STATISTICAL ANALYSIS ON

PROPERTY VALUE

GROUP PROJECT
DANA-4810

Dharunkumar D
100342119

Roberta Bukowski
100342032

Linebeth Ruales
100339962

Kirra Widjaja
100346067

Vedant Mayor
100345949

Jasmanpreet Kaur
100342127

INTRODUCTION

What are the components that influence the sale price of a home? Could it be the number of bathrooms, the square foot area of the house or even if it is a detached home? It seems that several aspects may influence the final sale price of a house. This leads us to a question: Can house prices be predicted using statistical techniques? What variables driven the price? Are the prices driven by rationality?

This project aims to understand house sales prices or at least have a glimpse of how the behavior. Predict sales price is a challenging problem in the US a company called Zillow has launched a $1 million competition called the Zillow Prize to create an algorithm to improve their accuracy in house price evaluations[1].

Buying a house is an important financial transaction, and in the housing market, the transactions are negotiated individually this leads the market to inefficiency. Also, due to asymmetrical information, inexperienced buyers pay more than the house is worth[2]. Likewise, there is a tendency for experienced property investors to consistently overestimate the worth of a property.[3]

The real estate market tends to be affected by factors external to the market itself, such as GDP growth, employment levels, uncertainties about the future, among others. There is a wide variety of research papers on the real estate market and how to predict house sale prices. This project will attempt to construct a model to predict house sales prices.

In this project we will be using The Ames Housing dataset the dataset is a modernized and expanded version of the Boston Housing dataset.

OBJECTIVE

The main motivation to complete this study is to predict the behaviour of house sales price.

RESEARCH QUESTIONS

This project will attempt to answer the following questions.

- Can statistical techniques explain the variation in house price based solely on the fundamental characteristics?
- How accurate and realistic is a model to predict house sales prices?
- What variable explain most the sale price of a house?
- The squared meters area of a house is strongly related to the house price?

DATASET

12 Variables

Dataset consists of 14 variables overall before variable screening

1 Response Variables

the property's sale price in dollars which is numerical. "SalePrice"

11 Independant Variables

which is a combination of categorical and numerical variables

2 Categorical Variables

We have 2 categorical variables Shape and Utilities containing 4 levels and 2 levels respectively.

9 Numerical Variables

Dataset consists of 9 variables that provide numerical information

4 Dummy Variables

For the categorical variables mentioned above. Shape: 3 and Utilities: 1

1320 Sample Size

For the categorical variables mentioned above. Shape: 3 and Utilities: 4

VARIABLES

LotArea — **Numerical**
the lot size in square feet

AgeOfHouse — **Numerical**
The age of the house

AllUtilities — **Dummy**
whether the house has all utilities installed or not

GarageArea — **Numerical**
the size of the garage in square feet (GarageArea)

PoolArea — **Numerical**
pool area in square feet (PoolArea),

TotalBsmntSF — **Numerical**
total square feet of basement area (TotalBsmntSF)

GrLivArea — **Numerical**
above grade (ground) living area square feet (GrLivArea),

FullBath — **Numerical**
number of full bathrooms above grade (FullBath)

KitchenAbvGrd — **Numerical**
number of kitchens above grade (KitchenAbvGrd),

TotRmsAbvGrd — **Numerical**
total number of rooms above grade (excluding bathrooms)

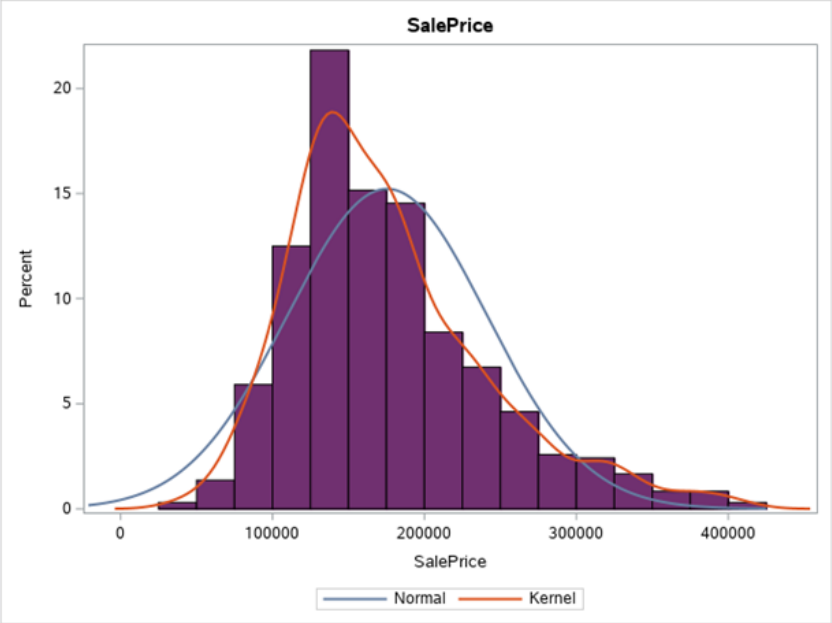
ShapeIR1 — **Dummy**
whether the shape of the property is slightly irregular

