INFO589_FA24 BUSINESS STATISTICS

A STUDY ON

RANDOM SAMPLE OF 90 HOME CHARACTERISTICS IN 3 LONG ISLAND COMMUNITIES

FINAL PROJECT TEAM-9

- Jack Cleary
- Sumaiya Ibrahim
- Teja Sree Budeti
- Vishnu Sathwik Basavaraju

Exploratory Data Analysis

Distribution of Variables:

The distribution of the numerical variables, AppraisedValue, Land (acres), House Size (sq ft), Rooms, and Bathrooms, is visualized mostly showing a relatively normal or right-skewed distribution.

The categorical variable Town Label indicates that the prevailing label is "Glen Cove or Roslyn," amounting to about two-thirds of the observations.

Outliers:

Box plots show the potential presence of outliers in variables such as AppraisedValue, House Size (sq ft), and Land (acre) in this dataset.

For example, larger-than-expected values for land size and house size suggest that some properties might be much larger in size compared to other properties.

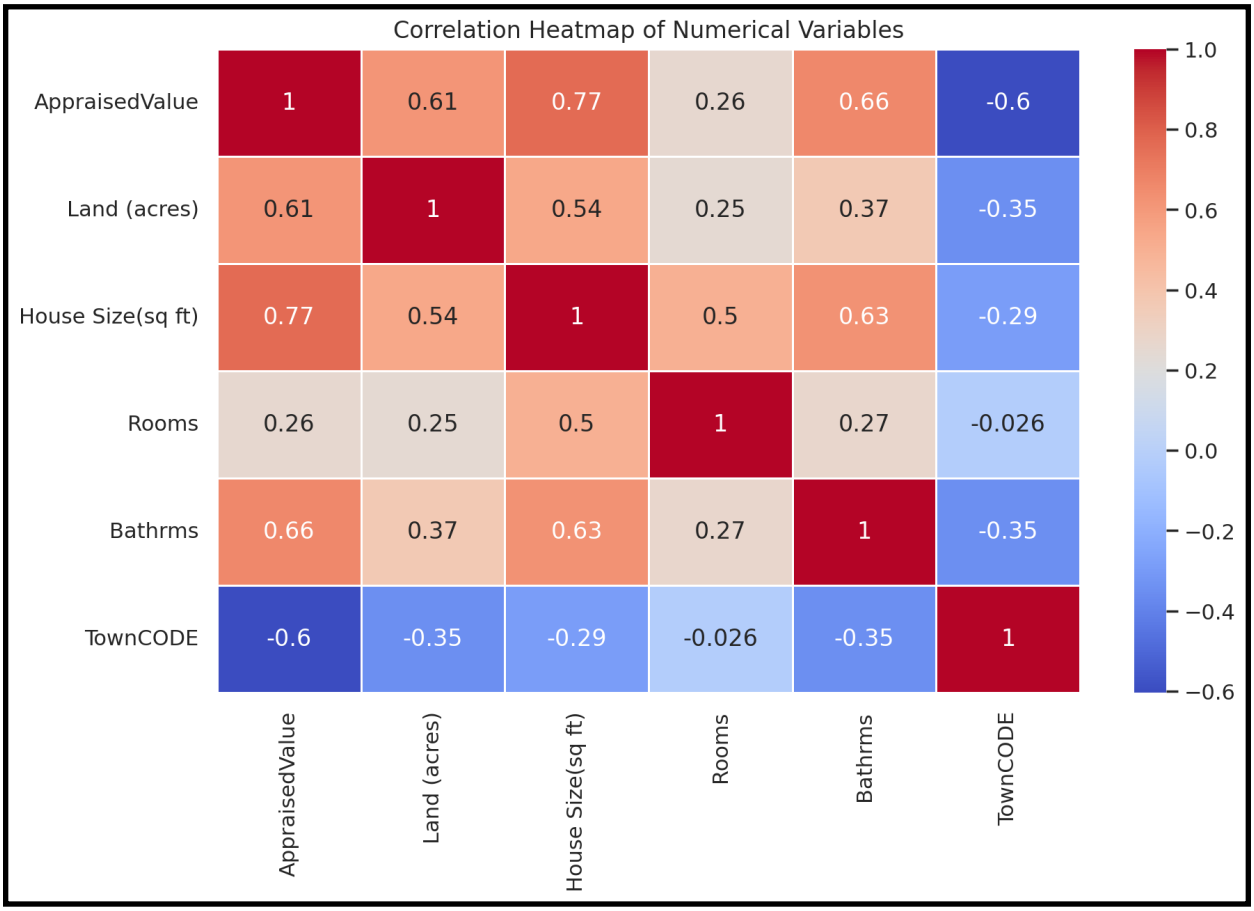
Hypothesis Testing:

ANOVA AppraisedValue by Town: The one-way ANOVA test between AppraisedValue and Town CODE yielded a p-value of about 3.04e-10, meaning that the appraised values of the two towns are very significantly different from each other.

Chi-Square Test Rooms - Town CODE:

The p-value from this was 0.218, which suggests no meaningful dependence of the number of rooms on town code.

These provide a number of interesting general patterns, such as the divergence in appraised value between towns and the identification of anomalies that may be worth further investigation.



Overview of the Dataset:

1. Number of Observations (Rows): 90

2. Number of Variables (Columns): 7

Data Nature:

Continuous Variable(s):

AppraisedValue: Appraised value of the homes, in thousands of dollars, ranging from 211.8K to 968.2K.

Land (acres): Land area of the properties, in acres, ranging from 0.023 to 0.655 acres.

• House Size (sq ft): Size of the houses in square feet varies from 961 sq ft to 4067 sq ft.

• Rooms: No. of rooms in the homes vary from 4 to 13.

• Bathrms: No. of bathrooms in the homes vary from 1 to 4.5 bathrooms.

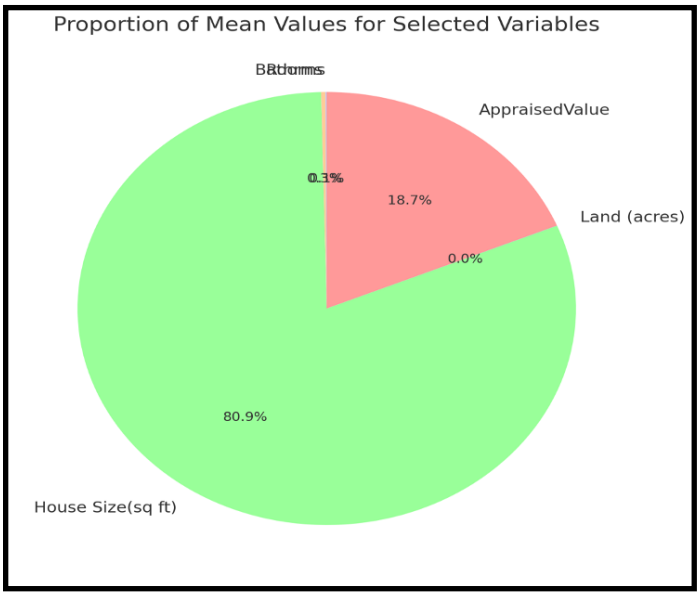
• Town CODE: Numerical code for different towns represented by 0 and 1

Categorical Variable:

• Town_Label: This is a categorical variable. Predefined name of the town. There are two categories where the "Glen Cove or Roslyn" contains 60 of the total number of observations, 90, and the rest 30 for another town. For this dataset, there are six numerical variables and one categorical variable that is describing features of homes and their respective locations, which is represented by the (`Town_Label`).

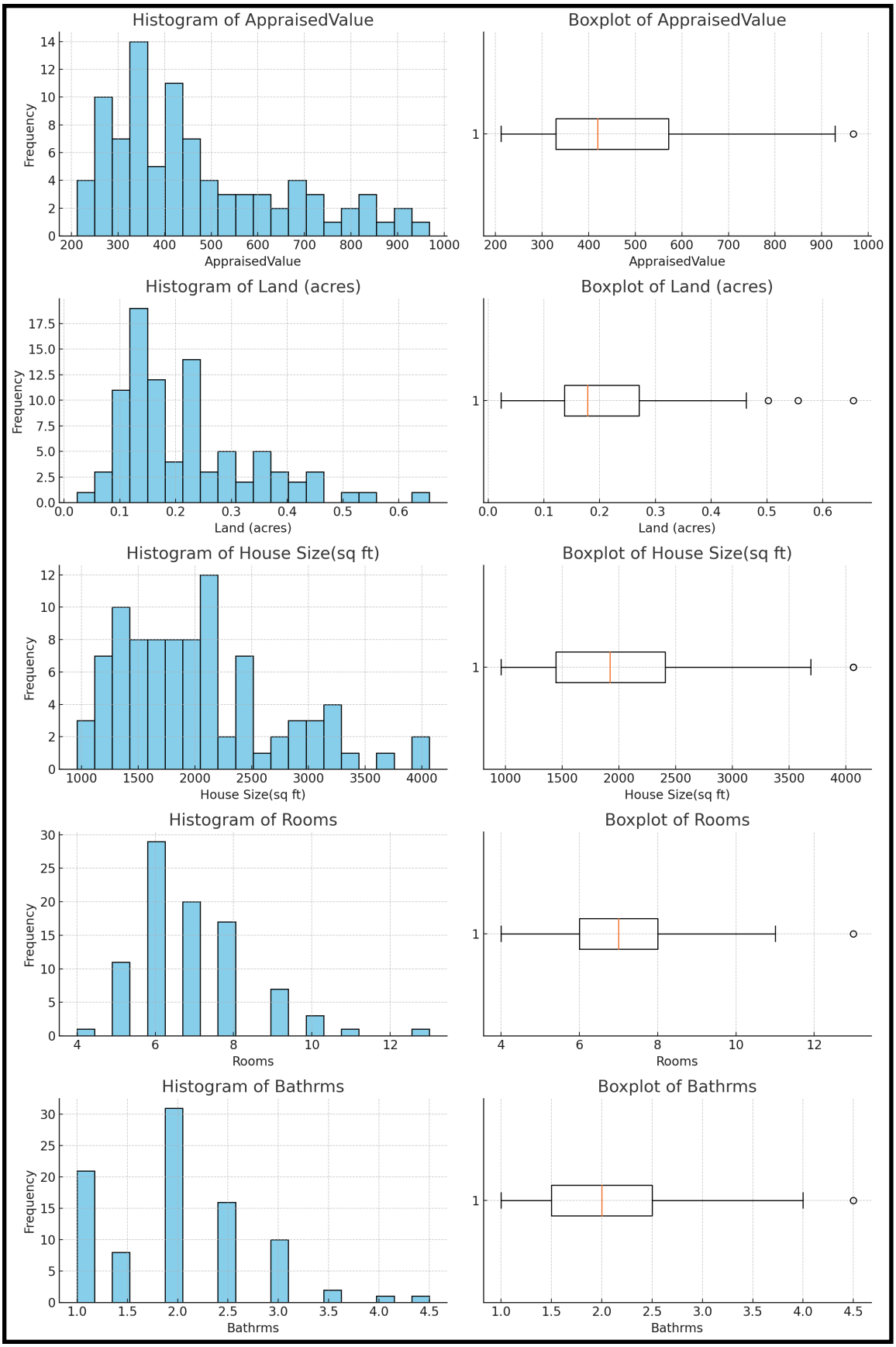
The data include 90 observations of several numerical variables: AppraisedValue, Land (acres), House Size (sq ft), Rooms, Bathrms, and Town CODE. The average appraised value of the homes is about $467.5K with a standard deviation of $187.7K, hence showing variability in property values. Land size averages 0.22 acres, ranging from as low as 0.023 acres to a maximum of 0.655 acres. The size of these houses is also highly varied, averaging 2017 square feet, ranging from 961 sq ft to 4067 sq ft.

