**1. Introduction:**

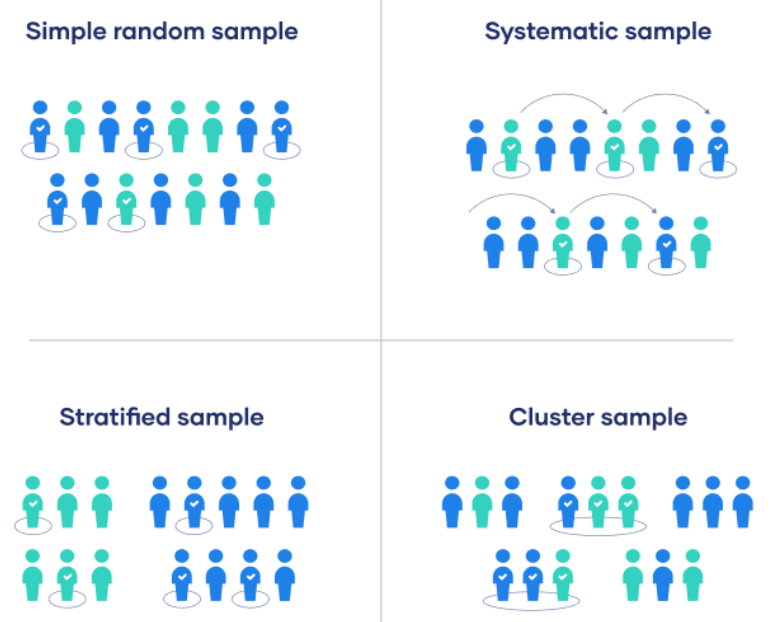
Data analysis plays a critical role in research, scientific experiments, and decision-making processes. One essential aspect of data analysis is data sampling, which determines the accuracy and precision of research or survey results. In this report, we will explore different sampling techniques such as probability and non-probability sampling methods using the Spring Houses dataset. We will also perform interval estimation, correlation, regression, and time-series analysis to gain insights into the dataset. We will interpret the results of the analyses, evaluate their usefulness in decision-making and discuss how decision-makers can mitigate the risks associated with the use of the data.

**2. Different sampling techniques:**

Data Sampling forms the essential part of most of the research, scientific and data experiments. It is one of the most important factors which determines the accuracy of our research or survey result. It can be classified in two ways as following

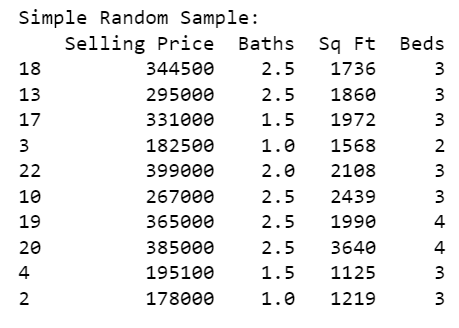
**2.1.** **Probability sampling methods:**

Probability sampling methods, such as simple random sampling, stratified sampling, systematic sampling, and cluster sampling, are designed to ensure that every member of the population has an equal chance of being selected in the sample. As a result, probability samples are generally considered to be representative of the population and can be used to make inferences about the population with a known degree of precision and confidence. By selecting a probability sample of houses from the Spring Houses dataset, we can be more confident that our findings from the sample can be generalized to the population of all houses for sale via that estate agent. Different types of probability sampling methods we can apply like



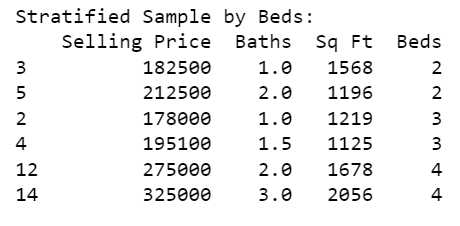
**Simple random sampling:**

In a simple random sample, every member of the population has an equal chance of being selected. Our sampling frame should include the whole population. To conduct this type of sampling, we have selected 10 records based entirely on chance from the data set.



