CAPSTONE PROJECT: BATTLE OF THE NEIGHBOURHOODS

Exploring Airbnb Listings in Sydney

1. Problem and background

Nowadays, people have become used to staying in even a stranger's place when traveling to a new place, as this gives them a greater feeling of being at home and allows them to experience a local's way of living. This has spawned the birth of many online renting platforms, among which Airbnb is undoubtedly a world leader, which has made it possible for travellers to find places to stay directly from individuals in thousands of cities around the world.

Airbnb was founded in 2008 with a few listings in the U.S. Since then, it has expanded globally to over 190 countries and regions with more than seven million listings worldwide (Airbnb, 2019). Many travellers, especially young travellers, prefer Airbnb for its flexibility and various options regarding price, place, type of accommodation, etc. Even business travellers are starting to choose Airbnb: the number of companies using Airbnb had increased from only 250 in 2015 to over 250,000 in 2017 (Zaleski, 2017).

Since I live in Sydney, which also happens to be a very popular tourist destination, I am interested in finding out what the listings are like in this city. Specifically, I would like to find out:

a) what are the busiest times of the year for Airbnb hosts
b) distribution of prices and number of listings in each neighbourhood
c) distribution of the most and least expensive listings geographically, and if possible, are there any patterns among these listings
d) create clusters for listings located in Sydney CBD area using Foursquare

Answers to these questions will provide valuable insights for people intending to travel to Sydney, as they will be able to make informed decisions about which neighbourhood to stay given their budget, preferred nearby facilities, etc. On the other hand, for prospective hosts, they will know when to put their properties on the market and what price to set for their listings so as to be competitive.

2. Description of the data

The dataset was obtained from Kaggle's Sydney Airbnb open data (web address here), originally sourced from publicly available information from the Airbnb site (web address here).

The dataset consists of several csv documents, but the ones of interest to me are the *calendar_dec18* and *listings_summary_dec18* files. These two contain all the essential information needed for this project, such as listing availability, name, location, room type, price, reviews per month, etc., shown in **Table 1**. Besides, the dataset contains one geojson file, which can be used to create choropleth graphs. A choropleth map is a type of thematic map in which areas are shaded in proportion to the density of a statistical variable.

Table 1: first five listings in each of the two datasets

	listing_id	date	available	price
0	14250	2019-12-06	t	$470.00
1	12351	2019-08-17	t	$110.00
2	12351	2019-08-16	t	$110.00
3	12351	2019-08-15	t	$110.00
4	12351	2019-08-14	t	$110.00

	id	name	neighbourhood	latitude	longitude	room_type	price	reviews_per_month	availability_365
0	12351	Sydney City & Harbour at the door	Sydney	-33.865153	151.191896	Private room	100	4.83	187
1	14250	Manly Harbour House	Manly	-33.800929	151.261722	Entire home/apt	471	0.03	321
2	15253	Stunning Penthouse Apartment In Heart Of The City	Sydney	-33.880455	151.216541	Private room	109	3.63	316
3	20865	3 BED HOUSE + 1 BED STUDIO Balmain	Leichhardt	-33.859072	151.172753	Entire home/apt	450	0.18	69
4	26174	COZY PRIVATE ROOM, GREAT LOCATION!	Woollahra	-33.889087	151.259404	Private room	62	0.45	140

Furthermore, as mentioned previously, I will divide the listings into different clusters based on their nearby facilities, which will involve using services from Foursquare. Foursquare is a platform that aggregates, by location, data of numerous venues including restaurants, museums, shopping malls, etc. The data available on Foursquare includes basic information such as venue's name, category, location latitude and longitude, which will aid our analysis, as well as other non-essential details such as menu, user rating and review, etc.

3. Exploratory Data Analysis

3.1 Pre-processing the dataset

Although both datasets are overall well-structured, there are a few minor issues that should be addressed before doing any further analysis.

