Question: Is there a difference in market value between occupied and vacant housing units? Values for processing:

- VALUE (market value)
- STATUS (occupied or vacant housing unit)

Analysis Plan:

1. Remove all suspicious market value figures (values less than $1000)
2. Collect statistical description of market value for occupied and vacant housing units
3. Interpret results and determine the presence of a difference between the market values of occupied and vacant housing units
4. Repeat the previous steps for all data sets
5. Make a final conclusion

Course of Analysis:

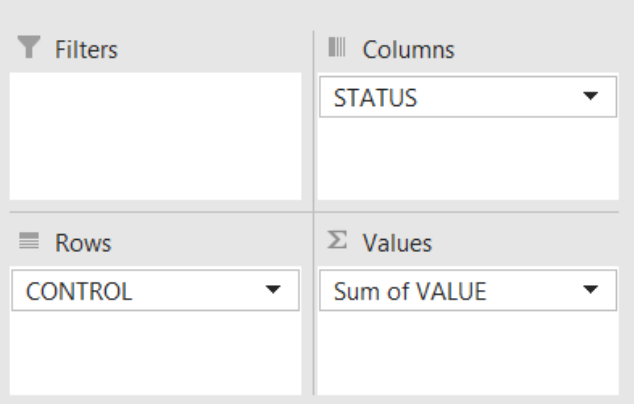
1. Removing Suspicious Values

За допомогою фільтрів для кожного набору даних було видалено усі рядки, у яких значення стовпця «VALUE» було менше за 1000$. Пробіли між рядками вдалось прибрати за допомогою сортування за стовпцем «CONTROL»

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CONTROL	AG	BEDRN	PI	REGIC	LMED	FM	IPOV	BUI	STATU	NUNI	TY	VALUE	ZINC2
5	'100007130148'	22	1	1	'3'	56785	519	9974	1980	'1'	16	1	-6	27040
6	'100007390148'	48	1	1	'3'	60308	600	9930	1985	'1'	32	1	-6	14000
8	'100008700141'	-9	2	-6	'4'	48751	702	-9	1980	'3'	8	1	-6	-6
9	'100009170148'	23	2	2	'2'	55986	546	12811	1985	'1'	24	1	-6	48000
22	'100028170140'	-9	1	-6	'3'	49799	531	-9	1980	'3'	24	1	-6	-6

2. Using Pivot Tables

Divided the data set into two separate columns: market value for occupied and market value for vacant housing units.



3. Data Separation

For each data set, transferred the values for occupied and vacant housing units to separate columns on the "Main" sheet.

	A	B
1	**Occupied**	**Vacant**
2	90000	500000
3	150000	525000
4	187000	130000
5	150000	350000
6	175000	200000
7	200000	290000

4. Creating an Excel Report

Created an Excel file for the report named Q1.excel. On the "Data descriptors" sheet, calculated basic statistical indicators for occupied and vacant housing units: average, median, mode, 5th and 95th percentile.

		Vacant						Occupied			
Year	Average	Median	Mode	5th percentile	95th percentile	Year	Average	Median	Mode	5th percentile	95th percentile
2005	229324	150000	154079	17000	653150	2005	247131	160000	200000	30000	700000
2007	289004.5	200000	182947	31800	725112	2007	278961	190000	200000	30000	750000
2009	248868.1	165000	200000	20800	745938	2009	247480	178000	200000	35000	650000
2011	222116.9	144450	200000	17998	625000	2011	258136	177000	200000	40000	740000
2013	251996.8	150000	150000	20000	750000	2013	249859	180000	150000	40000	650000

5. Forming Hypothesis

Based on the previously obtained data, the answer to the main question is not obvious. I have formulated a hypothesis that will help answer the question:

$$H_0: \mu_{Occupied} - \mu_{Vacant} = 0$$
$$H_A: \mu_{Occupied} - \mu_{Vacant} \neq 0$$

The hypothesis is based on the idea that the data we have is a **sample** from a larger **population**. Knowing its average, we know the sample average but not the population average.

According to the central limit theorem, the distribution of sample means approximates a normal distribution regardless of the population distribution. In other words, the sample average provides an **approximate** estimate of the population average. The basis of this hypothesis is that we can estimate with 95% probability whether the sample averages belong to the same true population value. If so, the difference is not statistically significant, with only a 5% probability that they differ.

6. Statistical Tests (sheet "Statistical tests")

Conducted hypothesis testing using a series of t-tests assuming different variances. The t-statistic was outside the confidence interval only twice.