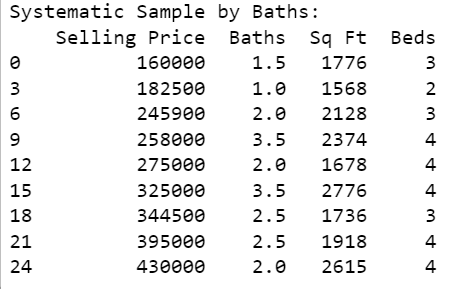
**Stratified sampling:**

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows us to draw more precise conclusions by ensuring that every subgroup is properly represented in the sample. To use this sampling method, I have divided the houses into subgroups based on the Beds then use randomly selected sample from each subgroup.



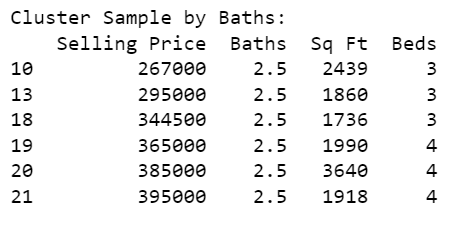
**Systematic sampling:**

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals. Here every 3rd row is selected from the data set.



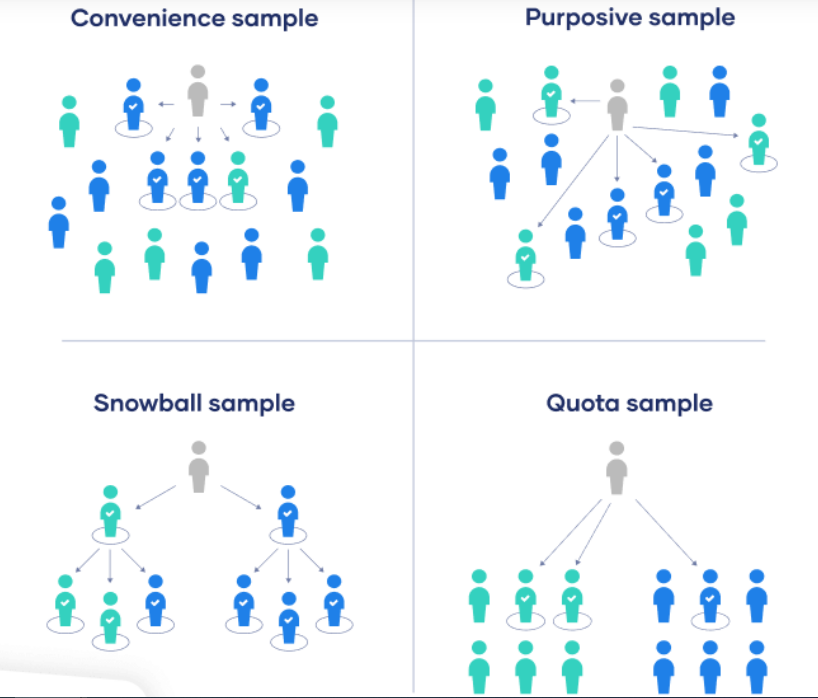
**Cluster sampling:**

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, we randomly select entire subgroups. To perform this type of sampling I did the grouping based on Baths.



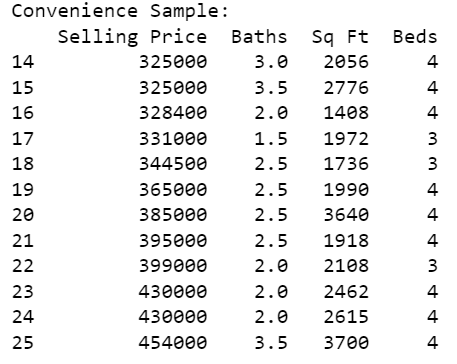
**2.2. Non Probability sampling methods:**

Non-probability sampling methods, such as convenience sampling, judgmental sampling, and quota sampling, do not ensure that every member of the population has an equal chance of being selected in the sample. As a result, non-probability samples may not be representative of the population, and the results obtained from them may not be generalizable to the larger population. However, non-probability sampling methods can still be useful in some cases where probability sampling is not feasible or practical. For example, if we want to explore the relationship between the selling price and square footage of houses for sale via a certain estate agent. Different types of non probability sampling methods we can apply like



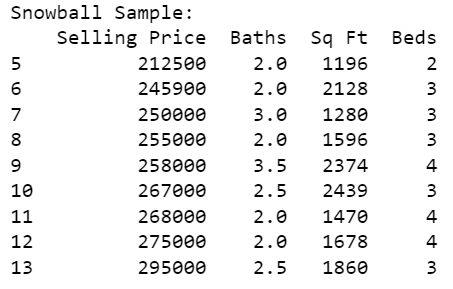
**Convenience sampling:**

A convenience sample simply includes the individuals who happen to be most accessible to the researcher. This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population. Here I have selected the records based on sampling which selling price is more than 300000.