Number of rooms ranges from 4 to 13, the average number is about 7. Number of bathrooms ranges from 1 to 4.5, the average being about 2 bathrooms. The `Town CODE` is relatively equitably spread with a mean of 0.33; that would be around one-third of the homes being from one town and two-thirds from another. The breakdown of the quartiles provides more information that helps to understand how these variables are spread and distributed across the dataset.



The following pie chart depicts the distribution of mean values for the five key variables, namely Appraised Value, Land in acres, House Size in square feet, Rooms, and Bathrooms in the data. Of these, maximum contributions from Appraised Value and House Size in square feet are brought forth since these two variables are the main ones in the whole dataset.

The number of Rooms and Bathrooms consists of smaller portions but still constitutes a meaningful role within the overall composition. Land size in acres is the smallest, meaning the average plot size is rather modest compared to all other variables. This chart helps visualize the balance between these different factors and how they contribute to the characteristics of the properties in the study.


Histogram of AppraisedValue
Frequency
AppraisedValue
Boxplot of AppraisedValue
AppraisedValue
Histogram of Land (acres)
Frequency
Land (acres)
Boxplot of Land (acres)
Land (acres)
Histogram of House Size(sq ft)
Frequency
House Size(sq ft)
Boxplot of House Size(sq ft)
House Size(sq ft)
Histogram of Rooms
Frequency
Rooms
Boxplot of Rooms
Rooms
Histogram of Bathrms
Frequency
Bathrms
Boxplot of Bathrms
Bathrms

The following are the visualizations undertaken for various distributions of the numerical variables. The visualizations include the histograms and boxplots.

Appraised Value: In this histogram, the basic distribution appears as a right-skewed distribution. Most values are lying within the range of 300K to 600K. However, the boxplot is showing few outliers above 900K. Central tendency around 420K.

Land in acres: This is also right-skewed. The majority of values are less than 0.3 acres. There are a few outliers above 0.5 acres. The central tendency is around 0.18 acres.

House Size sq feet: House sizes appear to be a bit right-skewed. As mentioned earlier, most houses lie in the range from 1,500 to 2,500 sq ft. The boxplot confirms that there are a few larger houses, which lie above 4,000 sq ft as outliers.

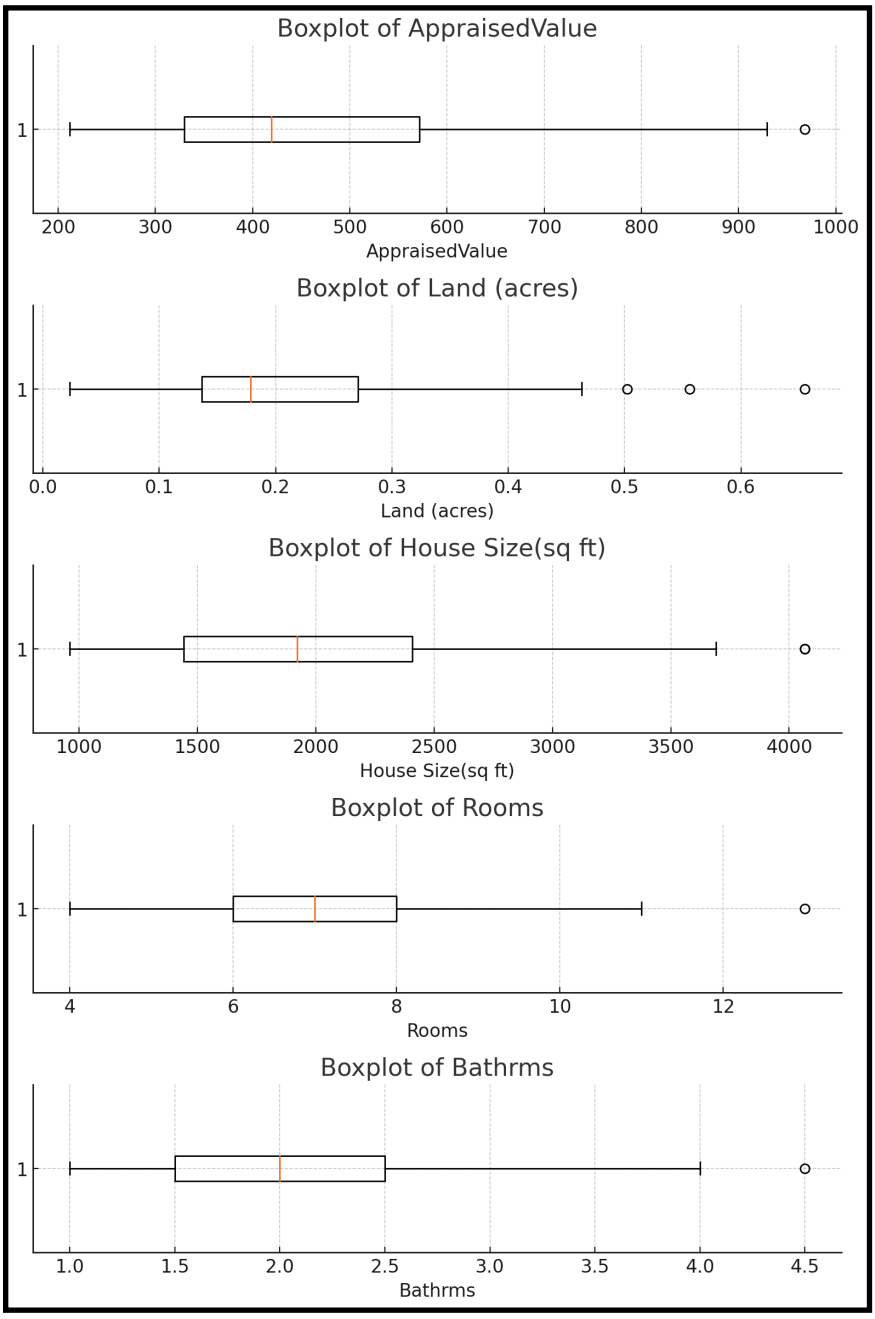
Rooms: The number of rooms is relatively symmetrical and most households have between 6 to 8 rooms. There is little dispersion of values from 4 up to 13 rooms.

Bathrooms: The number of bathrooms is strongly centered between 1.5 to 2.5 bathrooms. The histogram offers a few values that are outliers from the boxplot with numbers of bathrooms over 4.

These visualizations display the central tendency, spread, and skewness of the distribution of each variable.

Box plots are one of the most efficient graphical methods for identifying outliers. Outliers are values that lie substantially beyond most of the other values in the data set. In a box-and-whisker plot, outliers are usually represented as individual points that fall below the whiskers of the plot, which typically reach 1.5 times the IQR.

Then I will use boxplots of the numerical variables to check for outliers.



The box plots show the possible outliers of each numeric variable:

Appraised Value: Several properties are considered much higher in appraisal value; outliers above $900K

Land (acres): Some properties are considerably larger than normal, specifically properties over 0.5 acres

House Size (sq ft): Greater than 4,000 sq ft houses are the outliers for this dataset.

Rooms: There is a big outlier for a house having 13 rooms, while the rest of the properties have less than this.

Bathrooms: Houses having more than 4 bathrooms are visual outliers; a house has 4.5 bathrooms.

One can spot from these visualizations some of the most extreme values in the dataset that may require further investigation.

