COURSERA IBM MACHINE LEARNING – EDA – COURSE PROJECT

Author: Wojciech Gołębiowski

1. Brief description of the data set and a summary of its attributes

Dataset used in analysis presents flats available to buy in one of the neighbourhoods in Wrocław, Poland. In general the datasets consists of 182 records and presents variables such as:

- Offer title
- Price (PLN)
- Year built
- Area (m^2)
- Market (primary/secondary)
- Number of rooms
- Floor

Below you can find a summary of numeric attributes of the dataset:

	price	year_built	area_m2	rooms	floor
count	110.00	182.00	182.00	182.00	182.00
mean	1286511.76	2021.40	54.84	2.55	2.07
std	5841356.43	0.65	5.84	0.53	1.06
min	401436.00	2020.00	45.36	2.00	0.00
25%	448706.75	2021.00	49.77	2.00	1.00
50%	500633.50	2021.00	53.45	3.00	2.00
75%	537100.00	2022.00	60.28	3.00	3.00
max	44014575.00	2023.00	64.96	4.00	4.00

Figure 1. Summary of statistics

Mean price of a flat equals 1 286 511 PLN and it differs from the median of a price which is close to 500k. Data regarding year when the flat was built is coherent – all of the buildings were/will be built in a period 2020-2023. The area of a flats is very consistent as well – mean area in square meters equals 54.84 and standard deviation for this variable equals 5.84.

Number of rooms oscillates around 2 to 4 rooms per flat. When it comes to floor this number oscillates around 0 to 4 floor. Blocks built in this neighbourhood seems to be quite short.

2. Initial plan for data exploration

The plan of data exploration will contain the following steps:

- Initial transformation of variables
- Searching for duplicates
- Examining missing data and potential data imputation process (as present in the Figure 1 there is one variable (price) which seems to be missing many records of the data
- Examining outliers (as per Figure 1 the variable 'price' seems to contain outliers as there is a huge difference between mean and median. Additionally standard deviation for this variable is relatively big)
- Feature engineering – adding new columns and transforming existing ones.

3. Actions taken for data cleaning and feature engineering

First step was to initially transform data in variables as it came in different formatting. Problems which were encountered are: adding texts to numeric values which made Python read these variables as an objects. Using commas rather than points to separate decimal part. The necessary changes were applied to variables: floor, area_m2, year_built and price.

```
flats['floor'] = [int(0) if x == 'parter' else int(x) for x in flats.floor]
flats['area_m2'] = [float(x.replace(",",".").split(" ")[0]) for x in flats.area_m2]
flats['year_built'] = flats['year_built'].astype('int64')
flats['price'] = [float(x.replace('Zapytaj o cenę', '0').replace(" zł","").replace(" ","").\
                        replace(",","")) for x in flats.price]
flats.loc[flats['price']==0.0, 'price'] = np.nan
```

Second step was to look for duplicates. It turns out there are 3 duplicated entries in our dataset.

```
flats.loc[flats.duplicated(keep = False),:].sort_values(['title'])
```

	title	price	year_built	area_m2	market	rooms	floor
36	Jagodno / 2M, możliwe 3 / Bezpiecznie i Towarz...	442485.0	2021	50.14	pierwotny	2	2
85	Jagodno / 2M, możliwe 3 / Bezpiecznie i Towarz...	442485.0	2021	50.14	pierwotny	2	2
35	Jagodno / 3M, możliwe 4 / Bezpiecznie i Towarz...	499855.0	2021	59.33	pierwotny	3	1
86	Jagodno / 3M, możliwe 4 / Bezpiecznie i Towarz...	499855.0	2021	59.33	pierwotny	3	1
68	Olsztyn // 2 Pokoje // Balkon // Spokojna Oko...	44014575.0	2021	51.63	pierwotny	2	4
97	Olsztyn // 2 Pokoje // Balkon // Spokojna Oko...	44014575.0	2021	51.63	pierwotny	2	4

Figure 2. Duplicated entries

Interestingly one of the duplicated entries is one which has the maximum value of a price in the dataset. That might be one of the reasons of huge difference between mean and median and big value of standard deviation. All duplicated entries were removed from the dataset using drop_duplicates() function.

