**HOUSING DATA ANALYSIS**

**Alekhya Reddy Pasarakayala**

**TABLE OF CONTENTS**

**INTRODUCTION & METHODOLOGY**..................................................................................3

**Descriptive Statistics**.....................................................................................................................3

**ANOVA 1**.......................................................................................................................................4

**ANOVA 2**.......................................................................................................................................5

**ANOVA 3**.......................................................................................................................................6

**Linear Regression 1**

Model Accuracy..................................................................................................................8

**Linear Regression 2**

Model Accuracy................................................................................................................10

**Linear Regression 3**

Model Accuracy................................................................................................................13

**Analysis**........................................................................................................................................14

**Conclusion**....................................................................................................................................15

**INTRODUCTION & METHODOLOGY**

**ANOVA:**

The Analysis of variance is the full form of ANOVA. It is used to generate random variables based on a tool. To determine whether there is any existence of common tools this is used to relate to each other. Three ANOVA types exist. They are Random effect, fixed effect, and mixed effect. ANOVA has made some assumptions which mean the population of a condition, variance analysis, and the population parameters with the model structure and the exact description made. There is also a normality score which is given importance by the ANOVA. The correlations that come from the independent and related to each observation have given importance in the analysis.

**REGRESSION:**

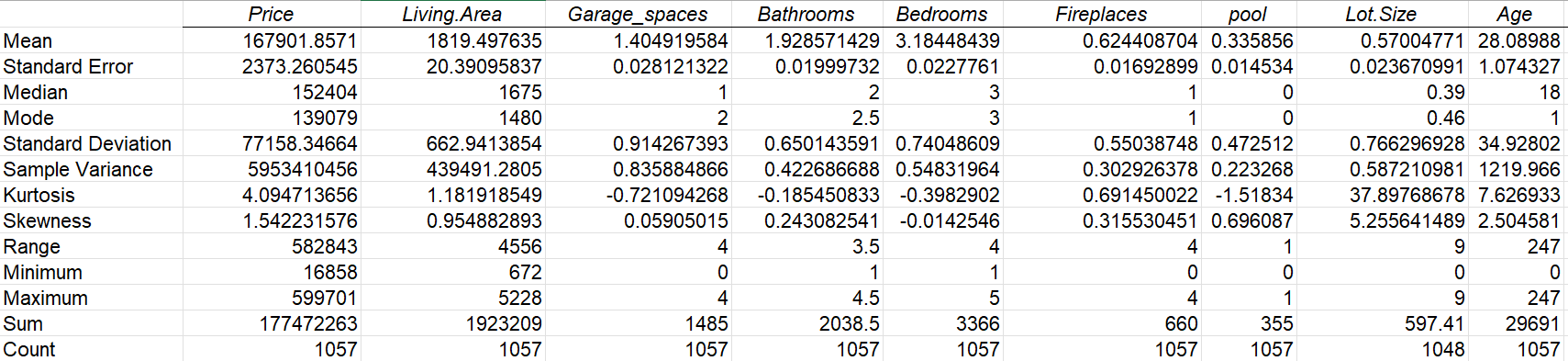
It is a powerful tool for determining the relation between variables that are effective. Only dependent variables have been done with the regression analysis and one independent variable. Here, in this regression analysis, there has been a qualitative variance present with the variables. The correlation coefficient square is equal to the shared proportion variance. There are many approaches that can be used to measure the variability of the amount which is given in several sets. All the methods mentioned are important to notice the desired first and last quartile, median, maximum, and minimum values of the analysis. Standard deviation, the sum of squared deviations, and multiple variance analysis are all important models for the analysis which can be considered rest are ignored.

Research Questions and Hypothesis This research study aims in identifying various attributes that contribute to predicting the price of houses. I have identified some research questions for this study.