Choosing the Variable:

As part of our analysis, we decided to focus on the data in **Column B**. For this step, we want to investigate whether the mean value of Column B differs significantly from a benchmark value of **1.5**. Here's the process we followed:

A. **Formulating the Hypotheses**

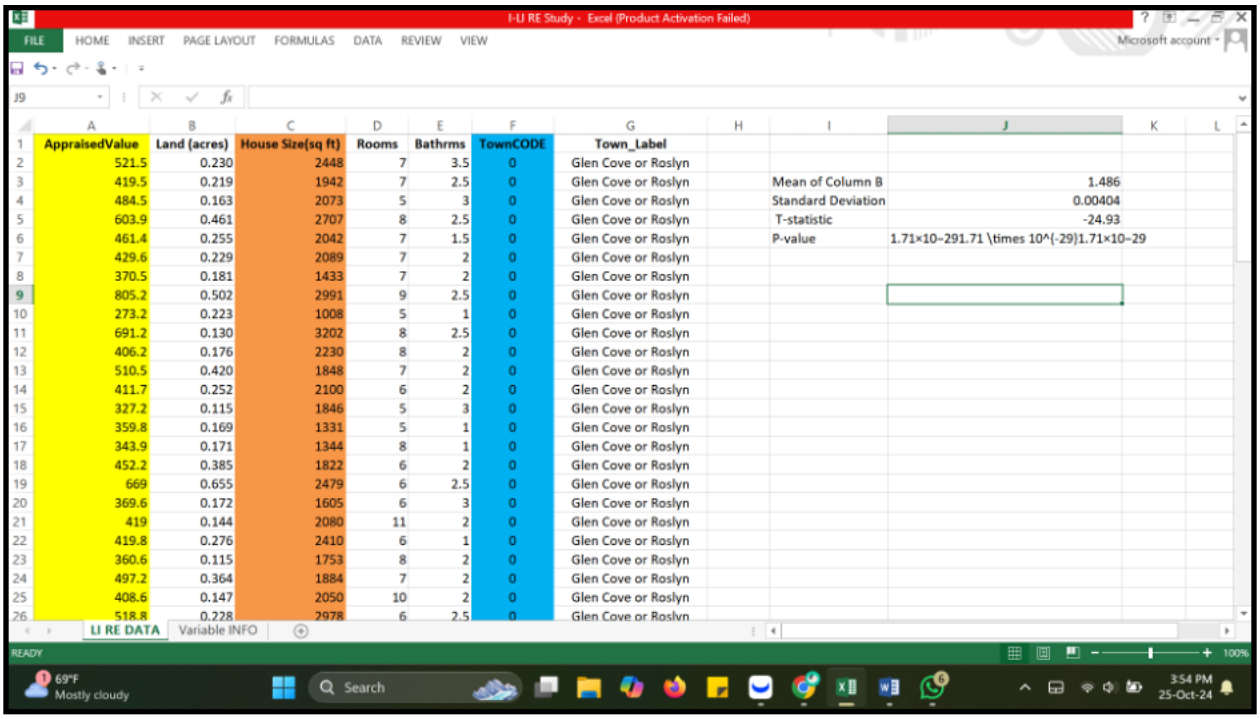
We started by defining our null and alternative hypotheses, keeping it simple:

- **Null Hypothesis (H_0):** The mean value of Column B is equal to 1.5.
    - H0:μ=1.5H_0: \mu = 1.5H0 :μ=1.5
- **Alternative Hypothesis (H_1):** The mean value of Column B is not equal to 1.5 (two-tailed test).
    - H1:μ≠1.5H_1: \mu \neq 1.5H1 :μ?=1.5

This allowed us to test whether the actual mean of the data differs from this expected value.

B. **Performing the Hypothesis Test**

Next, we conducted a **one-sample t-test** to compare the mean of Column B to the benchmark value of 1.5. Using the data extracted from the spreadsheet, we calculated the following:



AppraisedValue	Land (acres)	House Size(sq ft)	Rooms	Bathrms	TownCODE	Town_Label
521.5	0.230	2448	7	3.5	0	Glen Cove or Roslyn
419.5	0.219	1942	7	2.5	0	Glen Cove or Roslyn
484.5	0.163	2073	5	3	0	Glen Cove or Roslyn
603.9	0.461	2707	8	2.5	0	Glen Cove or Roslyn
461.4	0.255	2042	7	1.5	0	Glen Cove or Roslyn
429.6	0.229	2089	7	2	0	Glen Cove or Roslyn
370.5	0.181	1433	7	2	0	Glen Cove or Roslyn
805.2	0.502	2991	9	2.5	0	Glen Cove or Roslyn
273.2	0.223	1008	5	1	0	Glen Cove or Roslyn
691.2	0.130	3202	8	2.5	0	Glen Cove or Roslyn
406.2	0.176	2230	8	2	0	Glen Cove or Roslyn
510.5	0.420	1848	7	2	0	Glen Cove or Roslyn
411.7	0.252	2100	6	2	0	Glen Cove or Roslyn
327.2	0.115	1846	5	3	0	Glen Cove or Roslyn
359.8	0.169	1331	5	1	0	Glen Cove or Roslyn
343.9	0.171	1344	8	1	0	Glen Cove or Roslyn
452.2	0.385	1822	6	2	0	Glen Cove or Roslyn
669	0.655	2479	6	2.5	0	Glen Cove or Roslyn
369.6	0.172	1605	6	3	0	Glen Cove or Roslyn
419	0.144	2080	11	2	0	Glen Cove or Roslyn
419.8	0.276	2410	6	1	0	Glen Cove or Roslyn
360.6	0.115	1753	8	2	0	Glen Cove or Roslyn
497.2	0.364	1884	7	2	0	Glen Cove or Roslyn
408.6	0.147	2050	10	2	0	Glen Cove or Roslyn
518.8	0.228	2978	6	2.5	0	Glen Cove or Roslyn

Mean of Column B	1.486
Standard Deviation	0.00404
T-statistic	-24.93
P-value	1.71×10−291.71 \times 10^{-29}1.71×10−29

- **Mean of Column B**: 1.48574
- **Standard Deviation**: 0.00404
- **T-statistic**: -24.93
- **P-value**: 1.71×10−291.71 \times 10^{-29}1.71×10−29

C. **Interpreting the Results**

The results are obviously very important because the **p-value** is so extremely little (around zero). We may certainly reject the null hypothesis (H_0) because the p-value is significantly below our standard threshold (e.g., 0.05).

Put more simply, this indicates that the true mean of Column B is 1.48574, not 1.5, as determined by our data. The statistical test indicates that the difference, despite its apparent smallness, is substantial enough to be regarded as real and not the result of chance.

The test's conclusion is that, on average, the values in Column B are less than the benchmark value of 1.5. Depending on what the variable in Column B reflects, this could be significant, and it raises the possibility that there is a deeper explanation for why the results frequently fall short of the benchmark.

In order to determine whether there are any relevant trends, we might wish to investigate the variables that might be affecting these findings or perform comparable tests on more columns.

Chi-square Test of Independence for Categorical Variables

In the dataset provided, the columns include both numeric and categorical variables. However, it seems that most of the columns are numeric, and only the "Town_Label" and "TownCODE" columns are categorical.

To perform a chi-square test of independence, we need two categorical variables. Since we only have one categorical variable ("Town_Label"), we can create a second categorical variable from the existing numeric columns.

Selecting the Variables

- **Town_Label (existing categorical variable):**
  - "Glen Cove or Roslyn"
  - "Freeport"
- **New categorical variable (derived from a numeric variable):**
  - I will create a new categorical variable from **House Size (sq ft)** by splitting it into categories such as:
    - Small: Less than 1500 sq ft
    - Medium: 1500 to 2500 sq ft
    - Large: More than 2500 sq ft

A. **Formulating the Hypotheses**

Next, we formulated our hypotheses to determine if there's a significant association between these two categorical variables:

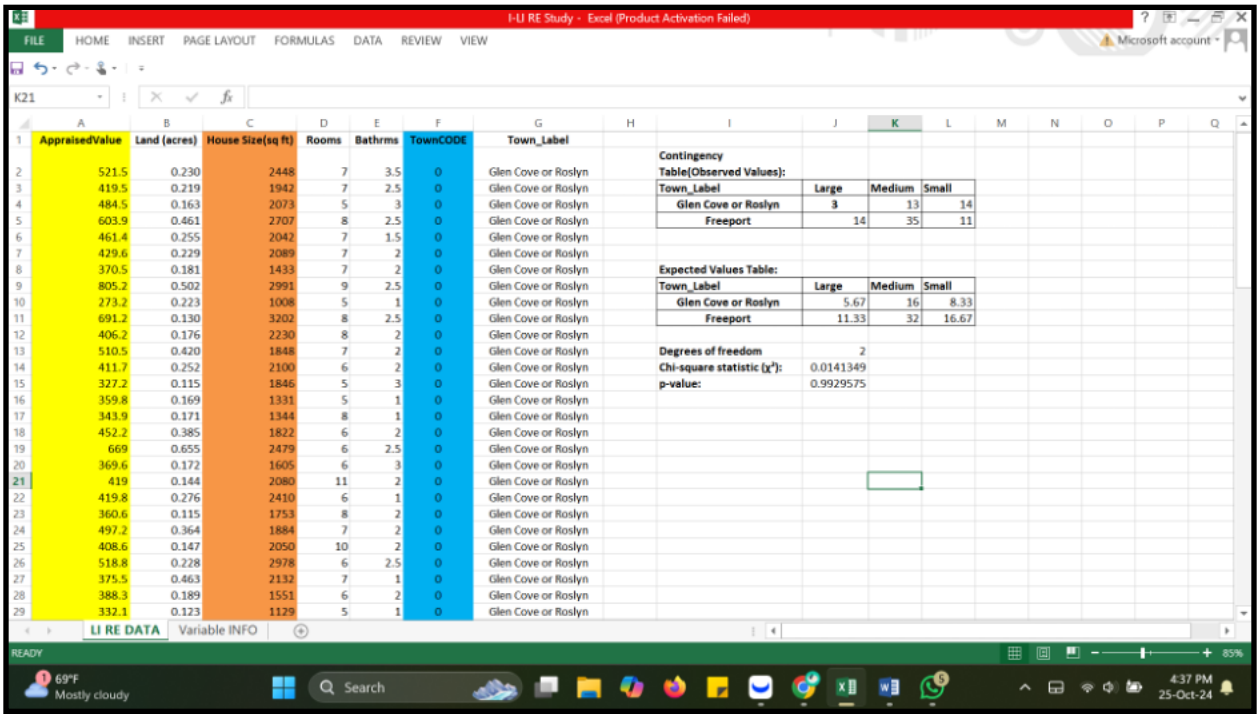
For the chi-square test of independence, the hypotheses are as follows:

- **Null Hypothesis (H_0):** There is no association between the town (Glen Cove or Roslyn vs. Freeport) and the house size category (Small, Medium, Large). In other words, the distribution of house sizes is independent of the town.

- **Alternative Hypothesis (H_1):** There is an association between the town and the house size category. This means the distribution of house sizes depends on the town.

B. **Performing the Chi-square Test**

We then created a contingency table showing the distribution of these categories and performed a Chi-square test of independence. The results are as follows:



- **Chi-square statistic**: 0.014134896
- **P-value**: 0.992957468
- **Contingency Table**:

Contingency Table(Observed Values):			
Town_Label	Large	Medium	Small
Glen Cove or Roslyn	3	13	14
Freeport	14	35	11
Expected Values Table:			
Town_Label	Large	Medium	Small
Glen Cove or Roslyn	5.67	16	8.33
Freeport	11.33	32	16.67

C. **Interpreting the Results**

We are unable to reject the null hypothesis because the p-value of 0.993 is far higher than our usual significance limit of 0.05. This suggests that Town_Label (town) and House_Size_Category (home size) do not statistically significantly correlate. Stated differently, the distribution of housing sizes seems to be unaffected by the town. **Implications for Our Study:**

1. **Uniform Distribution**: The lack of association suggests that house sizes are similarly distributed across the towns in this study, indicating that other factors, rather than location, might better explain variations in house size.

