Housing sales price and enterprises data analysis in Vilnius city

1 Introduction

1.1 Background

I have been living in Vilnius since 2013 and I have already witnessed true power of Lithuanian business capabilities. Vilnius is ever-growing city, bringing more and more technology based investments into the region, thus increasing demand for commercial real estate which leads into big urbanization problem in the districts of Vilnius. So far we already felt the need of residential housing, public schools and kindergartens. Too bad, Lithuanian economy growth is not reaching majority of the Vilnius city dwellers, which results in majority of young professionals not being able to afford their housing even in the suburbs of Vilnius for a decent amount of living cost. This problem leads a lot of talented people traveling away to distant countries find their way of living.

1.2 Problem

With this little project of mine, I am looking to achieve a clear understanding of Vilnius city economical state, by evaluating current enterprises running in Vilnius and housing sales price.

1.3 Interest

By no means, analyzing this type of data is beneficial for everyone living in a city of Vilnius, real estate dealers, residents, people who are researching demographics of Lithuania and of course fellow data scientists.

2 Data acquisition and preparation

2.1 Data sources

In many cases, finding raw, open-sourced data about housing market and current enterprises in Lithuania is really hard. For example biggest real estate advertising companies do not share their database data in .csv; .json format files which would make data analyst job easier, companies like https://rekvizitai.vz.lt/ which have well structured database (later as DB) on enterprises in Lithuania, do not share these DB's for free. Most of the time I had to leverage on my skills to try and gather data by using myself prepared algorithms.

So main sources of my data came from following websites:

a) Housing prices in Vilnius - https://www.aruodas.lt
b) State Social Insurance Fund's open database for current insurers (companies) - http://atvira.sodra.lt/imones/paieska/index.html
c) Other websites, containing additional data points, that previous data sets did not have

2.2 Data preparation

After scraping thousands of pages I encountered bunch of problems like some enterprises having changed their address, some of them not giving full detail and so on, which resulted in most of my time cleaning data sets instead of making actual data analysis.

Firstly I have to mention that publicly available data has a tendency to be super trashy and dis-organized. I had to spend considerable amount of time to get the data frames ready from State Social Insurance Fund’s open database. One of the biggest problems I faced was that, they do not provide company’s address which leaves a big hole in my problem solving because I meant to analyze spatial data and provide beautiful data visualizations on interactive map. For this case I needed to parse all the names of the companies I just scraped and scrape another pages to get address for them.

This led me to another problem, company names in database where typed in by super strict Lithuanian grammar rules which made my address search harder. So I had to fixed the names of the companies and then scrape for address’s.

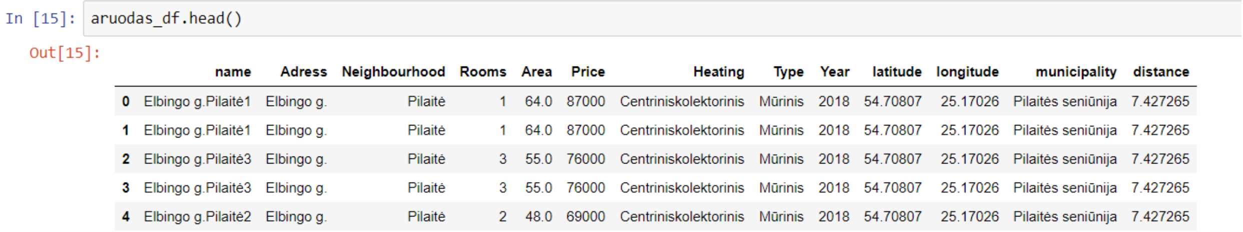
Thirdly, after I had address’s, I had to figure out a way to sort them into existing neighborhoods in Vilnius which led to more scraping. I wanted to get companies sorted by neighborhoods because data sets I scraped from housing price website contained neighborhood columns which seamed to be promising and easy to work with, instead of address’s.

Lastly for all of the address’s I had to use some libraries to get geo locations (latitude’s and longitudes) to be able to represent results on a map.

