

Predicting The Type & Location Of a New Venue

Piotr Zubrzycki

January 11, 2020

1. Introduction/Business Problem

1.1 Background

Decision what type of place to open, and where it should be located, is crucial when preparing business plan for a new project. It is hard to keep track and understand which places have the biggest financial potential. Each year new or different businesses might be more popular than the others. Without proper analysis of the trends over the years, it is almost impossible to be sure that some particular type of venue is potentially more profitable than the others. It is also extremely important to know, where the new venue, which we would like to invest in, should be located. Because it is impossible to have an access to the financial records of each venue, it could be assumed that the number of visitors at some particular place (popularity) corresponds to the income. Also by utilisation of location data, the optimal area in reference to city centre and main train station could be chosen.

1.2 Problem

Data that might contribute to determine what business should be opened, in the particular city, based on number of citizens, might include number of visitors and its location in reference to the city centre and main train station. This project aims to predict what type of venue investor should invest in.

1.3 Interest

Everyone who thinks about investing into new venue, would be highly interested to see, what types of places are the most popular/profitable, and what is their ideal location.

2. Data acquisition

2.1 Data sources

In different cities different venue types might be popular. This is why, in order to predict what is the best investment in the city of our interest, cities with similar size/population will be chosen for statistical analysis. Information about their population, is extracted from simplemaps.com (worldcities.csv). Because this file does not include data about size of the cities, it will be assumed that size correlates with population. Next the most popular venues will be selected based on their popularity. Such information can be extracted from foursquare.com. Additionally the optimal location will also be chosen based on the coordinates. This test will be run for several years in order to see what is the trend of the most popular venues.

2.2 Data examples

Population of the cities ("worldcities.csv")

	city	city_asc ii	lat	lng	country	iso 2	iso 3	admin_name	capital	population	id
0	Tokyo	Tokyo	35.6897	139.6922	Japan	JP	JPN	Tōkyō	primary	37977000.0	1392685764
1	Jakarta	Jakarta	-6.2146	106.8451	Indonesia	ID	IDN	Jakarta	primary	34540000.0	1360771077
2	Delhi	Delhi	28.6600	77.2300	India	IN	IND	Delhi	admin	29617000.0	1356872604
3	Mumbai	Mumbai	18.9667	72.8333	India	IN	IND	Mahārāshtra	admin	23355000.0	1356226629
4	Manila	Manila	14.5958	120.9772	Philippines	PHL	PHL	Manila	primary	23088000.0	1608618140

Selecting city we would like to invest in

```
city_name = "Krakow"
```

Searched city "Krakow" exists in "worldcities.csv" file

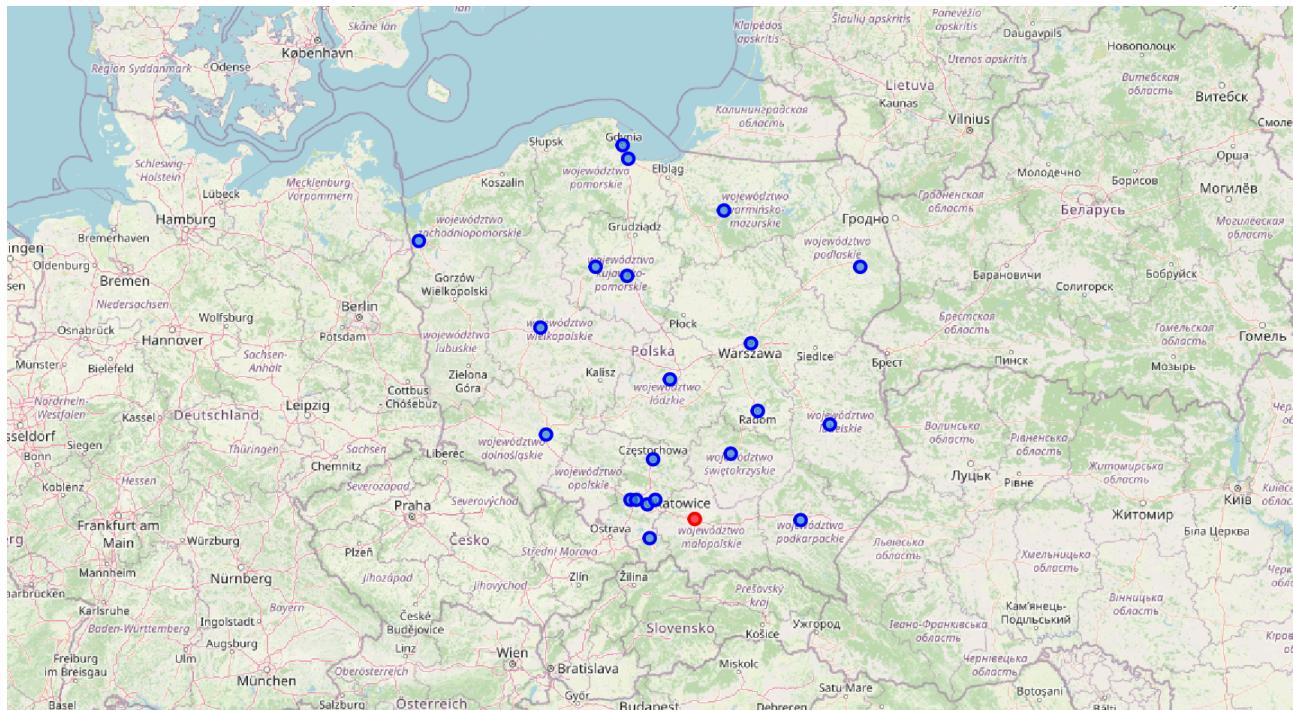
city	city_as_cii	lat	lng	country	iso2	iso3	admin_name	capital	population	id
Kraków	Krakow	50.0614	19.9372	Poland	PL	POL	Małopolskie	admin	766739.0	1616172264

