# Sport vs Crime

March 09, 2019

## 1. Introduction

### 1.1. Background

Crime and youth antisocial behavior are complex social issues. There one of the most important risk factors is living in a deprived area. People often fall into criminal companies from adolescence. And 70 % of teenagers believe antisocial behavior occurs because young people are bored, and six in ten say that there isn't enough for young people to do in their area [https://www.sportandrecreation.org.uk/pages/gol-anti-social].

Sport can attract people and help them solve a number of problems that push them into crime:

- Developing self-regulating and problem-solving abilities as a result of developing skills needed to sport activity.
- Adventurous sport can satisfy the thirst for risk.
- Sport helps people to socialize, playing sport, a person turns into a group, also can find friends and mentors, who provide positive role models.

### 1.2. Problem

The development of sports in the city requires large investments from business, municipal and federal governments. I wanted to show that sports venues, such as sports fields, swimming pools, sports schools, can reduce crime in the city. That is, such investments are profitable for everyone.

However, this is still our guess. **I needed to check whether there is an obvious relationship between sports venues in the city and the crime rate.**

Our **audience** is businessmen and municipal government interested in reducing crime for the long-term sustainable development of their city.

## 2. Data

### 2.1. Data sources

The data about sports venues of a certain city can be found in Foursquare. In Foursquare a venue specified by a category. There are the Foursquare Venue Category Hierarchy, but there is not a sports category group, so I needed to manually select sports categories: Basketball Courts, Baseball Field, Athletics & Sports, Climbing Gym, etc.

The data about crime rates by cities can be found in Wikipedia, for example, https://en.wikipedia.org/wiki/List_of_United_States_cities_by_crime_rate. Also, this page gets the population of cities.

To search venues in a city using Foursquare API, I apply the geolocator GeoPy to translate the state name + city name to geolocation (latitude, longitude).

### 2.2. Data cleaning

1. The table of crime rates has a few missing values. I dropped rows (cities) with gaps in this table.

2. Initially, I planned to use the rating of venues from Foursquare. However, it turned out that the venue rating is included in the premium calls. Free account tier provides only 50 premium calls / day. It is too little to get a rating of hundreds of venues.

3. Big cities have too much sports venues. So even for a specific category, the number of found venues exceeds the query limit. Therefore, I limited the research to cities with a population of up to 500 thousand.

### 2.3. Feature selection

As features (X) the following values are used:

- city population (thousands)
- count of sports venues / city population (thousands) by categories:
    - athletics stadium
    - badminton court
    - baseball field
    - basketball court
    - football field
    - gym / fitness
    - hockey field
    - rink
    - ski tracking
    - sport club
    - swimming school
    - tennis court
    - university gym
    - volleyball court
- total count of sports venues / city population (thousands)

The target (Y) is the crime rate of city = total number of crimes per year per number of inhabitants.

### 2.4. Struct of an instance

| Filed | Type | Used to create a model | Feature / Target |
|---|---|---|---|
| City | String | No | - |
| Population | Float | Yes | Feature |
| Number of athletics stadiums / populations | Float | Yes | Feature |
| Number of badminton courts / populations | Float | Yes | Feature |
| … | Float | Yes | Feature |
| Number of o volleyball courts / populations | Float | Yes | Feature |
| Crime rate | Float | Yes | Target |

