

**ASSIGNMENT**

**AQ049-3-M-MMDA**

**MULTIVARIATE METHODS FOR DATA ANALYSIS**

**HAND IN DATE: *16 November 2022***

**LECTURER NAME: LOW KOK SUN**

**STUDENT NAME: MOHAMED ZAHID**

**TP NUMBER: TP067803**

**HOUSE PRICE PREDICTION REGRESSION MODEL**

**TABLE OF CONTENTS**

LIST OF FIGURES.........................................................................................................................1

**SECTION - A**

PROBLEM STSTEMENT..............................................................................................................2

INTRODUCTION...........................................................................................................................2

RESEARCH OBJECTIVES............................................................................................................2

ANALYSIS, RESULTS, AND INTERPRETATIONS..................................................................4

CONCLUSIONS AND RECOMMENDATIONS........................................................................18

REFERENCES..............................................................................................................................18

**SECTION - B**

PURPOSE OF FACTOR ANALYSIS..........................................................................................19

EXCLUDE NON-METRIC VARIABLES...................................................................................19

FACTOR ANALYSIS ON SPSS..................................................................................................19

INTERPRETATIONS...................................................................................................................24

ASSESSMENT..............................................................................................................................24

**LIST OF FIGURES**

Fig.1 – Variable view of house price dataset in SPSS.....................................................................4

Fig.2 – Data view of house price dataset in SPSS...........................................................................5

Fig.3 – Pearson correlation coefficients...........................................................................................6

Fig.4 – P-P plot................................................................................................................................7

Fig.5 – Scatter plot of residuals.......................................................................................................7

Fig.6 – Checking for multicollinearity.............................................................................................8

Fig.7 – Model summary, ANOVA table, and variable particulars for variable x4 in step 1...........9

Fig.8 – Excluded variables in model 1...........................................................................................10

Fig.9 – Overall model summary of stepwise linear regression......................................................11

Fig.10 – Coefficients of variables in the model at each step.........................................................15

Fig.11 – ANOVA analysis for all models......................................................................................17

Fig.12 – Bartlett’s Test..................................................................................................................19

Fig.13 – Anti-image correlation.....................................................................................................20

Fig.14 – Eigen values.....................................................................................................................20

Fig.15 – Scree plot.........................................................................................................................21

Fig.16 – Communalities.................................................................................................................22

Fig.17 – Rotated component matrix...............................................................................................23

**SECTION - A**

1. **PROBLEM STSTEMENT**

This study's primary objective is to create a multiple linear regression model to forecast home prices based on a number of significant variables identified by a stepwise regression analysis. In addition, metric independent variable factor analysis will be performed to see how they might be grouped or decreased to a smaller number. The data set utilized includes 1460 records and 32 variables that are all of a numerical character from Kaggle.

1. **INTRODUCTION**

The house pricing data collection contains details about numerous homes, including the year of construction, the year of remodeling, the size of the garage, the square footage of the basement, the price of extra amenities, the month and year that the home was sold, and the sales price of the home. By using stepwise linear regression on SPSS, this study aims to identify which of these factors is linearly associated to the price. It is possible to think of the variables as belonging to several categories, and factor analysis may be performed.

The rising cost of housing makes the creation of a system that can predict future home prices necessary. Whether deciding when to buy a property, the developer and the customer can both benefit from the capacity to anticipate house values.

1. **RESEARCH OBJECTIVES**

The purpose of this study is to utilize stepwise linear regression to pinpoint the variables that, with a high degree of confidence, may account for the majority of variance in sales price of a home. Consequently, the goals of this investigation are:

1. To ascertain whether the size or area of the garage has an impact on the cost. The assumption is that a larger house will sell for more money.

2. To determine if number of fireplaces in a house is linearly related to the price of it. It is expected that more the number of fireplaces in a house, higher is its price.

3. To determine if above grade living area i.e., area of living that is above ground level, affects the overall sales price. It is expected that higher the value of this variable, the higher is the sales price.

As a result, the suggested null hypotheses for these three goals are shown below-

1. *Ho*:The area of garage has a negative linear relationship with sales price.

2. *Ho*:The number of fireplaces in a house has a negative linear relationship with sales price.

3. *Ho*:The area of living above grade has a negative linear relationship with sales price.

Alternative hypotheses can be put forth if significant results (p-value < 0.05) are obtained in opposition to these null hypotheses.

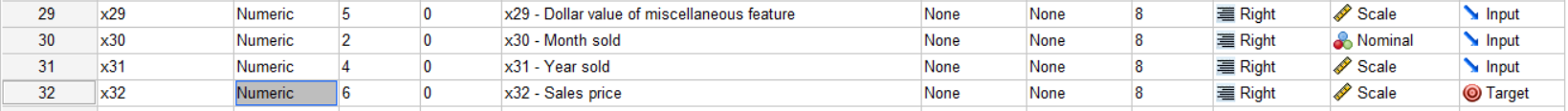
1. **ANALYSIS, RESULTS, AND INTERPRETATIONS**

**4.1. SUMMARY OF DATASET**

The dataset has 1460 rows and 31 columns, is entirely composed of numeric data, and lacks any category information. The variable names were provided from x1 to x32, and their labels were given in line with the attribute data received from the dataset's source after the cleaned dataset was loaded into the IBM SPSS statistics application for analysis. The variable view in SPSS is used to summarize this data.

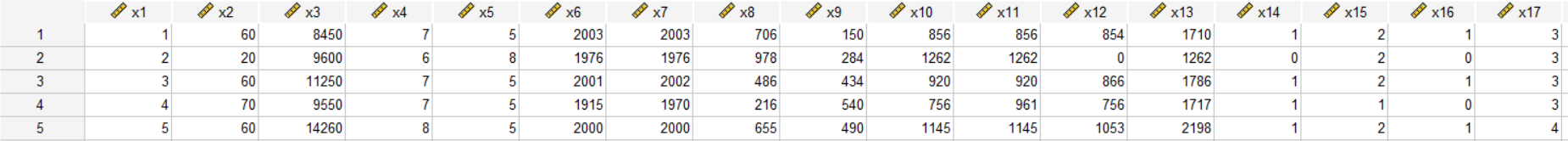
. Graphical user interface, application, table

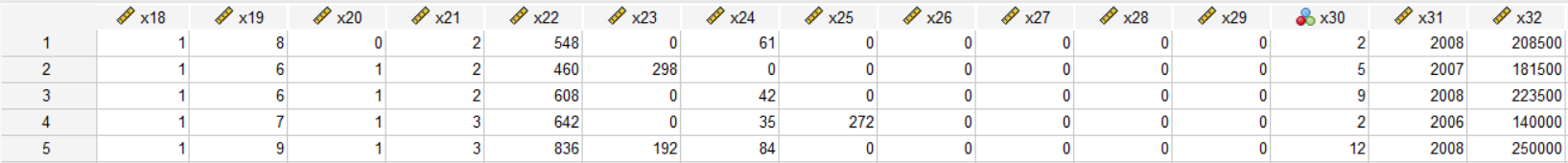
Description automatically generated



*Fig.1 –* *Variable view of house price dataset in SPSS.*

The first five rows of the dataset are shown below.