ShapeIR2 — **Dummy**
whether the shape of the property is moderately irregular

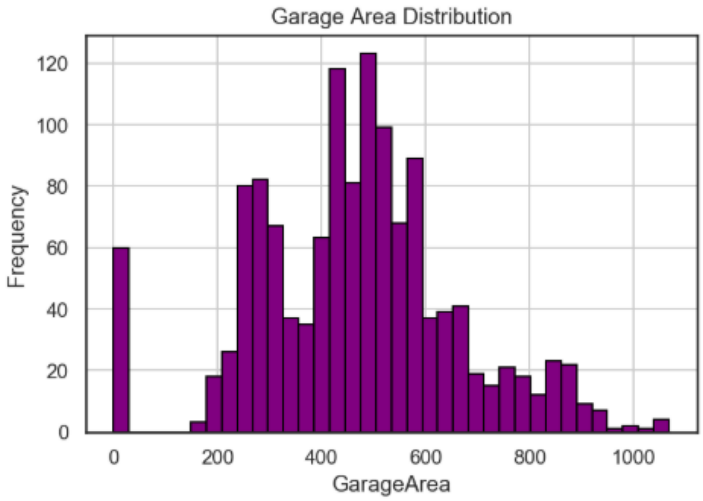
ShapeIR3 — **Dummy**
Whether the shape of the property is irregular

PRELIMINARY ANALYSIS

In this project, we are going to predict the housing sale prices (SalePrice in $USD). We have the below distribution of the sale price of houses. The average price of the houses is approximately 175,409 dollars. Most of the houses were approximately from 110,000 to 241,000 dollars. According to the histogram it looks like the sample is normally distributed for this variable.

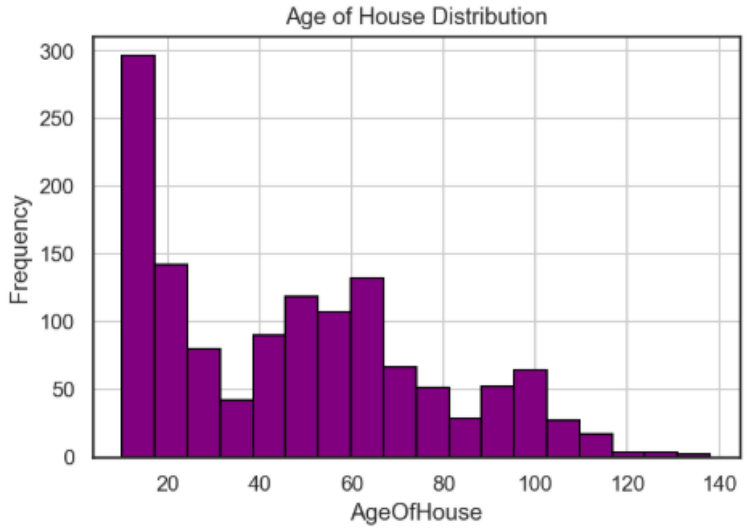


Univariate Analysis of Garage Area (in square feet).



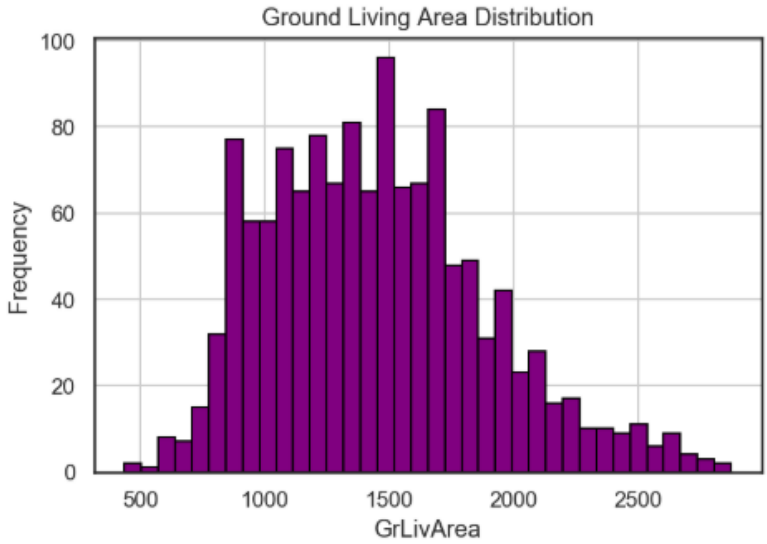
The average area of the garage of the houses is approximately 464 square feet. Around 68% of the houses has an area of between 267 to 661 square feet. According to the histogram it looks like the sample is normally distributed.

Univariate Analysis of Age of House (in years).



From the distribution, with an average of 47 years, approximately 68% of the houses are in between 17 and 77 years old. The sample is a little skewed for this.

Univariate Analysis of Ground Living Area (in square feet).



From the distribution, with an average area of 1,451 square feet the majority of the houses are between 1011 to 1891 square feet. According to the histogram it looks like the sample is normally distributed.