### 2.5. Example of an instance

'Boise', 225677, 0.005, , 0.08, , …, 0.06, **2741.97**

## 3. Exploratory Data Analysis

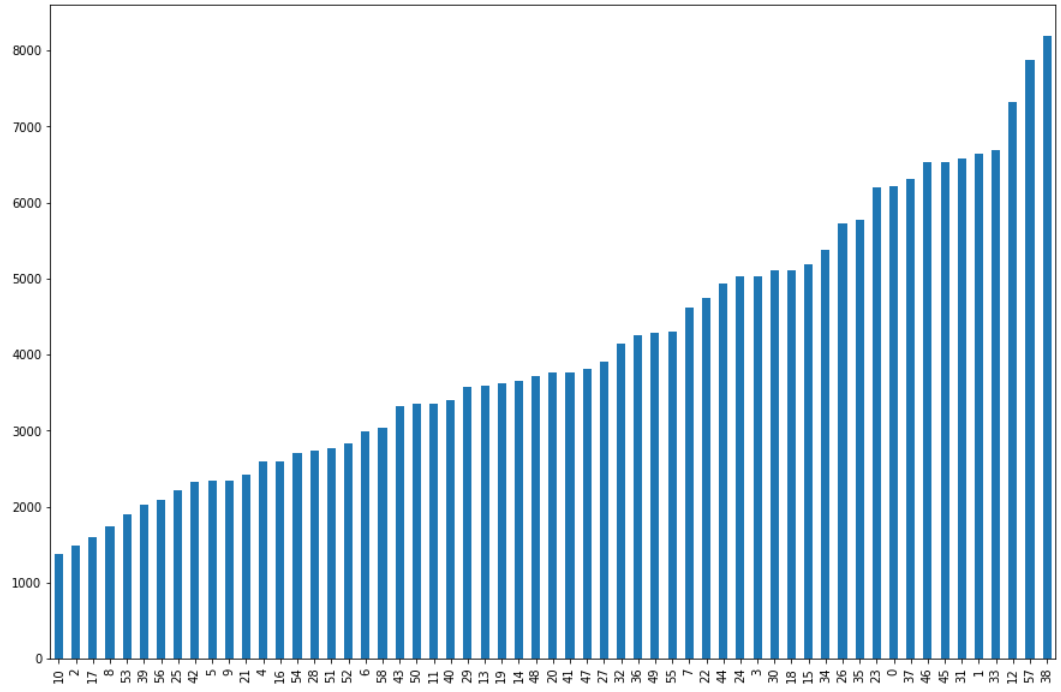
**For an impartial analysis, I do not further use the names of cities.**

After data cleaning, I get the following table (first 19 rows are listed).

