**Abstract**

The purpose of this research study is to compare the average sales in the housing prices in four US regions. Using the not seasonally adjusted quarterly data and all transaction home sales prices from Federal Reserve Economic Data(FRED), I adjusted for the effect of inflation using quarter 1 2010 dollars as the base year. In addition, through a variety of methods including, calculation of the descriptive statistics, time series analysis of the dataset for each region under consideration and performed an analysis of variance to find answers to study hypothesis, that the average sales price of houses over the last 20 years is the same across the four US regions.

This study also performs Bonferroni and TukeyHSD post-hoc test to find out where the difference in mean lies.

The result of the study showed that there is a significant difference in the average price of houses sold across the four regions with the difference lying between the Midwest and South region.

1. **Introduction:**

Housing decisions are one of the most important investments decisions individuals make in their lifetime. The housing market can also have a significant impact on the broader economy, as evidenced by the housing bubble that caused the 2007-2009 recession. In two ways, the housing market is incorporated into gross domestic product (GDP), the most important measure of economic activity. Expenditures on the construction of new structures, residential remodeling, and brokers’ fees, referred to as residential fixed investment are incorporated in GDP. “As of 2020, spending on residential fixed investment was about $885 billion, accounting for about 4.2% of GDP. Second, GDP includes all spending on housing services, which includes renters’ rents and utilities and homeowners’ imputed rent and utility payments. It is reported by the Congressional Research Service that in 2020, spending on housing services was about $2.8trillion, which was about 13.3% of GDP. In sum, spending within the housing market accounted for 17.5% of GDP in 2020”. (Congressional Research Service, 2021).

Over the years, the housing price in America has gone through various changes from the peak price in the year 2000 to early 2006, to the recession in 2007-2009, the housing market plays a significant role in the U.S. economy. As of 2005, when the housing bubble began to burst, the average home sales price was about $283,000. According to reports by Congressional Research Service, at the individual level, roughly 65% of occupied housing units are owner occupied, homes are often a substantial source of household wealth in the United States, and housing construction provides widespread employment. At the aggregate level, housing accounts for a significant portion of all economic activity, and changes in the housing market can have broader effects on the economy. This paper emphasizes both cross-sectional variation of house prices sold across US census regions and time series variation.

**1.1 Problem Description**

For each of the regions, this study paper would like to identify the if there is a difference between the average sales price of houses sold between four US regions: Midwest, South, West, and Northeast region.

The study wishes to confirm if there is a difference, then where does the difference in mean lie between the four regions under consideration.

1. **Literature Review:**

Previous research on housing prices have highlighted demographics, government policies and income trends as main drivers of housing price. Markets in particular geographical areas have exhibited still greater fickleness (cf. Himmelberg et al, 2005). These sustained changes have been said to be as a result of speculative bubbles, at least in specific localities (Case and Shiller, 2003), expansions and contractions in monetary policy (Iacoviello and Neri, 2006), and developments in financial markets-innovations such as new types of mortgage instruments, or breakdowns such as the recent subprime mortgage boom and bust. Such explanations suggest that market irrationality may have played an important role in housing sector fluctuations. According to Khan James, 2008 there is some indication that the increase in housing wealth does not exactly stem from an increase in the value of houses per se, but from the increase in the value of the land upon which they are built.

An index kept by Federal Housing Finance Agency has indicated that the prices of homes fell at an annual rate of 12.4 percent, from October 2010 to January while the government’s calculation of owners’ equivalent rent shows it increased at a yearly rate of 1.5 percent”- Fang Block, Realtor.com, Aug 2020.

There is the ideology that inflation rates might have been understated when it seemed home prices were rising. This might have been because of rising speculations-.- Floyd Norris comments on finance and the economy in his blog at nytimes.com/norris.

1. **Methodology:**

Data for this study was collected from Federal Reserve Economic Data (FRED). The time series data collected was quarterly data for average sales price of houses in four US census regions (Mid-west, South, West, and North-east) for the past 20 years; Year 2000-Year 2020. The year 2021 was excluded from this data. This covers 84 quarters of data.