Summary Table of Sale Price, Garage Area, Age of the House and Ground Living Area

	MEAN	STANDARD DEVIATION	MEDIAN	25%	50%	75%	MAX
Sale Price	175409.8	65532.6	34900.0	130000	163000	209625	415298
Garage Area	464.67	196.7	0	319.8	472.5	576.0	1069.0
Age of House	47.6	29.1	10	19	47	88	138
Ground Living Area	1451.2	440.1	438	1113	1423.5	1716.3	2872

Inferential Analysis

MODEL BUILDING

In this project we are working with multiple linear regression, however, the initial dataset contained many variables and identifying the ideal model could be somewhat challenging. After cleaning and selecting data, we have fifteen variables (one response variable, fourteen predictors variables where four are dummy variables and ten numerical variables) and we already knew that our variable response (dependent variable) would be the sale price. To select the best possible model, we used stepwise regression in SAS.

The Stepwise selection starts with a model with no predictors, then he adds the variable with the largest F-statistic and refit with this variable added, and so on. The Stepwise will use t-statistics to "search" for a model and based on t-statistic will add or delete a variable of the model this will be repeated till no variables can be added or deleted (based in the alpha that we set initially α=0.05).

After stepwise selection our selected model was a straight-line model for the response y in terms of x:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

Line of Means:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

The Prediction Equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$$

Where:

x1 = Lot Area; x2 = Age of house

x3 = Total square feet of basement area

x4 = Above grade (ground) living area square feet

x5 = Garage Area Size of garage in square feet.

Inferential Analysis

MODEL FITTING

The model was fit to the data using SAS:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4.520421E12	9.040842E11	1038.39	<.0001
Error	1314	1.144051E12	870662801		
Corrected Total	1319	5.664472E12			

Root MSE	29507
Dependent Mean	175410
R-Square	0.7980
Adj R-Sq	0.7973
AIC	28500
AICC	28500
SBC	27209

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	19853	4472.513970	4.44	<.0001
LotArea	1	0.839147	0.217129	3.86	0.0001
AgeOfHouse	1	-589.654407	34.055514	-17.31	<.0001
TotalBsmtSF	1	46.526419	2.666588	17.45	<.0001
GrLivArea	1	70.075748	2.195075	31.92	<.0001
GarageArea	1	55.396945	5.472223	10.12	<.0001

$$\widehat{y} = 19853 + 0.839147x_1 - 589.654407x_2 + 46.526419x_3 + 70.075748x_4 + 55.396945x_5$$

$$\widehat{saleprice} = 19853 + 0.839147_{lotarea} - 589.654407_{ageofhouse} + 46.526419_{totalbsmtsf} + 70.075748_{grlivarea} + 55.396945_{garagearea}$$

Inferential Analysis

Model Fitting(cont.d)

$$\widehat{saleprice} = 19853 + 0.839147_{lotarea} - 589.654407_{ageofhouse} + 46.526419_{totalbsmtsf} + 70.075748_{grlivarea} + 55.396945_{garagearea}$$

These estimates tell us about the relationship between the independent variables and the dependent variable. These estimates tell the amount of increase in sales price that would be predicted by a one unit increase in the predictor variable.

About Above grade concept- Above grade means the portion of a house that is above the ground. The term is usually used to describe a room or square footage. For example, two bedrooms above grade means two bedrooms that are not located in a basement.

Interpretations

β_0

the intercept of the model, which means that if all the other variables are zero, we can expect that the sale price of a house will be 19853(US dollars)

β_1

for every unit increase in lote area (squared meters) a 0.839147-unit increase (US dollars) in sale price is predicted, holding all other variables constant

β_2

for every unit increase in age of house (years) a 589.654407-unit decrease (US dollars) in sale price is predicted, holding all other variables constant.

β_3

for every unit increase in total basement area (squared meters) a 46.526419 unit increase in (US dollars) in sale price is predicted, holding all other ariables constant.

β_4

for every unit increase in total ground living area (squared meters) a 70.075748 unit increase in (US dollars) in sale price is predicted, holding all other variables constant

β_5

for every unit increase in Garage Area (squared meters) a 55.396945-unit increase (US dollars) in sales price is predicted, holding all other variables constant.

Inferential Analysis

MODEL EVALUATION

R- square and adjusted R square (Goodness of fit)

Root MSE	29507
Dependent Mean	175410
R-Square	0.7980
Adj R-Sq	0.7973
AIC	28500
AICC	28500
SBC	27209

The R-squared measure how close the data are to the fitted regression line. The selected model has an R2 of 0.7980 or 79.80%, which means that 79.80% of the house sales price is explained by the proposed model.

The adjusted R-squared value is 0.7973 or 79.73%, the interpretation is the same as R-squared, however, the adjusted R-squared discounts the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.