| | Population | swimming school | gym / fitness | football field | rink | baseball field | hockey field | badminton court | ski tracking | tennis court | athletics stadium | volleyball court | basketball court | sport club | total_sports | Crime_Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 248431 | 0.000000 | 0.120758 | 0.016101 | 0.000000 | 0.016101 | 0.000000 | 0.000000 | 0.004025 | 0.004025 | 0.000000 | 0.000000 | 0.008051 | 0.008051 | 0.177112 | 6217.02 |
| 1 | 296188 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 6640.04 |
| 2 | 242090 | 0.028915 | 0.413070 | 0.004131 | 0.020653 | 0.095006 | 0.000000 | 0.000000 | 0.057830 | 0.037176 | 0.037176 | 0.012392 | 0.070222 | 0.012392 | 0.788963 | 1483.75 |
| 3 | 249273 | 0.004012 | 0.148432 | 0.016047 | 0.000000 | 0.076222 | 0.000000 | 0.000000 | 0.016047 | 0.020058 | 0.016047 | 0.004012 | 0.024070 | 0.008023 | 0.332968 | 5037.85 |
| 4 | 492268 | 0.006094 | 0.203141 | 0.012188 | 0.004063 | 0.069068 | 0.004063 | 0.000000 | 0.028440 | 0.022346 | 0.020314 | 0.020314 | 0.048754 | 0.008126 | 0.446911 | 2592.49 |
| 5 | 251840 | 0.007942 | 0.397078 | 0.015883 | 0.003971 | 0.162802 | 0.007942 | 0.000000 | 0.138977 | 0.051620 | 0.031766 | 0.023825 | 0.035737 | 0.023825 | 0.901366 | 2338.38 |
| 6 | 353400 | 0.002830 | 0.282965 | 0.039615 | 0.014148 | 0.181098 | 0.005659 | 0.002830 | 0.076401 | 0.033956 | 0.056593 | 0.008489 | 0.053763 | 0.005659 | 0.764007 | 2997.74 |
| 7 | 381154 | 0.000000 | 0.220383 | 0.010494 | 0.010494 | 0.013118 | 0.002624 | 0.000000 | 0.002624 | 0.015742 | 0.039354 | 0.002624 | 0.013118 | 0.010494 | 0.341069 | 4618.34 |
| 8 | 271109 | 0.003689 | 0.276641 | 0.033197 | 0.003689 | 0.114345 | 0.000000 | 0.000000 | 0.044263 | 0.014754 | 0.011066 | 0.011066 | 0.055328 | 0.014754 | 0.582791 | 1738.79 |
| 9 | 236368 | 0.016923 | 0.423069 | 0.042307 | 0.016923 | 0.071922 | 0.000000 | 0.000000 | 0.059230 | 0.029615 | 0.029615 | 0.004231 | 0.046538 | 0.016923 | 0.757294 | 2343.80 |
| 10 | 276115 | 0.018108 | 0.362168 | 0.061569 | 0.007243 | 0.126759 | 0.000000 | 0.000000 | 0.112272 | 0.112272 | 0.054325 | 0.021730 | 0.105029 | 0.028973 | 1.010449 | 1381.31 |
| 11 | 471397 | 0.006364 | 0.212135 | 0.019092 | 0.012728 | 0.046670 | 0.000000 | 0.000000 | 0.012728 | 0.042427 | 0.021214 | 0.012728 | 0.031820 | 0.014849 | 0.432756 | 3357.25 |
| 12 | 424915 | 0.004707 | 0.235341 | 0.035301 | 0.004707 | 0.087076 | 0.000000 | 0.002353 | 0.087076 | 0.056482 | 0.037655 | 0.007060 | 0.075309 | 0.004707 | 0.637775 | 7328.29 |
| 13 | 328023 | 0.000000 | 0.228643 | 0.012194 | 0.006097 | 0.045729 | 0.000000 | 0.000000 | 0.015243 | 0.012194 | 0.027437 | 0.000000 | 0.027437 | 0.006097 | 0.381071 | 3589.38 |
| 14 | 499997 | 0.004000 | 0.200001 | 0.030000 | 0.006000 | 0.048000 | 0.000000 | 0.000000 | 0.028000 | 0.026000 | 0.026000 | 0.004000 | 0.028000 | 0.024000 | 0.424003 | 3654.02 |
| 15 | 217259 | 0.000000 | 0.303785 | 0.023014 | 0.004603 | 0.041425 | 0.000000 | 0.000000 | 0.018411 | 0.023014 | 0.013808 | 0.004603 | 0.046028 | 0.004603 | 0.483294 | 5189.66 |
| 16 | 335699 | 0.017873 | 0.297886 | 0.047662 | 0.011915 | 0.157880 | 0.002979 | 0.002979 | 0.077450 | 0.050641 | 0.062556 | 0.023831 | 0.059577 | 0.020852 | 0.834081 | 2596.67 |
| 17 | 216350 | 0.004622 | 0.462214 | 0.023111 | 0.004622 | 0.023111 | 0.000000 | 0.000000 | 0.046221 | 0.046221 | 0.027733 | 0.018489 | 0.018489 | 0.013866 | 0.688699 | 1592.79 |
| 18 | 309566 | 0.000000 | 0.171207 | 0.000000 | 0.003230 | 0.022612 | 0.000000 | 0.000000 | 0.006461 | 0.006461 | 0.019382 | 0.000000 | 0.016152 | 0.006461 | 0.251966 | 5109.09 |

