Housing data analysis

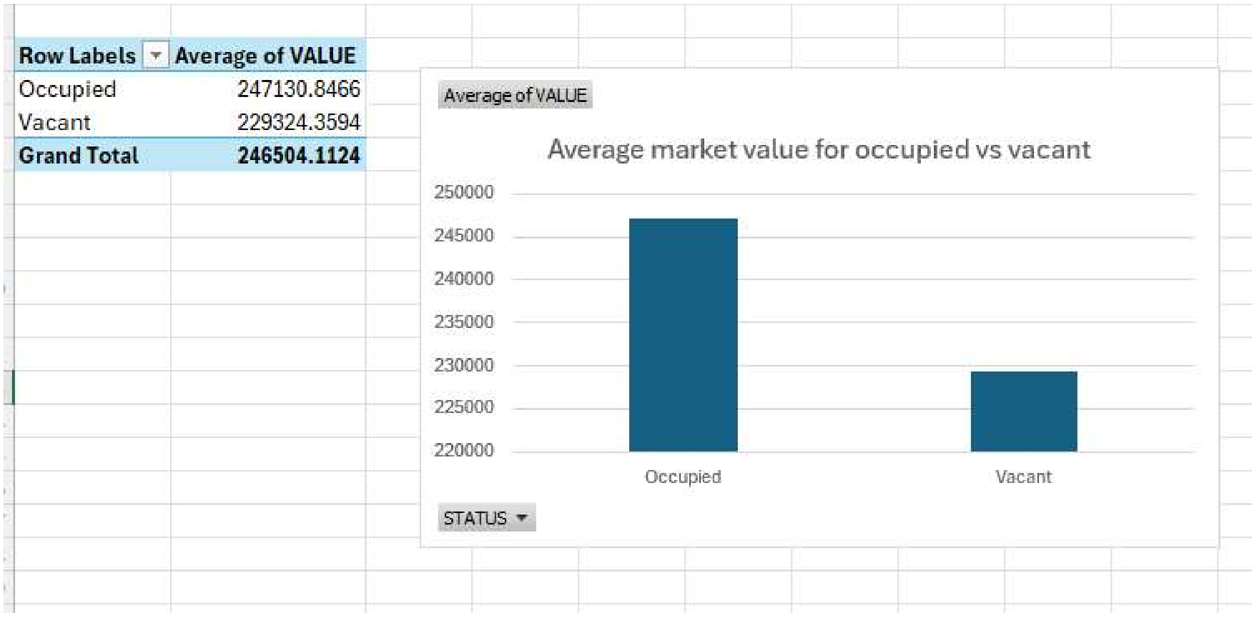
In this project I have done data cleaning, merging and analysis using Microsoft Excel on Housing affordability datasets of 2005, 2007, 2009, 2011, 2013 to answer some questions.

Source : https://www.huduser.gov/portal/datasets/hads/hads.html

1 - Are there some differences in the Market values of occupied versus not occupied housing units?

In 2005 :

Statistics	Occupied	Vacant
Average value	247,130.8466 $	229,324.3594 $
Max	1,540,794 $	1,540,794 $
Min	1000 $	1200 $
Count	29440	1074
stdev	281,859.6405	264,371.4834



Row Labels	Average of VALUE
Occupied	247130.8466
Vacant	229324.3594
Grand Total	**246504.1124**

Applying hypothesis testing

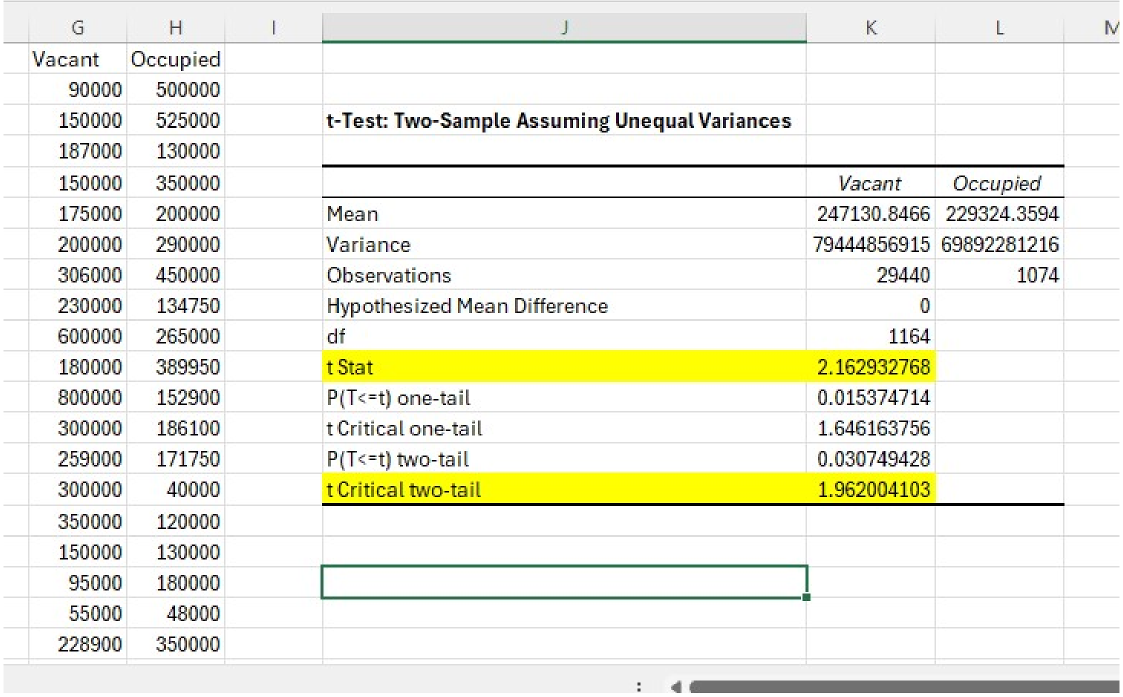
We'll compare the mean market value between the two groups using **two sample t-test**

