**Project Report**

**Objective:**

The project aims to provide a comprehensive understanding of the dataset through a series of analytical tasks. The goal is to uncover patterns, relationships, and insights within the real estate data.

1. **Detect and Handle Missing Data:**

In order to do detect the presence and total number of missing values within the dataset ‘data’, the is.na(data) and sum(is.na(data)) functions can be used. It can be observed that there are 13 missing values NA in the dataset and every variable except ‘TotalPrice’ have at least 1 missing value.

Further, data$ID[ (rowSums(is.na(data)) > 0)] can be used to retrieve the IDs corresponding to the rows with missing values. After running this, it can be seen that that IDs 1, 14, 51 and 80 contain observations with 1 or more missing values NA.

Finally, the na.omit(data) function can be used to remove observations with missing data to ensure the dataset is clean and ready for analysis.

1. **Variable Summarization:**

**Variable 1: SqFt**

As a continuous numerical variable, the statistical and visual summary of this data includes:

Minimum = 705, 1st Quartile = 1192, Median = 1574, Mean = 1726, 3rd Quartile = 2023, Maximum = 4650, Standard Deviation = 730.7355

A graph of a function

Description automatically generated with medium confidence

Interpretation of this distribution:

* Range: The range of SqFt is from 705 to 4650, indicating a substantial variability in property sizes within the dataset.
* Modality: The histogram suggests that the distribution of SqFt is unimodal, with one prominent peak indicating the most common range of property sizes between 500 to 1500 square feet.
* Quartiles: The 1st Quartile (25th percentile) is at 1192 square feet, indicating that 25% of the properties have a square footage less than or equal to this value. The Median (50th percentile) is at 1574 square feet, which signifies that half of the properties are below this square footage. The 3rd Quartile (75th percentile) is at 2023 square feet, meaning that 75% of the properties have a square footage less than or equal to this value.
* Central Tendency: The mean of the distribution is 1726 square feet, which is higher than the median (1574 square feet). This reinforces the observation that the distribution is right-skewed, with a tail of larger properties pulling the mean upwards.
* Variability: The standard deviation is approximately 731 square feet, which is quite large relative to the mean. This suggests there is a wide dispersion in property sizes, and the data points are spread out over a large range of values.
* Outliers: The maximum value of 4650 square feet is substantially higher than the mean and even the 3rd quartile. This suggests the presence of outliers on the higher end of the distribution, consistent with the long tail observed in the histogram. The properties that have a square feet value near 5000 could be considered outliers, as they are significantly larger than the majority of the observations.
* Interquartile Range (IQR): The IQR, calculated as 3rd Quartile minus the 1st Quartile, is 831 square feet. This range contains the middle 50% of the data and gives a sense of the typical spread of the central portion of the dataset.
* Skewness: The fact that the mean is greater than the median confirms the right skewness of the distribution, indicating a longer tail to the right. This is typical for data where a few high values stretch the mean.
* Gap in the Data: There appears to be a gap between 3000 and 4000 square feet in the histogram where no properties are present. This could indicate a sampling gap or a natural absence of properties within this range in the dataset.

**Variable 2: Fireplace**

As a categorical variable, the statistical and visual summary of this data includes:

FALSE = 27, TRUE = 69

A red rectangular bar graph

Description automatically generated

Interpretation of this distribution:

* Frequency Counts: The bar chart shows that there are 27 properties without a fireplace and 69 properties with a fireplace.
* Distribution Balance: A significantly higher number of properties feature a fireplace than those that do not. This suggests that having a fireplace may be a common feature in the properties within this dataset.
* Majority Preference: The presence of a fireplace in a majority of properties could imply a preference or a trend in the market where properties are being built with fireplaces, or it could suggest that properties with fireplaces are more likely to be put up for sale or rent.

**3. Price Comparison: Does the presence of a fireplace make a difference?**

The graphical and numerical summaries for data related to the comparison of property prices:   
Mean with fireplace = 247,782

Mean without fireplace = 295,198

Standard deviation with fireplace = 154,942

Standard deviation without fireplace = 923,479

A graph of a number of graphs

Description automatically generated with medium confidence

A graph of a graph with blue and red dots

Description automatically generated

Based on the results from the histograms and the scatterplot, combined with the numerical summaries provided, the analysis and interpretation regarding whether the presence of a fireplace is related to property price is as follows:

Histogram:

The histogram shows the distribution of property prices, segmented by the presence (TRUE) or absence (FALSE) of a fireplace. The prices are scaled down to millions for readability.