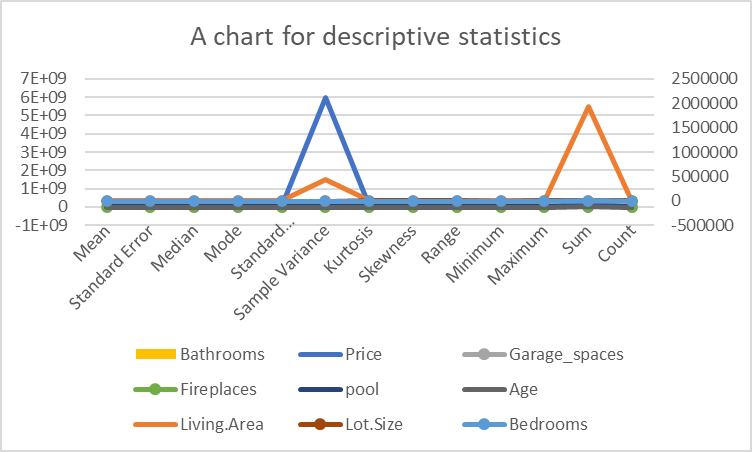
**ABOUT SAMPLE DATASET:**

The Housing dataset contains 1057 rows and 10 columns as attributes. Price, Living Area, Garage Space, Bathrooms, Bedrooms, Fireplaces, Pool, Lot Size, Age, Fireplaces (Boolean value). As a part of data cleaning, I have removed null values present in the Lot size and replaced Fireplaces with True and Fase as 0 and 1 respectively.

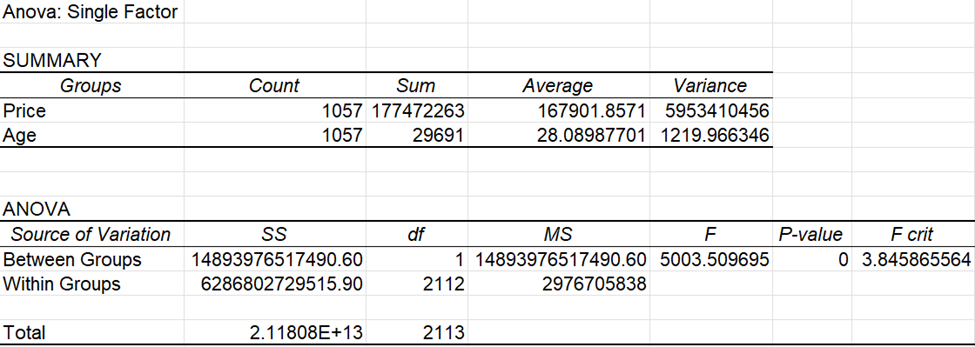
**DESCRIPTIVE STATISTICS:**



Descriptive Statistics provide a clear analysis of the whole data.



**1) Is there a significant difference in the average price of the houses based on the age of the building using Anova?**

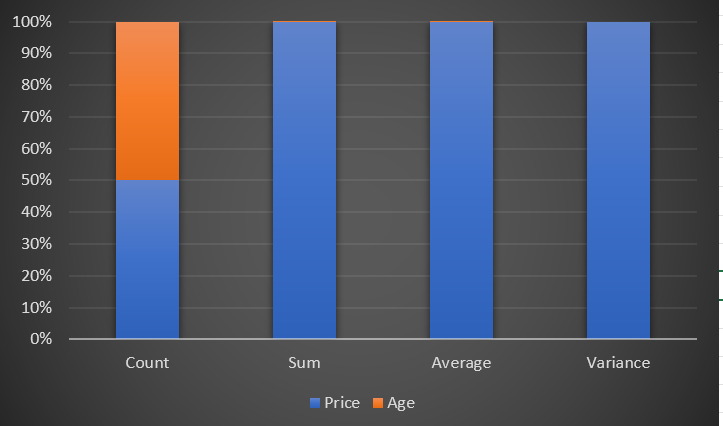


The Anova generated is for single factor. This test provides if there is a significant relation between the Price of property and the Age of the Building. The output says that there is a short significant relationship between means because the P-value is zero

