HOUSE PRICE PREDICTION MODEL:

TABLE OF CONTENTS

INTRODUCTION

This project aims to find how different features of houses affect sales prices. There are two datasets, one containing house features (housing_data) and the other containing sales details (price_sale). By analysing different features, we can find the most influential features on house prices. This analysis will help to identify key factors for house pricing and predicting house prices. It is valuable for the real estate industry and potential house buyers.

Data

Two data sets are combined based on the ID_House in the SAS. There are 200 rows, eight columns for housing and 210 rows, and five columns for pricing. The final dataset has 200 rows and 12 columns (ID needs to be deleted). The target variable price is numerical.

housing_data.xlsx

- **ID_House:** a unique identifier which corresponds to the house identification number
- **Living_area:** the floor area of the living room in square meters
- **Garage_area:** the floor area of the garage in square meters
- **Garage_type:** Garage location (NA means information 'not available')
- **Count_Bedroom:** a count of the number of bedrooms in the house
- **Count_Bathroom:** a count of the number of bathrooms in the house
- **Central_Air:** a binary variable that is Yes (Y) if the house has central air conditioning and is No (N) if it does not
- **Fireplaces:** Number of fireplaces

price_sale.xlsx

- **ID_House:** a unique identifier which corresponds to the house identification number in the study.
- **Overall_Qual:** Rating of the overall quality of the house (on a scale of 1 to 8)

- **Year_construction**: Year of the construction of the house
- **Year_Sold:** Year the house was sold
- **Price_of_Sale:** Sale price in £.

Exploratory Data Analysis (EDA)

Categorical

Garage_Type				
Garage_Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	42	21.00	42	21.00
Atached	2	1.00	44	22.00
Attached	81	40.50	125	62.50
Detached	52	26.00	177	88.50
NA	23	11.50	200	100.00

Count_Bathroom				
Count_Bathroom	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	92	46.00	92	46.00
2	91	45.50	183	91.50
3	16	8.00	199	99.50
4	1	0.50	200	100.00

Count_Bedroom				
Count_Bedroom	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	17	8.50	17	8.50
2	68	34.00	85	42.50
3	110	55.00	195	97.50
4	5	2.50	200	100.00

Central_Air				
Central_Air	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	33	16.50	33	16.50
Y	154	77.00	187	93.50
n	4	2.00	191	95.50
x	1	0.50	192	96.00
y	8	4.00	200	100.00

Overall_Qual				
Overall_Qual	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1	0.50	1	0.50
2	4	2.00	5	2.50
3	10	5.00	15	7.50
4	19	9.50	34	17.00
5	71	35.50	105	52.50
6	66	33.00	171	85.50
7	20	10.00	191	95.50
8	9	4.50	200	100.00

Fireplaces				
Fireplaces	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	136	68.00	136	68.00
1	64	32.00	200	100.00

Central_Air: A categorical variable needs to be encoded for modelling. Also, there are some data issues

- Values y and n should transfer to Y and N
- X is an invalid value that must be changed for data consistency.

Garage_type: Need to be encoded for modelling.

- Atached is typo error corrected to Attached.
- Blank shows missing values
- The NA category is used for no garage (does not have a garage).

Fireplaces: This variable is described as a number of fireplaces, while there are two categories which could be considered as yes or no (1 for 'yes', 0 for 'no'), where frequency is more meaningful.

These are ordinal categorical variables; in the project, they will be treated as numerical, but frequency has more value here.

Count_Bathroom: Just one observation has four bathrooms.

Count_Bedroom:55% of houses have three bedrooms.

Overall_Qual: qualities 5 and 6 have a higher frequency. This variable could be influenced by central air or other variables since it shows the overall quality of the houses; therefore, it needs to be checked.

Numerical (statistical analysis)

Descriptive statistics is generated Price_of_Sale, Living_Area, Garage_Area, Year_Construction, and Year_Sold, using PROC MEANS. This shows important characteristics and potential data issues.