The data used in this paper may be downloaded from

<https://fred.stlouisfed.org/series/ASPMW>

<https://fred.stlouisfed.org/series/ASPS>

<https://fred.stlouisfed.org/series/ASPW>

<https://fred.stlouisfed.org/series/ASPNE>

Accessed November 10th, 2021.

**3.1 Data Structuring:**

The initial dataset contained the average sales price (in dollars) for the four US regions, collected quarterly and not seasonally adjusted.

In order to clean up the data and adjust for inflation over the years, house price index was collected from Federal Reserve Economic Data (FRED) and adjusted in 2010 quarter 1 US dollars for each region.

USSTHPI Quarterly Data is collected from <https://fred.stlouisfed.org/series/USSTHPI#0>

Data collected on Nov 10th, 2021.

**3.1.2 Data Analysis:**

Seasonality is a characteristic of a time series data in which the data shows an anticipated trend each year. In order to adequately compare the average sales price of houses over the study years between these regions, a time series analysis was performed in order to take out the seasonality in each dataset.

The time series plot as depicted below showed oscillations in the sales prices for all the regions.

Graphical user interface, chart, line chart

Description automatically generated

*Figure 1: Average Sales price of houses sold for Midwest Census region.*

*Figure 2: Adjusted Average Sales Price of Houses sold in Midwest census region. (In 2010 Quarter 1US dollars)*



*Figure 3: Average Sales Price of Houses sold in Four Census Regions. (In 2010 Quarter 1US dollars), Year 2000-2020*

**.**



*Figure 4:Inflation Adjusted Average Sales Price of Houses sold in Four Census Regions. (In 2010 Quarter 1US dollars), Year 2000-2020.*

**3.2 Results and Discussions**

**3.2.1 Descriptive Statistics:**

A descriptive statistic was carried out to summarize the raw data. The mean, variance, and standard deviation for each sample region was found. The purpose of this is to highlight the potential relationships between the variables.

| Regions → | MW | S | W | NE | Pooled Total |
| --- | --- | --- | --- | --- | --- |
| observations N | 84 | 84 | 84 | 84 | 336 |
| sum ∑xi | 21,180.6800 | 21,090.6200 | 29,098.6900 | 35,655.0400 | 107,025.0300 |
| mean ¯xx¯ | 252.1510 | 251.0788 | 346.4130 | 424.4648 | 318.5269 |
| sum of squares ∑x2*i* | 5,394,418.9890 | 5,343,634.3134 | 10,130,920.4051 | 15,440,476.8178 | 36,309,450.5253 |
| sample variance s2 | 646.8717 | 581.0428 | 611.5253 | 3,688.7802 | 6,624.1856 |
| sample std. dev. Ss | 25.4337 | 24.1048 | 24.7290 | 60.7353 | 81.3891 |
| std. dev. of mean SE¯xSEx¯ | 2.7750 | 2.6301 | 2.6982 | 6.6268 | 4.4401 |

*Table 4: Summary descriptive statistics*

**3.2.2 Moving Average:**

**Simple Moving Averages**

The moving averages of the various regional prices was analyzed in order to understand the pattern of price movement across the quarters from year to year.

The 4-period moving average was seen to have the least mean average deviation (MAD) in the Mid-west region. Showing that it adequately tracked the price changes over the quarters each year under consideration.

In a twist, the 2-period moving average had the least MAD in the South and West region. In the Northeast, the MAD was quite high, the 3-period moving average had the least MAD.

**3.2.3 Weighted Moving Averages**

The weighted moving average was also carried out as it is understood to track prices more closely by giving more weight to recent data and less on past data.

For this study, the weights (0.5, 0.3, 0.2), (0.4, 0.4, 02), and (0.7, 0.2, 0.1) was used to compare the regions.

All regions had the least MAD with weight (0.5, 0.3, 0.2), except for the North-east region where the weight (0.7, 0.2, 0.1) had the least MAD when compared with the other regions.

**3.2.4 Time series Analysis:**

A time series analysis was carried out based on the inflation adjusted average sales prices for each of the regions in order to understand price oscillation and variation over time.