* For properties with a fireplace (TRUE), the distribution of prices is concentrated in the lower price range, with the majority falling below 0.5 million.
* For properties without a fireplace (FALSE), there is a significant spike at the lowest price range (just above 0 million), suggesting that a substantial portion of these properties are at the lower end of the market.

Scatterplot:

The scatterplot visualizes individual property prices, with blue dots representing properties with a fireplace and red dots representing properties without.

* The data points for both categories are dispersed throughout the price spectrum, suggesting variability in property prices regardless of fireplace presence.
* Despite this variability, properties featuring a fireplace tend to group towards the lower end of the price range, indicating that a fireplace may not substantially increase property value.
* Conversely, properties lacking a fireplace show a wider price distribution, including numerous properties at substantially higher price points, which may imply that factors other than a fireplace are influencing the higher property prices.

Numerical Summaries:

* The mean price for properties with a fireplace is approximately $247,782, while it is about $295,198 for properties without a fireplace.
* The standard deviation for properties with a fireplace is much lower ($154,942) compared to properties without a fireplace ($923,479), indicating more variability in the prices of properties without a fireplace.

Insights:

* Based on the mean values, properties without a fireplace tend to have a higher average price compared to those with a fireplace. This is somewhat counterintuitive since one might expect properties with a fireplace to be priced higher due to the added feature.
* The large standard deviation for properties without a fireplace suggests that this category has a wider range of property prices, including both lower-end and higher-end properties. It's important to consider other factors that might influence property prices, such as location (indicated by zip code), size, and age of the building, which could confound the relationship between fireplace presence and property price.
* The concentration of properties with a fireplace at the lower end of the price spectrum may suggest that fireplaces are more common in less expensive homes or that the value added by a fireplace does not significantly impact the overall price

Interpretation:

While the presence of a fireplace might be thought to contribute to a higher property price, the data does not show a clear trend supporting this. In fact, properties without a fireplace have a higher average price in this dataset. However, it's critical to perform a more comprehensive analysis that considers other influencing factors to draw a definitive conclusion about the impact of a fireplace on property prices.

**4. Numerical Relationship Exploration:**

The following are the continuous variables in the dataset:

- Story, SqFt, Acres, TotalPrice, LandPrice, BuildingPrice

There are several potential relationships among these variables, such as:

* SqFt and TotalPrice: Larger homes (more square footage) often have higher prices.
* Acres and TotalPrice: More land often correlates with a higher price.
* BuildingPrice and TotalPrice: The value of the building is a component of the total price

For this project, the pair of continuous variables ‘SqFt and TotalPrice’ may have a discernible relationship and can be further analysed. The rationale is that typically, the larger the home (in terms of square footage), the higher the price, assuming other factors are constant.

A graph with blue dots

Description automatically generated

Based on the scatter plot, the following interpretation can be made:

* Positive Correlation: The plot suggests a positive correlation between the square footage of a property and its total price, which is expected in real estate markets. Properties with more square footage generally command higher prices.
* Data Distribution: The concentration of data points towards the lower end of both axes indicates that a larger number of properties have smaller square footage and are priced lower. As the square footage increases, the number of properties seems to decrease.
* Variability in Price: There is variability in the total price for properties within similar square footage ranges. This suggests that while size is a factor, there are likely other significant factors affecting the total price, such as location, quality of construction, age, and amenities.
* Outliers: There are a few properties with higher total prices that deviate from the general trend, especially in the mid-to-high square footage range. These could be luxury properties, properties in highly desirable locations, or properties with features that are highly valued in the market.
* Unusual Patterns: The vertical 'stacking' of points at certain square footage values suggests that there are standard property sizes at which multiple properties are priced differently. This 'stacking' effect might also indicate the presence of multiple properties with the same square footage but varying other features, leading to a range of prices.