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Std Dev	Minimum	Maximum
Price_of_Sale	Price_of_Sale	199	1	107345.57	29568.99	31440.00	176000.00
Living_Area	Living_Area	200	0	109.0600000	90.0023249	0	1333.00
Garage_Area	Garage_Area	200	0	35.3300000	30.3185947	0	388.0000000
Year_Construction	Year_Construction	200	0	1964.18	27.2072702	1877.00	2009.00
Year_Sold	Year_Sold	198	2	2011.95	1.2814470	2010.00	2014.00

Living area: minimum of 0

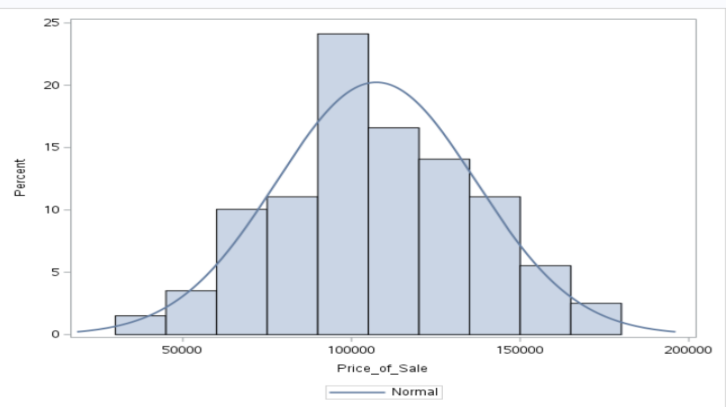
Garage_area: Averages 35 and a minimum of 0.

Year_Construction: Houses range from 1877 to 2009.

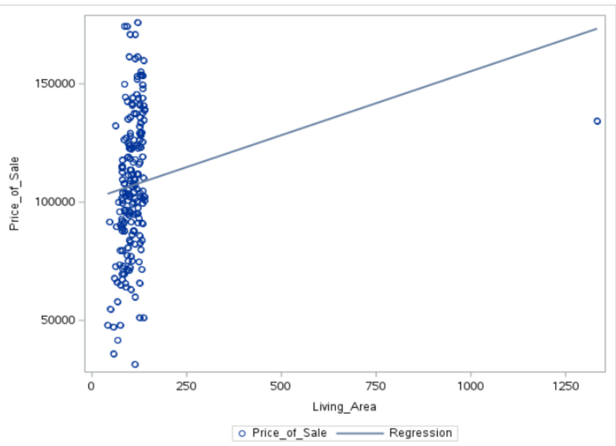
Year_Sold: Two missing values. Sold between 2010 and 2014.

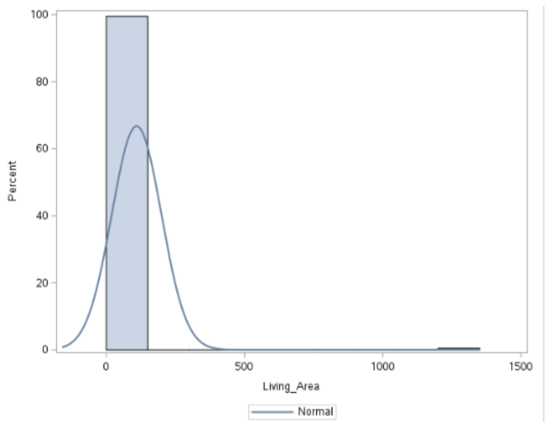
Price_of_Sale: 1 missing value. Averages around 107,346, ranging from 31,440 to 176,000, with a standard deviation 29,569.

Initial Visualization



This histogram shows a roughly normal distribution for Price_of_Sale. It could be suitable for modelling since it is not significantly skewed (probably affecting residual normality). 100,000 - 125,000 is the peak of the distribution.





The scatter plot shows that points are dense in lower living areas, and few points extend to higher regions. It also indicates a slightly linear relationship between target value and living area. For very small living areas, an increase in size might lead to a sharp rise in price. Living_Area is highly positively skewed in the histogram.

This variable may need transformation because of the intense skewness.

DATA CLEANING & TRANSFORMATION:

Id_house has been deleted**.**

Garage_type: Attached is corrected to Attached. Also, the missing area has another category as missing since they have a garage, but they do not know what type, and there are a high number, so they cannot be deleted.

Central_Air: y is transferred to Y, and n is transferred to N. X was assigned the value N, while the frequency of Y is higher, based on the Principle of Conservatism, has decided to assign value N to avoid biasing the proportion of Y**.**

Dummy variables: For nominal categorical variables, central area and garage type dummy encoding has been applied (creating 0 and 1 binary variables to avoid multicollinearity), providing meaningful coefficients and used in the model.

For garage_type, no_garage was selected as the reference category, which allows us to show how specific types of garages influence price relative to a non-existent garage.