*Fig.2 –* *Data view of house price dataset in SPSS.*

For the multiple linear regression to describe the dependent variable, price, y = x32, only continuous numerical features will be used as independent variables. **Therefore, every variable in the dataset is utilized, with the exception of x1 (Symbolling) and x30 (Month sold).**

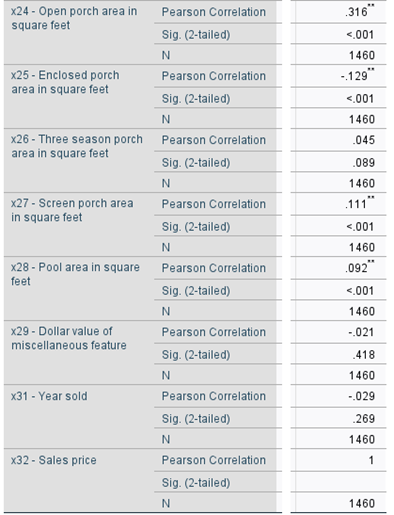
**4.2. STEPWISE MULTIPLE LINEAR REGRESSION**

**Correlation Analysis** - The sample appropriateness of the data must be evaluated before to doing multiple linear regression by computing the Pearson correlation coefficient between the dependent variable and independent variables. The bivariate Pearson correlation was run on the continuous variables.

Table

Description automatically generated Table

Description automatically generated



*Fig.3 – Pearson correlation coefficients. Correlation values greater than or equal to 0.5 are highlighted.*

The sample size of 1460 and 29 variables satisfies the minimal requirement of a ratio of 5:1 between observations and variables. **As a result, the sample was deemed to be considerably sufficient to move on to the following step of analysis.**

**Assumptions of Regression Analysis** – The residuals from regression should adhere to a normal distribution in order to draw reliable conclusions from it. The residuals are made up of the error terms, or the differences between the dependent variable's predicted and observed values. We may determine if the residuals are normally distributed by examining a typical Predicted Probability (P-P) plot. If so, they will do so by moving down the diagonal normality line seen in the plot.

Whether these residuals are homoscedastic or exhibit a tendency to cluster at certain values while dispersion broadly at other values is determined by homoscedasticity. When data resembles a shotgun discharge of randomly dispersed data, it is homoscedastic. The residuals and projected values are shown on a scatterplot to verify this premise.

When a regression is said to be linear, it signifies that the connection between the predictor variables and the result variable is linear. **Linearity is not a concern if the residuals are normally distributed and homoscedastic.**

Multicollinearity occurs when there is a significant connection between your predictor variables. This is an issue since the regression model won't be able to accurately connect the fluctuation in the predictor variable of your outcome variable to it, leading to muddled results and incorrect inferences. Only a multiple linear regression with several predictor variables may make use of this premise.

**It is sufficient to evaluate the regression's assumptions using a normal P-P plot, a scatterplot of the residuals, and VIF values.**

For normality test, we look at P-P plot as shown.

Chart, line chart

Description automatically generated

*Fig.4 – P-P plot.*

**There is a minor discrepancy, but it is acceptable. As long as there are no significant deviations, normality is presumed.**

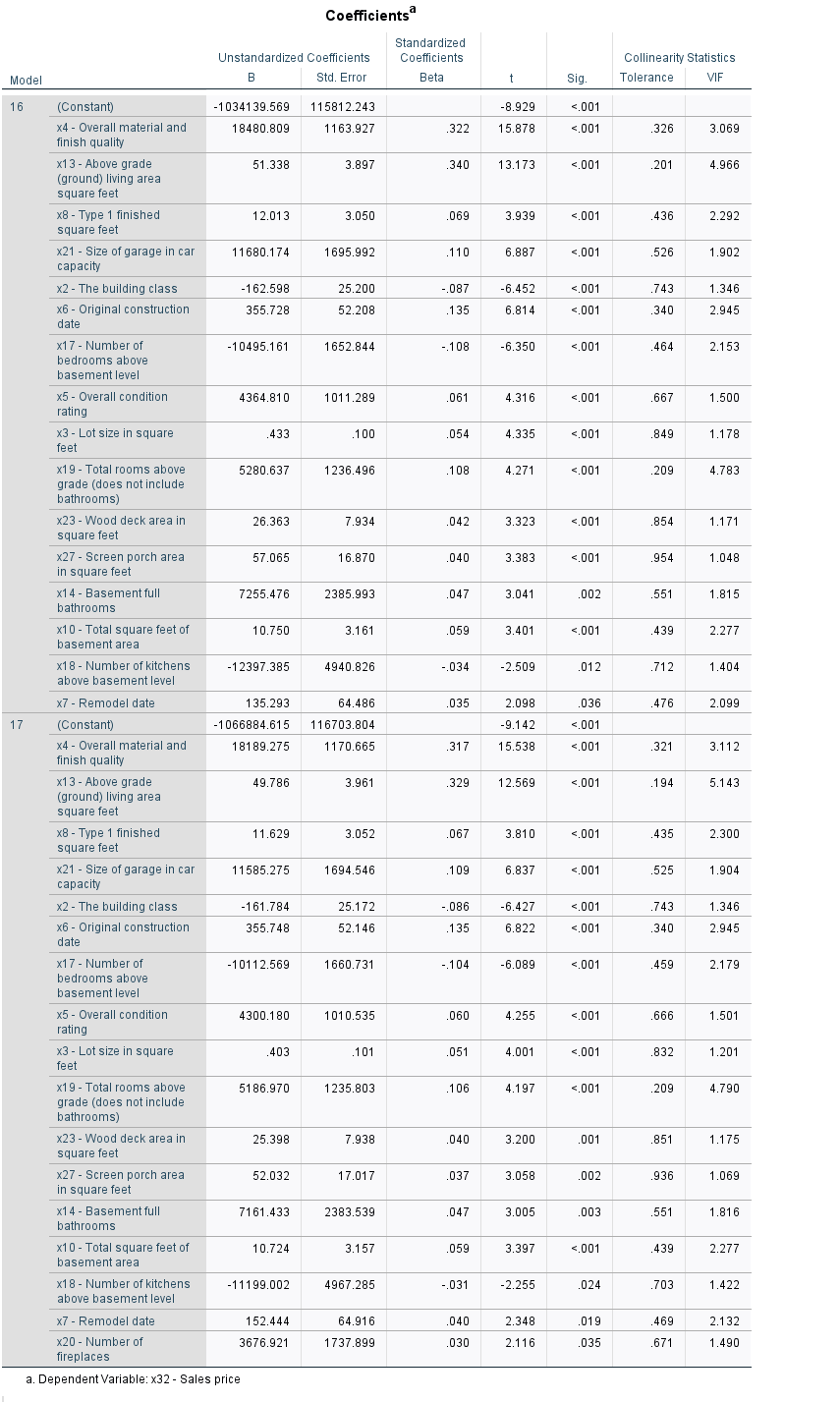
Homoscedasticity is the next presumption to be verified. The data appears to have been fired from a shotgun since **there is no discernible pattern**, and there are equally spaced dots on the X and Y axes above and below zero.