To evaluate the model utility we have to perform a global F-test

$$\begin{cases} H_0: & \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_A: & At\ least\ one\ \beta \neq 0\ for\ i = 1\ to\ 5 \end{cases}$$

From the SAS output, we can see that the p-value is lesser than 0.0001 which is lesser than our α = 0.05, therefore we reject the null hypothesis, this means that the model is statistically significant.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4.520421E12	9.040842E11	1038.39	<.0001
Error	1314	1.144051E12	870662801		
Corrected Total	1319	5.664472E12			

Inferential Analysis

MODEL EVALUATION

P-value check to find out if the interaction term is useful

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	4.587105E12	7.645174E11	931.73	<.0001
Error	1313	1.077367E12	820538632		
Corrected Total	1319	5.664472E12			

Root MSE	28645	R-Square	0.8098
Dependent Mean	175410	Adj R-Sq	0.8089
Coeff Var	16.33036		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	69966	7053.61125	9.92	<.0001
LotArea	1	0.84603	0.21079	4.01	<.0001
AgeOfHouse	1	-576.21243	33.09430	-17.41	<.0001
TotalBsmtSF	1	45.84881	2.58978	17.70	<.0001
GrLivArea	1	32.25130	4.70590	6.85	<.0001
GarageArea	1	-50.24083	12.86609	-3.90	<.0001
Interactionterm	1	0.07532	0.00835	9.01	<.0001

Alternate model for the response y in terms of x with an interaction term:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_4 x_5 + \varepsilon$$

Line of Means:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_{5+} + \beta_6 x_4 x_5$$

The Prediction Equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_4 x_5$$

Inferential Analysis

MODEL EVALUATION

P-value check to find out if the interaction term is useful

Now let us check the p value at alpha = 0.05 to find out if there is sufficient evidence to indicate that the interaction term between GrLivArea and GarageArea (β6) is a useful predictor of sale price.

$$\begin{cases} H_0: \beta_6 = 0 \\ H_A: \beta_6 \neq 0 \end{cases}$$

P value = <0.0001
P value < 0.05
We reject H0 and conclude that there is sufficient evidence to indicate that the interaction term is a useful predictor of sale price, adjusting for all other independent variables at alpha = 0.05.

$$\widehat{saleprice} = 19853 + 0.839147_{lotarea} - 589.654407_{ageofhouse} + 46.526419_{totalbsmtsf} + 70.075748_{grlivarea} + 55.396945_{garagearea} + 0.07532_{grlivingarea*garagearea}$$

