Quarterly Effect on Housing Market in USA

Multiple Regression with Dummy Variable and Contrast

Submitted By :

Disha Kapoor 200606410012
Mahak Saxena 20060641018
Mahima Vaidya 20060641019
Rutvij Joshi 20060641016

Introduction

Seasonality in the housing market depends on many factors and to be on the winning side of the seasonal trend one must know their local real estate market. Seasonality varies from location to location and every market has its own nuance. Some of the key factors that can affect seasonality in the housing market are school years, holidays and also weather.

This paper explains how to use multiple linear regression with dummy variables to identify the effect of seasonality in the housing market and take decisions based on the trends identified.

Problem Statement

Seasonality is a crucial factor to consider when taking decisions related to real estate. To determine the quarterly effect and observe patterns to predict future trends in the housing market is the aim of this project. Using this model can be beneficial for potential buyers and sellers in the housing market especially during this pandemic.

Data Collection

The data has been collected for 8 years from **redfin.com: Home Prices, Sales and Inventory.** The data is about Real Estate for the USA and there are various variables mentioned out of which **Region, Months and Houses Sold** have been used for our analysis. The region **"National"** has been selected and monthly data for houses sold has been used for the years 2012-2020 which amounts to 110 observations.

To show the effect of seasonality on real estate (house sales), average houses sold in each quarter were calculated for all years. Dummy Variables Q2, Q3, Q4 have been created and Q1 has been excluded to avoid the dummy variable trap where

Q2 = 1 if Month is {April, May, June} and 0 otherwise

Q3 = 1 if Month is {July, August, September} and 0 otherwise

Q4 = 1 if Month is {October, November, December} and 0 otherwise

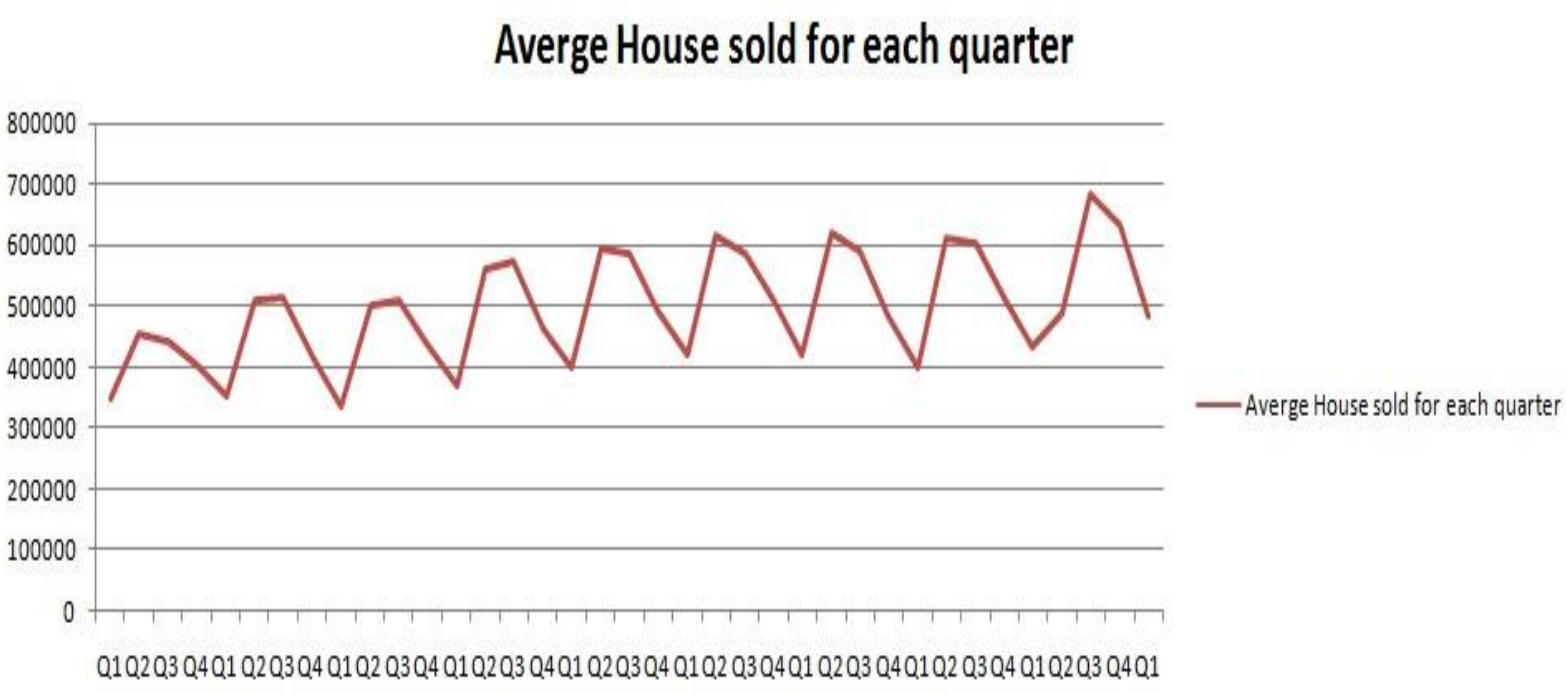
Region	Month of I	Homes Sold	Month	Quarter	Averge
National	Feb-12	306059	February	1	350919.5
National	Mar-12	395780	March	1	
National	Apr-12	406603	April	2	454049.7
National	May-12	465926	May	2	
National	Jun-12	489620	June	2	
National	Jul-12	452781	July	3	444613.7
National	Aug-12	487259	August	3	
National	Sep-12	393801	Septembe	3	
National	Oct-12	424766	October	4	405853
National	Nov-12	400096	November	4	
National	Dec-12	392697	December	4	
National	Jan-13	312816	January	1	352590.3
National	Feb-13	323301	February	1	
National	Mar-13	421654	March	1	
National	Apr-13	462105	April	2	509317