2.3 Feature selection

After all data cleaning and organizing I ended up having 2 data frames:

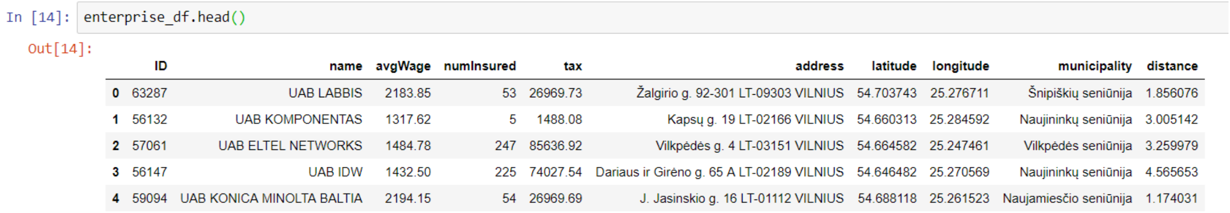
a) Dataframe for housing



```
In [15]: aruodas_df.head()
Out[15]:
```

	name	Adress	Neighbourhood	Rooms	Area	Price	Heating	Type	Year	latitude	longitude	municipality	distance
0	Elbingo g.Pilaitė1	Elbingo g.	Pilaitė	1	64.0	87000	Centriniskolektorinis	Mūrinis	2018	54.70807	25.17026	Pilaitės seniūnija	7.427265
1	Elbingo g.Pilaitė1	Elbingo g.	Pilaitė	1	64.0	87000	Centriniskolektorinis	Mūrinis	2018	54.70807	25.17026	Pilaitės seniūnija	7.427265
2	Elbingo g.Pilaitė3	Elbingo g.	Pilaitė	3	55.0	76000	Centriniskolektorinis	Mūrinis	2018	54.70807	25.17026	Pilaitės seniūnija	7.427265
3	Elbingo g.Pilaitė3	Elbingo g.	Pilaitė	3	55.0	76000	Centriniskolektorinis	Mūrinis	2018	54.70807	25.17026	Pilaitės seniūnija	7.427265
4	Elbingo g.Pilaitė2	Elbingo g.	Pilaitė	2	48.0	69000	Centriniskolektorinis	Mūrinis	2018	54.70807	25.17026	Pilaitės seniūnija	7.427265

b) Dataframe for enterprises



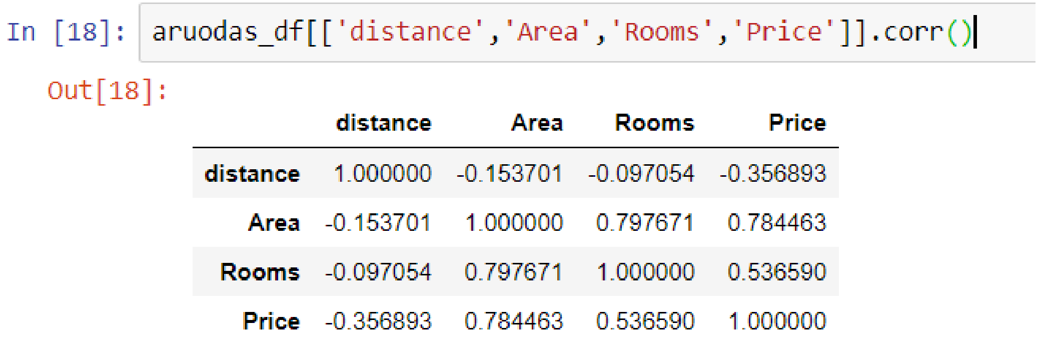
```
In [14]: enterprise_df.head()
Out[14]:
```

	ID	name	avgWage	numInsured	tax	address	latitude	longitude	municipality	distance
0	63287	UAB LABBIS	2183.85	53	26969.73	Žalgirio g. 92-301 LT-09303 VILNIUS	54.703743	25.276711	Šnipiškių seniūnija	1.856076
1	56132	UAB KOMPONENTAS	1317.62	5	1488.08	Kapsų g. 19 LT-02166 VILNIUS	54.660313	25.284592	Naujininkų seniūnija	3.005142
2	57061	UAB ELTEL NETWORKS	1484.78	247	85636.92	Vilkpėdės g. 4 LT-03151 VILNIUS	54.664582	25.247461	Vilkpėdės seniūnija	3.259979
3	56147	UAB IDW	1432.50	225	74027.54	Dariaus ir Girėno g. 65 A LT-02189 VILNIUS	54.646482	25.270569	Naujininkų seniūnija	4.565653
4	59094	UAB KONICA MINOLTA BALTIA	2194.15	54	26969.69	J. Jasinskio g. 16 LT-01112 VILNIUS	54.688118	25.261523	Naujamiesčio seniūnija	1.174031