Chart, scatter chart

Description automatically generated

*Fig.5 – Scatter plot of residuals.*

Using VIF values, you may check for multicollinearity.



*Fig.6 – Checking for multicollinearity.*

**Each value is below 10, indicating that the assumption is met.**

**And since your residuals are normally distributed and homoscedastic, linearity is satisfied.**

**Stepwise Regression Analysis** – **Step 1** – From figure 3, Overall material and finish quality, the variable with the highest zero order correlation, has a statistically significant multiple R value of 0.791. As a result, model 1 was produced when the variable was initially included to the regression model.

Table

Description automatically generated

Table

Description automatically generated

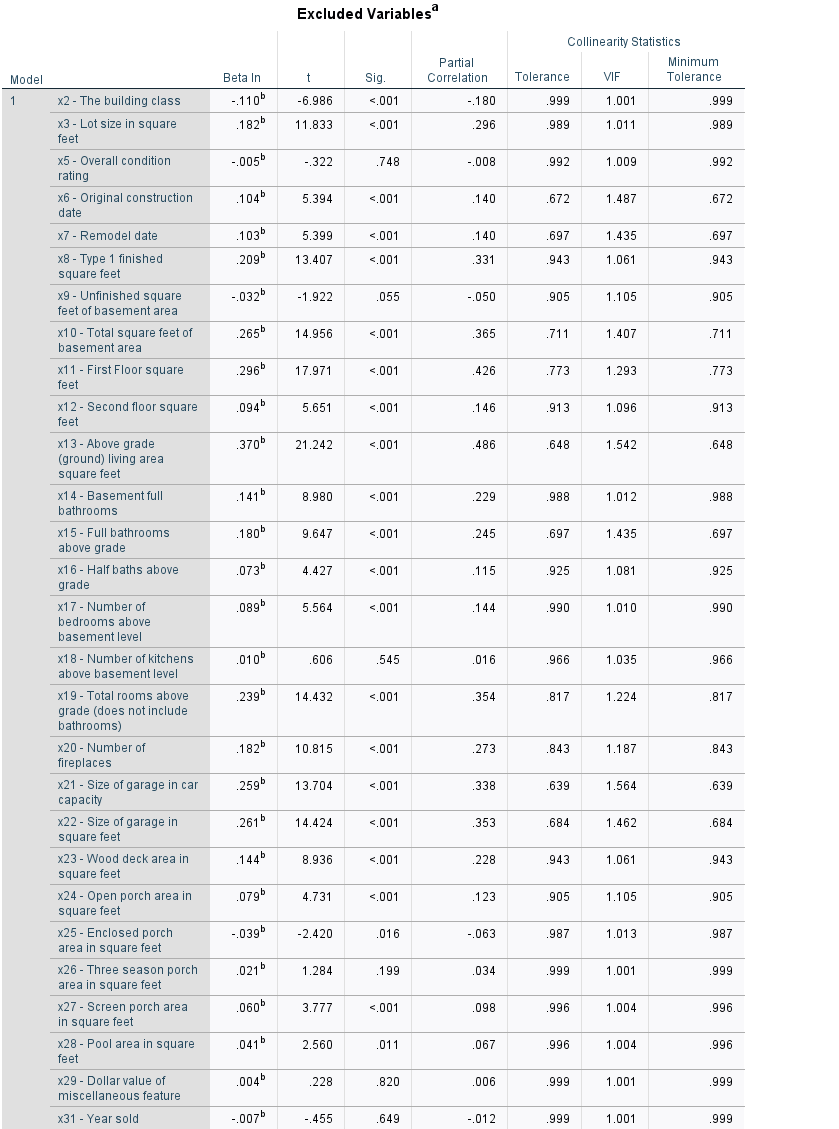
Table

Description automatically generated

*Fig.7 – Model summary, ANOVA table, and variable particulars for variable x4 in step 1.*

The modified r square value indicated that x4 alone could explain around 62.5% of the overall variance in x32 - sales price. With a F ratio of 2436.771, the ANOVA analysis shows that the model fits the data statistically significantly. Additionally given are the regression coefficients b0 and b4, -96206.080 and 45435.803. The coefficients can be interpreted as “how much of y changes to a unit change in x4”.

**Step 2** - The variables not taken into account in the partial correlation values of the equation are taken into consideration after the first stage of stepwise regression is finished. Partial correlation is important because it measures the variation in the model that has not yet been well explained by the earlier input variable. The excluded variables are as shown below.



