Capstone project

1) Introduction

discuss the business problem and who would be interested in this project.

When potential clients are going through real estate listings the only reference easily available is the average price in a given city or a district which is not always helpful because there are many factors to be considered when buying an apartment. I'm going to leverage information from Foursquare about a number of different types of venues nearby as well as information from local real estate listings website to predict prices of apartments and to identify which ones have most attractive price compared to the predicted price.

Target audience - The final result may be of value for people who are looking to buy an apartment as well as people who are selling an apartment and would like to know what price to ask for.

2) Data

describe the data that will be used to solve the problem and the source of the data.

There are three sources of the data which will be combined to solve the problem

City of choice

I'll analyze the city of Poznan in Poland, because this is where i live and therefore may one day need results of this project. https://en.wikipedia.org/wiki/Pozna%C5%84

Foursquare data

I will identify what kind of venues are within a distance of each apartment. Next I'll count how many venues there are by category. I'll use this information in regression models to predict apartment prices.

Top level categories of Foursquare venue[1]:
- Arts & Entertainment
- College & University
- Event
- Food
- Nightlife Spot
- Outdoors & Recreation
- Professional & Other Places
- Residence
- Shop & Service
- Travel & Transport

Data with real estate listings

I'll download all real estate listings in city of Poznan and use them in regression models to predict apartment prices. Types of data that I'm planning to use are
- No of rooms
- Building type
- Construction status
- Price per m

1 based on https://developer.foursquare.com/docs/build-with-foursquare/categories/

- Rent
- Built year
- Size
- Extras like lifts, garage etc.

Only secondary market - There are significant differences between secondary and primary real estate market, therefore this project will focus only on secondary real estate market.

Location data

To be able to explore venues around apartments, I'll also need location data. This information will be retrieved using OpenCage Geocoding API to get the longitude and latitude
https://opencagedata.com/api#forward-opt

After getting and cleaning the data we have the following columns

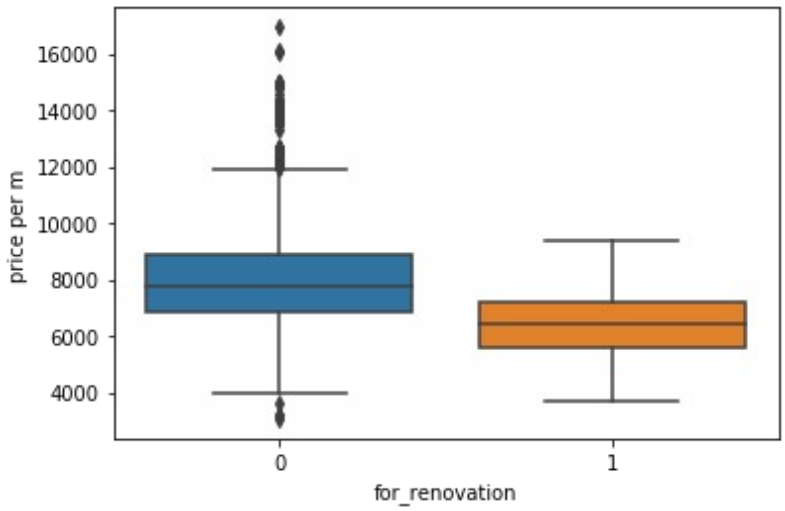
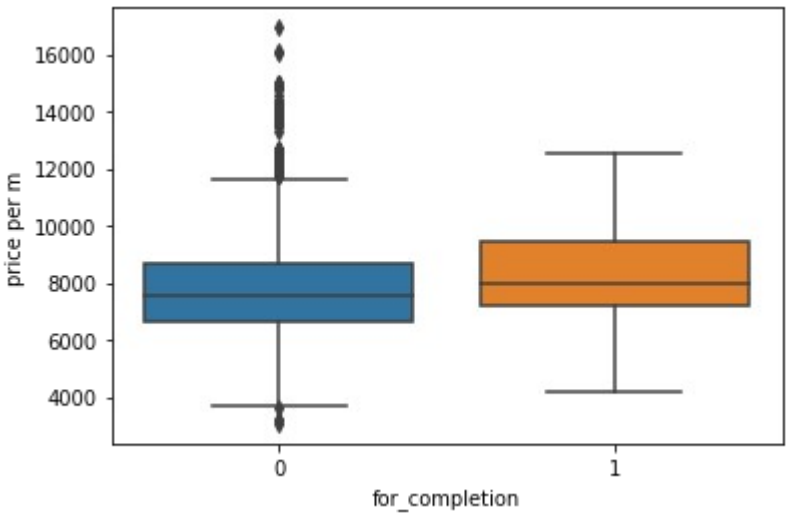
```
'id' – identifier of the listing from the real estate website
 'rooms' – number of rooms in the apartment
 'building_type' – type of the building (block, apartment, loft, house, etc.)
 'price per m' – price per square meter
 'rent' – cost of rent paid to the building administration for utilities, cleaning.
 'built year' – when building was constructed
 'size' – size of the apartment
 'floors' – number of floors in the building
 'floor no' – floor number of the apartment
 'district'
 'address'
 'url' – url to original listing
 'description' – text description
 'for_renovation' – does apartment require renovation
 'for_completion' – is apartment ready to move in
 'has_balcony'
 'has_basement',
 'has_garage',
 'has_garden',
 'has_terrace',
 'has_lift',
 'is_two_storey',
 'latitude',
 'longitude',
 'distance' – distance from the city center

Number of venues within 1000 meters from the apartment:
 'Arts & Entertainment'
 'College & University',
 'Food',
 'Nightlife Spot',
 'Outdoors & Recreation',
 'Professional & Other Places',
 'Shop & Service',
 'Travel & Transport'
```

