**STATS 10 FINAL PROJECT**

STATISTICAL ANALYSIS OF A REAL-ESTATE DATABASE

## 1. Introduction:

This statistical paper examines a real estate dataset that contains information on 100 residential properties randomly selected from a specific County registry. The dataset includes 10 variables, such as year built, square footage, adjusted land value, and zip code. The aim of the study is to gain insights into the dataset by employing various statistical measures and graphical tools.

The paper begins with data cleaning to ensure the dataset's integrity by detecting and handling missing data. After data cleaning, variable summarization is performed, and two variables of interest are selected for in-depth analysis. The study also explores whether the presence of a fireplace impacts property prices. The analysis employs numerical and graphical tools to unravel the nuances of the variables and identify any irregularities or patterns within the data.

Moreover, the paper also investigates potential relationships among continuous variables and identifies any hidden connections by employing graphical tools. The analysis culminates in a linear regression analysis that focuses on two selected variables. The estimated coefficients are presented and interpreted, and goodness-of-fit metrics are evaluated to shed light on the predictive power and significance of the identified relationship within the context of the dataset.

## 2. Methodology:

### 2.1 Data Overview

The data used in this paper is taken from a real estate database that includes information on 100 randomly selected residential properties in a specific County registry. Each property has a unique ID number and 10 variables recorded, covering diverse aspects such as year of construction, square footage, number of acres, zip code, number of stories, number of baths, total price, land price, building price, and whether it has a fireplace or not.

### 2.2 Detecting and Handling Missing Data

In its initial form, the dataset contains missing values in the form of NAs. The process begins with the loading of the dataset from the "data.csv" file into RStudio, establishing a foundation for subsequent analysis. To identify missing values within the dataset, the 'is.na(data)' function is employed, generating a logical matrix where 'TRUE' indicates the presence of missing values. The total count of missing values in the entire dataset is then computed using 'sum(is.na(data))' Which came out to be 13, Extending this analysis to individual columns, such as 'ID,' 'YearBuilt,' 'SqFt,' and others, allowed for a detailed examination of missing data across specific variables. In the next step, observations containing missing data are removed to ensure the dataset's integrity. The 'na.omit(data)' function is utilized, resulting in a refined

dataset named 'data_clean' that is devoid of incomplete observations. Finally, the cleanliness of this processed dataset is confirmed by verifying the absence of missing values through 'sum(is.na(data_clean)).' This rigorous approach ensures that the dataset is thoroughly cleansed and ready for subsequent in-depth analysis. The subsequent removal of these incomplete observations is imperative to guarantee a clean and unambiguous dataset, poised for insightful analysis.

Variables with missing data:

YearBuilt, SqFt, Story, Acres, N_Baths, Fireplace, LandPrice, BuildingPrice, and Zipcode.

IDs of observations with missing data:

1, 14, 51, 80

## 3. Variable Summarization:

The selection of Square Footage (SqFt) and Total Assessed Value (TotalPrice) as variables of interest is grounded in their significance in real estate analysis, offering distinct perspectives on a property.

Square Footage (SqFt): The square footage of a property is a fundamental measure directly linked to its physical size. Larger square footage often correlates with more spacious living areas, influencing property functionality and market appeal. Analyzing SqFt allows us to discern trends in property sizes within the dataset, which can be crucial for understanding potential market trends, and the overall composition of properties in the specific County registry.

Total Assessed Value (TotalPrice): Total Assessed Value, encompassing both land and building values, is a pivotal financial metric providing an overall valuation of a property. Analyzing TotalPrice offers insights into the distribution of property values in the dataset, shedding light on the economic landscape of the properties sampled.

The ‘summary()’ function is used to provide a summary of key statistics, including the mean, median, minimum, maximum, and quartiles, offering insights into the distribution and central tendency of square footage and total assessed value within the cleaned dataset.

```
> summary(data_clean$SqFt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    705    1192    1574    1726    2023    4650
> summary(data_clean$TotalPrice)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32184  122680  161998  261119  296585 4904102
```

Figure 1: Distribution of the two selected variables of interest

Analyzing the histograms of the two variables of interest provides a clear representation of the distribution of a continuous variable. It reveals the underlying shape of the distribution and assesses how common or rare specific ranges of values are in the dataset.

