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.

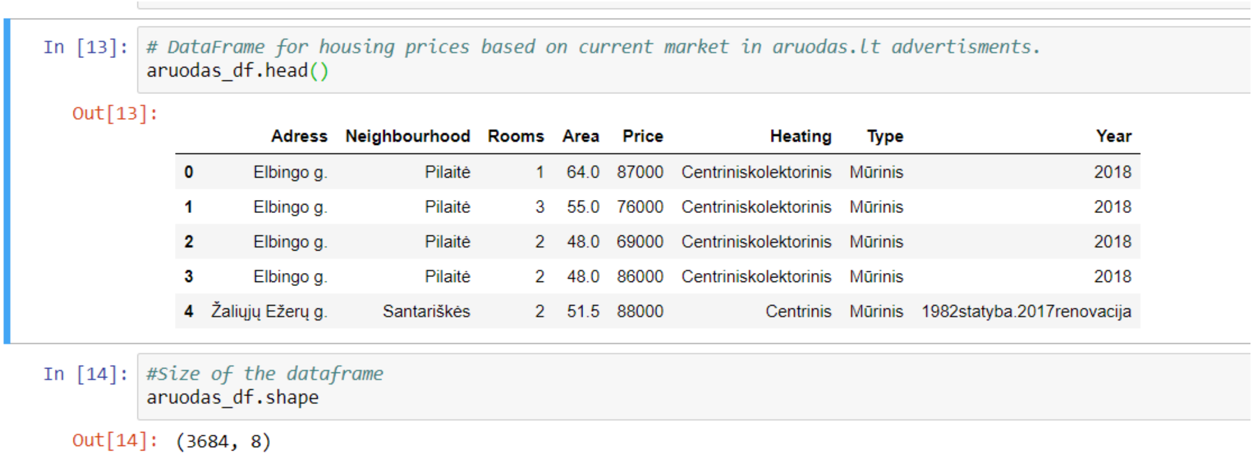
Thirdly, after I had address’s, I had to figure out a way to sort them into existing neighborhoods in Vilnius which led to more scraping. I wanted to get companies sorted by neighborhoods because data sets I scraped from housing price website contained neighborhood columns which seamed to be promising 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 [13]: # DataFrame for housing prices based on current market in aruodas.lt advertisments.
         aruodas_df.head()
```

Out[13]:

	Adress	Neighbourhood	Rooms	Area	Price	Heating	Type	Year
0	Elbingo g.	Pilaitė	1	64.0	87000	Centriniskolektorinis	Mūrinis	2018
1	Elbingo g.	Pilaitė	3	55.0	76000	Centriniskolektorinis	Mūrinis	2018
2	Elbingo g.	Pilaitė	2	48.0	69000	Centriniskolektorinis	Mūrinis	2018
3	Elbingo g.	Pilaitė	2	48.0	86000	Centriniskolektorinis	Mūrinis	2018
4	Žaliųjų Ežerų g.	Santariškės	2	51.5	88000	Centrinis	Mūrinis	1982statyba.2017renovacija

```
In [14]: #Size of the dataframe
         aruodas_df.shape

Out[14]: (3684, 8)
```

b) Dataframe for companies

3 Exploratory data analysis

3.1 Asdasd

3.2 Asdasdee

4 Predictive modeling

4.1 Regression models

4.2 Classification models

5 Conclusions

6 Future directions