Escola Tècnica Superior d'Enginyeria Informàtica

etsinf

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

STATISTICAL ANALYSIS OF THE FLATS FOR RENT MARKET IN WARSAW

Kacper Moll

Estadística (11539), English group

Valencia, 2022

Exercise 2

The starting data, which is used in this research, have been obtained from the Kaggle website. Below I present the link:

https://www.kaggle.com/beksultankarimov/warsaw-flat-rent-prices

Beksultan Karimov is an author of this dataset and he collected all data from otodom.pl – polish website to search the flat.

Original dataset contains 3472 observations with 85 columns each of them. Among the variables are: price (gross) per month for renting each flat, flat surface, district where the flat is located, year of built and a lot of others.

In the original dataset, each district formed a separate column containing as values:
0 - a flat is not located in this district;
1 - a flat is located in this district.
For the purpose of the task, all the district columns were merged into one. The same process was done with the columns concerning heating.

Variable name	Description		Symbol
area	Quantitative	numeric continuous variable representing the living area of the flat in in square meters	x_1
floor		a numeric, discrete variable, representing the number of the floor on which the flat is located	x_2
year_built		a numeric discrete variable representing the year in which the flat was built	x_3
gross_price		numeric, continuous variable, representing the gross price per month of renting the flat, price given in PLN	x_4
heating_type	Qualitative	a character variable indicating which type of heating is installed in the flat	F_1
district		a character variable representing the district in which the flat is located	F_2

Exercise 3:

Objectives of the work:

- Correlation between year of built and district, it should clearly show which area of the city is newer
- Comparison of flat prices by district
- Correlation between year of built and type of heating installed in that flat
- Comparison of rental price and the surface of the apartment
- Correlation between floor number, where the flat is located and price
- Comparison of flat surface and district, where the flat is located

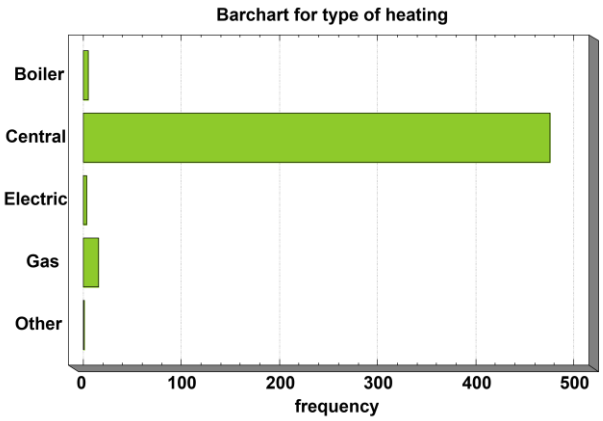
Exercise 4:

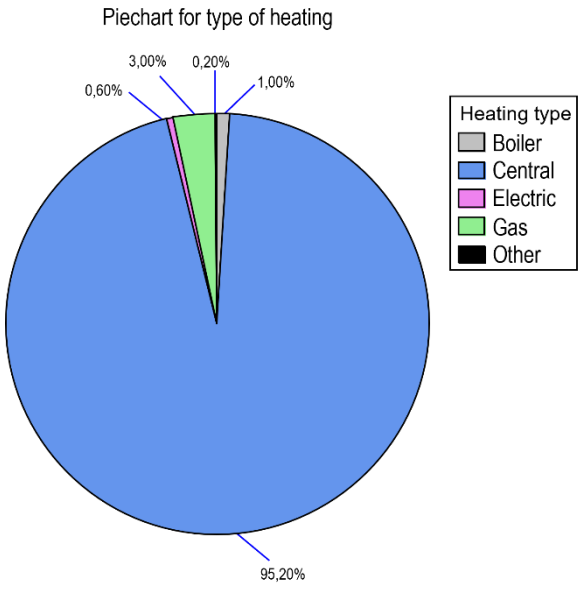
Discussion of the sample and population:

Since the population is a dataset of 3472 records, which present the situation of the market at a particular time (the exact information from what year the data is, was impossible to check, but the data cannot be older than 2 years, because it includes buildings built in 2021). The sample, I will work on, contains 500 observations. These records has been obtained by using random function in excel (as suggested in "Propuesta Trabajo Academico 2022 vs 1 febrer" file). Because of the random selection of observations, it is fair to say that the **selected sample is representative.**

Exercise 5:

Pie and Bar chart of heating type variable:





It may be observed from the provided sample that the most frequent type is central heating. This type of heating is installed in 95,2% percent of flats available for rent. The next most common type of heating is heating by using gas - 3%. Percentage of the other heating types reached less than 1% of the total sample. Other

heating types, which include boiler heating, electric heating and other heating types, reached the following values, respectively: 1%, 0,6% and 0,2%.