Chart, line chart

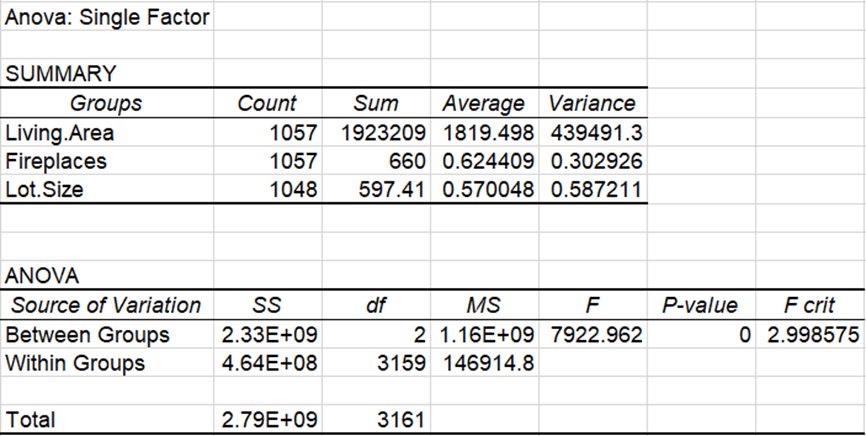
Description automatically generated

A graphical representation of between groups and within groups which says that highly significant



Here is the bar graph of the ANOVA analysis for the Price and Age where the count of these groups is equal which is 50%, and the sum, average, and variance represent the 100% analysis.

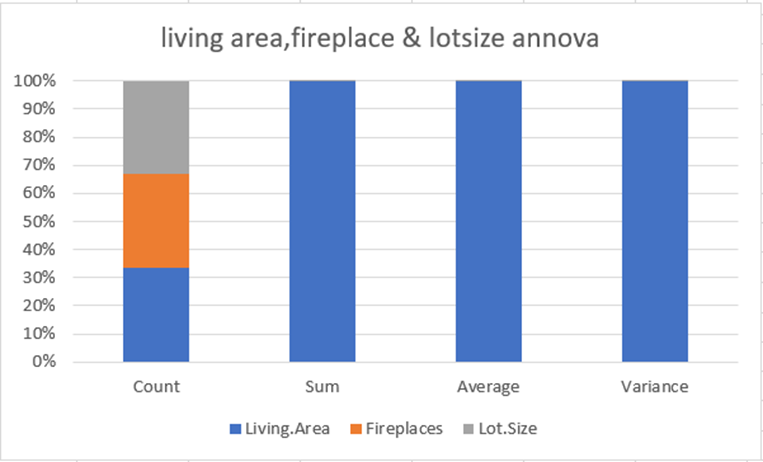
**2) Identify and analyze ANOVA for columns living area, fireplaces, and lot size**



Here is the single factor ANOVA analysis for the columns of living area, fireplaces, and the lot size where the summary output represents the count, sum, average, and variance of the living area, fireplaces, and the lot size.

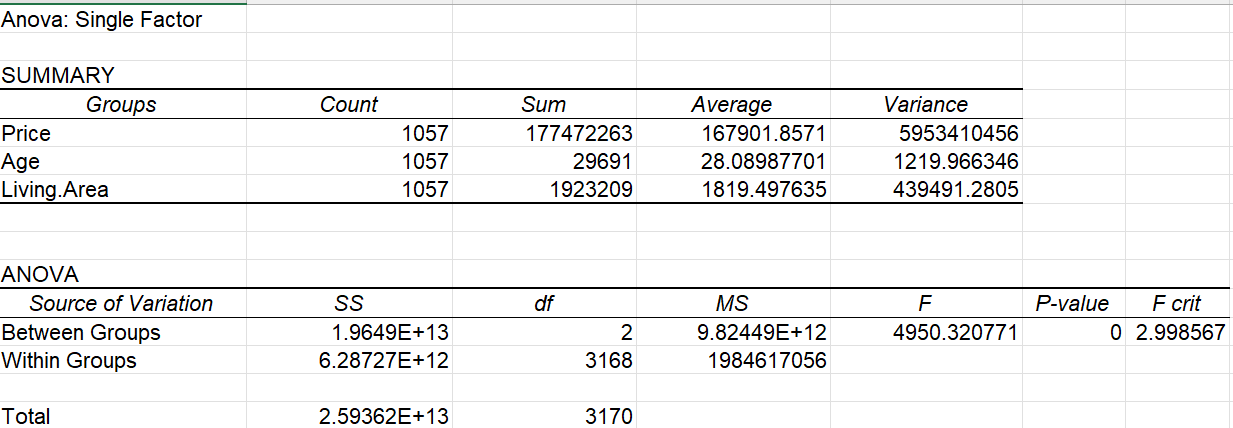
From the ANOVA output, I can observe the source of variation of both between groups and within groups with the sum of squares, degrees of freedom, mean square, p-value, and the F and F critic.