Third step was to examine missing data. In figure 3 I present the number of missing records for each of the analysed variables.

```
flats.isnull().sum()
```

```
title          0
price         72
year_built     0
area_m2        0
market         0
rooms          0
floor          0
dtype: int64
```

Figure 3. Missing values in variables

Price variable is the only one containing missing entries. There are 72 of them which equals 40% of analyzed data. This is a big proportion. Missing entries are caused by the flat offers which in price fields contained phrase “ask for bid” and then at the 1st step of data cleaning that was converted to NaN value. As this proportion of data is quite significant I will not remove these entries from dataset. As all of the analyzed flats are similar in terms of location, number of flats and area in square meteres I will imput the data using median method having in mind that this brings additional level of uncertainty into the analysis.

Fourth step is recognizing outliers. To do that boxplots for each of analysed variables were created. Below I present the boxplots generated using python’s seaborn package.

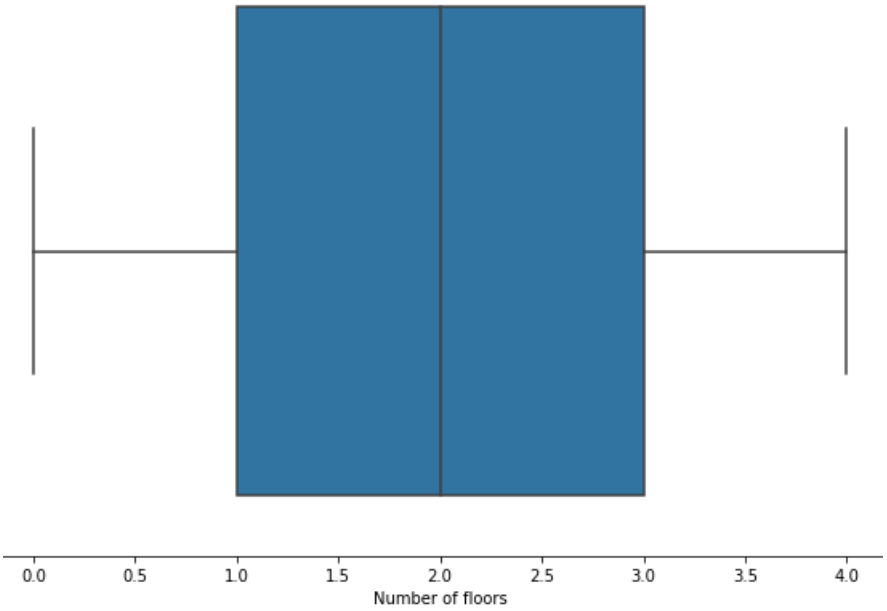

Figure 4. Boxplot for Floors variable

There were no outliers spotted for *floors* variable.

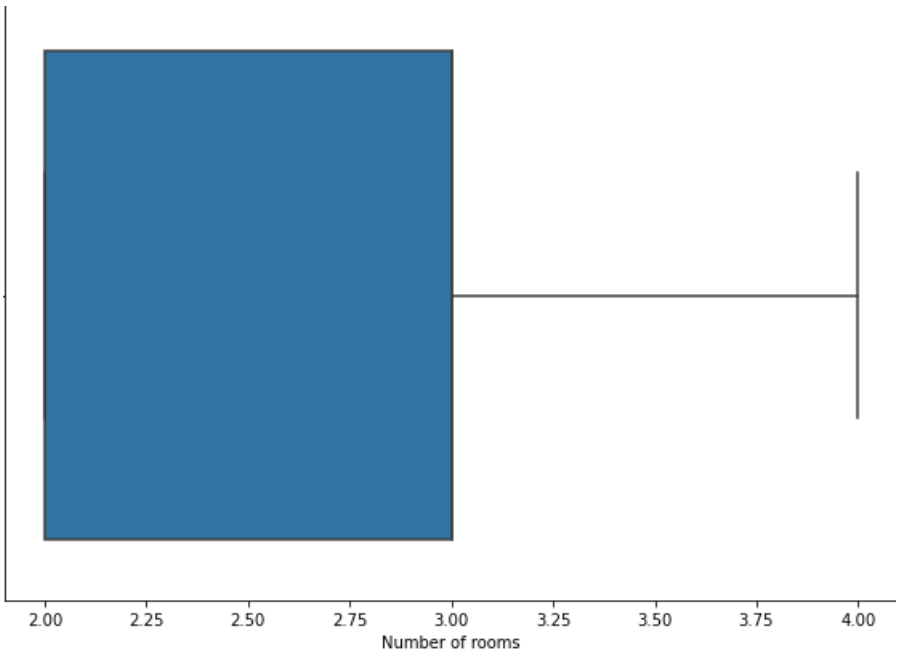

Figure 5. Boxplot for Rooms variable

There were no outliers spotted for *rooms* variable.