Exercise 6:

A frequency table of the variable F2

Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
Bemowo	24	0,048	24	0,048
Bialoleka	30	0,06	54	0,108
Bielany	38	0,076	92	0,184
Mokotow	139	0,278	231	0,462
Ochota	39	0,078	270	0,54
Praga-Polnoc	22	0,044	292	0,584
Praga-Poludnie	34	0,068	326	0,652
Srodmiescie	67	0,134	393	0,786
Ursynow	31	0,062	424	0,848
Wola	76	0,152	500	1

The table shows how many times each district value occurred (frequency column), as well as the percentages (relative frequency column) and total statistics (last two columns). For example, in 24 rows of the sample, the district was equivalent to **Bemowo**. This represents **0,0480 = 4,8%** of the 500 values in the sample. The last two columns give the total numbers (cumulative frequency) and percentages (cumulative relative frequency) from the top of the table downwards.

Exercise 7:

	Bemowo	Bialoleka	Bielany	Mokotow	Ochota	Praga-Polnoc	Praga-Poludnie	Srodmiescie	Ursynow	Wola	Row Total
Boiler	0	0	2	1	0	1	1	0	0	0	5
	0,00%	0,00%	5,26%	0,72%	0,00%	4,55%	2,94%	0,00%	0,00%	0,00%	1,00%
Central	23	24	36	137	35	21	32	64	30	74	476
	95,83%	80,00%	94,74%	98,56%	89,74%	95,45%	94,12%	95,52%	96,77%	97,37%	95,20%
Electric	0	1	0	0	1	0	0	1	0	0	3
	0,00%	3,33%	0,00%	0,00%	2,56%	0,00%	0,00%	1,49%	0,00%	0,00%	0,60%
Gas	1	5	0	1	2	0	1	2	1	2	15
	4,17%	16,67%	0,00%	0,72%	5,13%	0,00%	2,94%	2,99%	3,23%	2,63%	3,00%
Other	0	0	0	0	1	0	0	0	0	0	1
	0,00%	0,00%	0,00%	0,00%	2,56%	0,00%	0,00%	0,00%	0,00%	0,00%	0,20%
Column Total	24	30	38	139	39	22	34	67	31	76	500
	4,80%	6,00%	7,60%	27,80%	7,80%	4,40%	6,80%	13,40%	6,20%	15,20%	100,00%

The purpose of compiling these two variables is to examine the occurrence of a particular type of heating per specific district. Such a result can only be achieved by applying "**percentages per row**", that is why percentage per row provides more information in this table.

The **absolute frequency** describes the number of times the particular value for a variable has been observed and **relative frequency** is an absolute frequency in **relation** to the total number of values for that variable.

This table shows how often the 5 different heating types occur together with each of the 10 districts. The first number in each cell of the table is the count or frequency. For example, there were 23 times when type of heating equaled Central and district equaled Bemowo. The second number shows that cell's percentage of the column in which it falls. For our example it represents 95,83% of the 24 times when district equaled Bemowo.

Exercise 8:

Main statistics of the four numerical variables

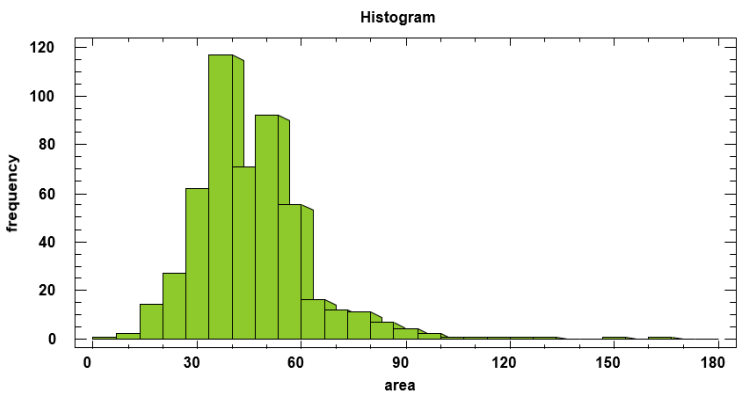
	area	floor	year of built	gross price
Range	163	10	118	8627
Interquartile range	17	4	28,5	2749,04
Mean	45,436	3,578	1995,4	764
Median	42	3	2003	2600
Variance	319,0224654	7,506942414	604,1764	832942,7996
Standard deviation	17,8612	2,73988	24,58	912,657
Coeff. of variation	39,31%	76,58%	1,23%	33,20%
Coefficient of skewness	18,5031	10,8206	-11,0094	23,632
Coefficient of kurtosis	38,7657	2,75014	3,63471	52,5494