Null hypothesis (H_0):
There is *no difference* in average market value between vacant and occupied houses.

$$H_0: \mu_{\text{vacant}} = \mu_{\text{occupied}}$$

Alternative hypothesis (H_1):
There *is* a difference.

$$H_1: \mu_{\text{vacant}} \neq \mu_{\text{occupied}}$$



Vacant	Occupied
90000	500000
150000	525000
187000	130000
150000	350000
175000	200000
200000	290000
306000	450000
230000	134750
600000	265000
180000	389950
800000	152900
300000	186100
259000	171750
300000	40000
350000	120000
150000	130000
95000	180000
55000	48000
228900	350000

t-Test: Two-Sample Assuming Unequal Variances

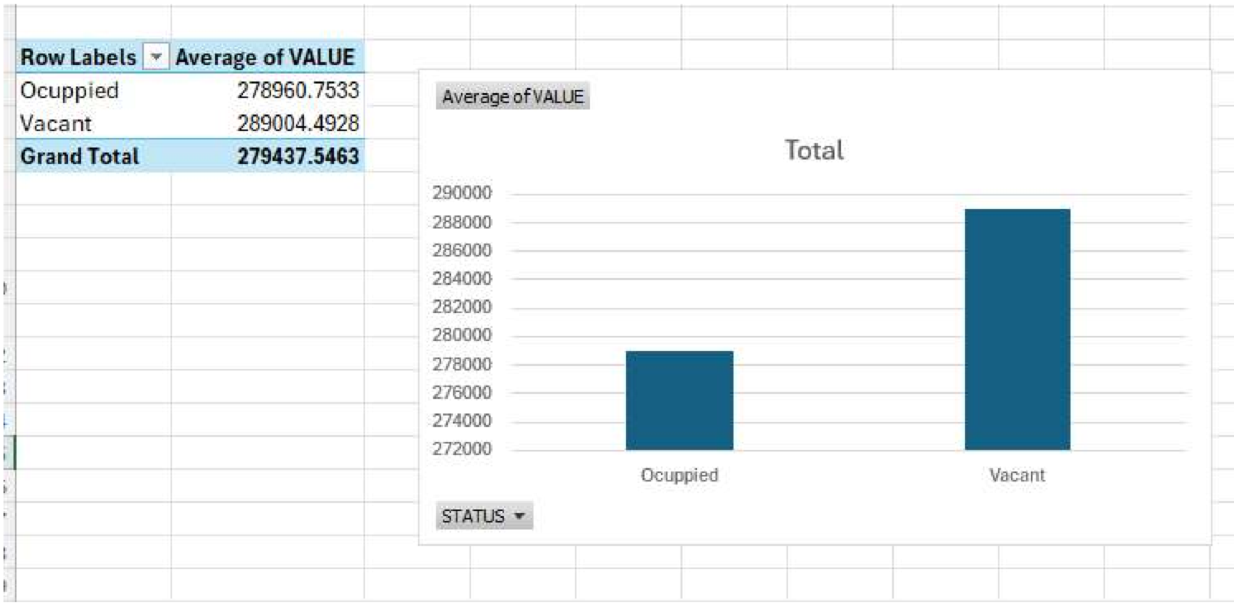
	Vacant	*Occupied*
Mean	247130.8466	229324.3594
Variance	79444856915	69892281216
Observations	29440	1074
Hypothesized Mean Difference	0	
df	1164	
t Stat	2.162932768	
P(T<=t) one-tail	0.015374714	
t Critical one-tail	1.646163756	
P(T<=t) two-tail	0.030749428	
t Critical two-tail	1.962004103	

t-stat is greater than t-critical, so we reject the null hypothesis.

In 2005 data There’s a statistically significant difference between vacant and occupied house values.

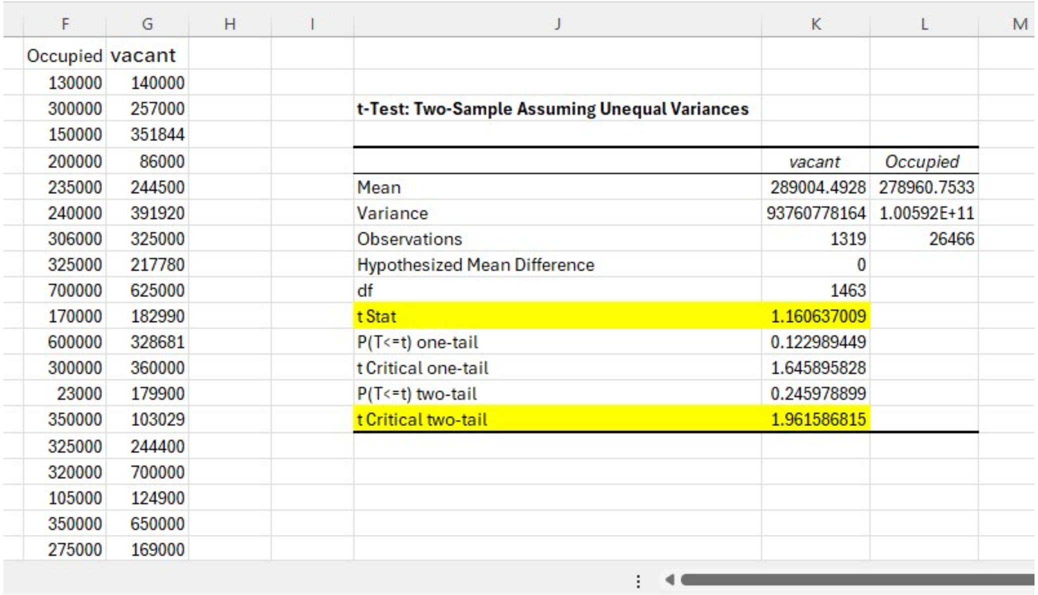
In 2007 :