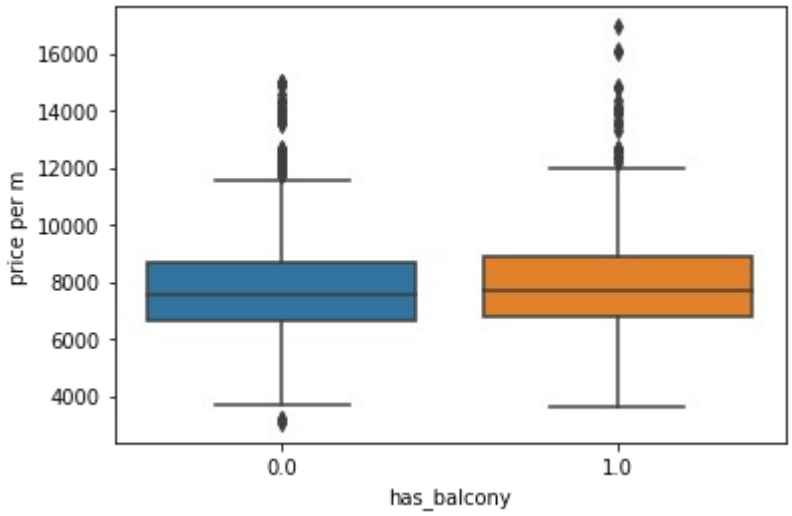
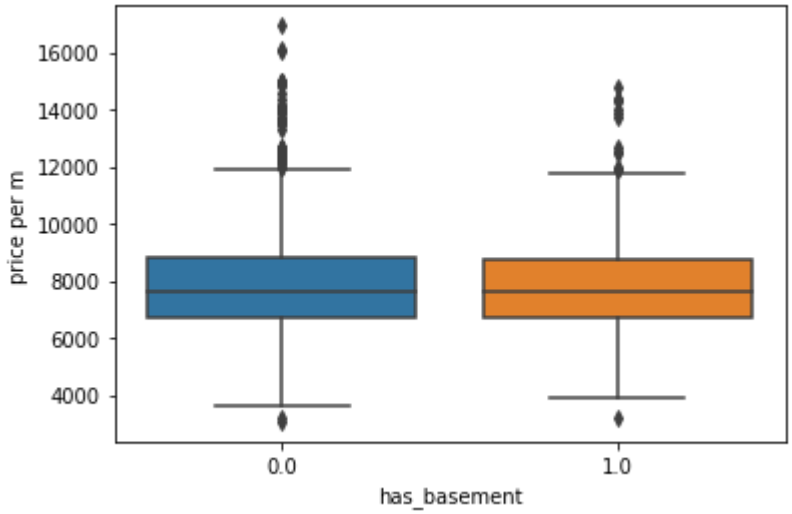
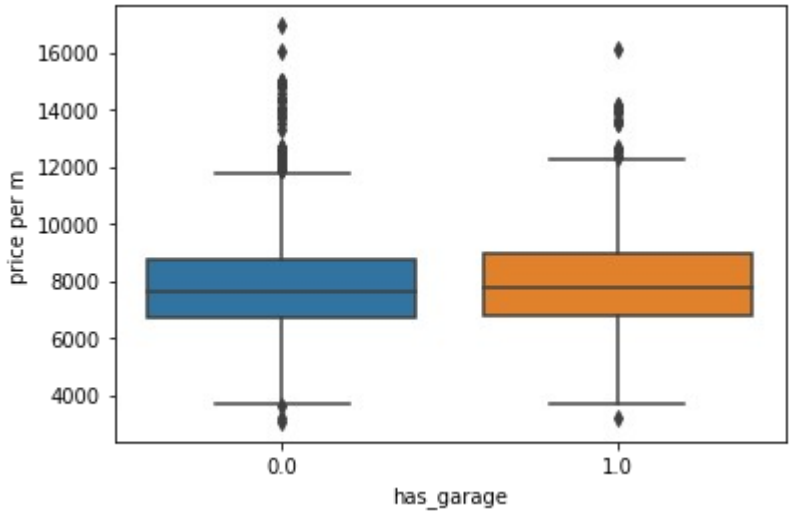
3) Methodology