As it is observed in Figure 2 the distribution of the variable square footage is unimodal and right skewed. This indicates that the mean will be greater than the median, and both will be to the right of the mode which is depicted in Figure 1. It also has some outliers present that deviate from the dataset and contribute to the rightward skewness. As the presence of outliers strongly influences the mean, the median is a better representation of the central tendency.

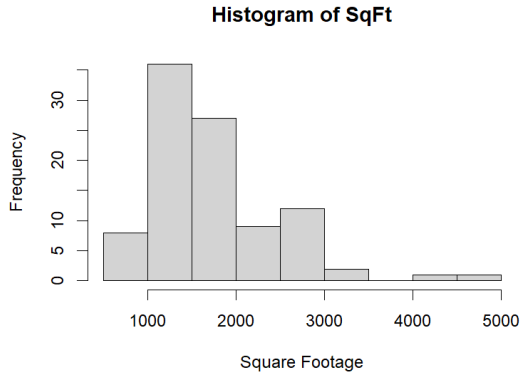



Figure 2: Histogram of Square footage

Similarly, As it is observed in Figure 3 the distribution of the total price variable is also unimodal and right skewed. This indicates that the mean will be greater than the median, and both will be to the right of the mode which is depicted in Figure 1. It also has some outliers present that deviate significantly from the dataset and contribute to the extreme rightward skewness.

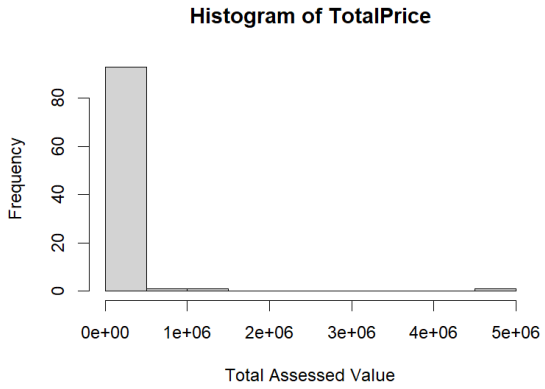



Figure 3: Histogram of Total Price

Boxplots are also valuable tools that offer a quick summary of the distribution of data, presenting key statistics such as the median, quartiles, and potential outliers. The position of the box in the boxplot corresponds to the median, providing a clear indicator of the central tendency of the data. The length of the box in the boxplot represents the interquartile range (IQR), offering insights into the spread or variability of the data. They also clearly depict outliers as individual points beyond the "whiskers" or lines extending from the box.