*Fig.8 – Excluded variables in model 1.*

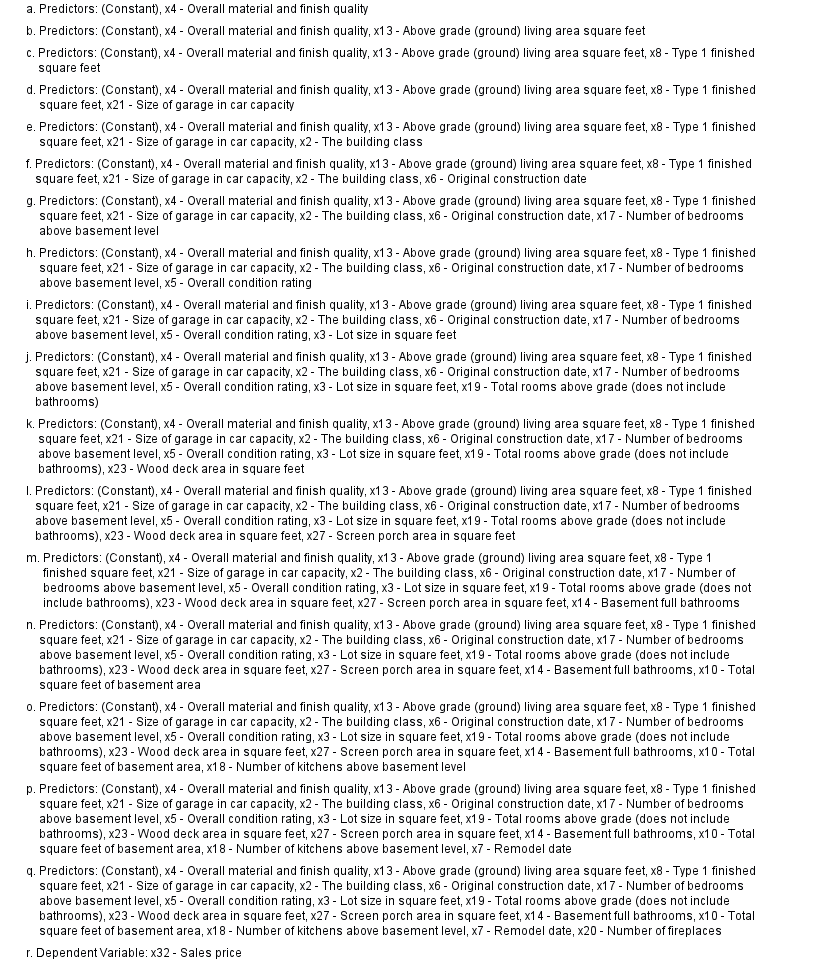
**We can see that x13 – Above grade living area square feet, has the highest partial correlation and was chosen for model 2. Until there are no longer any partial correlations at a significance threshold of 0.05, this stepwise process is repeated.**

**4.3. EXPLANATION AND INTERPRETATION OF MODEL SUMMARY RESULTS**

The model summary of regression analysis as determined by SPSS is as shown.