For the overall_qual variable, we can consider a dummy variable since it allows for a greater effect for each quality (1 to 8). This variable is identified as an influential predictor in the model section, and this approach could have a good effect. However, this approach was not used due to the increased number of variables and complexity.

Ziro numerical values: Living area is impossible, and there is one zero value in the living area, which is turned to “.”.

Missing values

Year_Sold: Two missing values. Inputting with the mean of 2012.

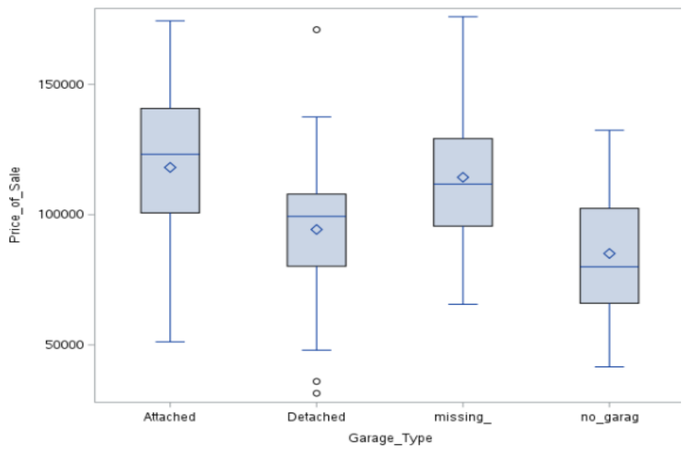
Price_of_Sale: Mean imputation for that single Price_of_Sale missing value 107345.57.

Living_Area: with transforming 0 to missing, it is inputting with 103.00.

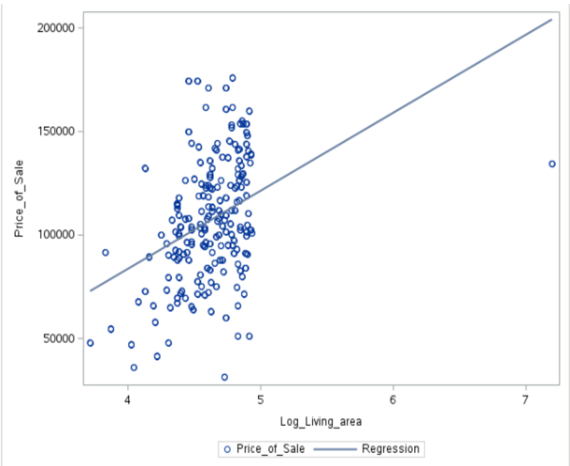
New variable: house age came from two variables (Year_Sold - Year_Construction). This reduces the number of variables for modelling. It is also meaningful and has a good correlation with the target variable.

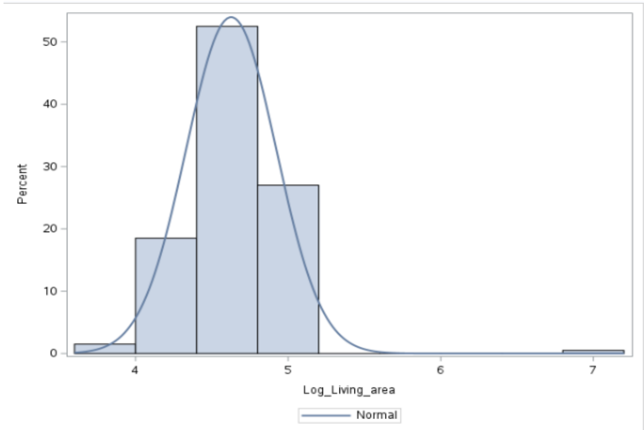
Transform: Transforming the living_area variable to a logarithm probably makes the distribution more symmetric and normal to gain a better linear model prediction. The main reason for this transformation was observing negative parameters estimated in the linear model, which was unrealistic (this problem may be because of other issues, but transforming this variable solved the problem). In addition, with the Pearson matrix, it has been seen that the transform variable of the living area has a stronger correlation with the target variable.

Visualisation after data cleaning



Attached garages tend to have the highest median sale prices**,** followed by Detached garages, and no_garage garages have the lowest median prices. It shows a strong relationship between the target variable and the garage type.





This shows a roughly normal distribution after log transformation. Normally distributed observations do not directly affect the assumption of the model. This was used since it gained better correlations with the target variable and better performance in the model, although this variable has been eliminated by the backwards technique.

