Applied Data Science Capstone Project — Sydney City Nightlife Choice

Prologue

In the recent months, many of my private plans, such as travelling to different states of Australia, were cancelled due to the Corona virus situation. Of course, this is the smallest problem one can have given the health-related, even life-threatening, and economic challenges many people face right now. I used some of the additional free time to understand the fundamentals of data science and learn how to practically use various tools and methods such as IBM Watson Studio, CRISP and Python with its diverse libraries, such as pandas, folium, numpy, geopy, Scikit learn etc. In my current positions as a Research and Development Intern at Naritas, I do not aim to regularly do data analyses and program machine learning applications myself, but knowing the basics is totally useful to work in this environment.

The final assignment of this course is the so-called "Capstone Project" in which many of the tools and methods learned throughout the recent months are applied in a self-chosen challenge around the general idea of a "Battle of Neighbourhoods", e. g. the analysis and useful preparation of data on venues in city environments. I chose the city of Sydney as a case. To pass the final assignment, creating a Jupyter Notebook with Python as the programming language and all the necessary code and comments is required, as are a blogpost or presentation and a final report. This is the required final report. Hope you like it.

Introduction/Business problem

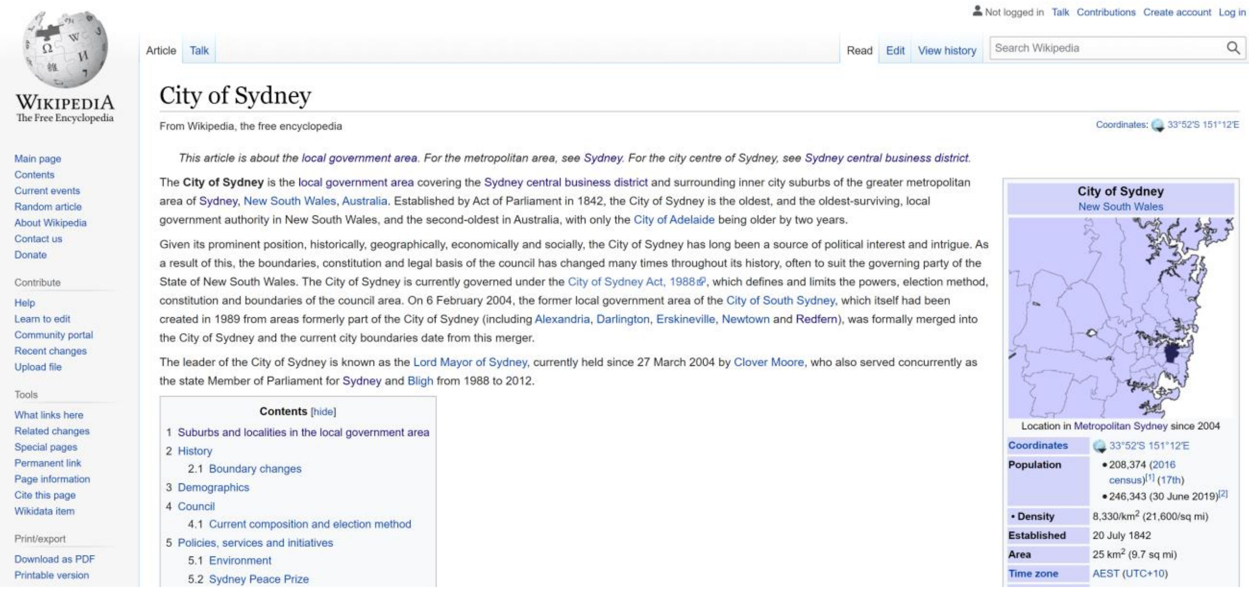
Sydney, the city the author lives in, attracts a huge number of tourists, not least due to its famous Opera House and the Harbour Bridge. Australia, where the city is situated, is not well known for nightlife. Finding the right spot to entertain at night in Sydney can be difficult and frustrating.