3 Exploratory data analysis

3.1 Correlations in data sets

It is widely accepted that apartments that are nearby city center and are bigger , they tend to be more expensive than a bit further smaller ones, but how much expensive? To start analyzing data, I simply started by checking out correlations in my data sets.



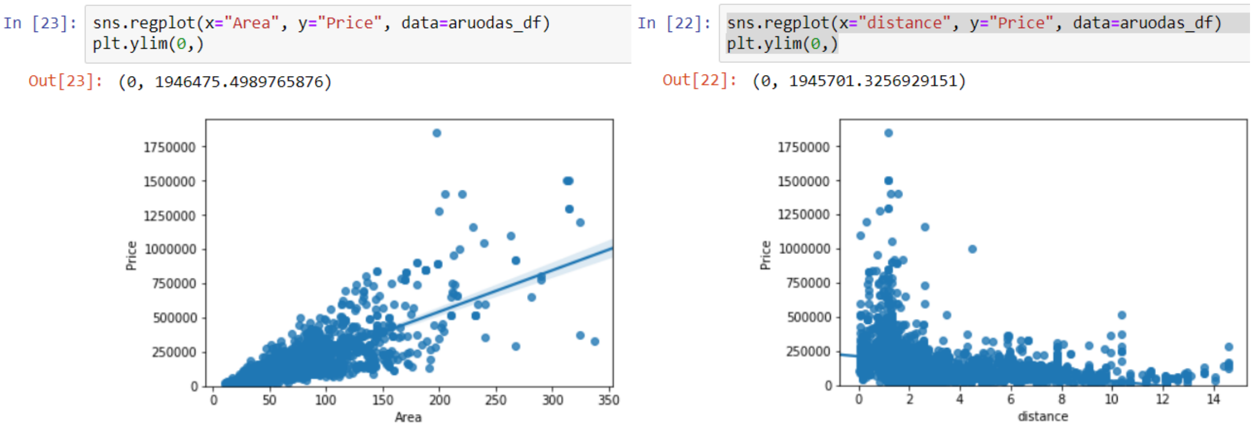
```
In [18]: aruodas_df[['distance','Area','Rooms','Price']].corr()

Out[18]:
```

	distance	Area	Rooms	Price
distance	1.000000	-0.153701	-0.097054	-0.356893
Area	-0.153701	1.000000	0.797671	0.784463
Rooms	-0.097054	0.797671	1.000000	0.536590
Price	-0.356893	0.784463	0.536590	1.000000

3.1 Correlation coeficients of aruodas_df data set float variables

Having these correlation parameters, I assumed that most of the analysis can be done based on these variables: "distance, Area, Price". I dropped out "Rooms" because of having multiple variables concerning 1 function area of the apartment (having more rooms means it has more area in an apartment, so it is obvious that it is goanna be positively correlated).



```
In [23]: sns.regplot(x="Area", y="Price", data=aruodas_df)
         plt.ylim(0,)

Out[23]: (0, 1946475.4989765876)
```

```
In [22]: sns.regplot(x="distance", y="Price", data=aruodas_df)
         plt.ylim(0,)

Out[22]: (0, 1945701.3256929151)
```