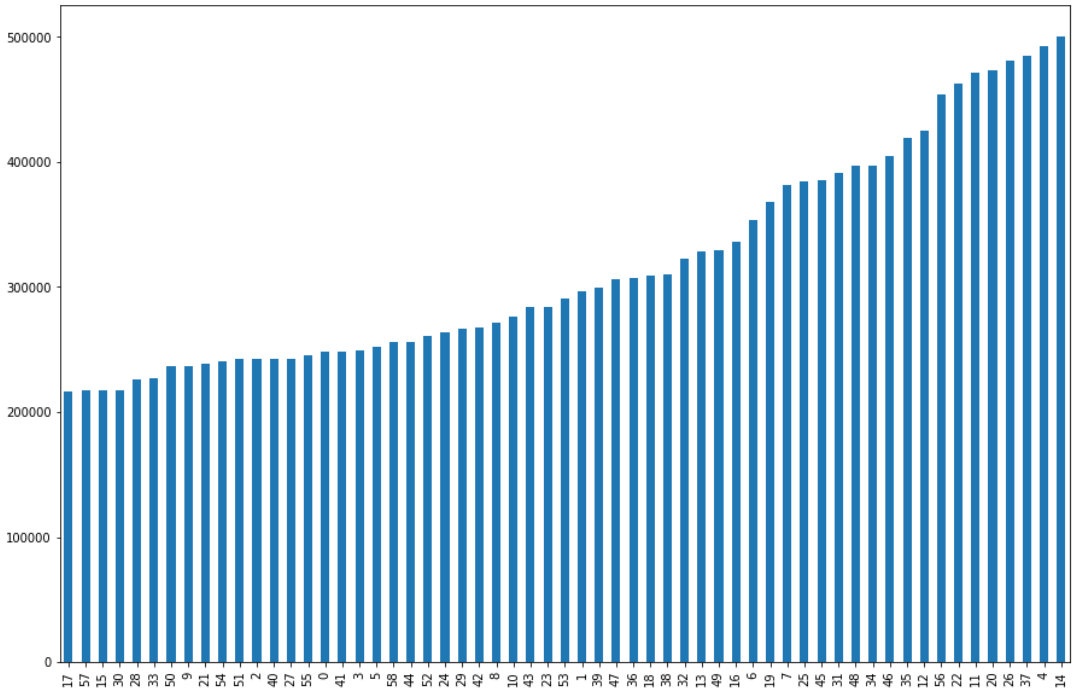
### 3.1. Histogram of the Crime Rate

The Crime Rate is unevenly distributed, the ratio of the maximum value to the minimum is almost 6 times.



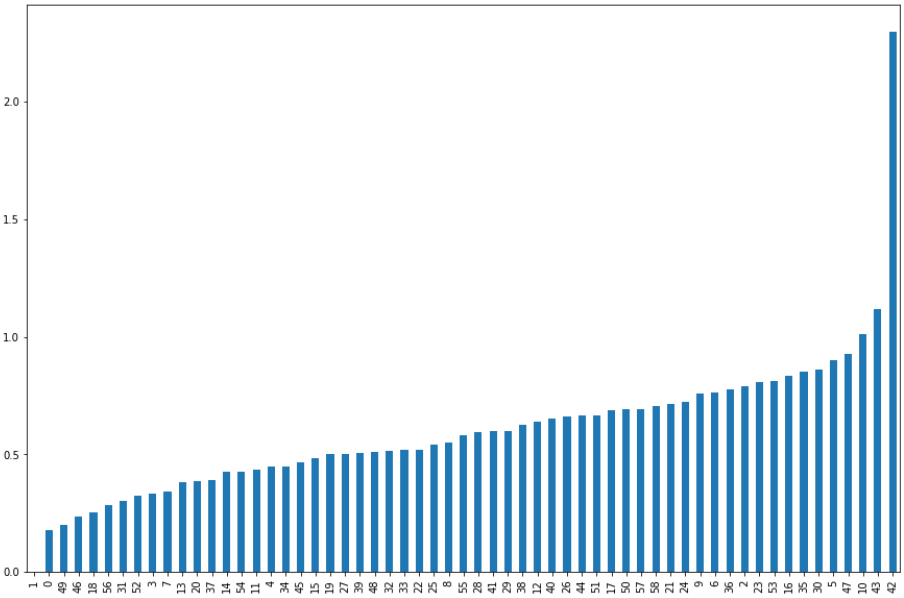

### 3.2. Histogram of Population

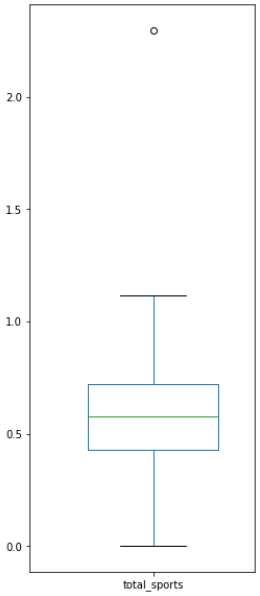
There are about 30 % of cities with the population about 220 - 250 thousand; the population linearly increases among other cities if sorted by this value.



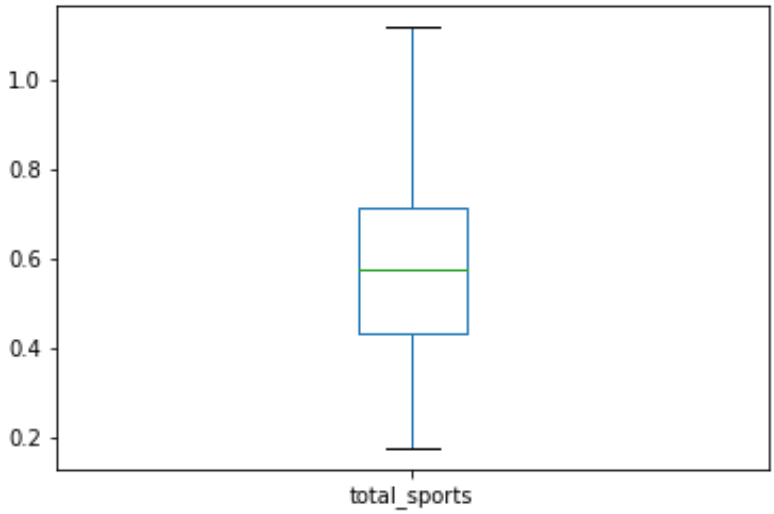
### 3.3. Histogram of Number of sports venues per inhabitant