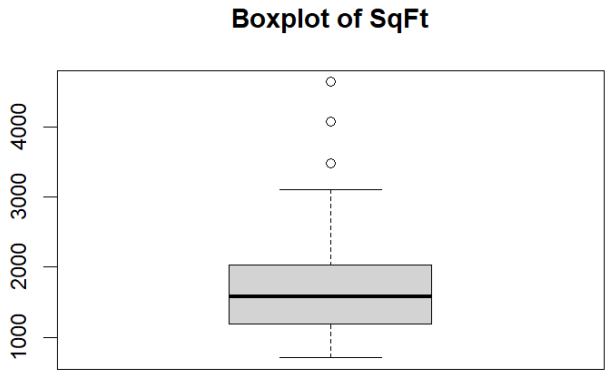



Figure 4: Boxplot of Square footage

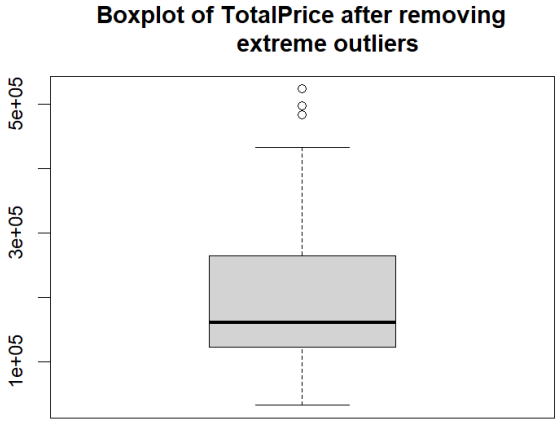
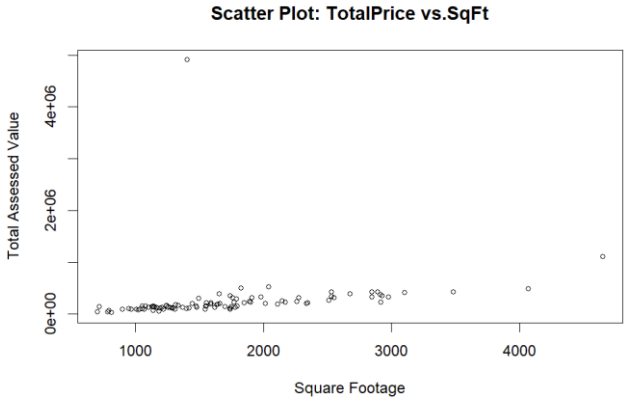



Figure 5: Boxplot of Total Price

Figure 4 indicates the presence of 3 outliers in square footage variables. As seen in Figure 3, the outlier is so large that the data is extremely close together and it is hard to visualize the distribution of the variable. There is a huge spread in the x-axis caused by the presence of one outlier. To obtain the Box Plot in Figure 5 a modified data frame was created in which the extreme outliers of the original data were emitted. This helped improve the visualization of the distribution.



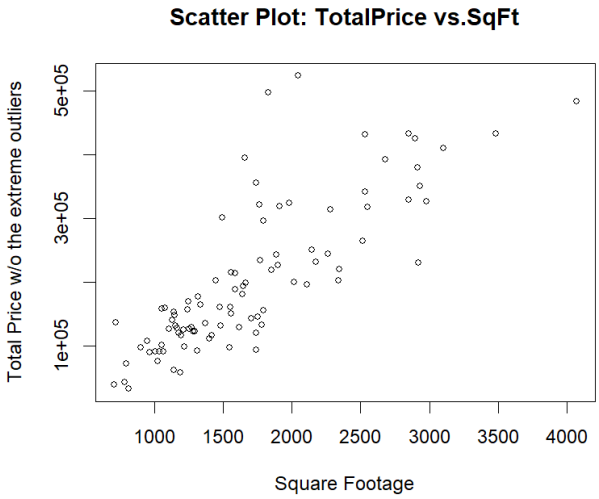



Figure 6: Scatter Plot of Total price against Sqft.

Scatter plots are particularly useful for understanding the relationship between two continuous variables. As observed in the plot present at the top in Figure 6 which is a Scatter plot of the variable total price against the variable square footage, there is a general upward trend from the bottom-left to the top-right, suggesting a positive linear correlation between the two variables. There is also a strong correlation between the two as the pattern is tightly clustered. A single data point that deviates significantly from the overall pattern may be affecting the scatter plot. To check the relation between the two variables another scatterplot is created which doesn't contain the extreme values. It can be seen that although the relationship between the two variables stays positive and linear, they are no longer strongly related.

Overall, the findings of the analysis reveal a positive correlation between Square Footage (SqFt) and Total Assessed Value (TotalPrice) in the real estate dataset. Larger square footage tends to be associated with higher assessed values, indicating the influence of property size on financial valuation. The distributions of both variables are unimodal and right-skewed, with outliers present, suggesting the presence of properties with unique characteristics impacting their assessed values. The boxplots and histograms highlight these distributional characteristics, while the scatter plot reinforces the observed positive linear correlation.

## 4. Price Analysis: Effect of the Presence of a Fireplace:

This section examines whether the presence of a fireplace influences property prices. The consideration of amenities, such as a fireplace, has been known to impact the perceived value and desirability of a property. To comprehensively explore this relationship, we will investigate whether properties with a fireplace differ in terms of their total assessed prices.

First, the cleaned dataset is divided into two subsets using logical conditions based on the presence or absence of a fireplace. The data_clean_fireplace subset contains observations where the variable "Fireplace" is true, indicating the presence of a fireplace in the property.

