Final summary

1 . House-and-Urban-development

The data sets are based on the American Housing Survey, AHS, national files from 1985 onwards. Housing-level variables include information on the number of rooms in the housing unit, the year the unit was built, whether it was occupied or vacant, whether the unit was rented or owned, whether it was a single unit or multi-unit structure, the number of units in the building, the current market value of the unit and measure of relative housing costs. The dataset also includes variables describing the number of people living in the household, household income and the type of residential area. Example, urban or suburban. These are made available to the public by the US department of housing and urban development. These data are made available every two years. The latest data being available for year 2013. For our project we'll be using data from year 2005 onwards. That is we will use the data from 2005, 2007, 2009, 2011 and 2013.

2. Variables:

Variables used in the analysis are listed in file: List-of-variables

3. Questions to be answered:

1. Whether there are some differences in the market values of occupied versus not occupied housing units.

2. Whether these differences have a pattern over the period 2005 through 2013

- I carried out a descriptive statistical comparison among the housing prices from 2005 to 2013

- Then I carried out t-Test of two-Sample Assuming Equal Variances to test the hypothesis

House price - unoccupied					
	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

Table 1 Statistical summary of housing price (occupied)

House price - occupied					
	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

Table 2 Statistical summary of housing price (unoccupied)

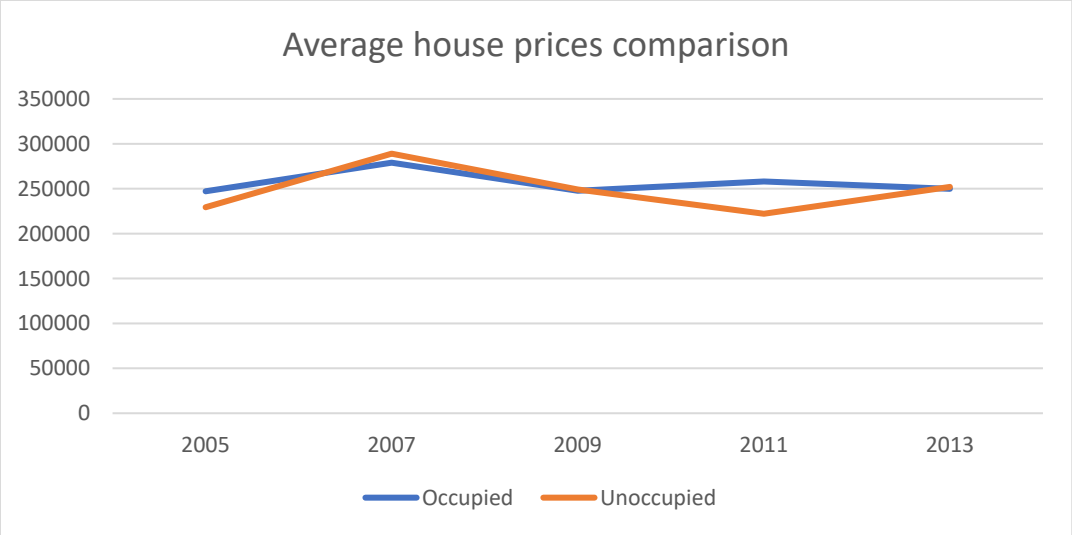


Figure 1 Comparison of housing price between occupied and unoccupied

2005 t-Test: Two-Sample Assuming Equal Variances		
	Occupied	Not occupied
Mean	247130.8466	229324.3594
Variance	79444856915	69892281216
Observations	29440	1074
Pooled Variance	79108926340	
Hypothesized Mean Dif	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	

2011 t-Test: Two-Sample Assuming Equal Variances		
	1'	3'
Mean	258136.2211	222116.855
Variance	90602120816	1.00069E+11
Observations	82078	2972
Pooled Variance	90932830060	
Hypothesized Mean Dif	0	
df	85048	
t Stat	6.396999818	
P(T<=t) one-tail	7.96396E-11	
t Critical one-tail	1.644871544	
P(T<=t) two-tail	1.59279E-10	
t Critical two-tail	1.959991878	