Statistics	Occupied	Vacant
Average value	278,960.7533 $	289,004.4928 $
Max	1,829,479 $	1,829,479 $
Min	1000 $	1000 $
Count	26466	1319
stdev	317,162.7659	306,203.818



Row Labels	Average of VALUE
Ocuppied	278960.7533
Vacant	289004.4928
Grand Total	**279437.5463**

Applying hypothesis testing



Occupied	vacant
130000	140000
300000	257000
150000	351844
200000	86000
235000	244500
240000	391920
306000	325000
325000	217780
700000	625000
170000	182990
600000	328681
300000	360000
23000	179900
350000	103029
325000	244400
320000	700000
105000	124900
350000	650000
275000	169000

t-Test: Two-Sample Assuming Unequal Variances

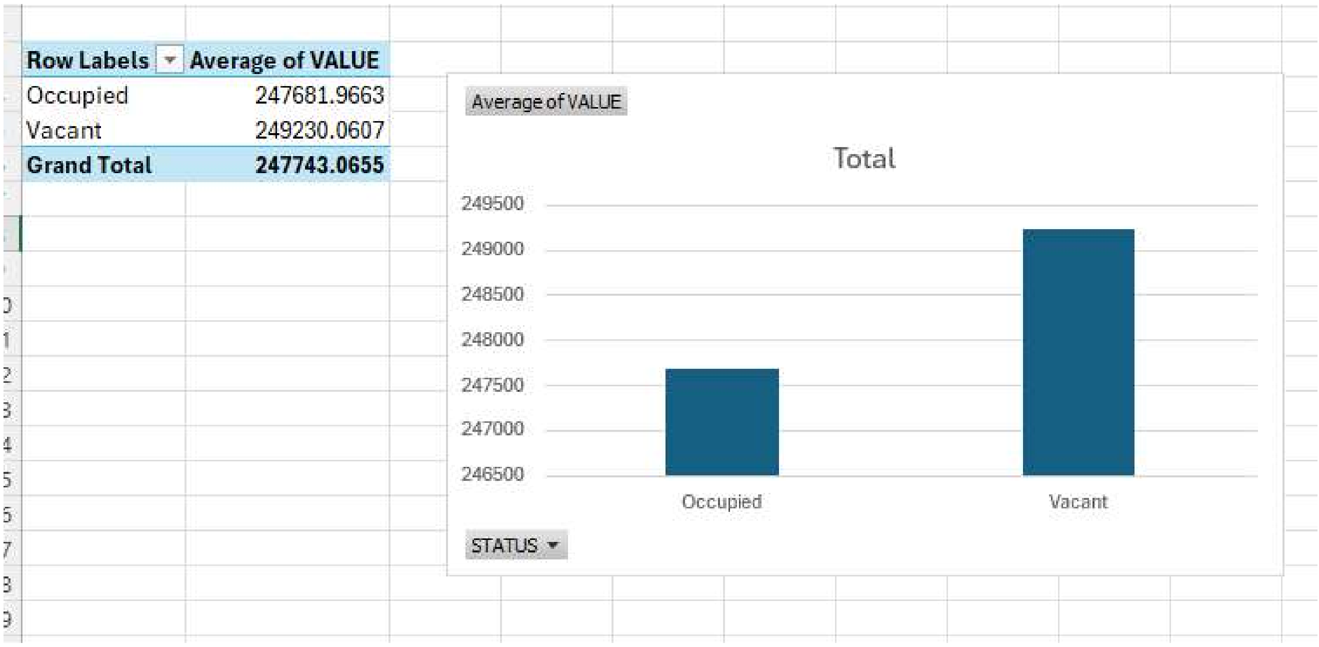
	vacant	*Occupied*
Mean	289004.4928	278960.7533
Variance	93760778164	1.00592E+11
Observations	1319	26466
Hypothesized Mean Difference	0	
df	1463	
t Stat	1.160637009	
P(T<=t) one-tail	0.122989449	
t Critical one-tail	1.645895828	
P(T<=t) two-tail	0.245978899	
t Critical two-tail	1.961586815	

t-stat is less than t-critical, so we will accept the Null hypothesis .

In 2007 data There’s no significant difference between vacant and occupied house values.