Thus, the problem I want to solve is to give a simple recommendation to tourists in Sydney: in which district of the city will you find the highest number of bars, pubs or lounges? The target audience are foreign tourists looking for an area in which to enjoy their nightlife.

Description of the data

I will, as requested by the assignment task, use foursquare data about nightlife in Sydney. Here is an example of a bar in Sydney on foursquare: https://foursquare.com/v/opera-bar/4b058760f964a520988e22e3 I will use data such as the bar name, ID, location, category of drinks (beers etc.?).

Also, I will use the overview of districts/city parts of Sydney from Wikipedia: https://en.wikipedia.org/wiki/City_of_Sydney.


Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

City of Sydney

From Wikipedia, the free encyclopedia

Coordinates: 33°52′S 151°12′E

This article is about the local government area. For the metropolitan area, see Sydney. For the city centre of Sydney, see Sydney central business district.

The **City of Sydney** is the local government area covering the Sydney central business district and surrounding inner city suburbs of the greater metropolitan area of Sydney, New South Wales, Australia. Established by Act of Parliament in 1842, the City of Sydney is the oldest, and the oldest-surviving, local government authority in New South Wales, and the second-oldest in Australia, with only the City of Adelaide being older by two years.

Given its prominent position, historically, geographically, economically and socially, the City of Sydney has long been a source of political interest and intrigue. As a result of this, the boundaries, constitution and legal basis of the council has changed many times throughout its history, often to suit the governing party of the State of New South Wales. The City of Sydney is currently governed under the City of Sydney Act, 1988, which defines and limits the powers, election method, constitution and boundaries of the council area. On 6 February 2004, the former local government area of the City of South Sydney, which itself had been created in 1989 from areas formerly part of the City of Sydney (including Alexandria, Darlington, Erskineville, Newtown and Redfern), was formally merged into the City of Sydney and the current city boundaries date from this merger.

The leader of the City of Sydney is known as the Lord Mayor of Sydney, currently held since 27 March 2004 by Clover Moore, who also served concurrently as the state Member of Parliament for Sydney and Bligh from 1988 to 2012.

Contents [hide]

1 Suburbs and localities in the local government area
2 History
2.1 Boundary changes
3 Demographics
4 Council
4.1 Current composition and election method
5 Policies, services and initiatives
5.1 Environment
5.2 Sydney Peace Prize

City of Sydney
New South Wales

Location in Metropolitan Sydney since 2004

Coordinates	33°52′S 151°12′E
Population	• 208,374 (2016 census)[1] (17th) • 246,343 (30 June 2019)[2]
• Density	8,330/km² (21,600/sq mi)
Established	20 July 1842
Area	25 km² (9.7 sq mi)
Time zone	AEST (UTC+10)

Here, you will find a section "Suburbs and localities in the local government area" which shows the 33 local government areas. I will use the areas and the data about nightlife in these areas from foursquare to show the density of entertaining place in them.

Methodology

In this section, I will describe the data analysis and how I used the data to yield the results.