Finding cities with similar population. For our test we will find 20 cities from the same geographical region/country, in this case Poland, with the closest number of citizens to "city_name" (Krakow)

	city_ascii	population	lat	lng
0	Warsaw	1790658.0	52.2167	21.0333
1	Lodz	690422.0	51.7769	19.4547
2	Wroclaw	638586.0	51.1077	17.0353
3	Poznan	538633.0	52.4000	16.9167
4	Gdansk	464254.0	54.3667	18.6333
5	Szczecin	403833.0	53.4247	14.5553
6	Bydgoszcz	358614.0	53.1167	18.0000
7	Lublin	339850.0	51.2333	22.5667
8	Bialystok	297288.0	53.1167	23.1667
9	Katowice	276499.0	50.2500	19.0000
10	Gdynia	247672.0	54.5189	18.5319
11	Czestochowa	228179.0	50.8000	19.1167
12	Radom	219703.0	51.4036	21.1567

13	Torun	202521.0	53.0167	18.6167
14	Sosnowiec	199974.0	50.3000	19.1667
15	Kielce	191605.0	50.8725	20.6319
16	Rzeszow	182548.0	50.0500	22.0000
17	Gliwice	179806.0	50.2976	18.6766
18	Zabrze	174349.0	50.3000	18.7833
19	Olsztyn	173599.0	53.7800	20.4942
20	Bielsko-Biala	172781.0	49.8225	19.0444

Depicting all cities on the map (Krakow marked on red):



This test was originally performed on the data from cities around the world. Unfortunately most of the data from "analysis cities" come from Asia/China what potentially could bias the result significantly. Thus decision was made to limit cities to the same country (alternatively test could be conducted for the cities from the same region/continent).

New we will check what are the most popular venues in selected Cities. In order to do so Foursquare API will be used to explore the cities and segment them. Venues from radius of 5000 meters will be searched. Data will be extracted for. Following time periods:

(Foursquare VERSION = Date format: YYYYMMDD)

VERSION: 20130731
 VERSION: 20131031
 VERSION: 20140731
 VERSION: 20141031
 VERSION: 20150731
 VERSION: 20151031
 VERSION: 20160731
 VERSION: 20161031
 VERSION: 20170731
 VERSION: 20171031
 VERSION: 20180731
 VERSION: 20181031
 VERSION: 20190731
 VERSION: 20191031

Below table consist of data examples for various cities and dates.