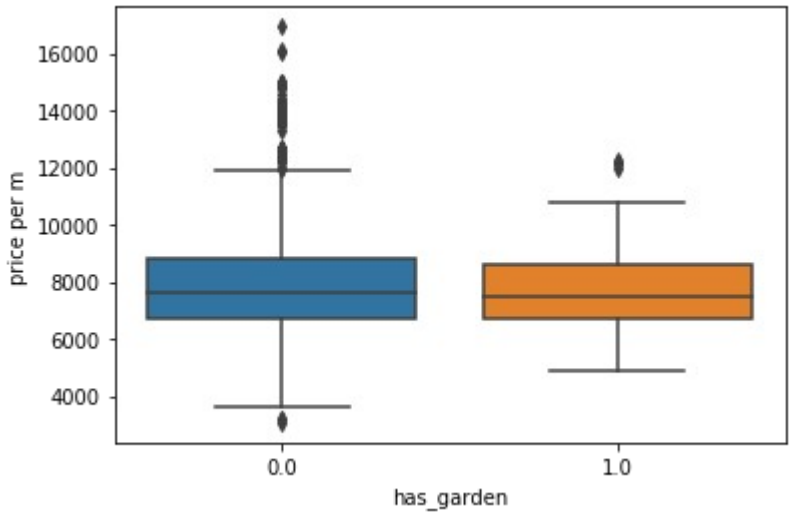
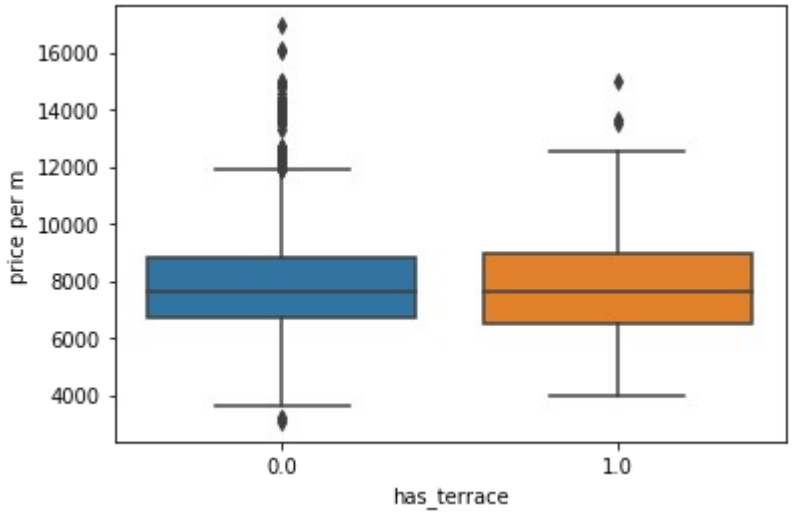
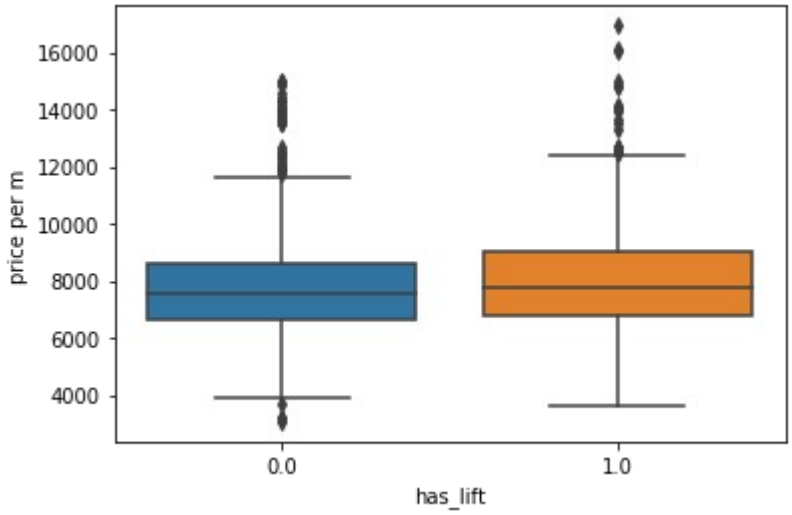
which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