Pearson Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

	Price_of_Sale	Living_Area	Log_Living_area	Garage_Area	House_Age	Count_Bedroom	Count_Bathroom	Fireplaces	Overall_Qual	CentralAir_Dummy	GarageType_Attached	GarageType_Detached
Price_of_Sale Price_of_Sale	1.00000	0.16382 0.0205	0.37777 <.0001	0.24652 0.0004	-0.61751 <.0001	0.20578 0.0035	0.31946 <.0001	0.12937 0.0679	0.71119 <.0001	0.39834 <.0001	0.30883 <.0001	-0.26245 0.0002
Living_Area Living_Area	0.16382 0.0205	1.00000	0.78893 <.0001	0.84850 <.0001	-0.12217 0.0848	0.27891 <.0001	0.33452 <.0001	0.17415 0.0137	0.17062 0.0157	0.11241 0.1130	0.17895 0.0112	-0.08537 0.2294
Log_Living_area	0.37777 <.0001	0.78893 <.0001	1.00000	0.66510 <.0001	-0.17493 0.0132	0.52121 <.0001	0.41467 <.0001	0.29396 <.0001	0.33594 <.0001	0.30942 <.0001	0.35470 <.0001	-0.18008 0.0107
Garage_Area Garage_Area	0.24652 0.0004	0.84850 <.0001	0.66510 <.0001	1.00000	-0.23214 0.0009	0.18936 0.0072	0.39712 <.0001	0.13251 0.0614	0.16135 0.0225	0.24635 0.0004	0.23307 0.0009	0.00785 0.9121
House_Age	-0.61751 <.0001	-0.12217 0.0848	-0.17493 0.0132	-0.23214 0.0009	1.00000	0.00812 0.9091	-0.25796 0.0002	0.04563 0.5211	-0.40540 <.0001	-0.28802 <.0001	-0.23897 0.0007	0.21855 0.0019
Count_Bedroom Count_Bedroom	0.20578 0.0035	0.27891 <.0001	0.52121 <.0001	0.18936 0.0072	0.00812 0.9091	1.00000	0.12460 0.0788	0.11010 0.1207	0.10488 0.1394	0.27094 0.0001	0.19625 0.0054	-0.06287 0.3765
Count_Bathroom Count_Bathroom	0.31946 <.0001	0.33452 <.0001	0.41467 <.0001	0.39712 <.0001	-0.25796 0.0002	0.12460 0.0788	1.00000	0.07712 0.2777	0.26567 0.0001	0.19477 0.0057	0.22949 0.0011	-0.10094 0.1550
Fireplaces Fireplaces	0.12937 0.0679	0.17415 0.0137	0.29396 <.0001	0.13251 0.0614	0.04563 0.5211	0.11010 0.1207	0.07712 0.2777	1.00000	0.12999 0.0666	0.08634 0.2241	0.27062 0.0001	-0.18669 0.0081
Overall_Qual Overall_Qual	0.71119 <.0001	0.17062 0.0157	0.33594 <.0001	0.16135 0.0225	-0.40540 <.0001	0.10488 0.1394	0.26567 0.0001	0.12999 0.0666	1.00000	0.28592 <.0001	0.18518 0.0087	-0.14964 0.0344
CentralAir_Dummy	0.39834 <.0001	0.11241 0.1130	0.30942 <.0001	0.24635 0.0004	-0.28802 <.0001	0.27094 0.0001	0.19477 0.0057	0.08634 0.2241	0.28592 <.0001	1.00000	0.27859 <.0001	-0.09066 0.2017
GarageType_Attached	0.30883 <.0001	0.17895 0.0112	0.35470 <.0001	0.23307 0.0009	-0.23897 0.0007	0.19625 0.0054	0.22949 0.0011	0.27062 0.0001	0.18518 0.0087	0.27859 <.0001	1.00000	-0.49925 <.0001
GarageType_Detached	-0.26245 0.0002	-0.08537 0.2294	-0.18008 0.0107	0.00785 0.9121	0.21855 0.0019	-0.06287 0.3765	-0.10094 0.1550	-0.18669 0.0081	-0.14964 0.0344	-0.09066 0.2017	-0.49925 <.0001	1.00000

The Pearson Correlation Coefficients table shows there is no significant multicollinearity among variables. In addition, Price_of_Sale has a stronger positive correlation with log_Living_Area rather than Living_Area. This table gives a good understanding of numerical variables and their correlation.