City Name	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Date
Warsaw	52.2167	21.0333	Palace On The Isle (Pałac Łazienkowski (Pałac ...	52.214860	21.035599	Palace	20130731
Czestochowa	50.8000	19.1167	Gourmet del Corso	50.812061	19.109529	Modern European Restaurant	20130731
Szczecin	53.4247	14.5553	Browar Stara Komenda	53.421642	14.550926	Brewery	20131031
Warsaw	52.2167	21.0333	Gringo Bar Burritos Tacos & More	52.201305	21.020496	Burrito Place	20140731
Czestochowa	50.8000	19.1167	M1	50.836951	19.116623	Shopping Mall	20140731
Szczecin	53.4247	14.5553	Perugia	53.441056	14.489863	Pizza Place	20141031
Warsaw	52.2167	21.0333	Egurrola Dance Studio	52.227045	20.999160	Dance Studio	20150731
Radom	51.4036	21.1567	Aldi	51.404122	21.140704	Market	20150731
Szczecin	53.4247	14.5553	Cukiernia-Piekarnia Woźniak	53.422653	14.491217	Bakery	20151031
Warsaw	52.2167	21.0333	Locale	52.192363	21.003873	Italian Restaurant	20160731

Torun	53.0167	18.6167	Hanza Café	53.010479	18.602070	Café	20160731
Bydgoszcz	53.1167	18.0000	Starbucks	53.115696	18.010736	Coffee Shop	20161031
Lodz	51.7769	19.4547	Angelo	51.768582	19.455180	Italian Restaurant	20170731
Torun	53.0167	18.6167	Hotel Spichrz	53.008747	18.608722	Hotel	20170731
Bydgoszcz	53.1167	18.0000	HTI PÓŁNOC	53.091638	18.057030	Miscellaneous Shop	20171031
Lodz	51.7769	19.4547	Park im. ks. Józefa Poniatowskiego	51.753155	19.442303	Park	20180731
Torun	53.0167	18.6167	Media Markt	53.025928	18.637707	Electronics Store	20180731
Lublin	51.2333	22.5667	Ogród Saski	51.248799	22.547158	Park	20181031
Lodz	51.7769	19.4547	Tel Aviv Urban Food	51.762699	19.457760	Vegetarian / Vegan Restaurant	20190731
Sosnowiec	50.3000	19.1667	McDonald's	50.318312	19.107605	Fast Food Restaurant	20190731

Now let's find out how many unique categories/venues there are in each city:

There are 209 uniques categories.

Next let's display the top 10 venues for 2 example cities and all dates:

City Name	Date	1st Most Comm on Venue	2nd Most Comm on Venue	3rd Most Comm on Venue	4th Most Comm on Venue	5th Most Comm on Venue	6th Most Comm on Venue	7th Most Comm on Venue	8th Most Comm on Venue	9th Most Comm on Venue	10th Most Comm on Venue
Bialystok	201307	Hotel	Pub	Restaurant	Pizza Place	Shopping Mall	Coffee Shop	Grocery Store	Fast Food Restaurant	Sushi Restaurant	Multiple x
Bialystok	201310	Hotel	Pub	Restaurant	Pizza Place	Shopping Mall	Coffee Shop	Grocery Store	Fast Food Restaurant	Sushi Restaurant	Multiple x
Bialystok	201407	Hotel	Pub	Restaurant	Pizza Place	Shopping Mall	Coffee Shop	Grocery Store	Fast Food Restaurant	Sushi Restaurant	Multiple x

Bialystok	201410	Hotel	Pub	Restaurant	Pizza Place	Shopping Mall	Coffee Shop	Grocery Store	Fast Food Restaurant	Sushi Restaurant	Multiple x
Bialystok	201507	Hotel	Pub	Restaurant	Pizza Place	Shopping Mall	Coffee Shop	Grocery Store	Fast Food Restaurant	Sushi Restaurant	Multiple x
...
Zabrze	201710	Fast Food Restaurant	Gas Station	Shopping Mall	Museum	Supermarket	Hotel	Polish Restaurant	Creperie	Clothing Store	Electronics Store
Zabrze	201807	Fast Food Restaurant	Gas Station	Shopping Mall	Museum	Supermarket	Hotel	Polish Restaurant	Creperie	Clothing Store	Electronics Store
Zabrze	201810	Fast Food Restaurant	Gas Station	Shopping Mall	Museum	Supermarket	Hotel	Polish Restaurant	Creperie	Clothing Store	Electronics Store
Zabrze	201907	Fast Food Restaurant	Gas Station	Shopping Mall	Museum	Supermarket	Hotel	Polish Restaurant	Creperie	Clothing Store	Electronics Store
Zabrze	201910	Fast Food Restaurant	Gas Station	Shopping Mall	Museum	Supermarket	Hotel	Polish Restaurant	Creperie	Clothing Store	Electronics Store