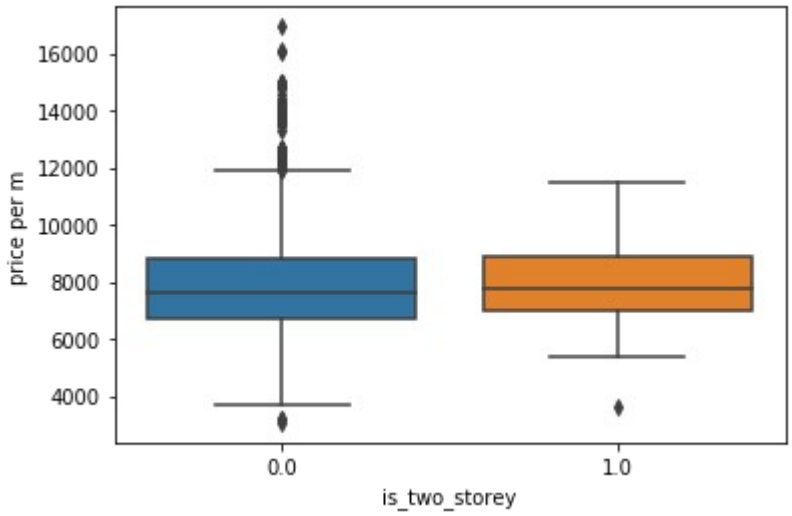
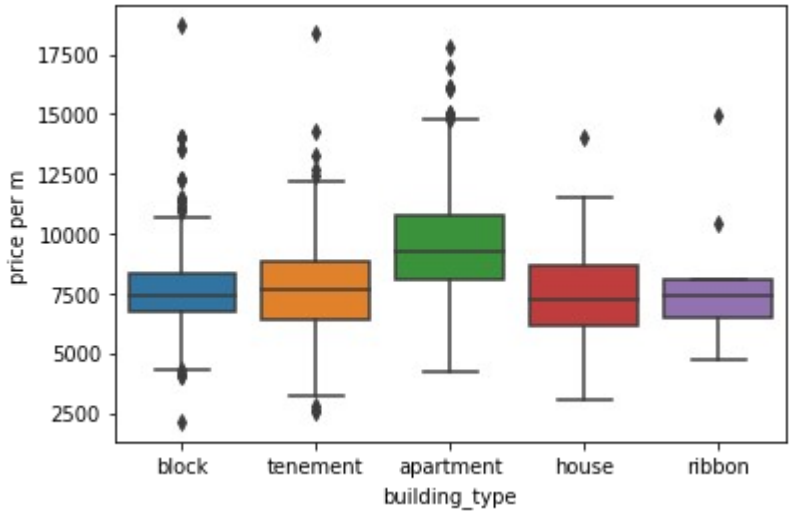
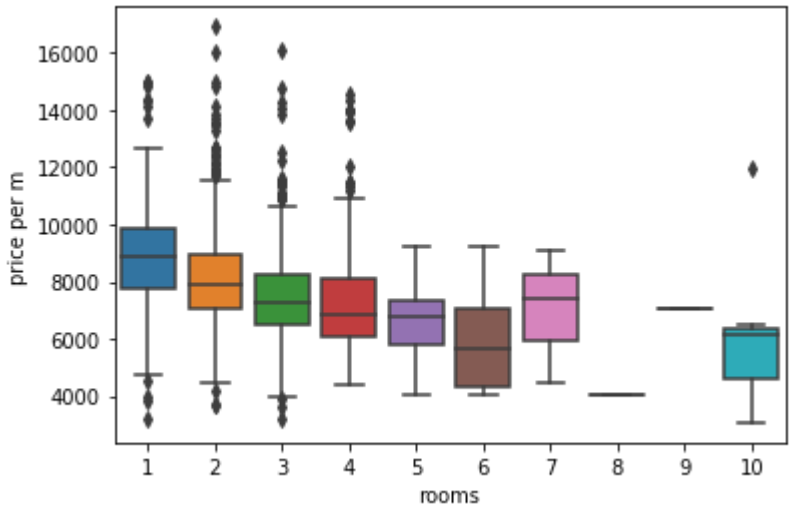
exploratory data analysis