The distribution of this indicator is very uneven. Perhaps this shows that in some cities the sports infrastructure is underdeveloped. But a zero value indicates an error when getting data from Foursquare.



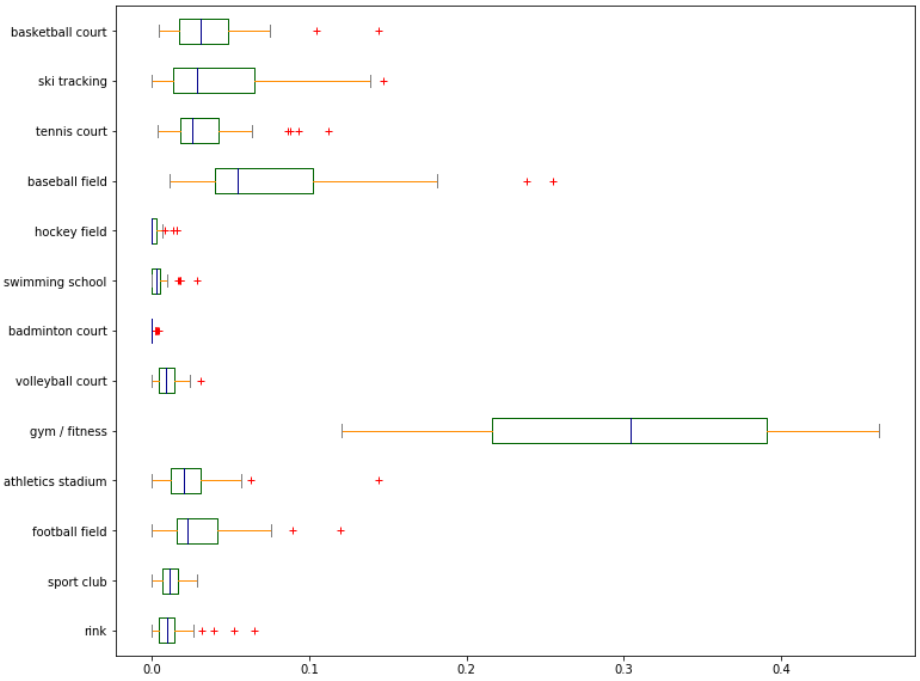



Then I removed outliners (<= 0. OR >= 2.0)



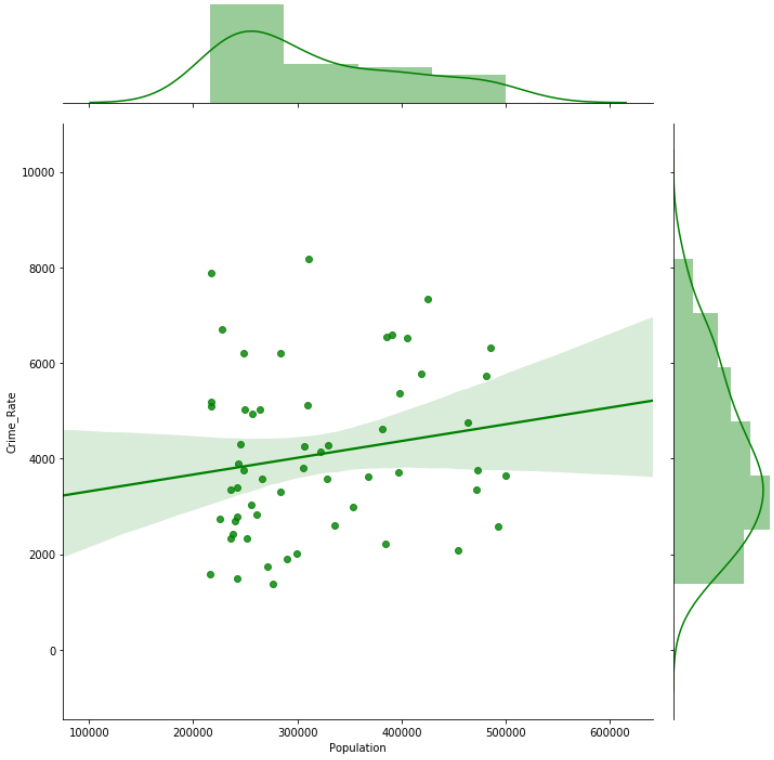
The boxes grouped by categories of venues are shown below.

