Housing Affordability

The Housing Affordability Data System. [HADS]
From 2005 to 2013

Introduction

This project answer these Questions

- Are there differences in the Market Values of occupied versus Unoccupied housing units?
- Is there a pattern in these differences over the period 2005 - 2013?
- Analysis of the differences in Fair Market Rents(FMR) across the various years.
- Building Prediction Model for House Market Value for year 2013.

Using “Housing Affordability Data System” of the U.S. Department of Housing and Urban Development. [HADS] Data-set

Data : The Housing Affordability Data System (HADS)

a set of files derived from the 1985 and later national American Housing Survey (AHS) and the 2002 and later Metro AHS.

This system categorizes housing units by affordability and households by income, with respect to the Adjusted Median Income, Fair Market Rent (FMR), and poverty income. It also includes housing cost burden for owner and renter households.

These files have been the basis for the worst case needs tables since 2001. The data files are available for public use, since they were derived from AHS public use files and the published income limits and FMRs.

We take data from year 2005 to 2013 as a sample for our Project.

Analysis and Methodology For Question 1 & 2

1. Are there differences in the Market Values of occupied versus Unoccupied housing units?
2. Is there a pattern in these differences over the period 2005 - 2013?

- Data Cleaning
- Descriptive statistics
- Plotting overall trend of market value for houses prices (2005-2013)
- Hypothesis testing for each year between Occupied and Unoccupied
- Summary & Answer For Question 1 & 2

Data Cleaning

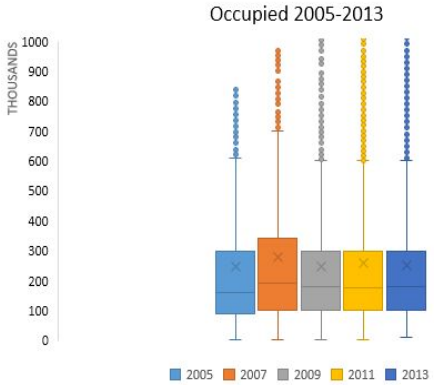
Using power query to

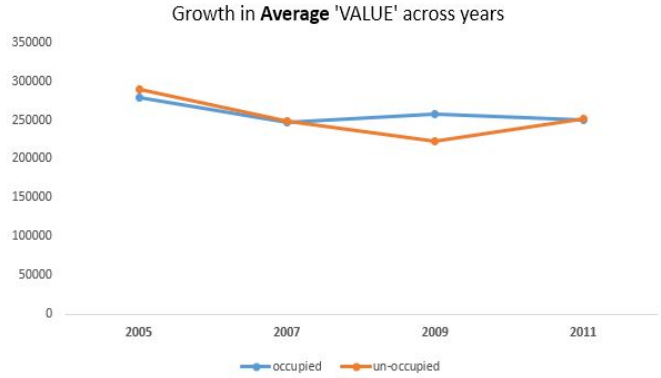
- Keep (Control, Status, Value) Columns and Remove the rest
- Remove all rows that has missing value or less than 1000
- Using table Filter on Status column to separate Occupied vs Unoccupied

We do this for each year from (2005 - 2013)

Descriptive Statistics & Plotting overall trend

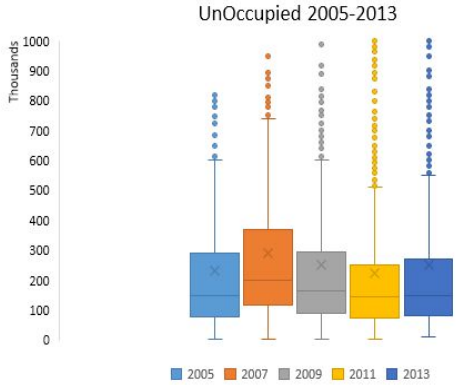
	2005	2007	2009	2011	2013
Mean	247130.8466	278960.7533	247681.9663	258136.2211	249858.5465
Standard Error	1642.721687	1949.56645	1577.650565	1050.644519	1499.97655
Median	160000	190000	179000	177000	180000
Mode	200000	200000	200000	200000	150000
Standard Deviation	281859.6405	317162.7659	273625.7419	301001.8618	282290.6451
Sample Variance	79444856915	1.00592E+11	74871046642	90602120816	79688008338
Kurtosis	11.02424529	12.86411398	31.88616457	52.89559501	33.43670103
Skewness	3.090205632	3.285310639	4.667869315	5.4596834	4.84535521
Range	1539794	1828479	2464647	5263699	2510000
Minimum	1000	1000	1000	1000	10000
Maximum	1540794	1829479	2465647	5264699	2520000
Sum	7275532125	7382975298	7450521228	21187304757	8849490000
Count	29440	26466	30081	82078	35418