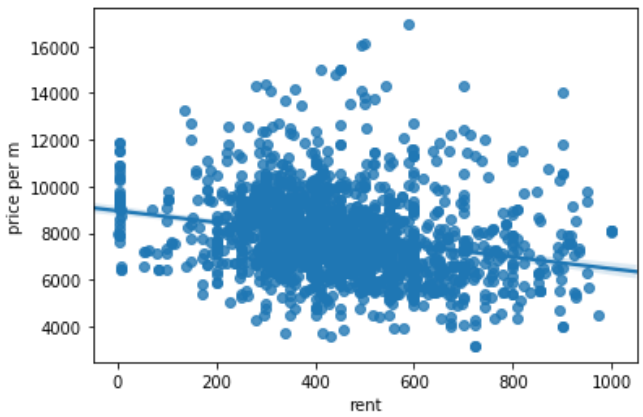
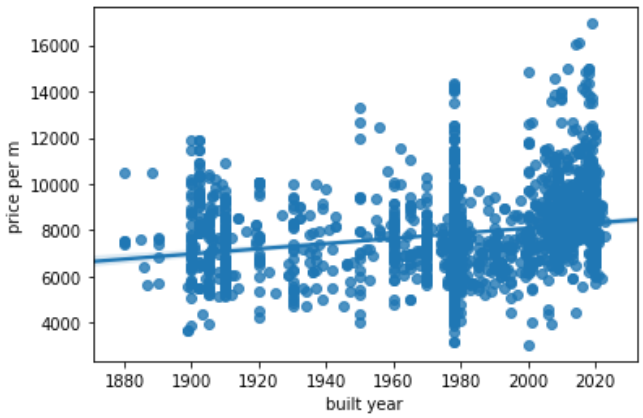
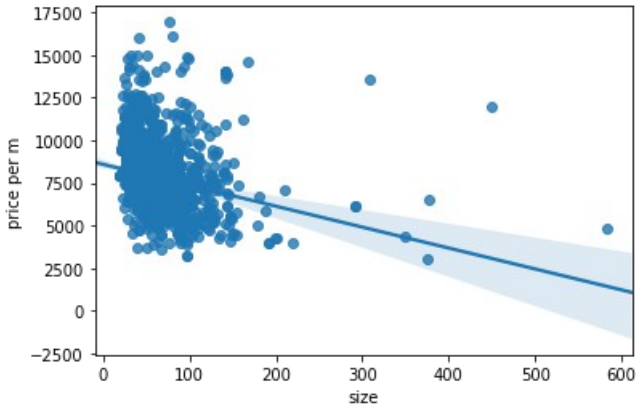
For zero/one values

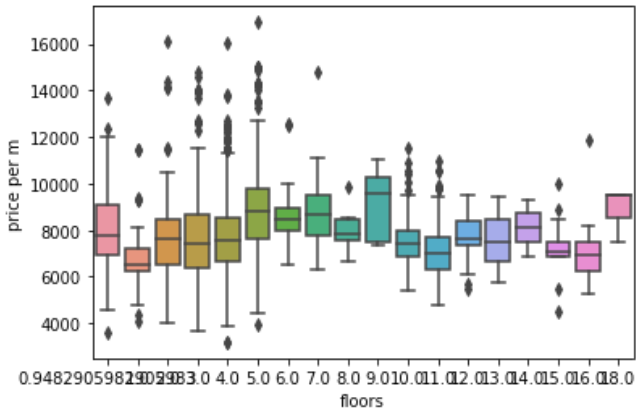
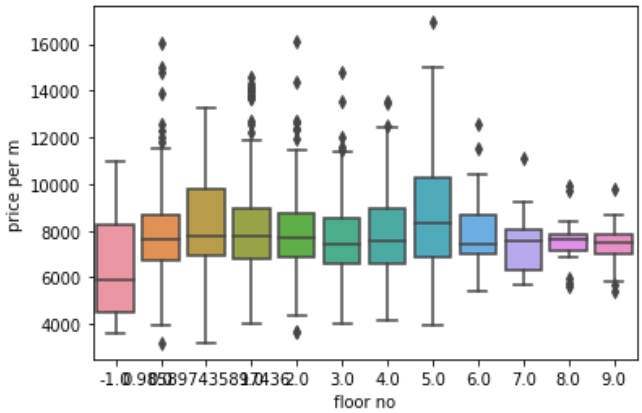
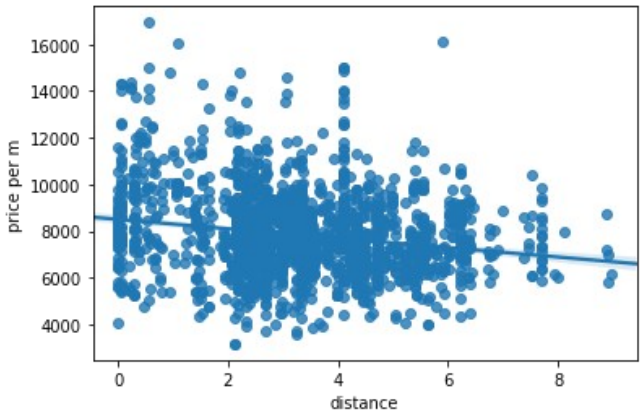
for_renovation - leave	
for_completion- leave	