Year 2005

t-Test: Two-Sample Assuming Unequal Variances

	Occupied	*Vacant*
Mean	247130.8	229324.4
Variance	7.94E+10	6.99E+10
Observations	29440	1074
Hypothesized Mea	0	
df	1164	
t Stat	2.162933	
P(T<=t) one-tail	0.015375	
t Critical one-tail	1.646164	
P(T<=t) two-tail	0.030749	
t Critical two-tail	1.962004	

Year 2011

t-Test: Two-Sample Assuming Unequal Variances

	Occupied	*Vacant*
Mean	258136.2	222116.9
Variance	9.06E+10	1E+11
Observations	82078	2972
Hypothesized Mea	0	
df	3169	
t Stat	6.108097	
P(T<=t) one-tail	5.65E-10	
t Critical one-tail	1.645335	
P(T<=t) two-tail	1.13E-09	
t Critical two-tail	1.960713	

7. Conclusion

Thus, only for 2005 and 2011, there is a statistically significant difference in the averages between occupied and vacant housing units.

8. Additional Hypothesis

Formulated another hypothesis questioning the nature of the difference: is the market value of occupied housing units higher or lower than that of vacant housing units?

$$H_0: \mu_{Occupied} - \mu_{Vacant} \geq 0$$
$$H_A: \mu_{Occupied} - \mu_{Vacant} < 0$$

9. Data Analysis

According to previous data**, in both years where the difference is statistically significant, the average of occupied housing units is greater or equal to the average of vacant housing units. The hypothesis that the average of occupied is less than the average of vacant was rejected.**

10. Graphs and Charts (sheet «Graphs and Charts»)

Created a series of line charts for the statistical indicators previously created:

11. Task of Creating a Graph of Relative Values

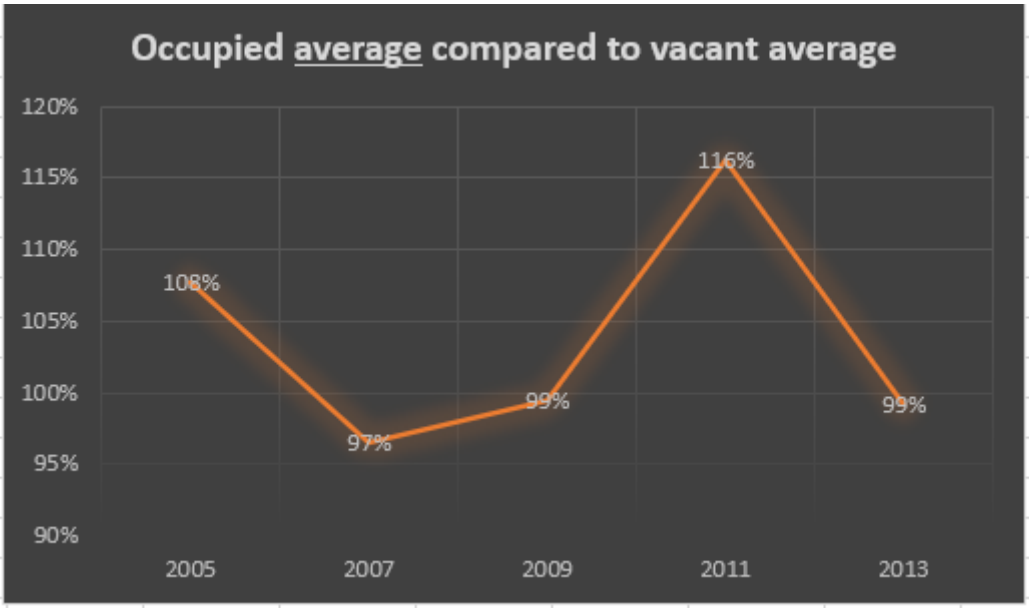
The aim was to visually indicate how much the market value of occupied homes was greater than that of vacant housing units. However, realized that absolute values do not clearly show the difference.