Source : https://www.redfin.com/news/data-center/

Year	Quarter	Averge House sold for each quarter	t	Q2	Q3	Q4
2012	1	350919.5	0	0	0	0
	2	454049.6667	1	1	0	0
	3	444613.6667	2	0	1	0
	4	405853	3	0	0	1
2013	1	352590.3333	4	0	0	0
	2	509317	5	1	0	0
	3	513818	6	0	1	0
	4	415390	7	0	0	1
2014	1	337159.3333	8	0	0	0
	2	500452.3333	9	1	0	0
	3	508863.3333	10	0	1	0
	4	436664.3333	11	0	0	1
2015	1	369670	12	0	0	0
	2	560437.6667	13	1	0	0
	3	573397.3333	14	0	1	0
	4	461585	15	0	0	1
2016	1	399453.6667	16	0	0	0
	2	595156	17	1	0	0
	3	585853.3333	18	0	1	0
	4	493932.6667	19	0	0	1
2017	1	422856.3333	20	0	0	0
	2	615730.3333	21	1	0	0
	3	587432	22	0	1	0
	4	509016	23	0	0	1
2018	1	423010.3333	24	0	0	0
	2	619039.6667	25	1	0	0
	3	590217.3333	26	0	1	0
	4	486777.6667	27	0	0	1

2019	1	402236.3333	28	0	0	0
	2	610384.6667	29	1	0	0
	3	602428.3333	30	0	1	0
	4	515037.6667	31	0	0	1
2020	1	433739.3333	32	0	0	0
	2	487255	33	1	0	0
	3	682926.3333	34	0	1	0
	4	633168.3333	35	0	0	1
2021	1	483723	36	0	0	0

Is there any seasonality?



By visualizing the data ,we can conclude that there is seasonality in the data series.

Regression with Dummy Variable

To capture both the seasonality and potential underlying trend in the data , we will rely on the regression analysis functionality using MS Excel.