Using the multiplicative decomposition model, a 4-period moving average was calculated to capture the four quarters of each year represented in the dataset.

*yt = Tt* x *St* x *Rt*

Where:

*yt*= The time series value at period *t*

*Tt* = The trend component at period *t*

*St* = The seasonal component at period *t*

*Rt* = The random component at period *t*

The purpose of calculating a four-period centered moving average is to remove the seasonal component, *St* , from the time series.

Averaging excludes the random component *Rt* , because the random fluctuations are expected to average out over time .

With seasonality and randomness removed, the centered moving average values represent trend (*Tt*)

*CMA* = *Tt*

The ratio-to-moving-averagerepresents the seasonal and random components in the original time series and was found by dividing the time series value by the centered moving average calculated.

*RMAt = St* x *Rt* x *yt* = *yt*

*Tt CMA*

To deseasonalize each of the time series, the seasonal component was used to take out the seasonality.

*yt* = *Tt* x *Rt*

*St*



*Figure 3: Comparison between adjusted average sales price and deseasonalized sale for*

*Midwest census region.*



*Figure 4: Comparison between adjusted average sales price and deseasonalized sale for*

*West census region.*



*Figure 5: Comparison between adjusted average sales price and deseasonalized*

*sale for West census region.*



*Figure 6: Comparison between adjusted average sales price and deseasonalized*

*sale for North-east census region.*

**3.2.5 Dummy Regression:**

Multiple regression allows the inclusion of a very useful technique to classify the data in an attempt to build a more reliable model and accurate prediction model.

The data was classified by using dummy variable. A dummy variable takes the value of 0 or 1.

With four seasons, Quarter 1(January) was used as the base year dummy; n-1. Three dummies were used to run the regression.

April Dummy, Denoted by AprD= 1 if APrD, otherwise, 0.

July Dummy, denoted by JulD = 1 if JulD, otherwise 0.

October Dummy, denoted by OctD = 1 if OctD, otherwise 0.

*Yt = β0 + β1(Trend)+ β2(Yt - 1)+ β3(DApr)+β4(DJul)+β5(DOct)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.866 |  |  |  |  |  |
| R Square | 0.750 |  |  |  |  |  |
| Adjusted R Square | 0.734 |  |  |  |  |  |
| Standard Error | 13.166 |  |  |  |  |  |
| Observations | 83 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 5 | 40096.814 | 8019.363 | 46.26452 | 7.88551E-22 |  |
| Residual | 77 | 13346.964 | 173.3372 |  |  |  |
| Total | 82 | 53443.777 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 31.978 | 15.185 | 2.106 | 0.038466 | 1.742 | 62.215 |
| Trend | 0.035 | 0.063 | 0.555 | 0.580404 | -0.090 | 0.160 |
| Adj Sales(-1) | 0.857 | 0.060 | 14.262 | 2.6E-23 | 0.737 | 0.977 |
| AprD | -1.091 | 4.116 | -0.265 | 0.791694 | -9.288 | 7.106 |
| JulD | 0.286 | 4.133 | 0.069 | 0.945006 | -7.945 | 8.517 |
| OctD | 9.735 | 4.150 | 2.345 | 0.021581 | 1.470 | 17.999 |

*Yt = 31.978 + 0.035(trend) + 0.857(AdjSales-1) -1.091(AprD)+0.286(JulD) + 9.735(OctD)*

*Table 1: Dummy variable for Midwest region.*

The overall model is significant.