For the calendar dataset, date is of type object, which should be converted to datetime. Besides, price information is displayed with a currency symbol and is treated as strings. Some values also have a comma between the numbers. So, we need to remove all these special characters and then convert the price to type float.

Regarding the second dataset, firstly, the monthly reviews information for some listings are not available. In other words, these listings have not received any reviews yet. But in the data frame, the inputs are expressed as NaNs, which need to be replaced with zeros. Secondly, a small number of listings have a price of zero, which obviously is not realistic. Thus, it is necessary to remove the corresponding rows. Once these initial cleanings are done, we can move onto performing some exploratory analysis.

3.2 Availability of listings in different months

The calendar dataset contains information on whether a listing is available or not on any day between Dec 2018 and Dec 2019. Letter t means available and f means not. We can replace t with number 1 and f with number 0 and calculate their respective counts in each month of the one-year period. This will inform us in which months the listings are easily available and what are the busiest months.

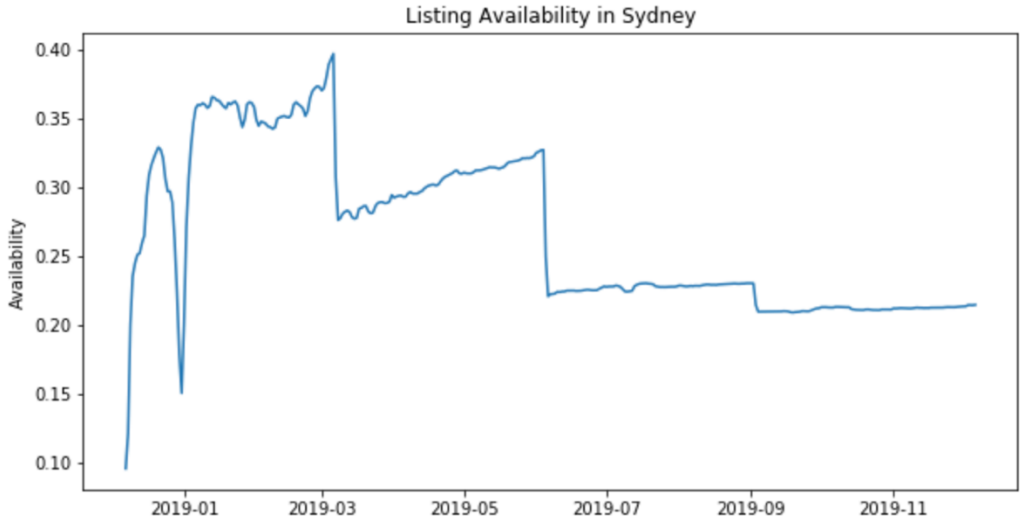


Figure 1

This implies that the second half of the year is actually the busiest time for Airbnb hosts, with most listings being booked out. And months from January to March are the easiest time to find a listing, indicating a low tourist season. This is counter-intuitive as Sydney's high season actually falls between Dec. and Feb, because of its location in the Southern hemisphere and tourists flood the city to take advantage of the warm weather here.

I am guessing although many of the listings are marked as unavailable from June 2019 and onwards, this is not because they have been all booked out. It might be simply for the reason that these dates are too far into the future and the listings are not open for booking yet. Remember for our dataset the starting date is Dec. 7 2018.

3.3 Plotting number of listings and average price in each neighbourhood

As of December 2018, there have been more than 36,000 listings in the city of Sydney, scattered across 38 different neighbourhoods. As the second part of our exploratory analysis, I decided to look at how the listings are spread across these neighbourhoods and what the average prices are like in each neighbourhood. This can be easily achieved via pandas' groupby and sort functions, and the visualization can be created using the matplotlib library, with the results shown below:

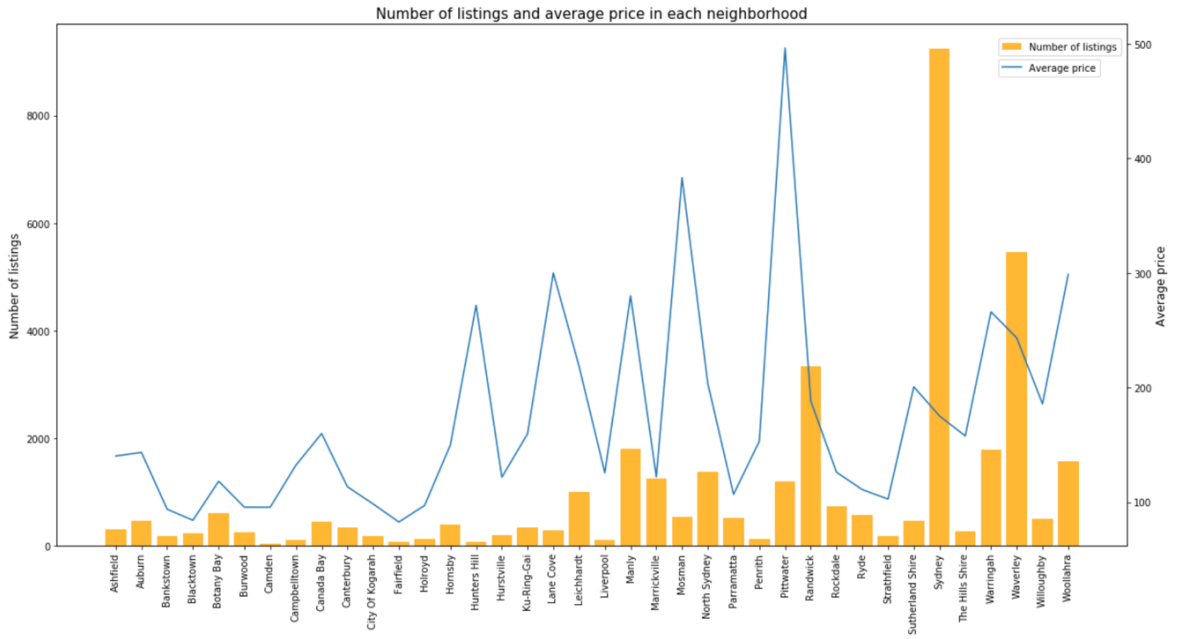


Figure 2

From this graph, we can see that Sydney, Waverley, and Randwick have the largest number of listings, whereas the average price is the highest in Pittwater, Mosman, and Land Cove.

Here, we can also make use of the geojson file contained in our dataset to visualize the above information, which is more straightforward and visually appealing. First, we create a choropleth graph showing the density of listings in each neighbourhood. Then we do the same for average price in each neighbourhood. Note that since the geojson file is a bit too large, the outputs cannot be rendered in jupyter notebook, so it is necessary to save the outputs to HTML files and then open them directly in a browser such as Chrome.