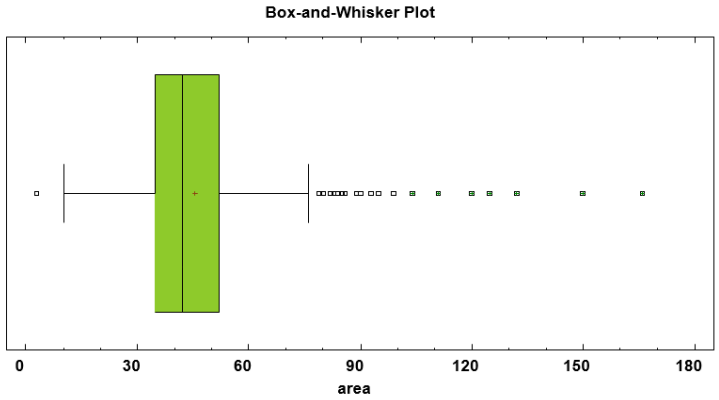
Parameters of:

- Position: mean, median;
- Dispersion: variance, standard deviation and coefficient of variation;
- Shape: coefficient of skewness and coefficient of kurtosis

Exercise 9:

Histogram and box-whisker plot of the variable X1





HISTOGRAM

1. X axis represents the area in square meters given and Y axis represents the number of flats with the given area.
2. The number of intervals should be lower, currently this number is equal 27 and shape of histogram is not proper, with the number of intervals equal 15, shape looks much better.
3. One of the advantages of a histogram is that each successive bar meets the vertical side of the previous one. This ensures that we do not miss data from any interval within the range of the data. On the other hand, the histogram allows for a subjective choice of the number of intervals, which is associated with the ambiguity of the shape of the diagram, so that conclusions can be manipulated.

BOX-WHISKER GRAPH

1. In a box-whisker graph, the box represents the difference between the first and third quartiles - the interquartile range. The mean and median are also visible in the box.
2. One of the advantages of box-whisper graph is fact that organising data into a box chart using five key concepts (median, the upper and lower quartiles, minimum and maximum data values) is an effective way to deal with big data that cannot be organised using other charts. However, the box plot does not keep the exact values and details of the distribution.

The histogram is the graph that best represents the characteristics of the distribution of a variable, as it clearly shows which range of flat's size is superior to the others.

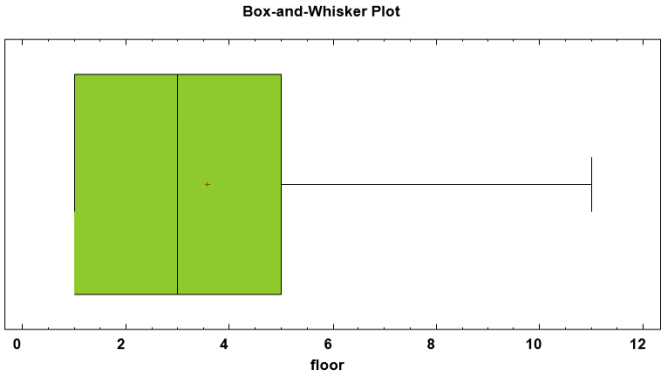
Exercise 10:

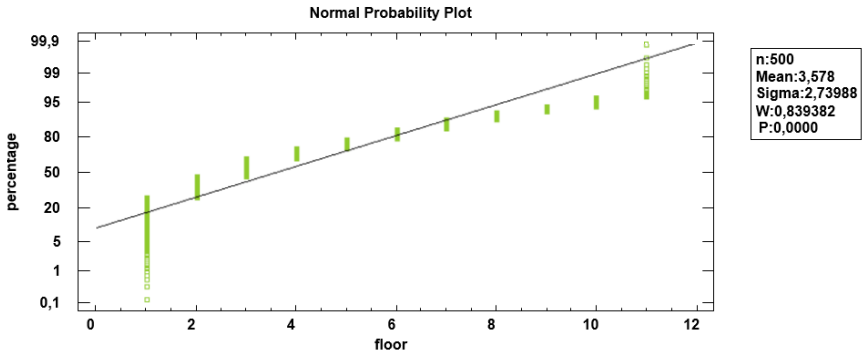
The area variable represents the size (in square meters) of each flat. In that sample:

- First quartile = 35
- Third quartile = 52
- Median = 42
- Average = 45,436
- Minimum value = **9,5** (35-1,5*17)
- Maximum value = **77,5** (35-1,5*17)
- Distribution: slightly asymmetric (slightly positively skewed)
- Data has a lot of outliers for that variable, what is easy to see from box-whisper graph. It means that these points should be discarded.

Exercise 11:

Box-and-whisker plot and normal probability plot of the variable X2





- The normal probability plot is formed by:
  - Vertical axis: percentile the data
  - Horizontal axis: values from dataset (indicating specific floors)

- One advantage of this technique is that you can see all outliers. Drawbacks may include that it is difficult to deduce which values are the most common.
- In that case normal probability plot gives us more information. Because from box-whisker plot we can deduce just more general information.