The Adjusted R square is 73.4%. This implies that is 73.4% variation in the dependent variable (Adjusted Sales for inflation) is explained by the data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.910 |  |  |  |  |  |
| R Square | 0.829 |  |  |  |  |  |
| Adjusted R Square | 0.817 |  |  |  |  |  |
| Standard Error | 10.362 |  |  |  |  |  |
| Observations | 83 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 5 | 39957.869 | 7991.574 | 74.429 | 0.000 |  |
| Residual | 77 | 8267.679 | 107.372 |  |  |  |
| Total | 82 | 48225.548 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 31.209 | 13.552 | 2.303 | 0.024 | 4.224 | 58.194 |
| Trend | 0.080 | 0.058 | 1.376 | 0.173 | -0.036 | 0.197 |
| Adj Sales(-1) | 0.861 | 0.058 | 14.816 | 0.000 | 0.745 | 0.977 |
| AprD | -0.806 | 3.238 | -0.249 | 0.804 | -7.253 | 5.642 |
| JulD | -0.773 | 3.239 | -0.239 | 0.812 | -7.223 | 5.676 |
| OctD | 2.496 | 3.243 | 0.770 | 0.444 | -3.962 | 8.953 |
| *Yt = 31.209 + 0.080(trend) + 0.861(AdjSales-1) - 0.806(AprD) - 0.773(JulD) + 2.496(OctD)* | | | | | | |

*Table 2: Dummy variable for South census region.*

The overall model is significant. The Adjusted R square is 81.7%. This implies that is 81.7% variation in the dependent variable (Adjusted Sales for inflation) is explained by the data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.836 |  |  |  |  |  |
| R Square | 0.698 |  |  |  |  |  |
| Adjusted R Square | 0.679 |  |  |  |  |  |
| Standard Error | 14.088 |  |  |  |  |  |
| Observations | 83 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 5 | 35376.099 | 7075.220 | 35.647 | 1.02596E-18 |  |
| Residual | 77 | 15282.842 | 198.478 |  |  |  |
| Total | 82 | 50658.940 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 54.832 | 23.183 | 2.365 | 0.021 | 8.668 | 100.996 |
| Trend | 0.044 | 0.070 | 0.632 | 0.529 | -0.095 | 0.183 |
| Adj Sales(-1) | 0.820 | 0.068 | 12.041 | 0.000 | 0.684 | 0.955 |
| AprD | 8.660 | 4.425 | 1.957 | 0.054 | -0.151 | 17.472 |
| JulD | 1.931 | 4.409 | 0.438 | 0.663 | -6.849 | 10.710 |
| OctD | 11.879 | 4.437 | 2.677 | 0.009 | 3.044 | 20.713 |
| *Yt = 54.832 + 0.044(trend) +0.820(AdjSales-1) +8.660(AprD) + 1.931(JulD) +11.879(OctD)* | | | | | | |

*Table 3: Dummy variable for West census region.*

The overall model id significant. The Adjusted R square is 67.9%. This implies that 67.9% variation in the dependent variable (Adjusted Sales for inflation) is explained by the data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.716 |  |  |  |  |  |
| R Square | 0.512 |  |  |  |  |  |
| Adjusted R Square | 0.481 |  |  |  |  |  |
| Standard Error | 43.886 |  |  |  |  |  |
| Observations | 83 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 5 | 155766.0202 | 31153.2 | 16.175 | 7.0772E-11 |  |
| Residual | 77 | 148299.708 | 1925.97 |  |  |  |
| Total | 82 | 304065.7282 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 252.453 | 41.910 | 6.024 | 5.466E-08 | 169.000 | 335.906 |
| Trend | 1.216 | 0.279 | 4.354 | 4.070E-05 | 0.660 | 1.772 |
| Adj Sales(-1) | 0.278 | 0.111 | 2.501 | 1.451E-02 | 0.057 | 0.500 |
| AprD | -10.030 | 13.743 | -0.730 | 4.677E-01 | -37.395 | 17.335 |
| JulD | 11.900 | 13.921 | 0.855 | 3.953E-01 | -15.820 | 39.621 |
| Oct D | 6.773 | 13.719 | 0.494 | 6.229E-01 | -20.545 | 34.090 |
| *Yt = 31.978 + 0.035(trend) + 0.857(AdjSales-1) -1.091(AprD)+0.286(JulD) + 9.735(OctD)* | | | | | | |

*Table 4: Dummy variable for North-east census region.*

The overall model is significant. The Adjusted R square is 48.1%. This implies that is 48.1% variation in the dependent variable (Adjusted Sales for inflation) is explained by the data.

The result of the dummy variables carried out for each of the regions showed that trend was significant only in the North-east region. The rest of the regions showed no significance in trend.