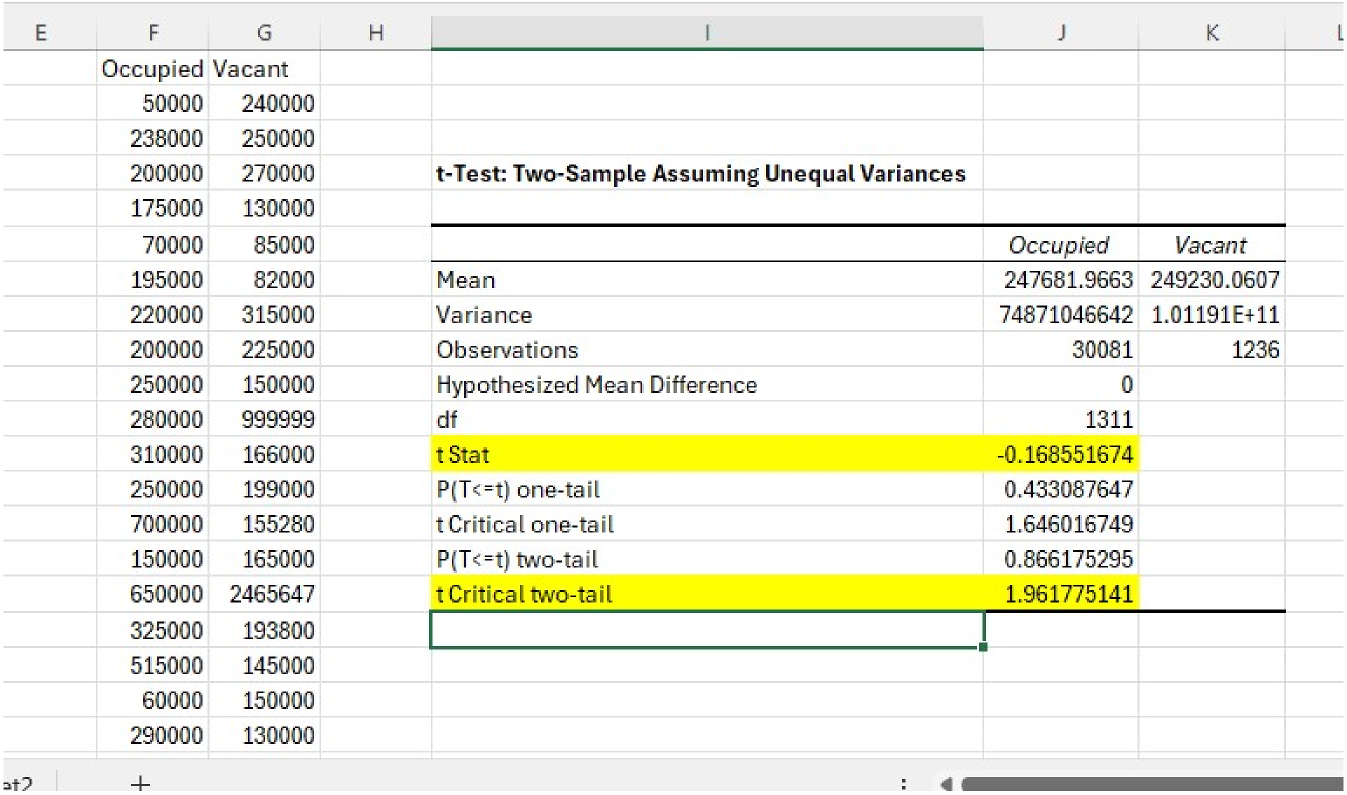
In 2009 :

Statistics	Occupied	Vacant
Average value	247,681.9663 $	249,230.0607 $
Max	2,465,647 $	2,465,647 $
Min	1000 $	1000 $
Count	30081	1236
stdev	273,625.7419	318,104.853



Row Labels	Average of VALUE
Occupied	247681.9663
Vacant	249230.0607
Grand Total	**247743.0655**

Applying hypothesis testing



Occupied	Vacant
50000	240000
238000	250000
200000	270000
175000	130000
70000	85000
195000	82000
220000	315000
200000	225000
250000	150000
280000	999999
310000	166000
250000	199000
700000	155280
150000	165000
650000	2465647
325000	193800
515000	145000
60000	150000
290000	130000

t-Test: Two-Sample Assuming Unequal Variances

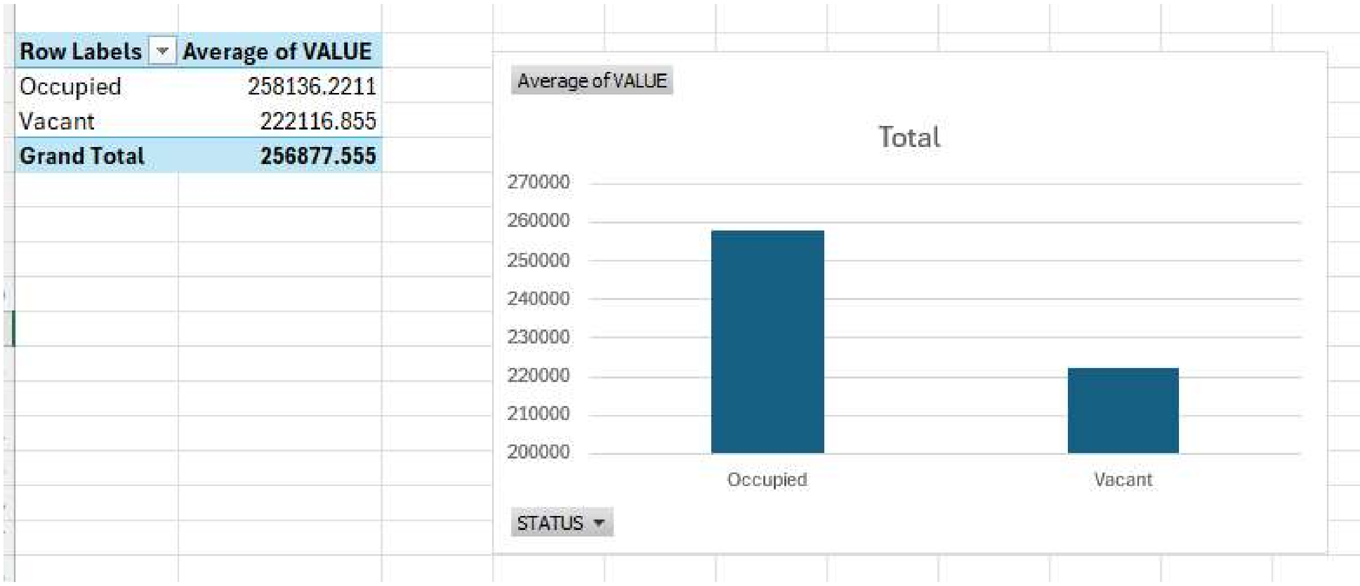
	Occupied	*Vacant*
Mean	247681.9663	249230.0607
Variance	74871046642	1.01191E+11
Observations	30081	1236
Hypothesized Mean Difference	0	
df	1311	
t Stat	-0.168551674	
P(T<=t) one-tail	0.433087647	
t Critical one-tail	1.646016749	
P(T<=t) two-tail	0.866175295	
t Critical two-tail	1.961775141	

t-stat is less than t-critical, so we will accept the Null hypothesis .