3.2 Scatter plots of Price/Area on left and Price/Distance data points.

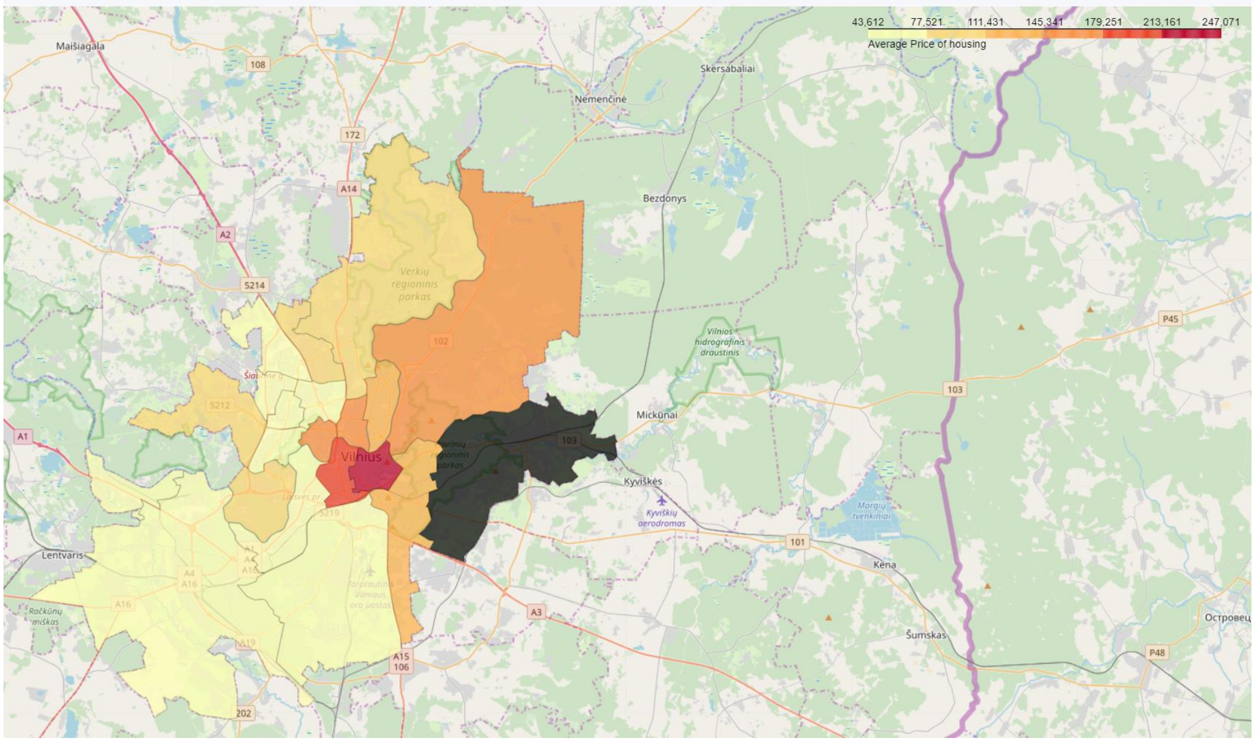
As it is shown in figure 3.2 we can't say for certain that variables are correlated similarly, Area/Price is sort of linear and Distance/Price is exponential from first look.