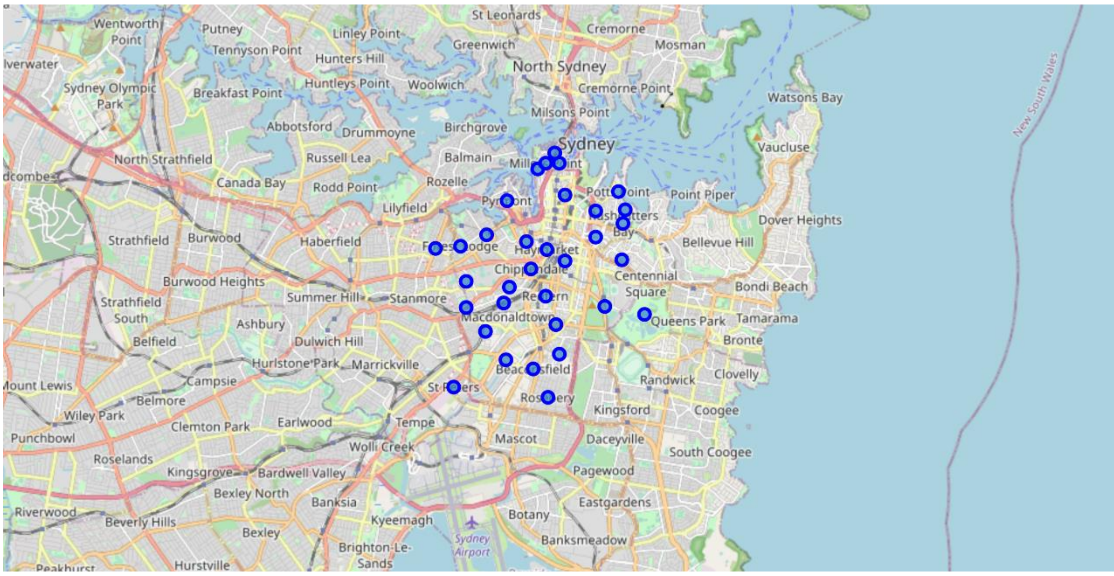
Starting out, I scraped data from Wikipedia to create a dataframe with the city districts of Sydney: https://en.wikipedia.org/wiki/City_of_Sydney. For this, I just enter the list of suburbs manually as it is not too long. The result is a nice data frame:

	City district
0	Alexandria NSW
1	Annandale NSW
2	Barangaroo NSW
3	Beaconsfield NSW
4	Camperdown NSW

Then, I enabled geopy functions by installing the conda-forge geopy package. I used the nominatim function to add geospatial data to the data frame, that is the latitude and the longitude seen on the right side of the following table.

	City district	Latitude	Longitude
0	Alexandria NSW	-33.909157	151.192128
1	Annandale NSW	-33.881224	151.170998
2	Barangaroo NSW	-33.861408	151.201688
3	Beaconsfield NSW	-33.911469	151.200315
4	Camperdown NSW	-33.889612	151.180099
5	Centennial Park NSW	-33.897778	151.233889
6	Chippendale NSW	-33.886329	151.199821
7	Darlinghurst NSW	-33.878338	151.219225
8	Darlington NSW	-33.890862	151.193216
9	Dawes Point NSW	-33.857222	151.206776

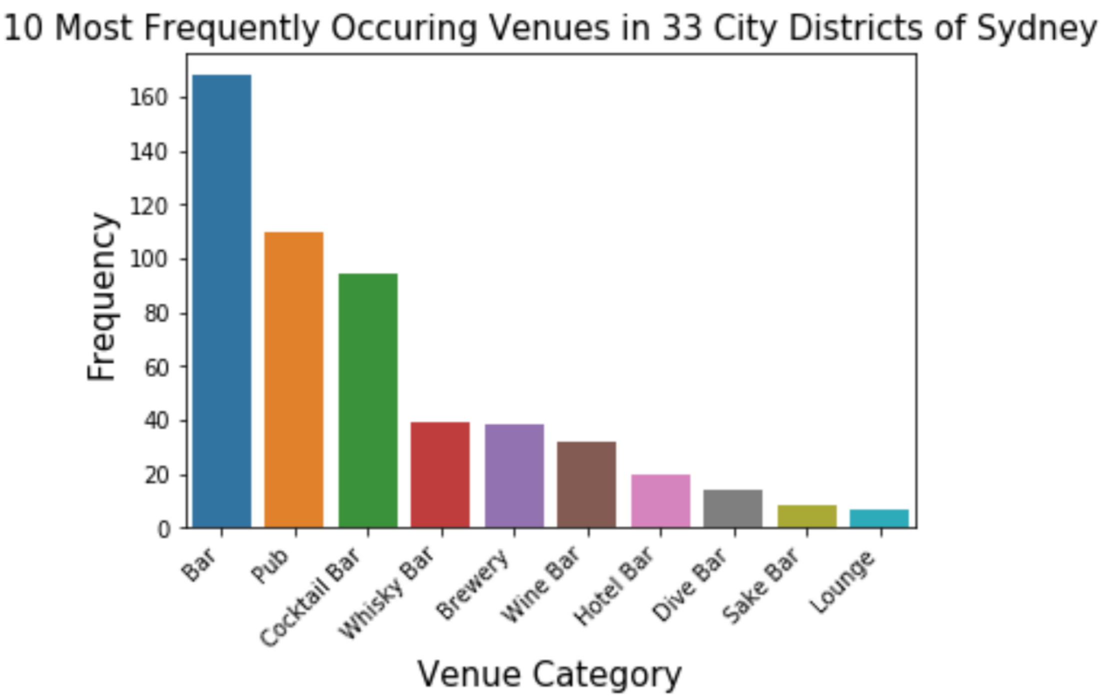
Using the folium package and my data frame, I then created a map with the 33 suburbs locations on it.



