**Contents**

[Research Questions & Hypotheses 2](#_Toc103208902)

[Introduction & Methodology 3](#_Toc103208903)

[Research Design 3](#_Toc103208904)

[Sample 4](#_Toc103208905)

[Buisness Understanding 5](#_Toc103208906)

[Results 5](#_Toc103208907)

[Descriptive Statistics 5](#_Toc103208908)

[Normality 7](#_Toc103208909)

[Correlation 8](#_Toc103208910)

[Multiple Linear Regression Model -1 9](#_Toc103208911)

[Limitations 15](#_Toc103208912)

[Recommendations 15](#_Toc103208913)

# Research Questions & Hypotheses

The main objective of this research analysis is to identify the factors contributing to the change in sales price of the houses. Being hired by the local real estate broker, my job is to analyze the strategies played in local housing market. I would be doing my research analysis on the below two questions and perform results that will help me in providing the best guidance to the buyers and sellers in real estate market.

1. How are the independent variables chosen as a preference to buy a house affect the sales price of the house?
2. How are the sale prices varied with respect to the years on which they were built?

Based on the above research questions, the below hypothesis is formed

**H1 (Hypothesis 1):** Lot Area, TotalBsmt SF, Gr Liv Area, Age of the house, Garage area, TotRmsAbvGrd determine the price of the house.

**H2 (Hypothesis 2):** Sales prices compared with respect to two years 2005 and 2006.

Sales prices compared with respect to the three years 2004,2005 and 2006 to observe the significant difference.

# Introduction & Methodology

## Research Design

The methodology of this research study is based upon a quantitative research design based on a critical realist approach. Since the research study considers a sample in real-time, therefore, a realist approach has been opted for, explaining the determinants of Ph.D. completion among candidates. Moreover, based on the approach, the study can be called deductive since it deduces generalizations based on results obtained through a proper detailed analysis provided in the report ahead. On the other hand, as mentioned before the research strategy that has been opted within this study is quantitative, using a more descriptive approach so that the determinants or predictors of Ph.D. completion among candidates can be discussed in detail. Instead of a longitudinal research study, a cross-sectional study has been considered to avoid any discrepancies such as differences in grading plans, etc.

For the data collection process, among the sample chosen, a survey was distributed which was filled by the students gradually. Therefore, a structured questionnaire was used for the survey. The survey was based on 18 variables including gender, age, GPA, GRE scores, letter of recommendations, motivation, stability, financial funding, marital status, age, social skills, hostility, and impression. Among all these variables Ph.D. completion is the dependent variable whereas GPA, GRE Scores, Motivation, Stability, Financial Resources, Hostility, Impression, Letter of Recommendation, and Social Abilities have been identified to be independent variables. All these variables have been measured on a 9-point Hedonic scale ranging from extremely low to extremely high.

## Sample

The housing data file contains 2360 observations on the attributes for buying and selling the house. Various attributes include categorical variables such as neighborhood, building type, Central air conditioning and housing style and numerical variables such as lot area, years built, years sold, total basement square feet, Garage area,Ground living area, full bath, half bath, bedrooms above ground etc.

**Data Cleaning**

Dataset has been cleaned by removing the blank values in the fields. There are some blank fields in the field Half bath which have been replaced with 0 instead of NULL to make the value numeric for building the models. My analysis required the age of the house field for building the model. The calculation has been made for adding a new variable age of the house which is year the house was sold subtracted with the year on which the house was built. Also, for the Air conditioning field, I converted the variable to numeric variable Y to 1 and N to 0 for building the regression model.

## Buisness Understanding

People often select various attributes particularly while buying a house of their own. Lot area, Age of the house, total basement square feet, Garage area, Ground living area, bedrooms above ground are identified as independent variables and sales price is considered as dependent variable on the mentioned independent variables. For Data Analysis, the multiple linear regression model has been utilized using the EXCEL software to avoid any kinds of human error within the analysis of the results.

.

