**The battle of Neighborhoods - Capstone Project at Coursera Applied Data Science**

**Table of contents**

* [Introduction. Business Problem](https://e-5d463095ec.cognitiveclass.ai/lab?#introduction)
* [Data](https://e-5d463095ec.cognitiveclass.ai/lab?#data)
* [Methodology](https://e-5d463095ec.cognitiveclass.ai/lab?#methodology)
* [Analysis](https://e-5d463095ec.cognitiveclass.ai/lab?#analysis)
* [Results and Discussion](https://e-5d463095ec.cognitiveclass.ai/lab?#results)
* [Conclusion](https://e-5d463095ec.cognitiveclass.ai/lab?#conclusion)

**Introduction. Business problem**

STEM education (where STEM stands for **S**cience, **T**echnology, **E**ngineering, **M**athematics) is not new for developed countries where technical education is already a part of studying at schools. However, in developing countries like Ukraine where I am from, functions of STEM education still rare at regular schools and delivered as services by Technic clubs. So it is up to parents/kids whether he would attend such courses.

On the other side such clubs themselves are commercial projects with profit as an aim. Though more and more clubs opening last years there is still great share of market of technical education for children to be covered. And the most lucrative is opening new Robotics club in Kyiv - the capital of Ukraine with more than 3 million people officially living there (real number much higher).

Starting such business is not hard - couple of franchise offers is readily available. But the main challenge that would define success of this business is **choosing proper place to start** - that is where Foursquare and data analysis come into force.

**Business problem**

Mistake in choosing the place in this case might cost existence of business and loss of large part of investment. Hence entrepreneur, planning to start this business, as the main stakeholder, should be desirable for support in this choice

To choose best place for starting Robotics club we need to take into account many aspects. Most important of them are:

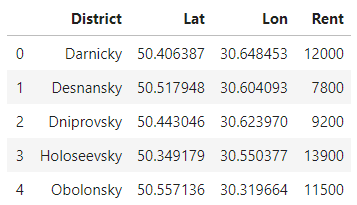
* How to divide the city into neighborhoods (no such information readily available so decision must be taken)
* Amount of schools in neighborhood (the more - the better)
* Amount of rivals (technical / math courses for school children, the less - the better)
* Amount of other not technical courses for same auditory (they might be interested in other education as well)
* Cost of rent (requires some work to get some indirect markers of price from Foursquare while rent prices not available there)

Obviously exploring Foursquare data for the city would answer most of above questions (and with some enhancements described in data section - all of them) so let us move on to define what data can we collect and which way to use it.

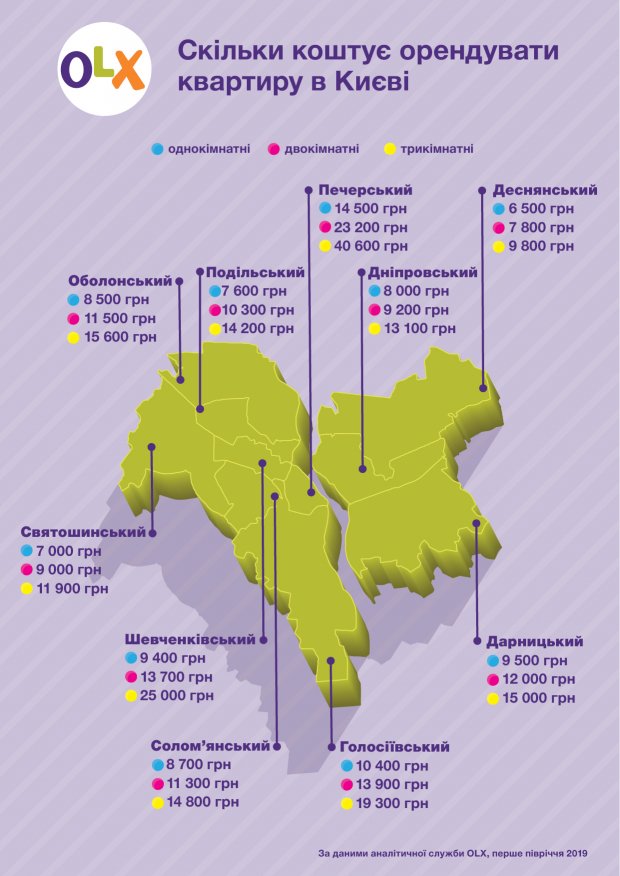
### Data - sourcing and usage

To answer main question of the project - i.e. choose best place to start with Robotics club following steps will be performed:

**Training data** We need to divide city into at least districts if not neighborhoods. Geospatial data for none of them is readily available so I have decided to use districts and pick one point for each as a center of a circle covering the most inhabited area (visible on satellite images). Main logic here is that starting club that works on weekends and working day evenings is better where most people live.