Using Quarter 1 as the dummy for the set,

In the Midwest and West regions, only October dummy was significant. In the South region and North-east regions, all the dummies were not significant.

**3.3 Hypothesis Testing:**

The α-level for this study is benched at 0.05. In order to test if there is a significant difference between the means of the different regions, t.test was carried out, and f.test was carried out to test for the equality of the population variance.

*Ho: µMW-µS =0*

*H1: µMW-µS0*

P-value *0.7795 > 0.05*. We fail to reject the null hypothesis that the average house sales price between Midwest and South Region are equal.

*Ho: µMW-µW =0*

*HMW: µA-µW0*

Because the *p-value 2.2e-16 < 0.05*, The average price sales of houses in the two regions is statistically significant. Therefore, I reject the null hypothesis that the average house price sales between the Midwest and West region is equal.

*Ho*: *µMW-µNE =0*

*H1*: *µMW-µNE0*

Because the p-value *2.2e-16 < 0.05*, the average price sales of houses in the two regions is statistically significant. Therefore, I reject the null hypothesis that the average house price sales between the Midwest and Northeast region is equal.

*Ho: µS-µW =0*

*H1: µS-µW0*

Because the *p-value* *2.2e-16 < 0.05*, the average price sales of houses in the two regions is statistically significant. Therefore, I reject the null hypothesis that the average house price sales between the South and West region is equal.

*Ho: µS-µNE =0*

*H1: µS-µNE0*

Because the p-value *2.2e-16 < 0.05*, the average price sales of houses in the two regions is statistically significant. Therefore, I reject the null hypothesis that the average house price sales between the South and Northeast region is equal.

*Ho: µW-µNE =0*

*H1: µW-µNE0*

Because the *p-value 2.2e-16 < 0.05*, the average price sales of houses in the two regions is statistically significant. Therefore, I reject the null hypothesis that the average house price sales between the West and Northeast region is equal.

**Test for Variance**

The variance is the average of the squared differences from the mean. It is a measurement of the spread between the numbers in the dataset.

Testing the hypothesis that the variance of **Midwest region and South region** are equal.

*H0*: =

*H1*: ≠

Because *p-value =* *0.626 > 0.05*. Therefore, we fail to reject the null hypothesis that the variance of average price sales between Midwest region and South region are equal. In other words, there is no significant difference between the two variances.

Testing the hypothesis that the variance of **Midwest region and West region** are equal.

*H0*: =

*H1*: ≠

Because *p-value =* *0.7986 > 0.05*. Therefore, we fail to reject the null hypothesis that the variance of average price sales between Midwest region and West region are equal. In other words, there is no significant difference between the two variances.

Testing the hypothesis that the variance of **Midwest region and Northeast region** are equal.

*H0*: =

*H1*: ≠

Because *p-value =* *7.489e-14 < 0.05*. Therefore, we reject the null hypothesis that the variance of average price sales between Midwest region and Northeast region are equal. In other words, there is a significant difference between the two variances.

Testing the hypothesis that the variance of **South region and West region** are equal.

*Ho*: =

*H1*: ≠

Because *p-value* *=* *0.8164 > 0.05*. Therefore, we fail to reject the null hypothesis that the variance of average price sales between South region and West region are equal. Therefore, we conclude that there is no significant difference between the two variances.

Test the hypothesis that the variance of **South region and Northeast region** are equal.

*Ho*: =

*H1*: ≠

Because *p-value = 2.995e-15 < 0.05*. Therefore, we reject the null hypothesis that the variance of average price sales between South region and Northeast region are equal. In other words, there is a significant difference between the two variances.

Test the hypothesis that the variance of **West region and Northeast region** are equal.

*Ho*: =

*H1*: ≠

Because *p-value = 1.408e-14 < 0.05*. Therefore, reject the null hypothesis that the variance of average price sales between West region and Northeast region are equal. In other words, there is a significant difference between the two variances.

**3.3.1 Analysis of Variance:**

*H0: µMW=µS = µW = µNE*

*H1: At least one mean is different from the others*

The null hypothesis is that there is not significant difference in the average sales of house price between the four regions.