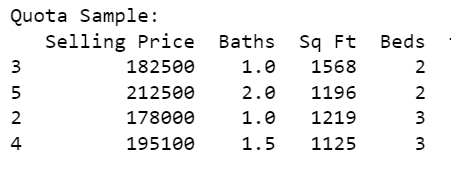
**Snowball Sampling:**

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people we have access to “snowballs” as we get in contact with more people.



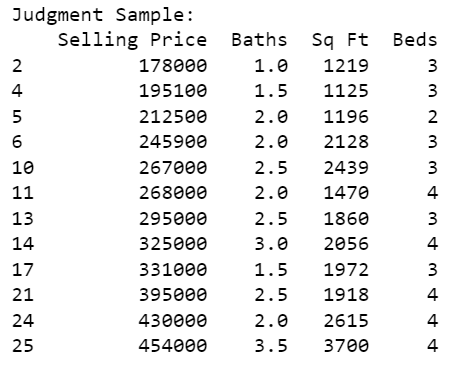
**Quota sampling:**

Quota sampling relies on the non-random selection of a predetermined number or proportion of units. This is called a quota. We first divide the population into mutually exclusive subgroups (called strata) and then recruit sample units until we reach our quota, here I have created the subgroups based on 2 and 3 Beds.



**Judgmental sampling:**

This type of sampling also known as purposive sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research. To perform this type of sampling I have selected the rows as per my judgment.



**3. Interval estimation:**

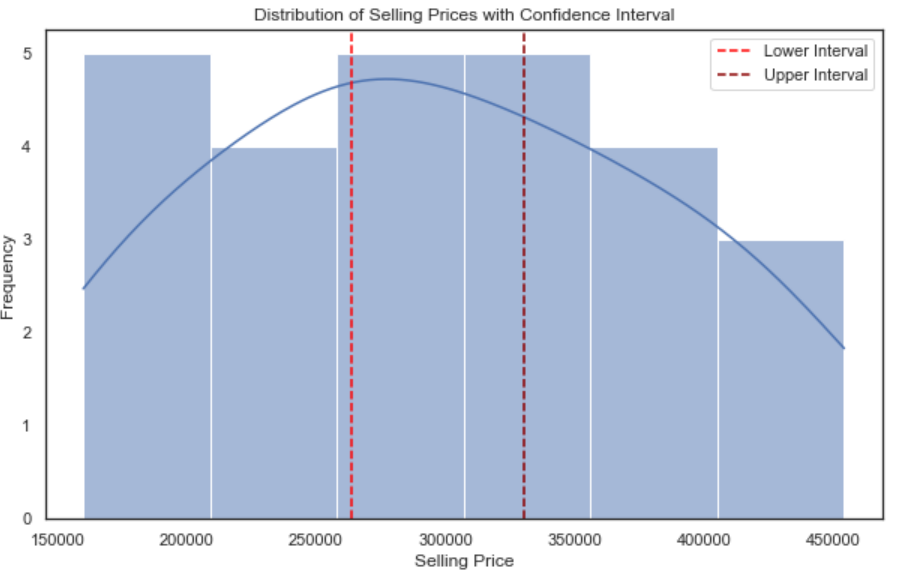
Interval estimation is a statistical technique used to estimate the range of values within which a population parameter is likely to fall, based on a sample of data. By computing a confidence interval for the mean selling price of houses in the dataset, we can estimate the range of selling prices that is likely to be representative of the population.

If we consider α which is the probability of a confidence interval will not include the population parameter is 0.05 then Z score will be 1.96 from Z score table for 95% (1-0.05 = 0.95) confidence interval.