# Results

## Descriptive Statistics

Around 100 Ph.D. candidates were engaged within this research study to provide an effective representation of the small sample considered. Among the entire sample, there were around 64% Females i.e., 64 and 36% Males i.e., 36 out of a sample of 100 as displayed in Table 1. Table 1. Gender

**Gender**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Frequency | Percent | Valid Percent | Cumulative  Percent |
|  | Female | 64 | 64.0 | 64.0 | 64.0 |
| Valid | Male | 36 | 36.0 | 36.0 | 100.0 |
|  | Total | 100 | 100.0 | 100.0 |  |

Similarly, most of the candidates were single i.e., 60% whereas the rest were married accounting to be only 40% of the entire sample as displayed in Table 2.

Table 2. Marital Status

**Marital Status**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Frequency | Percent | Valid Percent | Cumulative  Percent |
|  | Married | 40 | 40.0 | 40.0 | 40.0 |
| Valid | Single | 60 | 60.0 | 60.0 | 100.0 |
|  | Total | 100 | 100.0 | 100.0 |  |

As a Ph.D. is a higher degree, the age bracket starts from 20 years of age. Around 76% of the sample was between the ages of 20 to 30, followed by 20% within 31-40 and only 4 were either 41 or above. Table 3 displays the categorical distribution of the sample by age. Table 3. Age

**Age at Entry**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative  Percent |
| 20-30 | | 76 | 76.0 | 76.0 | 76.0 |
|  | 31-40 | 20 | 20.0 | 20.0 | 96.0 |
| Valid | 41 and Above | 4 | 4.0 | 4.0 | 100.0 |
|  | Total | 100 | 100.0 | 100.0 |  |

Table 4. Ph.D. completion status

**Ph.D. completion**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Frequency | Percent | Valid Percent | Cumulative  Percent |
|  | Completed | 50 | 50.0 | 50.0 | 50.0 |
| Valid | Incomplete | 50 | 50.0 | 50.0 | 100.0 |
|  | Total | 100 | 100.0 | 100.0 |  |

Coming towards the main sample in consideration, the entire sample is divided into Ph.D. candidates who have completed their degree and who have not. Therefore, an equal sample is considered so that an equal representation can be given to each, based on which, 50% of the sample has completed their degree and 50% of the sample has not as displayed in Table 5.

## Normality

Table 5. Descriptive Statistics

**Descriptive Statistics**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | N | Minimum | Maximu m | Mean | Std. Deviation | Skewness | | Kurtosis | |
| Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std.  Error | Statistic | Std.  Error |
| 1 Letter of  Recommendation | 100 | 4 | 9 | 6.94 | 1.324 | -.101 | .241 | -.844 | .478 |
| 2 Letter of  Recommendation | 100 | 4 | 9 | 7.00 | 1.303 | -.168 | .241 | -.768 | .478 |
| 3 Letter of  Recommendation | 100 | 4 | 9 | 7.06 | 1.324 | -.219 | .241 | -.783 | .478 |
| Student Motivation | 100 | 6 | 9 | 7.82 | .957 | -.334 | .241 | -.848 | .478 |
| Emotional Stability | 100 | 4 | 9 | 6.38 | 1.668 | .121 | .241 | -1.211 | .478 |
| Financial Resources | 100 | 3 | 9 | 5.78 | 1.685 | .044 | .241 | -.754 | .478 |
| Interpersonal Skills | 100 | 4 | 9 | 6.58 | 1.365 | -.123 | .241 | -.639 | .478 |
| Hostility | 100 | 1 | 5 | 2.60 | 1.044 | .217 | .241 | -.422 | .478 |
| Selectors Impression of Applicant | 100 | 5 | 9 | 7.08 | 1.203 | -.227 | .241 | -.970 | .478 |
| College GPA | 100 | 2.75 | 3.97 | 3.5130 | .26591 | -.567 | .241 | .278 | .478 |
| Major GPA | 100 | 3.20 | 4.00 | 3.7778 | .19812 | -.811 | .241 | .222 | .478 |
| GRE Specialty | 100 | 520 | 790 | 652.20 | 70.504 | -.041 | .241 | -.889 | .478 |
| GRE Quantitative | 100 | 550 | 787 | 688.49 | 63.906 | -.343 | .241 | -.899 | .478 |
| GRE Verbal | 100 | 470 | 780 | 631.80 | 71.596 | -.051 | .241 | -.652 | .478 |
| Valid N (listwise) | 100 |  |  |  |  |  |  |  |  |

Since the Skewness and Kurtosis of the data are between +/-1, the data is considered normal as displayed above in Table 5.

