Question

Is there a difference in FMR (Fair Market Rent) over the years?

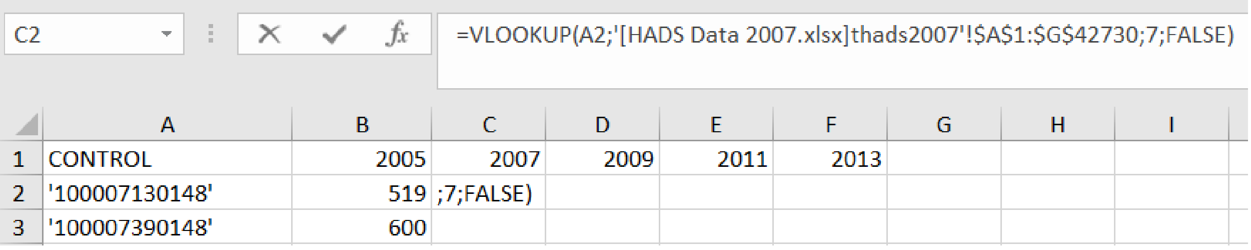
Analysis Plan

1. Merge the data on housing units across different years and retain only those that were observed in each year.
    a. Keep only the variables CONTROL and FMR.
    b. After merging, obtain values of only the variables CONTROL and FMR for all five years.
    c. Delete records that contain incorrect FMR values (missing or negative).
2. Provide descriptive statistics on FMR. Numerical as well as graphical.
3. Do appropriate analysis to compare the pairwise differences in Fair Market Rents (FMR) across the five years, 2005 through 2013.
4. Prepare a summary report.

Course of Analysis

1. Data merging and cleaning

- I copied all the CONTROL and FMR values from 2005 into the report file on the 'Data' sheet.
- For the columns of each subsequent year, I used the VLOOKUP formula: (set the search value as the CONTROL variable from 2005; set the search location as the data set table of the respective year; set the column number that points to FMR in the data set; used exact match).



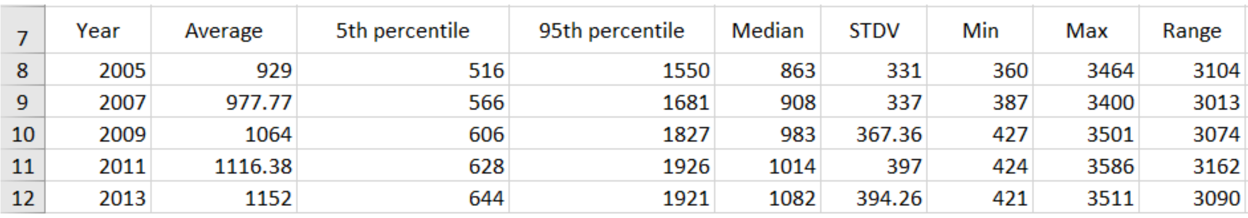
```
=VLOOKUP(A2;'[HADS Data 2007.xlsx]thads2007'!$A$1:$G$42730;7;FALSE)
```

	A	B	C	D	E	F	G	H	I
1	CONTROL	2005	2007	2009	2011	2013			
2	'100007130148'	519	;7;FALSE)						
3	'100007390148'	600							

- Using filters for each column, I identified all values that were missing or less than 0 and deleted them.
- After disabling the filters, I used sorting by the CONTROL column to remove the empty rows.

2. Gathering Numerical Descriptive Statistics

On the 'Data descriptors' sheet, I calculated for each year the following indicators: mean, median, 5th and 95th percentile, standard deviation, minimum and maximum values, and range. These are many measures of data dispersion, as FMR in this context can be considered a continuous type of data.