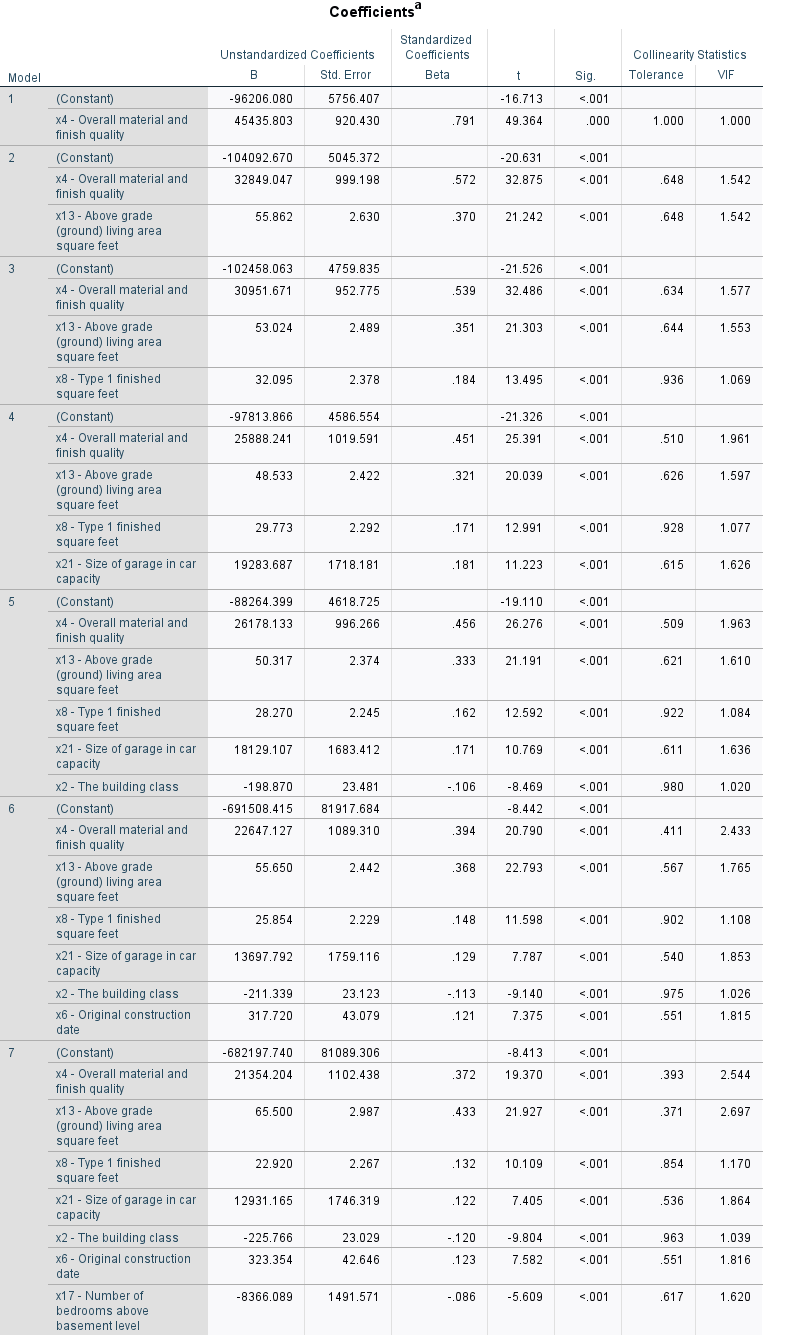
Table

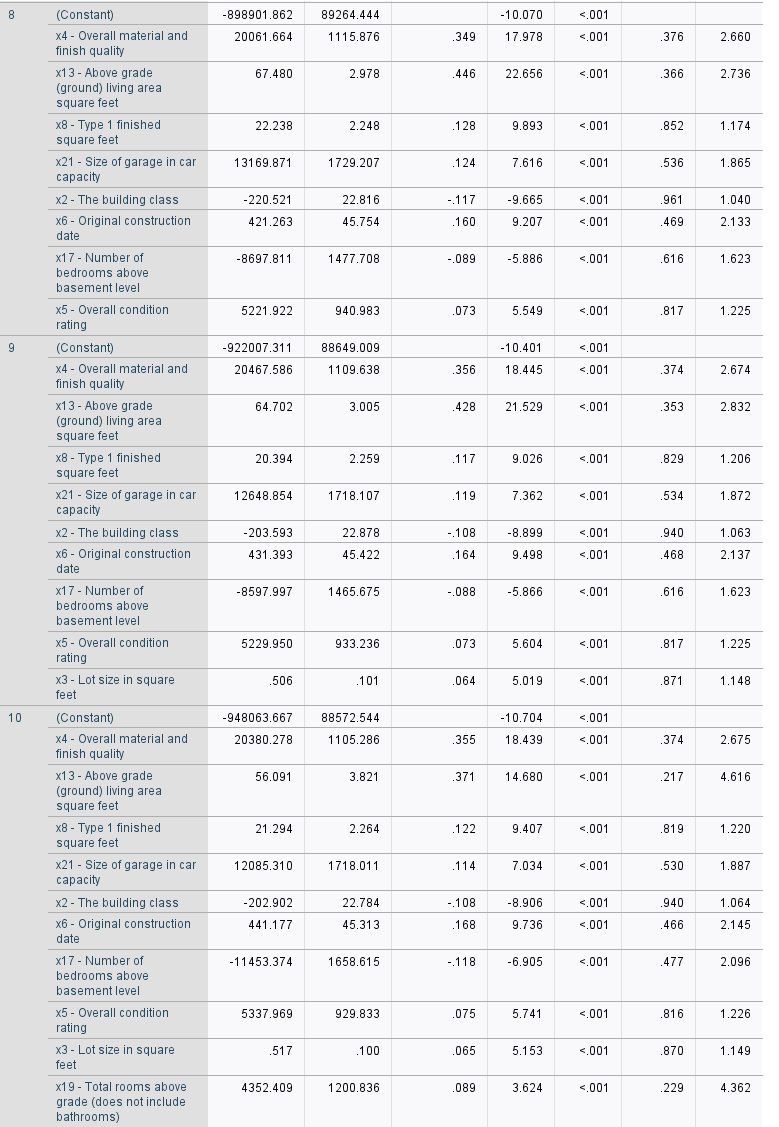
Description automatically generated

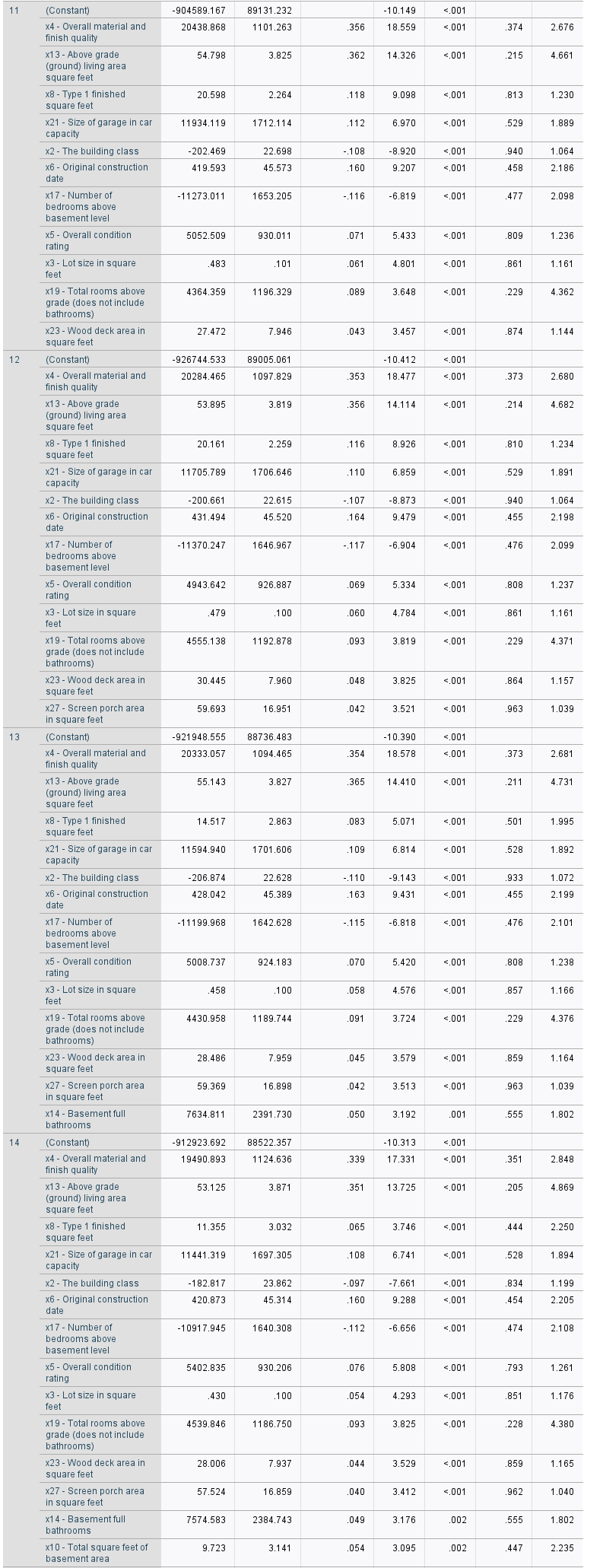


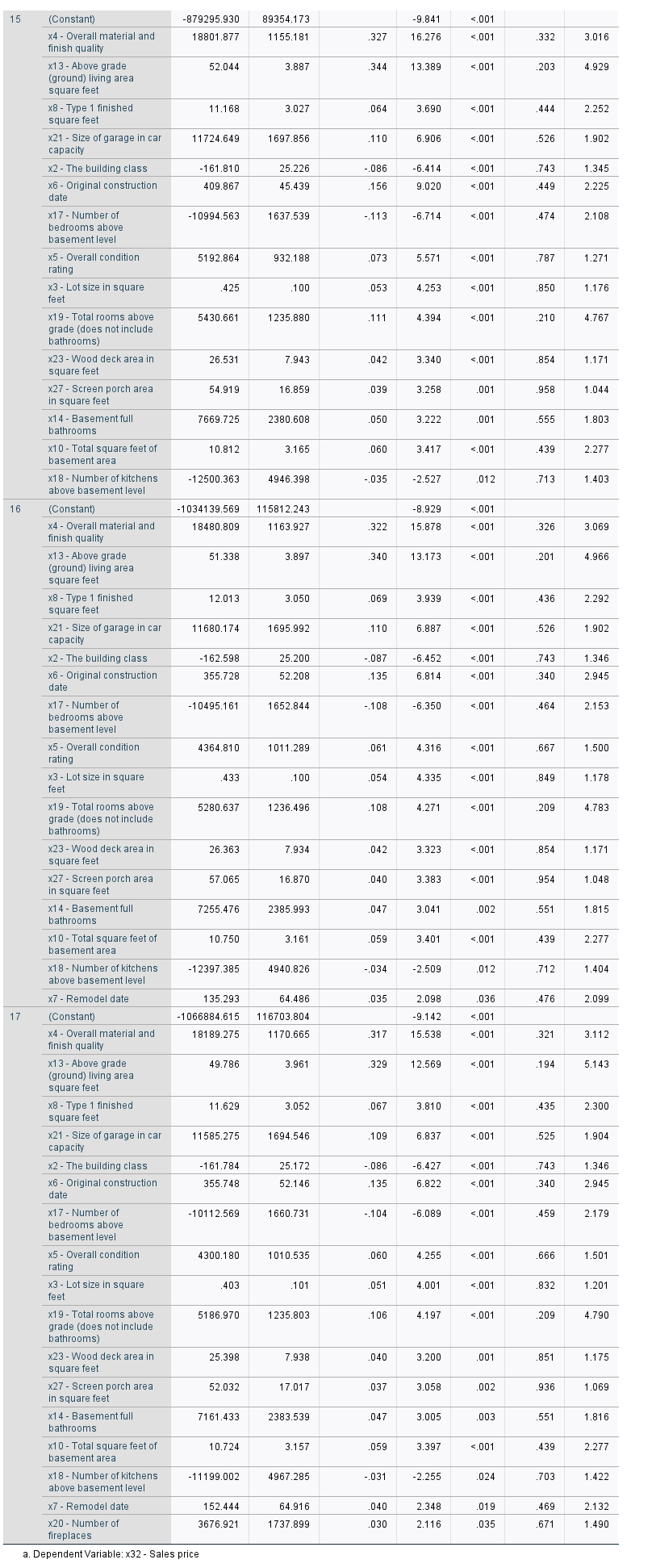
*Fig.9 – Overall model summary of stepwise linear regression.*

According to the model summary, 17 stages were completed before halting, with one variable being added to the overall model at each stage. With each new step, the modified R square values either rise or stay the same. The dependent variables' best linear fit is provided by the final model.