Table 3 and 4 t-Test of housing prices between occupied and unoccupied

2007 t-Test: Two-Sample Assuming Equal Variances

	Occupied	*Not occupied*
Mean	278960.7533	289004.4928
Variance	1.00592E+11	93760778164
Observations	26466	1319
Pooled Variance	1.00268E+11	
Hypothesized Mean Dif	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	

2013 t-Test: Two-Sample Assuming Equal Variances

	Occupied	*Not occupied*
Mean	249858.5465	251996.8178
Variance	79688008338	1.5183E+11
Observations	35418	1257
Pooled Variance	82158756559	
Hypothesized Mean Dif	0	
df	36673	
t Stat	-0.259914481	
P(T<=t) one-tail	0.3974656	
t Critical one-tail	1.644895178	
P(T<=t) two-tail	0.7949312	
t Critical two-tail	1.960028674	

2009 t-Test: Two-Sample Assuming Equal Variances

	Occupied	*Not occupied*
Mean	247681.9663	249230.0607
Variance	74871046642	1.01191E+11
Observations	30081	1236
Pooled Variance	75909040217	
Hypothesized Mean Dif	0	
df	31315	
t Stat	-0.193604738	
P(T<=t) one-tail	0.423243322	
t Critical one-tail	1.644902288	
P(T<=t) two-tail	0.846486644	
t Critical two-tail	1.960039743	

Conclusion:

- 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 'Not-Occupied' units. For the remaining years there is no significant difference in the market value across 'Occupied' and 'Not-Occupied' units.
- The pattern discernable is that the Market value of 'Occupied' units is never less than that for 'Not-Occupied' units. It is either greater (as in years 2005 and 2011) or equal (as in the remaining years).

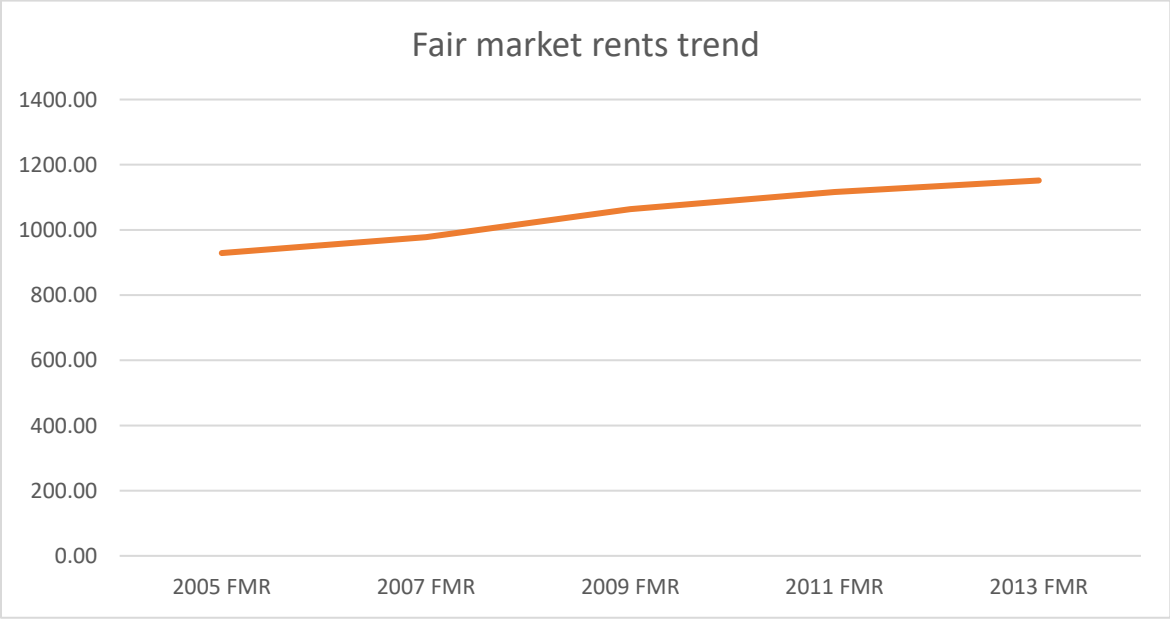
3. Is there a way to see if there was an effect of the 2008 subprime mortgage crisis, particularly, an effect on housing rent?