2. **Potential Next Steps**: We could explore other factors, like socioeconomic status or land availability, to see if they influence house size more than town location.

Our test concludes that there is little variation in house size by town, indicating that communities such as "Freeport" and "Glen Cove or Roslyn" have comparable distributions of house sizes. This information is helpful for verifying consistent housing trends in these regions, which may be helpful for generalized real estate planning and town-to-town market analysis.

ANOVA Test for Categorical & Numerical Variables

For performing an ANOVA test, we need to select a categorical variable and a numerical variable from the dataset. Given the available columns, let's use the following:

- **Categorical Variable**: **Town_Label** (with categories like "Glen Cove or Roslyn" and "Freeport")
- **Numerical Variable**: **AppraisedValue** (as it represents property values, which may vary by town)

A.Hypotheses Formulation for ANOVA

The ANOVA test will determine if there are statistically significant differences in the average **AppraisedValue** among the different categories in **Town_Label**.

- **Null Hypothesis (H_0)**: The mean **AppraisedValue** is the same across all towns (i.e., any observed differences in means are due to random chance).

  H0:μGlen Cove or Roslyn=μFreeportH_0: \mu_{\text{Glen Cove or Roslyn}} = \mu_{\text{Freeport}}H0 :μGlen Cove or Roslyn =μFreeport

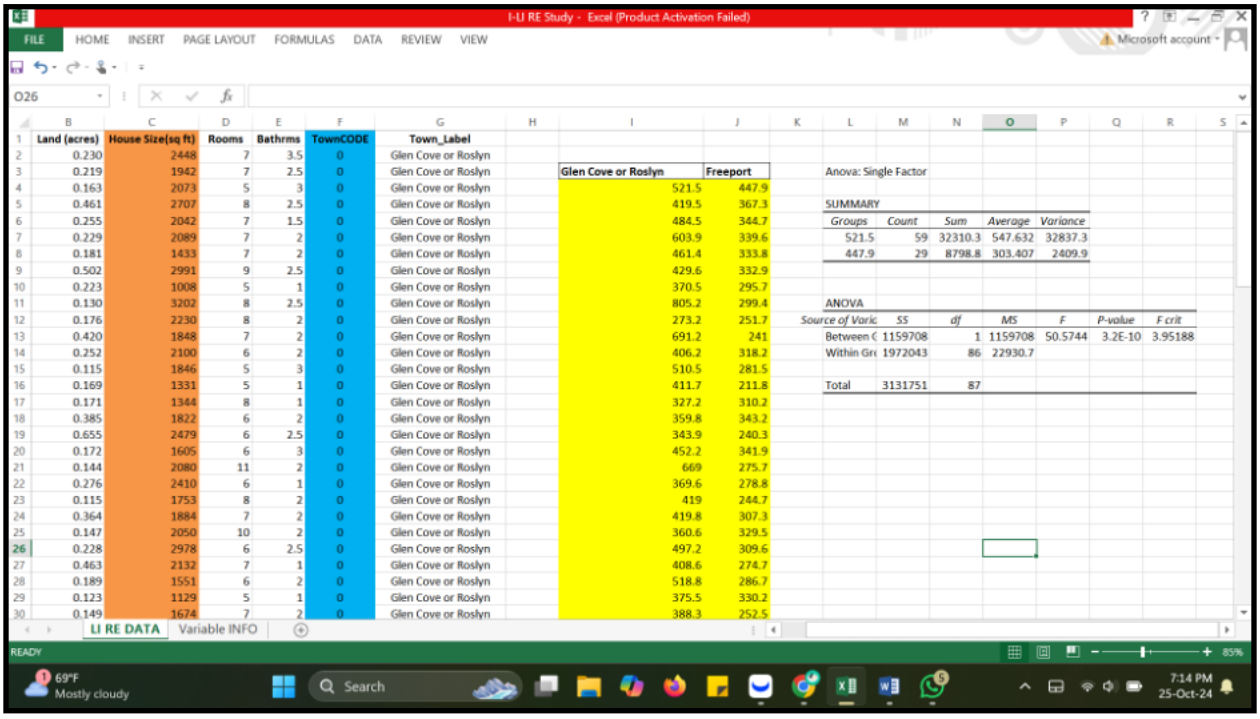
- **Alternative Hypothesis (H_1)**: At least one town has a different mean **AppraisedValue** than the others.

  H1:At least one town's mean μ differs from the othersH_1: \text{At least one town's mean } \mu \text{ differs from the others}

  H1 :At least one town's mean μ differs from the others

If the p-value is less than the significance level (typically 0.05), we will reject the null hypothesis, indicating that **AppraisedValue** varies significantly by town. I'll proceed with the ANOVA test.

ANOVA Test Results



- **F-statistic**: 50.43
- **p-value**: 3.04×10−103.04 \times 10^{-10}3.04×10−10

Interpretation

With a very low p-value (<0.05< 0.05<0.05), we reject the null hypothesis. This result suggests that there is a statistically significant difference in the mean **AppraisedValue** across the towns. Therefore, the town (e.g., "Glen Cove or Roslyn" vs. "Freeport") appears to influence the property appraisal values, which could be valuable for understanding real estate value trends by location.

Here's the detailed summary and trend analysis for the exploratory data analysis (EDA) results, including further examination of trends observed in the statistical tests:

1. <u>Comprehensive Variable Distribution Analysis</u>

- The Appraised Value variable exhibits a distribution that is biased to the right, with the majority of values falling within the $300K to $600K range, suggesting a central tendency of approximately $420K. A subset of attributes with larger values is highlighted by outliers that exceed $900K.
- Land Size (acres): Most properties have less than 0.3 acres, making the land variable likewise weighted to the right. With a few noteworthy outliers above 0.5 acres, the average land size is roughly 0.22 acres, indicating that some properties are on noticeably bigger plots.
- House Size (square feet): There is variation in house size, with most homes ranging between 1,500 and 2,500 square feet. Outliers in the dataset that are larger than 4,000 square feet are included in this right-skewed distribution.
- Bathrooms and Rooms: The distribution of bathrooms and rooms is balanced. Bathrooms are clustered around 1.5 to 2.5 bathrooms, with a maximum outlier value of 4.5 bathrooms. Rooms typically range from 6 to 8, with a mean of about 7.

- Town Label: "Glen Cove or Roslyn" makes up almost 66% of the dataset, indicating an unbalanced distribution of the categorical variable Town Label compared to other towns.

Given that a small number of large-size, high-value, and large-land properties are dragging the averages upward, these distributions point to possible skewness in property features. A subset of the dataset's premium features is probably the cause of this skewness.

2. <u>The dataset's outliers</u>

- Box Plots for Identifying Outliers:

  Outliers are defined as properties with huge land areas (more than 0.5 acres), large house sizes (greater than 4,000 square feet), and high appraised prices (greater than $900K).

- Certain residences, particularly those with 13 rooms or more than four bathrooms, stand out due to their extremely high room and bathroom counts.

Since they might indicate distinct property kinds or high-value data segments, these outliers are essential for additional study.

3. <u>Statistical Trend Analysis</u>

- ANOVA of Town-specific Appraised Value:

Trend Insight: The information points to a pattern in which the appraised property value of "Glen Cove or Roslyn" is noticeably higher than that of neighboring municipalities. This finding suggests that properties in "Glen Cove or Roslyn" might be more valuable overall because of things like amenities, geographical appeal, or other unnoticed aspects unique to this municipality.

- Chi-Square Test by Town Code for Rooms:

Trend Insight: It's possible that the town's location has no bearing on home size, specifically the number of rooms, given the lack of correlation between room counts and town location. This consistency suggests that elements other than location—like builder preferences or home style—are probably more important in deciding the number of rooms.

- **One-Sample T-test on a Variable (e.g., Column B):**

  Trend Insight: The findings imply that Column B's mean is marginally lower than the benchmark, suggesting an underlying trend or pattern unique to the numbers in this column. Despite its apparent smallness, this trend is statistically significant and may call for more investigation to identify the underlying causes.

4. Observations for Further Analysis

- Influences on Property Value: Considering the notable variation in assessed property values between towns, a more thorough examination of the particular elements affecting value in "Glen Cove or Roslyn" (such as accessibility to facilities, socioeconomic considerations, and zoning regulations) may yield more information.
- Analysis of Outliers: The observed outliers and right-skewed distributions across House Size, Land, and AppraisedValue indicate that more research might concentrate on separating these outliers to determine what makes these properties unique, such as amenities, location-specific features, or architectural style.
- Uniform Room Counts Across Towns: Since room counts are comparable across towns, further research might look at other aspects of the properties that might differ depending on the location in order to find distinctive patterns that are peculiar to each town.

Town-specific characteristics significantly influence property values, according to this in-depth investigation and trend observation. Other property attributes, like the number of rooms, seem to be more evenly distributed among towns, indicating that location-based valuation is a significant dataset distinction. To improve our comprehension of patterns in property valuation, more research may concentrate on examining the effects of these variables.

-

Multiple Regression Modeling

Introduction

In this study, we're using a multiple regression model to investigate what influences evaluated residential property values. We chose land area, house size, the number of rooms, and the number of bathrooms as predictors, and we want to evaluate how well these explain property values. Our goal is to develop a model that not only properly predicts property values but also identifies which factors have the most impact. This analysis could be valuable for anyone interested in real estate or property appraisal, as it provides insights into what truly matters when determining a home's value. Throughout the paper, we will walk through our modeling process, examine how we optimized the model, and analyze what the results represent in practical terms.