Using this data, we will carry out the following:

- Perform regression analysis through Excel and interpret the results.
- Draw the fitted regression line on the scatter plot. Does the regression line appear to be a good fit here?
- Obtain the residual and draw the normal probability plot.
- Forecast 2021(Q_2,Q_3,Q_4) , 2022(Q_1,Q_2,Q_3,Q_4) , 2023 (Q_1) quarters

Model: $Y = \beta_0 + \beta_1 t + \beta_2 Q_2 + \beta_3 Q_3 + \beta_4 Q_4 + \varepsilon$

Where, Y = Average house sold for each quarter

Q_2 = Dummy variable of the 2nd quarter of the year; 1 for {April, May, June}and 0 for others.

Q_3 = Dummy variable of the 3rd quarter of the year; 1 for {July, August, September}and 0 for others.

Q_4 = Dummy variable of the 4th quarter of the year; 1 for {October, November, December}and 0 for others

t = Period t is the fourth variable in the regression model and will represent our time series
These dummies have been created to inspect the impact of quarterly effect on housing markets over the years.

To avoid Dummy Variable Trap Q_1 have not been included in our regression model . Therefore Q_2, Q_3, Q_4 are the dummy variables.

Data Interpretation and Analysis

There are four components for the regression output :

- Regression Statistical table
- ANOVA table
- Regression Coefficient table
- Residual Output

1. Interpretation of the Regression Statistics Table:

Figure 1

Regression Statistics	
Multiple r	0.9214
R square	0.84892
Adjusted R square	0.83019
Standard Error	37225.038

- R^2= 0.848 which means that 84.8% of the variation in Y is explained by the predictors X_i's i.e 84.8% of the variability in Average houses sold each quarter is accounted for by the regression model.
- The value of standard error here refers to the estimated standard deviation of the error term e i.e $\hat{\sigma} = 37225.038$.It is sometimes called the standard error of the regression.

2. Interpretation of the ANOVA Table:

Figure 2

ANOVA					
Source of Variation	df	SS	MS	F	Significance F
Regression	4	2.493E+11	62325010092	44.97716185	0.00
Residual	32	44342511643	1385703489		
Total	36	2.93643E+11			

- Testing the significance of the regression model
  The output given in figure 2 shows the ANOVA table for the Houses sold data.
  If we consider,
  H_0 : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
  H_1:Atleast one of the regression coefficient is not zero
  $\alpha = 0.05$

  Conclusion :From the ANOVA table given in Fig 2, g p-value is 0. Since , the p-value is l less than α.Thus we may conclude that there is a linear relationship between houses sold and the combination of quarters.

3. Interpretation of the Coefficient table

Figure 3

Regression Coefficient Table								
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	318400.6238	15667.40167	20.32	0.00	286487.1711	350314.0765	286487.1711	350314.0765
t	4396.399603	574.3948138	7.65	0.00	3226.39566	5566.403546	3226.39566	5566.403546
Q2	157063.0644	17113.36707	9.18	0.00	122204.2766	191921.8523	122204.2766	191921.8523
Q3	167969.7019	17103.72483	9.82	0.00	133130.5546	202808.8491	133130.5546	202808.8491
Q4	82226.08003	17113.36707	4.80	0.00	47367.29217	117084.8679	47367.29217	117084.8679

The regression output depicts the regression parameters and the associated output.

- Fitting the regression model
  From above figure , we have β_0=318400.6238, β_1 =4396.399603, β_2 = 157063.064 , β_3 =167969.7019, β_4 = 82226.08003

  Hence, the fitted regression equation is
  $\widehat{Y} = 318400.6238 + 4396.399603t + 157063.064Q2 + 167969.7019Q3 + 82226.08003Q4$
- Standard error of the regression coefficients
  The standard errors of β_0, β_1, β_2 ,β_3, β_4 are SE(β_0)=15667.40167, SE(β_1)=574.3948138, SE(β_2)=17113.36707 ,SE(β_3)=17103.72483,

SE(β_4)=17113.36707

- Hypothesis Testing for the intercept
  To test the hypothesis that the intercept is equal to zero i.e
  $H_0 : \beta_0 = 0$
  $H_1: \beta_0 \neq 0$ ie the line of regression is not passing through origin for given data.
  $\alpha = 0.05$
  Conclusion : We use a p-value approach for decision making . The p-value is 0. Since p-value is less than $\alpha = 0.05$. Thus we have sufficient evidence to reject H_0. Hence, the intercept is not equal to 0 i.e. the line of regression is not passing through origin.