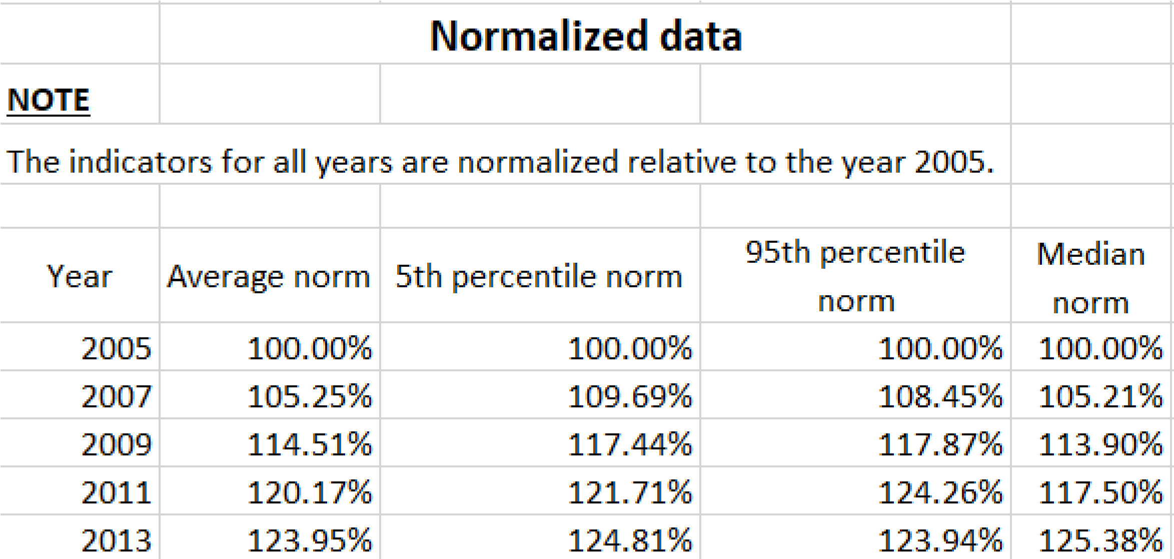
7	Year	Average	5th percentile	95th percentile	Median	STDV	Min	Max	Range
8	2005	929	516	1550	863	331	360	3464	3104
9	2007	977.77	566	1681	908	337	387	3400	3013
10	2009	1064	606	1827	983	367.36	427	3501	3074
11	2011	1116.38	628	1926	1014	397	424	3586	3162
12	2013	1152	644	1921	1082	394.26	421	3511	3090

These indicators allow for a quick and compact characterization of most of the data's features. Of course, for detailed analysis, they might not be entirely suitable, but they provide a good understanding of the general nature of the data.

3. Data Normalization

For further graphical representation, it was necessary to normalize the data to make the difference in indicators of different years visually apparent. For this purpose, I normalized the data for all indicators relative to the indicators of the year 2005.

I did not normalize the years relative to each other, as I needed to observe the long-term trend to understand the changes in the context of a larger time span. I did this by dividing each indicator by the corresponding indicator from the year 2005.



Normalized data

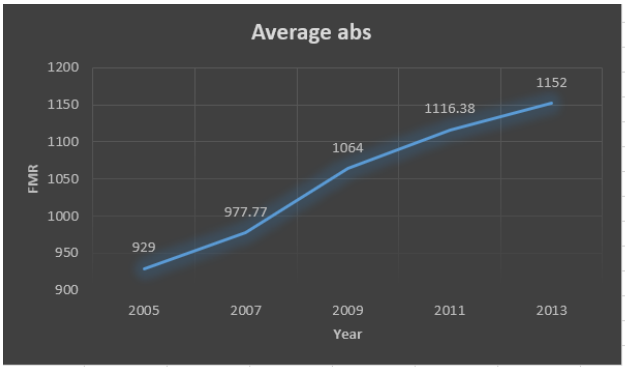
NOTE

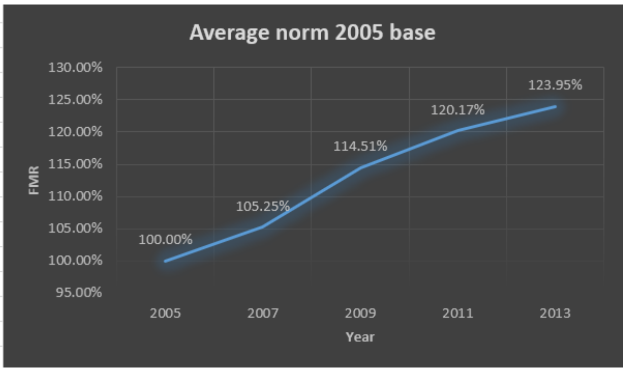
The indicators for all years are normalized relative to the year 2005.