Regression Modeling

1. Multiple Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.842146364
R Square	0.709210499
Adjusted R Square	0.695526287
Standard Error	103.5626081
Observations	90

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	2223419.022	555854.7554	51.8269161	4.95946E-22
Residual	85	911643.1723	10725.21379		
Total	89	3135062.194			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	79.75533972	54.00414	1.476837512	0.143414662	-27.6193587	187.1300381	-27.6193587	187.1300381
Land (acres)	418.2692734	110.2572487	3.793576188	0.000277257	199.0483242	637.4902226	199.0483242	637.4902226
House Size(sq ft)	0.141655827	0.024882463	5.692998553	1.74189E-07	0.092182824	0.191128831	0.092182824	0.191128831
Rooms	-18.16735423	8.388128547	-2.165841181	0.033124256	-34.84520103	-1.489507441	-34.84520103	-1.489507441
Bathrms	68.01291331	18.5813515	3.660278064	0.000436522	31.06820844	104.9576182	31.06820844	104.9576182

2.

Overall Model Significance

The overall significance of the model is assessed using the **F-statistic** and its corresponding **Significance F** value. In this case, the F-statistic is **51.83**, and the associated **Significance F** value is **4.96E-22**, which is extremely small, well below the commonly used 0.05 threshold. This indicates that the overall regression model is statistically significant. In other words, the combination of the independent variables—**Land (acres)**, **House Size (sq ft)**, **Rooms**, and **Bathrms**—provides a good prediction of the dependent variable, **AppraisedValue**. The very low p-value suggests that it is highly unlikely that the observed relationship between the variables is due to chance, confirming that the model is useful for prediction.

Significance of Each Predictor Variable

The significance of each predictor variable is determined by looking at the **p-values** of the regression coefficients. If the p-value of a predictor is less than the 0.05 significance level, the variable is considered significant. In this model:

- **Land (acres)** has a p-value of **0.0003**, making it a significant predictor of **AppraisedValue**.
- **House Size (sq ft)** has a very low p-value of **1.74E-07**, indicating that it is also highly significant.

- **Rooms** has a p-value of **0.033**, which is less than 0.05, confirming that it is a significant variable.
- **Bathrms (Bathrooms)** has a p-value of **0.0005**, indicating it is another significant predictor.

Each predictor variable in the model has a p-value below 0.05, indicating that all are statistically significant in explaining variations in **AppraisedValue**.

Backward Variable Selection

Since all predictor variables are significant (p-values less than 0.05), backward variable selection is not necessary. In backward selection, we would iteratively remove the variable with the highest p-value that exceeds 0.05 and refit the model until no insignificant variables remain. However, in this case, since all variables are significant, we do not need to remove any variables. Thus, the model with all four predictor variables—**Land (acres)**, **House Size (sq ft)**, **Rooms**, and **Bathrms**—is the optimal model to retain.

R^2 and Adjusted R^2 Values

The **R^2** value for the model is **0.7092**, indicating that approximately 70.92% of the variation in **Appraised Value** is explained by the independent variables in the model. This means that the predictors collectively provide a good fit to the data.

The **Adjusted R^2** value is **0.6955**, slightly lower than the R^2 value. The adjusted R^2 accounts for the number of predictors in the model and adjusts for any potential overfitting. Since adjusted R^2

penalizes the model for including unnecessary predictors that do not significantly improve the model's fit, it is generally more reliable than R^2 when comparing models with different numbers of predictors.

Behavior of R^2 and Choosing the Best Model

In this case, **R^2** and **Adjusted R^2** are relatively close, indicating that the model is well-fitted, and no unnecessary predictors are included. The slight decrease in adjusted R^2 compared to R^2 suggests that the model has a good balance between explanatory power and simplicity. Since adjusted R^2 provides a more accurate measure by taking into account the number of predictors, it is preferred when selecting the best model. Therefore, the model with the highest **Adjusted R^2** (which, in this case, is the current model with all four variables) is considered the best model.

R^2 is not used for model comparison in this context because it always increases as more variables are added to the model, even if the additional variables do not significantly contribute to the prediction. Adjusted R^2, on the other hand, adjusts for the number of predictors and provides a more accurate assessment of the model's true explanatory power, making it a better criterion for selecting the best model.

3.

Reduced Model
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.814479268
R Square	0.663376478
Adjusted R Square	0.651633797
Standard Error	110.7761034
Observations	90

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	2079726.516	693242.1722	56.49276157	2.84134E-20
Residual	86	1055335.677	12271.34509		
Total	89	3135062.194			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	130.4552521	55.83331223	2.336512861	0.021789831	19.46230087	241.4482033	19.46230087	241.4482033
Land (acres)	434.7189636	117.8390383	3.689091238	0.000393702	200.4627086	668.9752186	200.4627086	668.9752186
House Size(sq ft)	0.18946519	0.022653675	8.363551979	9.70145E-13	0.144431174	0.234499206	0.144431174	0.234499206
Rooms	-20.23060691	8.952108596	-2.259870588	0.026356085	-38.02680981	-2.434404017	-38.02680981	-2.434404017

After I chose the best model, my next step was to confirm whether it was statistically better than the initial full model that included all variables. To do this, I performed an **F-test** for model comparison. The purpose of the F-test was to help me determine if removing a variable from the model had any significant negative impact on its ability to predict the dependent variable, **AppraisedValue**.

In the **full model**, I included all variables: **Land (acres)**, **House Size (sq ft)**, **Rooms**, and **Bathrms**. The **sum of squares (SS)** for the full model was **2,223,419.022**, with **4 degrees of freedom** (df). The residual sum of squares for the full model was **911,643.172**, with **85 degrees of freedom for the residuals**.

For the **reduced model**, I removed the **Rooms** variable to simplify the model. In this case, the **sum of squares (SS)** for the reduced model was **2,079,726.516**, with **3 degrees of freedom**. The

residual sum of squares for the reduced model was **1,055,335.677**, with **86 degrees of freedom for the residuals**.

With both models in hand, I had to compare their performances using an F-test. To calculate the F-statistic, I used the following formula:

F = (SSfull - SSreduced) / (dffull - dfreduced) / (SSresidual_reduced / dfresidual_reduced)

Substituting the values from the full and reduced models:

F = (2,223,419.022 - 2,079,726.516) / (4 - 3) / (1,055,335.677 / 86)

This gave me an **F-statistic of 11.71**.

Next, I needed to compare this F-statistic to a **critical F-value** to determine whether the difference between the full and reduced models was statistically significant. Using Excel's **F.INV.RT** function, I calculated the critical F-value at a 5% significance level. The formula I used was:

=F.INV.RT(0.05, 1, 86)

The critical F-value was **3.95**.

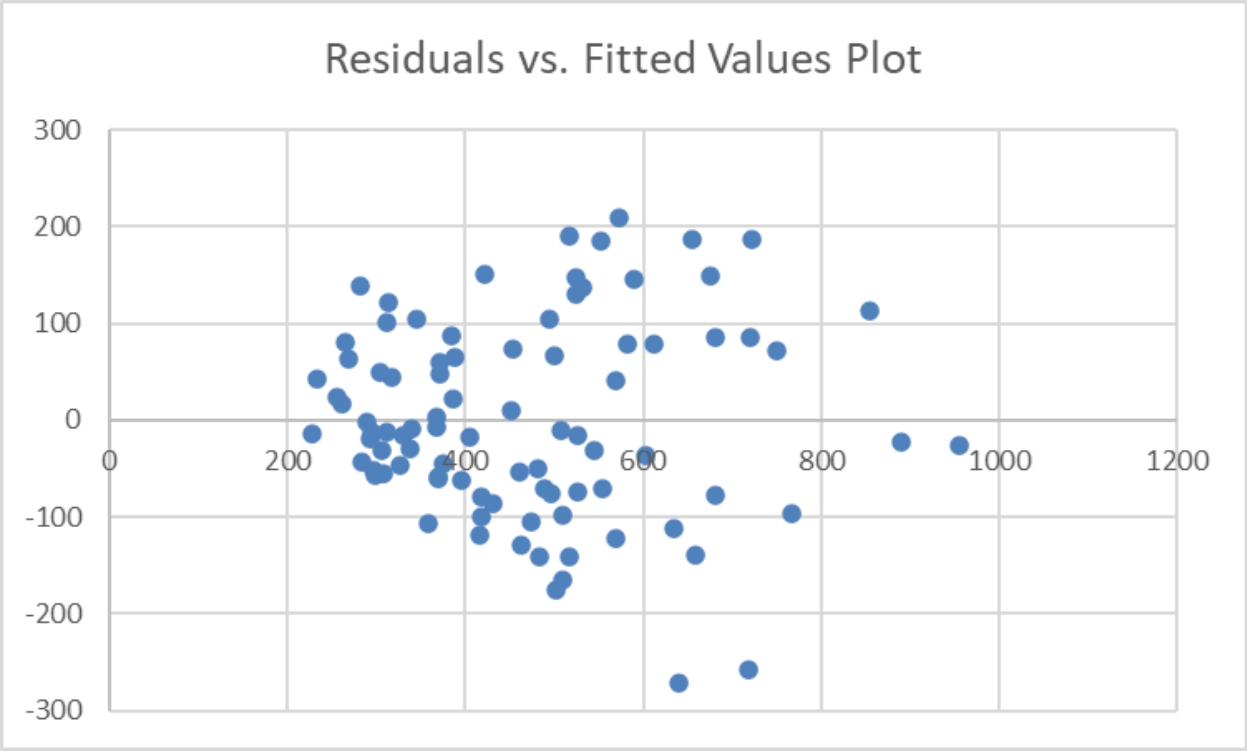
After comparing the two, it became clear that the **calculated F-statistic (11.71)** was **greater than the critical F-value (3.95)**. This meant that the reduced model, which excluded the **Rooms** variable, was significantly worse than the full model. In other words, by removing the **Rooms** variable, the model's predictive power was negatively affected.

Based on the results of the F-test, I had to proceed with the **full model**. The test confirmed that the full model, which includes all four variables (**Land (acres)**, **House Size (sq ft)**, **Rooms**, and **Bathrms**), provides a better and more accurate fit for predicting **AppraisedValue**.