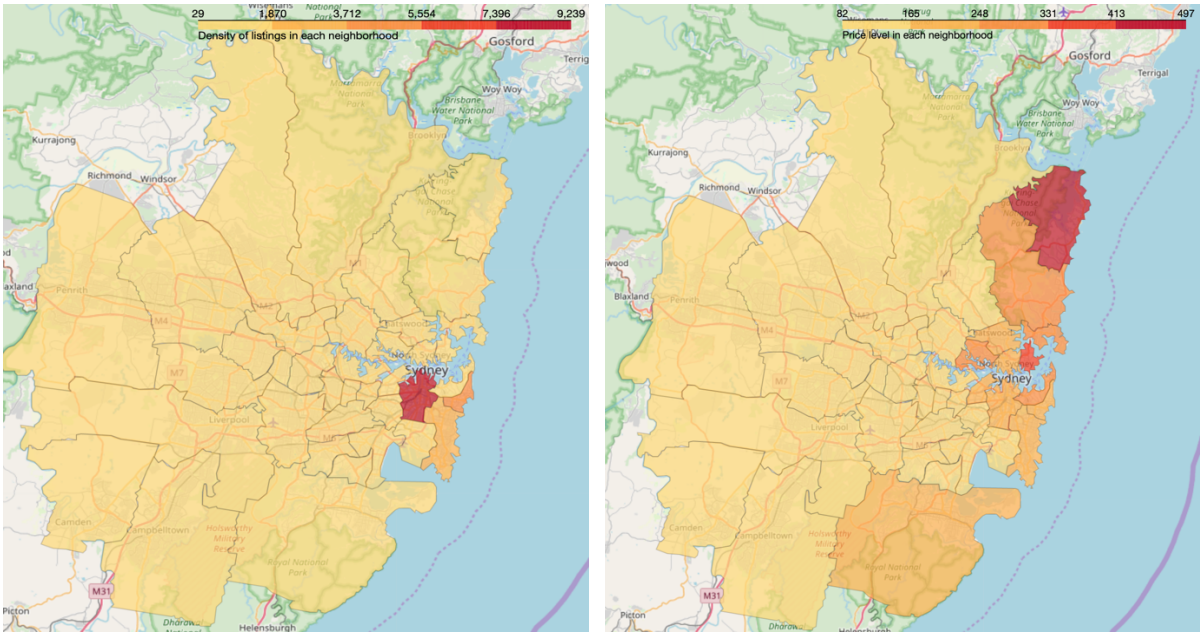


Figure 3: Density of listings (Left) and price level (Right)

3.4 Preliminary clustering based on price, reviews, and room type

In the next step, my objective is to run a preliminary clustering on the pre-processed dataset using the k-means algorithm. I compared the results for different cluster numbers and found out that setting cluster number to be 4 gives the most interpretable outcome.

By examining the features for each cluster, we can roughly conclude that each cluster has the following characteristics, respectively:

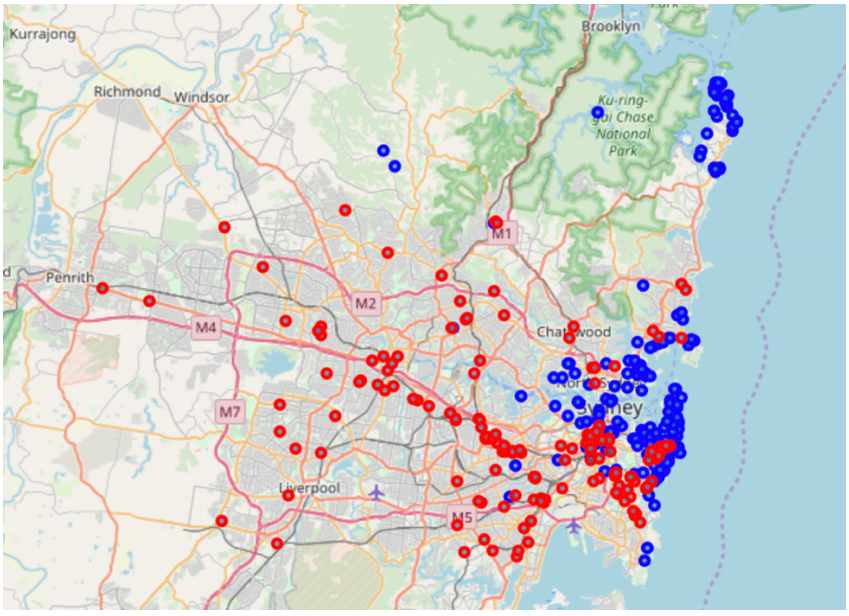
- a mix of private rooms and shared rooms, most affordable price, but also not many reviews from tenants
- mostly entire home/apt, price between affordable and expensive, but do not have many reviews
- also a mix of entire home/apt and private room, extremely expensive, and with the least number of reviews
- a mix of entire home/apt and private room, affordable price, and with most reviews