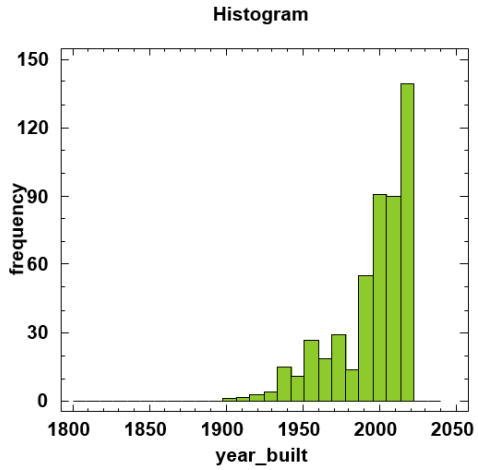
Exercise 12:

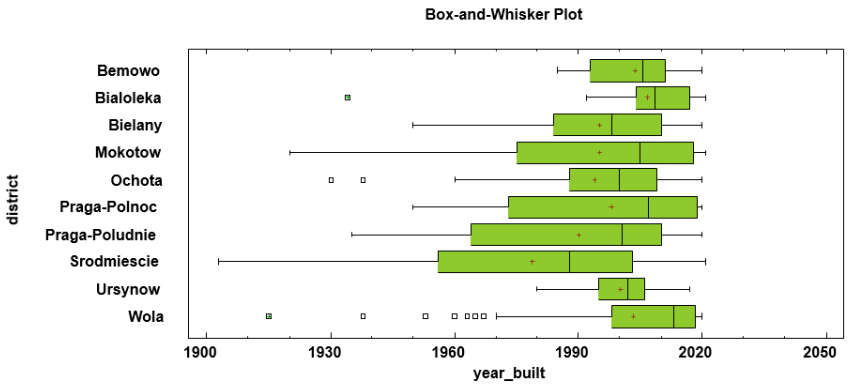
The most appropriate position and dispersion parameters for the variable X2

- Range = 10
- Interquartile range = 4 (Q3-Q1=5-1=4)
- Average = 3,578
- Median = 3
- Standard Deviation = 2,73988

Exercise 13:

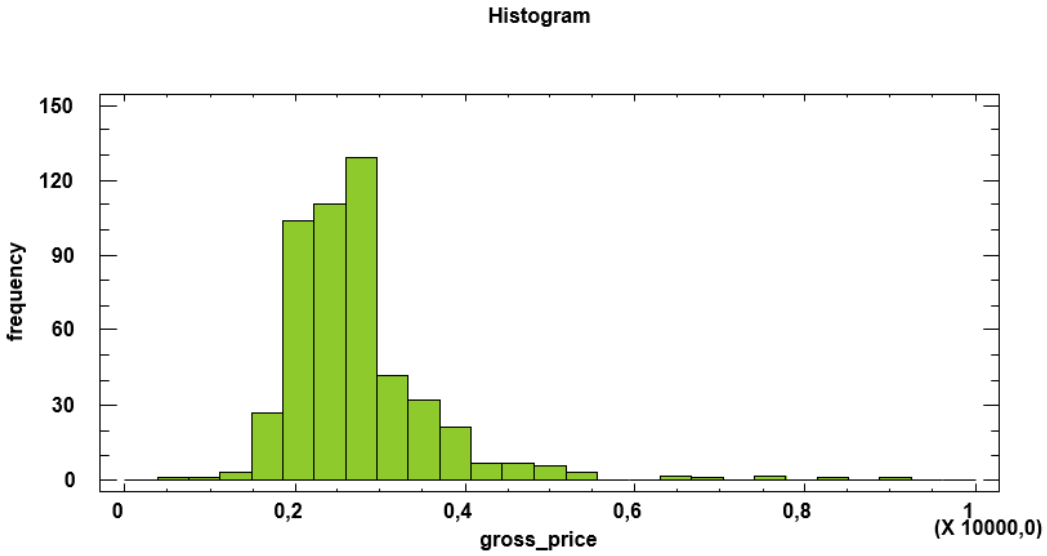
Histogram of the variable X3 and a multiple box-plot, as a function of the factor F2





Exercise 15:

Graphical description of variability of the variable X4



A histogram was chosen to represent the distribution of the given variable. This type of graph shows well which prices dominate in the market, which prices the market focuses on.

It is easy to see that the vast majority of flat prices are within the range of about PLN 1900 to PLN 3000.

A box-whisper plot would be best for determining what data should be disregarded when drawing conclusions from the survey, but it is also easy to see from the histogram that flats above PLN 6000 are the exception.