Now, foursquare data comes into play. I first did a view try-out for the city district "Alexandria NSW", which I know well, to see if the venues retrieved from foursquare seem reasonable and correct. That was the case.

Then, retrieved the foursquare data for all venues on foursquare with a distance of less than 3000 meters from each centre of each city district, as indicated as blue dots in the map above. The result was a list of 3300 venues all over Sydney city. Out of these 3300 venues, 534 where providing nightlife services. These 534 locations come from 12 unique nightlife categories, such as bar, pub or lounge.

I plotted a bar chart with the frequency of the 10 most frequently occurring nightlife in the whole city, using seaborn/matplotlib packages. We can see that bars, pubs and cocktail bars are the most frequently occurring locations for nightlife activities in Sydney.



To find clusters of nightlife types in the different city districts, I first transformed the data frame with the nightlife venues, associated to city districts, by one-hot encoding (0/1), as seen in the picture below.

	Neighborhood	Bar	Brewery	Cocktail Bar	Comedy Club	Dive Bar	Hotel Bar	Lounge	Pub	Sake Bar	Tiki Bar	Whisky Bar	Wine Bar
1	Alexandria NSW	1	0	0	0	0	0	0	0	0	0	0	0
2	Alexandria NSW	0	1	0	0	0	0	0	0	0	0	0	0
3	Alexandria NSW	1	0	0	0	0	0	0	0	0	0	0	0
4	Alexandria NSW	1	0	0	0	0	0	0	0	0	0	0	0
5	Alexandria NSW	0	0	0	0	0	0	0	1	0	0	0	0

Next, I used grouping to show the frequency of each category of restaurants in each city district.