- Hypothesis Testing for the slope
  $H_0 : \beta_1 = 0$
  $H_1: \beta_1 \neq 0$ ie time period affects the average houses sold for each quarter
  $\alpha = 0.05$
  Conclusion : We use a p-value approach for decision making . Since p-value is less than $\alpha = 0.05$. Thus we have sufficient evidence to reject H_0. Hence, time period affects the average houses sold for each quarter.

  $H_0 : \beta_2 = 0$
  $H_1: \beta_2 \neq 0$ ie Second Quarterly effect affects the average houses sold for each quarter
  $\alpha = 0.05$
  Conclusion : We use a p-value approach for decision making . Since p-value is less than $\alpha = 0.05$. Thus we have sufficient evidence to reject H_0. Hence, the second Quarterly effect affects the average houses sold for each quarter.

  $H_0 : \beta_3 = 0$
  $H_1: \beta_3 \neq 0$ ie Third Quarterly effect affects the average houses sold for each quarter
  $\alpha = 0.05$
  Conclusion : We use a p-value approach for decision making . Since p-value is less than $\alpha = 0.05$. Thus we have sufficient evidence to reject H_0. Hence, the third Quarterly effect affects the average houses sold for each quarter.

  $H_0 : \beta_4 = 0$
  $H_1: \beta_4 \neq 0$ ie Fourth Quarterly effect affects the average houses sold for each quarter
  $\alpha = 0.05$
  Conclusion : We use a p-value approach for decision making . Since p-value is less than $\alpha = 0.05$. Thus we have sufficient evidence to reject H_0. Hence, the fourth Quarterly effect affects the average houses sold for each quarter.

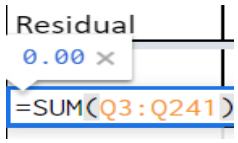
- Confidence intervals for regression coefficients
  We obtain the 95% confidence interval for intercept β_0 . It is (286487.1711,350314.0765).
  95% confidence interval for β_1 is (3226.39566,5566.403546)

95% confidence interval for β_2 is (122204.2766,191921.8523)
95% confidence interval for β_3 is (133130.5546,202808.8491)
95% confidence interval for β_4 is (47367.29217,117084.8679)

4. Interpretation of the Residual plot

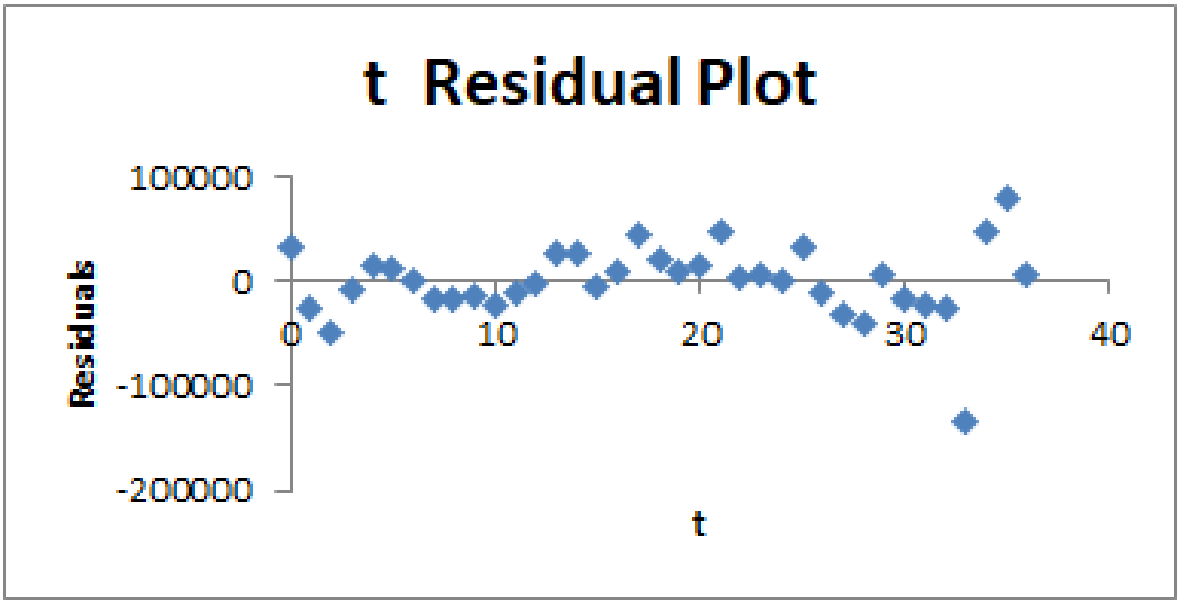
We can check the adequacy of the fitted regression model by making use of residuals .