Conversely, the data_clean_no_fireplace subset includes observations where the variable "Fireplace" is false, indicating the absence of a fireplace.

The subsequent step involves generating summary statistics for the "TotalPrice" variable within each subset. The summary() function provides key statistical measures such as the mean, median, minimum, maximum, and quartiles. Analysing the summary statistics for "TotalPrice" in both subsets allows to discern any notable differences in the distribution of property prices based on the presence or absence of a fireplace.

```
summary(data_clean_fireplace$TotalPrice)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
89871  142576  201733  247783  320419 1113750
 summary(data_clean_no_fireplace$TotalPrice)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
32184   82387  100207  295199  145710 4904102
```

Figure 7: Summary of the data.

It can be observed in Figure 7 that the mean TotalPrice for properties with a fireplace is $247,783, while for properties without a fireplace, it is higher at $295,199. This suggests that, on average, properties without a fireplace have a higher assessed value. The range between the minimum and maximum TotalPrice is wider for properties without a fireplace, indicating greater variability in assessed values within this group. The median TotalPrice is lower for properties without a fireplace ($100,207) compared to those with a fireplace ($201,733). The upper quartile (3rd Qu.) for properties without a fireplace is also lower, indicating that a significant proportion of these properties have lower assessed values. The maximum TotalPrice for properties without a fireplace is notably higher ($4,904,102) than for those with a fireplace ($1,113,750). This suggests the presence of potential outliers, indicating a few properties without fireplaces may have exceptionally high assessed values. The summary data suggests a notable difference in the distribution of TotalPrice between properties with and without fireplaces. While properties with fireplaces have a lower mean assessed value, they have higher variability and upper quartile in comparison to properties without fireplaces, indicating a potentially wider range of assessed values and a group of higher-valued properties without fireplaces.


Figure 8: Side by Side boxplot to compare the distribution of "TotalPrice" for properties with and without a fireplace.

Figure 8 depicts a side-by-side boxplot to visually compare the distribution of "TotalPrice" for properties with and without a fireplace. It excludes the values which were the extreme outliers in the original data set as their presence was making the graphs very hard to visualize and interpret. The overall box length and whisker spread indicate that the properties with fireplaces have a higher variability. This indicates that properties with fireplaces exhibit a broader range of assessed values. The presence of outliers in both groups implies that the mean may be influenced by these extreme values. Therefore, the median is considered a more robust measure of central tendency in this context. The median represents the middle value of the distribution and is less sensitive to the presence of outliers. The line inside the box being at a higher position, indicates that the properties with a fireplace have a higher median. The higher median suggests a tendency toward higher total prices within this subgroup.

Using the tally function, it is found that in the data set 72% of properties have fireplaces whereas the remaining 28% don’t. This indicates that fireplaces are a relatively common feature among the sampled residential properties. This can be visualized in the bar chart below.

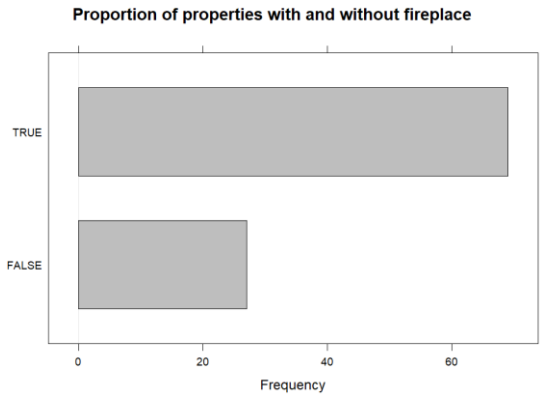



Figure 9: Distribution of properties with and without fireplace.

Overall, there is evidence to suggest that the presence of a fireplace is related to property price. The higher median total price in houses with fireplaces supports this association, emphasizing the potential impact of this amenity on the perceived value of a property. However, the presence of outliers and the variability in both groups also highlight the influence of other factors on property prices. Hence, we cannot establish a causal relationship between the presence of a fireplace and the total price of the property due to the presence of confounding variables but we can say that they are associated with each other.

