Scaling offline business or where to place a new shop/bar/something else

IBM Applied Data Science Capstone Project report

by: Ivan Ilukhin

2020

Introduction

Many businesses very often face the next problem: when it begins to grow up, owners start to think about: where to open new outlets or found another place where you satisfy your customers' needs. This issue usually relates to offline companies like barbershops, veterinary clinics, retail stores, etc. There is a lot of parameters that impact on the decision: monthly rent, competitors, crime rates, a wealth of the people around and many others. This little research is focusing on the similarity of the boroughs or **How to split all districts into groups.**

The idea is based on the underlying assumption that we should found spots in similar places.

Business Problem

The objective of this research is to find similar districts and extract attributes that distinguish them. I selected the Greater Sydney Area for the region of study. It's one of the biggest agglomerations in the world, consists of almost about 650 suburbs that sprawl about 70 km to the west, 40 km to the north, and 60 to the south and generates approximately 24.1% of Australia's GDP.

Target audience of this project

The result of this project will be interesting, primarily, for the business owners and people who choose the location for expanding. It may also be of interest to investors, especially when you are presenting your startup or when you ask for money to grow your business.

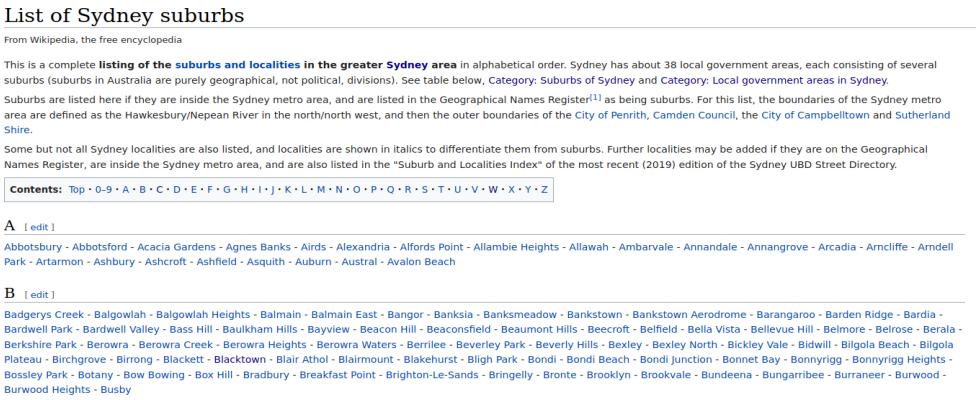
Data

For classifying suburbs, we should get features for each district. In this research, I will use the next:

- latitude and longitude to show them on a map;
- venue data, amount of each type of venues;
- crime data, about of crimes grouped by type(murder, robbery, carjacking;
- distance from the Sydney Central Business District (CBD).

Data sources

List of the suburbs will be obtained from the relevant wiki page - https://en.wikipedia.org/wiki/List_of_Sydney_suburbs



List of Sydney suburbs

From Wikipedia, the free encyclopedia

This is a complete **listing of the suburbs and localities in the greater Sydney area** in alphabetical order. Sydney has about 38 local government areas, each consisting of several suburbs (suburbs in Australia are purely geographical, not political, divisions). See table below, Category: Suburbs of Sydney and Category: Local government areas in Sydney.

Suburbs are listed here if they are inside the Sydney metro area, and are listed in the Geographical Names Register[1] as being suburbs. For this list, the boundaries of the Sydney metro area are defined as the Hawkesbury/Nepean River in the north/north west, and then the outer boundaries of the City of Penrith, Camden Council, the City of Campbelltown and Sutherland Shire.

Some but not all Sydney localities are also listed, and localities are shown in italics to differentiate them from suburbs. Further localities may be added if they are on the Geographical Names Register, are inside the Sydney metro area, and are also listed in the "Suburb and Localities Index" of the most recent (2019) edition of the Sydney UBD Street Directory.