In 2009 data There’s no significant difference between vacant and occupied house values.

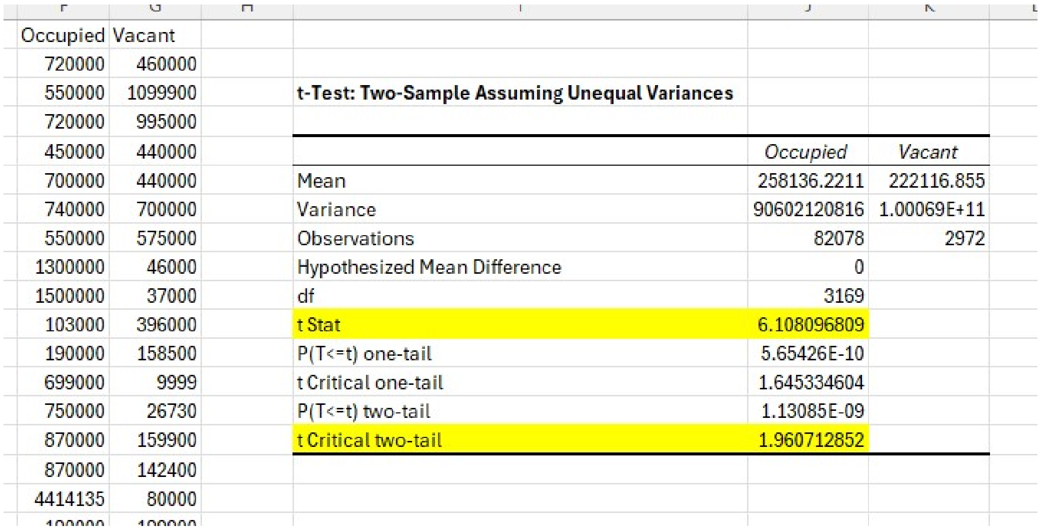
In 2011 :

Statistics	Occupied	Vacant
Average value	258,136.2211 $	222,116.855 $
Max	5,264,699 $	4,414,135 $
Min	1000 $	1000 $
Count	82078	2972
stdev	301,001.8618	316,336.8786



Row Labels	Average of VALUE
Occupied	258136.2211
Vacant	222116.855
Grand Total	**256877.555**

Applying hypothesis testing



Occupied	Vacant
720000	460000
550000	1099900
720000	995000
450000	440000
700000	440000
740000	700000
550000	575000
1300000	46000
1500000	37000
103000	396000
190000	158500
699000	9999
750000	26730
870000	159900
870000	142400
4414135	80000

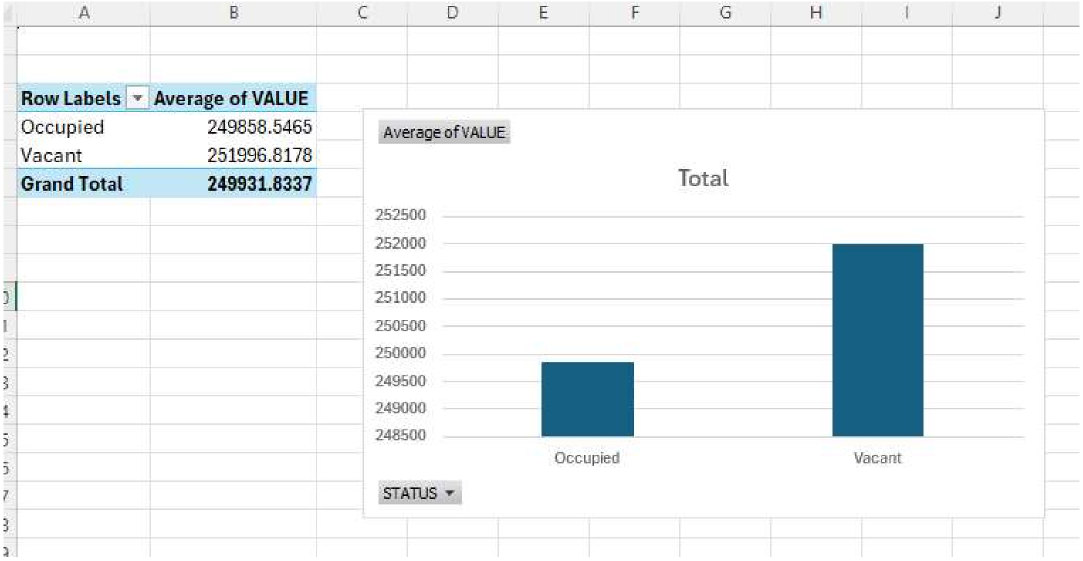
t-Test: Two-Sample Assuming Unequal Variances

	Occupied	*Vacant*
Mean	258136.2211	222116.855
Variance	90602120816	1.00069E+11
Observations	82078	2972
Hypothesized Mean Difference	0	
df	3169	
t Stat	6.108096809	
P(T<=t) one-tail	5.65426E-10	
t Critical one-tail	1.645334604	
P(T<=t) two-tail	1.13085E-09	
t Critical two-tail	1.960712852	

t-stat is greater than t-critical, so we reject the null hypothesis. In 2005 data There’s a statistically significant difference between vacant and occupied house values.