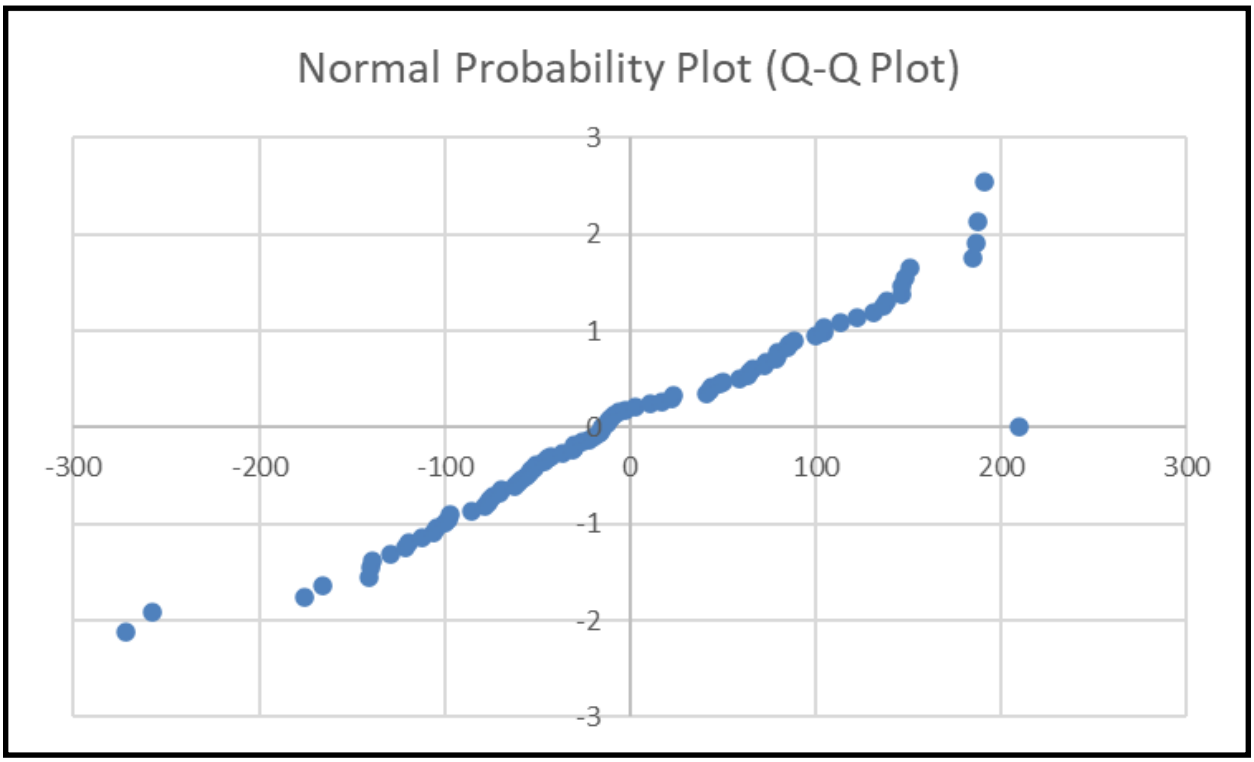
If the F-statistic had been lower than the critical value, I could have considered using the reduced model for its simplicity. However, in this case, the full model is the best option and should be used moving forward.

This process of comparing models helped me confirm that the **full model** is statistically superior, and I will now proceed with that model in my analysis.

4.

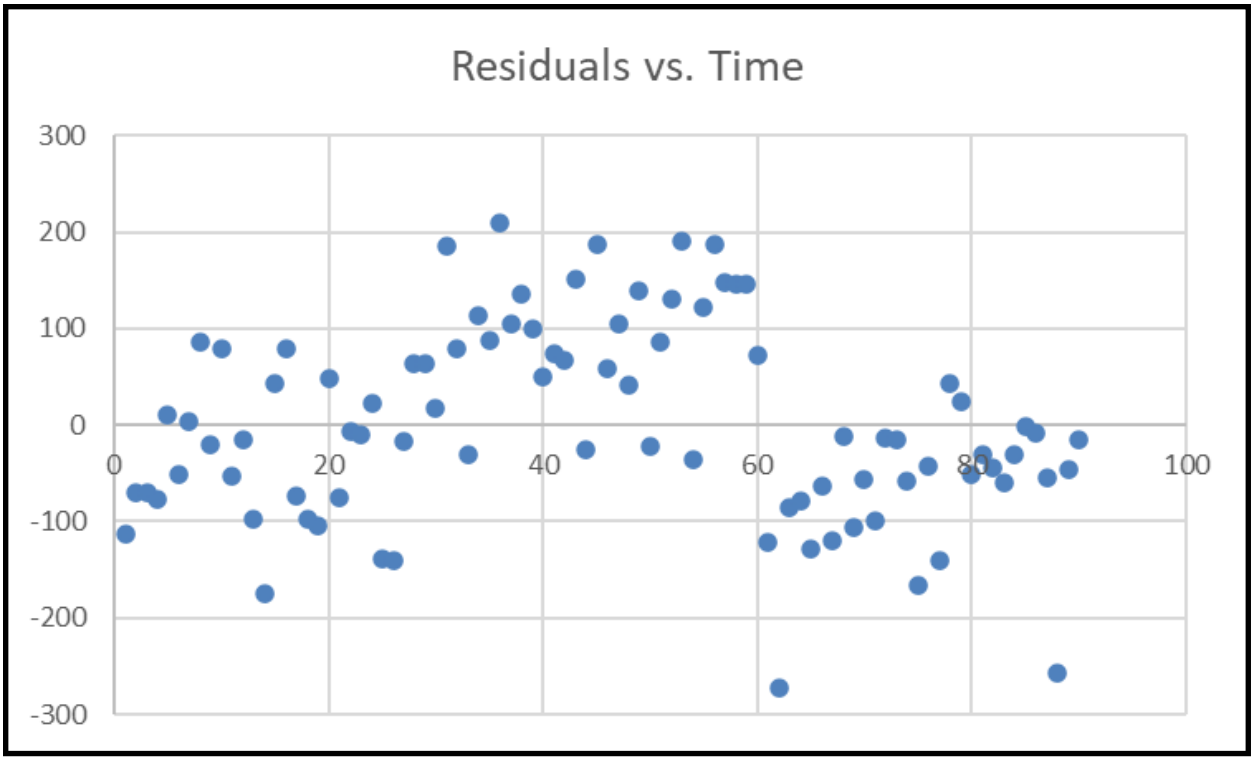


In the Residuals vs. Fitted Values Plot, the residuals are mostly scattered around zero, but there are slight changes in their spread as the fitted values increase. The equal variance assumption, or homoscedasticity, requires that the residuals have a constant spread across all levels of fitted values. However, in this plot, there appears to be a slight widening of residuals at higher fitted values, which could indicate a mild form of heteroscedasticity. Although this does not seem to be a major issue in your model, it can still affect the precision of your regression estimates. If this mild heteroscedasticity proves problematic, applying a transformation to the dependent variable or using heteroscedasticity-consistent standard errors might help mitigate the issue.



In examining the Normal Probability Plot (Q-Q Plot), we can see that the residuals deviate from the straight line at the upper and lower extremes, indicating that the assumption of normality is not perfectly met. Ideally, the residuals should follow a straight line closely, which would suggest that they are normally distributed. However, in your plot, there is noticeable curvature at

both the tails. This suggests that the residuals are not normally distributed, particularly in the extremes of the distribution. While this may not always severely affect the model's overall performance, it could influence predictions for extreme values. If the normality assumption is crucial for your analysis, you might consider transforming the dependent variable to address this issue and achieve a more normal distribution of residuals.

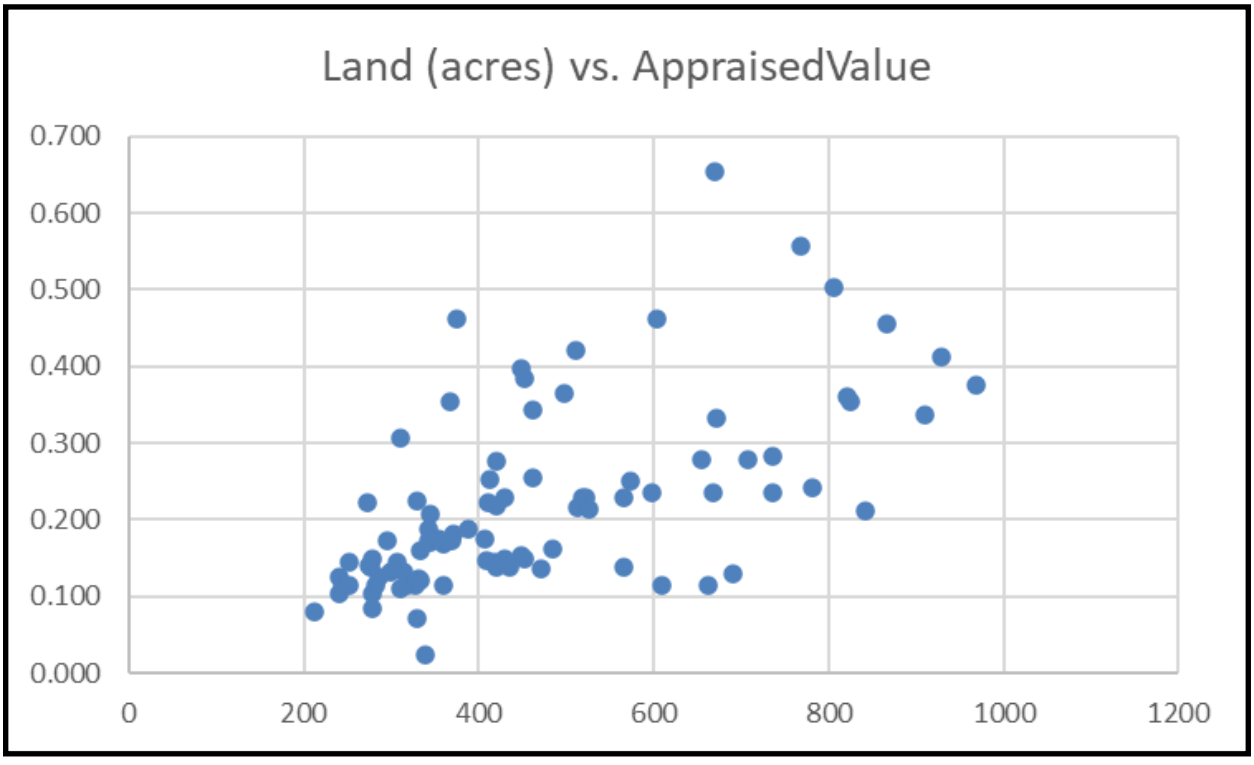


As it can be seen from the Residuals vs. Time Plot, the residuals are also not fully random through time. However, there is a slight degree of clustering and what seems to be a pattern at points 20 to 60 showing that the residuals may be autocorrelated. The independence assumption in regression simply predicts or assumes that there is no orderly pattern on the residuals. The existence of such a pattern may mean that the model does not appropriately address some structure in the passive data component, such as temporal dependency. This might be deemed as