- I carried out a descriptive statistical comparison among the fair market rents from 2005 to 2013

- Then I carried out t-Test of faired two samples for means to test the hypothesis

Table 8 Statistical summary of FMR

	2005 FMR	***2007 FMR***	***2009 FMR***	***2011 FMR***	***2013 FMR***
Mean	929.04	977.77	1063.87	1116.38	1151.57
Standard Error	2.04	2.08	2.26	2.44	2.43
Median	863.00	908.00	983.00	1014.00	1082.00
Mode	679.00	738.00	941.00	966.00	1032.00
Standard Deviation	331.02	337.06	367.36	396.70	394.26
Sample Variance	109572.68	113606.38	134955.84	157373.27	155443.11
Kurtosis	2.99	2.44	2.31	2.07	1.75
Skewness	1.42	1.35	1.30	1.30	1.16
Range	3104.00	3013.00	3074.00	3162.00	3090.00
Minimum	360.00	387.00	427.00	424.00	421.00
Maximum	3464.00	3400.00	3501.00	3586.00	3511.00
Sum	24501566.00	25786724.00	28057338.00	29442330.00	30370333.00
Count	26373.00	26373.00	26373.00	26373.00	26373.00



	One-tail	**Two-tailed**
H0:	After - before >=0	after - before =0
Ha:	After - before < 0	after - before <> 0

2007 vs. 2005 t-Test: Paired Two Sample for Means

	2007 FMR	*2005 FMR*		
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		not rejected	rejected
P(T<=t) two-tail	0			
t Critical two-tail	1.960053943		rejected	not rejected

2009 vs. 2007 t-Test: Paired Two Sample for Means

	2009 FMR	*2007 FMR*		
Mean	1063.865999	977.7698404		
Variance	134955.8426	113606.3846		
Observations	26373	26373		
Pearson Correlation	0.9526349			
Hypothesized Mean Difference	0			
df	26372			
t Stat	124.3221662			
P(T<=t) one-tail	0			
t Critical one-tail	1.644911409		not rejected	rejected
P(T<=t) two-tail	0			
t Critical two-tail	1.960053943		rejected	not rejected

t-Test: Paired Two Sample for Means

	2011 FMR	*2009 FMR*		
Mean	1116.381527	1063.865999		
Variance	157373.2707	134955.8426		
Observations	26373	26373		
Pearson Correlation	0.957568947			
Hypothesized Mean Difference	0			
df	26372			
t Stat	74.15124024			
P(T<=t) one-tail	0			
t Critical one-tail	1.644911409		not rejected	rejected
P(T<=t) two-tail	0			
t Critical two-tail	1.960053943		rejected	not rejected

t-Test: Paired Two Sample for Means

	2013 FMR	*2011 FMR*		
Mean	1151.569143	1116.381527		
Variance	155443.1071	157373.2707		
Observations	26373	26373		
Pearson Correlation	0.969212055			
Hypothesized Mean Difference	0			
df	26372			
t Stat	58.21091561			
P(T<=t) one-tail	0			
t Critical one-tail	1.644911409		not rejected	rejected
P(T<=t) two-tail	0			
t Critical two-tail	1.960053943		rejected	not rejected

Conclusion:

- As seen by the various statistical tests (t-tests for differences in means) it can be seen that the Fair Market Rents continuously rose across these various years. Further if we calculate the percentage increases across years, the highest increase was observed from 2007 to 2009, the period overlapping the subprime mortgage crisis.

3. What factors predict or contribute to the housing or the current market value of single-family units? (Only single-family units based the data of 2013)

- Use the data of 2013 to create a multi-linear regression model

- Test and transform the model to select the model with best R square: all the steps are recorded in the excels of Q3.

STATISTICAL SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.739160495
R Square	0.546358237
Adjusted R Square	0.546156413
Standard Error	0.538652869
Observations	31483

