Title: **PREDICTING HOUSE PRICES USING LINEAR REGRESSION IN SAS**
Author: **Valentine A**
Date: **May 2022**

The Case Study:

The aim of this project is to investigate how the housing prices are affected based on a set of characteristics available for this analysis, such as the overall quality of the house, the number of bedrooms, the living area in square metres, etc., using SAS programming.

Data Preparation:

The data we are using for this project are datasets which publicly are available here. There are two datasets in .csv formats, housing_charact.csv and sales_price.csv. The dataset proves to be reliable, original, comprehensive, current, and cited.

house_charact.csv: This dataset consists of 200 observations and 8 variables. With 2 categorical variables (**Garage_Type** & **Air_Cond**), and 6 numeric variables (**ID, Living_Area, Garage_Area, Nr_Bedroom, Nr_Bathroom, Fireplaces**).

sales_price: This dataset consists of 210 observations and 5 numerical variables (**ID, Year_Built, Yr_Sold, Price_of_Sale, Overall_Qual**).

Data Processing:

Importing the house_charact.csv dataset.

The CONTENTS Procedure

Data Set Name	MYSAS.HOUSE	Observations	200
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	09/05/2022 19:13:04	Observation Length	64
Last Modified	09/05/2022 19:13:04	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
7	Air_Cond	Char	1	$1.	$1.
8	Fireplaces	Num	8	BEST12.	BEST32.
3	Garage_Area	Num	8	BEST12.	BEST32.
4	Garage_Type	Char	8	$8.	$8.
1	ID	Num	8	BEST12.	BEST32.
2	Living_Area	Num	8	BEST12.	BEST32.
6	Nr_Bathroom	Num	8	BEST12.	BEST32.
5	Nr_Bedroom	Num	8	BEST12.	BEST32.

Importing the sales_price.csv dataset.

The CONTENTS Procedure

Data Set Name	MYSAS.SALES	Observations	210
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	09/05/2022 19:24:26	Observation Length	40
Last Modified	09/05/2022 19:24:26	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
1	ID	Num	8	BEST12.	BEST32.
5	Overall_Qual	Num	8	BEST12.	BEST32.
4	Price_of_Sale	Num	8	BEST12.	BEST32.
2	Year_Built	Num	8	BEST12.	BEST32.
3	Yr_Sold	Num	8	BEST12.	BEST32.

Next, we combine the housing_charact.csv and sales_price.csv datasets using the **ID** variable and the Inner Join statement. The output is a new dataset with 210 observations and 12 variables (2 categorical variables and 10 numerical variables)

The CONTENTS Procedure

Data Set Name	MYSAS.JOIN_HOUSE_SALES	Observations	210
Member Type	DATA	Variables	12
Engine	V9	Indexes	0
Created	09/05/2022 19:55:43	Observation Length	96
Last Modified	09/05/2022 19:55:43	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
7	Air_Cond	Char	1	$1.	$1.
8	Fireplaces	Num	8	BEST12.	BEST32.
3	Garage_Area	Num	8	BEST12.	BEST32.
4	Garage_Type	Char	8	$8.	$8.
1	ID	Num	8	BEST12.	BEST32.
2	Living_Area	Num	8	BEST12.	BEST32.
6	Nr_Bathroom	Num	8	BEST12.	BEST32.
5	Nr_Bedroom	Num	8	BEST12.	BEST32.
12	Overall_Qual	Num	8	BEST12.	BEST32.
11	Price_of_Sale	Num	8	BEST12.	BEST32.
9	Year_Built	Num	8	BEST12.	BEST32.
10	Yr_Sold	Num	8	BEST12.	BEST32.

Data Cleaning:

Identify missing values and invalid data

To identify the variables with invalid and missing values, we used **PROC FREQ** with the **NOCUM** and **NOPERCENT** options to list the frequencies of the variables. The output shows that the **Garage_Type** variable has 45 missing data and the **Air_Cond** variable has invalid data "n" and "y" which will need recoding.