## 5. Numerical Relationship Exploration:

Continuous variables are those that can take an infinite number of values within a given range. These variables are measured and can have fractional values. They are characterized

by an unbroken continuum and can be subdivided into smaller and smaller units without limit. Continuous variables are typically associated with measurements rather than counts.

The continuous variables present in this data set are:

- SqFt.: The area of the floor plan in square feet (in square feet).
- Acres: How many acres are included in the property.
- TotalPrice: The total assessed value of the property (in dollars).
- LandPrice: The assessed value of the land (in dollars).
- BuildingPrice: The assessed value of the building (in dollars).

To explore the potential relationship between the different continuous variables, we have used a correlation matrix. The values range from -1 to 1, indicating the strength and direction of the linear relationship between the variables.
A positive value close to 1 indicates a strong positive correlation (as one variable increases, the other tends to increase).
A negative value close to -1 indicates a strong negative correlation (as one variable increases, the other tends to decrease).
A value close to 0 suggests a weak or no linear correlation.

| | SqFt | Acres | TotalPrice | LandPrice | BuildingPrice |
|---|---|---|---|---|---|
| SqFt | 1.00000000 | -0.02391775 | 0.2003797 | 0.04440322 | 0.90864763 |
| Acres | -0.02391775 | 1.00000000 | 0.9574992 | 0.98805174 | -0.01678308 |
| TotalPrice | 0.20037973 | 0.95749918 | 1.0000000 | 0.98500166 | 0.24596406 |
| LandPrice | 0.04440322 | 0.98805174 | 0.9850017 | 1.00000000 | 0.07503101 |
| BuildingPrice | 0.90864763 | -0.01678308 | 0.2459641 | 0.07503101 | 1.00000000 |

Figure 10: Correlation matrix to explore the relation between different continuous variables.

The correlation matrix reveals insights into the relationships among key variables in the dataset. Square footage (SqFt) exhibits a weak positive correlation with both the total assessed price (TotalPrice) and the assessed value of the land (LandPrice), indicating that larger properties tend to have higher values. Interestingly, there is a strong positive correlation between square footage and the assessed value of the building (BuildingPrice), emphasizing that larger square footage is associated with higher building values. A notable negative correlation is observed between acres and the assessed value of the building, though this relationship is weak. Acres and total assessed price exhibit a very strong positive correlation, implying that larger land areas are strongly associated with higher property values. The relationship between the total assessed price and the assessed value of the land is also very strong. However, the correlation between the total assessed price and the assessed value of the building is weaker, suggesting that while land value strongly contributes to the overall assessed price, the building value has a comparatively smaller impact.

### 5.1 Building Price vs Square Footage

It is speculated that the pair of continuous variables "Square Footage" (SqFt) and "Building Price" may exhibit a discernible relationship. The rationale behind this speculation is supported by the strong positive correlation coefficient of 0.909 observed in the correlation matrix. A correlation coefficient close to 1 indicates a high positive linear relationship between the two variables. In this context, it implies that as the square footage of a property increases, there is a corresponding increase in the assessed value of the building. This aligns with the intuitive understanding that larger properties often entail more substantial structures or features, contributing to higher building values. The Scatter Plot of Building Price against Square Footage (Figure 11) further underscores this correlation, depicting a clear positive linear trend and a strong association between the two variables. This statistical evidence substantiates the hypothesis of a discernible relationship, indicating that the size of a property, as measured by square footage, plays a significant role in influencing the assessed value of the building.

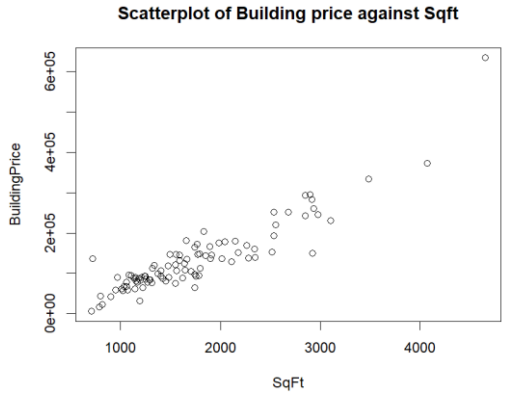



Figure 11: Scatter plot of Building Price against Square footage.