As you can see I have already added column of Rent rate (it is in UAH (Ukrainian Hryvnas) per 2 room flat). This is not commercial real estate prices which I have not found on the web, however it would give a clue on what districts are more expensive when building a model. As a source of information I have used infographics from olx.ua. In case you do not understand Cyrillic letters - there are districts of Kyiv and average prices for one, two and three room flats during 1H 2019.

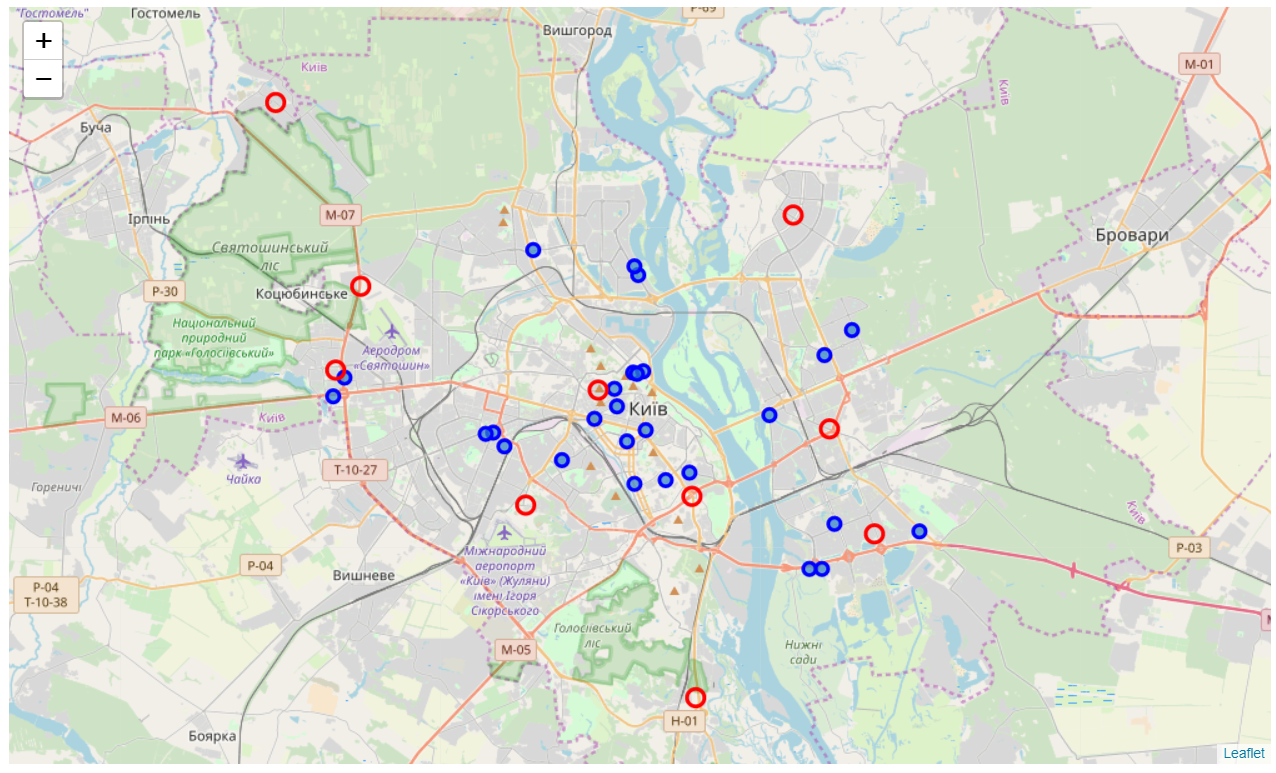


**Parameters of model** Main parameters of model that would help us score alternatives are Primary schools / Rivals / Other courses / Rent level. All this can be answered using Foursquare data. As it is can be seen request for search\_query = 'primary school' in Kyiv, returns only couple primary schools and rest is plenty of various Courses / Language Schools / Coworking etc.

1. Using category we might filter only schools and for each Neighborhood record quantity of schools in 3 km radius
2. Using category we might as well filter rivals (if it is technical education) and other courses for same age group (non-technical education). Those quantities would give us rivals and complements.
3. Finally we already have rent prices for districts though for living space. So we should normalize it and use in a model.

