value moves up upon removing all SqFt outliers.

Since both variables no longer contain their outliers, they follow a normal distribution. Now we can recalculate the correlation coefficient for SqFt and BuildingPrice and see if it changes by a significant margin compared to the correlation coefficient that factored in their outliers:

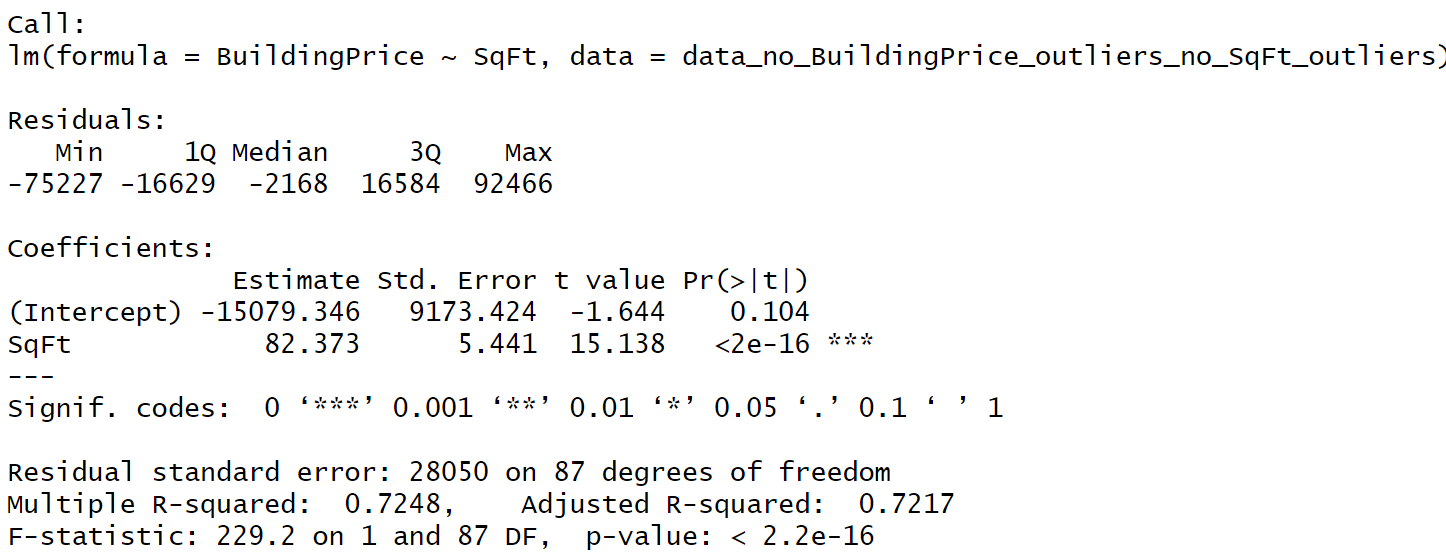


The new correlation coefficient is not far off from our original value, which indicates this new dataset will produce similar results when testing the validity of the regression line from the values these two variables carry in our linear regression analysis. To prepare for this next step, we present a scatter plot of BuildingPrice against SqFt:



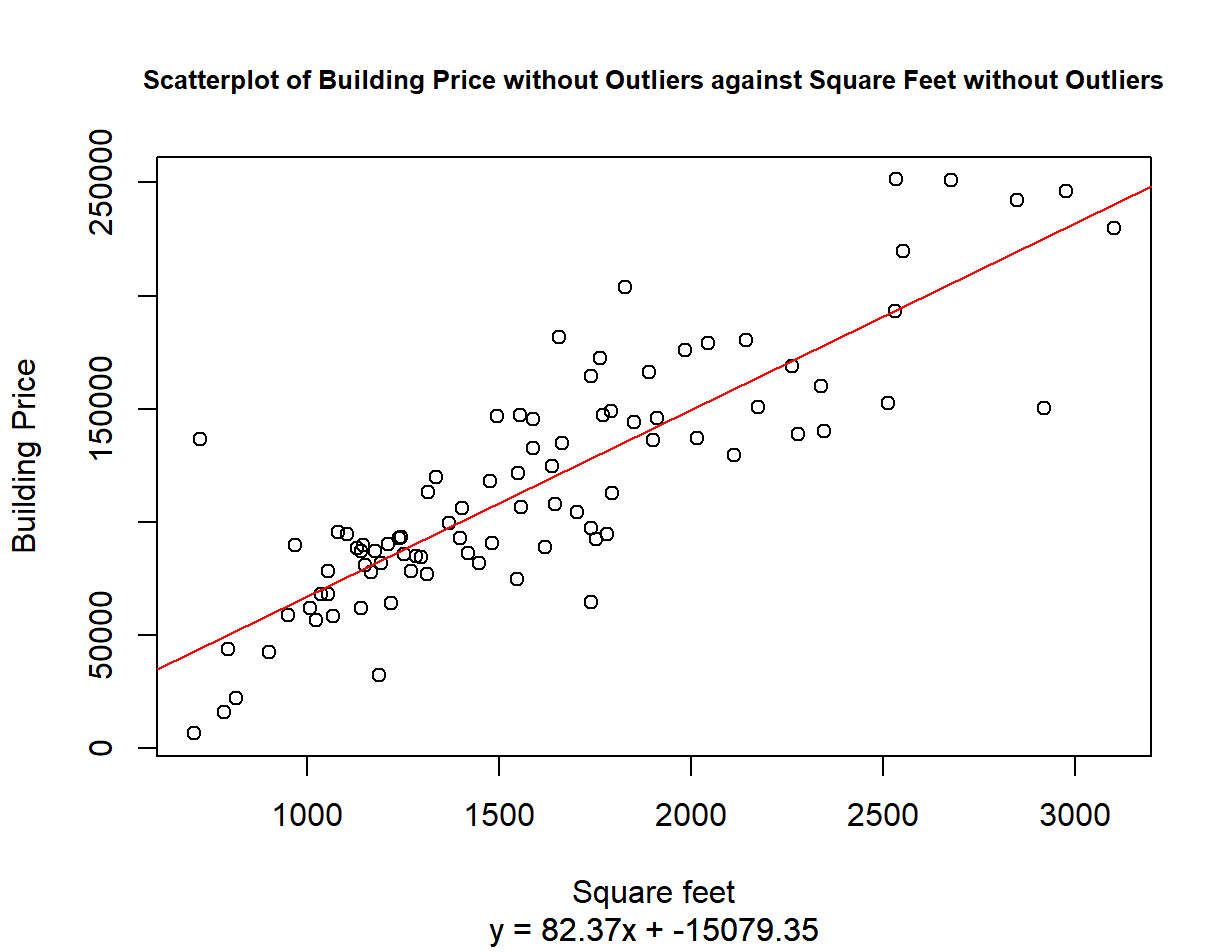
**Linear Regression Analysis**

We begin our linear regression analysis by constructing a linear model for BuildingPrice against SqFt via the lm() function. Performing the summary() function on this linear model returns many statistical values that determine how strong this model relates both variables:



The *p*-value of this model is *p* < 2(10-16), which is below the significance threshold of *p* < 0.05, indicating the model provides an accurate relationship between SqFt and BuildingPrice: it is a good fit of this linear relationship.

Since this linear model plots BuildingPrice against SqFt, the coefficient of the regression equation is 82.373 and the intercept is -15079.346. Reflecting a negative price, this intercept may scratch some heads, but there may be particular economic circumstances where it makes sense to give away a house, especially if its square footage is so little that it makes no sense to live in a house that tiny. Upon making sense of the regression model, we then plot its equation upon the scatter plot we produced from the previous part of this report:



**Conclusion**

Upon finalizing our linear regression analysis, we found that, on average, houses that occupy square feet have a higher building price based upon our finding that there is a significant linear relationship between building price and square footage based upon correlation coefficient (*r*2) and *p*-values alongside a plethora of visualizations.

**Appendix**

We submitted our R code as a separate file in addition to this report, but we also include our code below in case technical difficulties arise with our R file:

# STATS 10 Fall 2023 Project

# Edmund Leong 206049891

# Set working directory to Downloads

setwd("C:/Users/leong/Downloads")

# Read "data.csv" into data

data <- read.csv("data.csv", header = TRUE)

## Part 1: Data Handling

# Print a handful of rows of data

head(data)

# Return a matrix of TRUE and FALSE values that determine which values are NA

NA\_matrix <- is.na(data)

# Calculate the sum of TRUE values in each row of NA\_matrix

TRUE\_NA <- rowSums(NA\_matrix)