3.2 Housing price map

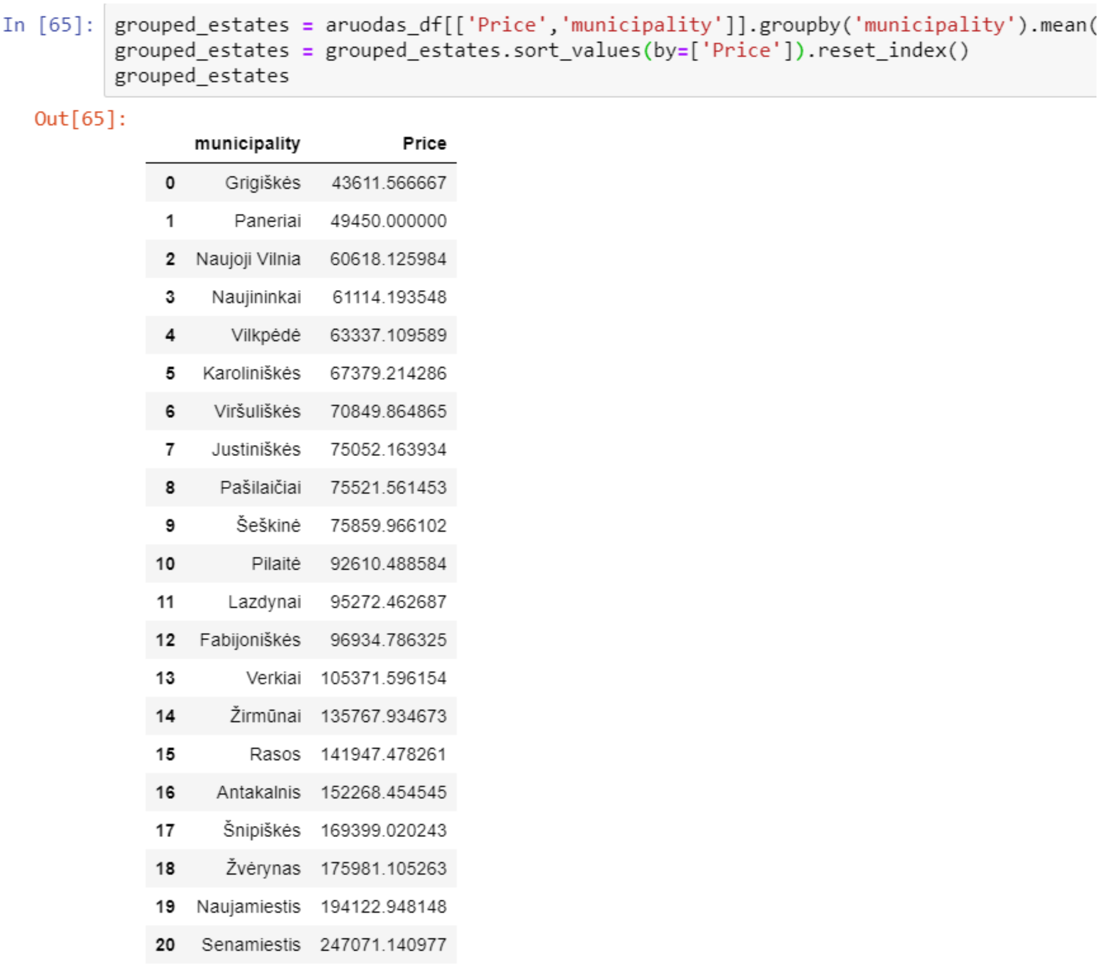
For exploratory purposes, I have create choropleth map of average housing price based on municipalities of Vilnius city.

It shows us, that regions further away from city center region, are way cheaper than in a city center municipality called Senamiestis. It confirms our 3.2 scatter plot of Distance/Price correlation, it is actually exponential on the map.

There’s only one issue with this map, that one of the regions has no info and has a black shade of it.



3.3 Average housing price in Vilnius city heat map.



In [65]:

```
grouped_estates = aruodas_df[['Price','municipality']].groupby('municipality').mean(
grouped_estates = grouped_estates.sort_values(by=['Price']).reset_index()
grouped_estates
```

Out[65]:

	municipality	Price
0	Grigiškės	43611.566667
1	Paneriai	49450.000000
2	Naujoji Vilnia	60618.125984
3	Naujininkai	61114.193548
4	Vilkpėdė	63337.109589
5	Karoliniškės	67379.214286
6	Viršuliškės	70849.864865
7	Justiniškės	75052.163934
8	Pašilaičiai	75521.561453
9	Šeškinė	75859.966102
10	Pilaitė	92610.488584
11	Lazdynai	95272.462687
12	Fabijoniškės	96934.786325
13	Verkiai	105371.596154
14	Žirmūnai	135767.934673
15	Rasos	141947.478261
16	Antakalnis	152268.454545
17	Šnipiškės	169399.020243
18	Žvėrynas	175981.105263
19	Naujamiestis	194122.948148
20	Senamiestis	247071.140977