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	14	10996.43	785.459	2707.108	0
Residual	31468	9130.343	0.290147		
Total	31482	20126.77			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	4.139606855	0.266845	15.51314	4.49E-54	3.61658	4.662634	3.61658	4.662634
Status	-0.362997797	0.028613	-12.6864	8.66E-37	-0.41908	-0.30691	-0.41908	-0.30691
Metros	-0.092263227	0.007592	-12.1529	6.6E-34	-0.10714	-0.07738	-0.10714	-0.07738
Age_of_unit	-0.002269527	0.000125	-18.1605	2.52E-73	-0.00251	-0.00202	-0.00251	-0.00202
Northeast	-0.128429018	0.011348	-11.3174	1.23E-29	-0.15067	-0.10619	-0.15067	-0.10619
Midwest	-0.308286814	0.011938	-25.8244	1.6E-145	-0.33169	-0.28489	-0.33169	-0.28489
South	-0.265315003	0.010468	-25.3452	2.6E-140	-0.28583	-0.2448	-0.28583	-0.2448
LN_LMED	0.266467837	0.031478	8.465275	2.66E-17	0.20477	0.328165	0.20477	0.328165
Bedrooms	-0.072841671	0.006666	-10.9277	9.54E-28	-0.08591	-0.05978	-0.08591	-0.05978
LN_FMR	0.545873658	0.022835	23.90464	3.6E-125	0.501115	0.590632	0.501115	0.590632
LN_UTILITY	-0.013634995	0.007241	-1.88312	0.059694	-0.02783	0.000557	-0.02783	0.000557
ROOMS	0.102739075	0.002907	35.34478	2.2E-268	0.097042	0.108436	0.097042	0.108436
ZSMHC	0.000217401	3.52E-06	61.77267	0	0.000211	0.000224	0.000211	0.000224
LN_OTHERCOST	0.232581694	0.003856	60.32102	0	0.225024	0.240139	0.225024	0.240139
ZADEQ	0.182283378	0.020526	8.880568	7E-19	0.142051	0.222515	0.142051	0.222515

- Summary: interpret the managerial meanings of the regression model

Conclusion:

A Regression Model for the Market Value of Housing Units (using data for Year 2013).

Main points to be covered in the summary...

a) We use the model that gives the best R-Square: **0.55**

b) Note that the 'Y' variable has a natural logarithm taken and also some of the 'X' variables have a natural logarithm taken. So the interpretations will be either semi-log or the log-log interpretations.

c) All p-values are below .05 (except LN_utility is close to .05), indicating that all coefficients are statistically significant.

Interpretation of coefficients

β_0: No managerially relevant interpretation, since talking about a situation when all 'X' variables are zero does not make managerial sense.

β_1: When the status of the housing unit is occupied, then the market value of the unit is lower by **36.3%,** all other variables being kept at the same level.

β_2: 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 **9.23%,** all other variables being kept at the same level.

β_3: Every unit is one year old, thus the value drops by **0.23%,** all other variables being kept at the same level.

β_4 When the housing unit is in the Northeast region of the country, then the market value tends to be lower by **12.85%** as compared to a similar housing unit being in the West region, all other variables being kept at the same level.

β_5 When the housing unit is in the Midwest region of the country, then the market value tends to be lower by **30.83%** as compared to a similar housing unit being in the West region, all other variables being kept at the same level.

β_6 When the housing unit is in the South region of the country, then the market value tends to be lower by **26.53%** as compared to a similar housing unit being in the West region, all other variables being kept at the same level.

β_7: For every one percentage increase in the area median income, the market value increases **by 0.27%,** all other variables remaining at the same level.

β_8: Every additional bedroom corresponds to a **7.28%** decrease in the market value of the housing unit, all other variables being kept at the same level.

β_9: For every one percentage increase in the fair market rent, the market value increases by **0.55%,** all other variables remaining at the same level.

β_{11}: For every one percentage increase in the monthly utility costs, the market value increases by **0.014%,** all other variables remaining at the same level.

β_{10}: One additional room corresponds to a **10.28%** increase in the market value of the housing unit, all other variables being kept at the same level.

β_{11}: For every one dollar increase in the monthly 'housing cost', the market value increases by **0.02%,** all other variables remaining at the same level.

β_{12}: For every one percentage increase in the monthly 'other' costs, the market value increases by **0.23%,** all other variables remaining at the same level.

β_{13}: When the status of unit adequacy is adequate , the market value increases by **18.23%** compared to not adequate , all other variables remaining at the same level.

Interpretation of R-square

The R-square and adjusted R-square are about **0.55**, indicating that the model explains about 55 percentage of variation in the market value of housing units.