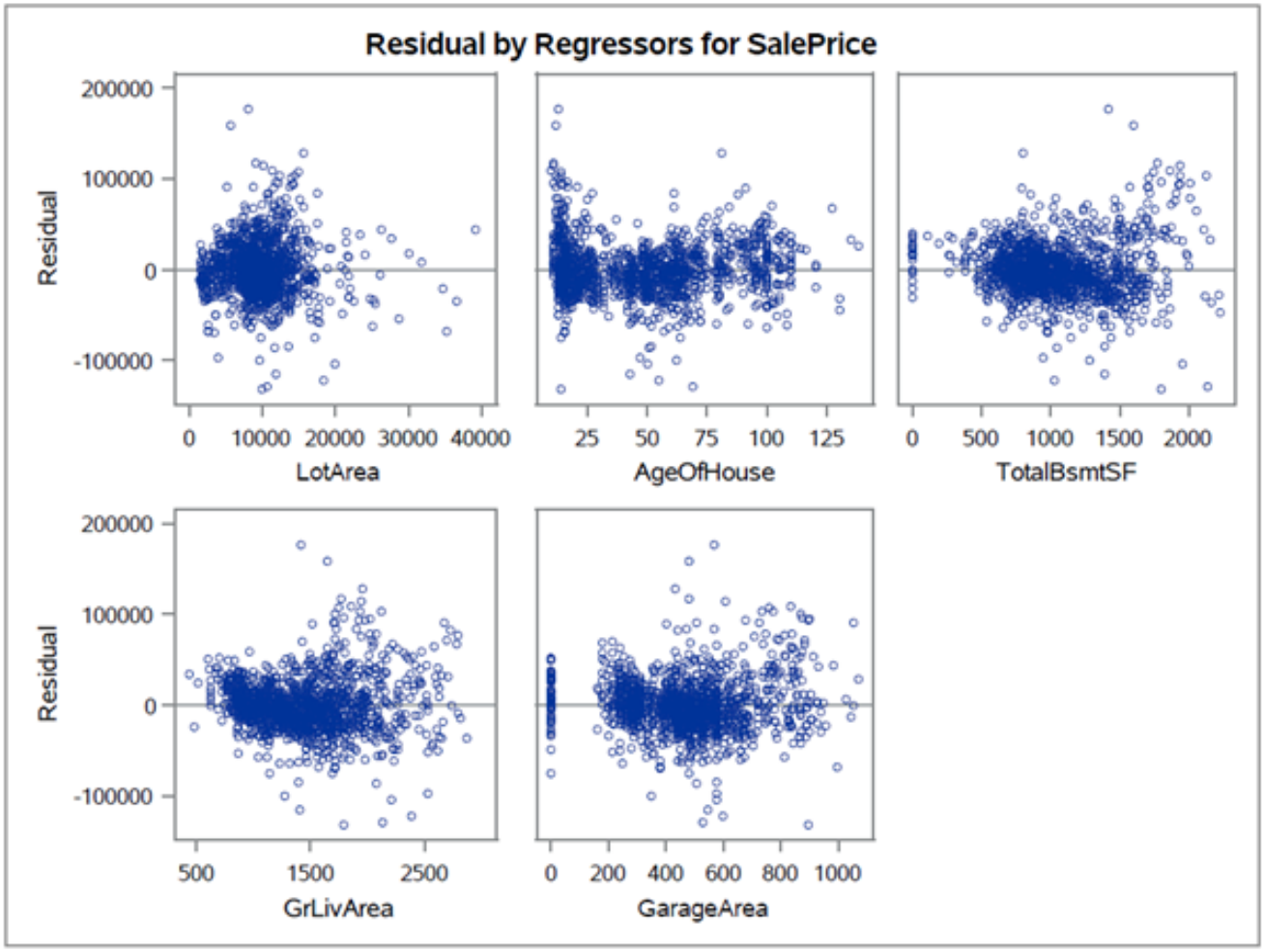
Inferential Analysis

RESIDUAL ANALYSIS

$$\widehat{y} = 19853 + 0.839147x_1 - 589.654407x_2 + 46.526419x_3 + 70.075748x_4 + 55.396945x_5$$

When analyzing the residuals three assumptions need to be satisfied, one is linearity also known as lack of fit, the second is homoscedasticity the third is normality.

Lack of Fit:

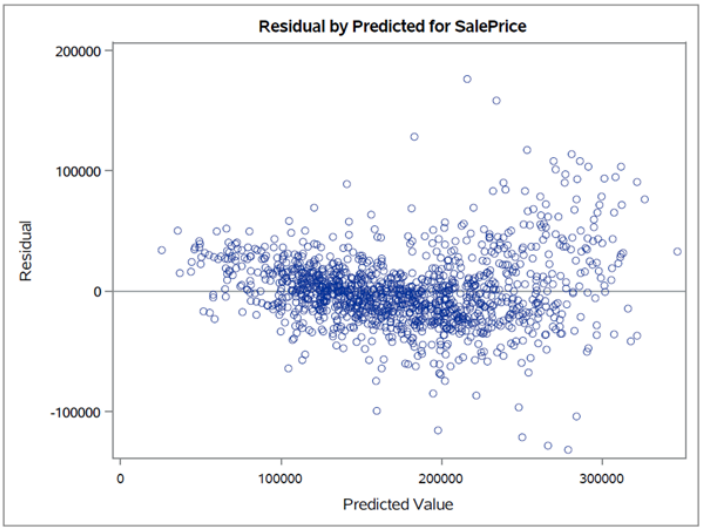


The residuals plotted against the predictors doesn't reveal a pattern is expected that part of the data be concentrated around the middle.

Inferential Analysis

RESIDUAL ANALYSIS

Homoscedasticity



The graph seems to show a type of pattern, however, we don't have enough evidence to say that if there is heteroscedasticity just by examining the graph. We performed the Breusch-Pagan test to check if the residuals are heteroskedastic.