*Fig.10 – Coefficients of variables in the model at each step.*

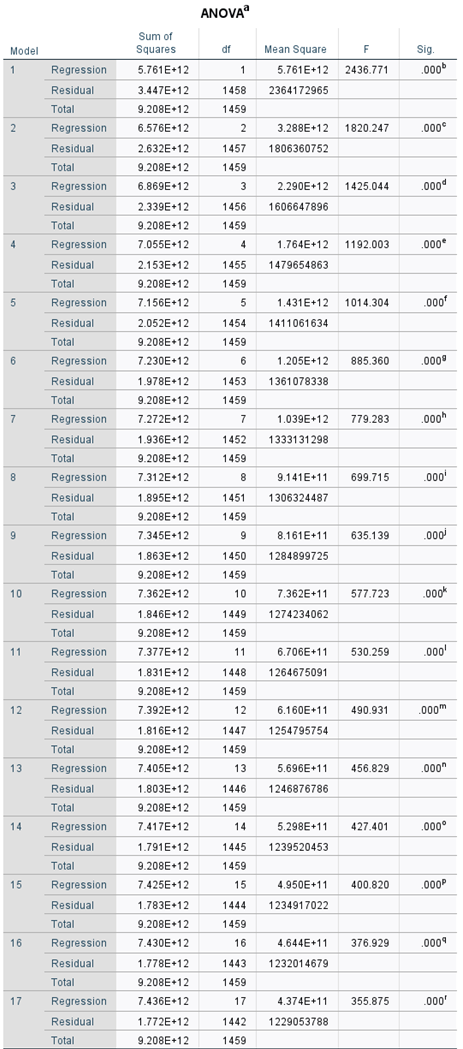
The final model, uses 17 variables that can explain the variation in sales price using the linear equation-

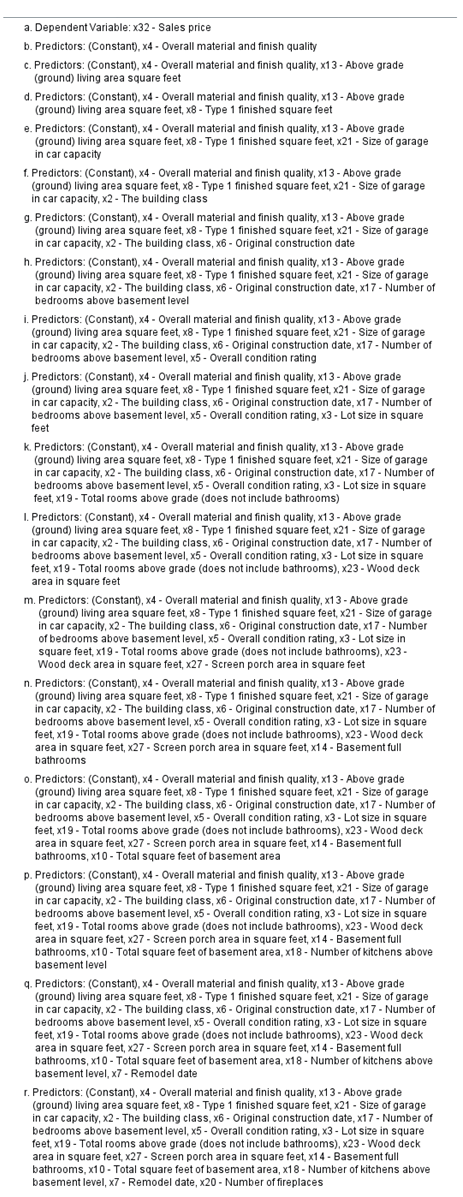
Y = -1066884.615 + 18189.275(x4) + 49.786(x13) + 11.629(x8) + 11585.275(x21) – 161.784(x2) + 355.748(x6) – 10112.569(x17) + 4300.180(x5) + 0.403(x3) + 5186.970(x19) + 25.398(x23) + 52.032(x27) + 7161.433(x14) + 10.724(x10) – 11199.002(x18) + 152.444(x7) + 3676.921(x20)