Garage_Type	Frequency
Attached	89
Detached	53
NA	23
Frequency Missing = 45	

Air_Cond	Frequency
N	33
Y	167
n	4
y	6

We recoded the missing **Garage_Type** as NA and recoded the "n" and "y" **Air_Cond** values as "N" and "Y" respectively. After recoding the variables, we have the output below:

Air_Cond	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	37	17.62	37	17.62
Y	173	82.38	210	100.00

Garage_Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Attached	89	42.38	89	42.38
Detached	53	25.24	142	67.62
NA	68	32.38	210	100.00

Identify and remove duplicate data

We checked and removed true duplicates from the data by running the **PROC SORT** command with the **NODUPRECS** option using **ID** as the **BY** variable. The output removed 10 duplicates, leaving us with 200 observations.

```
NOTE: There were 210 observations read from the data set MYSAS.RECODE2.
NOTE: 10 duplicate observations were deleted.
NOTE: The data set WORK.NO_DUPLICATES has 200 observations and 12 variables.
```

Transforming Variables

To get the number of years between when the house was built and when it was sold, we need to create a new variable **"Years-Before_Sale"** by subtracting the **"Year_Built"** from the **"Yr_Sold".**

The transformed data head is shown below:

ID	Living_Area	Garage_Area	Garage_Type	Nr_Bedroom	Nr_Bathroom	Air_Cond	Fireplaces	Year_Built	Yr_Sold	Price_of_Sale	Overall_Qual	Years_Before_Sale
27145	119	47	Attached	2	2	Y	0	1992	2010	153200	8	18
27146	80	0	NA	3	1	Y	1	1971	2010	92000	4	39
27147	133	37	Attached	3	2	Y	1	2000	2010	144000	7	10
27148	70	25	Attached	2	2	Y	0	1984	2010	100000	6	26
27149	114	45	Attached	1	2	Y	1	1978	2010	160800	8	32
27150	107	45	Attached	3	2	Y	1	1964	2010	144240	5	46
27151	100	22	NA	2	2	Y	0	1950	2010	113700	5	60
27152	118	23	Attached	2	1	Y	1	1954	2010	105600	5	56
27153	105	29	Attached	2	2	Y	1	1966	2010	141840	6	44
27154	99	30	Detached	2	1	Y	0	1952	2010	107920	5	58

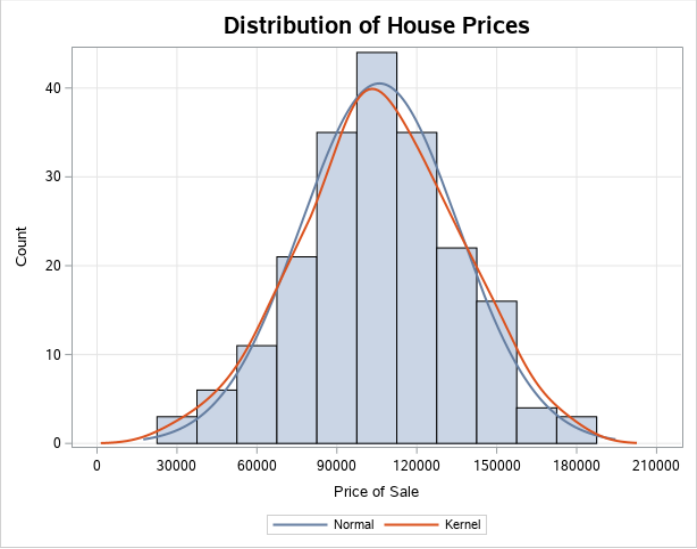
<u>Descriptive Statistics for Numerical Variables:</u>