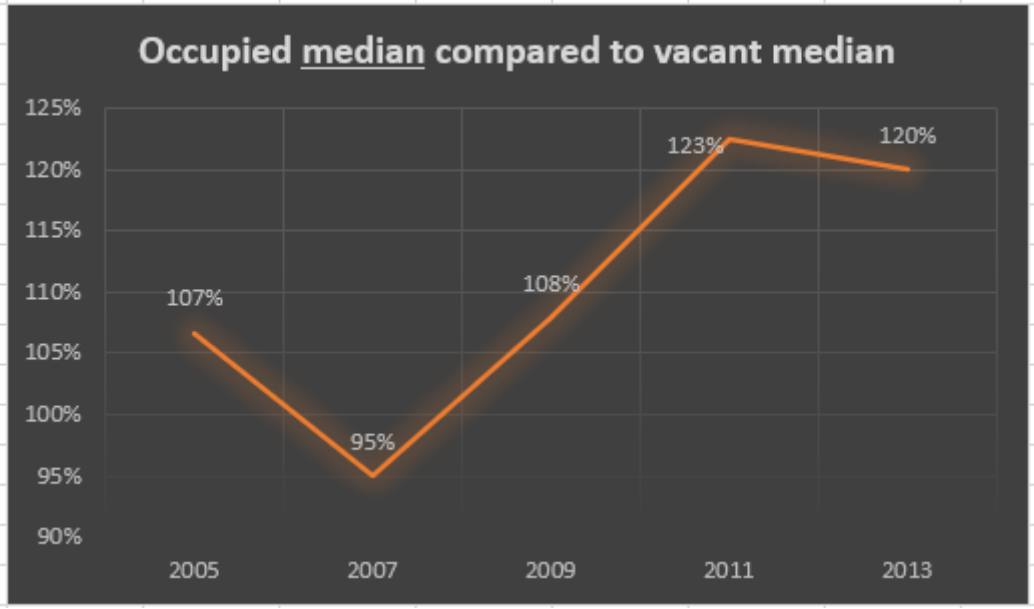
To accomplish this, it was necessary to normalize the indicators, which was done on the "Statistical descriptors" sheet.

Increase in **occupied** compared to vacant housing units				
Year	Average	Median	5th percentile	95th percentile
2005	108%	107%	176%	107%
2007	97%	95%	94%	103%
2009	99%	108%	168%	87%
2011	116%	123%	222%	118%
2013	99%	120%	200%	87%

Each indicator for occupied housing units was divided by the indicator of vacant housing units for the same year.

12. Graph Creation





The resulting graphs confirm the earlier conclusion: **A significant difference in averages exists only for 2005 and 2011.** Although the median fluctuated significantly throughout the entire time.

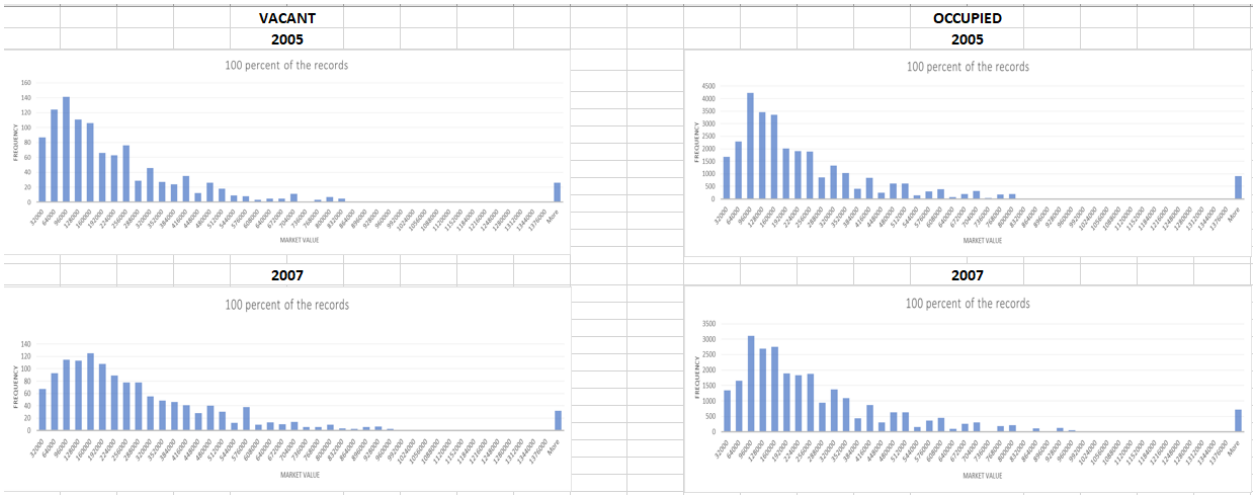
Further Analysis and Conclusion

The analysis could have concluded that only in 2005 and 2011 there is a difference in averages between occupied and vacant housing units, with the average of occupied being higher than vacant. However, the graph of medians inspired further research to find out why the average remains the same while the median fluctuates significantly from year to year.

1. Histograms of Market Value Distribution

Decided to create histograms of the market value distribution for occupied and vacant homes for each year on the "Vacant and Occupied 100%" sheet.