# Return the indices of the rows of TRUE\_NA that contain NA values

TRUE\_NA\_indices <- which(TRUE\_NA > 0)

# Return the ID column of data for TRUE\_NA\_indices

ID\_TRUE\_NA\_indices <- data$ID[TRUE\_NA\_indices]

# Homes with ID 1, 14, 51, or 80 have at least one NA value

# Remove homes with NA values

data\_no\_NA <- data[complete.cases(data),]

# Print a handful of rows of data\_no\_NA

head(data\_no\_NA)

# Return which rows of data\_no\_NA have duplicate ID values

row\_duplicate\_ID <- duplicated(data\_no\_NA$ID)

# Return all duplicate ID values in data\_no\_NA

duplicate\_ID <- data\_no\_NA[row\_duplicate\_ID, "ID"]

# Return data\_no\_NA without duplicate ID values

data\_no\_NA\_no\_duplicate\_ID <- data\_no\_NA[!row\_duplicate\_ID,]

# Print a handful of rows of data\_no\_NA\_no\_duplicate\_ID

head(data\_no\_NA\_no\_duplicate\_ID)

## Part 2: Variable Summarization

# Return a five-number summary of building prices

summary(data\_no\_NA\_no\_duplicate\_ID$BuildingPrice)

# The mean home price is $133577

# The median home price is $110260

Q1\_BuildingPrice <- 84680

Q3\_BuildingPrice <- 154377

# Calculate the IQR for BuildingPrice

IQR\_BuildingPrice <- Q3\_BuildingPrice - Q1\_BuildingPrice

# Calculate the lower bound for BuildingPrice outliers

lower\_bound\_BuildingPrice <- Q1\_BuildingPrice - 1.5 \* IQR\_BuildingPrice

# lower\_bound\_BuildingPrice < 0, so only an upper bound for BuildingPrice outliers exist

# Calculate the upper bound for BuildingPrice outliers

upper\_bound\_BuildingPrice <- Q3\_BuildingPrice + 1.5 \* IQR\_BuildingPrice

# The lower bound for BuildingPrice outliers is negative, so only consider the upper bound

outliers\_BuildingPrice <- sum(data\_no\_NA\_no\_duplicate\_ID$BuildingPrice > upper\_bound\_BuildingPrice)

# 7 homes have a building price that is an outlier compared to the entire dataset

# Create a boxplot of BuildingPrice

boxplot(data\_no\_NA\_no\_duplicate\_ID$BuildingPrice, main = "Boxplot of Building Prices", ylab = "Building Price ($)", side = 4, las = 1, yaxt = "n")

# Create a sequence of labels for the y-axis of the BuildingPrice boxplot

y\_labels\_BuildingPrice <- seq(0, 650000, 50000)

# Add the y-axis labels to the BuildingPrice boxplot

axis(2, at = y\_labels\_BuildingPrice, labels = format(y\_labels\_BuildingPrice, big.mark = ","), las = 1)

# Create a histogram of BuildingPrice

hist(data\_no\_NA\_no\_duplicate\_ID$BuildingPrice, main = "Histogram of Building Prices", xlab = "Building Price ($)", ylab = "Frequency")

# The BuildingPrice distribution of building prices is right-tailed to a trivial degree, indicating the median is a better measure of the typical building price than the mean

# It is better to say the typical building price is moreso the median price of $110260 than $133577

# Return the five-number summary of acres

summary(data\_no\_NA\_no\_duplicate\_ID$Acres)

# The mean house occupies 0.8065 acres

# The median house occupies 0.3000 acres

Q1\_Acres <- 0.2000

Q3\_Acres <- 0.5025

# Calculate the IQR for Acres

IQR\_Acres <- Q3\_Acres - Q1\_Acres

# Calculate the lower bound for Acres outliers

lower\_bound\_Acres <- Q1\_Acres - 1.5 \* IQR\_Acres

# The lower bound for Acres outliers is negative, so only consider the upper bound

upper\_bound\_Acres <- Q3\_Acres + 1.5 \* IQR\_Acres

# Calculate the number of outliers for Acres

outliers\_Acres <- sum(data\_no\_NA\_no\_duplicate\_ID$Acres > upper\_bound\_Acres)