METHODOLOGY & ANALYSIS QUESTIONS:

To achieve the project's aim and develop a statistical model, the following research questions need to be addressed.

- How do a single or multiple independent variables (Living area, overall quality, and the presence of central air) influence the sale price of a house?
- Within the given dataset, which features are the most statistically significant predictors for the target variable?
- Do all the independent variables explain the target variable efficiently?
- Can a multiple linear regression model effectively predict the target variable, and how good a fit is it?

Based on the analysis, there is more than one variable to explain this project's target variable, so that multiple linear regression could be more suitable.

For efficiency and simplicity, the backwards technique has been used since it is visible from all variables and the criteria for each step to compare the functionality of the model. All variables were used in the zero step, and then each one was deleted based on the p-value and its influence.

5.4. Multiple Linear Regression with Backward Elimination

The REG Procedure
Model: MODEL1
Dependent Variable: Price_of_Sale Price_of_Sale

Number of Observations Read	200
Number of Observations Used	200

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.6805 and C(p) = 11.0000
The model is not of full rank. A subset of the model which is of full rank is chosen.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1.17813E11	11781300599	40.26	<.0001
Error	189	55303324502	292610183		
Corrected Total	199	1.731163E11			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	36447	28616	474666207	1.62	0.2044
Log_Living_area	1288.23261	7390.74400	8889999	0.03	0.8618
Garage_Area	5.44115	59.10898	2479497	0.01	0.9268
Count_Bedroom	4498.11614	2209.41333	1212821323	4.14	0.0432
Count_Bathroom	1886.99016	2135.36599	228498872	0.78	0.3780
Fireplaces	2155.54901	2823.01955	170599386	0.58	0.4461
CentralAir_Dummy	7246.95869	3489.34550	1262156065	4.31	0.0392
Overall_Qual	11670	1156.05560	29819943517	101.91	<.0001
House_Age	-387.49614	53.08100	15593588727	53.29	<.0001
GarageType_Attached	1510.53253	3105.10132	69246535	0.24	0.6272
GarageType_Detached	-4692.60966	3313.03135	587039338	2.01	0.1583

Step zero of the multiple linear regression model with backwards elimination shows the model with all variables has R-Square = 0.6805(nearly 68.05% Price_of_Sale can be explained by the independent variables). C(p) = 11.0000: A criterion shows a relationship with the number of independent variables (the lower the number, the better the effect).

Backward Elimination: Step 5

Variable Count_Bathroom Removed: R-Square = 0.6759 and C(p) = 3.7466

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1.170093E11	23401864789	80.92	<.0001
Error	194	56107006541	289211374		
Corrected Total	199	1.731163E11			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	43216	8087.34070	8258299378	28.55	<.0001
Count_Bedroom	5173.54046	1837.72149	2292090060	7.93	0.0054
CentralAir_Dummy	8114.18716	3387.73028	1659158056	5.74	0.0176
Overall_Qual	12038	1085.54641	35564915226	122.97	<.0001
House_Age	-392.85422	50.57068	17453426186	60.35	<.0001
GarageType_Detached	-6038.67402	2820.89591	1325331188	4.58	0.0335