**Alternatives** Using same webpage olx.ua I have picked three (sometimes district did not have enough offers so sometimes it was only 1-2) alternatives of possible office spaces that can be used for starting Robotics club. I have chosen them using filter on square 90-130 m2, ready to be used (does not need renovation) and has easy access.

Let us see all these alternatives on the map of Kyiv



**Foursquare - getting the key**

So now we have coordinates of centers in all districts (call it train data) and coordinates of alternatives as well as information in the rent rate. On the next step we will explore Foursquare data for each place and its neighborhoods to find number of Schools, Entertaiments, Other courses. (I've decided not to use Restaurants/Cafe as proxy as it gives back a lot of venues quickly decreasing free daily limit)

Necessary information on categories ID were found at <https://developer.foursquare.com/docs/resources/categories>

For schools they are:

* Preschool : 52e81612bcbc57f1066b7a45
* Elementary School : 4f4533804b9074f6e4fb0105
* Middle School : 4f4533814b9074f6e4fb0106

For possible rivals they are:

* Entertainment Service : 56aa371be4b08b9a8d573554

For alternative courses:

* Language School : 52e81612bcbc57f1066b7a48
* Recreation Center : 52e81612bcbc57f1066b7a26
* Arts and Entertainment : 4d4b7104d754a06370d81259

I have found with some surprise that rivals (Technical education clubs) are presented on Foursquare not in an education related category but as an entertainment service. So I had to adopt my initial plan to build model not for rivals only expected but for overall entertainment services venues amount expected for this surrounding and than comparing it with actual. Then I would simply check all exact direct rivals locations and eliminate alternatives where rivals are less than 2 km away (distance on surface, actual via roads might be higher) from the alternative coordinates.

As for analysis we need rather count or frequency rather than list of venues lets group those dataframes by point names and venues categories.

Summarizing data section we got lists of venues of certain category which might have correlation with number of entertainment places around. However, it can be seen that Entertainment service is rare category and within districts theres 0 Entertainment services. It means that we cannot use it as a variable to predict. So I have decided to use data on schools , recreation centers and language schools to define corresponding amount of Art/Entertainment places.

Now we move to the Methodology section where I would explain how I plan to use this data for model building and detecting the best of alternatives.

### Methodology

In this section I will briefly explain how data we collected will be used to choose the best of alternatives

**Step 1.** We have count and rent rate in absolute numbers. So on the first step we divide all values by maximum expected for this column.

**Step 2.** In DF Districts we will have our "train data" and in DF Alternatives - our "test data". But before we move on we should check whether all input parameters (venue categories) make any use for regression model. We do this by checking correlation matrix

**Step 3.** Create, train and review Multiple Regression model to forecast Art&Entertainment venues basis on determinants we have selected on Step 2.

**Step 4.** Apply model to alternatives to get expected amount of Art/Entertainment venues around each of alternatives. Leave top 5 variants with highest ratio of expected / actual count of Entertainment venues. These alternatives should be the most interesting and used to choose best one of them

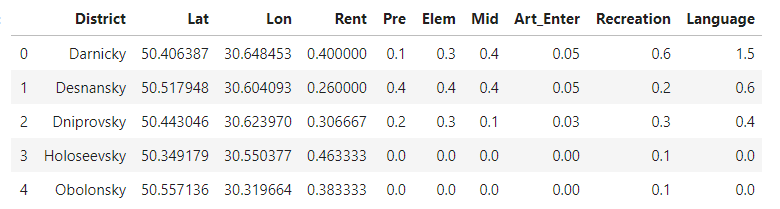
**Step 5.** Have you noticed that we have not ever mentioned direct rivals and used them in our comparison. I have collected database of existing Robotics Clubs in Kyiv and we need to define function that would calculate distance to nearest Robotics Club and apply it to alternatives. Among two equivalent alternatives one is best which has higher distance to direct competitor. Results will be discussed in corresponding section

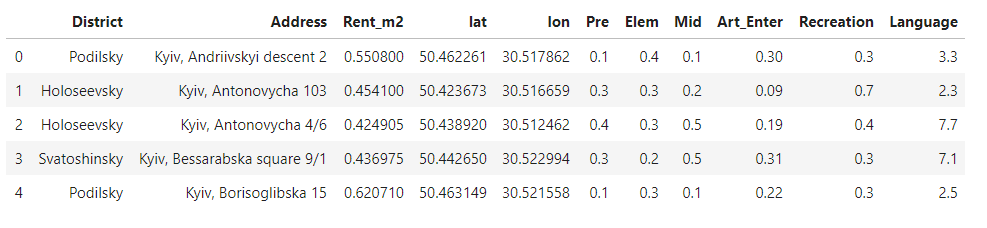
Let us now apply above methodology to our data

### Analysis

**Step 1. Input/Ouptut values**

Lets first once again look at data we collected using describe() method to check the range of values to define what how we should bring them to 0-1 range





**Step 2. Correlation matrices**

Using corr() method we can get correlation matrices but we should pay attention only to Art\_Enter and its correlation with other venues data.