## Correlation

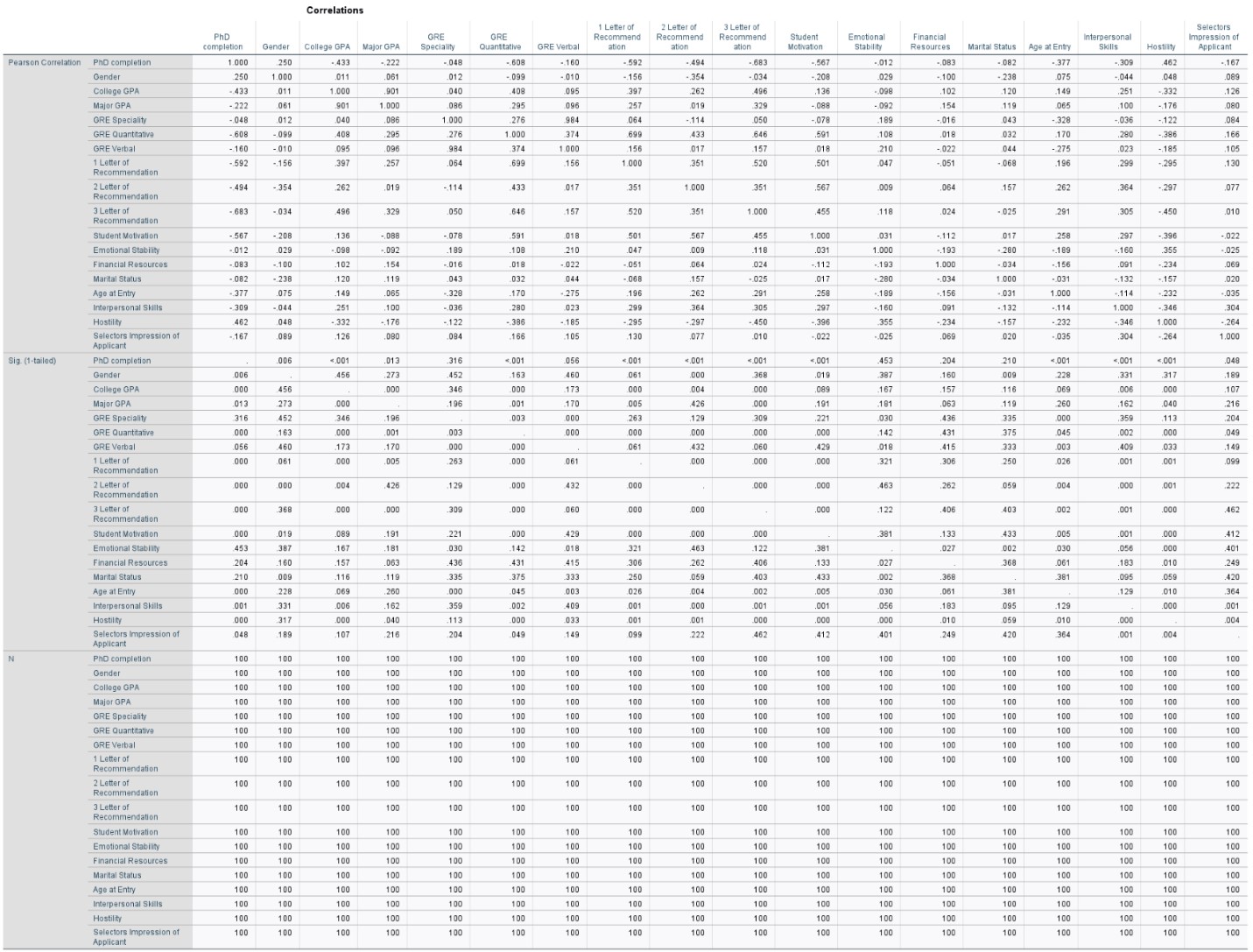
Table 6 displays that Ph.D. completion significantly correlates with the GRE score of the quantitative portion i.e., -0.608, along with Letter of recommendation-1, Letter of Recommendation-3, Student motivation with a correlation coefficient of -0.592, –0.683, and 0.567 respectively. All these variables are highly correlated with Ph.D. and have a strong negative relationship. College GPA, Letter of Recommendation-2, and Age at entry are moderately correlated and have a moderate negative relationship with Ph.D. with correlation coefficient –0.433, –0.494, and –0.377. From these correlation values, we can assume that these factors determine the Ph.D. completion of a candidate, however, there is no significant relation reported in results with Impression of the Selector.