Interpreting the Breusch-Pagan/Cook Weisberg test

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of saleprice

         chi2(1)      =    260.40
         Prob > chi2  =    0.0000
```

The null hypothesis is that the error variances are all equal versus the alternative that the error variances are a multiplicative function of one or more variables. In the test shown above the alternative hypothesis states that the error variances increase (or decrease) as the predicted values of Y increase, (the bigger the predicted value of Y, the bigger the error variance is).

A large chi-square indicates that heteroskedasticity is present, so by the previous test we can conclude that there is heteroskedasticity present.

One thing to notice is that the Breusch-Pagan test specified is a test for linear forms of heteroskedasticity, (goes up, the error variances follow the same movement). This test does not perform well for non-linear forms of heteroskedasticity since our graph looks oval, we going to apply another test.

Inferential Analysis

RESIDUAL ANALYSIS

White's General Test

White’s general test is a special case of the Breusch-Pagan test, where the assumption of normally distributed errors has been relaxed

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(20)      =    217.98
         Prob > chi2   =    0.0000

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	217.98	20	0.0000
Skewness	25.29	5	0.0001
Kurtosis	-252813.54	1	1.0000
Total	-252570.26	26	1.0000

As we can see from the output above the p-value is lesser than 0.0001 which is lesser than the default $\alpha = 0.05$ this means that we can reject the null hypothesis and conclude that there is heteroskedasticity present.

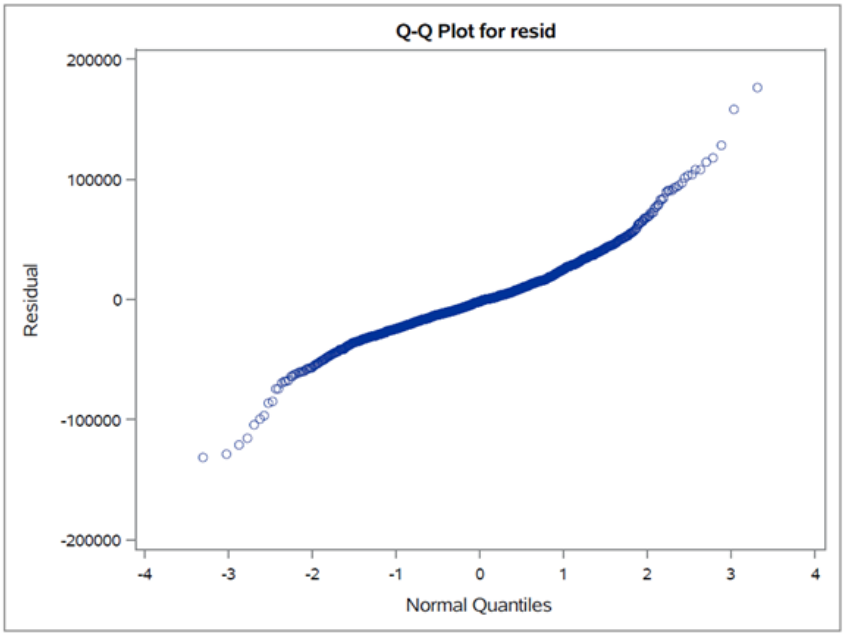
Therefore, the model doesn’t meet the assumption of homoscedasticity in the residuals.

Inferential Analysis

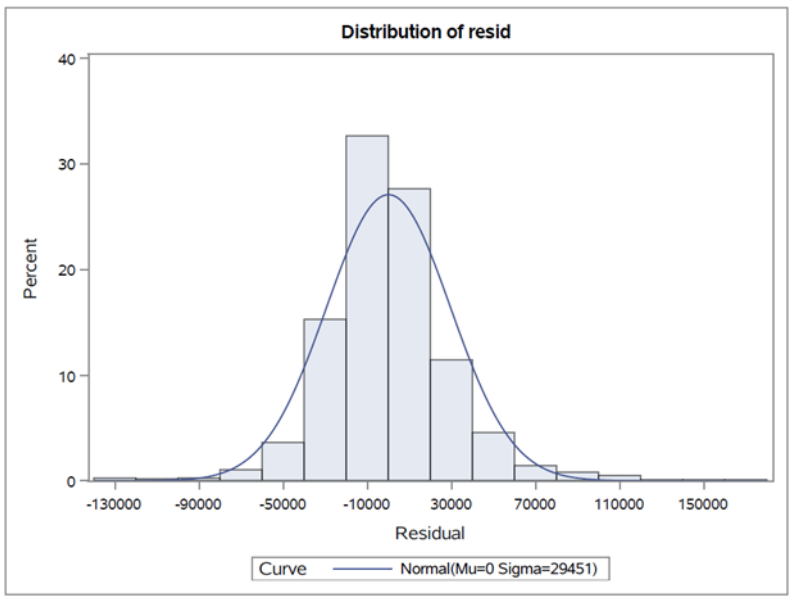
RESIDUAL ANALYSIS

Normality

Normal probability plot of the residual: This is a graph designed so that the cumulative normal distribution will plot as a straight line.



From this graph, we can access that the residuals are normally distributed. Basically, what you are looking for here is the data points closely following the straight line at a 45% angle upwards (left to right). We can see also that there is not a pattern or any fun shape.



The Residuals are normally distributed

Correlation Analysis

MULTICOLLINEARITY CHECK

In this session we start to explore whether or not our chosen version is suffering effects of multicollinearity. To do this we will begin analyzing the Pearson Correlation Coefficients Matrix.

Pearson Correlation Coefficients, N = 1320
Prob > |r| under H0: Rho=0

	SalePrice	LotArea	AgeOfHouse	TotalBsmtSF	GrLivArea	TotRmsAbvGrd	GarageArea
SalePrice	1.00000	0.34202 <.0001	-0.59472 <.0001	0.63754 <.0001	0.73966 <.0001	0.57259 <.0001	0.66187 <.0001
LotArea	0.34202 <.0001	1.00000	-0.03587 0.1928	0.27544 <.0001	0.34608 <.0001	0.32908 <.0001	0.27142 <.0001
AgeOfHouse	-0.59472 <.0001	-0.03587 0.1928	1.00000	-0.41673 <.0001	-0.29043 <.0001	-0.18347 <.0001	-0.51604 <.0001
TotalBsmtSF	0.63754 <.0001	0.27544 <.0001	-0.41673 <.0001	1.00000	0.36974 <.0001	0.23347 <.0001	0.47710 <.0001
GrLivArea	0.73966 <.0001	0.34608 <.0001	-0.29043 <.0001	0.36974 <.0001	1.00000	0.81368 <.0001	0.47138 <.0001
TotRmsAbvGrd	0.57259 <.0001	0.32908 <.0001	-0.18347 <.0001	0.23347 <.0001	0.81368 <.0001	1.00000	0.34796 <.0001
GarageArea	0.66187 <.0001	0.27142 <.0001	-0.51604 <.0001	0.47710 <.0001	0.47138 <.0001	0.34796 <.0001	1.00000

After reviewing these results, we can check if any of the variables included have a high correlation coefficient about 0.8 or higher – with any other variable. Reviewing this correlation matrix, there is only one variable highly correlate to other (Total Rooms Above Grade and Ground Living Area which is expected since a bigger house will have more rooms and bigger ground living area but because of this criteria the variable was left out our regression model) does not appear to be any variables with a particularly high correlation. Moving forward we will examine multicollinearity through the Variance Inflation Factor and Tolerance.