It can be seen that correlation between different factors is quite strong in most cases in training (Districts) dataset. Somewhat surprisingly for me is that in both sets there is almost no correlation between rent and count of Entertainment venues. This once again assures me in need to pay attention to rent price in results section. However, for building model we would use all factors.

**Step 3. Multiple linear regression model**

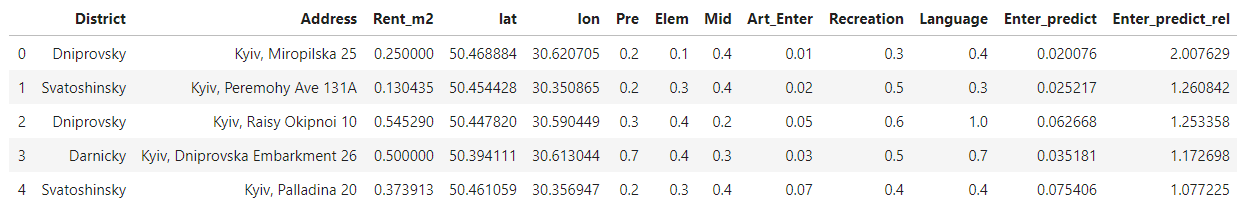
First we need to prepare train and test data and train model. Quick glance at coefficients:

regr.coef\_ 0.17295088, -0.11417027, 0.21067547, 0.17305665, -0.08236479, -0.00157314

regr.intercept\_ -0.06527889

**Step 4. Expected amount of Art/Entertainment venues**

Now we get expected Art\_Enter for each alternative and finally add in alternatives\_venues column with ratio of expected to actual Art\_Enter



**Step 5. Distance to rivals**

I will use Equirectangular approximation to calculate approximate distance between to points on a circle (Earth surface):

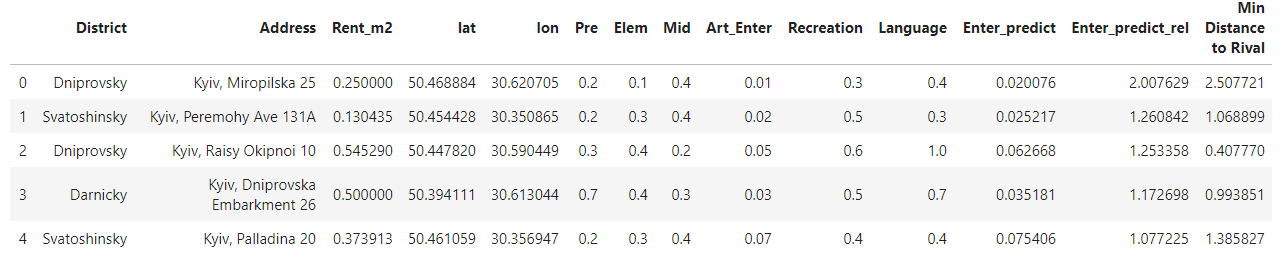
x = (λ2-λ1) \* Math.cos((φ1+φ2)/2);

y = (φ2-φ1);

dist = Math.sqrt(x*x + y*y) \* R;

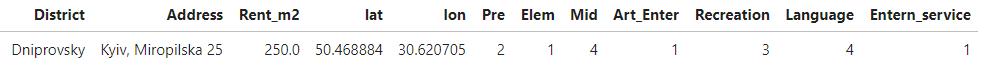
Where λ = lon.toRadians & φ = lat.toRadians

Lets look how top 5 look now and wheter leader has changed



### Results and Discussion

At the moment alternative "Kyiv, Miropilska 25" looks like the best choice with high distance to rivals and very high ratio of expected entertainment venues to actual. Both parameters are twice higher than 2nd alternative has so lets have closer look to this result.