**4.4. SIGNIFICANCE OF MODEL**

As can be seen from the modified R square value, the final model 17 accounts for exactly 80.5% of the difference in sales price. Additionally, it shows that the model is generic and there is no overfitting. The F ratio in the ANOVA table below is 355.875, and the F ratios up to model 17 were shown to decline, suggesting that including each variable in the model enhances fit and reduces error.

**Therefore, model is significant.**





*Fig.11 – ANOVA analysis for all models.*

**4.5. HYPOTHESIS TESTING FOR THE COEFFICIENTS OF REGRESSION**

Looking at the t values in figure-10 allowed us to perform the hypothesis testing for the coefficients of regression. The unstandardized coefficient is divided by its standard error to determine the T value. As an illustration, in the finished model, t value for x4 is given as 18189.275/1170.665 and is given as 15.538. This is a test against the null hypothesis and is known as a "t test." With 199 residual degrees of freedom and a significance threshold of 0.05, the crucial t value was estimated to be 1.653. Using an internet calculator, this was determined.

Since all of the coefficients in model 17 had absolute t values that are more than the threshold value for a significant level, the null hypothesis was rejected and the alternative hypothesis was accepted.

1. **CONCLUSIONS AND RECOMMENDATIONS**

The goal of this study was to create a stepwise multiple linear regression model on the home pricing dataset to investigate how different factors affect sales price. Three hypotheses were put forth, and the other hypotheses can be adopted based on figure-3 zero order correlation and figure-10 t-value test results. Additionally, it was determined that the final model was capable of statistically explaining 80.5% of the variation in home price.

A number of changes might be made to this study. Data exploration may be done in greater depth. Outliers can be found and considered. Before developing a model, methods like feature scaling or standardization might be used. Additionally, offered hypotheses can be thoroughly investigated.

1. **REFERENCES**

Alfiyatin, A. N., Febrita, R. E., Taufiq, H., & Mahmudy, W. F. (2017). Modeling house price prediction using regression analysis and particle swarm optimization case study: Malang, East Java, Indonesia. International Journal of Advanced Computer Science and Applications, 8(10).

Moran, M. (2021, August 11). Testing Assumptions of Linear Regression in SPSS. Statistics Solutions. <https://www.statisticssolutions.com/testing-assumptions-of-linear-regression-in-spss/>

Statistics Solutions. (2021, August 2). Conduct and Interpret a Factor Analysis. <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis-2/>

**SECTION - B**

1. **PURPOSE OF FACTOR ANALYSIS**

The goal of factor analysis is to consolidate several independent variables into a small number of dimensions. Data may be made simpler via factor analysis, for example, by lowering the number of variables in regression models. The house price dataset used for this study has 30 variables excluding the target variable and variable used for symbolling. Ideally for model building, having too many variables is a disadvantage. Therefore, for easier understanding of the data and to reduce the commonality between variable terms used to build a model, factor analysis is performed on the house price dataset.

1. **EXCLUDE NON-METRIC VARIABLES**

The dataset used for study does not contain any non-metric variables. However, as general rule, factor analysis presupposes normalcy, homoscedasticity, and linearity between variables, hence non-metric variables are discarded. Additionally, correlations must be determined, which is again inappropriate for non-metric variables.

1. **FACTOR ANALYSIS ON SPSS**

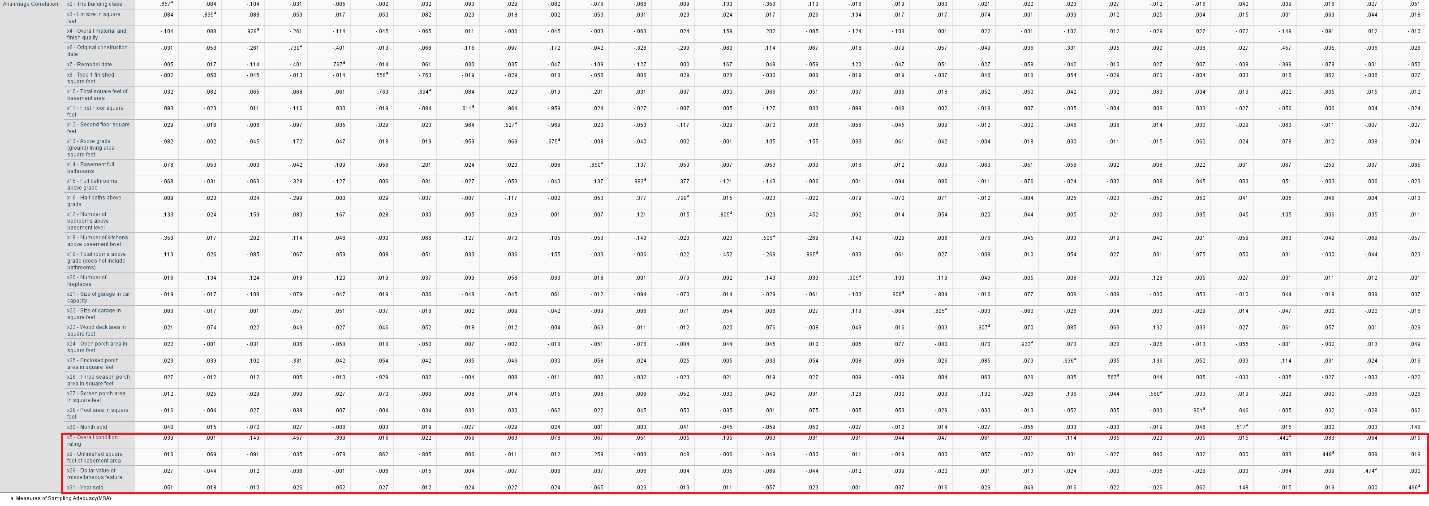
Factor analysis on the 30 independent variables was conducted on SPSS using principal component analysis and varimax rotation, producing output as shown

Table

Description automatically generated

*Fig.12 – Bartlett’s Test.*

Looking at individual MSA values for each variable, we exclude variables that have MSA < 0.5.