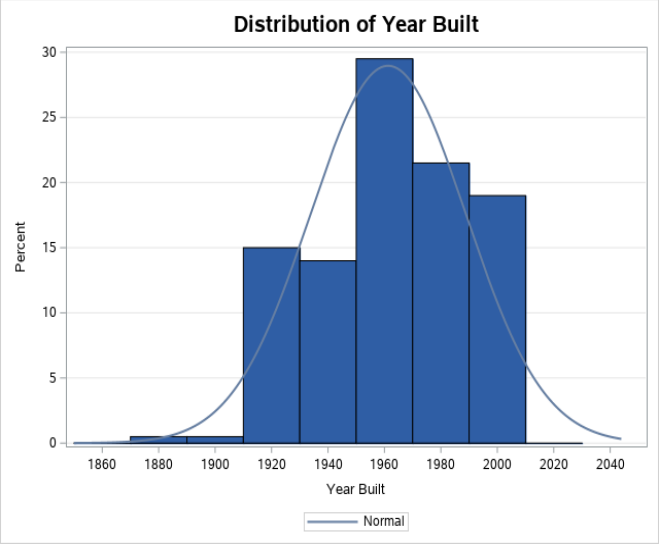
The MEANS Procedure

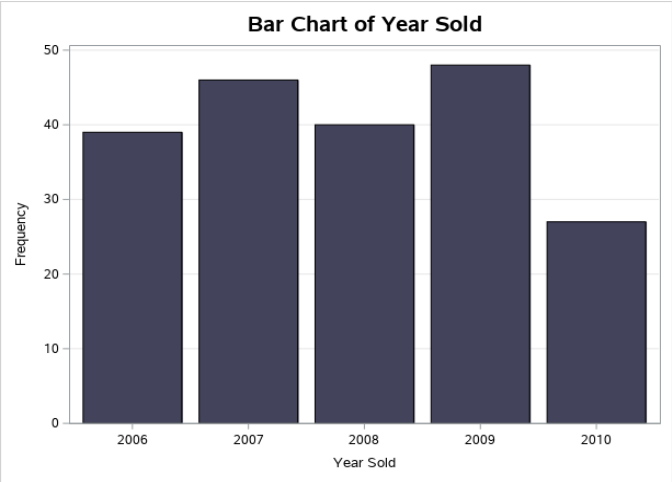
Variable	N	N Miss	Minimum	Maximum	Mean	Median	Std Dev
ID	200	0	27145.00	27344.00	27244.50	27244.50	57.88
Living_Area	200	0	13.00	1330.00	109.11	103.00	89.72
Garage_Area	200	0	0.00	380.00	35.29	35.50	29.85
Nr_Bedroom	200	0	1.00	4.00	2.51	3.00	0.68
Nr_Bathroom	200	0	1.00	3.00	1.62	2.00	0.63
Fireplaces	200	0	0.00	2.00	0.36	0.00	0.56
Year_Built	200	0	1875.00	2007.00	1961.31	1963.00	27.55
Yr_Sold	200	0	2006.00	2010.00	2007.89	2008.00	1.34
Price_of_Sale	200	0	28000.00	176000.00	105977.64	104600.00	29538.97
Overall_Qual	200	0	1.00	8.00	5.34	5.00	1.24
Years_Before_Sale	200	0	1.00	135.00	46.59	45.50	27.48

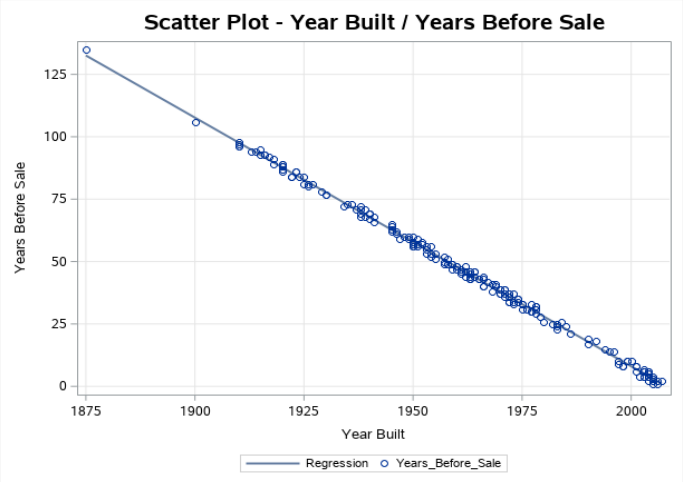
From the descriptive statistics using **PROC MEANS**, we can see that the mean and median of the numerical variables have close values. This is an indication that the data is likely to have a normal distribution.