Descriptive stats for Un-Occupied Houses

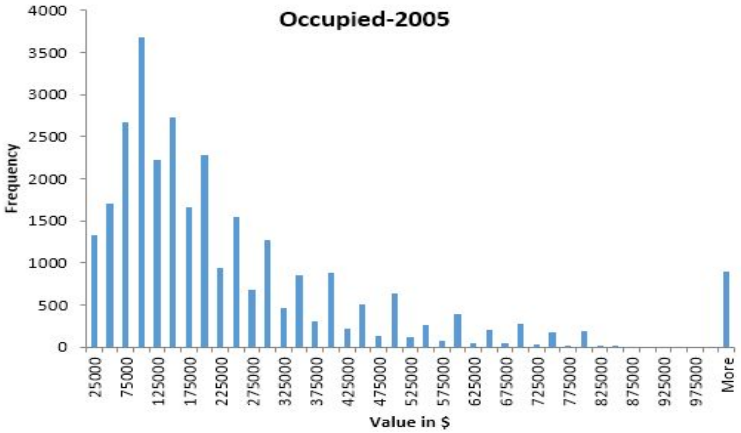
	2005	2007	2009	2011	2013
Mean	229324.3594	289004.4928	249230.0607	222116.855	251996.8178
Standard Error	8067.007619	8431.176942	9048.176411	5802.637262	10990.32368
Median	150000	200000	165000	144450	150000
Mode	1540794	1829479	200000	200000	150000
Standard Deviation	264371.4834	306203.818	318104.853	316336.8786	389653.0876
Sample Variance	69892281216	93760778164	1.01191E+11	1.00069E+11	1.5183E+11
Kurtosis	12.37474982	13.09387441	26.91874803	41.28221051	23.01107322
Skewness	3.165239742	3.217984913	4.538695664	5.428402134	4.526607886
Range	1539594	1828479	2464647	4413135	2510000
Minimum	1200	1000	1000	1000	10000
Maximum	1540794	1829479	2465647	4414135	2520000
Sum	246294362	381196926	308048355	660131293	316760000
Count	1074	1319	1236	2972	1257

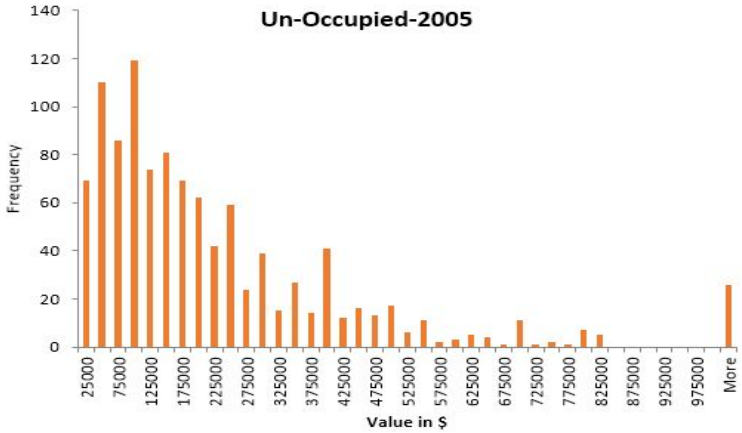


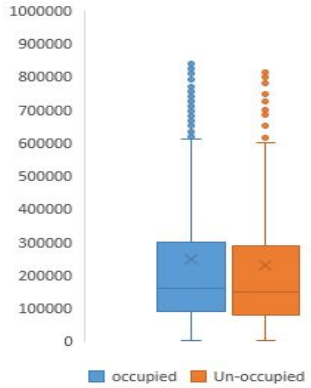
Summary

The houses prices are going down as a trend
With a difference in year 2009
where
The Occupied houses keep its value
Then the Unoccupied follow it up again

Hypothesis test for 2005







t-Test: Two-Sample Assuming Equal Variances

	occupied	*Un-occupied*
Mean	247130.8466	229324.3594
Variance	79444856915	69892281216
Observations	29440	1074
Pooled Variance	79108926340	
Hypothesized Mean Difference	0	
df	30512	
t Stat	2.03791929	
P(T<=t) one-tail	0.020783306	
t Critical one-tail	1.644903568	
P(T<=t) two-tail	0.041566612	
t Critical two-tail	1.960041736	