**5. Linear Regression Analysis:**

Utilizing insights from the two selected variables SqFt and TotalPrice, the following graphs conduct a linear aggression analysis:

A graph of a number of blue dots

Description automatically generated with medium confidence

A graph with blue dots

Description automatically generated

The regression model can be represented by the equation:

Total Price = 242.32 + 0.84 × Square Footage

The interpretation of the estimated coefficients are:

* The constant term (intercept value) is approximately 242.32. This represents the estimated starting value of the total price when the square footage is zero, although this is a theoretical value since a square footage of zero is not practical in this context.
* The coefficient for Square Footage is approximately 0.84, suggesting that for every additional square foot, the total price increases by an estimated 0.84 units.

The goodness of fit of a regression model can be assessed using the metrics of R-squared value and the residual plot:

1) R-squared Value (0.859):

* The R-squared value is a measure of how well the regression model fits the data. It is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).
* An R-squared value of 0.859 means that approximately 85.9% of the variation in the total price can be explained by the square footage of the properties in the model. This suggests a strong relationship between the two variables; however, it doesn't mean the model is perfect. An R-squared value closer to 1 would indicate a better fit.
* It's also important to note that a high R-squared value does not necessarily imply causation or that the model will make accurate predictions.

2) Residual Plot:

* A residual plot shows the residuals (the differences between observed and predicted values) on the vertical axis and the independent variable (square footage) on the horizontal axis.
* Ideally, if the model is a good fit, the residuals should be randomly scattered around the horizontal axis (zero line), with no discernible pattern. This indicates that the variance is constant and that the model is capturing all the relevant information.
* In the residual plot above, it can be observed that the residuals do not form any particular pattern as they are relatively scattered around the zero line. However, there are a few points with larger residuals, which might be outliers or leverage points. This suggests that while the model generally fits well, there may be a few instances where it does not predict accurately.
* Based on the R-squared value and the residual plot, the model appears to have a good fit to the data, but one must be cautious about outliers or influential points that could affect the model's performance.