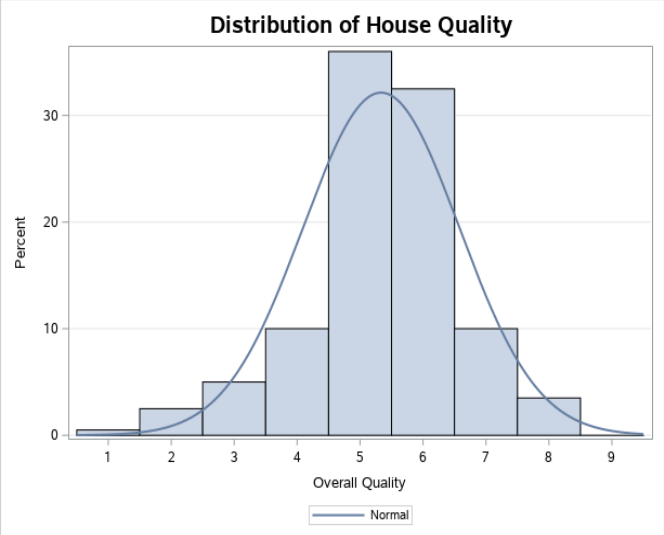
Data Visualization:

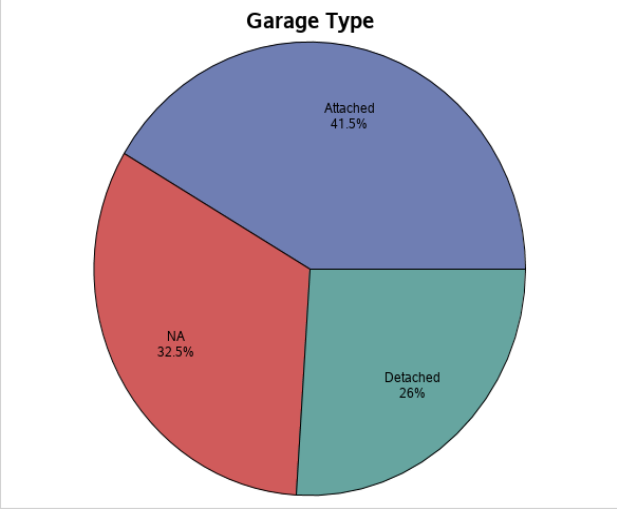






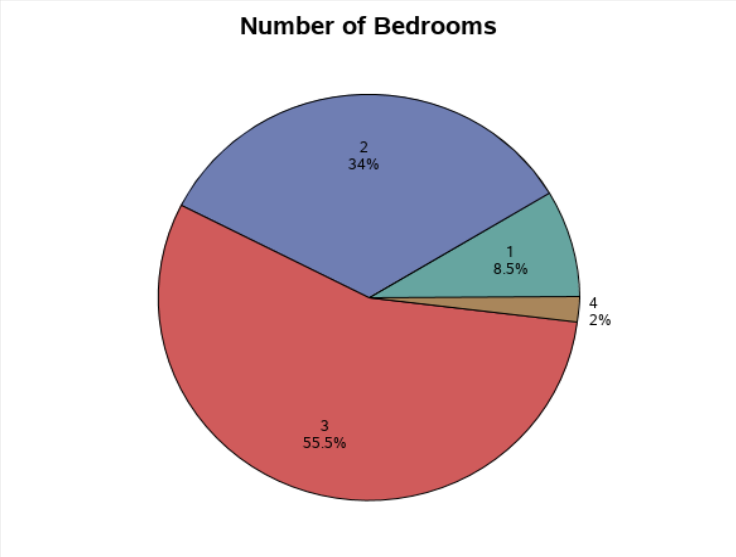






Visualization Summary

a) The House Prices and Year Built show a normal distribution.
b) The Year Sold chart peaked in 2009 and is lowest in 2010.
c) The House Quality also shows a normal distribution.
d) We can see a strong negative correlation between the Year Built and the Years Before Sale.
e) Houses with 3 bedrooms make up 55.5% of total houses.