Table 2: outcome for preliminary clustering

Cluster	price	reviews_per_month	Entire home/apt	Private room	Shared room
0	-0.384190	-0.284680	0.000000	0.953613	0.046387
1	0.173831	-0.294503	0.998106	0.001075	0.000819
2	6.803815	-0.439544	0.949206	0.050794	0.000000
3	-0.153340	2.397487	0.784171	0.211558	0.004271

3.5 Visualizing the top and bottom 200 listings by price

Another interesting aspect to explore is the geographical distribution of the 200 most and least expensive listings around the city. I first used the sort function in pandas to extract these listings from the original dataset. Then the folium library in Python is deployed to generate the listings together on a map, colored in blue and red respectively.



Map 1: distribution of the top and bottom 200 listings

This map clearly shows that listings on the high end of the price spectrum are generally located close to the coastline, especially in the north-eastern tip of the Sydney region, as well as in the eastern suburbs, ranging across most of the regions north of Coogee. The cheapest listings, on the other hand, are scattered across mainly the western and southern parts of the city, with some exceptions near the Bondi Junction area.

The results are intuitive, since being close to the coastline means much better scenery, and this can definitely push up property prices. This also coincides with the fact that properties in these areas are much more expensive than areas in the rest of the city.

One thing that's beyond our expectation is that we see quite a few cheap listings close to the Sydney downtown area, in suburbs like Chippendale and Ultimo. I am guessing these are probably shared rooms. Let's take a closer look at these listings to see if my intuition is true.

room_type	id	name	neighbourhood	latitude	longitude	price	reviews_per_month	availability_365
Entire home/apt	2	2	2	2	2	2	2	2
Private room	4	4	4	4	4	4	4	4
Shared room	33	33	33	33	33	33	33	33

Table 3: composition of room types

This indicates that among the 39 cheapest listings in the Sydney CBD area, 33 of them are shared rooms, which confirms my initial prediction.

4. Clustering listings using Foursquare

The main purpose of this project is to find out similar listings according to features like price and nearby facilities, so that travelers can know what other listings are available when one specific listing has been booked out.

Our original dataset contains more than 36k listings, so obviously this is not practical to explore using API from Foursquare. It will take a tremendous amount of time to return nearby venues for so many listings. Thus, to make the analysis more feasible, I decided to apply some filters to limit listings to only those that meet specific criteria. Here, the criteria I used include being located in the neighbourhood of Sydney and having at least 6 monthly reviews. After applying the filters, I was left with a dataset that contains only 252 listings, which is much more practical for conducting further analysis. Among the 252 listings, 190 of them fall in the entire home/apt category, and 62 fall in the private room category.

For each of these listings, I used API from Foursquare to get the top 30 nearby venues within a radius of 400 meters. Then I applied the one hot encoding technique on the returned results to find the top 5 most popular categories of venues for each listing. The price information was then added to the results, before the k-means algorithm was applied, setting the number of clusters equal to 3. The final results are presented in **Tables 4-6**. I have hidden the feature *reviews_per_month*, since it does seem to show any difference across the three clusters.

Table 4: First Five Listings in Cluster 1

name	room_type	price	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2 BED WITH FREE PARKING IN CBD	Entire home/apt	235	Thai Restaurant	Coffee Shop	Chinese Restaurant	Café	Hostel
2 BED WITH FREE PARKING IN CBD	Entire home/apt	226	Thai Restaurant	Coffee Shop	Chinese Restaurant	Café	Hostel
2 BR Home by Fish Market, ICC & Darling Harbour	Entire home/apt	199	Fish Market	Bar	Café	Hotel	Seafood Restaurant
2 BR Unit 10 Mins CBD, Near Grounds of Alexandria	Entire home/apt	186	Café	Pub	Grocery Store	Rugby Pitch	Thai Restaurant
2 Beds Cozy Apartment in CBD 1mins to QVB	Entire home/apt	192	Café	Coffee Shop	Japanese Restaurant	Shopping Mall	Hotel

Discussion: Listings in this category are mainly entire homes, and priced at around 200 Australian dollars per night. Common nearby facilities are coffee shops, bars, pizza places, and Asian restaurants, in particular Thai restaurants. Thus, these listings are fairly suitable for people with a decent amount of budget and who enjoy Asian cuisines.