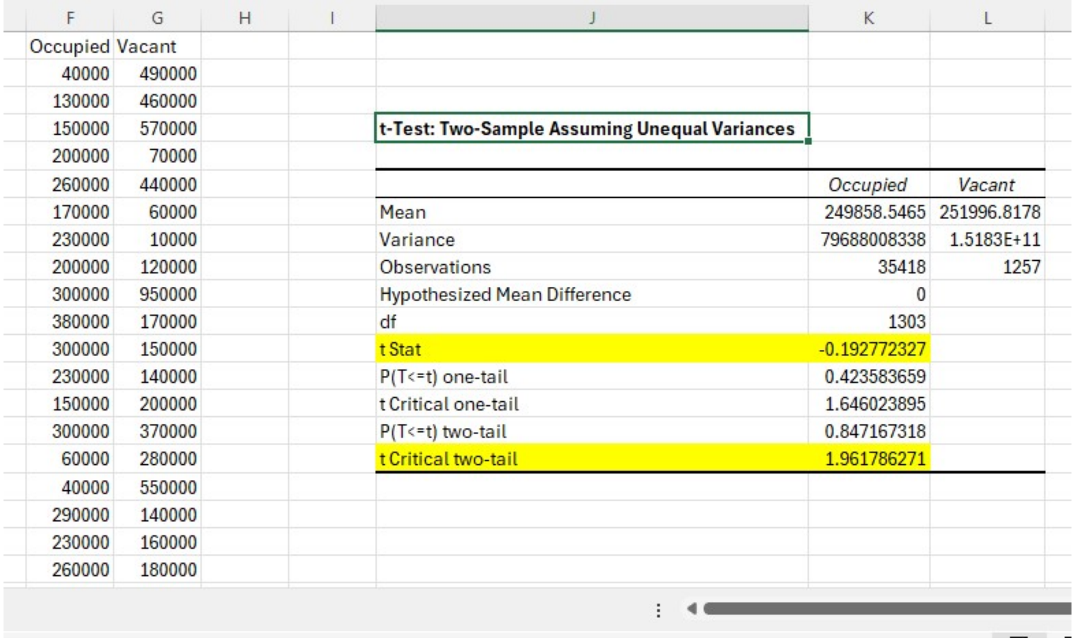
In 2013 :

Statistics	Occupied	Vacant
Average value	249,858.5465 $	251,996.8178 $
Max	2,520,000 $	2,520,000 $
Min	10,000 $	10,000 $
Count	35418	1257
stdev	282,290.6451	389653.0876



Row Labels	Average of VALUE
Occupied	249858.5465
Vacant	251996.8178
Grand Total	**249931.8337**

Applying hypothesis testing



Occupied	Vacant
40000	490000
130000	460000
150000	570000
200000	70000
260000	440000
170000	60000
230000	10000
200000	120000
300000	950000
380000	170000
300000	150000
230000	140000
150000	200000
300000	370000
60000	280000
40000	550000
290000	140000
230000	160000
260000	180000

t-Test: Two-Sample Assuming Unequal Variances

	Occupied	*Vacant*
Mean	249858.5465	251996.8178
Variance	79688008338	1.5183E+11
Observations	35418	1257
Hypothesized Mean Difference	0	
df	1303	
t Stat	-0.192772327	
P(T<=t) one-tail	0.423583659	
t Critical one-tail	1.646023895	
P(T<=t) two-tail	0.847167318	
t Critical two-tail	1.961786271	

t-stat is greater than t-critical, so we reject the null hypothesis. In 2005 data There's a statistically significant difference between vacant and occupied house values.