Most of all, as expected, gyms / finesses. Moreover, there is a large discrepancy. But in this category, there are no such outliers as for all other.



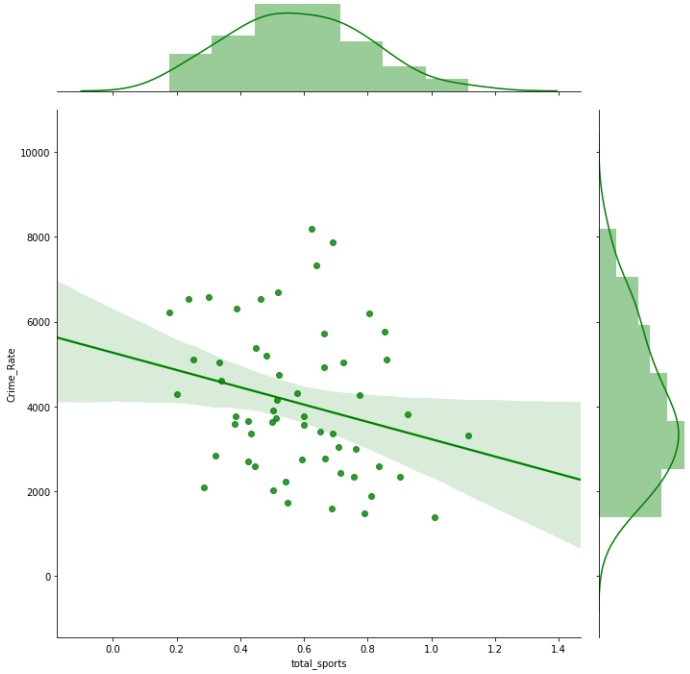

### 3.4. Relationships between crime rate and other features

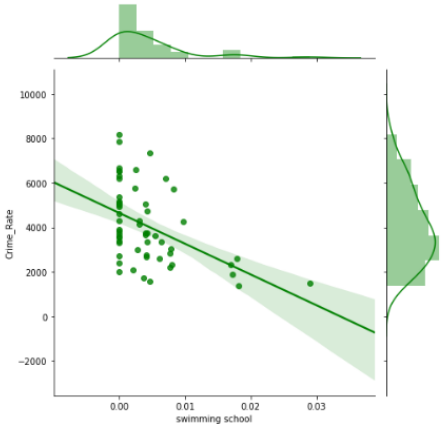
**Crime rate / population**. Dependence to a small extent can be traced, but too little data for any conclusions.

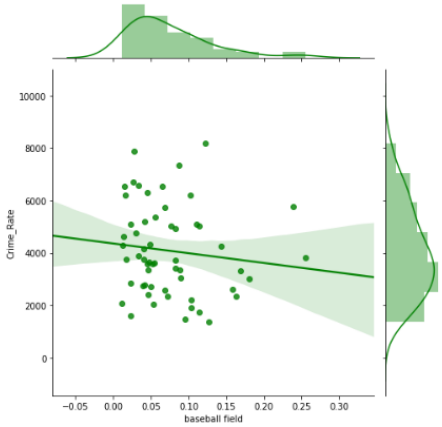


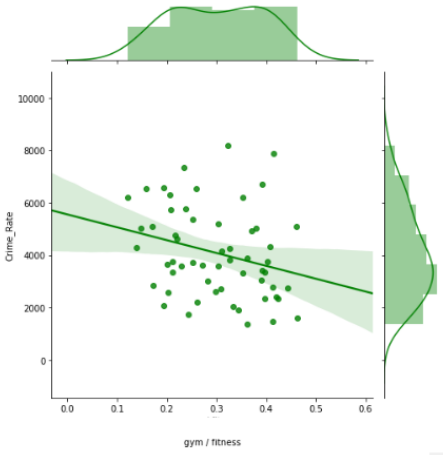
**Crime rate / Total sports venues**. It's the first result about the relationship between crime rate and the number of sports venues per inhabitant. The weak relationship can be traced. Although, of course, I was hoping for more.