If we do not reject the null hypothesis, no further analysis is required.

If the null hypothesis is rejected, then further analysis is required to help identify the group with a different mean from the four groups under analysis.

***Homoscedasticity:*** The assumption of equal variances. I assumed that the different samples have the same variance, even if they came from different populations.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Anova: Single Factor | |  |  |  |  |  |
|  |  |  |  |  |  |  |
| SUMMARY |  |  |  |  |  |  |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |  |  |
| MW | 84 | 21180.61 | 252.150 | 646.890 |  |  |
| S | 84 | 21090.61 | 251.079 | 581.044 |  |  |
| W | 84 | 29098.72 | 346.413 | 611.510 |  |  |
| NE | 84 | 35655.07 | 424.465 | 3688.762 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *Df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 1760278 | 3 | 586759.4 | 424.557 | 3E-113 | 2.632 |
| Within Groups | 458841 | 332 | 1382.051 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 2219119 | 335 |  |  |  |  |

*Table 5:Analysis of Variance*

Where SS: Sum of squares

SS between groups: 1760278, quantifies the variability between groups of interest, the regions.

SS within groups :458841, quantifies the variability within the groups of interest.

df between: Number of groups in the analysis less 1.

Df within: Number of observations in the analysis – number of groups.

MS: Mean square. This is the average variation either between or within the groups.

*MS = SS/df*

F=Ratio of the MS between groups to the MS within groups.

*F statistic = 424.557 > F critical= 2.632*

P:value 3.044E-113 < the alpha level, α shows that the there is a significant difference between the groups.

Because F statistic *= 424.557* > F critical *= 2.632,* we can reject the null hypothesis that the means are the same and conclude that there is a statistical difference in average house sales prices between the Midwest, South, West, and Northeast regions.

**3.3.1.1 Growth rate of Prices:**

Present prices are quite likely dependent on past prices in time-series data. That is, they are autocorrelated; as a result, lag values of the trend must be included as explanatory variables.

Is it possible that the growth rate of housing prices differs in various regions, or are the housing price growing at the same rate across the regions?

In order to do this, this study will be treating the data from the regions as a cross-sectional model.

To do this, an ANOVA was carried out on the percentage growth rate of the data. This was carried out to compare how they relate over time with present price.

*H0: µMW=µS = µW = µNE*

*H1: At least one mean is different from the others*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Anova: Single Factor | |  |  |  |  |  |
|  |  |  |  |  |  |  |
| SUMMARY |  |  |  |  |  |  |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |  |  |
| MW Growth rate | 83 | -0.400 | -0.005 | 33.016 |  |  |
| S Growth rate | 83 | 8.256 | 0.099 | 17.599 |  |  |
| W Growth rate | 83 | 6.174 | 0.074 | 19.982 |  |  |
| NE Growth rate | 83 | 70.710 | 0.852 | 139.722 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 39.893 | 3 | 13.298 | 0.253 | 0.859 | 2.632 |
| Within Groups | 17246.166 | 328 | 52.580 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 17286.060 | 331 |  |  |  |  |

*Table 6:Growth rate Analysis of Variance*

The result of the ANOVA carried out on the growth rate for the regions showed that there was no statistically significant difference in the growth rate of house prices across the regions. Therefore, we fail to reject the null hypothesis and conclude that the house prices across regions are growing at the same rate. In other words, the analysis suggests that there is no strong evidence to support the claim that at least one of the regions has a significantly different growth rate in house prices compared to the others.

The average sales of houses sold might differ in various regions based on geographical locations and other factors, however, the prices are growing at the same rate.

The p-value corresponding to the F-statistic of one-way ANOVA is lower than 0.05, suggesting that the one or more treatments are significantly different. The ANOVA test cannot tell which groups are significantly different from the other. In order to investigate this further, this study tries to find out where the significant differences in mean lie. That is, which specific regions are significantly different from the other.

**3.3.1.2 Post-hoc Tests**

The Bonferroni and Tukey HSD multiple comparison tests follow. These post-hoc tests would likely identify which of the pairs of mean prices are significantly different from each other.