**Graph:**



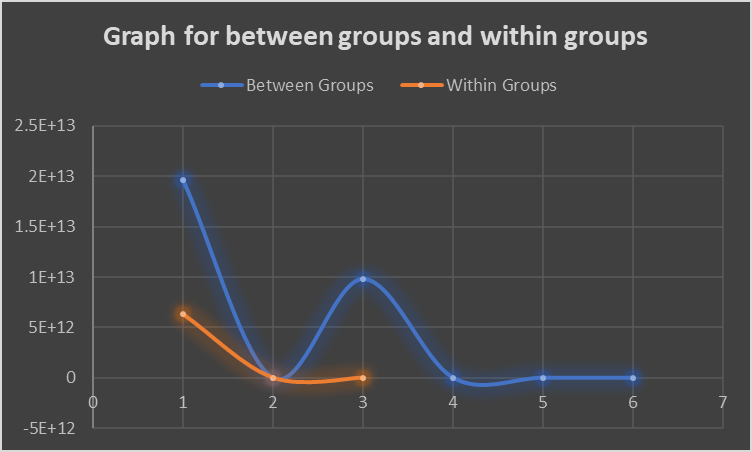
Here is the bar graph of the ANOVA analysis for the living area, fireplace, and lot size where the count of these groups is equal for all three which is 33.3%, and the sum, average, and variance represent the 100% analysis.

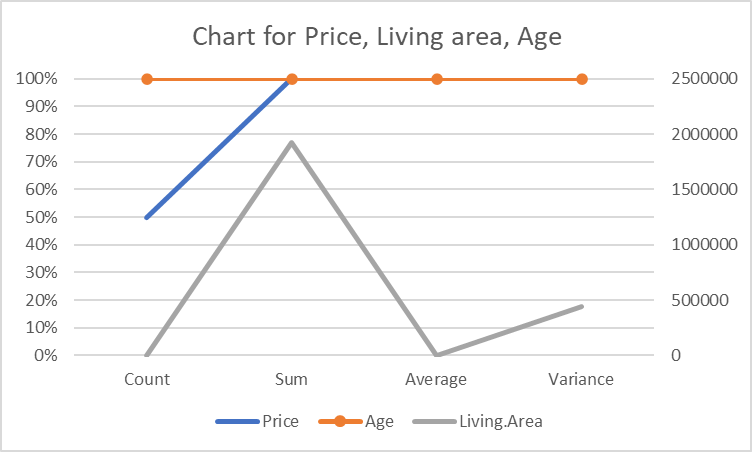
**3) Price of property depends on living area or age or both, is there any significant relation between mean price of property in age or living area groups?**



The Anova generated is for single factor. This test provides if there is a significant relation between the Price of property, Age, Living area. The output says that there is a short significant relationship between mean because P-value is zero.

A graphical representation of between groups and within groups which says that between groups highly significant compared to within groups.