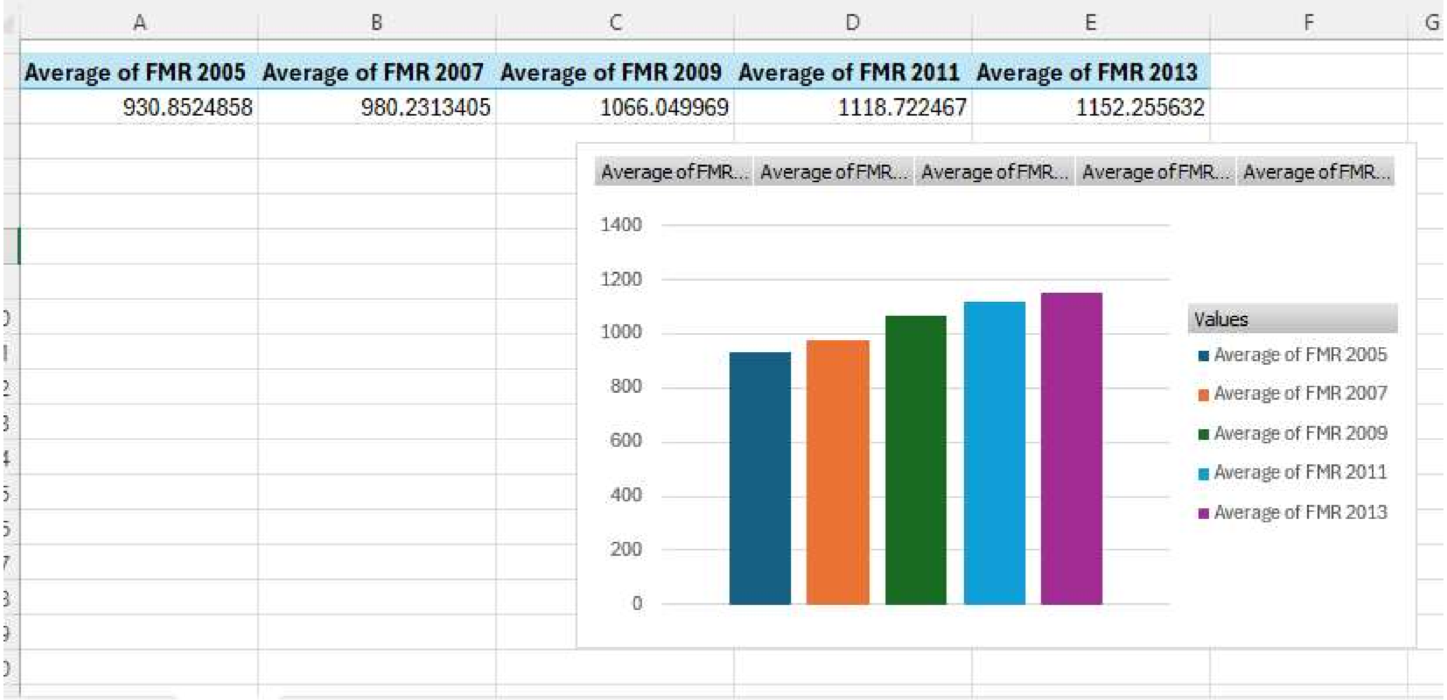
Summary

Difference in the Market Values is significant in 2005 and 2011. In these years the market value of Occupied units was greater than vacant units.

For the remaining years there is no significant difference in the market value across Occupied and vacant units

2 – Is housing rent was affected by the mortgage crisis occurred in the US in 2008?

To answer this question, we need to analyze data that precedes that year and data following that year.



Average of FMR 2005	Average of FMR 2007	Average of FMR 2009	Average of FMR 2011	Average of FMR 2013
930.8524858	980.2313405	1066.049969	1118.722467	1152.255632

From this data we see that the increase in FMR(Fair market monthly rent) average from 2007 to 2009 is significantly higher than the increase from 2005 to 2007, from 2009 to 2011, and from 2011 to 2013.

Applying hypothesis testing

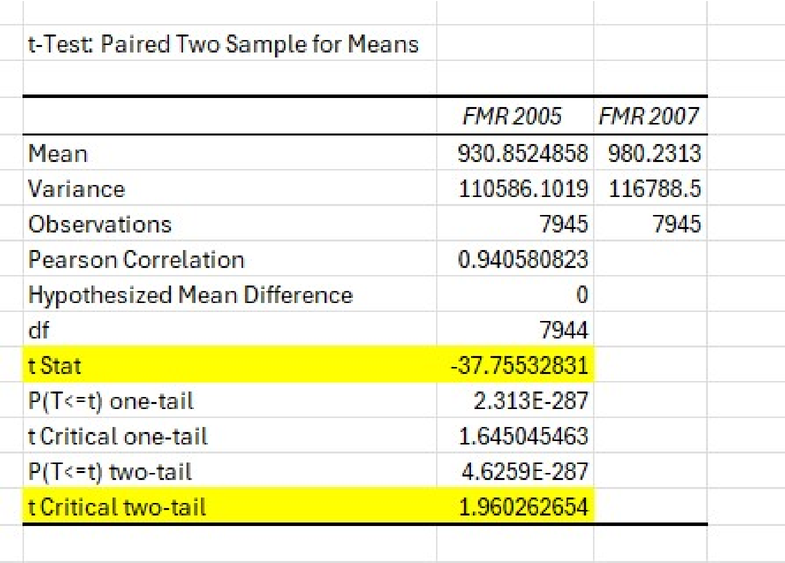
We'll compare the mean FMR value between each two consecutive years using **paired t-test**

Null hypothesis (H_0):
There is *no difference* in average FMR value for different years.

$$H_0: \mu_{\text{FMR7}} = \mu_{\text{FMR9}}$$

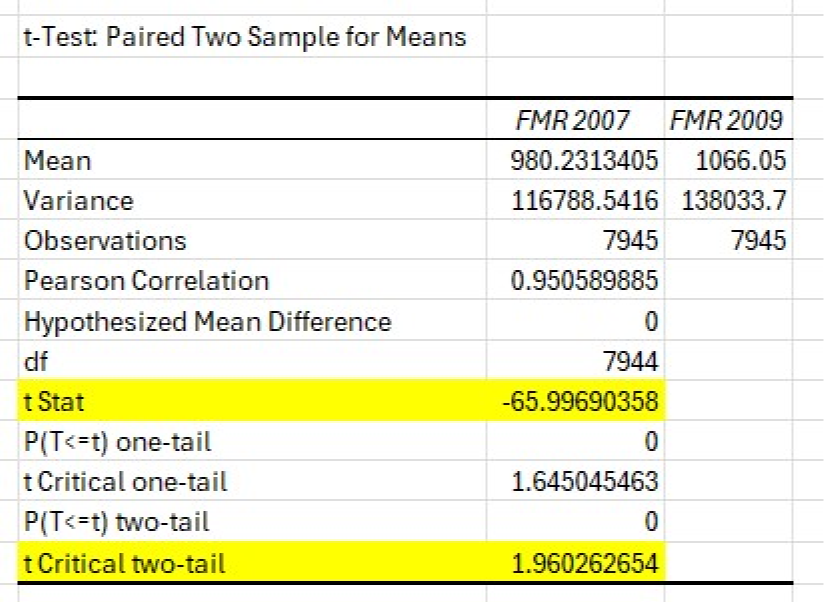
Alternative hypothesis (H_1):
There *is* a difference.