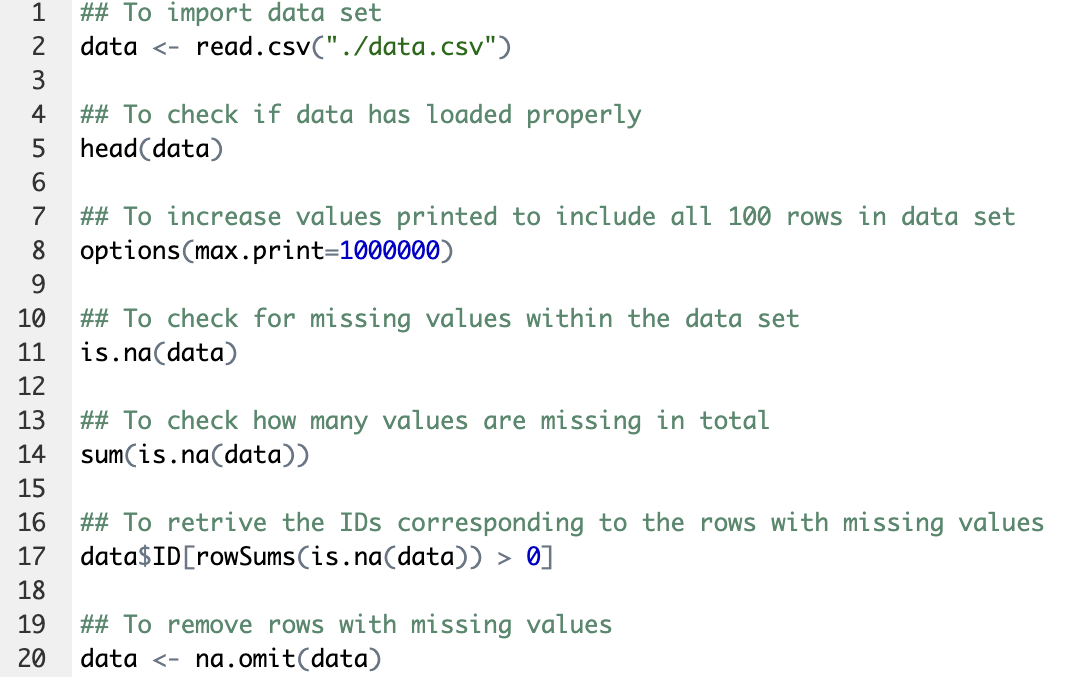
**Conclusion:**

The project's final analysis provides a detailed understanding of the real estate market, specifically how property characteristics like square footage and the presence of amenities such as fireplaces can influence market prices. After a thorough clean-up of the data, the key variables were summarized and a linear regression model was developed that accounted for a significant portion of price variations—about 85.9%. The model showed a strong correlation between a property's size and its price. The high R-squared value confirmed the model's accuracy, though it also pointed out some outliers. This study not only clarified trends in property valuation but also offered a solid approach for data-driven decision-making in the real estate field.

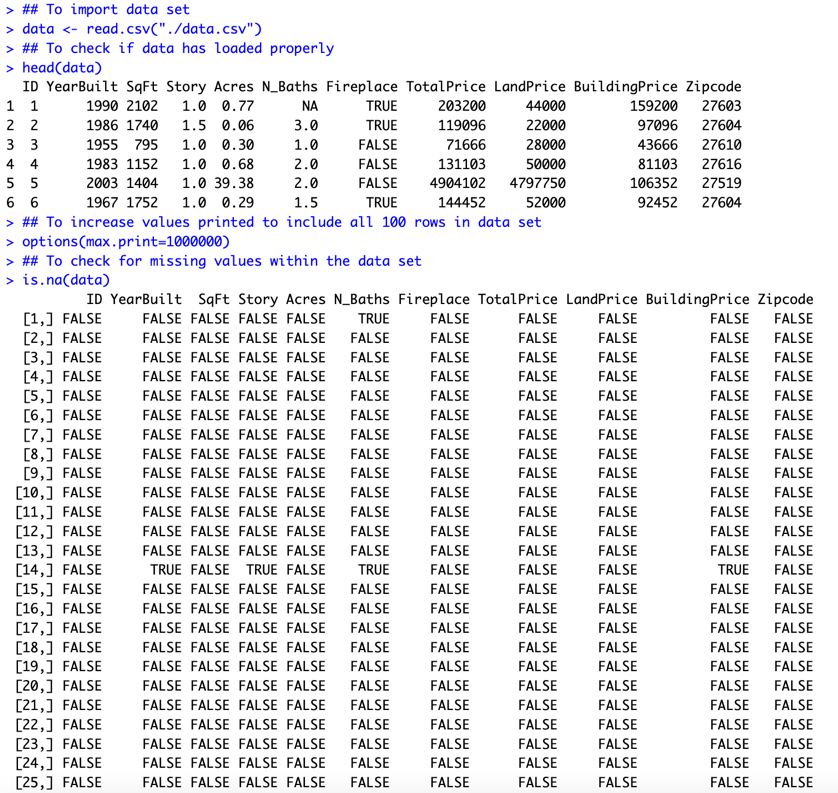
**APPENDIX (R CODE)**

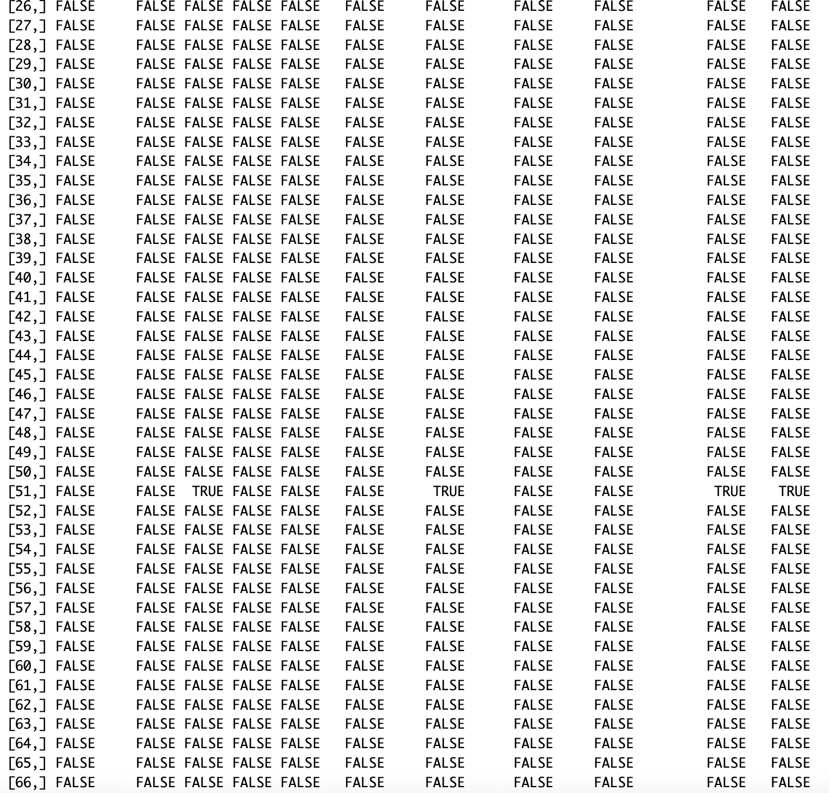
Task 1:

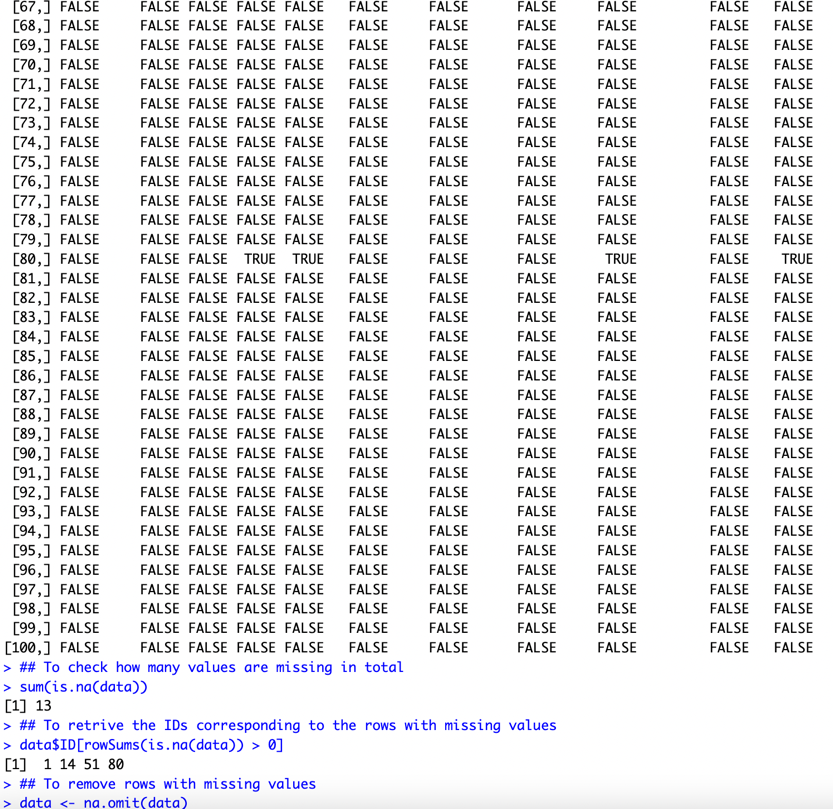
**Input**



**Output**







Task 2:

**Input**

A computer screen shot of a blue and green code

Description automatically generated

**Output**

A white background with blue text

Description automatically generated

🡪 Graphs in main report

Task 3:

**Input**

**A screenshot of a computer program

Description automatically generated**

**Output**

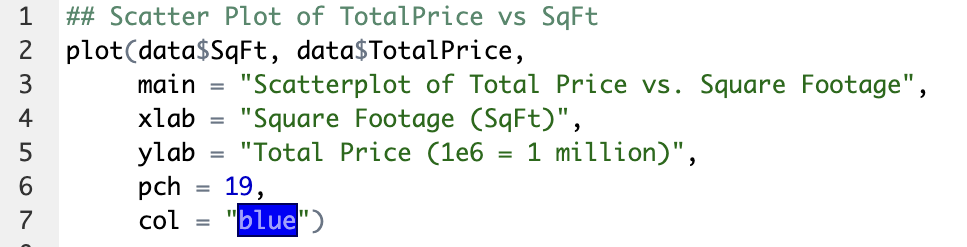
**A screenshot of a computer

Description automatically generated**

🡪 Graphs in main report

Task 4:

**Input**

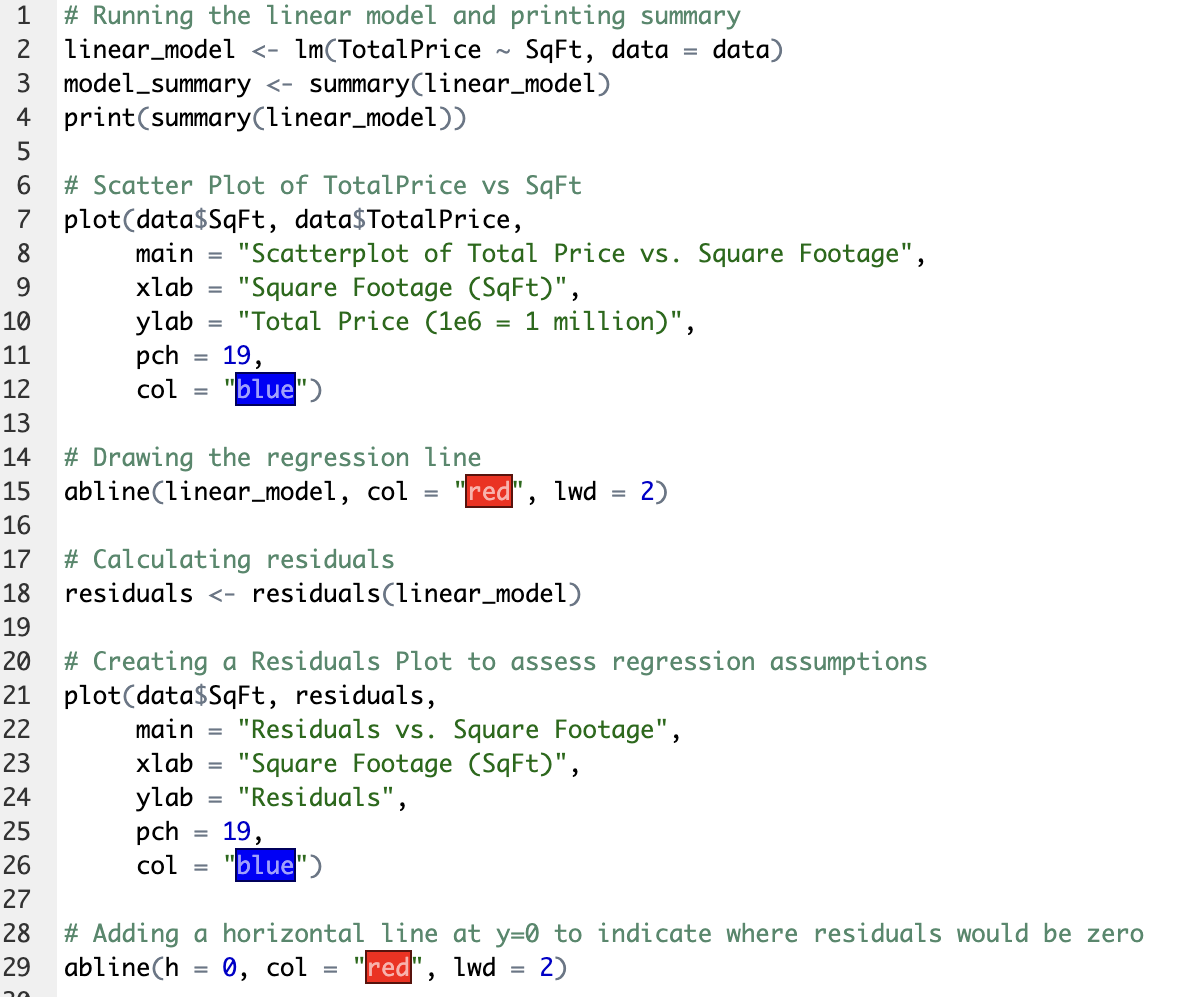
****

**Output**

🡪 Graph in main report

Task 5:

**Input**



**Output**

🡪 Graphs in main report