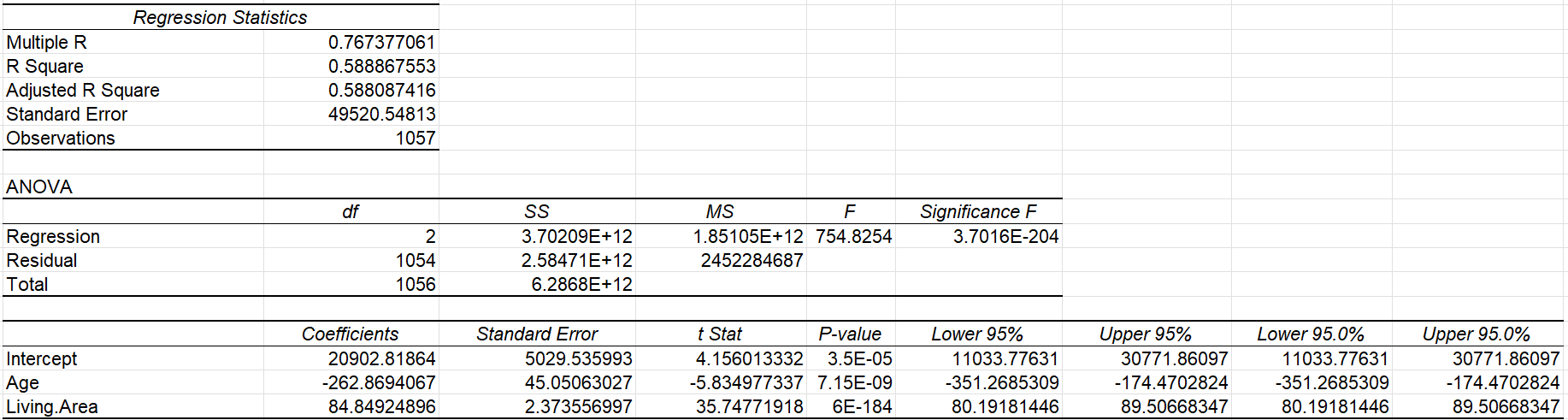


**4) How are age and living area related to price of house. Conclude with a linear regression for price based on age and living are for 95%.**

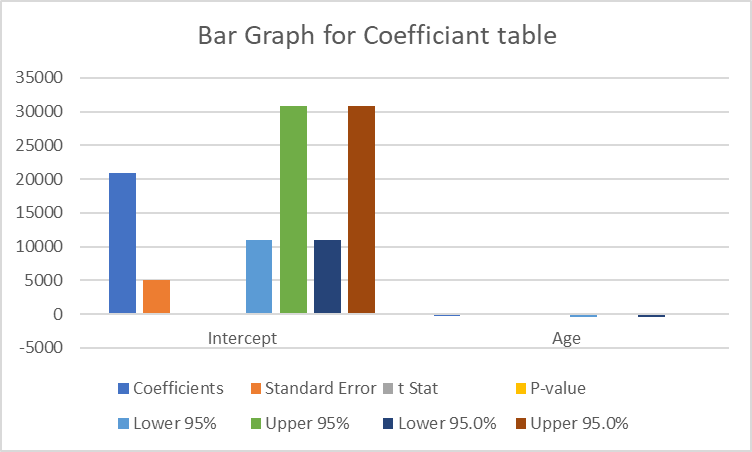
Here is a linear regression summary of Age, living area and price with 95% confidence level.

|  |  |
| --- | --- |
| SUMMARY OUTPUT |  |
|  |  |
| *Regression Statistics* | |
| Multiple R | 0.767377061 |
| R Square | 0.588867553 |
| Adjusted R Square | 0.588087416 |
| Standard Error | 49520.54813 |
| Observations | 1057 |

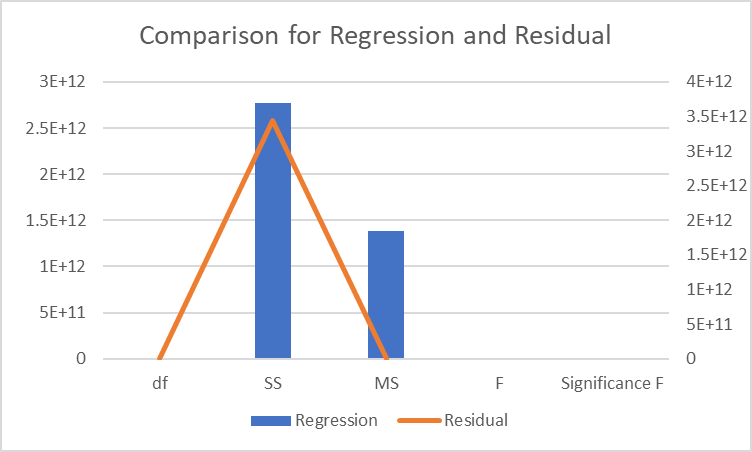
This explains the relationship between price of property and 2 independent variables Age, living area. The dataset consists of 1057 observations. The Multiple R value of 0.7674 says that there is a strong correlation between Age, living area and Price of property. The R-square value is 58.9% which means that the model is a good fit as the difference between R square and Adjusted R square is negligible.