$$H_1: \mu_{\text{FMR7}} \neq \mu_{\text{FMR9}}$$



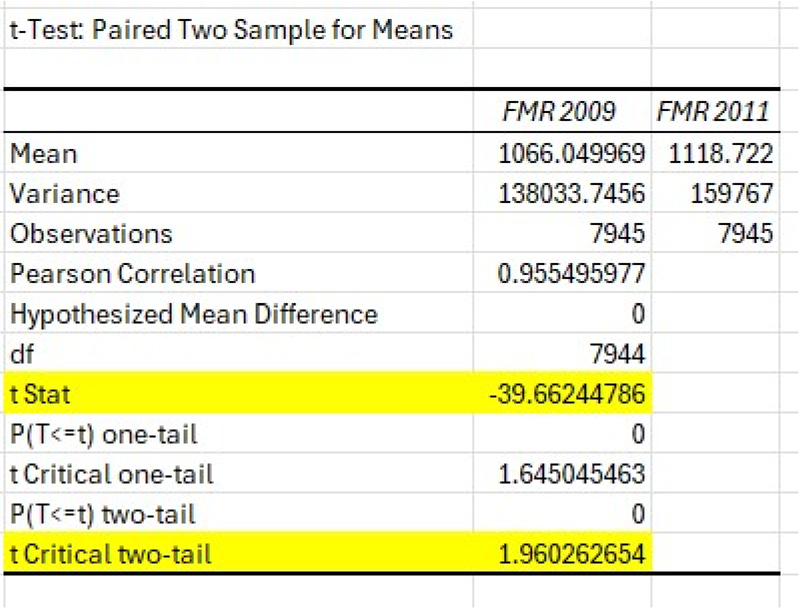
t-Test: Paired Two Sample for Means

	FMR 2005	*FMR 2007*
Mean	930.8524858	980.2313
Variance	110586.1019	116788.5
Observations	7945	7945
Pearson Correlation	0.940580823	
Hypothesized Mean Difference	0	
df	7944	
t Stat	-37.75532831	
P(T<=t) one-tail	2.313E-287	
t Critical one-tail	1.645045463	
P(T<=t) two-tail	4.6259E-287	
t Critical two-tail	1.960262654	



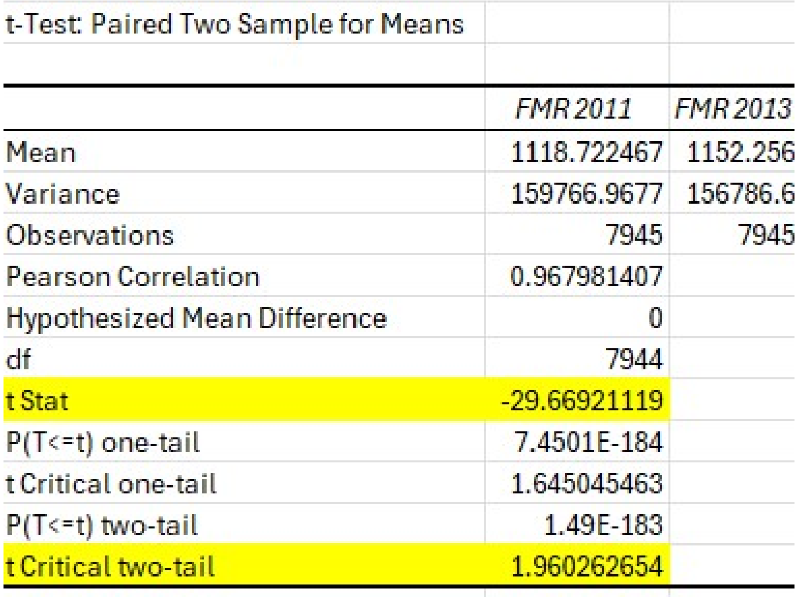
t-Test: Paired Two Sample for Means

	FMR 2007	*FMR 2009*
Mean	980.2313405	1066.05
Variance	116788.5416	138033.7
Observations	7945	7945
Pearson Correlation	0.950589885	
Hypothesized Mean Difference	0	
df	7944	
t Stat	-65.99690358	
P(T<=t) one-tail	0	
t Critical one-tail	1.645045463	
P(T<=t) two-tail	0	
t Critical two-tail	1.960262654	



t-Test: Paired Two Sample for Means

	FMR 2009	*FMR 2011*
Mean	1066.049969	1118.722
Variance	138033.7456	159767
Observations	7945	7945
Pearson Correlation	0.955495977	
Hypothesized Mean Difference	0	
df	7944	
t Stat	-39.66244786	
P(T<=t) one-tail	0	
t Critical one-tail	1.645045463	
P(T<=t) two-tail	0	
t Critical two-tail	1.960262654	



t-Test: Paired Two Sample for Means

	FMR 2011	*FMR 2013*
Mean	1118.722467	1152.256
Variance	159766.9677	156786.6
Observations	7945	7945
Pearson Correlation	0.967981407	
Hypothesized Mean Difference	0	
df	7944	
t Stat	-29.66921119	
P(T<=t) one-tail	7.4501E-184	
t Critical one-tail	1.645045463	
P(T<=t) two-tail	1.49E-183	
t Critical two-tail	1.960262654	

For all these tests t-stat value is higher than the critical value so we can reject the null hypothesis, Every year the FMR is significantly different from the previous year, but the most extreme value was the t-stat of 2007 vs 2009 before and after the crisis in 2008.