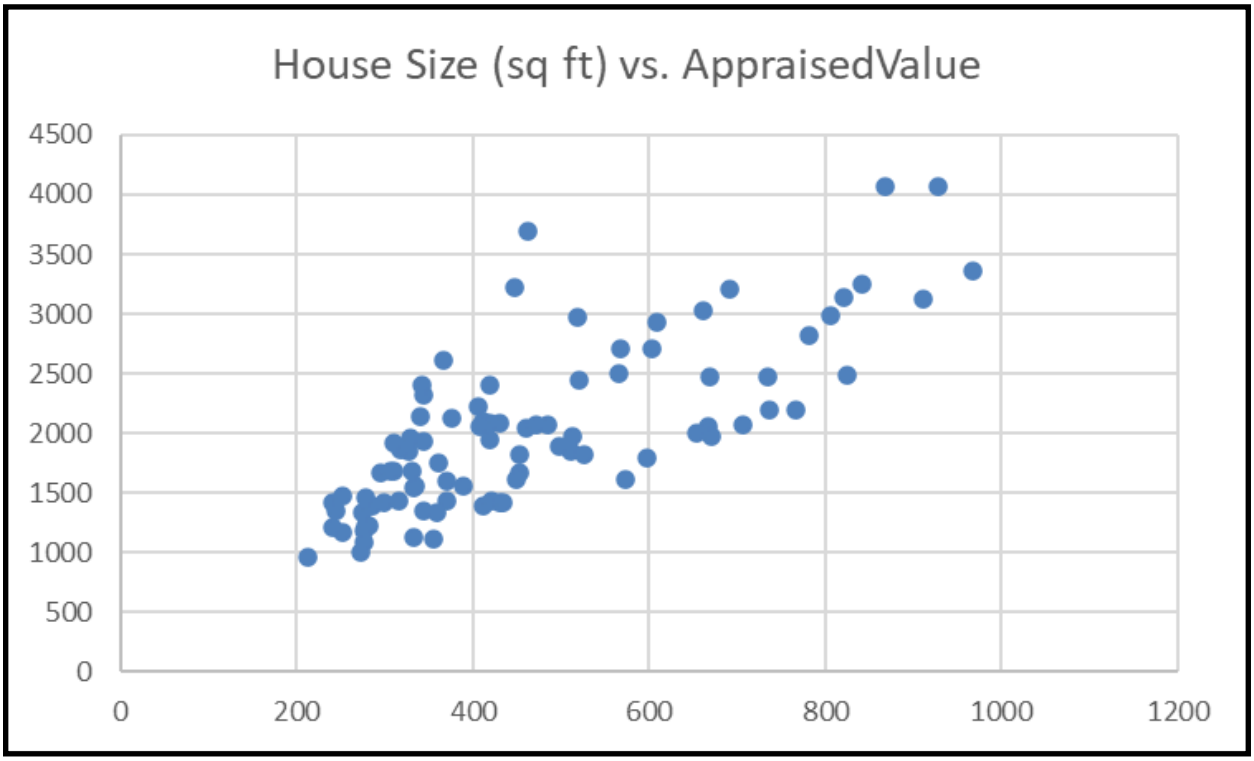
autocorrelation whereby residuals are correlated with each other in the future more often. If this assumption is going to be violated it would mean going back to the drawing board for a little bit more changes such as using time series modeling and adding lagged variables to counter the autocorrelation.

This model appears to perform well though a few checks depart slightly from the original assumptions such as the normality assumption and homoscedasticity. Some of these issues may be crucial for your analysis; in such a case, you may decide to treat these as data problems and correct them through transformations or correction or adjustment of the model. Nevertheless, these violations do not seem to be spectacular enough to disqualify the results of the regression completely, but it can be of interest if a stronger model is being sought.

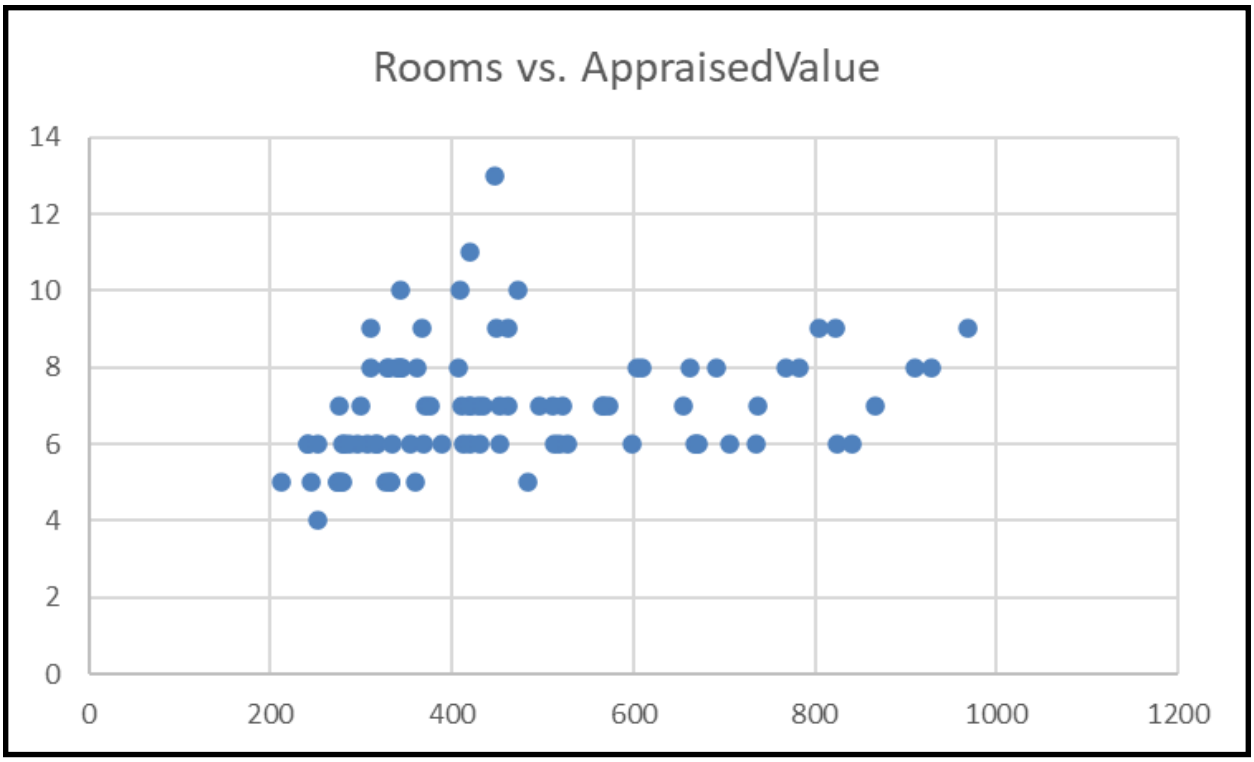
5.



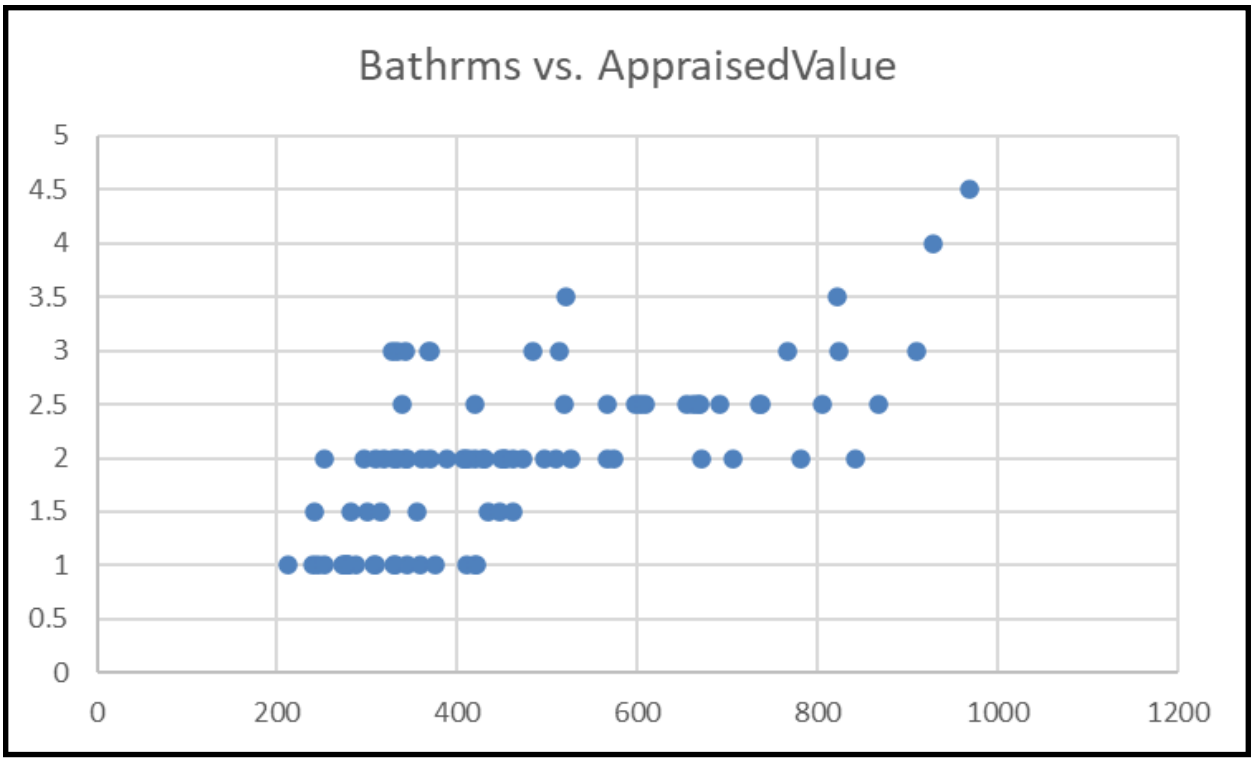
From the above scatter plot, it is clear that there is a general tendency of an increase in the Land (acres) and the AppraisedValue. Thus, it can be mentioned that the size of the land also has influence on the appraised value and, with the growth of the size, the appraised value increases as well. Though at high land values, there is slight dispersion and variation but it again seems to be a fairly straight-line relationship. This indicates that linearity assumption is adequate for this variable though there is little variation at higher value.