Contributing factors to the marketing value of the housing unit:

If the housing unit is unoccupied, the region is in the west region of the city, it is not in the central city, the unit is not adequate, each more additional room will make the unit more expensive. However, one more additional bedroom depreciates the price. Thus, the factors of status of occupied or not, region, metros, unit adequacy and rooms contribute positive to the value of the housing unit and bedrooms contributes the negative influence.

The regression model:

*LN_housing_value = 4.1396 - 0.3630*Status - 0.0923*Metros - 0.0023*Age_of_unit - 0.1284*Northwest - 0.3083*Midwest - 0.2653*South + 0.2665*LN_LMED - 0.0728*Bedrooms + 0.5459*LN_FMR - 0.014*LN_Utility + 0.1027*ROOMS + 0.0002*ZSMHC + 0.2326*LN_Othercost + 0.1823*ZADEQ*

4. Predict future market value for the same properties.

- Use the data of 2011 to feed the regression model

STATISTICAL
SUMMARY
OUTPUT

Regression Statistics	
Multiple R	0.669505
R Square	0.448237
Adjusted R Square	0.447865
Standard Error	0.572834
Observations	20775

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	14	5533.997	395.2855	1204.631	0
Residual	20760	6812.152	0.328138		
Total	20774	12346.15			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.494377	0.311088	1.589185	0.112034	-0.11538	1.104134	-0.11538	1.104134
Northeast	-0.22801	0.013962	-16.3305	1.41E-59	-0.25537	-0.20064	-0.25537	-0.20064
Midwest	-0.4406	0.012199	-36.117	4.4E-277	-0.46451	-0.41668	-0.46451	-0.41668
South	-0.33311	0.011835	-28.1466	4.3E-171	-0.35631	-0.30992	-0.35631	-0.30992
LN_LEMD	0.959676	0.027891	34.4086	2.2E-252	0.905009	1.014344	0.905009	1.014344
LN_FMR	5.38E-05	1.5E-05	3.598681	0.000321	2.45E-05	8.31E-05	2.45E-05	8.31E-05
ZADEQ	0.168466	0.027752	6.070468	1.3E-09	0.114071	0.222862	0.114071	0.222862
ZSMHC	0.000181	4.62E-06	39.23348	0	0.000172	0.00019	0.000172	0.00019
Status	-0.42912	0.042444	-10.1103	5.65E-24	-0.51231	-0.34593	-0.51231	-0.34593
Age_unit	-0.00298	0.000165	-18.1146	8.89E-73	-0.0033	-0.00266	-0.0033	-0.00266
Bedrms	0.019441	0.007225	2.690728	0.007135	0.005279	0.033602	0.005279	0.033602
rooms	0.097001	0.003832	25.31181	3.1E-139	0.08949	0.104513	0.08949	0.104513
LN_utility	0.0109	0.009973	1.09295	0.274429	-0.00865	0.030447	-0.00865	0.030447
LN_othercost	0.152017	0.005635	26.97506	1.5E-157	0.140971	0.163063	0.140971	0.163063
Metro	-0.0636	0.009904	-6.4216	1.38E-10	-0.08301	-0.04419	-0.08301	-0.04419

- Carry out a handout analysis by selection 1000 data at random to evaluate the model by its average mean difference

Mean(abs)difference between predicted and actual housing prices of 2013	**Avaerage actual price of 2013**
32,475.30210	**257,050.00000**

(All the calculations are recorded in the excels of Q4.)

Conclusion:

A Regression Model to predict the Market Value of Housing Units in Year 2013.

Main points to be covered in the summary...

a) We use the same model from Assignment 3, except that the dependent variable is the 2013 Market Value while the independent variab;es are from the year 2011.

b) We could also use an additional independent variable which is the Market Value of the housing unit in year 2011.

c) We need to hold out 1000 housing units from the data and estimate the regression model on remaining data.

d) The coefficients from the regression model are to be then used to predict the Market Value in the 'Hold-out Data'.

c) As a metric of prediction, we need to calculate the Mean Absolute Deviation (MAD) for our predictions.

Interpretation of coefficients

Interpretation of coefficients is not required since a similar interpretation was done in Assignment 3.

Prediction in the 'Hold-out' Data

The regression model now has a R-square of 0.45 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 **$32,475.3**. This seems ok given that the average Market Value is around $**257,050.00.**

Please see the worksheet 'Statistical Tests' for various calculations