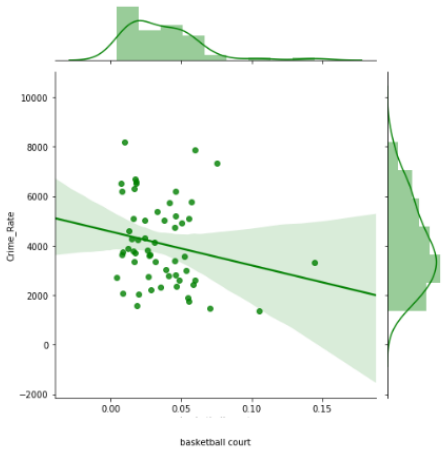
Also, I see relationships between the crime rate and separate categories.



### 3.5. Correlation

The relationships obtained is not very indicators. We can calculate the formal correlation coefficients, using Pearson product-moment correlation coefficients and NumPy library.

| Feature | Correlation |
|---|---|
| **swimming school** | **-0,47** |
| **gym / fitness** | **-0,28** |
| **total sports** | **-0,26** |
| **volleyball court** | **-0,26** |
| **basketball court** | **-0,20** |
| athletics stadium | -0,18 |
| sport club | -0,18 |
| football filed | -0,16 |
| tennis court | -0,14 |
| baseball field | -0,12 |
| ski tracking | -0,06 |
| badminton court | 0,01 |
| hockey field | 0,13 |
| rink | 0,14 |

There are weak negative correlations between the crime rate and total sports and almost all categories of sports venues. Also, there is a moderate negative correlation between the number of swimming schools and the crime rate. Of course, correlation does not mean cause-and-effect relationships, one should not forget about it.

## 4. Predictive Modeling

### 4.1. Regression problem

Initially, I was going to use regression models to predict the crime rate using statistical data about sports venues. However, this first stage of modeling shows that the amount of input data is not large enough and the correlations are not so strong to create satisfactory regression models. I tried several types of algorithms (MLP, SVR, kNN), but the results were too weak ($R^2$ score was about 0).

### 4.1. Classification Modeling

Therefore, I changed the initial task to the classification task. The features (X) now are the same as shown at 2.3. Feature selection; the target (y) is “whether the crime rate of a city is higher than the median”. Thus, I tried to predict whether the crime rate of a city is higher than the median for the dataset using population and numbers of sports venues.

I split dataset from 60 % to train and 40 % to test.