*Fig.13 – Anti-image correlation.*

From figure, x5, x9, x29 and x31 are excluded after stepwise removal and rechecking for individual MSA value.

Table

Description automatically generated

*Fig.14 – Eigen values.*

From figure-14, SPSS shows us the suggestion that, with 8 factors, 66.41% of the total variance can be obtained. For simplicity, we can trade off 18 variables that only contribute to 33.59% of total variance. This can be represented in a graphical form using a scree plot as shown below.

Chart, line chart

Description automatically generated

*Fig.15 – Scree plot.*

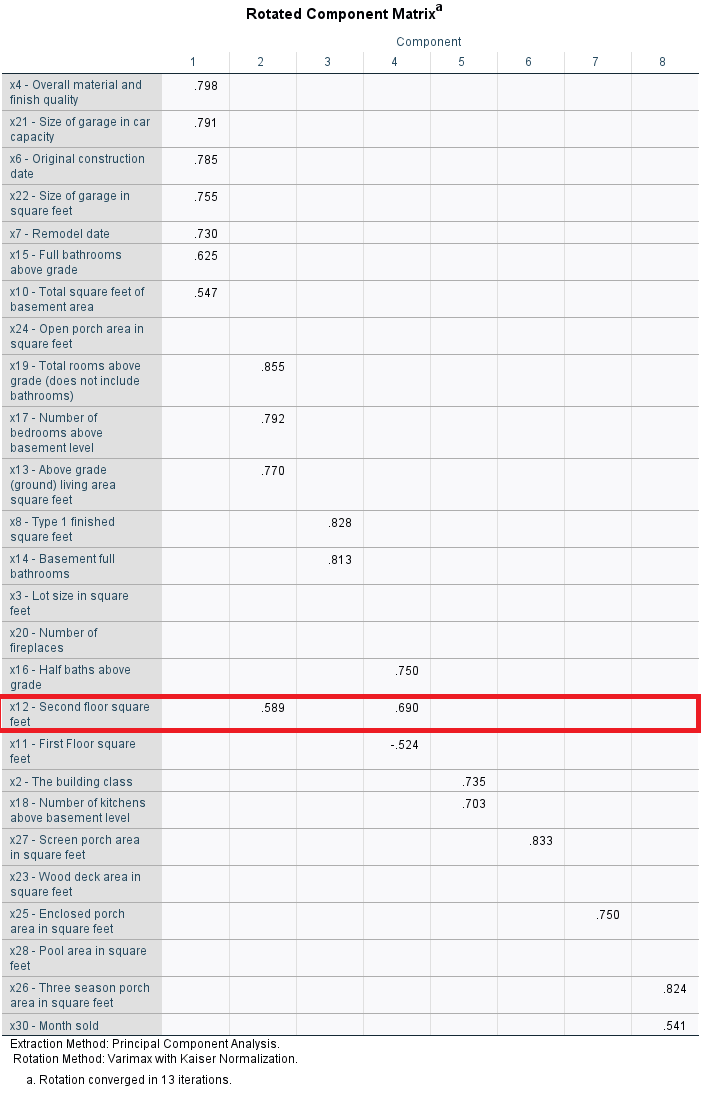
Another important table that **shows the proportion of each variable’s variance that can be explained by the principal components is the communalities table**, as shown below.

Table

Description automatically generated

*Fig.16 – Communalities.*

Looking at rotated component matrix, which is the output for factor analysis, we notice cross loading in variable x12 using **absolute value 0.5**. Therefore, this variable must be excluded. **Since absolute value of 0.5 is used for rotated matrix, some variables in the components do not have their value shown (values are below 0.5).**



*Fig.17 – Rotated component matrix.*

Deletion of this variable reduces the number of components and results in loss of information. Therefore, it is better to ignore the discrepancy.

**This is the output of factor analysis.**

1. **INTERPRETATIONS**

**4.1. COMMUNALITY**

As mentioned previously, proportion of each variable’s variance that can be explained by the principal components is known as communality. Referring back to figure-16, we see that, lowest number of variations by factor analysis was obtained for x24 – Open porch area in square feet (29.1%) and x28 – Pool area in square feet (30.3%). The highest variations were explained for variable x13 – Above grade living area in square feet (91.3%) and x12 – Second floor square feet (90%).

**4.2. EIGEN VALUE**

From figure-14, we see that if all the variables were combined to one component, the eigen value will be 6.280 and in general, higher the eigen value, the better. This component will explain a cumulative of 24.155%.

As the components are increased, we see that, at eight components, the eigen value is closest to one at 1.007 and variation explained is 66.41%. Increasing components after this would be detrimental to simplicity and therefore, factor analysis is stopped at 8 components.

1. **ASSESSMENT**

**5.1. HOW CAN FACTORABILITY BE IMPROVED PRIO TO ANALYSIS?**

Factorability tests such as the Kaiser–Meyer–Olkin (KMO) test can be conducted. A statistical test called the Kaiser-Meyer-Olkin (KMO) test evaluates the suitability of data for factor analysis. The test evaluates whether sampling is enough for both the entire model and individual variable.

**5.2. WHAT IS CROSS LOADING?**

It is difficult to name all the components that share the same variable, making it difficult to make those factors unique and reflect different ideas. **This is known as cross-loading** and occurs when a variable is determined to have more than one significant loading (depending on the sample size).

Example, in figure-17, we see that the variable x12 is cross loaded between components 2 and 4.

**5.3. HOW TO REDUCE CROSS LOADING?**

* Eliminate the variable.
* Use a different rotation technique.
* Decrease the number of factors.