- Verification of residual property : To verify the residual property that the sum of residuals is equal to zero.



Hence , the property $\sum_{i=1}^{n} e_i = 0$ is verified.

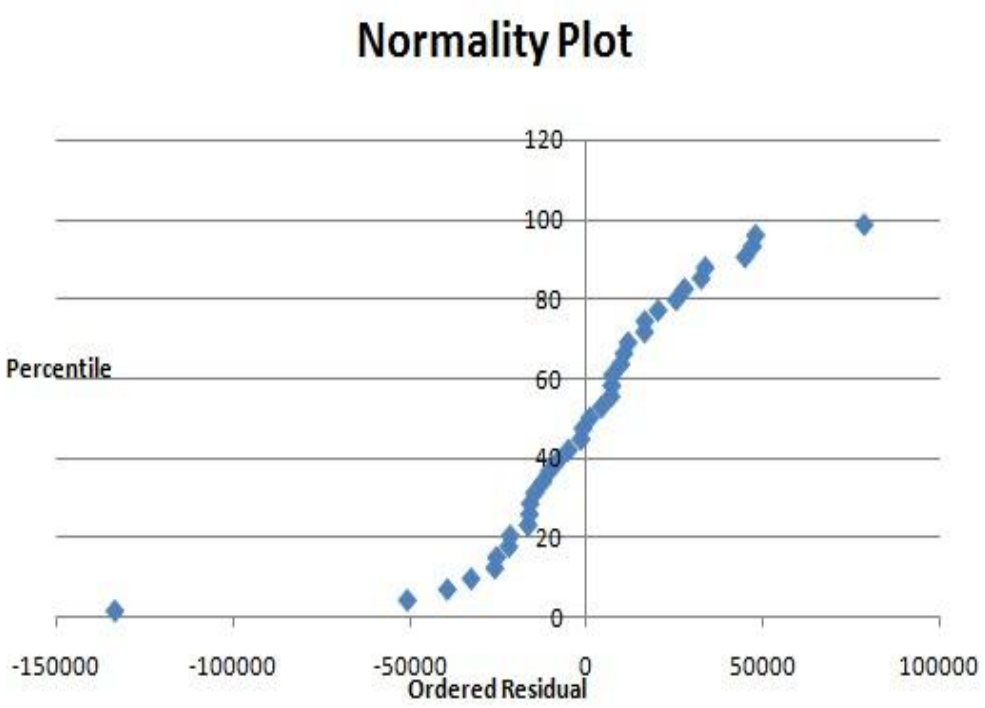
- Residual Plot : To conduct residual analysis , we plot the predicted sales vs standardised residuals .



Conclusion: The standard residuals appear to be fairly randomly scattered throughout a horizontal band around 0 (Y axis).Hence , the assumption of linear regression is valid or we can say that the Multiple with dummy variable regression model fitted well for the given data.

5. Interpretation of the Normal Probability Plot

We check the normality assumption as the t-test , F-test and confidence interval depend on the normality assumption of the residuals.



Conclusion: The normal probability plot reveals that the resulting points lie approximately along a straight line. It indicates that the distribution of error terms (residuals)is approximately normally distributed.