	Neighborhood	Bar	Brewery	Cocktail Bar	Comedy Club	Dive Bar	Hotel Bar	Lounge	Pub	Sake Bar	Tiki Bar	Whisky Bar	Wine Bar
0	Alexandria NSW	0.388889	0.111111	0.166667	0.000000	0.055556	0.000000	0.000000	0.222222	0.055556	0.000000	0.000000	0.000000
1	Annandale NSW	0.428571	0.142857	0.071429	0.000000	0.000000	0.000000	0.000000	0.285714	0.000000	0.000000	0.000000	0.071429
2	Barangaroo NSW	0.133333	0.066667	0.333333	0.000000	0.000000	0.133333	0.000000	0.266667	0.000000	0.000000	0.066667	0.000000
3	Beaconsfield NSW	0.437500	0.062500	0.062500	0.000000	0.062500	0.000000	0.000000	0.250000	0.062500	0.000000	0.062500	0.000000
4	Camperdown NSW	0.352941	0.117647	0.176471	0.000000	0.058824	0.000000	0.000000	0.176471	0.058824	0.000000	0.000000	0.058824
5	Centennial Park NSW	0.222222	0.000000	0.000000	0.111111	0.000000	0.000000	0.111111	0.333333	0.000000	0.000000	0.000000	0.222222
6	Chippendale NSW	0.411765	0.000000	0.235294	0.000000	0.058824	0.000000	0.000000	0.117647	0.000000	0.000000	0.117647	0.058824
7	Darlinghurst NSW	0.166667	0.000000	0.250000	0.000000	0.000000	0.083333	0.000000	0.166667	0.000000	0.000000	0.083333	0.250000
8	Darlington NSW	0.416667	0.083333	0.166667	0.000000	0.041667	0.000000	0.000000	0.166667	0.041667	0.000000	0.083333	0.000000
9	Dawes Point NSW	0.133333	0.066667	0.333333	0.000000	0.000000	0.133333	0.066667	0.200000	0.000000	0.000000	0.066667	0.000000
10	Elizabeth Bay NSW	0.142857	0.000000	0.285714	0.000000	0.000000	0.071429	0.000000	0.214286	0.000000	0.000000	0.071429	0.214286
11	Erskineville NSW	0.346154	0.192308	0.115385	0.000000	0.038462	0.000000	0.000000	0.192308	0.038462	0.000000	0.000000	0.076923
12	Eveleigh NSW	0.458333	0.083333	0.166667	0.000000	0.041667	0.000000	0.000000	0.125000	0.041667	0.000000	0.041667	0.041667
13	Forest Lodge NSW	0.500000	0.142857	0.142857	0.000000	0.000000	0.000000	0.000000	0.142857	0.000000	0.000000	0.000000	0.071429
14	Glebe NSW	0.312500	0.062500	0.250000	0.000000	0.000000	0.000000	0.000000	0.187500	0.000000	0.000000	0.125000	0.062500
15	Haymarket NSW	0.400000	0.000000	0.133333	0.000000	0.000000	0.066667	0.000000	0.133333	0.000000	0.000000	0.200000	0.066667
16	Millers Point NSW	0.125000	0.062500	0.312500	0.000000	0.000000	0.125000	0.062500	0.250000	0.000000	0.000000	0.062500	0.000000

I used this information to create a data frame in which you can see the most common nightlife venue types for each city district.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8
0	Alexandria NSW	Bar	Pub	Cocktail Bar	Brewery	Sake Bar	Dive Bar	Wine Bar	
1	Annandale NSW	Bar	Pub	Brewery	Wine Bar	Cocktail Bar	Whisky Bar	Tiki Bar	
2	Barangaroo NSW	Cocktail Bar	Pub	Hotel Bar	Bar	Whisky Bar	Brewery	Wine Bar	
3	Beaconsfield NSW	Bar	Pub	Whisky Bar	Sake Bar	Dive Bar	Cocktail Bar	Brewery	
4	Camperdown NSW	Bar	Pub	Cocktail Bar	Brewery	Wine Bar	Sake Bar	Dive Bar	
5	Centennial Park NSW	Pub	Wine Bar	Bar	Lounge	Comedy Club	Whisky Bar	Tiki Bar	
6	Chippendale NSW	Bar	Cocktail Bar	Whisky Bar	Pub	Wine Bar	Dive Bar	Tiki Bar	
7	Darlinghurst NSW	Wine Bar	Cocktail Bar	Pub	Bar	Whisky Bar	Hotel Bar	Tiki Bar	
8	Darlington NSW	Bar	Pub	Cocktail Bar	Whisky Bar	Brewery	Sake Bar	Dive Bar	
9	Dawes Point NSW	Cocktail Bar	Pub	Hotel Bar	Bar	Whisky Bar	Lounge	Brewery	

Now, with all this data, I could finally run an unsupervised machine learning algorithm, more specifically, a k-means clustering algorithm from the scikit-learn package. One could use the elbow method to systematically define the k value, but I simply chose k to be 5, having been inspired by one of the Coursera courses to do so.