Bounds on condition number: 1.2994, 29.65

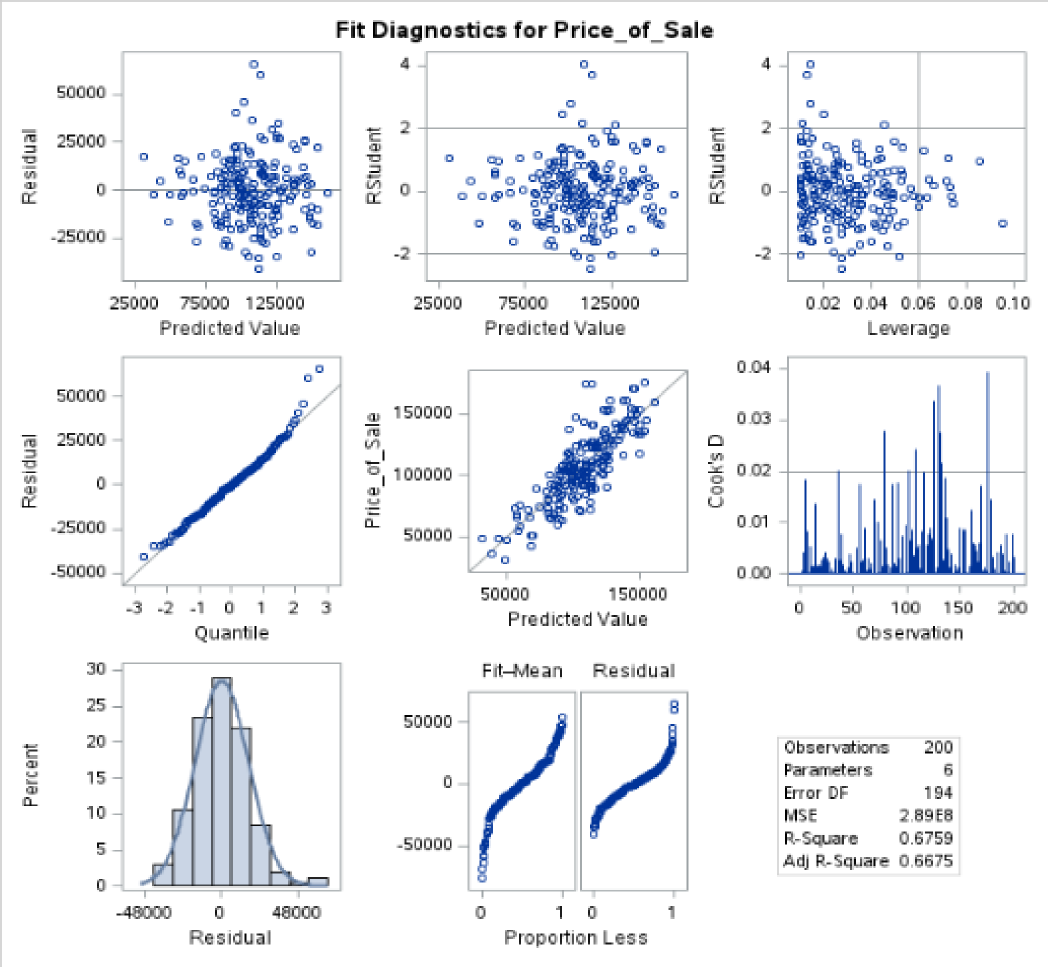
All variables left in the model are significant at the 0.0500 level.

The final stage of the model is after variable selection. The backwards technique eliminates all the variables with a non-significant value. The model includes five independent variables (DF for Model = 5). All variables are statistically significant predictors of the target variable (under 0.05 p value), also, Model F Value = 80.92, Pr > F = <.0001 shows it is a perfect fit for the data. A low condition number shows that multicollinearity is not concerning since the maximum range is under 30. In addition, C(p) is 3.7466 for five independent variables, which is a perfect number (showing simplicity and efficiency).

Compared to other steps, it still has a sophisticated R-Square, approximately 67.59%. Price_of_Sale can be explained by these five independent variables.

5.4. Multiple Linear Regression with Backward Elimination

The REG Procedure
Model: MODEL1
Dependent Variable: Price_of_Sale Price_of_Sale



These plots aim to show assumptions of linear regression and the overall fit for linear regression.

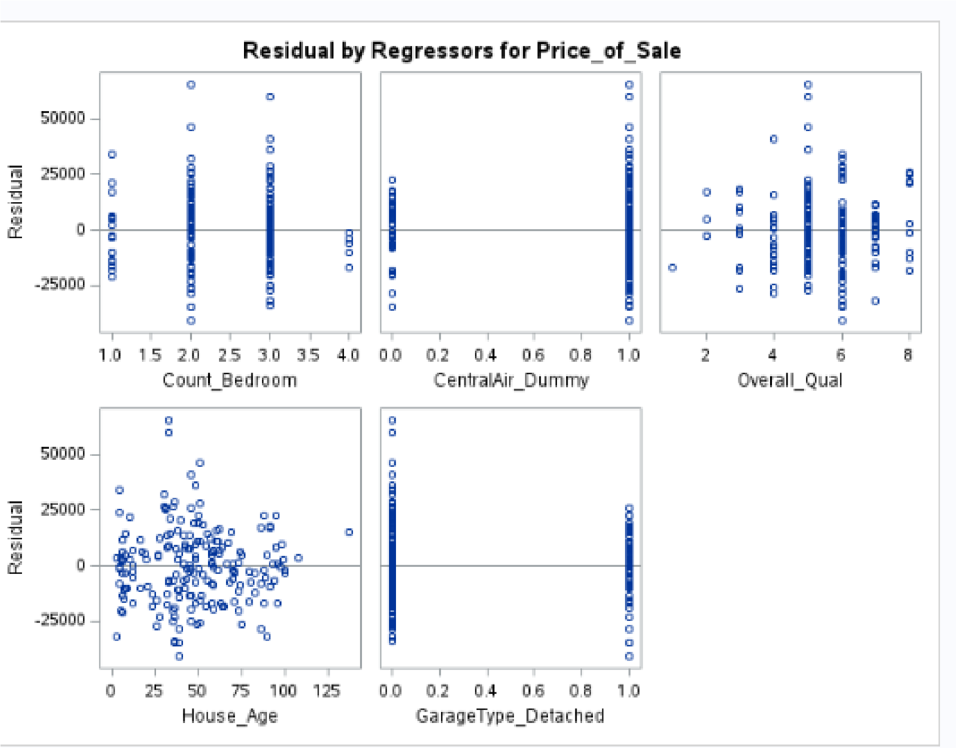
Residual vs. Predicted Value Plots: No pattern is shown, and there is no issue with linearity. Also, residuals are scattered around zero, showing unbiasedness across the range of predicted values.