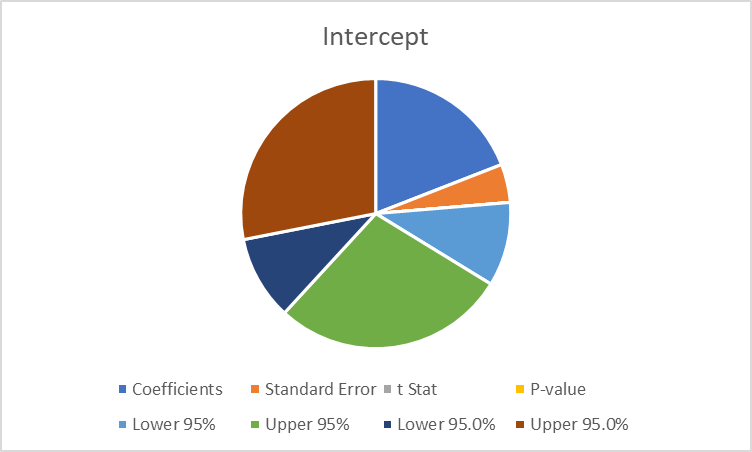
Here is the output of Regression Statistics which provides analysis of independent variables Age, living area and dependent variables Price of property. The coefficient table explains the values of coefficients, Standard Error, t Stat, P-value, Lower 95%, Upper 95% values. Age is the negative relationship against price and Living area is acting as positive relation.



A Chat for linear regression which explains the comparison between different variables.



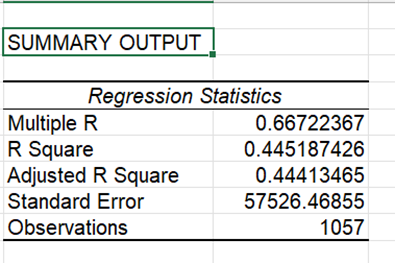
A Graphical representation of Regression and Residual. From this graph I can understand that Regression is higher compared to Residual.



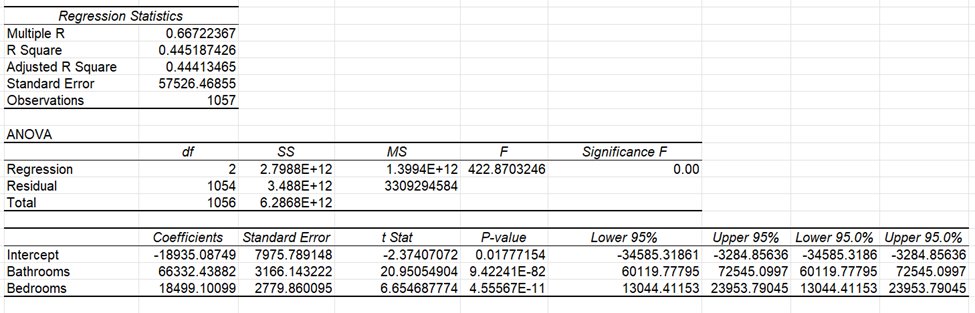
A pie chart which shows the intercept values of all values together.

**5) How well does a linear regression model predict the prices of houses based on the number of bedrooms and bathrooms?**

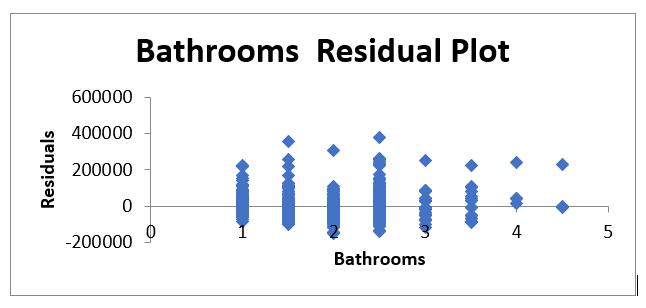
Here is a linear regression summary of Bedrooms, Bathrooms, and Price with 95% confidence level



This explains the relationship between price of property and 2 independent variables Bedroom and Bathrooms. The dataset consists of 1057 observations. The Multiple R value of 0.6672 says that there is a strong correlation between Bedrooms, Bathrooms and Price of property. The R-square value of 44.5% means that the model is a good fit as the difference between R square and Adjusted R square is negligible.