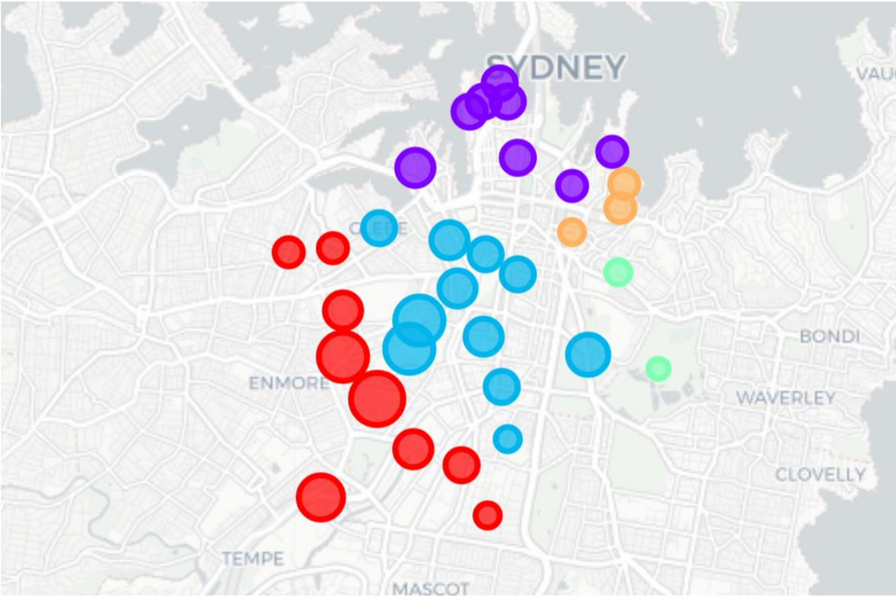
Results

And here already comes the result:

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8tl Common
0	0	Alexandria NSW	Bar	Pub	Cocktail Bar	Brewery	Sake Bar	Dive Bar	Wine Bar	Whis
1	0	Annandale NSW	Bar	Pub	Brewery	Wine Bar	Cocktail Bar	Whisky Bar	Tiki Bar	Sa
2	1	Barangaroo NSW	Cocktail Bar	Pub	Hotel Bar	Bar	Whisky Bar	Brewery	Wine Bar	T
3	0	Beaconsfield NSW	Bar	Pub	Whisky Bar	Sake Bar	Dive Bar	Cocktail Bar	Brewery	Wi
4	0	Camperdown NSW	Bar	Pub	Cocktail Bar	Brewery	Wine Bar	Sake Bar	Dive Bar	Whis
5	3	Centennial Park NSW	Pub	Wine Bar	Bar	Lounge	Comedy Club	Whisky Bar	Tiki Bar	Sa
6	2	Chippendale NSW	Bar	Cocktail Bar	Whisky Bar	Pub	Wine Bar	Dive Bar	Tiki Bar	Sa
7	4	Darlinghurst NSW	Wine Bar	Cocktail Bar	Pub	Bar	Whisky Bar	Hotel Bar	Tiki Bar	Sa
8	2	Darlington NSW	Bar	Pub	Cocktail Bar	Whisky Bar	Brewery	Sake Bar	Dive Bar	Wi

What we see in the table are the city districts and their most common venues, and they now have been assigned five different cluster labels from 0 to 4.

We can now use the cluster labels to show the city districts marked with a cluster-specific colour on a map, again using folium:



You will see 33 bubbles for the 33 city areas, with five different colours for the five different clusters.

Now, what is the result of this exercise? We now can show five clusters of nightlife type concentrations for the city of Sydney, which I named according to the nightlife concentration the data shows.