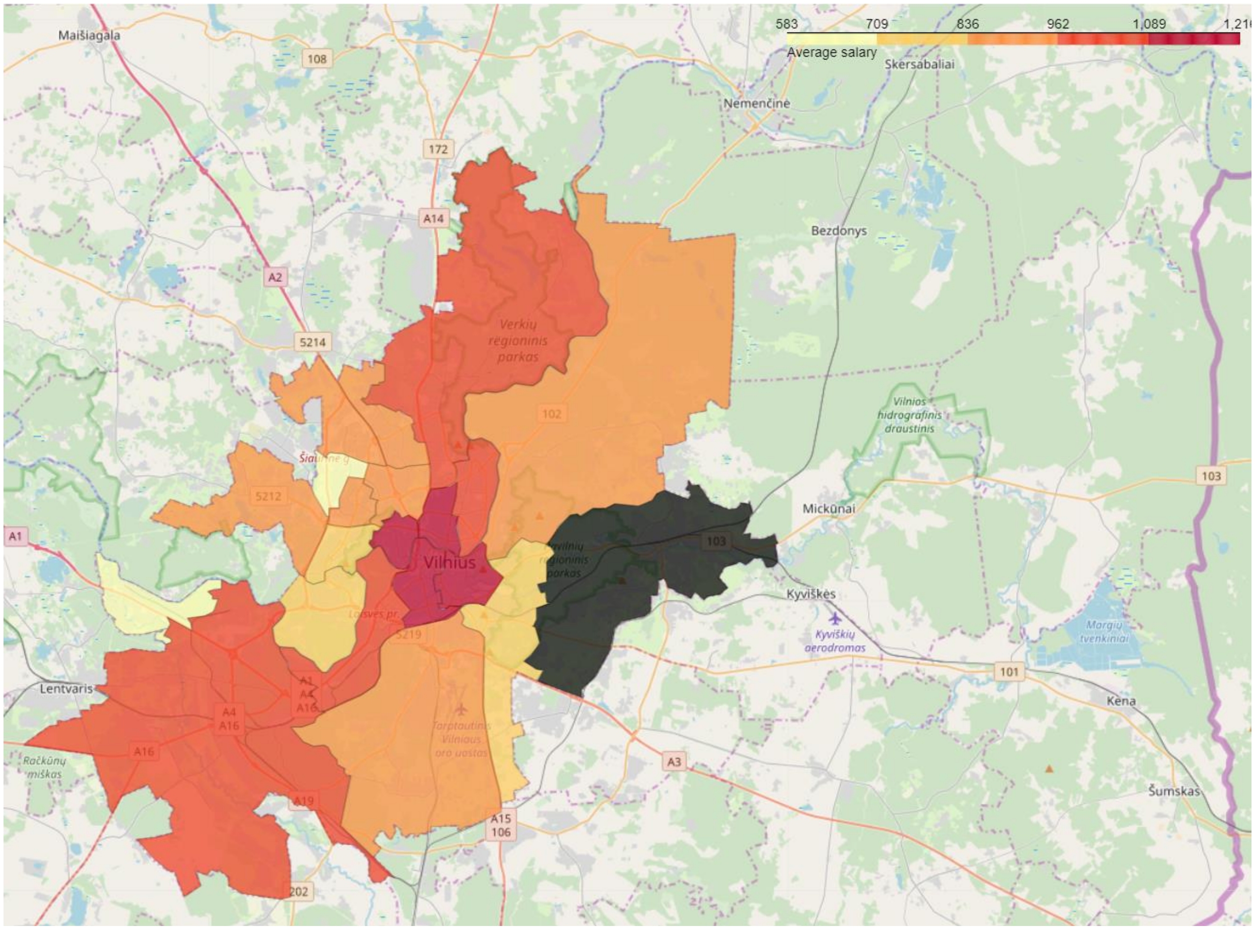
3.4 Dataframe, that shows average housing pricing in municipalities.

3.3 Enterprises average wage heat map

In this section in order to analyze average wage situation in region, I have created choropleth map for average wages in municipalities.

As we can see, city center contains biggest values of average salary, however it is not changing so rapidly as housing prices, it tends to stretch out in north and south.

There’s only one issue with this map, that one of the regions has no info and has a black shade of it.