Statistical Analysis and Modelling:

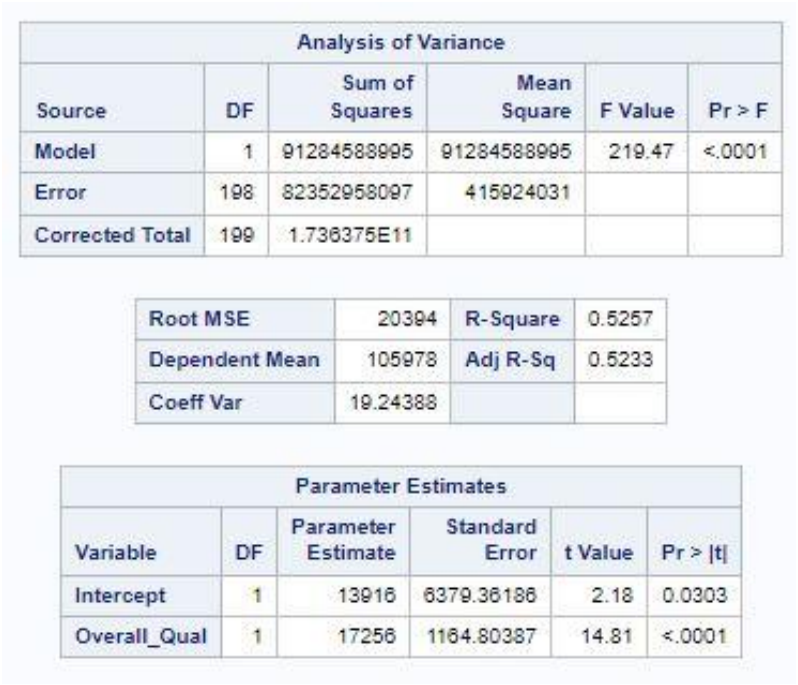
Pearson Correlation Coefficient.