3. Data Analysis

Main issue with extracted data from foursquare.com is that they do not consist of number of "check ins" into particular venue (I can't see such information in json file). It significantly limits whole analysis. Because we do not have this data, it will be assumed, that most viable option from the business perspective, will correlate with the venue which statistically occurs the most in analysed cities.

3.1 Data Manipulation

Checking what are statistically the most popular venues, in all analysed cities, for each 'Date'. This way we can see, how trend changes over time

Below to 10 venues are listed for each date

Date	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
20130731	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza
20131031	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza
20140731	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza
20141031	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza
20150731	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza
20151031	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza
20160731	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza
20161031	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza
20170731	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza
20171031	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza
20180731	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza

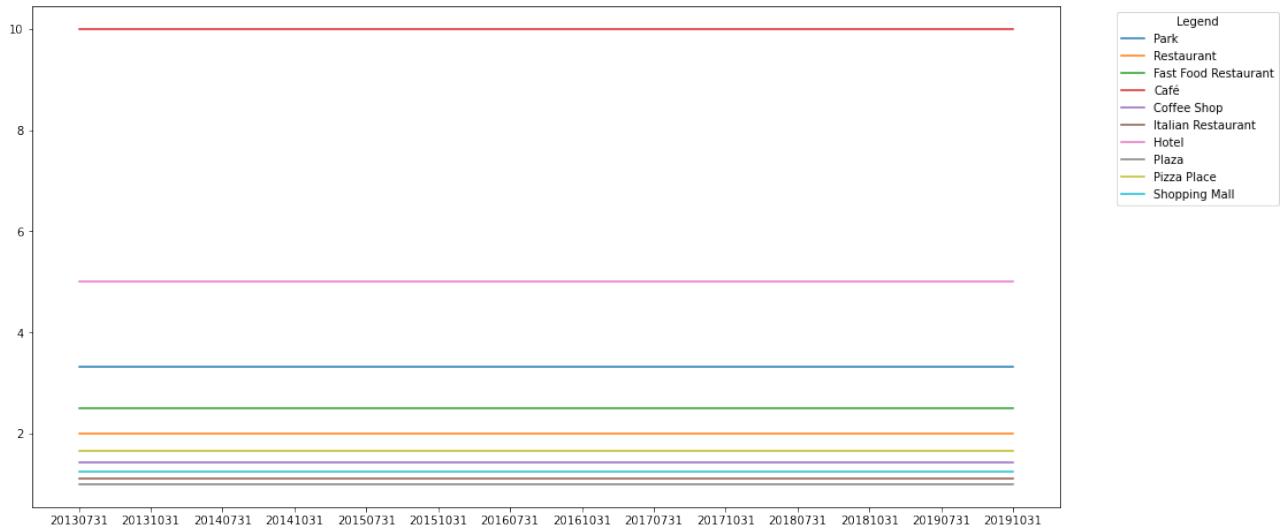
20181031	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza
20190731	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza
20191031	Café	Hotel	Park	Fast Food Restaurant	Restaurant	Pizza Place	Coffee Shop	Shopping Mall	Italian Restaurant	Plaza

From above table we can see that for each date there are exactly the same top 10 venues. That can indicate that foursquare.com database does not consist of data for each time period but just the latest.

To plot data easier, numbers from 1 to 10 will be assigned to the venues from above table, based its popularity:

Date	Park	Restaurant	Fast Food Restaurant	Café	Coffee Shop	Italian Restaurant	Hotel	Plaza	Pizza Place	Shopping Mall
20130731	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20131031	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20140731	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20141031	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20150731	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20151031	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20160731	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20161031	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20170731	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20171031	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20180731	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20181031	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20190731	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3
20191031	3.3	2.0	2.5	10.0	1.4	1.1	5.0	1.0	1.7	1.3

Plotting the result



3.2 Observations

From above graph it can be noticed that the most popular venue is of type "Cafe". Based on above data it looks like there is no change in "popularity" over time among Venue Types. Looks like "foursquare" database does not consist of the data from various periods of time for analysed region.