Cluster 1 - the Southern Pub Cluster (Red)

	Latitude	2nd Most Common Venue	3rd 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
0	-33.909157	Pub	Cocktail Bar	Brewery	Sake Bar	Dive Bar	Wine Bar	Whisky Bar	Tiki Bar	Lounge
1	-33.881224	Pub	Brewery	Wine Bar	Cocktail Bar	Whisky Bar	Tiki Bar	Sake Bar	Lounge	Hotel Bar
3	-33.911469	Pub	Whisky Bar	Sake Bar	Dive Bar	Cocktail Bar	Brewery	Wine Bar	Tiki Bar	Lounge
4	-33.889612	Pub	Cocktail Bar	Brewery	Wine Bar	Sake Bar	Dive Bar	Whisky Bar	Tiki Bar	Lounge
11	-33.902172	Pub	Brewery	Cocktail Bar	Wine Bar	Sake Bar	Dive Bar	Whisky Bar	Tiki Bar	Lounge
13	-33.880556	Pub	Cocktail Bar	Brewery	Wine Bar	Whisky Bar	Tiki Bar	Sake Bar	Lounge	Hotel Bar
18	-33.896113	Pub	Brewery	Wine Bar	Tiki Bar	Dive Bar	Cocktail Bar	Whisky Bar	Sake Bar	Lounge
23	-33.918586	Pub	Sake Bar	Dive Bar	Brewery	Wine Bar	Whisky Bar	Tiki Bar	Lounge	Hotel Bar
25	-33.915947	Pub	Bar	Cocktail Bar	Wine Bar	Sake Bar	Lounge	Whisky Bar	Tiki Bar	Hotel Bar

Cluster 2 - the Northern Pub Cluster (Purple)

	Latitude	2nd Most Common Venue	3rd 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
2	-33.861408	Pub	Hotel Bar	Bar	Whisky Bar	Brewery	Wine Bar	Tiki Bar	Sake Bar	Lounge
9	-33.857222	Pub	Hotel Bar	Bar	Whisky Bar	Lounge	Brewery	Wine Bar	Tiki Bar	Sake Bar
16	-33.859913	Pub	Hotel Bar	Bar	Whisky Bar	Lounge	Brewery	Wine Bar	Tiki Bar	Sake Bar
20	-33.867080	Pub	Whisky Bar	Bar	Hotel Bar	Wine Bar	Tiki Bar	Sake Bar	Lounge	Dive Bar
21	-33.869214	Cocktail Bar	Bar	Whisky Bar	Hotel Bar	Brewery	Wine Bar	Tiki Bar	Sake Bar	Lounge
27	-33.867957	Pub	Bar	Whisky Bar	Hotel Bar	Brewery	Wine Bar	Tiki Bar	Sake Bar	Lounge
28	-33.859992	Pub	Hotel Bar	Bar	Whisky Bar	Lounge	Brewery	Wine Bar	Tiki Bar	Sake Bar
31	-33.871876	Bar	Whisky Bar	Pub	Wine Bar	Hotel Bar	Tiki Bar	Sake Bar	Lounge	Dive Bar

Cluster 3 - the Central Cluster (Blue)

	Latitude	2nd Most Common Venue	3rd 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
6	-33.886329	Cocktail Bar	Whisky Bar	Pub	Wine Bar	Dive Bar	Tiki Bar	Sake Bar	Lounge	Hotel Bar
8	-33.890862	Pub	Cocktail Bar	Whisky Bar	Brewery	Sake Bar	Dive Bar	Wine Bar	Tiki Bar	Lounge
12	-33.895000	Cocktail Bar	Pub	Brewery	Wine Bar	Whisky Bar	Sake Bar	Dive Bar	Tiki Bar	Lounge
14	-33.877778	Cocktail Bar	Pub	Whisky Bar	Wine Bar	Brewery	Tiki Bar	Sake Bar	Lounge	Hotel Bar
15	-33.881441	Whisky Bar	Pub	Cocktail Bar	Wine Bar	Hotel Bar	Tiki Bar	Sake Bar	Lounge	Dive Bar
17	-33.895833	Pub	Wine Bar	Whisky Bar	Lounge	Dive Bar	Comedy Club	Cocktail Bar	Tiki Bar	Sake Bar
22	-33.893104	Pub	Whisky Bar	Cocktail Bar	Wine Bar	Dive Bar	Tiki Bar	Sake Bar	Lounge	Hotel Bar
26	-33.884512	Whisky Bar	Pub	Cocktail Bar	Wine Bar	Hotel Bar	Dive Bar	Comedy Club	Tiki Bar	Sake Bar
29	-33.879473	Whisky Bar	Pub	Cocktail Bar	Hotel Bar	Wine Bar	Tiki Bar	Sake Bar	Lounge	Dive Bar
30	-33.900276	Pub	Whisky Bar	Cocktail Bar	Dive Bar	Brewery	Wine Bar	Tiki Bar	Sake Bar	Lounge
32	-33.907662	Whisky Bar	Pub	Dive Bar	Cocktail Bar	Brewery	Wine Bar	Tiki Bar	Sake Bar	Lounge