Contents: Top · 0–9 · A · B · C · D · E · F · G · H · I · J · K · L · M · N · O · P · Q · R · S · T · U · V · W · X · Y · Z

A [edit]

Abbotsbury - Abbotsford - Acacia Gardens - Agnes Banks - Airds - Alexandria - Alfords Point - Allambie Heights - Allawah - Ambarvale - Annandale - Annangrove - Arcadia - Arncliffe - Arndell Park - Artarmon - Ashbury - Ashcroft - Ashfield - Asquith - Auburn - Austral - Avalon Beach

B [edit]

Badgerys Creek - Balgowlah - Balgowlah Heights - Balmain - Balmain East - Bangor - Banksia - Banksmeadow - Bankstown - Bankstown Aerodrome - Barangaroo - Barden Ridge - Bardia - Bardwell Park - Bardwell Valley - Bass Hill - Baulkham Hills - Bayview - Beacon Hill - Beaconsfield - Beaumont Hills - Beecroft - Belfield - Bella Vista - Bellevue Hill - Belmore - Belrose - Berala - Berkshire Park - Berowra - Berowra Creek - Berowra Heights - Berowra Waters - Berrilee - Beverley Park - Beverly Hills - Bexley - Bexley North - Bickley Vale - Bidwill - Bilgola Beach - Bilgola Plateau - Birchgrove - Birrong - Blackett - Blacktown - Blair Athol - Blairmount - Blakehurst - Bligh Park - Bondi - Bondi Beach - Bondi Junction - Bonnet Bay - Bonnyrigg - Bonnyrigg Heights - Bossley Park - Botany - Bow Bowing - Box Hill - Bradbury - Breakfast Point - Brighton-Le-Sands - Bringelly - Bronte - Brooklyn - Brookvale - Bundeena - Bungarribee - Burraneer - Burwood - Burwood Heights - Busby

Wiki page with Sydney's suburbs

Crime data are available as csv table where stored crimes count from 1995 to 2019 splitted by month for all New South Wales suburbs. Available there.

Suburb	Offence category	Subcategory	Jan 1995	Feb 1995	Mar 1995	Ap
Aarons Pass	Homicide	Murder *	0	0	0	
Aarons Pass	Homicide	Attempted murder	0	0	0	
Aarons Pass	Homicide	Murder accessory, conspiracy	0	0	0	
Aarons Pass	Homicide	Manslaughter *	0	0	0	
Aarons Pass	Assault	Domestic violence related assault	0	0	0	
Aarons Pass	Assault	Non-domestic violence related assault	0	0	0	
Aarons Pass	Assault	Assault Police	0	0	0	
Aarons Pass	Sexual offences	Sexual assault	0	0	0	
Aarons Pass	Sexual offences	Indecent assault, act of indecency and other sexual offences	0	0	0	
Aarons Pass	Abduction and kidnapping	Abduction and kidnapping	0	0	0	
Aarons Pass	Robbery	Robbery without a weapon	0	0	0	
Aarons Pass	Robbery	Robbery with a firearm	0	0	0	
Aarons Pass	Robbery	Robbery with a weapon not a firearm	0	0	0	
Aarons Pass	Blackmail and extortion	Blackmail and extortion	0	0	0	
Aarons Pass	Intimidation, stalking and harassment	Intimidation, stalking and harassment	0	0	0	
Aarons Pass	Other offences against the person	Other offences against the person	0	0	0	
Aarons Pass	Theft	Break and enter dwelling	0	1	0	

Crimes table

Coordinates of suburbs will be obtaned from the the Google Maps Geocoder API. We should make request with a subirb's name and the country.

Distance from Centrall Busines District will be calculated by applying distanse method from the geopy package on the previously getted coordinates.

Data cleaning and visualization

Let's transform the all above sources into datasets.

1. Suburbs names from the wiki page

	suburb
0	Abbotsbury
1	Abbotsford
2	Acacia Gardens
3	Agnes Banks
4	Airds

```
suburbs_df.shape
```