Year	Average norm	5th percentile norm	95th percentile norm	Median norm
2005	100.00%	100.00%	100.00%	100.00%
2007	105.25%	109.69%	108.45%	105.21%
2009	114.51%	117.44%	117.87%	113.90%
2011	120.17%	121.71%	124.26%	117.50%
2013	123.95%	124.81%	123.94%	125.38%

4. Graphical Representation of Descriptive Statistics

For graphical representation, I used only line graphs for key descriptive indicators: mean, median, 5th, and 95th percentile. Pie charts or bar charts are not suitable here, as they are more focused on category analysis, which is not applicable for continuous data.





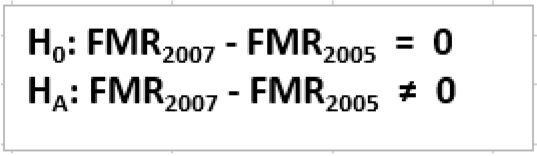
I did not use histograms, as they are quite difficult to compare with each other when dealing with a large number of data sets (and I have 5). I also did not use box plots for comparing distributions, as the question does not involve analyzing specific data segments.

The graphs can be found on the 'Graphs and Charts' sheet.

5. Comparing FMR Among Different Years

To determine whether the difference in FMR is statistically significant, I used paired t-tests.

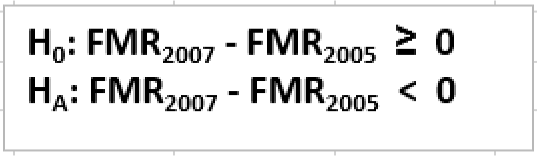
Formulated hypothesis:



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

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

I also formulated a hypothesis to identify the nature of the difference if it is statistically significant:



$H_0: FMR_{2007} - FMR_{2005} \geq 0$

$H_A: FMR_{2007} - FMR_{2005} < 0$

Starting from 2007, I compared each year with the previous one using paired t-tests. The results are contained in the “Pairwases” sheet.

6. Conclusion

In conclusion, each test refuted the first null hypothesis and confirmed the second null hypothesis. This means that **the difference in FMR between each year is statistically significant, with each year having a higher FMR than the previous one**.

Among all the changes, **the increase in FMR in 2007 is particularly notable**, being significantly greater than in other years. There was an almost 10% increase in the average, and the median increased by 8%, while changes in other years fluctuated within the range of 3-6%.

This period coincides with the mortgage lending crisis in the USA.

These conclusions are recorded on the 'Summary' sheet.

Problems Encountered During the Work

1. Merging Data from Five Tables: Initially, it was necessary to combine data from 5 tables into one, specifically for housing units observed in all years. After this, it was required to filter out records with incorrect FMR entries (negative or missing). While Excel allows this through filters, validating values for each year's column separately is quite routine.

2. Calculation of Averages, Medians, and Percentiles: Excel certainly provides tools for analyzing this statistics, but again, it is quite routine as you only need to change the column name from one formula to another.

3. Constructing Graphs: Initially, it was necessary to manually normalize the data due to Excel's limitations, and then to construct 2 graphs for each indicator (a total of 8 graphs). Although there were only 2 types of graphs, manually redefining values for each graph was also quite routine.

4. Further Analysis: For a deeper analysis, one could build 5 histograms to examine the distribution of FMR over the years. However, given the specifics of Excel, this is a time-consuming and routine activity.

5. Although Excel provides many tools for data analysis, working in it is quite routine, especially during detailed analysis. Python and its libraries for data analysis and neural networks offer much more in terms of automation capabilities.

Acquired Skills and Knowledge

- Understanding the Importance of Data Normalization: On a graph of absolute values, the extent of growth between values is often not obvious. Percentage values are more intuitively understandable and visually clear.

- Realizing the Importance of Evaluating Various Statistical Indicators Together: When several related indicators show the same trend, we gain more confidence in its validity. Including additional indicators that make the data 'cleaner' (by eliminating extreme values) allows for further verification of the initial assessment of the main indicators and possibly leads to additional conclusions. The more indicators analyzed, the more perspectives we have to view the data from, although it's important to remain judicious in this and choose only the most important for the current analysis.