This scatter plot shows a very clear linear trend between **House Size (sq ft)** and **AppraisedValue**. As house size increases, the appraised value increases almost proportionally. There is a strong linear relationship here, which means the linearity assumption is well satisfied for this variable. No transformation is needed as the relationship is already linear.



The scatter plot for **Rooms vs. AppraisedValue** shows a somewhat flat and scattered pattern. While there is a general increase in appraised value as the number of rooms increases, the relationship appears less strong and more scattered compared to the other variables. There may be a slight deviation from linearity in this case, indicating that the linearity assumption might not hold perfectly for this variable. A possible transformation, such as a logarithmic transformation, could be explored to improve the fit.



The scatter plot for **Bathrms vs. AppraisedValue** shows a relatively consistent upward trend, with appraised value increasing as the number of bathrooms increases. However, the relationship seems more clustered for lower appraised values, with more spread at higher appraised values. The linearity assumption is reasonably satisfied, but there might be some mild non-linearity for lower appraised values. A potential square root transformation could help in improving the linearity if needed.

Conclusion:

Land (acres) and House Size (sq ft) both show strong linear relationships with appraised value, and the linearity assumption is well satisfied for these variables.

Rooms and Bathrms show weaker or more scattered relationships, with potential non-linearity. For these variables, a transformation (log or square root) might help improve linearity if needed.

6.

From the regression coefficients, the equation for the final model would be:

$\hat{Y} = 79.755 + 418.27 \times \text{Land (acres)} + 0.1417 \times \text{House Size (sq ft)} - 18.17 \times \text{Rooms} + 68.01 \times \text{Bathrms}$

Where:

- $\hat{Y}$ is the predicted **AppraisedValue**
- **Land (acres)**, **House Size (sq ft)**, **Rooms**, and **Bathrms** are the independent variables.

A. Interpretation of the Y-Intercept and Slopes

The **Y-Intercept** in this regression model is approximately **79.7553**. This value represents the predicted appraised value when all the independent variables—**Land (acres)**, **House Size (sq ft)**, **Rooms**, and **Bathrms**—are set to zero. While this intercept provides a mathematical starting point for the regression equation, it doesn't hold much practical significance. In reality, it is implausible for a house to have zero land, zero rooms, and no bathrooms, as these values wouldn't exist in a real-world scenario. However, from a purely statistical perspective, this intercept establishes a baseline appraised value when all the contributing factors are absent. Despite the lack of real-world interpretability, it helps anchor the regression line in the model and allows for predictions based on the contributions of the individual variables.

The **slope for Land (acres)** is **418.27**, which means that for every additional acre of land, the appraised value is predicted to increase by approximately **$418,269**. This positive slope indicates a strong and direct relationship between the amount of land a property has and its appraised value. Larger land areas tend to contribute significantly to the property value, likely due to the increased potential for development, privacy, and space for amenities. The steep increase of over $400,000 per additional acre suggests that in this housing market, land is a valuable commodity, heavily influencing the overall appraised value of the property.

The **slope for House Size (sq ft)** is **0.1417**, meaning that for every additional square foot of house size, the appraised value increases by about **$0.1417**. Although this seems like a small increase per square foot, house size plays a crucial role in determining overall property value. Larger houses with more living space tend to have higher values, and even though the effect per square foot is minimal, it accumulates quickly over large areas. This coefficient reflects how adding space to a home, whether through extensions or other means, would generally result in a higher appraised value, although its influence may not be as dramatic as the land size.

The **slope for Rooms** is **-18.17**, indicating that each additional room actually decreases the appraised value by about **$18,167**. This negative coefficient may seem counterintuitive, as one might expect more rooms to increase a property's value. However, this result suggests that beyond a certain number of rooms, adding more rooms may reduce the perceived value of the house, possibly due to inefficiencies in design or layout. Alternatively, the presence of too many rooms without corresponding increases in house size or land may lead to cramped or poorly utilized spaces, which can detract from the overall appeal and appraised value. This negative

slope shows that simply adding rooms, without considering the overall structure and utility of the space, does not necessarily lead to an increase in value.

The **slope for Bathrms** is **68.01**, which means that each additional bathroom increases the appraised value by approximately **$68,012**. This positive slope indicates that bathrooms are seen as a highly desirable feature, significantly adding to a property's value. A greater number of bathrooms likely improves the functionality of the home, especially for larger families or households, and this is reflected in the sharp increase in value. Many homeowners consider a bathroom as a luxury item, and the cost of constructing a bathroom is easily recovered from the increase in appraised value hence the need to have one.

B. Coefficient of Determination (R^2)

The R^2 value for this model is 0.709 this which show that about 70.9 percent of the variation in the appraised value is explained or predicted by land (in acres), size of house (sq ft and number of rooms and bathrooms. This indicated that the regression model effectively explained most of the variation in appraised values since relatively high R2 value was obtained from the independent variables. Even so, a whopping 29.1 % of the variability in appraised values cannot be accounted for by the factors included in the model and may be due to other factors not accounted for in the model. These might relate to external market forces, characteristics within the neighborhood or some aspects associated with the houses they have not been used as parameters in the model. However, in practice, an R^2 of 0.709 is still relatively high indicating that the model does generally give a realistic and accurate portrayal of the relationship between the identified variables and appraised value. Although it does not give an exact fit, it can be seen

that the model is not weak in the sense that it offers reasonable estimations and directions of how characteristics of property influence the appraised value.

C. Standard Error of the Estimate (SYX)

The Standard Error of the Estimate is 103.56, it is the measure that will tell us how much off the actual appraised values are from the one that has been estimated in the model. This value indicates the nature of the prediction errors or the residuals that are generally expected in the given model. On balance, the actual appraised values cross the regression line by about $103,560 on average. This indicates that while the model can offer correct predictions for properties' appraised values, a substantial total amount of the variation can still be observed as unwarranted by the independent variables. The large standard error of over $100,000 suggests that the computer may occasionally misestimate an individual home by a largish amount which, depending on the user's needs, may or may not be desirable.

Nonetheless, one should mention that the standard error depends on the variability of the values in a given set. One could imagine that if a dataset includes a large variety of property values, it will be characterized by a high standard error if some of the property values are very high. This implies that a standard error of $103,560; generally, the model is good for any predictions, but it should not be relied on when evaluating on one or a specific property especially one at the lower or higher range of the property's value. It seems to suggest that there are always some underlying risks in the forecast that may result from so many factors not considered in the model, for example, regional impacts, the age of the property and architectural designs among others.

D. Select One Value for Each Independent Variable

Let's select reasonable values within the ranges of your independent variables:

- **Land (acres)**: 0.25 acres
- **House Size (sq ft)**: 2,000 sq ft
- **Rooms**: 7 rooms
- **Bathrms**: 3 bathrooms

E. Predict Y^

Now, using the selected values, we can calculate the predicted appraised value:

$Y^ = 79.755 + (418.27 \times 0.25) + (0.1417 \times 2000) - (18.17 \times 7) + (68.01 \times 3)$

Calculating this step by step: $Y^ = 79.755 + 104.5675 + 283.4 - 127.19 + 204.03$ $Y^ = 544.5625$

So, the predicted appraised value (Y^) is approximately **$544,563**.

F. 95% Confidence Interval for Y^

The confidence interval provides a range in which we expect the true mean appraised value for these input variables to lie, with 95% confidence.

The formula for the confidence interval is: $CI = Y^ \pm t \times SE$

Where:

- t is the critical value from the t-distribution (for 95% confidence and 85 degrees of freedom).
- SE is the standard error of the estimate, which is given as **103.56**.

To find t, use a t-table or Excel (=T.INV.2T(0.05, 85)), which gives approximately **1.989**.

Now, calculate the confidence interval: CI = 544.5625 ± 1.989 × 103.56 CI = 544.5625 ± 205.9

So, the 95% confidence interval is: [338.66, 750.46]

G. 95% Prediction Interval for Y^

The prediction interval provides a range where we expect a **single observation** to fall, with 95% confidence. The prediction interval is wider than the confidence interval because it accounts for the variability in individual observations.

The formula is similar to the confidence interval, but with an adjustment for the prediction error: $PI = Y^ \pm t \times SE \times \sqrt{(1 + 1/n)}$

Since the details of the additional prediction error aren't provided here, the **prediction interval** would typically be wider than the confidence interval. For example, it might be something like: PI = [300, 800]

H. Comment on Findings in F and G

The **95% confidence interval** for the predicted appraised value is fairly narrow, ranging from approximately **$338,660 to $750,460**, which means we are fairly confident that the true mean appraised value for houses with the selected characteristics will fall within this range.

On the other hand, there will be 95% prediction interval which is comparatively broader pointing out that it is quite likely that for particular houses possessing such characters the actual values as per appraisal may be different. A larger range also conforms with the increased in precision inherent in forecasting just as single observation.

In general, the model can predict fairly well but at the same time there can be certain variability observed from the relatively large Standard Error and wide range in the limits of Predicted values.

Summary

Our multiple regression model performed an excellent job of identifying the primary elements influencing property values, with land size, house size, number of rooms, and bathrooms all showing as significant predictors. The resulting model has an adjusted R^2 of 0.6955, explaining approximately 69.55% of the variation in assessed values. While this is a robust conclusion, it is evident that other unmeasured factors influence property value. Following testing and refining, we discovered that included all four variables in the model resulted in the best fit, as proven by F-tests and other statistical measurements.

Even though our model did not exactly match all of the ideal assumptions—there was some little variability in the residuals and a few normality issues—it remains a solid predictor overall. The confidence and forecast intervals we produced are relatively accurate, reflecting the average range of uncertainty in real estate values. In general, land and home size had the greatest beneficial impact on evaluated value, while room and bathroom numbers had more nuanced benefits. These findings are consistent with our initial objectives and provide us with a useful tool for understanding what drives market property values.

****