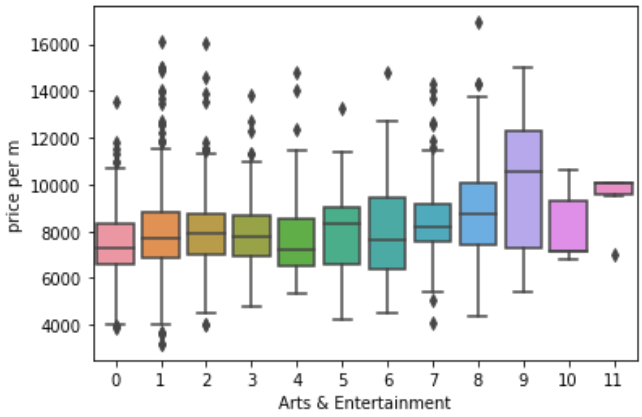
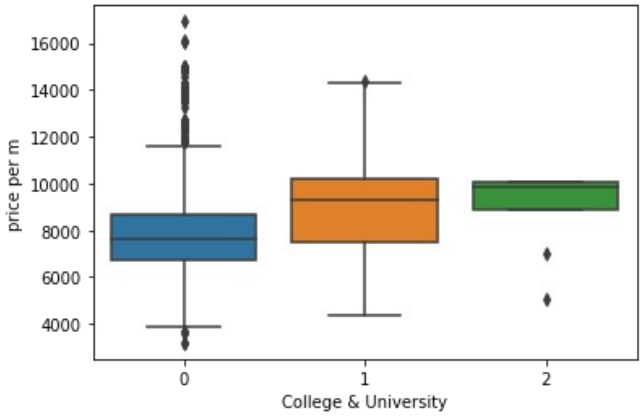
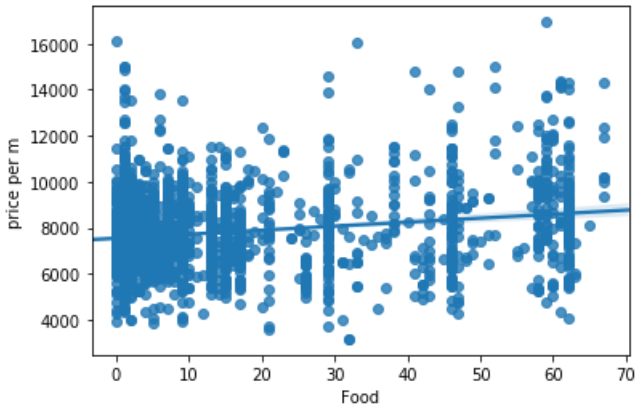
has_balcony – **REMOVE** as there is no difference in price based on this value	
has_basement – **REMOVE** as there is no difference in price based on this value	
has_garage – **REMOVE** as there is no difference in price based on this value	

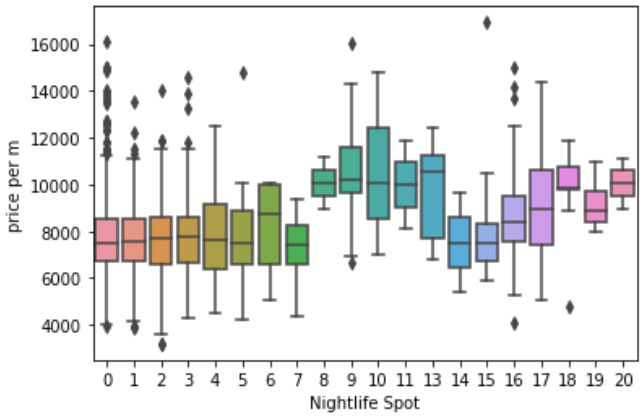
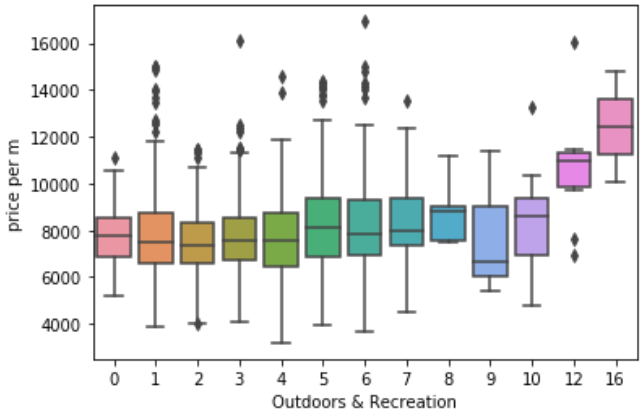
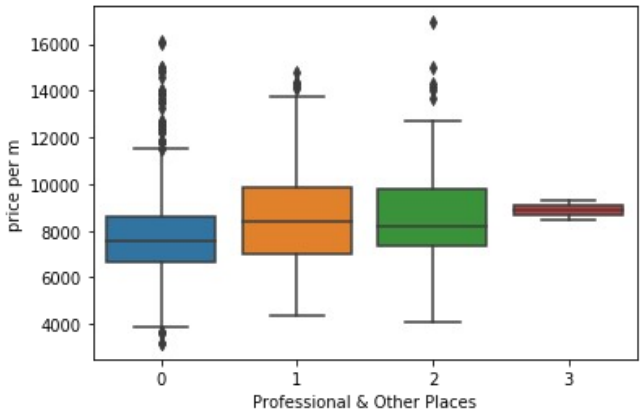
has_garden – **REMOVE** as there is no difference in price based on this value	
has_terrace – **REMOVE** as there is no difference in price based on this value	
has_lift – **REMOVE** as there is no difference in price based on this value	