Parameter Estimates

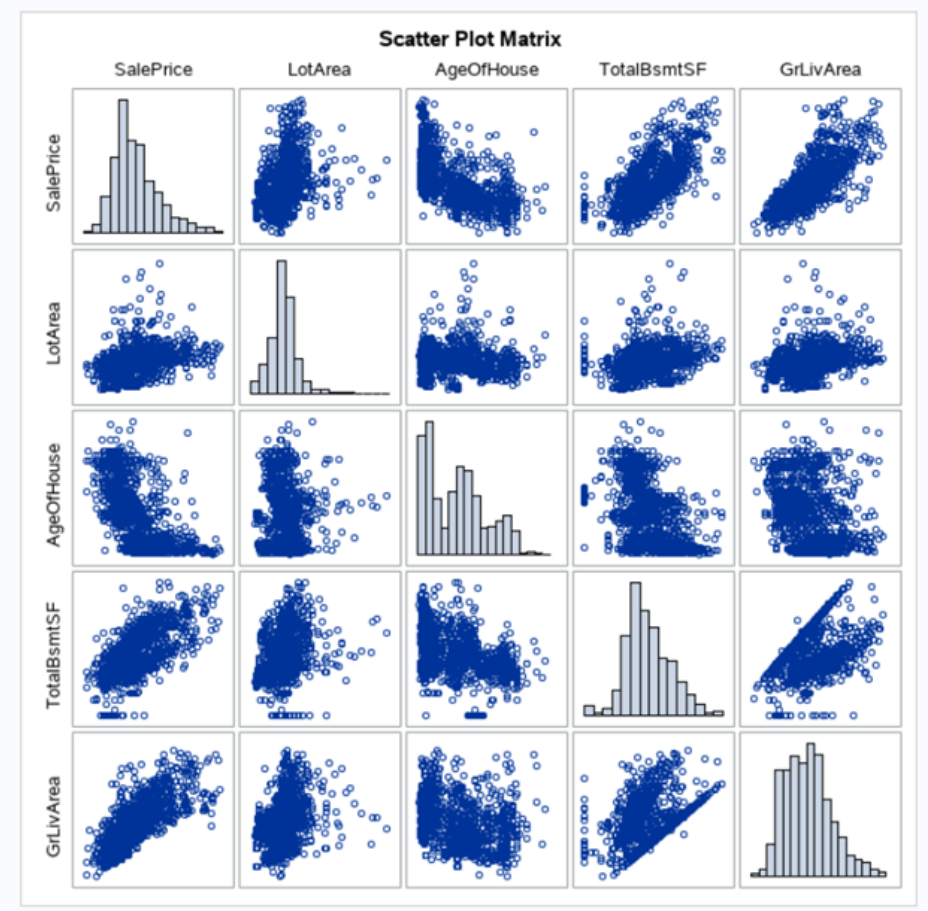
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	19853	4472.51397	4.44	<.0001	.	0
LotArea	1	0.83915	0.21713	3.86	0.0001	0.82199	1.21657
AgeOfHouse	1	-589.65441	34.05551	-17.31	<.0001	0.67260	1.48677
TotalBsmtSF	1	46.52642	2.66659	17.45	<.0001	0.68923	1.45089
GrLivArea	1	70.07575	2.19507	31.92	<.0001	0.70726	1.41391
GarageArea	1	55.39694	5.47222	10.12	<.0001	0.56956	1.75573

Correlation Analysis

MULTICOLLINEARITY CHECK

In reviewing tolerance, we need to make sure that no values fall below 0.1. In our output, the lowest tolerance value is 0.56956, so there is no threat of multicollinearity indicated through our tolerance values. For variance inflation, we should not have anything above the value of 10.

From the values indicated in this column, our highest value sits at 1.75573, indicating a lack of multicollinearity,
according to these results.



Limitations & Conclusion

DISCUSSION ABOUT HETEROSKEDASTICITY IN THE MODEL

It is more common to find heteroskedasticity in datasets that have a large range between the largest and smallest observed values. There are several reasons why heteroskedasticity happens, one possible explanation is that the error variance changes proportionally with a factor. This factor might be a variable that is included in the model

One of the assumptions made about residuals in Ordinary Least Squares regression is that the residuals have the same but unknown variance. The constant variance means homoscedasticity. When this assumption is violated leads to heteroskedasticity.

The violation of the homoscedasticity assumption (Heteroskedasticity) leads to the following consequences

Heteroskedasticity does not result in biased parameter estimation

- However, Ordinary Least Square estimates are no longer BLUE (Best Linear Unbiased Estimators). Hence, Ordinary Least Squares does not provide the estimate with the smallest variance. Depending on the nature of the heteroskedasticity, significance tests can be misleading (high or low).
- The standard errors are biased when heteroskedasticity is present. This, in turn, leads to bias in test statistics (t-test, F-test) and confidence intervals.
- Unless heteroskedasticity is too high significance tests are virtually unaffected, and thus OLS estimation can be used without concern of serious distortion. But, severe heteroskedasticity can sometimes be a huge problem.

As possible solutions to heteroskedasticity you can, use robust regression, standardized your variables, do variable transformations that could change the residuals among others. However, it is not easy to figure out what to do, sometimes, requires a lot of effort figuring out the variables that are warming your model, so in case of non-severe heteroskedasticity sometimes the decision
is doing nothing about the issue.

Limitations & Conclusion

ROBUST REGRESSION

All statistical software includes options with most routines for estimating robust standard errors, they are also to as Huber/White estimators or sandwich estimators of variance. As previously discussed, heteroskedasticity causes standard errors to be biased. Ordinary Least Squares assumes that errors are both independent and identically distributed; robust standard errors Heteroskedasticity relax these assumptions. When heteroskedasticity is present, robust standard errors tend to be more trustworthy.

The use of robust standard errors does not change coefficient estimates, but (because the standard errors are changed) the test statistics give us more accurate p values. The use of Weighted Least Squares also corrects the bias in the standard errors, hence giving more efficient estimates.