## 6. Linear Regression Analysis:

Building on insights gained from preliminary explorations and correlation assessments, the chosen variables for this analysis are "Square Footage" (SqFt) and "Building Price." The linear regression model generated herein aims to elucidate the quantitative association between the size of a property, as measured by square footage, and its assessed building value.

The linear regression analysis focuses on exploring the relationship between "Square Footage" (SqFt) and "Building Price." The estimated coefficients from the model reveal that the intercept is -$52,514.780, indicating the expected building price when the square footage is zero. Since a property's square footage cannot be zero, the intercept does not have a practical interpretation in this context. The coefficient for SqFt is 107.798, suggesting that, on average, the building price increases by 107.798 units for each additional square foot.

The linear regression model for the relationship between "Square Footage" (SqFt) and "Building Price" is expressed as follows:

$$\textit{Building Price} = -52{,}514.780 + 107.798\ X\ \textit{SqFt}$$

The above equation showcases the quantitative association between square footage and building price.

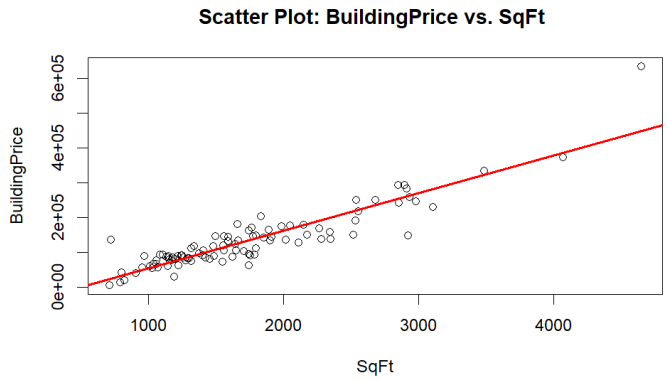



Figure 12: Scatter plot of Building Price against Square footage and overlaying regression line.

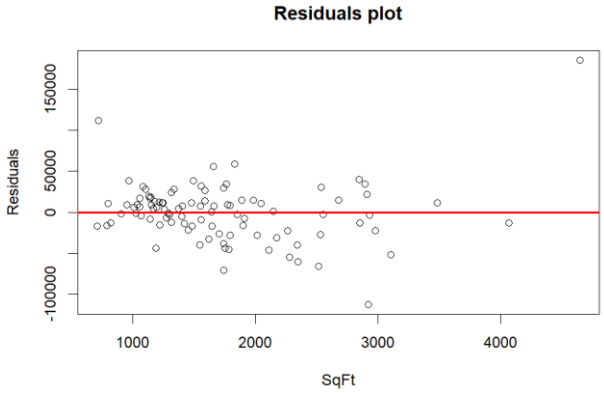



Figure 13: Residuals Plot.

The linearity condition is met as observed in Figure 11. The two variables have a linear relationship since most of the data is close to the regression line. Based on the residuals plot in Figure 12, it seems like the residuals are not scattered symmetrically across the zero line (positive and negative residuals do not line up with one another). So, the symmetry condition is not met. Moreover, looking at the Building Price vs SqFt plot the points in the bottom left corner are more closely clustered than the points in the top right. Also, in the residuals plot, we can see that the residuals are close to zero on the left and more spread out on the right. The x-value is directly proportional to variance. As we traverse along the x-axis, we see an increase in residual variance. This indicates that the variance is not equal across all values of the explanatory variable.

The coefficient of determination found is 0.8256. It represents the proportion of variance in the response variable (BuildingPrice) that is explained by the explanatory variable (SqFt). In this case, approximately 82.56% of the variance in BuildingPrice is explained by the size of the property (SqFt). A higher R-squared indicates a better fit of the model to the data.

Although two out of the three assumptions are not fully met, they are not violated to a great extent. Hence, the linear regression model is a good fit for data.

## 7. Conclusion:

In conclusion, this statistical analysis of a real estate dataset offers insights into the dynamics of residential properties within a specific County registry. The exploration began with data cleaning, ensuring the dataset's integrity by handling missing values. Variable summarization highlighted the significance of Square Footage (SqFt) and Total Assessed Value (TotalPrice) in understanding property features and economic landscapes. The investigation into the effect of the presence of a fireplace on property prices unveiled an association, emphasizing the potential impact of amenities on perceived property value.