# 9 homes occupy a quantity of acres that are outliers compared to the entire dataset

# Create a boxplot of Acres

boxplot(data\_no\_NA\_no\_duplicate\_ID$Acres, main = "Boxplot of Acres Occupied per House", ylab = "Acres Occupied", side = 4, las = 1, yaxt = "n")

# Create a sequence of labels for the y-axis of the Acres boxplot

y\_labels\_Acres <- seq(0, 40, 0.50)

# Add the y-axis labels to the Acres boxplot

axis(2, at = y\_labels\_Acres, labels = format(y\_labels\_Acres, big.mark = ","), las = 1)

# The Acres distribution is right-tailed to a trivial degree, indicating the median is a better measure of the amount of acres occupied by the typical home than the mean

# It is better to say the typical amount of acres occupied by a home is moreso the median coverage of 0.3000 acres than the mean coverage of 0.8065 acres

## Part 3: Price Comparison

# Calculate the correlation between Fireplace and TotalPrice

correlation\_Fireplace\_TotalPrice <- cor(data\_no\_NA\_no\_duplicate\_ID$Fireplace, data\_no\_NA\_no\_duplicate\_ID$TotalPrice)

# The correlation is near-identical to 0, which shows that the presence of a fireplace has little effect on a house's total price

# Create a boxplot of TotalPrice for homes with and without a fireplace

boxplot(data\_no\_NA\_no\_duplicate\_ID$TotalPrice ~ data\_no\_NA\_no\_duplicate\_ID$Fireplace, main = "Boxplot of Total Price by Fireplace", xlab = "Does the house have a fireplace?", ylab = "Total Price ($)")

# The boxplot displays miniscule, yet also visible, differences in summary statistics for TotalPrice by Fireplace

# Return all homes without a fireplace

houses\_without\_fireplace <- data\_no\_NA\_no\_duplicate\_ID[data\_no\_NA\_no\_duplicate\_ID$Fireplace == FALSE,]

# Return the five-number summary of TotalPrice for houses\_without\_fireplace

summary(houses\_without\_fireplace$TotalPrice)

median\_TotalPrice\_without\_fireplace <- 100207

Q1\_TotalPrice\_without\_fireplace <- 82387

Q3\_TotalPrice\_without\_fireplace <- 145710

# Calculate the IQR of TotalPrice for homes without a fireplace

IQR\_TotalPrice\_without\_fireplace <- Q3\_TotalPrice\_without\_fireplace - Q1\_TotalPrice\_without\_fireplace

# Calculate the lower bound for TotalPrice outliers for homes without a fireplace

lower\_bound\_TotalPrice\_without\_fireplace <- Q1\_TotalPrice\_without\_fireplace - 1.5 \* IQR\_TotalPrice\_without\_fireplace

# Calculate the upper bound for TotalPrice outliers for homes without a fireplace

upper\_bound\_TotalPrice\_without\_fireplace <- Q3\_TotalPrice\_without\_fireplace + 1.5 \* IQR\_TotalPrice\_without\_fireplace

# Count the number of TotalPrice outliers for homes without a fireplace

outliers\_TotalPrice\_without\_fireplace <- sum(houses\_without\_fireplace$TotalPrice < lower\_bound\_TotalPrice\_without\_fireplace) + sum(houses\_without\_fireplace$TotalPrice > upper\_bound\_TotalPrice\_without\_fireplace)

# Since the lower bound is negative, consider only the upper bound

# This means the 2 outliers for TotalPrice for homes without a fireplace are above the upper bound

# Return all homes with a fireplace

houses\_with\_fireplace <- data\_no\_NA\_no\_duplicate\_ID[data\_no\_NA\_no\_duplicate\_ID$Fireplace == TRUE,]

# Return the five-number summary of TotalPrice for houses\_with\_fireplace

summary(houses\_with\_fireplace$TotalPrice)

median\_TotalPrice\_with\_fireplace <- 201733

# Calculate the change in TotalPrice with the addition of a fireplace

median\_TotalPrice\_with\_fireplace - median\_TotalPrice\_without\_fireplace

Q1\_TotalPrice\_with\_fireplace <- 142576

Q3\_TotalPrice\_with\_fireplace <- 320419

# Calculate the IQR of TotalPrice for homes with a fireplace

IQR\_TotalPrice\_with\_fireplace <- Q3\_TotalPrice\_with\_fireplace - Q1\_TotalPrice\_with\_fireplace