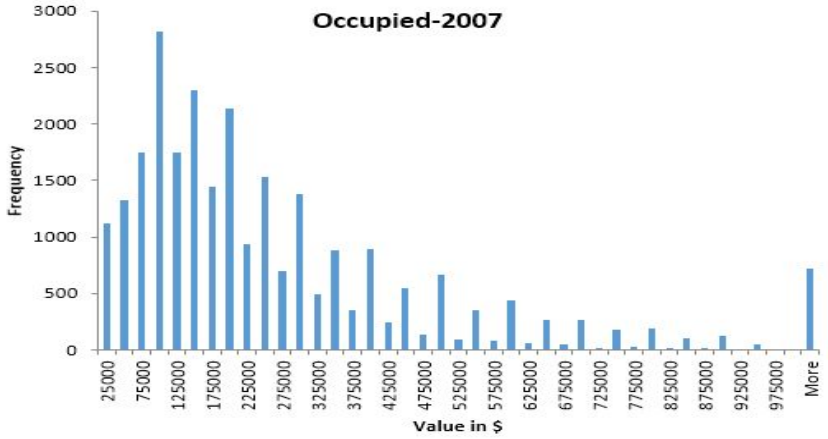
H_0: $\mu_{Occupied} - \mu_{Not\text{-}Occupied} = 0$ **Reject**

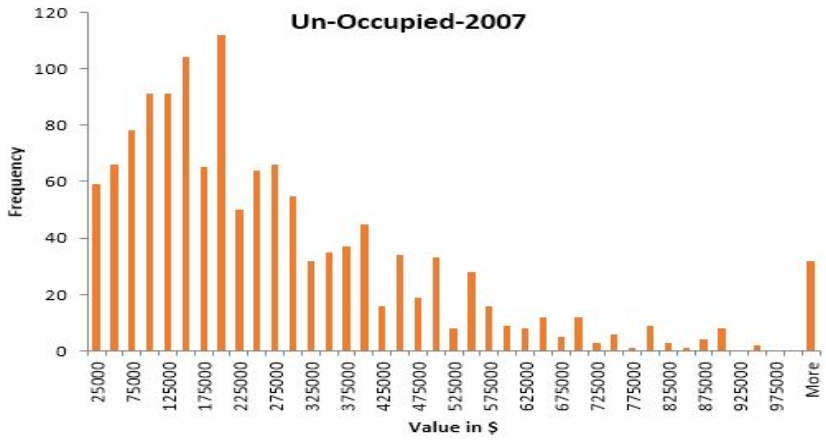
H_A: $\mu_{Occupied} - \mu_{Not\text{-}Occupied} \neq 0$

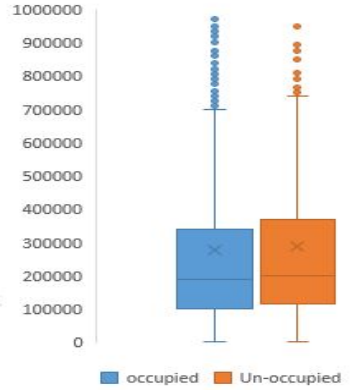
Summary

We Reject the Null Hypothesis which means that
The difference in market value for occupied houses is greater than not occupied
and, the difference is statistically significant

Hypothesis test for 2007







t-Test: Two-Sample Assuming Equal Variances

	occupied	*Un-occupied*
Mean	278960.7533	289004.4928
Variance	1.00592E+11	93760778164
Observations	26466	1319
Pooled Variance	1.00268E+11	
Hypothesized Mean Difference	0	
df	27783	
t Stat	-1.12428212	
P(T<=t) one-tail	0.130451537	
t Critical one-tail	1.644908474	
P(T<=t) two-tail	0.260903074	
t Critical two-tail	1.960049374	

H_0: $\mu_{Occupied} - \mu_{Not\text{-}Occupied} = 0$ **Fail to Reject**

H_A: $\mu_{Occupied} - \mu_{Not\text{-}Occupied} \neq 0$

Summary

We Fail to Reject the Null Hypothesis which means that The difference in market value for houses in 2007 is NOT statistically significant

Summary & Answer For Question 1 & 2

- **Are there some differences in the Market values of occupied versus Unoccupied housing units?**

  Difference in the Market Values is significant only for years 2005 and 2011.
  In these years the market value of 'Occupied' units was greater than 'Un-Occupied' units.
  For the remaining years there is no significant difference in the market value across 'Occupied' and 'Un-Occupied' units

- **Do these differences have a pattern over the period 2005 through 2013?**

  The pattern discernable is that the Market value of 'Occupied' units is never less than that for 'Un-Occupied' units.
  It is either greater (as in years 2005 and 2011) or equal (as in the remaining years).

Analysis and Methodology For Question 3

Analysis of the differences in Fair Market Rents(FMR) across years.