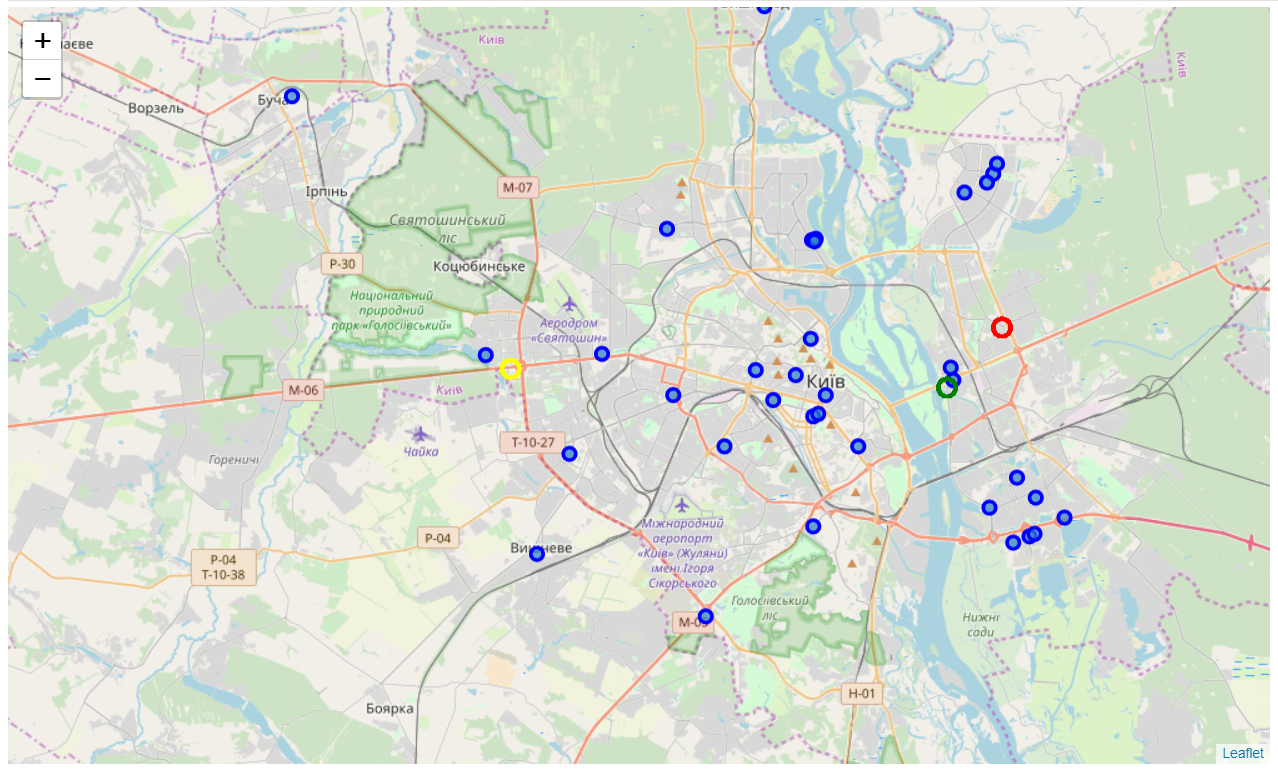
Lets put this three alternatives on Map with all real rivals (robotics clubs)

On the map below Red circle is alternative 1, Yellow circle is alternative 2 and Green circle is alternative 3 while all blue circles are existing Robotics Clubs

This once again proved that first alternative is the best one. Alternative 3 (Green) is very close to two competitors. Even more, the rent rate in alternative 3 is twice more than in alternative one.

Alternative two might be viable and it has lower rent rate and many schools in surrounding. This alternative might be use if 1 for some reason cannot be realized.

And final thoughts regarding alternative 1. It is situated far from competitors, it's a district where people spend most evenings and weekends (living area) and it has 2 preschool and 4 mid school near. This means that best choice is to open Robotics club for the full range of ages from 4 y.o. to 15 y.o.



### Conclusion

In this work choice of alternatives had to be made. Among possible 27 location across capital of Ukraine the best was chosen. To do this first Folium was used to represent alternatives on Map. Than couple of points in each district was taken to build regression model used for "scoring" alternatives. Foursquare data using https requests was used to get list (and then number) of venues in those categories that might have (and actually have) strong correlation with amount of entertainment venues.

Basis on that data and rent rates multiple linear regression model was trained on selected data to predict amount of Entertainment venues basis count of venues in other categories and rent rate.

On the final stage function to determine distance between two points on the surface of Earth was defined and used in couple with the list of existing Robotics clubs to find minimal distance to nearest rival.

Only one of alternatives happened to be outstanding, i.e. having twice larger distance to rivals and twice less entertainment venues than it should be expected for such surrounding. Moreover, the rent rate is admirable.

So the outcome of this analysis is recommendation for investor to start new Robotics Club at address: Kyiv, Miropilska 25

Good Luck!