The Weighted Least Squares gives estimates that have the smallest possible standard errors. However, Weighted Least Squares needs more assumptions being difficult to implement, therefore, robust standard errors are a more used method for dealing with issues of heteroskedasticity.

To run the robust regression we used the software Stata, where robust standard errors are computed via the addition of two parameters, robust and cluster. The robust relaxes the assumption that the errors are equally distributed, while cluster relaxes the assumption that the error terms are independent of each other.

The output above is from a robust regression:

Linear regression

Number of obs	= 1,320
F(5, 1314)	= 757.61
Prob > F	= 0.0000
R-squared	= 0.7980
Root MSE	= 29507

saleprice	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lotarea	.8391475	.2476564	3.39	0.001	.3533022	1.324993
ageofhouse	-589.6544	33.86494	-17.41	0.000	-656.0897	-523.2192
totalbsmtsf	46.52642	3.303012	14.09	0.000	40.04667	53.00617
grlivarea	70.07575	2.47535	28.31	0.000	65.21968	74.93182
garagearea	55.39694	5.562677	9.96	0.000	44.48425	66.30964
_cons	19852.72	5081.091	3.91	0.000	9884.785	29820.66

Limitations & Conclusion

CONCLUSION

There could be an unknown variable that is confounding with one of the response variables. One lurking variable that could be confounding with an independent variable could be the furniture quality in the garage. If people with bigger garages have more money to get more expensive furniture for the garage. Then we can't tell whether the higher sale price of the property could be because of the bigger garage size or because the garages have expensive furniture.

The final sale price of a house could be influenced by several distinct factors, our model provides a good R2 0.7980, which means that 79.80% of the sale price is explained by our predictors. discussion).

Among our predictors, the one that mostly explains the sale price is Ground Living Area (in squared meters) that basically could be understood as the common use area of a house, the Pearson correlation coefficient between Ground Living Area and Sales Price is 0.73966 it is a fairly strong positive correlation.

There are, however, some limitations in this paper. Firstly, we did find a problem among heteroscedasticity that could influence in the goodness of our model, this could be from model misspecification or just of the fact that the dataset should've been standardized. (refer to heteroscedasticity session to a separated discussion about it)

Finally, the house price could be affected by some other externalities (such as exchange rate and interest rate) that are not included in the estimation so this would be omitted variables. Also, there are psychological aspects of the seller and the buyer that are practically impossible to be modeled that could impact the decision to buy, hence impacting the final sales price.

$$\widehat{saleprice} = 19852.72 + 0.8391475_{lotarea} - 589.6544_{ageofhouse} + 46.52642_{totalbsmtsf} + 70.07575_{grlivarea} + 55.39694_{garagearea}$$

Appendix

CITATIONS

[1] Zillow, "Zillow Prize", Zillow Promotions, 2018. [Online]. Available:https://www.zillow.com/promo/zillow-prize/. [Accessed: 02 -Apr-2020].

[2] L. Ackert, B. Church and N. Jayaraman, "Is There a Link Between Money Illusion and Homeowners' Expectations of Housing Prices?", RealEstate Economics, vol. 39, no. 2, pp. 251-275, 2011

[3] A. Ising, "Pompian, M. (2006): Behavioral Finance and Wealth Management – How to Build Optimal Portfolios That Account for Investor Biases", Financial Markets and Portfolio Management, vol. 21, no. 4, pp.491-492, 2007.

[4] Dean de Cock https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview[Accessed: 04-Apr-2020]