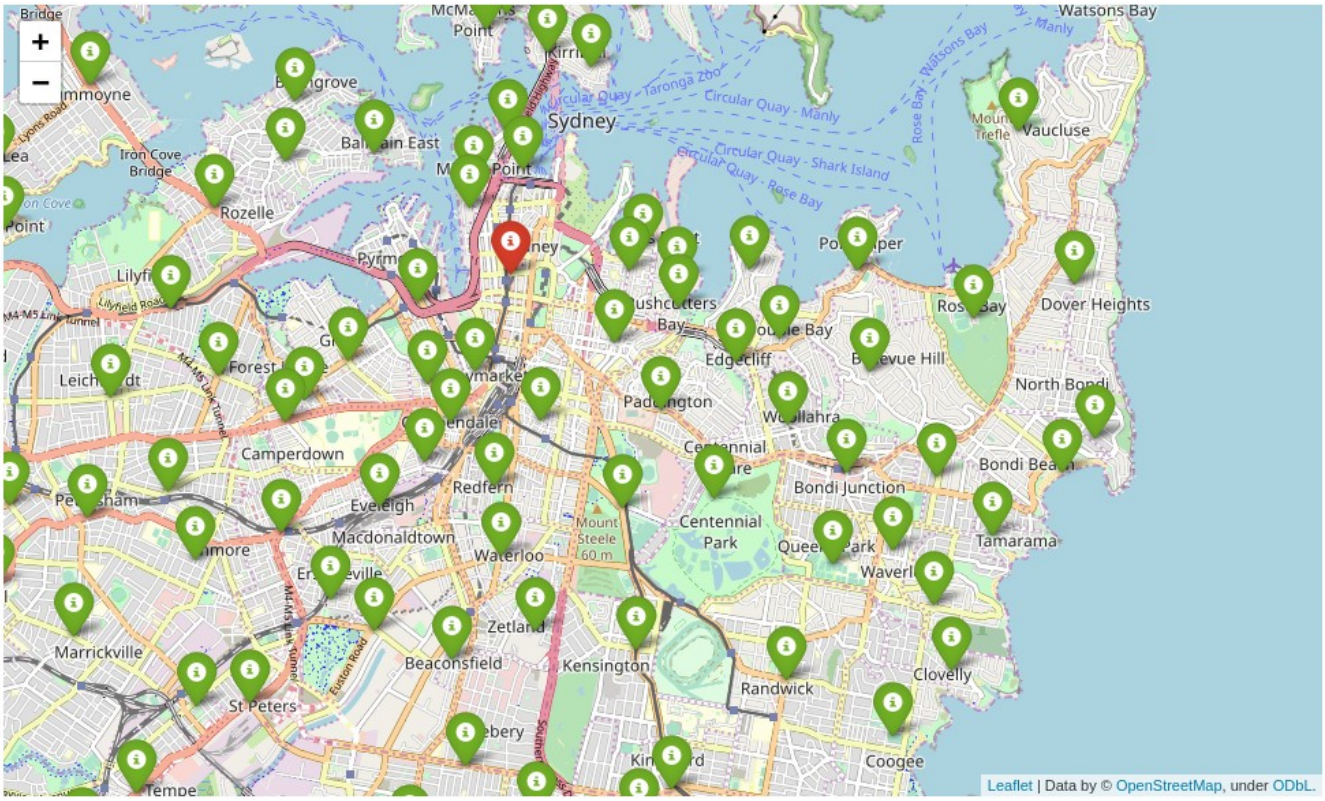
```
(689, 1)
```

2. Then get the locations using the Google Maps Geocoder API

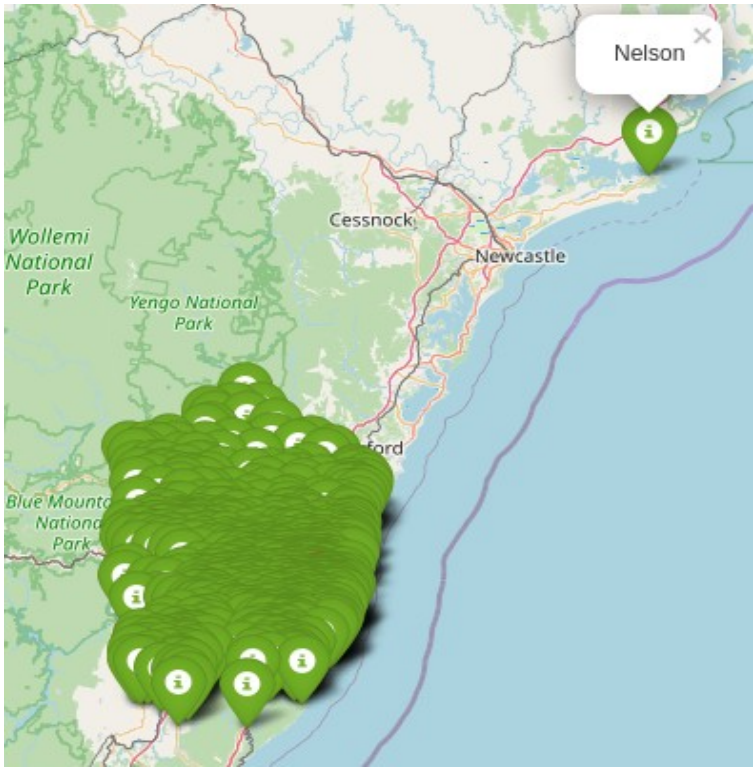
suburb	lat	lng
Abbotsbury	-33.87500	150.86200
Abbotsford	-33.85215	151.12726
Acacia Gardens	-33.73220	150.91700
Agnes Banks	-33.61529	150.71616
Airds	-34.09160	150.82490

Suburbs with latitudes and longitudes

And show all suburbs on the map.