# Calculate the lower bound for TotalPrice outliers for homes with a fireplace

lower\_bound\_TotalPrice\_with\_fireplace <- Q1\_TotalPrice\_with\_fireplace - 1.5 \* IQR\_TotalPrice\_with\_fireplace

# Calculate the upper bound for TotalPrice outliers for homes with a fireplace

upper\_bound\_TotalPrice\_with\_fireplace <- Q3\_TotalPrice\_with\_fireplace + 1.5 \* IQR\_TotalPrice\_with\_fireplace

# Count the number of TotalPrice outliers for homes with a fireplace

outliers\_TotalPrice\_with\_fireplace <- sum(houses\_with\_fireplace$TotalPrice < lower\_bound\_TotalPrice\_with\_fireplace) + sum(houses\_with\_fireplace$TotalPrice > upper\_bound\_TotalPrice\_with\_fireplace)

# Because the set of homes without fireplaces has more extreme outliers for TotalPrice than the set of homes with fireplaces, the distribution of TotalPrice for homes without fireplaces will be more right-tailed than for homes with fireplaces

## Part 4: Numerical Relationship Explanation:

# Calculate the correlation between YearBuilt and SqFt

correlation\_YearBuilt\_SqFt <- cor(data\_no\_NA\_no\_duplicate\_ID$YearBuilt, data\_no\_NA\_no\_duplicate\_ID$SqFt)

# Calculate the correlation between YearBuilt and Story

correlation\_YearBuilt\_Story <- cor(data\_no\_NA\_no\_duplicate\_ID$YearBuilt, data\_no\_NA\_no\_duplicate\_ID$Story)

# Calculate the correlation between YearBuilt and Acres

correlation\_YearBuilt\_Acres <- cor(data\_no\_NA\_no\_duplicate\_ID$YearBuilt, data\_no\_NA\_no\_duplicate\_ID$Acres)

# Calculate the correlation between YearBuilt and N\_Baths

correlation\_YearBuilt\_N\_Baths <- cor(data\_no\_NA\_no\_duplicate\_ID$YearBuilt, data\_no\_NA\_no\_duplicate\_ID$N\_Baths)

# Calculate the correlation between YearBuilt and TotalPrice

correlation\_YearBuilt\_TotalPrice <- cor(data\_no\_NA\_no\_duplicate\_ID$YearBuilt, data\_no\_NA\_no\_duplicate\_ID$TotalPrice)

# Calculate the correlation between YearBuilt and LandPrice

correlation\_YearBuilt\_LandPrice <- cor(data\_no\_NA\_no\_duplicate\_ID$YearBuilt, data\_no\_NA\_no\_duplicate\_ID$LandPrice)

# Calculate the correlation between YearBuilt and BuildingPrice

correlation\_YearBuilt\_BuildingPrice <- cor(data\_no\_NA\_no\_duplicate\_ID$YearBuilt, data\_no\_NA\_no\_duplicate\_ID$BuildingPrice)

# Calculate the correlation between SqFt and Story

correlation\_SqFt\_Story <- cor(data\_no\_NA\_no\_duplicate\_ID$SqFt, data\_no\_NA\_no\_duplicate\_ID$Story)

# Calculate the correlation between SqFt and Acres

correlation\_SqFt\_Acres <- cor(data\_no\_NA\_no\_duplicate\_ID$SqFt, data\_no\_NA\_no\_duplicate\_ID$Acres)

# Calculate the correlation between SqFt and N\_Baths

correlation\_SqFt\_N\_Baths <- cor(data\_no\_NA\_no\_duplicate\_ID$SqFt, data\_no\_NA\_no\_duplicate\_ID$N\_Baths)

# Calculate the correlation between SqFt and TotalPrice

correlation\_SqFt\_TotalPrice <- cor(data\_no\_NA\_no\_duplicate\_ID$SqFt, data\_no\_NA\_no\_duplicate\_ID$TotalPrice)

# Calculate the correlation between SqFt and LandPrice

correlation\_SqFt\_LandPrice <- cor(data\_no\_NA\_no\_duplicate\_ID$SqFt, data\_no\_NA\_no\_duplicate\_ID$LandPrice)

# Calculate the correlation between SqFt and BuildingPrice

correlation\_SqFt\_BuildingPrice <- cor(data\_no\_NA\_no\_duplicate\_ID$SqFt, data\_no\_NA\_no\_duplicate\_ID$BuildingPrice)