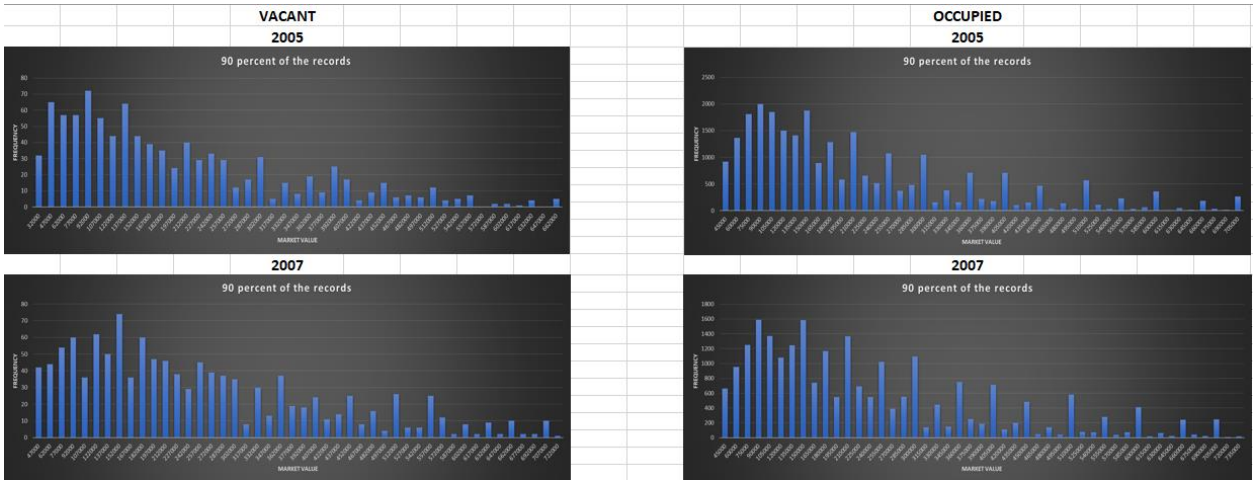
This involved a lot of routine work but also led to a better understanding of the problems of working with histograms and their comparison. To compare them, the grains must be the same or very similar. It's also important that the size of the histograms themselves is identical.



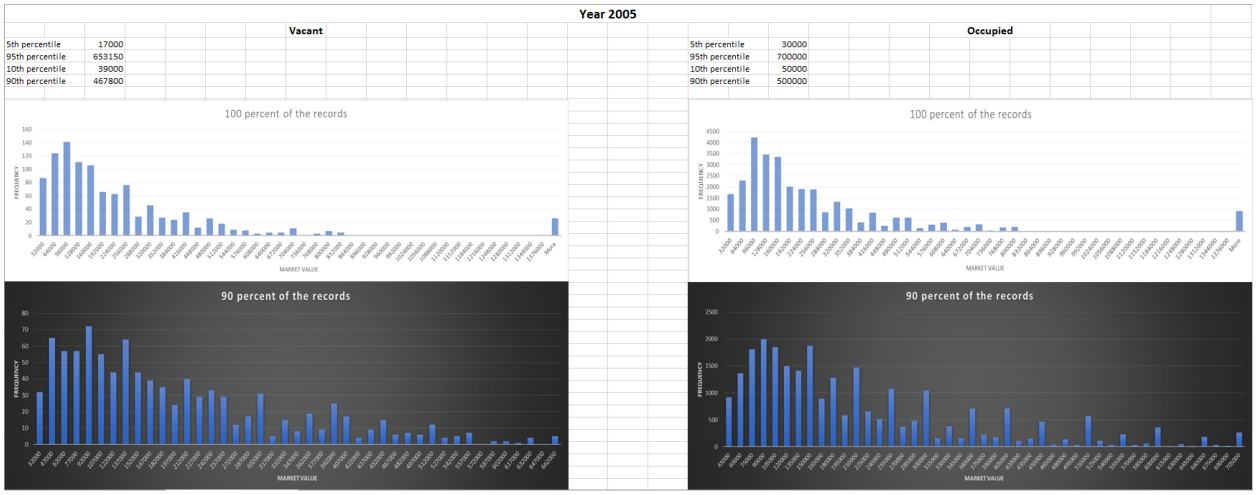
2. Histograms Analysis

The histograms across the entire data set only give a rough understanding of the distribution in the most demanded segments, which account for 90% of all observations.

Therefore, on the "Vacant and Occupied 90%" sheet, created histograms starting from the 5th to the 95th percentile, i.e., the 90% most frequent observations.



3. These histograms together are also present on the 'Histograms by years' sheet and in the file of each data set.



4. Conclusion

The median of occupied housing units in 2005 was 7% higher than the median of vacant housing units. In 2007, on the contrary, the median of occupied was 5% lower than the median of vacant housing units. And in 2009, the median of occupied was again higher by 8%.

However, these histograms show that the distribution of market value for occupied housing units from 2005 to 2009 did not change significantly. Therefore, **the fluctuations in the median are caused by changes in the market of vacant housing units.**