The center of Sydney. Red mark is Cenral Business District.



All suburbs

On this map we see the outliner and immediately delete it.

3. Crude distance from the central district will be calculated based on the centres of suburbs, got on the second step.

suburb	lat	lng	dist_from_cbd
East Kurrajong	-33.515278	150.767778	56.711739
Yarramundi	-33.622200	150.670900	56.843291
Grose Wold	-33.598611	150.684722	57.072646
Tennyson	-33.536944	150.737500	57.172222
Wisemans Ferry	-33.381944	150.985000	58.018323
Grose Vale	-33.584000	150.674000	58.782859
The Slopes	-33.532801	150.706912	59.649490
Kurmond	-33.551000	150.690000	59.650771
Kurrajong	-33.550000	150.666700	61.464918
Kurrajong Hills	-33.533333	150.650000	63.804356

Farest from the CBD districts after removing the outliner

4. Monthly crime reports based on the data from the Bureau of Crime Statistics and Research.

suburb	sum_crimes	lat	lng	dist_from_cbd
Abbotsbury	3715	-33.87500	150.86200	31.953973
Abbotsford	6009	-33.85215	151.12726	7.693985
Acacia Gardens	1716	-33.73220	150.91700	30.971957
Agnes Banks	947	-33.61529	150.71616	53.617942
Airds	22071	-34.09160	150.82490	42.992898

Suburbs with coodinates, distanse and summary amount of crimes

5. And the list of venues using Foursquare API in 800 meters radius. This dataset contains suburbs with at least one venue.

suburb	ATM	Accessories Store	Advertising Agency	Afghan Restaurant	Airport	Airport Terminal	American Restaurant	Aquarium	Arcade	Arepa Restaurant	...	Wine Bar	Wine Shop	W
Abbotsbury	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
Abbotsford	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	
Acacia Gardens	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
Airds	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
Alexandria	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Yagoona	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
Yarrawarrah	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
Yennora	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
Yowie Bay	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
Zetland	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	

Venues count dataframe

6. Also add total venues to the total dataset

suburb	lat	lng	dist_from_cbd	total_venues	sum_crimes
Central Business District	-33.870846	151.207330	0.000000	241.0	354404
Haymarket	-33.880920	151.202940	1.188894	206.0	153156
Darlinghurst	-33.878018	151.220444	1.450814	184.0	95930
Barangaroo	-33.863794	151.202230	0.913537	179.0	508
Chippendale	-33.886300	151.199900	1.846810	170.0	23291

Total dataframe

7. And finally let’s create combined dataframe

suburb	lat	lng	dist_from_cbd	total_venues	sum_crimes	ATM	Accessories Store	Advertising Agency
Annangrove	-33.652550	150.940960	34.572020	1.0	938	0.0	0.0	0.0
Milsons Passage	-33.517000	151.176000	39.354880	1.0	111	0.0	0.0	0.0
Canoelands	-33.508240	151.061780	42.422889	1.0	99	0.0	0.0	0.0
Werrington	-33.759430	150.747080	44.367803	1.0	17041	0.0	0.0	0.0
Bardia	-33.978217	150.857865	34.438619	1.0	726	0.0	0.0	0.0

Methodology

Finding correlations

Main goal of this research is to find the best place for the certain kind of business. For it we will use Correlation Analysis. Pandas provides fantastic method corr for it that computes pairwise correlation of columns. I will use Pearson corellation coefficient to compare it.