Numerical relationship exploration through a correlation matrix provided an understanding of the interplay between continuous variables. The focus on the SqFt and Building Price relationship, supported by a strong positive correlation, was further validated through a linear regression analysis. The model's estimated coefficients and goodness-of-fit metrics revealed a quantitative association, explaining approximately 82.56% of the variance in Building Price based on SqFt.

The analysis illuminated diverse facets of the real estate dataset, from property size and amenities to the relationships among continuous variables. While the presence of outliers and variability poses challenges, it adds richness to the dataset, emphasizing the need for a nuanced interpretation. Overall, this statistical paper contributes to a deeper understanding of real estate dynamics, providing insights for stakeholders and researchers alike. Further exploration and refinement of analyses could unravel additional layers of complexity within the dataset, fostering continuous improvement in real estate research methodologies.

## 8. Appendix:

```
data <- read.csv("./data.csv")

# Section 2

# Detect any missing values within the dataset and provide
#   the IDs of observations with missing data.
data <- read.csv("./data.csv")
is.na(data)
sum(is.na(data))
sum(is.na(data$ID))
sum(is.na(data$YearBuilt))
sum(is.na(data$SqFt))
sum(is.na(data$Story))
sum(is.na(data$Acres))
sum(is.na(data$N_Baths))
sum(is.na(data$Fireplace))
sum(is.na(data$TotalPrice))
sum(is.na(data$LandPrice))
sum(is.na(data$BuildingPrice))
sum(is.na(data$Zipcode))

# Remove observations with missing data before proceeding with the analysis.
# Remove observations with missing data
data_clean <- na.omit(data)



#
#Ensure the dataset is clean and ready for analysis.
# Check for missing values after removal
sum(is.na(data_clean))

# Check for missing values in the entire dataset
missing_values <- data[!complete.cases(data), ]

# Get the IDs of observations with missing data
missing_ids <- missing_values$ID
missing_ids



##section 3


#variables of interest
#Square Footage (SqFt): Square footage is a fundamental metric in real estate
#Sand can be indicative of the size of a property. Analyzing this variable may
#Sreveal trends in property sizes within the dataset.

#STotal Assessed Value (TotalPrice): The total assessed value of a property is
#Sa key financial metric. It encompasses both the land and building values,
#providing an overall valuation. Analyzing this variable can help understand
#the distribution of property values in the dataset.

##
```

```r
#Summarize each variable with relevant statistics and graphical tools.
summary(data_clean$SqFt)
summary(data_clean$TotalPrice)


# Histogram for SqFt
hist(sqft, main = "Histogram of SqFt", xlab = "Square Footage")

# Histogram for TotalPrice
hist(total_price, main = "Histogram of TotalPrice",
     xlab = "Total Assessed Value")

# Scatter plot

plot(sqft, total_price, main = "Scatter Plot: TotalPrice vs.SqFt",
     xlab = "Square Footage", ylab = "Total Assessed Value", cex=0.75)
#scatter plot without the extreme outlier in the total prive variable

plot(data_clean_tp$SqFt, data_clean_tp$TotalPrice, main = "Scatter Plot: TotalPrice vs.SqFt",
     xlab = "Square Footage", ylab = "Total Price w/o the extreme outliers", cex=0.75)

# IS THE DISTRIBUTION UNIMODAL, SKEW, OUTLIERS

tally(~sqft | total_price, data = NCbirths, format = "proportion")

#Identify any unusual data.
boxplot

##code to check if the graphical representation of total price variable becomes
#easier to read after removing the outliers

# Calculate the interquartile range (IQR)
total_price_iqr <- IQR(data_clean$TotalPrice)

# Calculate the first quartile (Q1)
q1 <- quantile(data_clean$TotalPrice, 0.25)