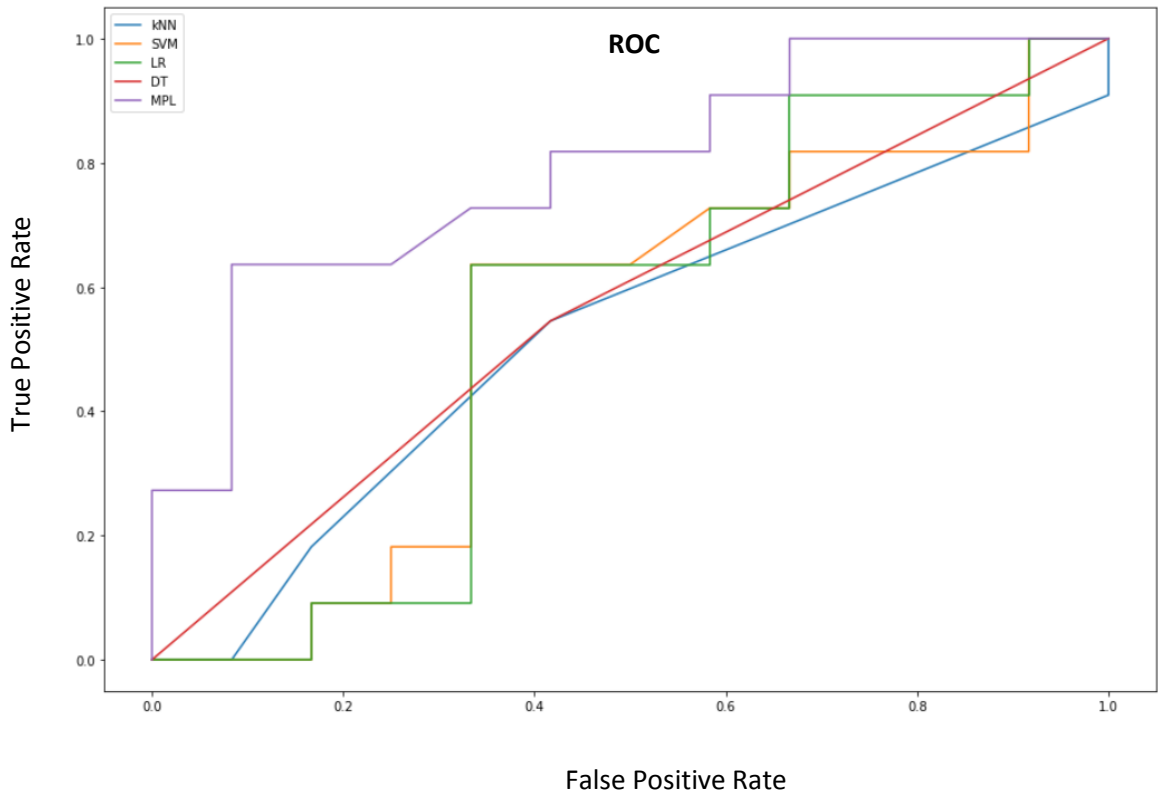
### 4.2. Results of models

I used different type of ML algorithms (implementations were provided by scikit-learn):

- k-Nearest Neighbors (kNN)
  - knn_model = KNeighborsClassifier(n_neighbors = 4).fit(X_train,y_train)
- Support Vector Machine (SVM)
  - svm_model = svm.SVC(kernel = 'rbf', degree = 3, probability=True).fit(X_train,y_train)
- Decision Tree (DT)
  - tree_model = DecisionTreeClassifier(max_depth = 2).fit(X_train, y_train)
- Logistic Regression (LR)
  - lr_model = LogisticRegression(solver = 'lbfgs', C = 0.2).fit(X_train, y_train)
- Multilayer Perceptron (MLP)
  - mpl_model = MLPClassifier(hidden_layer_sizes = (4, 1), activation = 'relu', random_state = 0).fit(X_train, y_train)

I obtained the following results on test set.

| Classifier | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| **kNN** | 57 % | 58 % | 58 % | 58 % |
| **SVM** | 65 % | 67 % | 67 % | 67 % |
| **Decision Tree** | 56 % | 58 % | 58 % | 58 % |
| **Logistic Regression** | 65 % | 67 % | 67 % | 67 % |
| **Multilayer Perceptron** | **73 %** | **71 %** | **83 %** | **77 %** |



## 5. Discussion

The study has a number of simplifications.

1. The number of cities is not large enough.
2. Some venues could be counted several times in different categories.
3. Rating of each venue is not taken into account.
4. The severity of the different types of crimes is not taken into account.

Eliminating some of the shortcomings listed above can be accomplished with improved data pre-processing. Some drawbacks are associated with the restriction of access to Foursquare using free account tier.

Nevertheless, the study shows that there is definitely a correlation between the number of sports facilities and the crime rate in the city. Of course, this factor is not the main factor, but it is logical to assume that the availability of sports facilities can divert some children and adults from various sources of negativity and even crime.

Moreover, in this study, not expensive stadiums were considered, but venues much cheaper to create: volleyball and basketball courts, gyms, etc. The strongest correlation is between sports that are available all year round and does not require costs from those who want to play sports (volleyball and basketball courts, gyms, BUT NOT ski tracking hockey field, rink).

Data on the number of sports venues allows determining whether the city's crime rate is higher than median crime rate. Quite simple classifiers such as a Logistic Regression or a Multilayer perceptron with the 4-neurons hidden layer can predict this with an accuracy of 65-75%.

## 6. Conclusion & Future directions

Despite the fact that so far only a small exploration work has been carried out in this area of the research, I dare to formulate the following conclusion.

Based on the results obtained, it should be recommended to invest in low-cost sports facilities for the long-term sustainable development of their city.

Thus, I think that the initial phase of this study presented here gives showed the promise of this study. It can be continued to produce more reasonable results.

In the future, the study can be significantly deepened by expanding the sample of output data, using not only the number of institutions but also their rating, as well as the division of crime statistics by age categories.