First we should normalize our venues dataset – transform amount of venues of the certain type: leave zero values and remove all non zero values to 1.

```
normalized_sum_venues = sum_venues.applymap(lambda x: (1, 0)[x>0])
normalized_sum_venues.describe()
```

c

venue_category	ATM	Accessories Store	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Terminal	I
count	647.000000	647.000000	647.000000	647.000000	647.000000	647.000000	647.000000	6
mean	0.998454	0.998454	0.998454	0.998454	0.998454	0.995363	0.998454	
std	0.039314	0.039314	0.039314	0.039314	0.039314	0.067988	0.039314	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	
50%	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	
75%	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

also normalize and remove not using features from *suburbs_total_df* dataset.

suburb	dist_from_cbd	total_venues	sum_crimes
Kurrajong Hills	1.000000	0.0	0.000683
Maroota	0.791063	0.0	0.001388
Kentlyn	0.571608	0.0	0.005327
Blairmount	0.682550	0.0	0.002918
Schofields	0.556274	0.0	0.020996

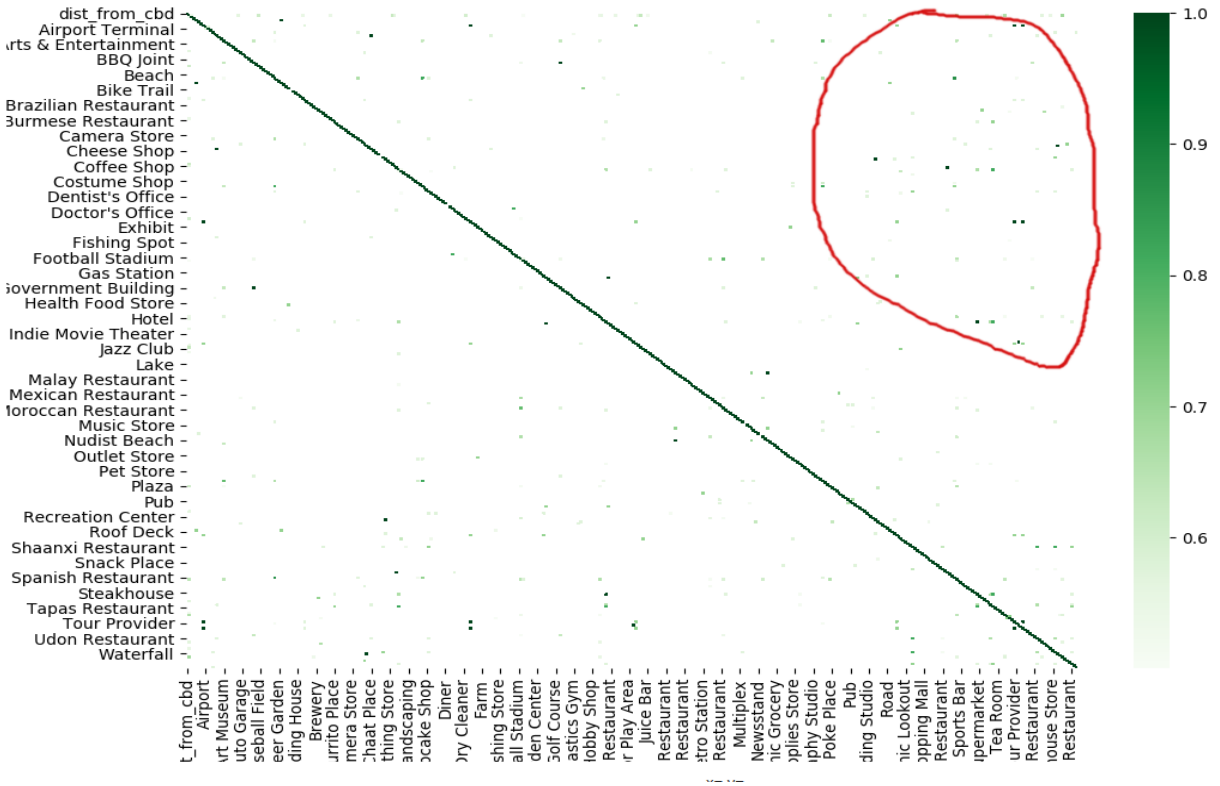
And merge them and build corellation matrix

	dist_from_cbd	total_venues	sum_crimes	ATM	Accessories Store	Advertising Agency	Afghan Restaurant	African Restaurant
dist_from_cbd	1.000000	0.476712	0.120894	0.071550	0.045454	0.046907	0.003398	0.051680
total_venues	0.476712	1.000000	0.529710	0.018692	0.023865	0.026702	0.013935	0.152954
sum_crimes	0.120894	0.529710	1.000000	0.012050	0.007774	0.013198	0.085331	0.088701
ATM	0.071550	0.018692	0.012050	1.000000	0.001610	0.001610	0.001610	0.001610
Accessories Store	0.045454	0.023865	0.007774	0.001610	1.000000	0.001610	0.001610	0.001610

Fragment of the corellation dataframe

Let's take a look at the selection. We can see that amount of crimes doesn’t depend on distance from the city center but depends from the total venues. This is fairly obvious: **more places → more people density → more crimes.**

If the Pearson correlation coefficient is between 0.5 and 1.0 it is said to be a strong correlation, so let’s drop all values smaller than 0.5 and plot a heatmap.