Confidence Interval = mean ± z\_score \* standard\_error

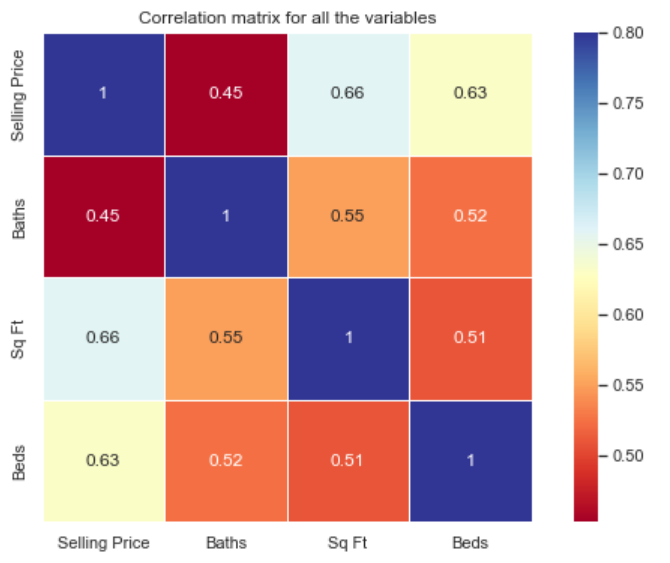
Based on the equation we can calculate the interval estimation for population mean Selling Price of the house using software tool.





**4. Correlation:**

Correlation analysis can help us understand the relationship between the variables in the dataset. For example, we can examine the correlation between the selling price and other variables such as the number of bathrooms, square footage, and bedrooms to see if there are any strong relationships between these variables.



Correlation is a statistical measure that describes the strength and direction of a linear relationship between two variables. Correlation values range from -1 to +1, where -1 indicates a perfectly negative correlation, +1 indicates a perfectly positive correlation, and 0 indicates no correlation.

In the given correlation matrix, the correlation values between Selling Price and the other variables are:

* Baths: 0.453097 (moderate positive correlation)
* Sq Ft: 0.658877 (moderately strong positive correlation)
* Beds: 0.630955 (moderate positive correlation)

These correlation values indicate that there is a positive relationship between Selling Price and each of these variables, but the strength of the relationship varies. Sq Ft has the strongest correlation with Selling Price, followed by Beds and then Baths.

**5. Regression analysis:**

Regression analysis can be used to create a model that predicts the selling price based on other variables in the dataset. By analyzing the coefficients of the model, we can understand which variables have the most significant impact on the selling price.

**5.1. Simple linear regression:**

At first, we are going to create model of simple linear regression model between Sq Ft and selling Price as this two variable have highest correlation coefficient (0.658877).

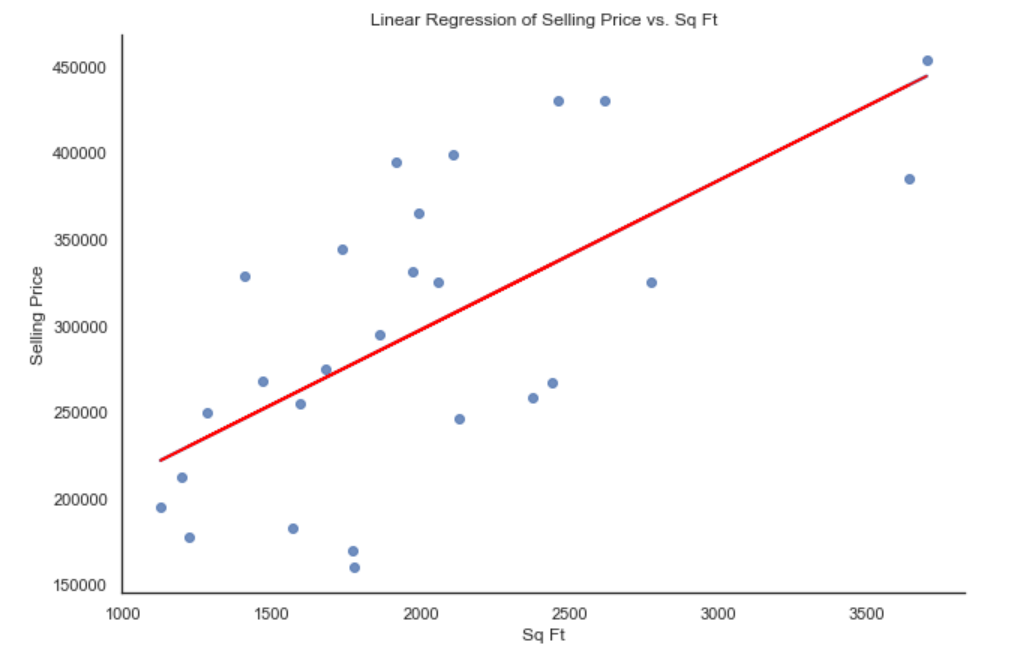
We can set the Hypothesis test this above simple linear regression model as