Cluster 4 - the Eastern Bar Cluster (Green)

	Latitude	2nd Most Common Venue	3rd 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
5	-33.897778	Wine Bar	Bar	Lounge	Comedy Club	Whisky Bar	Tiki Bar	Sake Bar	Hotel Bar	Dive Bar
19	-33.884157	Pub	Bar	Whisky Bar	Lounge	Cocktail Bar	Tiki Bar	Sake Bar	Hotel Bar	Dive Bar

Cluster 5 - the North-eastern Cluster (Yellow)

	Latitude	2nd Most Common Venue	3rd 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
7	-33.878338	Cocktail Bar	Pub	Bar	Whisky Bar	Hotel Bar	Tiki Bar	Sake Bar	Lounge	Dive Bar
10	-33.871691	Wine Bar	Pub	Bar	Whisky Bar	Hotel Bar	Tiki Bar	Sake Bar	Lounge	Dive Bar
24	-33.875037	Wine Bar	Cocktail Bar	Bar	Whisky Bar	Hotel Bar	Tiki Bar	Sake Bar	Lounge	Dive Bar

Interestingly, it is possible to define clusters of certain services in Sydney city. People living in Sydney will probably agree that these clusters sound reasonable and are not too far away from what you would have expected.

Discussion

If I reflect the work necessary to create these results, what comes to my mind is that for typical ways of scraping, cleaning, handling, transforming and visualizing data, all the tools are simply there. We just have to get to know the available open-source packages and learn how to use them. What I find fantastic is that nearly all of them are free of charge. Also, a simple notebook computer is enough: in my case, I used a Dell XPS 15 9570, more than three years old. All the rest is concentrated, creative, interesting, sometimes hard work and searching for hints, tips, examples, explanations etc. in the web. With these tools, many exciting data science use cases can be created, for all kinds of useful purposes.

Conclusion

We achieved the goal presented at the outset of this report: tourists can see in the results which city districts best match their nightlife activities regarding of distances. This is just one example of fantastic data science uses cases one can realize applying technology which is available for free today! What a time to be alive.

Acknowledgement & sources

Several publications have inspired this piece of work and helped me develop the skills to run this analysis and the difficult coding behind. Also, when running into difficulties, it is common to borrow a few fragments of code, as long as you fully understand them, change and apply them to your needs, and name the source - at least if we are talking about non-commercial use for qualification purposes. Amongst these sources are:

The courses of the IBM Data Science Professional Certificate itself and the plethora of hours I spent with them: https://www.Coursera.org/professional-certificates/ibm-data-science Especially, courses number 7 "Data Visualization with Python" and 8 "Machine Learning with Python" played an important role here.

The examples for outstanding solutions for the capstone project mentioned on Coursera and others I found in the web where also inspiring and helpful: https://www.linkedin.com/pulse/housing-sales-prices-venues-data-analysis-ofistanbul-sercan-y%C4%B1ld%C4%B1z/, https://medium.com/@radialee/capstone-project-the-battle-of-neighborhoods-in-tokyo-restaurants-45a503e65ff, and several more.

Of course, also a number of cheat sheets I found on github (https://github.com/), stack overflow (https://stackoverflow.com/), simplyanalytical (https://simpleanalytical.com/) etc. helped.

Thus, I thank the overall, worldwide data science and machine learning community for sharing so much useful stuff openly with the public and me!