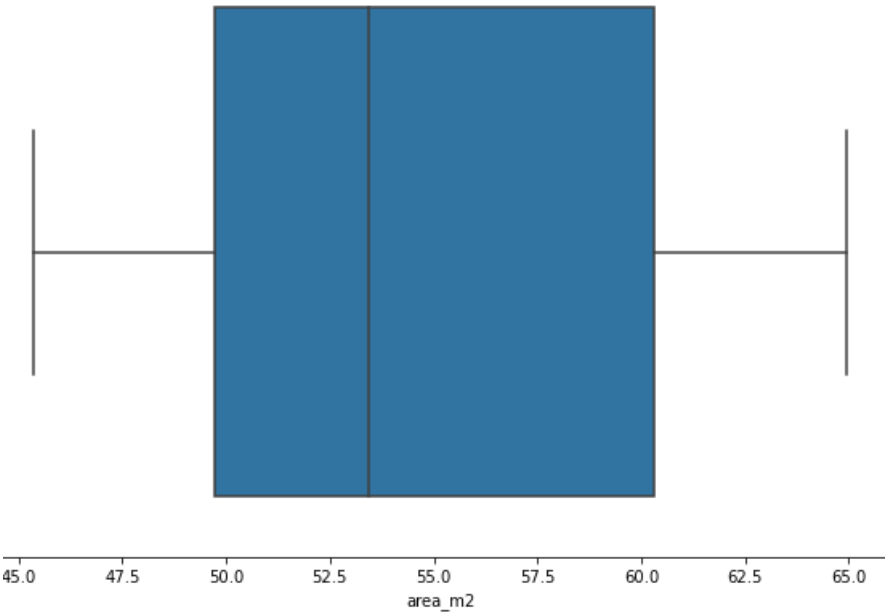

Figure 6. Boxplot for Area is squared meters

There were no outliers spotted for *area_m2* variable.

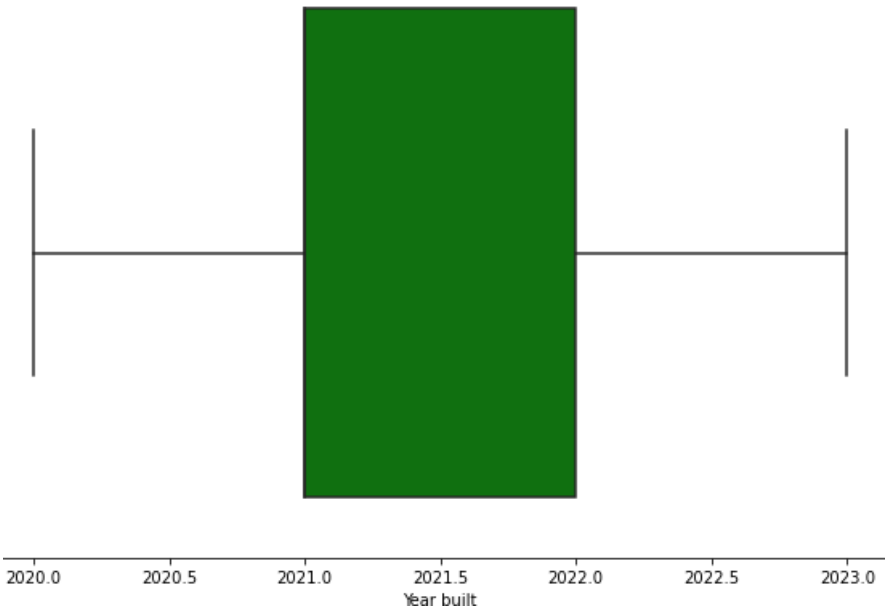

Figure 7. Boxplot for Year built

There were no outliers spotted for *year_built* variable.