To do this, a separate t-test was performed and then controlled for multiple comparison between the regions, using a post-hoc test via the Bonferroni correction.

The α-level of the one-way ANOVA was 0.05. This means we could conclude that the test was significant if p-value ≤ 0.05. In carrying out the Bonferroni correction, α-level is adjusted to account for the multiple hypothesis being performed and correct type-I error.

Bonferroni corrected α = α/k

Where k = Number of post-hoc test

In this case, we are comparing the means of four U.S. regions, so there are six possible pairwise comparisons to compare every region to every other region. These comparisons are:

Region 1 vs. Region 2 (MW vs S)

Region 1 vs. Region 3 (MW vs W)

Region 1 vs. Region 4 (MW vs NE)

Region 2 vs. Region 3 (S vs W)

Region 2 vs. Region 4 (S vs NE)

Region 3 vs. Region 4 (W vs NE)

= 0.05 = 0.008

6

|  |  |
| --- | --- |
| ALPHA |  |
| *Test* | *Alpha* |
| ANOVA | 0.05 |
| Post-hoc test(Bonferroni corrected) | 0.008 |
|  |  |

*Table 6: Bonferroni Correction α-level.*

If p-value ≤ 0.008, test is significant

If p-value > 0.008, test is not significant.

|  |  |  |
| --- | --- | --- |
| POST-HOC TEST |  |  |
| *Groups* | *P-value* | *Significance* |
| MW & S | 0.7796 | No |
| MW & W | 1.1147E-56 | Yes |
| MW & NE | 8.0637E-56 | Yes |
| S & W | 7.5103E-59 | Yes |
| S & NE | 1.3424E-56 | Yes |
| W & NE | 3.1199E-21 | Yes |

*Table 7: Post-hoc test: Bonferroni Test*

The post-hoc test confirms that there is a significant difference in the average sales price of houses between the four US regions. The significant difference in mean lies between the Midwest and West region, Midwest and Northeast region, South and West region, South and Northeast region, and West and Northeast region as confirmed by the post-hoc test.

**3.3.1.3 TukeyHSD Test:**

The p-value corresponding to the F-statistic of one-way ANOVA is lower than 0.01 which strongly suggests that one or more pairs of treatments are significantly different. There are k=4 treatments, for which the Tukey's HSD test will be applied to each of the 6 pairs to find which of them exhibits statistically significant difference.

We first establish the critical value of the Tukey-Kramer HSD QQ statistic based on the k=4 treatments and ν =332 degrees of freedom for the error term, for significance level

 α= 0.01 and 0.05 (p-values) in the Studentized Range distribution.

The critical values for QQ, for αα of 0.01 and 0.05, are obtained as Qα=0.01,k=4,ν=332 critical α=0.01,k=4,df=332 = 4.4365 and Qα=0.05, k=4, df=332. criticalα=0.05,k=4,ν=332 = 3.6518, respectively. These critical values may be verified at several published tables of the inverse Studentized Range distribution.

| Groups pair | Tukey HSD Q statistic | Tukey HSD p-value | Tukey HSD inference |
| --- | --- | --- | --- |
| MW vs S | 0.2643 | 0.8999947 | Insignificant |
| MW vs W | 23.2388 | 0.0010053 | \*\* p<0.01 |
| MW vs NE | 42.4812 | 0.0010053 | \*\* p<0.01 |
| S vs W | 23.5031 | 0.0010053 | \*\* p<0.01 |
| S vs NE | 42.7455 | 0.0010053 | \*\* p<0.01 |
| W vs NE | 19.2424 | 0.0010053 | \*\* p<0.01 |

*Table 8: TukeyHSD Test*

The TukeyHSD confirms that there is a difference in mean in the population means. Therefore, we reject the null hypothesis.

**4.0 CONCLUSION:**

An analysis of the growth rate on the average house price data for four US regions provided coupled with existing literature reveals that there is no significant difference in the growth rate of average house sales prices between the Midwest, South, West, and Northeastern region.