# Calculate the third quartile (Q3)
q3 <- quantile(data_clean$TotalPrice, 0.75)

# Calculate the lower whisker
lower_whisker <- q1 - 1.5 * total_price_iqr
lower_whisker
# Calculate the upper whisker
upper_whisker <- q3 + 1.5 * total_price_iqr
upper_whisker
# Select variables of interest
sqft <- data_clean$SqFt
total_price <- data_clean$TotalPrice

# Remove rows where TotalPrice is greater than 557443.9
data_clean_tp <- data_clean[data_clean$TotalPrice <= 557443.9, ]

# Check the first few rows of the modified dataframe
head(data_clean_tp)
boxplot(data_clean_tp$TotalPrice, main = "Boxplot of TotalPrice after removing
        extreme outliers")
```

```r
##  section 4
#Price Comparison: Does the presence of a fireplace make a difference?



# Create data_clean_fireplace only for observations where Fireplace is true
data_clean_fireplace <- subset(data_clean, Fireplace == TRUE)

# Create data_clean_no_fireplace only for observations where Fireplace is false
data_clean_no_fireplace <- subset(data_clean, Fireplace == FALSE)

# Check the first few rows of the new dataset
head(data_clean_fireplace)
head(data_clean_no_fireplace)

#We check the summary to see the difference in trend of total price
summary(data_clean_fireplace$TotalPrice)
 summary(data_clean_no_fireplace$TotalPrice)

 #Create side-by-side boxplots to visually compare the distribution of
 #property prices for homes with and without a fireplace.

boxplot(TotalPrice ~ Fireplace, data = data_clean,
        main = "Side by Side Boxplot of TotalPrice by Fireplace", ylab = "Total Price")
## there is a very large outlier affecting the graph, remove that data point and rerun it
 # to see if the boxplot is more readable

boxplot(data_clean_tp$TotalPrice ~ data_clean_tp$Fireplace, data = data_clean,
        main = "Side by Side Boxplot of TotalPrice by Fireplace", ylab = "Total Price", xlab="Fireplace")
```

```r
install.packages("mosaic")
library(mosaic)
tally(data_clean$Fireplace, format = "percent")

##BAR CHART
barchart(data_clean$Fireplace, col="grey", main="Proportion of properties with and without fireplace", xlab="Frequency" )

#visualize the relationship between the presence of a fireplace and property prices.
plot(data_clean$Fireplace, data_clean$TotalPrice,
     xlab = "Fireplace Presence (0: No, 1: Yes)", ylab = "Total Price")
```

```r
                          #SECTION 5
                  #Numerical Relationship Exploration:
##CONTINUOUS VARIABLES
#should just use the class command

#SqFt, Acres, TotalPrice, LandPrice, BuildingPrice
## a discrete variable can go up and down by a fixed amount
## cont- go up and down by any amount
class(data_clean$ID)
class(data_clean$YearBuilt)
class(data_clean$SqFt)
class(data_clean$Story)
class(data_clean$Acres)
class(data_clean$N_Baths)
class(data_clean$Fireplace)
class(data_clean$TotalPrice)
```

```r
class(data_clean$BuildingPrice)
class(data_clean$Zipcode)


correlation_matrix <- cor(data_clean[c( "SqFt", "Acres", "TotalPrice", "LandPrice", "BuildingPrice")])
correlation_matrix


#building price and sqft

plot(BuildingPrice~SqFt, data= data_clean, main="Scatterplot of Building price against Sqft")
cor(data_clean$BuildingPrice, data_clean$SqFt)


#section 6
##Produce a summary of a linear model of Building Price against SqFt
linear_model <- lm(BuildingPrice ~ SqFt, data = data_clean)
#Building Price is the response variable and SqFt is the explanatory variable.
summary (linear_model)
linear_model

##Plot the data and overlay the regression line.
plot(data_clean$SqFt, data_clean$BuildingPrice,xlab = "SqFt", ylab = "BuildingPrice",
     main = "Scatter Plot: BuildingPrice vs. SqFt ")
abline(linear_model, col= "red", lw=2)


##Plot the residuals of the model over sqft and include the zero line.
plot(linear_model$residuals ~ data_clean$SqFt,xlab="SqFt", ylab= "Residuals",
     main= "Residuals plot")
abline(a=0, b=0, col= "red", lw=2)
```