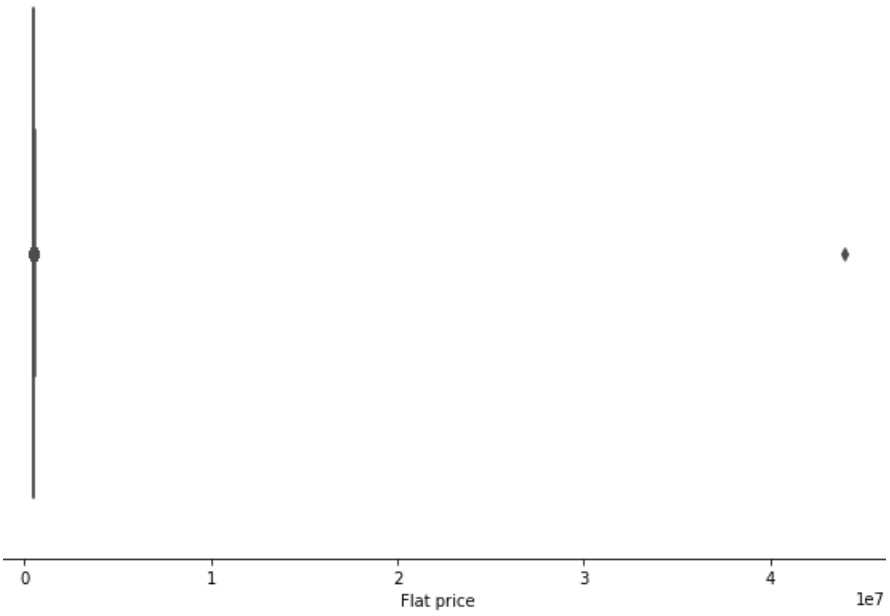

Figure 8. Boxplot for Price

There is at least one record which seems to be an outlier. There is a huge difference between that one point and the rest of the observations therefore it might has a substantial impact on further analysis. Going forward we should exclude it from the analysis as this record might be faulty one.

Figure 9 shows the boxplot for Price variable after removing this one observation from the dataset.

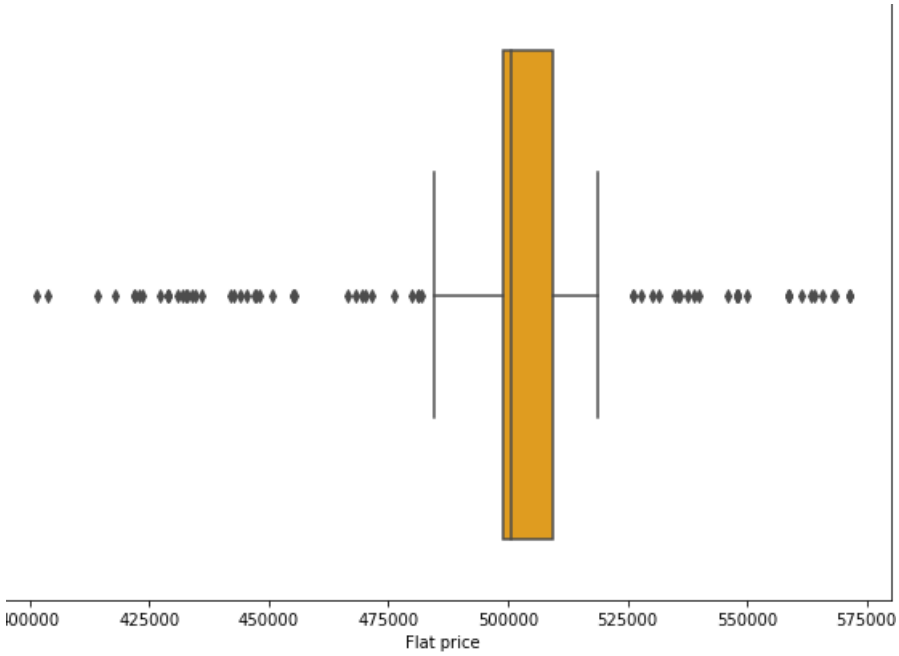

Figure 9. Boxplot for Price after removing one outlier

After removing outliers there is still significant number of records which exceeds the whiskers of boxplot. Nevertheless the difference between the prices is not that significant. Some of the outliers presents flats which are just 75 000 PLN more expensive than the average flat. This difference might be caused by some additional features of the flat. Additionally median and quartile1 and quartile3 seems to be quite flat now. Most of the home prices oscillates around 500 000 PLN. As the difference between outliers and median flat price is not huge I would keep these records for further analysis as they might bring some additional information. Maybe some of the flats have or are missing additional features which makes them more/less expensive.

The last – fifth step – is feature engineering. In the current stage the only action which I will take is creating a two new variables showing whether the flat was already built and price per squared meter which will be calculated by dividing flat price by its area in squared meters. Statistics for newly created price per square meter variable you can find in Figure 10.

count	178.000000
mean	9131.495724
std	809.555549
min	7202.246125
25%	8589.996339
50%	8889.441350
75%	9682.069588
max	10987.275121

Figure 10. Statistics for Price per square meter

Maximum price per square meter in the neighbourhood equals almost 11 000 PLN whereas the minimal value equals 7 200 PLN. As all of the flats are located in the same neighbourhood this difference seems to be big.