**The Null Hypothesis (H0):** There is no significant linear relationship between Selling Price and Sq Ft

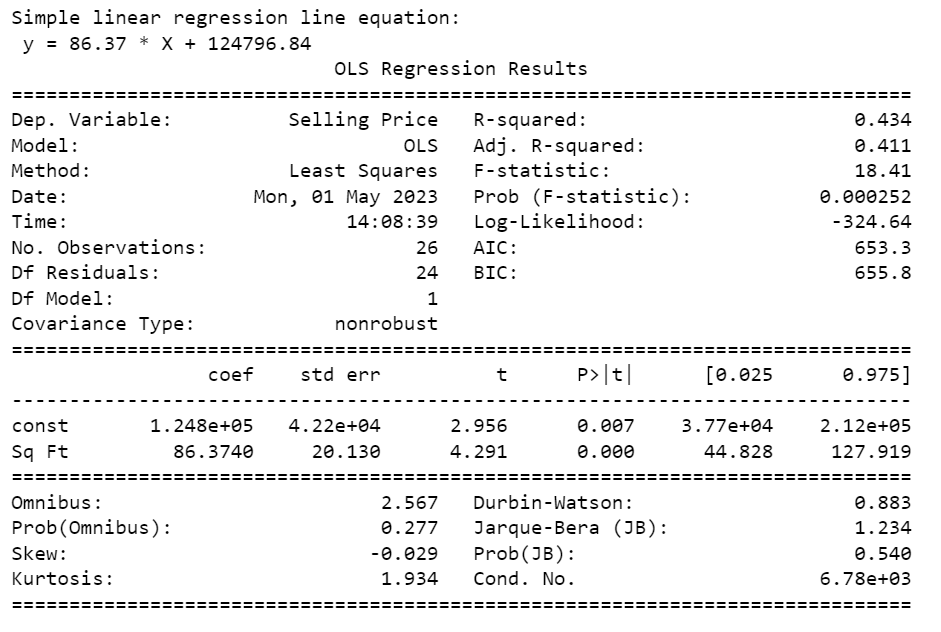
**The Alternative Hypothesis (Ha):** There is a significant linear relationship between Selling Price and Sq Ft

To test the hypotheses, we can use p-value and α value (for our analysis we made it as 0.05). The null hypothesis is true if p-value is greater than or equal to α value. If the p-value is less than the significance level which is α value, we will reject the null hypothesis and conclude that there is a significant linear relationship between the two variables.

Based on simple regression model I have plotted it below with best possible Least square regression line



Now we can check the summary of this model to validate the hypothesis test and evaluate model performance using r squared value and anova test.



From this result we can extract the following insights

* The p-value for the coefficient of Sq Ft is less than the significance level of 0.05, indicating that there is a statistically significant linear relationship between Selling Price and Sq Ft. Therefore, we can reject the null hypothesis that there is no relationship between the two variables, and accept the alternative hypothesis that there is a linear relationship between Selling Price and Sq Ft.
* The R-squared value of 0.434 indicates that approximately 43.4% of the variation in Selling Price can be explained by Sq Ft. This means that Sq Ft is a significant predictor of Selling Price, but there are other factors that are also important in determining Selling Price.
* The F-statistic of 18.41 and its associated p-value of 0.000252 indicate that the overall model is statistically significant, which means that the linear relationship between Selling Price and Sq Ft is not due to chance.
* The model intercept, or constant, is 1.248e+05. This means that if the Sq Ft of a property is zero, the predicted Selling Price would be approximately $124,800.
* The coefficient for Sq Ft is 86.3740. This means that for every additional square foot of living space, the predicted Selling Price increases by approximately $86. The coefficient's 95% confidence interval is (44.828, 127.919), which means that we can be 95% confident that the true population coefficient lies within this interval.
* Overall, the model provides a reasonable fit to the data, as indicated by the statistically significant p-value and F-statistic. However, the low R-squared value suggests that Sq Ft alone may not be sufficient to accurately predict Selling Price, and there may be other important factors that need to be included in the model.