- Data Cleaning
- Descriptive statistics
- Plotting overall trend of FMR average price (2005-2013)
- Hypothesis testing between each year and the earlier one
- Summary & Answer For Question 3

Data Cleaning

Using power query to

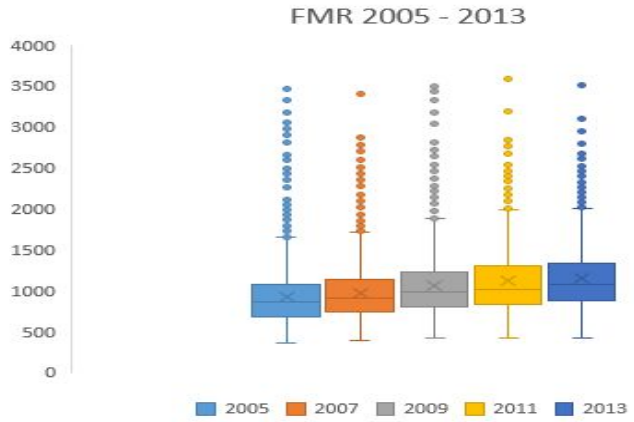
- Keep (Control, FMR) Columns and Remove the rest
- Remove all rows that has missing or negative value
- Using Vlookup to merge all the years in one table and dropping missing values

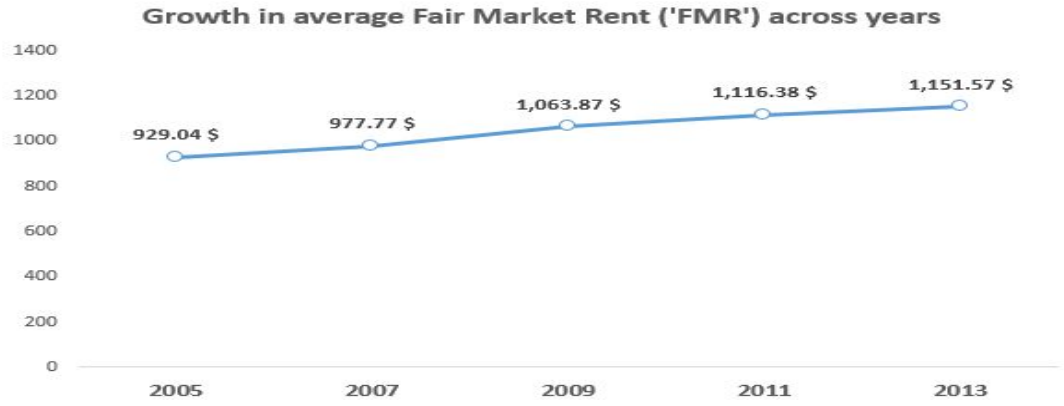
We do this for each year from (2005 - 2013)

Descriptive Statistics and Trend Line for FMR

Descriptive stats for FMR

	FMR_2005	FMR_2007	FMR_2009	FMR_2011	FMR_2013
Mean	929.039776	977.76984	1063.866	1116.38153	1151.569143
Standard Err	2.03831535	2.07549457	2.26212228	2.44278766	2.427761204
Median	863	908	983	1014	1082
Mode	679	738	941	966	1032
Standard Dev	331.017643	337.055462	367.363366	396.703001	394.2627387
Sample Varia	109572.68	113606.385	134955.843	157373.271	155443.1071
Kurtosis	2.99056658	2.43556763	2.31299575	2.06693658	1.754256298
Skewness	1.41977542	1.35066093	1.30453136	1.3030014	1.163717371
Range	3104	3013	3074	3162	3090
Minimum	360	387	427	424	421
Maximum	3464	3400	3501	3586	3511
Sum	24501566	25786724	28057338	29442330	30370333
Count	26373	26373	26373	26373	26373





Year	Total FMR	% increase
2005	$24,501,566	-
2007	$25,786,724	5.25%
2009	$28,057,338	8.81%
2011	$29,442,330	4.94%
2013	$30,370,333	3.15%

Summary

The trend for Fair Market Rent is going up Throughout the data with significant increase Each year with high increase at 2009.

Hypothesis test for 2007 vs 2005 (FMR)

Year 2007 Vs 2005

t-Test: Paired Two Sample for Means

	2007	*2005*
Mean	977.7698404	929.0397755
Variance	113606.3846	109572.6799
Observations	26373	26373
Pearson Correlation	0.942043935	
Hypothesized Mean Difference	0	
df	26372	
t Stat	69.49039888	
P(T<=t) one-tail	0	
t Critical one-tail	1.644911409	
P(T<=t) two-tail	0	
t Critical two-tail	1.960053943	

H_0: $FMR_{2007} - FMR_{2005} = 0$ **Reject**

H_A: $FMR_{2007} - FMR_{2005} \neq 0$

Summary

The difference is statistically significant in this Years and all other

Summary & Answer For Question 3

Analysis of the differences in Fair Market Rents(FMR) across years.