# There appears to be a strong, positive relationship between SqFt and BuildingPrice

# Return a five-number summary of SqFt

summary(data\_no\_NA\_no\_duplicate\_ID$SqFt)

Q1\_SqFt <- 1193

Q3\_SqFt <- 2023

# Calculate the IQR for SqFt

IQR\_SqFt <- Q3\_SqFt - Q1\_SqFt

# Calculate the lower bound for SqFt values

lower\_bound\_SqFt <- Q1\_SqFt - 1.5 \* IQR\_SqFt

# Calculate the upper bound for SqFt values

upper\_bound\_SqFt <- Q3\_SqFt + 1.5 \* IQR\_SqFt

# Create a boxplot of SqFt

boxplot(data\_no\_NA\_no\_duplicate\_ID$SqFt, main = "Boxplot of Square Footage", ylab = "Area (square ft)", side = 4, las = 1, yaxt = "n")

# Remove BuildingPrice outliers from data\_no\_NA\_no\_duplicate\_ID

data\_no\_BuildingPrice\_outliers <- data\_no\_NA\_no\_duplicate\_ID[data\_no\_NA\_no\_duplicate\_ID$BuildingPrice >= lower\_bound\_BuildingPrice & data\_no\_NA\_no\_duplicate\_ID$BuildingPrice <= upper\_bound\_BuildingPrice,]

# Remove SqFt outliers from data\_no\_NA\_no\_duplicate\_ID

data\_no\_BuildingPrice\_outliers\_no\_SqFt\_outliers <- data\_no\_BuildingPrice\_outliers[data\_no\_BuildingPrice\_outliers$SqFt >= lower\_bound\_SqFt & data\_no\_BuildingPrice\_outliers$SqFt <= upper\_bound\_SqFt,]

# Return the five-number summary for SqFt in the new dataset

summary(data\_no\_BuildingPrice\_outliers\_no\_SqFt\_outliers$SqFt)

# Create a boxplot of SqFt from the new dataset

boxplot(data\_no\_BuildingPrice\_outliers\_no\_SqFt\_outliers$SqFt, main = "Boxplot of Square Footage except Outliers", ylab = "Area (square ft", side = 4, las = 1, yaxt = "n")

# Return the five-number summary for BuildingPrice from the new dataset

summary(data\_no\_BuildingPrice\_outliers\_no\_SqFt\_outliers$BuildingPrice)

# Create a boxplot of BuildingPrice from the new dataset

boxplot(data\_no\_BuildingPrice\_outliers\_no\_SqFt\_outliers$BuildingPrice, main = "Boxplot of Building Price except Outliers", ylab = "Building Price ($)", side = 4, las = 1, yaxt = "n")

# Calculate the correlation between SqFt and BuildingPrice from the new dataset

correlation\_SqFt\_BuildingPrice\_no\_outliers <- cor(data\_no\_BuildingPrice\_outliers\_no\_SqFt\_outliers$SqFt, data\_no\_BuildingPrice\_outliers\_no\_SqFt\_outliers$BuildingPrice)

# Create a scatterplot of BuildingPrice against SqFt from the new dataset

plot(data\_no\_BuildingPrice\_outliers\_no\_SqFt\_outliers$SqFt, data\_no\_BuildingPrice\_outliers\_no\_SqFt\_outliers$BuildingPrice, main = "Scatterplot of Building Price without Outliers against Square Feet without Outliers", cex.main = 0.8, xlab = "Square feet", ylab = "Building Price")

## Part 5: Linear Regression Analysis

# Create a linear model for SqFt and BuildingPrice

model\_SqFt\_BuildingPrice <- lm(BuildingPrice ~ SqFt, data = data\_no\_BuildingPrice\_outliers\_no\_SqFt\_outliers)

# Return a summary of the model

summary(model\_SqFt\_BuildingPrice)

# Return the slope of the model

slope <- round(coef(model\_SqFt\_BuildingPrice)[2], 2)

# Return the intercept of the model

intercept <- round(coef(model\_SqFt\_BuildingPrice)[1], 2)

# Plot the regression line on the scatterplot

abline(model\_SqFt\_BuildingPrice, col = "red")

# Add the equation of the regression line to the scatterplot

equation <- paste0("y = ", slope, "x + ", intercept)

mtext(equation, side = 1, line = 4)