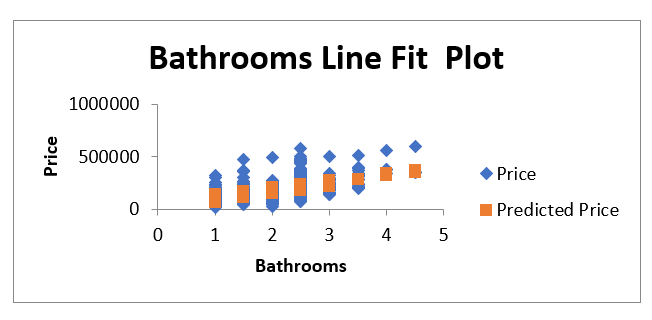
Here is the output of Regression Statistics which provides analysis of independent variables Bathrooms, Bedrooms, and dependent variables Price of property. The coefficient table explains the values of coefficients, Standard Error, t Stat, P-value, Lower 95%, Upper 95% values.



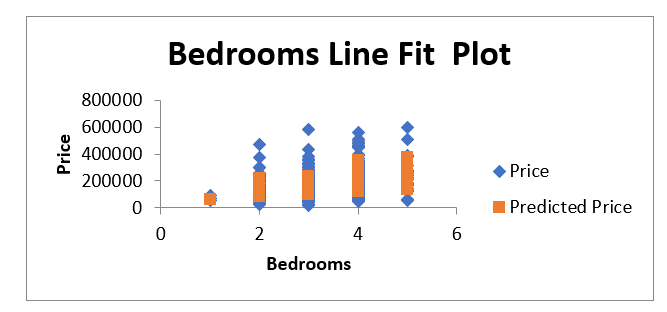
Here is the line chart of the linear regression analysis of the Bathrooms residual plot where the Bathrooms are represented on the x-axis and the Residuals on the y-axis.



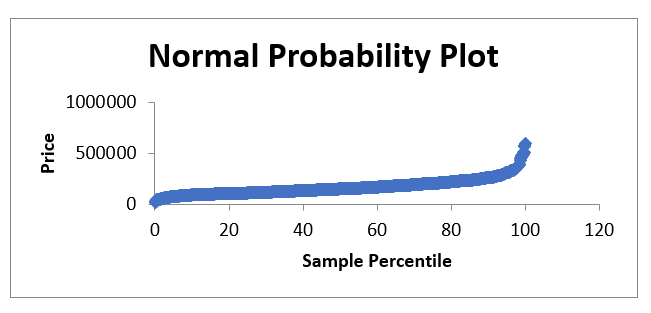
Here is the line chart of the linear regression analysis of the Bedrooms residual plot where the Bedrooms are represented on the x-axis and the Residuals on the y-axis.



Here is the Bathrooms line fit plot of a linear regression analysis where the Bathroom represent the x-axis and the Price on the y-axis

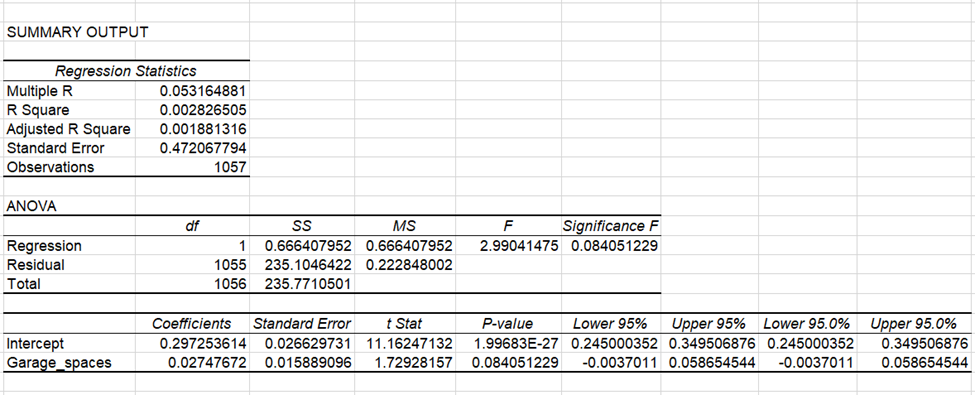


Here is the Bedrooms line fit plot of a linear regression analysis where the Bathroom represent the x-axis and the Price on the y-axis.



Here is the normal probability plot of a linear regression analysis with the sample percentile on the x-axis and Price on the y-axis.