As seen by the various statistical tests (t-tests for differences in means) across years,

the highest increase was observed from 2007 to 2009, the period overlapping the subprime mortgage crisis.

it can be seen that the Fair Market Rents continuously rose across these various years.

Furthermore, when calculate the percentage increases across years, the highest increase was observed

from 2007 to 2009, The period overlapping the subprime mortgage crisis.

Analysis and Methodology For Question 4

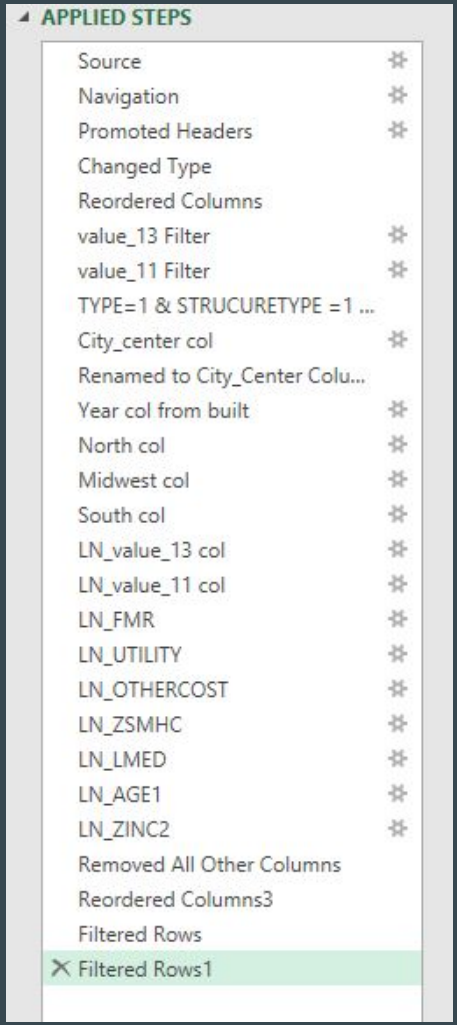
Building Prediction Model for House Market Value for year 2013.

- Data Cleaning
- Splitting the data to Training data and testing data
- Variable Definition table
- Descriptive statistics for all variables
- Building a Regression Model
- Regression Model Equation & it's Meaning
- Holdout Analysis & Mean Absolute Deviation (MAD)
- Actual value vs Predicted & Residual plots
- Summary & Answer For Question 4

Data Cleaning

- Using Vlookup to add Value 2013 to 2011 file and make 'raw_data_PredictionModel.xlsx'
- Using Power Query to Remove all missing data or values less than 1000
- Filter the data to include single family house (TYPE = 1 , STRUCTERUPE = 1)
- Adding dummy Variables (City_center, North , Midwest, South)
- Calculate Year variable from Built column
- Taking LN( ) Some Variables to get the perfect linear relationship
- Remove all other columns
- After power query i get non-numeric data so i Checked every column using ISNUMBER( ) function in Excel and remove them

Check the power query at 'Prediction_model.xlsx' for equation for each step



Splitting the data to Training data and testing data

- Using Rand( ) function in Excel to make a random column
- Sort ascending according to random column
- Take first 1000 rows and separate them as a test data
- Using the rest as training data

See (raw_data + Random ,Training data , test data) worksheets at Prediction_model.xlsx

Variable Definition

Value_2013	Market value in 2013	AGE1	Age of head of household
Value_2011	Market value in 2011	Year	How old was this building in 2011?
FMR	Fair Market Monthly Rent	City_Center	at the city center or not
UTILITY	Monthly utilities cost	North	North area (dummy variable)
OTHERCOST	Sum of 'other monthly costs'	Midwest	Midwest area (dummy variable)
ZSMHC	Monthly housing costs	South	South area (dummy variable)
LMED	Area Median Income	ROOMS	How many rooms are in the unit
ZINC2	Annual Household income	BEDRMS	How many bedrooms are in the unit?