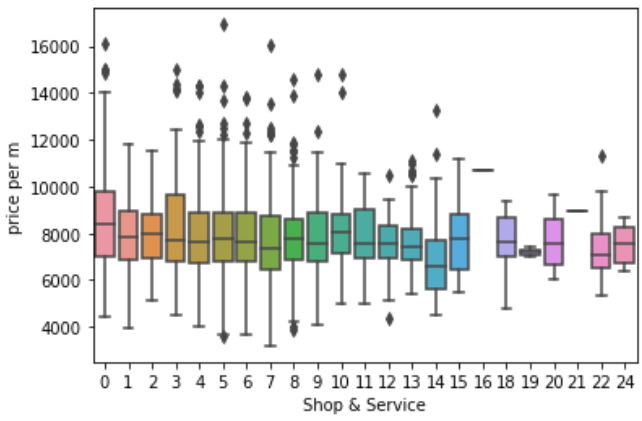
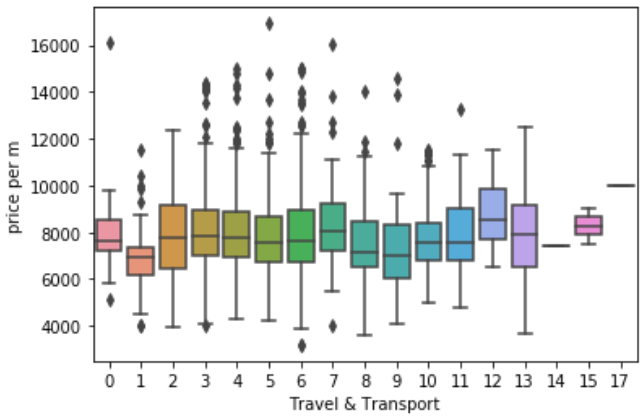
is_two_storey – **REMOVE** as there is no difference in price based on this value	
building_type – leave – apartment type of building has higher prices	
Rooms – leave – the more rooms the lower the price	

Rent – leave – the lower the rent the higher the price	
built year – leave – the newer the building the higher the price	
Size – leave – and remove outliers	

Floors – leave – the fewer floors the better the price	
floor no – leave – different floors favor different prices	
Distance – leave – the closer to the city center the better the price	

Arts & Entertainment – leave – the more the higher the price	
College & University – leave – as number of universities has impact on price	
Food – leave – as we can see the number of food places has an weak correlation with the price	

Nightlife Spot - leave – there is positive correlation between prices	
Outdoors & Recreation – leave – there is positive correlation between prices	
Professional & Other Places - leave – there is positive correlation between prices	

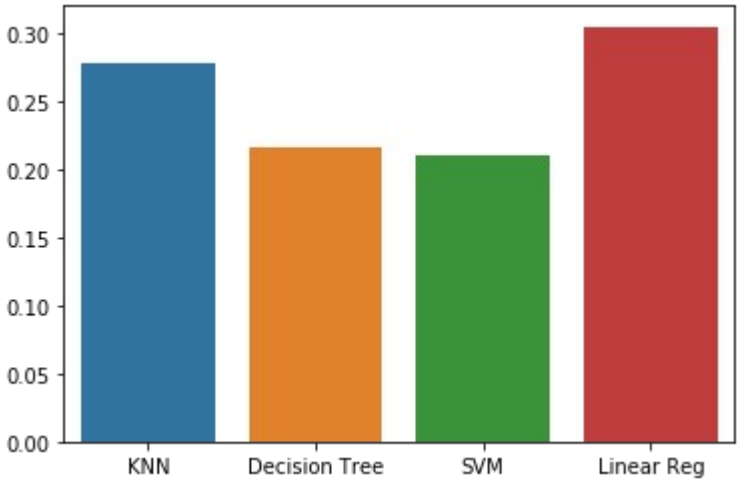
Shop & Service – REMOVE – no visible difference in price	
Travel & Transport – REMOVE – no visible difference in price	

Different types of regression models to predict prices of apartments were used:

- K Nearest Neighbor (KNN)
- Decision Tree
- Support Vector Machine
- Linear regression