From the correlation table, we can also observe that the independent variables are highly correlated with each other and have a strong positive relationship between them. This high correlation between the independent variables leads to a multicollinearity issue. The following are the highly correlated independent variables. College GPA and Major GPA with a correlation coefficient of 0.901. GRE quantitative correlates with Letter of Recommendation-1 and Letter of Recommendation- 3 whose correlation coefficients are 0.699 and 0.646. GRE specialty and GRE

Verbal with correlation 0.984, GRE quantitative, and student motivation with correlation 0.591. Letter of Recommendation-1, Letter of Recommendation- 2 with student motivation with correlation coefficients 0.501 and 0.567respectively. Letter of Recommendation-1 and Letter of

Recommendation- 3 with correlation 0.520.

Table 6. Correlation



### Multiple Linear Regression Model -1

I have performed linear regression using Lot Area, Total Bsmt SF, Gr Liv Area, Age of the house, Garage Area, TotRMS Abvgrd as my independent variables to predict SalePrice and from the regression analysis output I can say that the model is statistically significant. since the p values of all variables are less than 0.05, we reject the null hypothesis.

The R-Square values and Adj R-square values are 0.7345 and 0.7339 respectively. From the R-square we can say that 73.45% of the variance in the Y (salePrice) is explained by my independent variables x.

From the regression output, I can say that

For 1 unit change in the Age of the house, SalePrice is negatively impacted by -707.642 holding all other variables constant.

For 1 unit change in Garage Area, Saleprice is positively impacted by 64.156 holding all other variables constant.

For unit change in Lot Area, Saleprice increases by 0.380 holding all other variables constant.

For one unit change in Gr Living Area, salesprice increases by 79.487 keeping all other independent variables constant.

For 1 unit change in Totrms Abbvgrd, saleprice decreases by 3519.477 keeping all other variables constant.

The formula for predicting the price of the house is

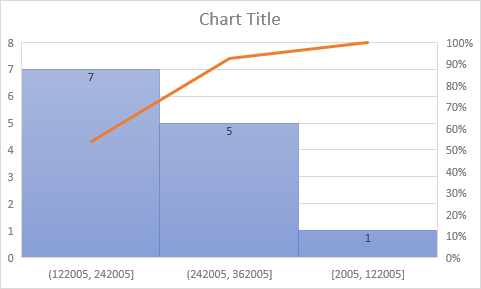
Intercept+ Lot Area\*2500+Total Bsmt SF\*1300+Gr Liv Area\*1100+ Age of the house\*5+Garage area\*300+TotRms AbvGrd\*6=

=M22+ M23\*2500+M24\*1300+M25\*1100+M26\*5+M27\*300+M28\*6

Which is 169047.3502

ANOVA

ANOVA was made by comparing the sale prices with respect to the field years on they were built.