Descriptive statistics for all variables

	LN_value_13	LN_value_11	LN_FMR	LN_UTILITY	LN_OTHERCO	LN_ZSMHC	LN_LMED	LN_ZINC2	LN_AGE1	Year	City_Center	North	Midwest	South	ROOMS	BEDRMS
Mean	12.1206797	12.1137446	7.0355646	5.5322344	4.2285499	6.991726	11.09213	10.979064	3.9690033	41.617743	0.2166955	0.1717609	0.2693195	0.3610631	6.6328335	3.252451
Standard Error	0.00538452	0.00523523	0.0021539	0.0031519	0.0054633	0.0050628	0.001181	0.0067184	0.0020511	0.1827873	0.0028562	0.0026148	0.0030753	0.0033298	0.0111418	0.0057489
Median	12.1007121	12.0951411	6.9902565	5.5373343	4.1997051	7.0483864	11.05889	11.089393	4.0073332	36	0	0	0	0	6	3
Mode	11.9183906	12.2060726	6.8731638	5.4380793	3.912023	5.888878	11.047806	11.512565	3.988984	61	0	0	0	0	6	3
Standard Deviation	0.77671654	0.75518074	0.3107015	0.4546616	0.7880787	0.7303012	0.1703534	0.9691266	0.2958681	26.367035	0.4120033	0.3771815	0.4436169	0.4803203	1.6072068	0.8292739
Sample Variance	0.60328859	0.57029794	0.0965354	0.2067171	0.6210681	0.5333399	0.0290203	0.9392063	0.0875379	695.22053	0.1697467	0.1422659	0.196796	0.2307076	2.5831138	0.6876952
Kurtosis	0.89533508	1.02287142	0.0564819	0.9759001	8.4781335	-0.343252	0.7502978	7.8944282	-0.307025	-0.873543	-0.108324	1.0299647	-0.918283	-1.665413	1.0037294	1.0032467
Skewness	0.03789606	0.08958285	0.496353	-0.221544	-0.885583	-0.137921	0.5617951	-1.510925	-0.449604	0.3700914	1.3753857	1.7406509	1.0400987	0.5785735	0.7668521	0.4655226
Range	5.52942909	7.98710206	2.1469217	5.0218263	9.847023	5.6133769	1.3235347	14.219342	1.4880771	92	1	1	1	1	13	8
Minimum	9.21034037	7.31322039	6.0378709	2.2064412	-2.484907	3.7841896	10.425253	0	3.0445224	0	0	0	0	0	1	0
Maximum	14.7397695	15.3003224	8.1847927	7.2282675	7.3621164	9.3975665	11.748788	14.219342	4.5325995	92	1	1	1	1	14	8
Sum	252207.104	252062.799	146396.03	115114.73	87987.665	145483.84	230805.05	228452.37	82587.021	865982	4509	3574	5604	7513	138016	67677
Count	20808	20808	20808	20808	20808	20808	20808	20808	20808	20808	20808	20808	20808	20808	20808	20808

Building a Regression Model

- I made a model that gives the Higher R squared and Lower Mean Absolute Deviation (MAD) ,
  used the dependent variable is the 2013 Market Value while all the independent variables are from the year 2011.

- I included 14 independent variable from the year 2011 and choose them according to logic which one affect the market Value the most and these are :

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.776744994
R Square	0.603332786
Adjusted R Square	0.603032159
Standard Error	0.488703587
Observations	19808

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	15	7189.709704	479.314	2006.915	0
Residual	19792	4726.947034	0.238831		
Total	19807	11916.65674			

		Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
β_0	Intercept	-0.368042809	0.294318831	-1.25049	0.211135	-0.944932396	0.208846778	-0.944932396	0.208846778
β_1	LN_value_11	0.566754999	0.006780631	83.5844	0	0.553464393	0.580045604	0.553464393	0.580045604
β_2	LN_FMR	0.349978777	0.024329367	14.38503	1.11E-46	0.302291177	0.397666376	0.302291177	0.397666376
β_3	LN_UTILITY	0.020956547	0.009335326	2.244865	0.024788	0.002658526	0.039254568	0.002658526	0.039254568
β_4	LN_OTHERCOST	0.000469211	0.005247125	0.089422	0.928747	-0.009815595	0.010754016	-0.009815595	0.010754016
β_5	LN_ZSMHC	0.062411984	0.00727663	8.577046	1.04E-17	0.04814918	0.076674788	0.04814918	0.076674788
β_6	LN_LMED	0.154706185	0.034471586	4.487933	7.23E-06	0.087138986	0.222273384	0.087138986	0.222273384
β_7	LN_ZINC2	0.04315068	0.004252194	10.14786	3.88E-24	0.034816022	0.051485338	0.034816022	0.051485338
β_8	LN_AGE1	0.100895383	0.013580919	7.429201	1.14E-13	0.074275642	0.127515123	0.074275642	0.127515123
β_9	Year	-0.001807096	0.000146427	-12.3413	7.29E-35	-0.002094105	-0.001520088	-0.002094105	-0.001520088
β_{10}	City_Center	-0.043772403	0.008687241	-5.0387	4.73E-07	-0.060800124	-0.026744682	-0.060800124	-0.026744682
β_{11}	North	-0.076846437	0.013242364	-5.80308	6.61E-09	-0.102802581	-0.050890294	-0.102802581	-0.050890294
β_{12}	Midwest	-0.100403523	0.013892343	-7.22726	5.11E-13	-0.127633679	-0.073173367	-0.127633679	-0.073173367
β_{13}	South	-0.103501453	0.011283328	-9.17295	5.04E-20	-0.125617722	-0.081385184	-0.125617722	-0.081385184
β_{14}	ROOMS	0.045015847	0.003497083	12.8724	9.13E-38	0.038161271	0.051870423	0.038161271	0.051870423
β_{15}	BEDRMS	-0.036715456	0.007422271	-4.94666	7.61E-07	-0.051263729	-0.022167183	-0.051263729	-0.022167183