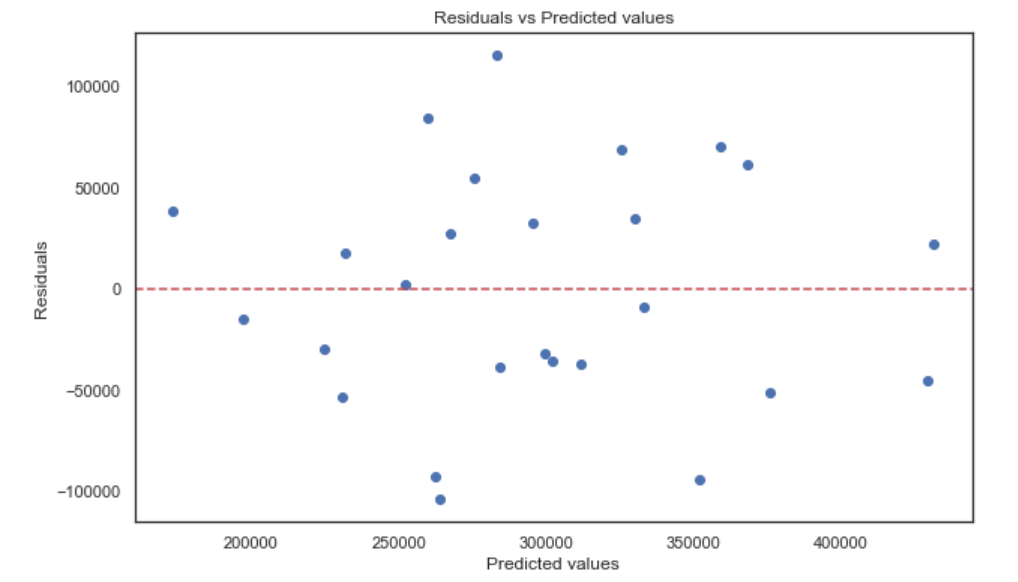
**5.2. Multiple linear regression:**

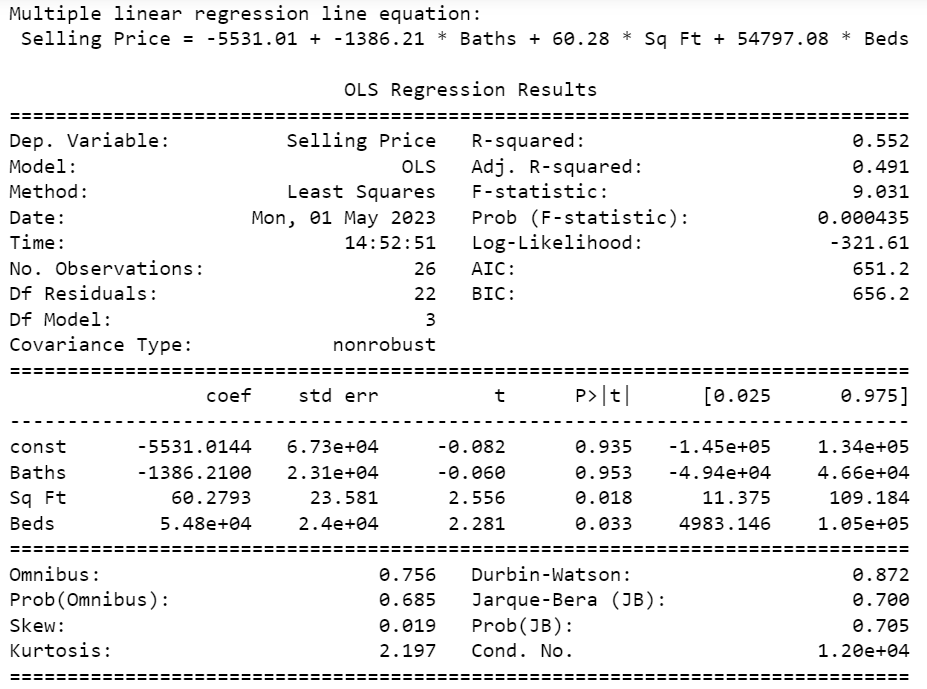
As our previous analysis we came to know that only Sq Ft will not be sufficient to predict Selling Price, we will need to perform multiple linear regression including other predictor variables Beds and Baths as well.