Normal Quantile Plot of Residuals: Residuals are roughly normally distributed since they follow a straight line.

Observed vs. Predicted Values: Points along the diagonal line show the model is a good fit.

Cook's D: Most values are low, demonstrating no influential observation affecting the coefficients.

Residual Histogram: R-Square = 0.6759, Adj R-Square = 0.6675, shows around 67.6% of Price_of_Sale is explained by the model (good fit).



These plots check the assumption for linearity and constant variance of errors for the final influential variables.

The residuals are scattered randomly around the horizontal line (0) for all plots, with no pattern. This shows that each variable's linearity and constant variance assumptions are met. It should be considered that ordinal and dummy variables are dense in their category values which is normal.

CONCLUSION AND RECOMMENDATIONS:

A multiple linear regression model has been developed for this project to predict house sale prices. It explained around 67.6% of the variability in Price_of_Sale (Adj R-Square = 0.6675. House_Age (negative correlation), GarageType_Detached (negative correlation) Count_Bedroom, CentralAir_Dummy, and Overall_Qual are determined as significant predictors. In addition, the model's assumptions were checked.

However, small dataset size and a lack of variables, such as location, could be considered for future projects to improve. Future projects could explore more advanced models, such as tree-based models.

This analysis provides excellent insight for understanding house pricing, which could be helpful in real estate and for prospective buyers.

REFERENCES:

- Field, A. (2018) *Discovering Statistics Using SAS*. 5th edn. London: SAGE Publications.
- Hill, R.C., Griffiths, W.E. and Lim, G.C. (2018). *Principles of Econometrics*. 5th edn. New York: John Wiley & Sons.
- Neter, J., Kutner, M.H., Nachtsheim, C.J. and Li, W. (2005) *Applied Linear Regression Models*. 4th edn. Boston: McGraw-Hill/Irwin.

- SAS Institute Inc. (no date) *SAS 9.4 Documentation and Support Resources***.** Available at: https://support.sas.com/en/software/visual-analytics-support.html#documentation (Accessed: 25 May 2025).
- SAS Institute Inc. (no date) *SAS 9.4 Documentation and Support Resources***.** Available at: https://support.sas.com/en/software/base-sas-support.html (Accessed: 25 May 2025).