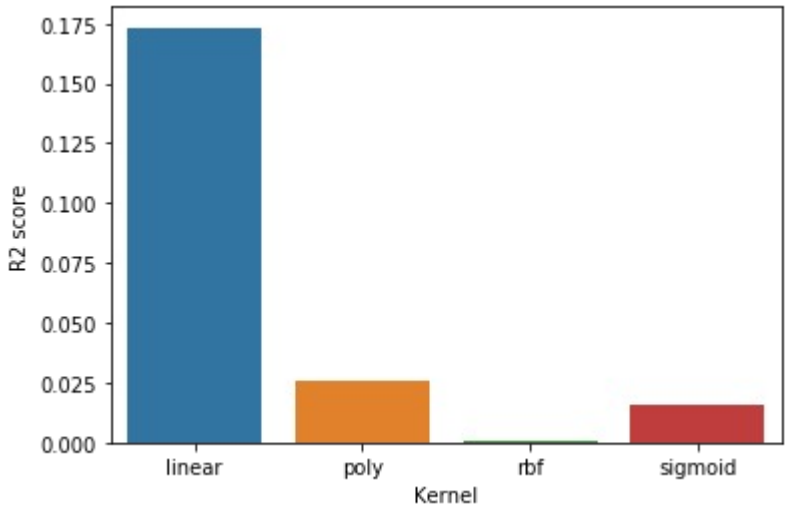
It was validated how well are they predicting prices by using R2 score.
To do this I've split the data into:

- test and train sets - used to fine tune each model before validation
- validation set - used to compare models



The winner model was **Linear regression.** But R2 was very low even for best model which suggests that there is more information influencing prices of apartments.

I didn't try polynomial regression as all data indicated linear relationship. Furthermore I've checked how SVM performed with different kernels, and linear was far better.



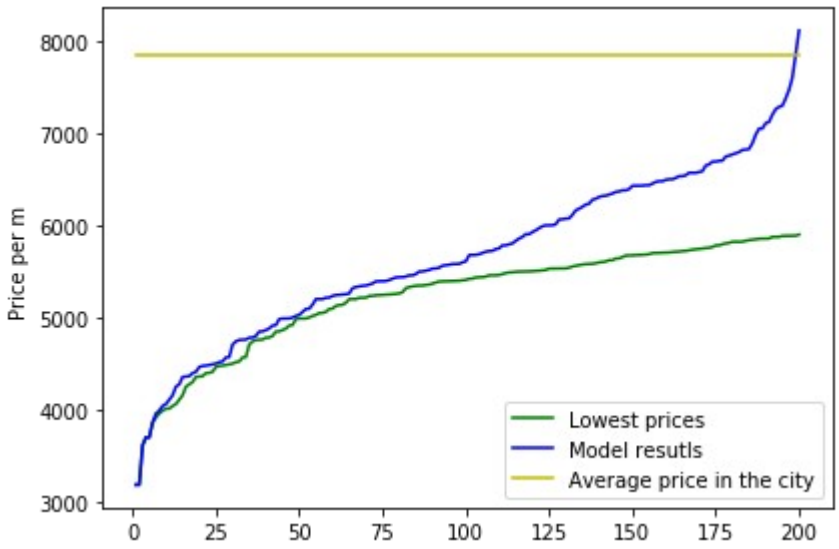
Finally I've used whole data set to predict apartment values with the best model and used difference between predicted value and apartment value as a filter for potentially good deals.

4) Results

where you discuss the results.

R2 score for models were rather low, but is the final result helpful at all? Maybe a simpler alternative could be just to review only the cheapest listings? I've compared 200 listings (10% of all data) with biggest difference in price when compared to the **model** with the list of 200 listings with lowest **prices**. A result was a list in which almost 60% of listings were the same.

When we visualize our results w can see that model is more likely to suggest listings with prices closer to average price in the city than simply listing apartments with lowest prices.



5) Discussion

where you discuss any observations you noted and any recommendations you can make based on the results.

Based on the results of the project we can see that there is a lot o information impacting the price but missing in the data. Unfortunately I was not able to parse this information because it was either unavailable at all or could be inferred from picture or descriptions only.

More information could be added like: standard of the apartment, building condition, how well rooms are arranged. Missing data which were replaced during preparation could be filled in.

Both of the above would have to be done manually but this should improve results of models.

In future an ability to add data incrementally about new listings and venues would be a plus as this project was based only on active listings. My hope is that with more historical data, we could get better predictions.

6) Conclusion

here you conclude the report.

Based on Linear Regression Model and Foursquare, geo and real estate data, I was able to identify potential good deals in a price ranges closer to average price of an apartment in the city. This can be considered as satisfactory result as such information is not available when filtering only by lowest price.

The goal of this project was to suggest potential good deals, but making the final decision on purchase consists of more criteria that need to be evaluated by the buyer.