We can set the Hypothesis test this above multiple linear regression model as

**The Null Hypothesis (H0):** The coefficients of all the predictor variables (Baths, Sq Ft, Beds) are equal to zero, implying that there is no significant linear relationship between Selling Price and any of the predictor variables.

**Alternative Hypothesis (Ha):** At least one of the coefficients of the predictor variables (Baths, Sq Ft, Beds) is not equal to zero, implying that there is a significant linear relationship between Selling Price and at least one of the predictor variables. I have plotted Residuals from the multiple regression model against Predicted value below.





From the above result we can extract the following insights:

* The p-value for the F-statistic is 0.000435, which is less than the significance level of 0.05. Therefore, we can reject the null hypothesis and conclude that there is a statistically significant relationship between the predictor variables (Baths, Sq Ft, Beds) and the target variable (Selling Price).
* The R-squared value of 0.552 indicates that the model explains 55.2% of the variability in the target variable.
* The adjusted R-squared value of 0.491 suggests that the model is not a perfect fit, but still reasonably good. Also multiple liner regression model R-squared value (0.491) greater than previously designed simple liner regression model R-squared value (0.411) which indicate that multiple liner regression model is better suited in terms of predicting selling price.
* The p-values for the individual predictor variables indicate that Sq Ft and Beds are statistically significant, while Baths is not.
* Overall, the model appears to have some predictive power, but further analysis is needed to evaluate its performance.

**6. Time series analysis:**

Time series analysis is a statistical technique used to analyze time series data, which are observations collected at regular intervals over time. It involves studying the patterns and trends in the data to make predictions about future values. Time series forecasting is the process of using a model to predict future values of a time series based on past observations.

Unfortunately, the Spring Houses dataset does not contain any time series data, so it would not be appropriate to apply time series analysis techniques to this dataset. It requires data to be collected at regular intervals over time to identify trends, patterns, and cycles in the data which is missing in spring house data set.

**7. Risk analysis by decision makers:**

Based on the analysis of the Spring Houses dataset, decision-makers can take several steps to mitigate the risks associated with using the dataset:

* Ensure data quality: Decision-makers should verify the accuracy and completeness of the dataset before using it for any analysis. They should also check for any outliers or missing values in the dataset that can impact the analysis results. This can be done by performing data cleaning and validation steps.
* Use multiple analysis techniques: To mitigate the risk of making incorrect conclusions, decision-makers should use multiple analysis techniques on the Spring House dataset. By applying different statistical methods, they can cross-validate their findings and ensure that their conclusions are robust and reliable.
* Incorporate domain knowledge: Decision-makers should incorporate domain knowledge when analyzing the Spring House dataset. By combining their expertise in the domain with the insights gained from the analysis, they can make more informed decisions. This can be done by involving subject matter experts in the analysis process or by consulting relevant literature to gain additional context.

**8. Conclusion:**

This report explains the importance of interval estimation, correlation analysis, and regression analysis to identify the relationship between variables and predict future values. The results of the analyses suggest that multiple linear regression with predictor variables Sq Ft and Beds is a better predictor of selling price than simple linear regression with only Sq Ft as a predictor. The article also highlights the importance of evaluating the model's performance and the potential risks associated with using data, such as sampling biases and insufficient data for analysis. Decision-makers can mitigate these risks by selecting appropriate sampling techniques and validating the results with additional analyses. Overall, these statistical techniques can provide valuable insights for decision-making, such as forecasting and predicting future trends about selling price of houses.