3.3 Finding the best location for the "most popular" Venue

Now when we know what is the most popular Venue let's find its average distance form the city centre.

Table below contains few examples of the city centre and cafe coordinates:

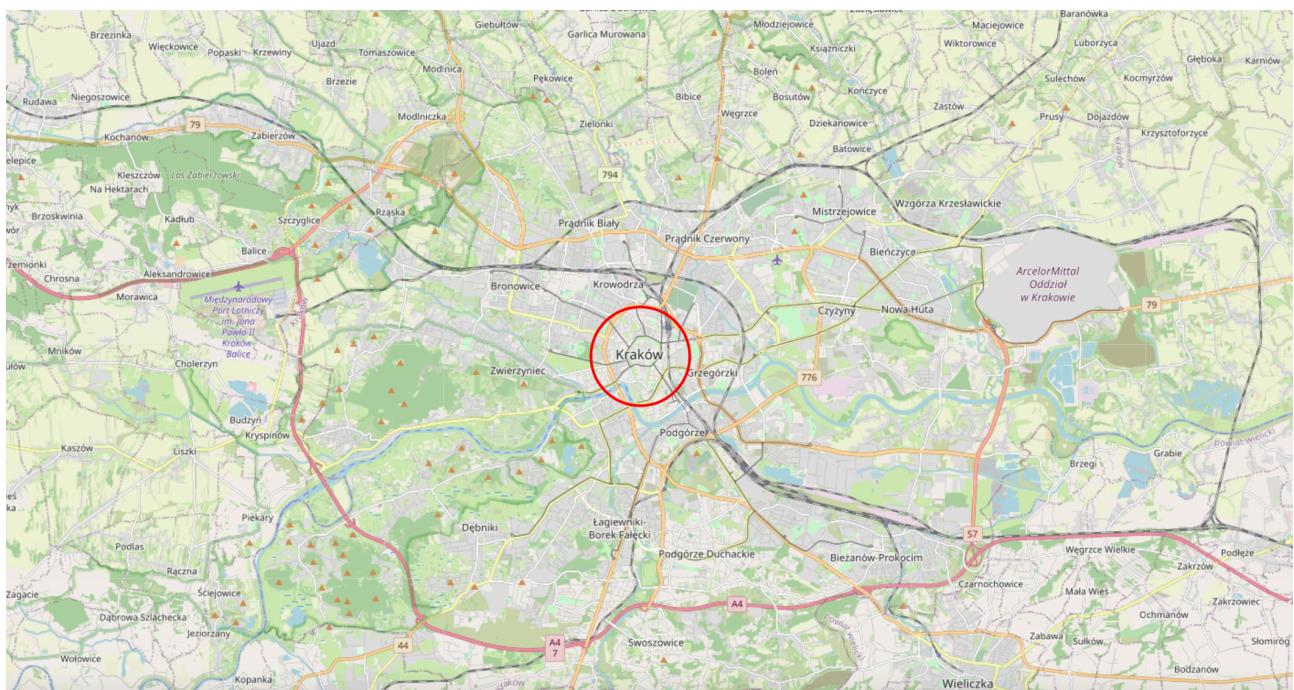
Venue Category	City Latitude	City Longitude	Venue Latitude	Venue Longitude
Café	52.2167	21.0333	52.225629	21.019006
Café	52.2167	21.0333	52.236939	21.024575
Café	52.2167	21.0333	52.233204	21.010875
Café	52.2167	21.0333	52.239527	21.022618
Café	52.2167	21.0333	52.222851	20.988527

In order to calculate the distance between city centre and "Venue Category" (Cafe) the "Haversine" formula will be used. Next based on all distances median and mean values can be calculated:

Median of the distance: 1434.04

Mean of the distance: 1618.35

3.4 Printing the area of interest over the city of interest (Krakow)



4. Conclusion

Based on the above analysis, it can be concluded, that the venue which we should invest in, is of type "Cafe". Its ideal location should be close to the city center.

Please note that due to limited number od data included in foursquare database for the analysed cities, it wasn't viable to check the trend in popularity over time, of different venues. Thus no regression was used in order to predict the best investment for the future!!!.