APPENDIX:

```
LIBNAME MYSAS '/home/u64149066/sasuser.v94/MM711/MYSAS';

PROC IMPORT DATAFILE="/home/u64149066/sasuser.v94/MM711/MYSAS/price_sale.xlsx"
        OUT=MYSAS.PRICE
        DBMS=XLSX
        REPLACE;

    GETNAMES=YES;


LIBNAME MYSAS '/home/u64149066/sasuser.v94/MM711/MYSAS';
PROC IMPORT DATAFILE="/home/u64149066/sasuser.v94/MM711/MYSAS/Housing_data.xlsx"
        OUT=MYSAS.housing
        DBMS=XLSX
        REPLACE;

    GETNAMES=YES;
RUN;


/*merge data*/
/*this approach is chosen because avoid warnig for doplicate id with sql code*/
proc sort data=mysas.housing; by ID_House; run;
proc sort data=mysas.price; by ID_House; run;

data CombinedData;
   merge mysas.housing(in=a) mysas.price(in=b);
   by ID_House;
   if a and b;
run;



/*explanatory analysis 1*/

/* View column types and structure */
proc contents data=CombinedData; run;

/* Descriptive statistics */
proc means data=CombinedData n nmiss mean std min max;
   var Price_of_Sale Living_Area Garage_Area  Year_Construction Year_Sold;
run;

/* Frequency tables for categorical variables */
proc freq data=CombinedData;
   tables Garage_type Central_Air Fireplaces Count_Bathroom Count_Bedroom Overall_Qual/ missing;
run;


/* Histogram of Sale Price */
proc sgplot data=CombinedData;
   histogram Price_of_Sale;

   density Price_of_Sale / type=normal;
run;
proc sgplot data=CombinedData;
```

```
   HISTOGRAM LIVING_AREA;
   density LIVING_AREA / type=normal;
run;

/* Scatter plot of Living Area vs Price */
proc sgplot data=CleanData;
   scatter x=Living_area y=Price_of_Sale;
   reg x=Living_area y=Price_of_Sale;
run;

/* Standardize Central_Air values */

data CleanData;
    set CombinedData(drop=ID_House);
   Central_Air = upcase(Central_Air);
   if Central_Air not in ("Y", "N") then Central_Air = "N"; /* assume invalids are no */

   /* Correct common typos in Garage_type */
   if strip(Garage_type) = "Atached" then Garage_type = "Attached";

    if Garage_type = "Attached" then Garage_type = "Attached"; /* Correct the common typo */
   if Garage_type = "NA" then Garage_type = "no_garage"; /* IMPORTANT: 'NA' explicitly means 'no garage' */
   if Garage_type = "" then Garage_type = "missing_info"; /* IMPORTANT: Blank means truly missing information */

   /* Handle missing numeric values , preventing skewing statistical analyses.*/
   if Living_area <= 0 then Living_area = .;
   if Garage_area < 0 then Garage_area = .;/*for future if there is any invalid number*/

  /* Year_Sold (2 missing values) */
   if Year_Sold = . then Year_Sold = 2012;

  /* Price_of_Sale (1 missing value) */
    if Price_of_Sale = . then Price_of_Sale = 107345.57;
     /* Add imputation for Living_Area (now 1 missing value) */
    if Living_area = . then Living_area = 103.00;


  /* Create dummy variables for regression */
   if Central_Air = "Y" then CentralAir_Dummy = 1; else CentralAir_Dummy = 0;
    /* Create dummy variables for regression */
   if Garage_type = "Attached" then GarageType_Attached = 1; else GarageType_Attached = 0;
   if Garage_type = "Detached" then GarageType_Detached = 1; else GarageType_Detached = 0;
   if Garage_type = "missing_info" then GarageType_MissingInfo = 1; else GarageType_MissingInfo = 0;

    /* Create new variables */
   /* Create House Age */
   House_Age = Year_Sold - Year_Construction;

/* Log Transformations for Skewed Variable */
   if Living_area > 0 then Log_Living_area = log(Living_area); else Log_Living_area = .;
run;

/* Histogram of Sale Price */

proc sgplot data=CleanData;

   HISTOGRAM log_LIVING_AREA;

   density log_LIVING_AREA / type=normal;

run;

/* Boxplot of Price by Garage Type */

proc sgplot data=CleanData;
```

```
    vbox Price_of_Sale / category=Garage_type;
run;
/* Scatter plot of Living Area vs Price */
proc sgplot data=CleanData;
    scatter x=log_Living_area y=Price_of_Sale;
    reg x=log_Living_area y=Price_of_Sale;
run;
proc corr data=CleanData;
    var Price_of_Sale Living_area Garage_Area House_Age Count_Bedroom Count_Bathroom Fireplaces Overall_Qual CentralAir_Dummy GarageType_Attached GarageType_Detached GarageType_MissingInfo;
    title "Correlation Matrix of Key Variables";
run;
proc reg data=CleanData;
    model Price_of_Sale = log_Living_area Garage_area Count_Bedroom Count_Bathroom Fireplaces
                          CentralAir_Dummy Overall_Qual House_Age
                          GarageType_Attached GarageType_Detached GarageType_MissingInfo /
                          selection=backward slstay=0.05;
    title "5.4. Multiple Linear Regression with Backward Elimination";
run;
quit;
```