The second variable refers to whether flat was already built or not. Statistics for newly created categories you can find in Figure 11.

	price		year_built		area_m2		rooms		floor		price_m2	
	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median
built												
built	496243.515789	500800.0	2020.852632	2021	57.060316	59.33	2.684211	3	2.115789	2	8718.332283	8648.989899
to be built	499509.277108	500800.0	2022.036145	2022	52.371446	52.04	2.421687	2	1.975904	2	9604.393639	9535.415080

Figure 11. Statistics for built/to be build categories

There are two interesting things you can see there:

- Flats not yet build tends to be smaller (5 square meters difference in terms of average)
- Flats not yet built tends to be more expensive (almost 1000 PLN more expensive when it comes to price per square meter)

4. Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner

Key findings from Exploratory Data Analysis:

- Data regarding number of rooms, floor number and area is quite flat.
- For many of the offers the developer doesn't present the price. These records were replaced by the median value of price of flats.
- There was one outlier showing price of more than 4000000 PLN which seems to be wrong. The hunch is that too many 0 were entered into the offer. This record was excluded.
- Some of the flats differs from the mean price by maximum of 100000 PLN. This might be caused by some additional facilities present/missing in the flat or by number of square meters.
- The price per square meter shows a big difference in pricing. All of the flats are located in the same neighbourhood, though for some of them the price per square meters reaches 11 000 PLN where for the others it is around 7 500 PLN.
- Flats not yet build tends to be smaller but more expensive (they are around 5 square meters smaller but 1000 PLN more expensive in regards to price per square meter)

5. Formulating at least 3 hypothesis about this data

Based of the key findings presented in point 4 we can enrich the conclusions by formulating the below hypothesis:

- 1st hypothesis: There is a difference in price per square meter between flats already built and ones to be built
- 2nd hypothesis: There is a difference in number of square meters in flats already built and ones to be built

- The mean price of flats located on 1st, 2nd, 3rd floor are the same

6. Conducting a formal significance test for one of the hypotheses and discuss the results

I will conduct a hypothesis test for the first hypothesis presented in point 5. First I will define hypothesis:

H0: There is no difference in price per square meter between flats already built and ones to be built

H1: There is a difference in price per square meter between flats already built and ones to be built

Next, using pythons ttest_ind function, we will conduct a t-test at a alpha = 0.05 significance level. Results of the test are presented in Figure 12.

```
built_hypo = flats.loc[flats.built == 'built','price_m2']
not_built_hypo = flats.loc[flats.built == 'to be built','price_m2']

alpha=0.05
t_value, p_value = stats.ttest_ind(built_hypo, not_built_hypo)
print("t_value1 = ",t_value, ", p_value1 = ", p_value)
```

```
t_value1 =  -8.680985090393076 , p_value1 =  2.592678631797329e-15
```

Figure 12. Results of t-test conducted for formualted hypothesis

As the p-value is less than alpha (0.05) we reject the null hypothesis that there is no difference in price per square meter between flats already built and flats to be built. This confirms our assumption that developers are asking for higher bids for flats which are currently under construction.

7. Suggestions for next steps in analyzing this data

In our analysis we formulated very interesting points and discovered powerful insights regarding local housing market in Wrocław. To enhance our knowledge regarding the reasons staying behind higher prices for some of the flats in the next steps of the analysis we might need to gather additional data regarding existence/non-existence of additional facilities in flats e.g.

pools, solar systems, balcony, terrace etc. This binary variables might show why some of the flats are priced higher than the others. Having this binary variables as well as already possessed ones we would be able to predict the price of a new flat by features such as floor number, number of rooms, year build and existence of facilities. That might help people considering buying a new flat in the future in preparing sufficient money contribution and evaluating bank mortgage offers.

8. A paragraph that summarizes the quality of this data set and a request for additional data if needed

Analysed data presents good quality. One of the bottle-necks is a missing price information from the developers who sometimes put “ask for bid” category under price field in an offer. This forced as to replace such an entries with median price. Getting more exact data from the developers might improve our analysis and possibly change some of the presented metrics. Additionally developers in the future might put additional information regarding additional facilities attached to a given flat. That might give us more knowledge about the reasons of high/low flat price.