According to P-value :

All of (X variables) is statistically significant Except (OTHERCOST, Intercept) in our Model

Interpretation of R-square:

The R-squared and adjusted R-square are about 0.60, indicating that the model explains about 60 percentage of variation in the market value of housing units.

Regression Model Equation & it's Meaning

$$\ln(VALUE\_13) = \beta 0 + \beta 1 \ln(Value\_11) + \beta 2 \ln(FMR) + \beta 3 \ln(UTILITY) + \beta 4 \ln(OTHERCOST) + \beta 5 \ln(ZSMHC) + \beta 6 \ln(LMED) + \beta 7 \ln(ZINC2) + \beta 8 \ln(AGE1) + \beta 9 YEAR + \beta 10 CITY\_CENTER + \beta 11 NORTH + \beta 12 MIDWEST + \beta 13 SOUTH + \beta 14 ROOMS \beta 14 BEDRMS$$

β0 -0.37 No managerially relevant interpretation, since talking about a situation when all 'X' variables are zero does not make managerial sense.

β1 0.57 For every one percentage increase in the value 2011 , the market value increases by 57 %, all other variables remaining at the same level.

β2 0.35 For every one percentage increase in the fair market rent, the market value increases by 35 %, all other variables remaining at the same level.

β3 0.02 For every one percentage increase in the Utility cost, the market value increases by 2 %, all other variables remaining at the same level

β4 0.00 For every one percentage increase in the Other cost, the market value does not affect besides its statistically insignificant

β5 0.06 For every one percentage increase in the Monthly housing costs , the market value increases by 6 %, all other variables remaining at the same level.

β6 0.15 For every one percentage increase in Area Median Income , the market value increases by 15 %, all other variables remaining at the same level.

β7 0.04 For every one percentage increase in Annual Household income, the market value increases by 4 %, all other variables remaining at the same level.

β8 0.10 For every one percentage increase in Age of head of household, the market value increases by 10.8 % , all other variables remaining at the same level.

β9 0.00 For every one year increase in Age of the house, the market value decreases by 18 %, all other variables remaining at the same level.

β10 -0.04 When the geographical location of the Housing unit is classified as 'Central CIty' area, then the market value of the housing unit is lower by 4.37%, all other variables being kept at the same level.

β11 -0.08 When the housing unit is in the Northeast region of the country, then the market value tends to be lower by 7.68 % as compared to a similar housing unit being in the West region, all other variables being kept at the same level. β12, and β13 can be similarly interpreted

β14 0.04 One additional room corresponds to a 4.5% increase in the market value of the housing unit, all other variables being kept at the same level

β15 -0.03 One additional bedroom corresponds to a 3.6% decrease in the market value of the housing unit, all other variables being kept at the same level.

Holdout Analysis & Mean Absolute Deviation (MAD)

At this stage we will use our Model and test its accuracy by

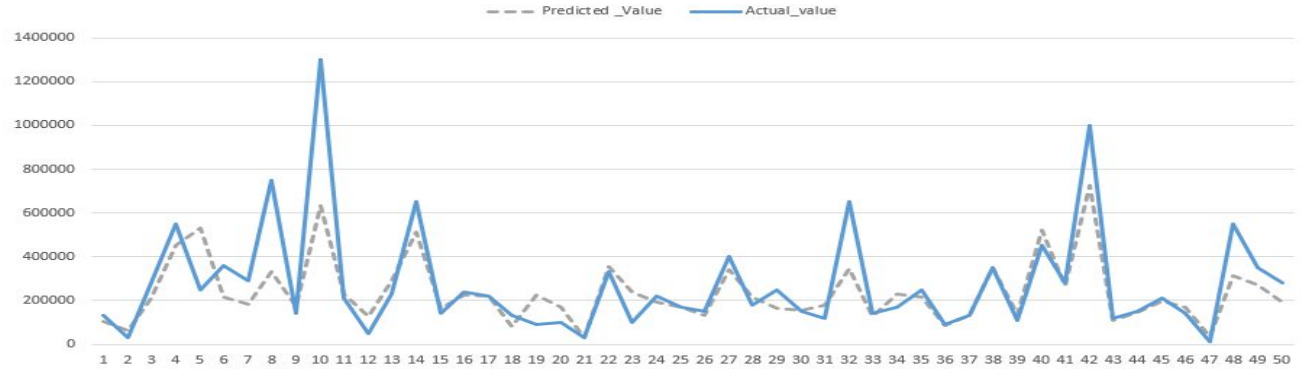
- Using test data from (Test worksheet) to predict the house value for 2013.
- Compare a predicted value and Actual value by calculating the Mean Absolute Deviation (MAD)