However, the analysis of variance on the prices itself showed a statistical significance on the average sales price of houses in the regions. Through a variety of methods including descriptive statistics, time series analysis, analysis of variance and post-hoc tests via Bonferroni and TukeyHSD tests.

Below is a summary of the conclusions drawn from the analysis conducted in this project.

Housing price is quite diverse and is dependent on location. Although the average sales price of houses has been said to have increased over the last 20 years, this study shows that there is no significant difference in the growth rate across regions, as shown in the inflation adjusted sales prices across the regions.

It is important to note that if inflation is not given attention, this study would have perhaps presented a different result. This can be tricky because of the changes in the industry over the years and changes in price. However, adjusting for inflation wiped out the trend. This can be seen in the graphs presented in *figure 3 - figure 6* and as confirmed by the seasonal dummies created to see if there is a seasonal effect in the data. The Midwest, South and West regions showed no trend. However, the Northeast region showed that trend was significant.

The essential value of a house involves factors like its proximity to good-paying jobs, aspects of its location such as climate, good school district, size per square of the house, age of the house and trendy architecture

According to a report by realtor.com, the northeast dominated the 2020 hottest market, and the region has half of the country’s 10 fastest-moving markets. This can be as a result of the region’s population density and has many of the most expensive cities as compared to other regions.

The analysis of variance rejected the null hypothesis and concluded that there was indeed a statistical difference when the average sale price of houses are compared between the regions.

This may be as a result of population size, square-foot size per house, density, or the need for luxury in the different regions. Notably, the crucial price of a house is based on supply and demand and demand for houses in a particular location: If it is appealing to live in that area, if it is appealing to live in that area and there aren't enough homes for newcomers, the fundamental value of each house will increase.

Overall, the Northeast is seen to have differed from most statistical analysis carried out.

Due to time limitation, this study was unable to dive deeper into the cause of these difference in average prices in the regions. Future research should look into the cause of this difference in these four regions and why the housing industry having been adjusted for inflation the growth rates appear to be relatively similar across the regions based on the data and the analysis conducted.

**REFERENCES**

*U.S. Census Bureau and U.S. Department of Housing and Urban Development, Average Sales Price of Houses Sold for the West Census Region [ASPW], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/ASPW, November 10, 2021.*

*James A. Kahn “What Drives Housing Prices?” Federal Reserve Bank of New York Staff Reports, no. 345 September 2008 JEL classification: E22, E32, O41, O51*

*Attanasio, O., L. Blow, R. Hamilton, A. Leicester, “Consumption, House Prices and Expectations,” Bank of England Working Paper Series no. 271 (2005).*

*Case, K.E., and Shiller, R.J., “Is There a Bubble in the Housing Market?” Brookings Papers on Economic Activity, No. 2, 2003*

*Himmelberg, C., C. Mayer, and T. Sinai, “Assessing High House Prices: Bubbles: Fundamentals, and Misperceptions” Federal Reserve Bank of New York Staff Report no. 218, September 2005.*

*Iacoviello, M., and Neri, S., “The Role of Housing Collateral in an Estimated Two-Sector Model of the US Economy,” manuscript, 2006.*

*Introduction to US Economy: Housing market* [*https://crsreports.congress.gov/product/pdf/IF/IF11327/8*](https://crsreports.congress.gov/product/pdf/IF/IF11327/8)

*The builder’s daily policy. From Peak to trough to rebound in 60 days: The Covid-19 recission* [*https://www.thebuildersdaily.com/from-peak-to-trough-to-rebound-in-60-days-the-covid-recesison/#:~:text=As%20of%202020%2C%20spending%20on%20residential%20fixed%20investment,utilities%20and%20homeowners%E2%80%99%20imputed%20rent%20and%20utility%20payments*](https://www.thebuildersdaily.com/from-peak-to-trough-to-rebound-in-60-days-the-covid-recesison/#:~:text=As%20of%202020%2C%20spending%20on%20residential%20fixed%20investment,utilities%20and%20homeowners%E2%80%99%20imputed%20rent%20and%20utility%20payments)*.*

*Fang Block, Realtor.com, Aug 2020.*