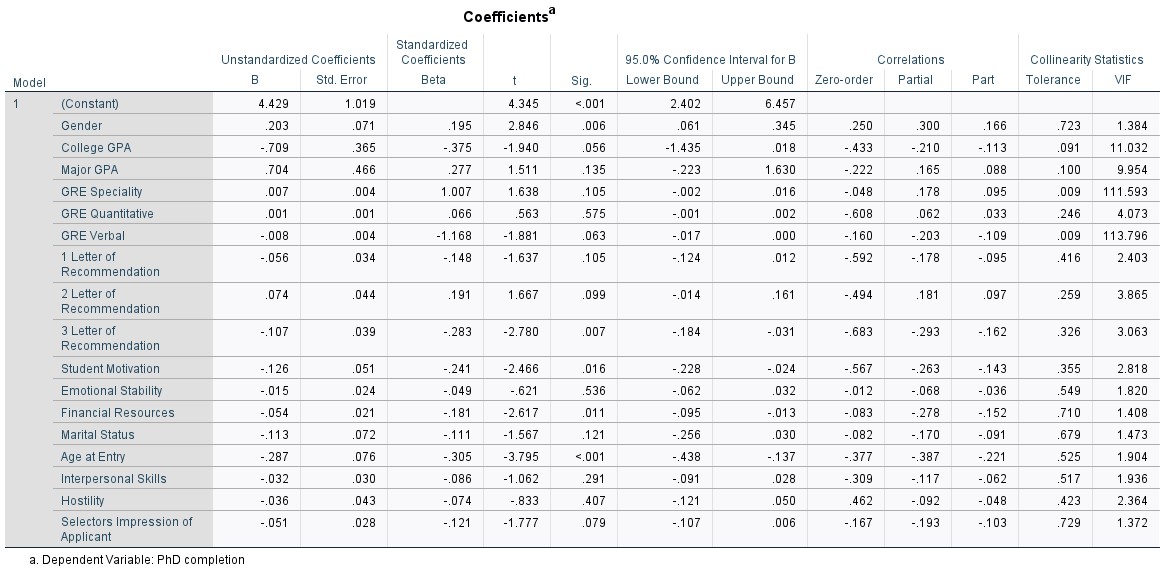


First, I compared sale prices of both the years 2005 and 2006 for houses,

The below analysis tells us that the F<critical and P value> alpha which is 0.05. This indicates that the both the values are similar and we accept the null hypothesis.

For the second analysis, when the sale prices are compared for three years 2005,2006 and 2008.

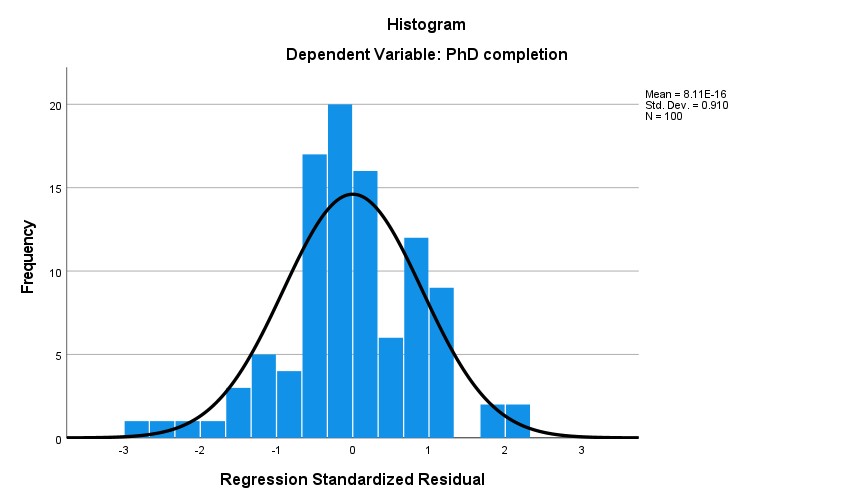
We can observe that F> critical and P value<0.05 which means that the sales prices are not similar and we can reject the null hypothesis.

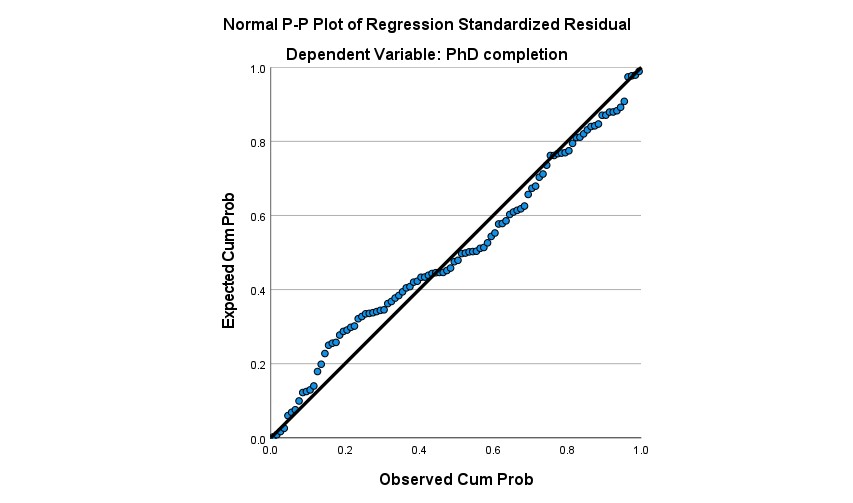


As we have 100 observations, so the degrees of freedom df= number of observations - 1 = 99.