**6) Find out the Linear regression analysis for the pool and garage based on price with some graphs?**



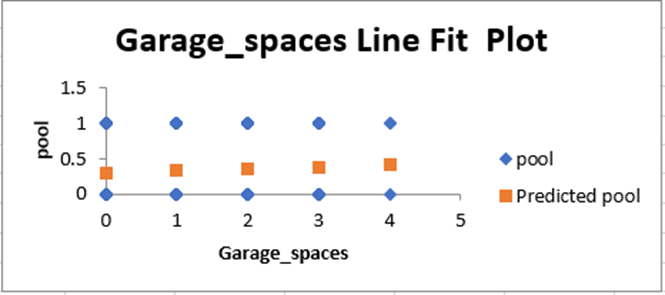
Here, I have performed a linear regression analysis for the pool and garage based on the price of the building. From the regression statistics, I can observe that I have a total of 1057 observations and the values I got from the multiple R, R square, Adjusted R square, and Standard error says the correlation between pool and garage spaces. And the Anova output represents the output of regression and residual of degrees of freedom, the sum of squares, mean squares, and the significance of F.

The regression statistics output provides the analysis of coefficients, standard error, t stat, p-value, lower 95%, upper 95%, lower 95.0%, and upper 95.0% for both the intercept and garage spaces.

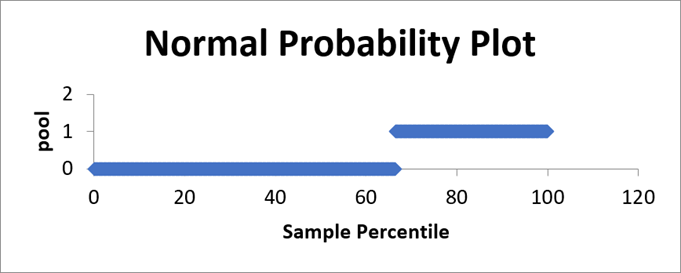
**Graphs:**



Here is the line chart of the linear regression analysis of the Garage spaces residual plot where the Garage spaces are represented on the x-axis and the Residuals on the y-axis.



Here is the Garage spaces line fit plot of a linear regression analysis where the garage spaces represent the x-axis and the pool on the y-axis.



Here is the normal probability plot of a linear regression analysis with the sample percentile on the x-axis and pool on the y-axis.

**ANALYSIS:**

Based on complete analysis I can say that our models are significantly different. Anova provides an understanding of various factors regarding property prices. Based on our first analysis comparison between groups and within groups the output says that the significance of these factors is based on property price. Later I conducted Anova analysis on living area, fireplaces, and lot size which reveals that variation between groups and within groups are significant of these factors for determining property price.

Linear regression analysis predicts that a strong correlation between living area, age, property price. The model is a good fit as R square and Adjacent R square value are having same value around 58.8%. As per the next regression analysis relation between price, bedrooms and bathrooms are significantly correlated with a value of 44.5%. The visualization plots explain the comparison of different independent variables with respect to dependent variables. Analysis of pool and garage spaces assess their impacts on property price. The graphs provide insights on coefficients, standard error and significance of these variables.

Descriptive Statistics is used for summarizing the data and providing a clear understanding of Dataset. Count, sum, average and variance are used to say about features and distribution of variables. The graphical visualization is used to illustrate different Anova and linear regression outputs. Overall, the analysis provides how different attributes are impacting property prices.

**CONCLUSION**

I have proposed different models to check whether there is a significant relation between the different attributes using ANOVA analysis and multiple regression and all our models are found to be considerably significant.

There is a significant relation between Price and Age with F-statistic of 5003.50 and p-value as 0.

The F-statistic and p-values of model with Living area, fireplaces and Lot size is 7992.96, 0 respectively and I can say that the model is significant.

The model to check whether the Price of property depends on Living area, age or both has F-statistics as 4950.32 and p-value as 0, and I can conclude the model is significant.

**REFERENCES:**

1. Freund, R. J., Wilson, W. J., & Sa, P. (2006). *Regression analysis*. Elsevier.
2. Cardinal, R. N., & Aitken, M. R. (2013). *ANOVA for the behavioral sciences researcher*. Psychology Press.