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

	ID	Living_Area	Garage_Area	Nr_Bedroom	Nr_Bathroom	Fireplaces	Year_Built	Yr_Sold	Price_of_Sale	Overall_Qual	Years_Before_Sale
ID	1.00000	-0.08229 0.2467	-0.05579 0.4326	-0.01916 0.7878	-0.09955 0.1608	-0.02798 0.6941	-0.11746 0.0976	-0.97750 <.0001	-0.12452 0.0789	-0.08852 0.2126	0.07020 0.3232
Living_Area	-0.08229 0.2467	1.00000	0.84689 <.0001	0.19201 0.0065	0.13327 0.0599	0.15874 0.0248	0.14397 0.0420	0.06031 0.3962	0.23581 0.0008	0.24551 0.0005	-0.14140 0.0458
Garage_Area	-0.05579 0.4326	0.84689 <.0001	1.00000	0.10608 0.1349	0.22800 0.0012	0.12119 0.0874	0.30788 <.0001	0.03014 0.6718	0.39405 <.0001	0.30116 <.0001	-0.30720 <.0001
Nr_Bedroom	-0.01916 0.7878	0.19201 0.0065	0.10608 0.1349	1.00000	0.09096 0.2002	0.08311 0.2420	-0.02632 0.7114	0.00122 0.9864	0.27094 0.0001	0.21338 0.0024	0.02645 0.7100
Nr_Bathroom	-0.09955 0.1608	0.13327 0.0599	0.22800 0.0012	0.09096 0.2002	1.00000	0.03367 0.6359	0.48001 <.0001	0.08728 0.2191	0.55085 <.0001	0.31116 <.0001	-0.47698 <.0001
Fireplaces	-0.02798 0.6941	0.15874 0.0248	0.12119 0.0874	0.08311 0.2420	0.03367 0.6359	1.00000	0.03169 0.6559	0.01965 0.7824	0.33937 <.0001	0.21660 0.0021	-0.03082 0.6649
Year_Built	-0.11746 0.0976	0.14397 0.0420	0.30788 <.0001	-0.02632 0.7114	0.48001 <.0001	0.03169 0.6559	1.00000	0.07652 0.2815	0.61606 <.0001	0.41000 <.0001	-0.99882 <.0001
Yr_Sold	-0.97750 <.0001	0.06031 0.3962	0.03014 0.6718	0.00122 0.9864	0.08728 0.2191	0.01965 0.7824	0.07652 0.2815	1.00000	0.09344 0.1882	0.05867 0.4093	-0.02806 0.6932
Price_of_Sale	-0.12452 0.0789	0.23581 0.0008	0.39405 <.0001	0.27094 0.0001	0.55085 <.0001	0.33937 <.0001	0.61606 <.0001	0.09344 0.1882	1.00000	0.72506 <.0001	-0.61308 <.0001
Overall_Qual	-0.08852 0.2126	0.24551 0.0005	0.30116 <.0001	0.21338 0.0024	0.31116 <.0001	0.21660 0.0021	0.41000 <.0001	0.05867 0.4093	0.72506 <.0001	1.00000	-0.40818 <.0001
Years_Before_Sale	0.07020 0.3232	-0.14140 0.0458	-0.30720 <.0001	0.02645 0.7100	-0.47698 <.0001	-0.03082 0.6649	-0.99882 <.0001	-0.02806 0.6932	-0.61308 <.0001	-0.40818 <.0001	1.00000

For this project, we will demonstrate a linear regression model, using **Price_of_Sale** as the response variable. The most correlated predictor with **Price_of_Sale** from the correlation table above is **Overall_Qual** (0.72508). Other predictors that are significantly correlated with **Price_of_Sale** include **Year_Built** (0.61606), **Years_Before_Sale** (-0.61308), **Nr_Bathroom** (0.55085).

Simple Linear Regression



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	91284588995	91284588995	219.47	<.0001
Error	198	82352958097	415924031		
Corrected Total	199	1.736375E11			

Root MSE	20394	R-Square	0.5257
Dependent Mean	105978	Adj R-Sq	0.5233
Coeff Var	19.24388		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13916	6379.36186	2.18	0.0303
Overall_Qual	1	17256	1164.80387	14.81	<.0001

Explaining the Simple Linear regression

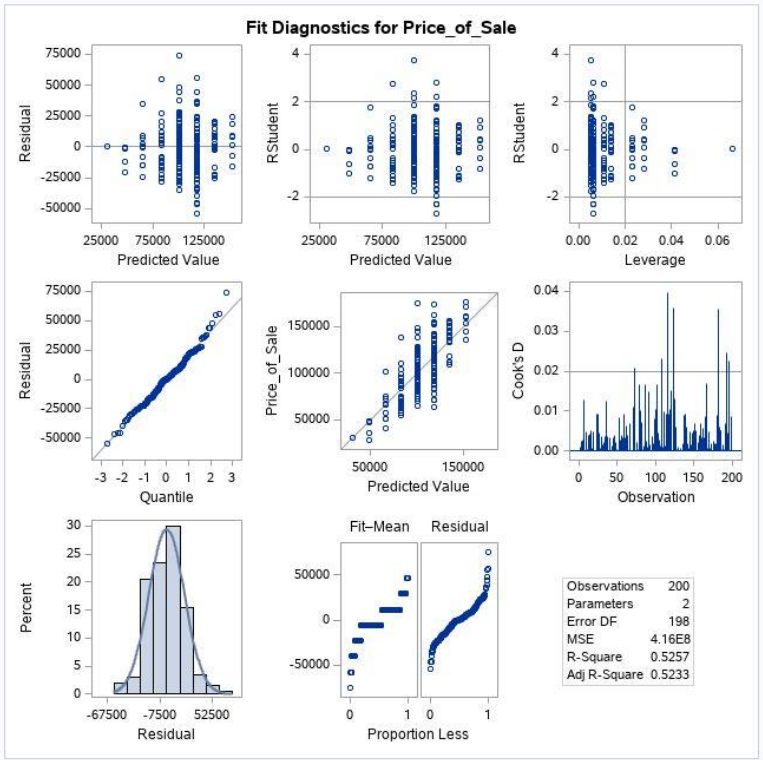
The Overall average of **Price_of_Sale** is 105978.