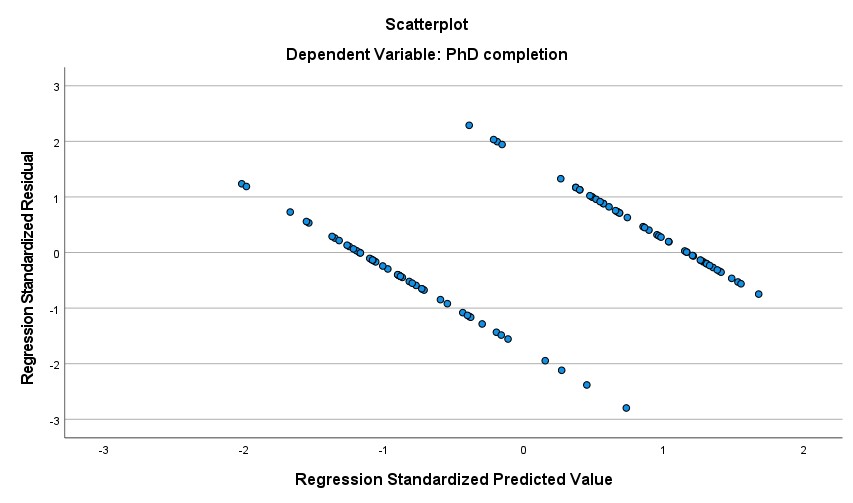
With 95% of confidence interval and df=99, p-value =0.05, so the critical t-value is ±1.984. Based on corresponding t-values and p-values we can tell which independent variables are good for our model. This can be done by comparing the critical t-value against the calculated t-value, the calculated t-value should not fall in the interval of critical t-value, also checking the significance level of the independent variables whose p-value<0.05. This helps in interpreting which independent variables to be considered as a good fit for our model.

Upon comparison based on the above discussion, we can understand that the independent variables College GPA, Major GPA, GRE specialty, GRE verbal, GRE quantitative, 1 Letter of recommendation, 2 Letter of recommendation, emotional stability, Marital status, interpersonal skills, Hostility, selectors impression of applicants have p-value more than 0.005 and t-calculated values fall in the interval of ±1.98 which is the t-critical value. Hence, indicating that these variables are not statistically significant and may overfit our model. So, there is a need to drop these variables and rerun our regression model without considering those variables which violates our assumption of statistical significance.





From the histogram and the normal P-P plot, we can infer that the residuals(errors) are not much normally distributed as the residuals do not fall along with the linear line in the normal P-P plot, and the distribution of residuals in the histogram also does not seem normally distributed violating the normally distributed errors assumption.



Homoscedasticity, independent errors, Linearity, normally distributed errors Assumptions: A regression plane shown above tells us about the linearity of the multiple regression with the standardized predictor and residual variables on the x-axis and y-axis respectively which are cantered around Zero (0). We observe a pattern of dots rather than being randomly scattered, showing that the successive residuals are correlated, and the errors are not normally distributed proving that linearity assumption, independent errors, and normally distributed errors assumption are violated. Because of the dots not being normally distributed, it may indicate that the variances of the residuals are not constant which violates the Homoscedasticity assumption.

# Limitations

Only few independent variables have been chosen for the regression analysis. The output results of the data would not fairly represent the sales price prediction but only with respect to the selected attributes. Also, the variance in sales prices is only identified with the selected years would not actually help to predict the sales price of the house in general. These results only indicate the prices for the short sample but not for the large samples.

# Recommendations

The data set is highly flexible with various categorical and numerical variables where we can find a greater number of ways to determine the sales prices for the houses. Data can be grouped into different sets to determine the sales price. For instance, data can be divided into samples based on the neighborhood and building styles and types. This would help in extracting the output results in specific.