Regression with Contrast

A contrast is essentially a difference in regression coefficients. We have seen that the regression coefficients can express a difference in means or a single mean, as well as the slope and intercept of a line. A contrast is a way of testing more general hypotheses about population means.
Suppose we have *p* different populations (treatments) and we have measured the expression level of the gene in a sample from each population. The mean in population *j* is μ_j and the sample mean is $\overline{Y}_j$ .Suppose c1,c2,…,cp are numbers and the $\sum c_j=0$. $\sum c_j\mu_j$ is called a contrast of the population means. Notice that if all the means are equal, then any contrast is zero.
We can estimate the population contrast by plugging in the sample means:

$$c_1\overline{y}_1 + c_2\overline{y}_2 + c_3\overline{y}_3 + c_4\overline{y}_4$$

H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \boldsymbol{\mu}_4 \Leftrightarrow \beta_1 = \beta_2 = \beta_3 = 0$

H_1: At Least one $\beta_{i,\, i=1,2,3}$ is not equal to zero

SUMMARY OUTPUT

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	1.68121E+11	56040352196.28	14.7331867	0.00
Residual	33	1.25521E+11	3803681679		
Total	36	2.93643E+11			

Since the p-value of our Model for testing group means is 0 (less than 0.05) , we will reject H_0 i.e. all group means are not the same. Since there is evidence that not all population means are equal, we need to find out which group means differ. We will use the concept of contrast in order to find this.

Given :

	Q2	Q3	Q4	Q1
Mean	550202.4815	565505.5185	484158.2963	397535.817
ybar1	397535.8167			
ybar2-ybar1	152666.6648			
ybar3-ybar1	167969.7019			
ybar4-ybar1	86622.47963			

1. Model 1 : Quarter 2 vs Quarter 1

Model 1					
H0:	mu2-mu1=0				
H1:	mu2-mu1>0				
SUMMARY OUTPUT					
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	1.68121E+11	56040352196	14.7331867	0.00
Residual	33	1.25521E+11	3803681679		
Total	36	2.93643E+11			
coefficient=	(-1,1,0,0)				
c=	152666.6648				
sp=	61673.99516				
SEc=	28337.2452				
t=	5.3874914				
p-value=	0.00				
95% CI	95138.99867	210194.331			

Conclusion : For the 1st Model, we are testing the difference between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Since the p-value is less than 0.05, we reject H_0 and hence $\boldsymbol{\mu}_2 > \boldsymbol{\mu}_1$ i.e treatment 2 is greater than treatment 1.

2. Model 2 : Quarter 3 vs Quarter 1

Model 2					
H0:	mu3-mu1=0				
H1:	mu3-mu1>0				
SUMMARY OUTPUT					
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	1.68121E+11	56040352196	14.7331867	0.00
Residual	33	1.25521E+11	3803681679		
Total	36	2.93643E+11			
coefficient=	(-1,0,1,0)				
c=	167969.7019				
sp=	61673.99516				
SEc=	28337.2452				
t=	5.927524029				
p-value=	0.00				
95% CI	110442.0357	225497.368			