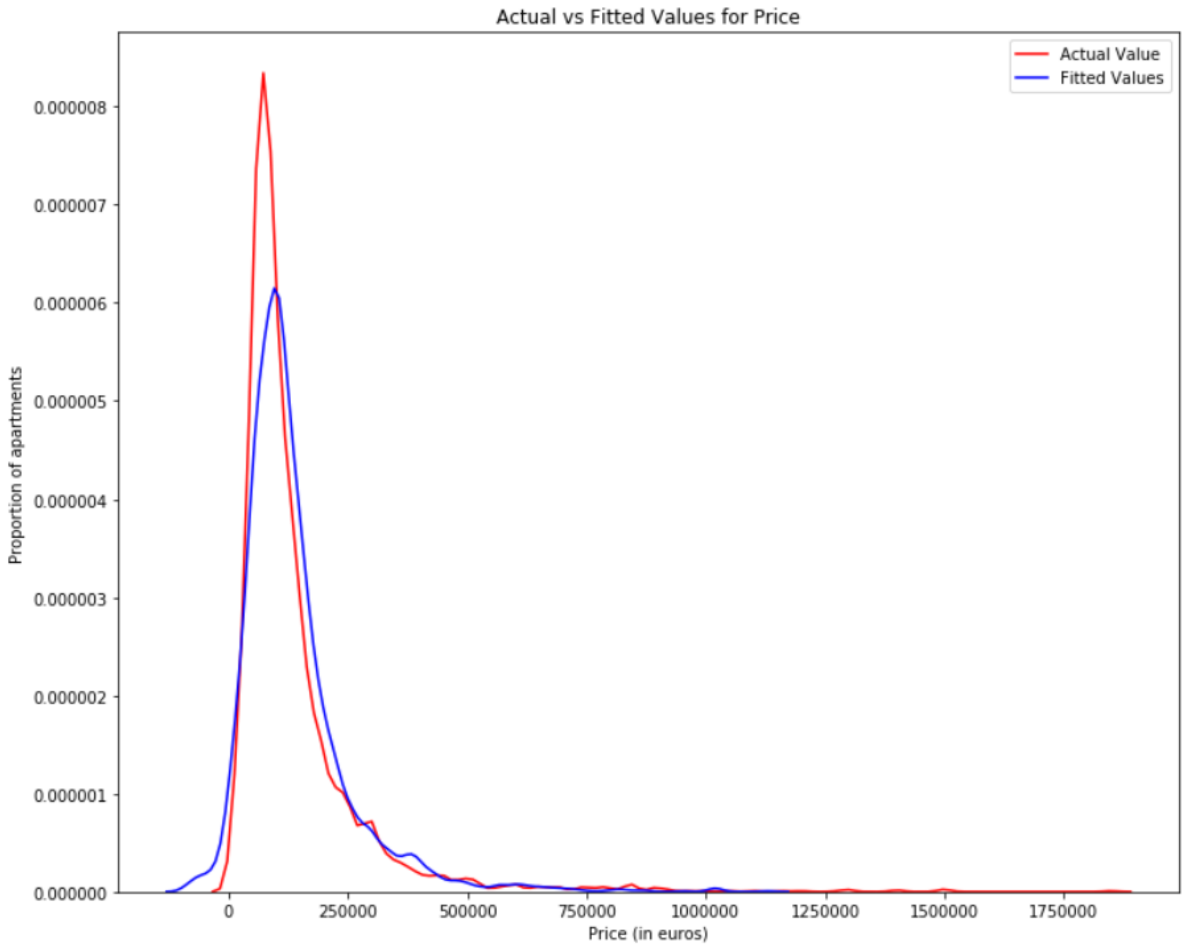
3.5 Average salaries in Vilnius city municipalities

4 Predictive modeling

4.1 Multi linear regression model for housing

After exploratory analysis of current housing listings in Vilnius city, I created multi linear model to predict Price of apartment based on distance to city center, area of apartment and number of rooms.

After training model and testing it, I have managed to achieve R^2 coefficient score of roughly 0.7.