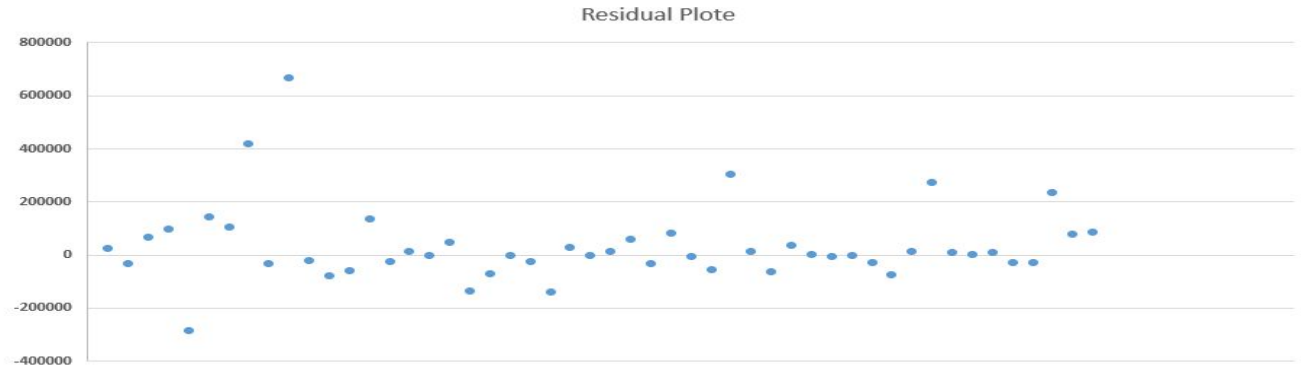
	Intercept	LN_value_11	LN_FMR	LN_UTILITY	LN_OTHERCOST	LN_ZSMHC	LN_LMED	LN_ZINC2	LN_AGE1	Year	City_Center	North	Midwest	South	ROOMS	BEDRMS
Coefficients	-0.368042809	0.566754999	0.349978777	0.020956547	0.000469211	0.062411984	0.154706185	0.04315068	0.100895383	-0.001807096	-0.0437724	-0.07684644	-0.10040352	-0.10350145	0.045015847	-0.03672
	-0.37	0.57	0.35	0.02	0.00	0.06	0.15	0.04	0.10	0.00	-0.04	-0.08	-0.10	-0.10	0.05	-0.04

Predicted _LN(Value)	Predicted _Value	Actual_value	ABS(Difference)	Mean abs(Diff) [MAD]
12.19675094	198144.3212	240000	41855.67878	72,856.32
11.61067647	110268.8189	120000	9731.181082	
11.79841232	133040.9594	130000	3040.959355	
11.73423927	124771.4907	140000	15228.50926	
12.89927114	400020.5272	90000	310020.5272	

See (test data) worksheets at Prediction_model.xlsx for formula and calculation

Actual value vs Predicted & Residual plots





Summary

The difference between predicted values and Actual value is not that match

Residual plot shows that
The errors centered around zero line
With some outliers

In general the Model perform well and we can use it to get a good answers.

Check the 'Plotting predicted vs Actual worksheet at

Prediction_model.xlsx' for equation and graphs

Summary & Answer For Question 4

The regression model now has a R-square of 0.60 since we added the Market Value for year 2011 as an additional 'X' variable.

Using the coefficients from this regression model and using the set of 'X' variables in the hold out data we make predictions of the Market Value for the 1000 housing units held out.

The MAD statistic (Mean Absolute Deviation) for the prediction turns out to be $ 72,856.32 .
This seems ok given that the average Market Value is around $252,262.03.

Please see 'Prediction_model.xlsx' for various calculations

Conclusion

This Case study gives us some insights

- the Market Values of occupied always more than Unoccupied houses with the all Market value going down as it gets older.
- The Rents is always going up.
- We can Predict the House Market Value and Rents Using this data.