The R-Square value is 0.5257. This shows that the variability of **Overall_Qual** is explaining about 52% of the variability in **Price_of_Sale**.

From the F value and the independent variable, the p-value is <0.0001, this is significantly different from 0, which means that we have a significant result. Hence, we must reject the Null Hypothesis which states that the simple linear regression model does not fit the data better than the baseline model.

The coefficient of the independent variable **Overall_Qual** is 17256, this means that for every time that the **Overall_Qual** increase by 1 unit, there will be a 17256 increase in the **Price_of_Sale.**

Model: MODEL1
Dependent Variable: Price_of_Sale



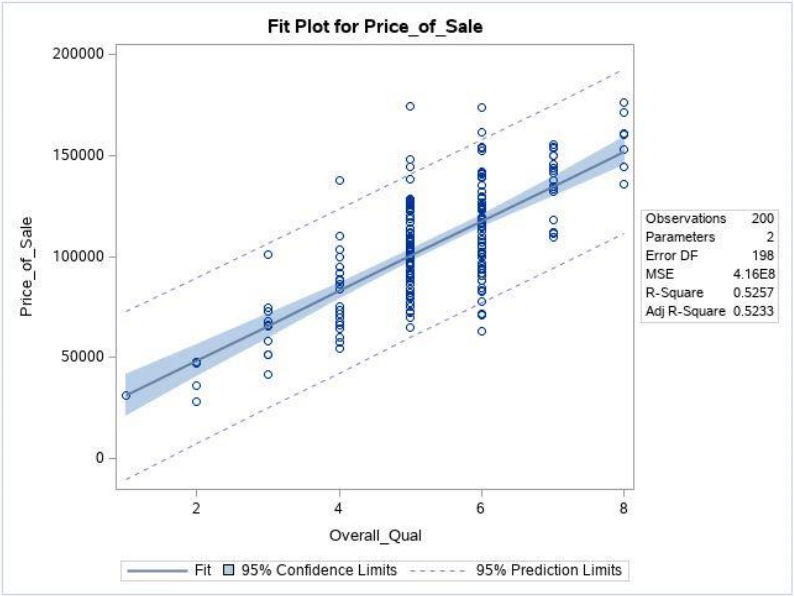
Using the Fit Diagnostics to Validate Assumptions for Linear regression

Linear: There's a linear relationship between the target variable and the input variable. A straight line connects the response variable at each value of the predictor variable

Independence: We used residual to validate the independence. From the Fit Diagnostics above we can confirm that the observations are scattered and independent. This means that knowing something about one point cannot tell us anything about the next point.

Normal Distribution of Errors: The Quantile plot can be used to validate the normal distribution errors by observing that the observations are fitted on the diagonal line. The histogram can also be used to validate the normal distribution of errors. And from the chart above it looks relatively normal.

Equal Variability of Errors: This means that there should be no patterns in the variability of errors. The residual above shows that the observations are completely random



We have a positive relationship from the Fit plot for Price_of_Sale, with a 95% confidence interval.

Conclusion:

During the project, we were able to prepare and process the data, remove duplicates, and identify missing and invalid data. We used data visualization to view the distribution of the variables. Finally, we fitted a simple linear regression model to predict the sale price from the overall quality, with a 95% confidence interval, and validated our assumptions for the linear regression.