Heatmap for all corellations > .5

Selected area shows that there is not too much but high enough corellation between venues. Convert the matrix to a more convenient structure

category_x	category_y	correlation
Accessories Store	Beer Store	1.000000
	Rock Club	0.706537
Advertising Agency	Newsstand	0.576420
African Restaurant	Boutique	0.576420
	Egyptian Restaurant	1.000000
	Israeli Restaurant	0.706537
	Sake Bar	0.706537
	Tiki Bar	1.000000
	Trade School	1.000000
Airport	Airport Terminal	0.576420
Airport Terminal	Airport	0.576420
Aquarium	Bed & Breakfast	0.575017
	Candy Store	0.575017
	Water Park	0.575017
	Zoo	0.575017
Arcade	Gymnastics Gym	0.526315
Arepa Restaurant	Chaat Place	1.000000

Transformed dataframe

At this table we see what venues usually placed closer to the venue from the *category_x.* It's looks pretty obvious that near aquarium we can find some places for children like a candy store or a zoo.

Results

Applied Pearson correlation to our datasets we've got the following result.

There is a correlation between the type of the venue and other venues that are placed nearby, so we can choose the suburb for new business based on how many correlated places in the suburb.

For example you want to open a new Wine Bar. Let's look at these related venues.

Wine Bar	**Breakfast Spot**	0.534640
	Hostel	0.551575
	Speakeasy	0.568502
	Theater	0.578253
	Vegetarian / Vegan Restaurant	0.565202
	total_venues	0.691454

We see that usually in this suburb placed Hostels, Theaters, Breakfast Spots, Speakeasy and Vegan Restaurants. Also it depends on total_venues, that's mean that we should place it in a popular/touristic district, where are a lot of venues.

```
possible_suburbs = all_data_df[(all_data_df['Theater']>0) &
                               (all_data_df['Speakeasy']>0) &
                               (all_data_df['Breakfast Spot']>0) &
                               (all_data_df['Hostel']>0)]
possible_suburbs[['Theater', 'Hostel', 'Breakfast Spot', 'Wine Bar', 'Speakeasy']]
```

suburb	Theater	Hostel	Breakfast Spot	Wine Bar	Speakeasy
Elizabeth Bay	1.0	1.0	1.0	2.0	1.0
Dawes Point	4.0	1.0	1.0	2.0	1.0
Rushcutters Bay	1.0	1.0	2.0	3.0	2.0
Millers Point	1.0	1.0	1.0	1.0	3.0
Surry Hills	1.0	3.0	2.0	1.0	1.0
Woolloomooloo	1.0	2.0	1.0	3.0	2.0
Haymarket	2.0	3.0	1.0	2.0	1.0

We found the 7 suburbs, but unfortunately they all already have wine bars. Let's remove two of them type of venue from it. Speakeasy and Hostel for example.

suburb	Theater	Hostel	Breakfast Spot	Wine Bar	Speakeasy	total_venues
Pyrmont	2.0	0.0	2.0	0.0	0.0	115.0
Redfern	1.0	0.0	1.0	0.0	0.0	118.0

We found the two free from Wine bars suburbs. Total venues are almost the same so we can open a new bar in both.

Discussion

For further improvements, we can add census data for each suburb because many parameters depend on it. Age and mean suburb wealth are essential parameters that can strongly persuade on the decision. Also, we didn't count how long the analysed venues exist to except recently created businesses.

Venues types absolutely don't depend on a crime rate in the suburbs, but the amount of crimes depends on the venues count.

Conclusion

In this research were showed how to find a new place for your business using the correlation analysis.

Even the simplest method can solve your problem, not only huge neural networks and combinations of complex ML-algorithms, just Pearson correlation, map and scatterplot charts.