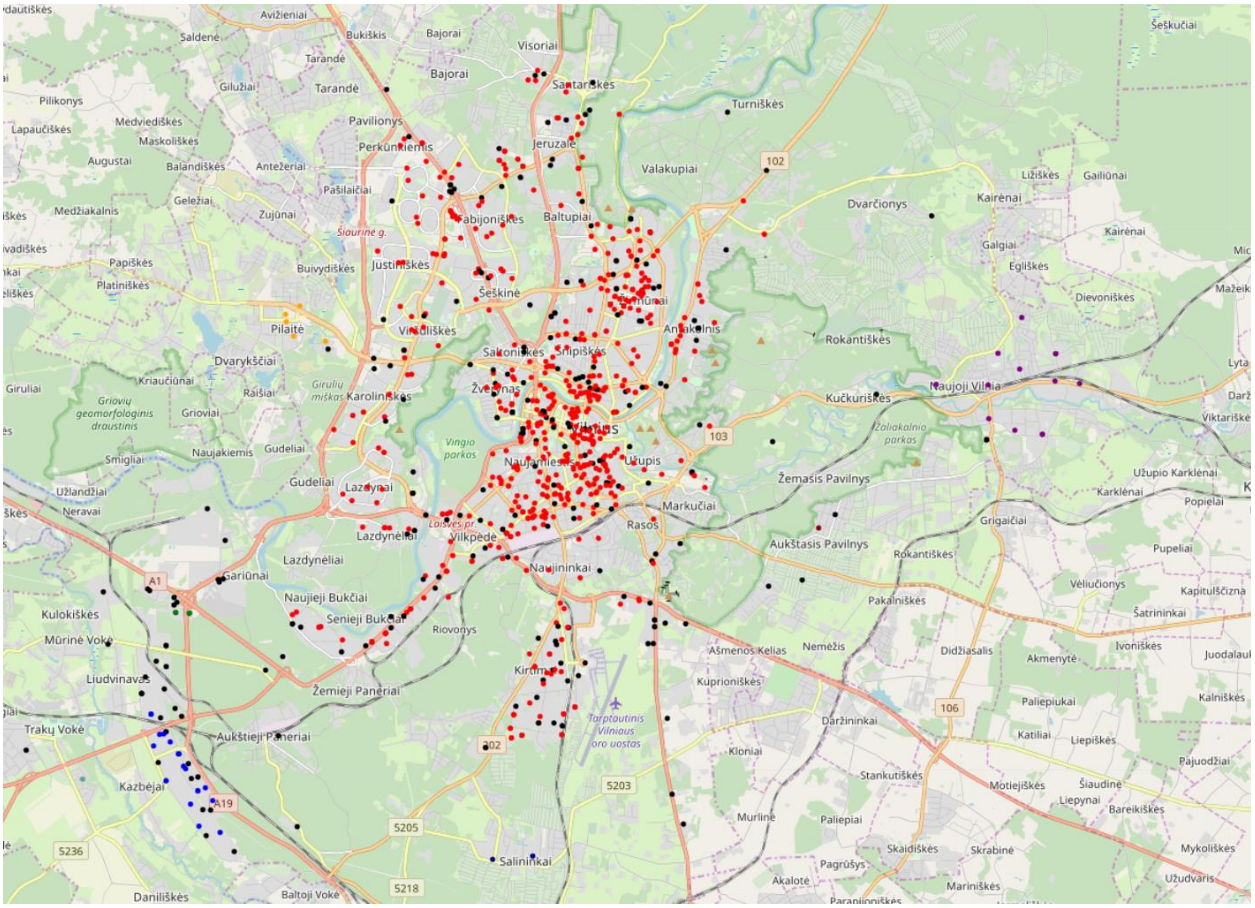


4.1 Actual vs fitted prices based on multilinear model.

4.2 Classification models

In this section I analyzed enterprises in Vilnius city, by clustering them based on coordinates of enterprise headquarters, average wage, distance to city center, employee number and payed tax to government.

After clustering all data, I represent it in folium map. Only one issue with creating map was that I can not represent all the dataframe companies I have, because notebook crashes, so I have limited my self of showing of only 1000 enterprises.



4.2 Clustered map of Vilnius city enterprises.

5 Conclusions

After quick analysis on created maps, we can certainly say that the further You get from city center, economical stance of city changes super fast:

- Average wage drops faster going east and west of the city, than north and east
- Average housing price drops dramatically

Based on these assumptions, we can justify the motion of people living outside of the city, but working in city center.

6 Future directions

Having generated maps for enterprises, we should look more into what are the factors that makes these flows between areas of average wages. What makes enterprises to pay bigger salaries based on it’s headquarters?

I have only touched base with data science curriculum and I wish to continue exploring all interesting patterns in our community.