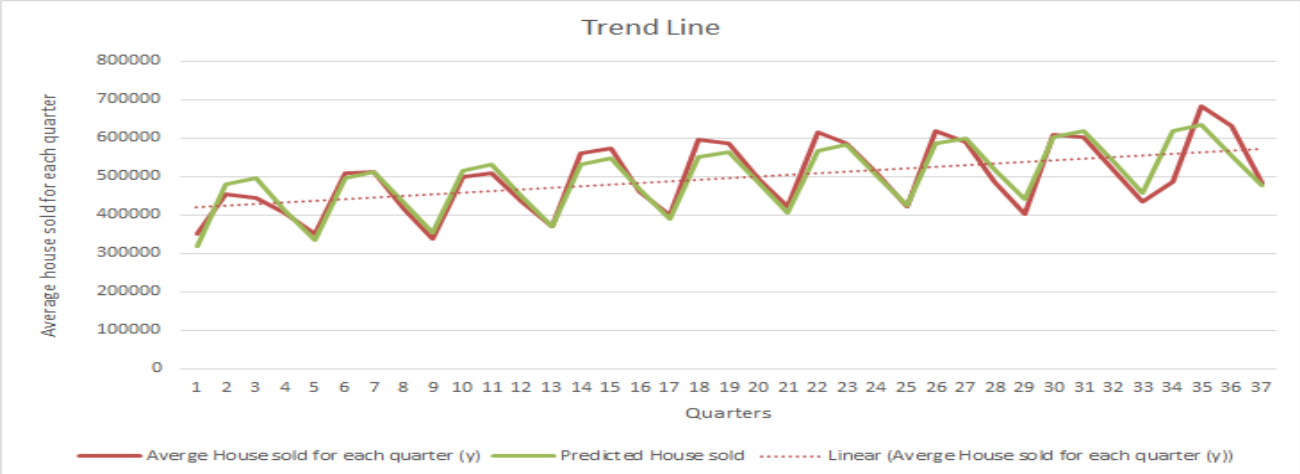
Conclusion : For the 2nd Model, we are testing the difference between μ_3 and μ_1. Since the p-value is less than 0.05, we reject H_0 and hence $\mu_3 > \mu$ i.e treatment 3 is greater than treatment 1.

3. Model 3 : Quarter 4 vs Quarter 1

Model 3					
H0:	mu4-mu1=0				
H1:	mu4-mu1>0				
SUMMARY OUTPUT					
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	1.68121E+11	56040352196.28	14.7331867	0.00
Residual	33	1.25521E+11	3803681679		
Total	36	2.93643E+11			
coefficient=	(-1,0,0,1)				
c=	86622.47963				
sp=	61673.99516				
SEc=	28337.2452				
t=	3.056841941				
p-value=	0.00				
95% CI	29094.81348	144150.1458			

Conclusion : For the 3rd Model, we are testing the difference between μ_4 and μ_1. Since the p-value is less than 0.05, we reject H_0 and hence $\mu_4 > \mu_1$ i.e treatment 4 is greater than treatment 1.

Fitted Regression lines



The fitted line of regression is given by

$\widehat{Y} = 318400.6 + 4396.4t + 157063.1Q2 + 167969.7Q3 + 82226.08Q4$.

Hence, the fitted regression line is a good fit.

Conclusion

- From the Regression table, the correlation between average houses sold quarterly is strongly-positively associated with the independent variables. We have R^2 of 0.848 and R^2adj is 83.02% which indicates that 83.02% of variation in Average Houses Sold is explained by the regression model.
- From the ANOVA table we conclude that the regression model is a good fit and there is a significant relationship between the independent variables and the dependent variable.
- From the analysis of intercept & slope we can conclude that the line of regression is not passing through the origin and all independent variables affect the dependent variable.
- From the residual plot, we see that the residuals are randomly scattered throughout a horizontal band around 0 i.e. Y-Axis. Hence, the regression model is well fitted.
- From the Normal probability plot, we see the points lie approximately along the straight line which indicates that the errors are normally distributed.

Recommendation

- It can be observed that the 2nd and 3rd quarter of the year have the highest average sale of houses in the year. Ideally the housing market seasonality can be said to be a full-proof trend and from the regression model and observation of trend it can be said that potential sellers must consider it a good time to sell houses during the 2nd and 3rd quarter i.e from April to September.
- Similarly we can say that since the 2nd and 3rd quarters of the year is the busiest time for sales and most people are buying houses during this time, buyers can find raised prices and more competition across the board. But in order to optimise their purchase of the house the Winter season can be considered as a good time to buy since the demand is less in the 1st and 4th Quarter of the year.
- With respect to the market situation in the pandemic, the trend is still showing an increase in the house sales in 2020. The sales are higher in 2020 as compared to 2019 and follow the same seasonal pattern as the other years. Hence we can say that the real estate boom is far from over

and that 2021 might witness a similar pattern as compared to other years. This information can be used to take decisions regarding the buying or selling of houses.

The above graph represents the forecast of 2021(Q_2,Q_3,Q_4) , 2022(Q_1,Q_2,Q_3,Q_4) , 2023 (Q_1) quarters