Table 5: First Five Listings in Cluster 2

name	room_type	price	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
"Brand New" 3 Bedroom apartment at 'PIER 99'	Entire home/apt	282	Seafood Restaurant	Café	Fish Market	Pub	Sushi Restaurant
2BR near Harbour Bridge, Opera House & the Rocks	Entire home/apt	300	Pub	Hotel	Park	Planetarium	Sandwich Place
2BR+2BA APT at The Rocks+parkng+ Bridge views+wifi	Entire home/apt	300	Café	Pub	Hotel Bar	Brewery	Hotel
5 Bedroom Terrace house crown st Central station	Entire home/apt	400	Café	Japanese Restaurant	Pizza Place	Italian Restaurant	Bakery
Apartment in heart of city Sydney CBD-World Square	Entire home/apt	300	Japanese Restaurant	Thai Restaurant	Korean Restaurant	Coffee Shop	Hotel

Discussion: Listings in cluster 2 are also mostly entire apartments, but have a much higher price level, generally being above 300 dollars per night. Surrounding facilities are a bit similar to cluster one, but frequently we can see seafood restaurants and cocktail bars scattered around these listings, revealing that the listings are probably closer to river banks. This implies that the listings may have better ocean views, which explains their higher cost. Obviously, these places are an ideal choice for people with a generous budget, possibly experience professionals, or people who travel with their family.

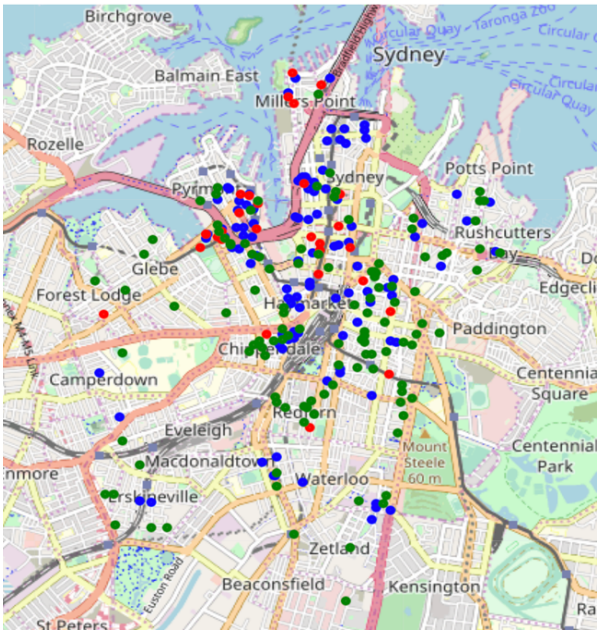
Table 6: First Five Listings in Cluster 3

name	room_type	price	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
"THE GEM'" Sydney (Affordable-Serviced-Simple)	Entire home/apt	118	Italian Restaurant	Bar	Café	Cocktail Bar	Burger Joint
$25 | 5mins Central St | Hike Bondi to Coogee	Entire home/apt	91	Café	Hotel	Art Gallery	Bar	Yoga Studio
+++ LUX WAREHOUSE IN BEST LOCATION, AS SEEN ON TV!	Entire home/apt	140	Café	Bar	Art Gallery	Pub	Yoga Studio
1 Studio in the heart of Sydney-Darlinghurst CBD	Entire home/apt	115	Italian Restaurant	Bar	Burger Joint	Japanese Restaurant	Café
10 minute walk to the city	Private room	105	Italian Restaurant	American Restaurant	Tapas Restaurant	Japanese Restaurant	Wings Joint

Discussion: Common characteristics in the last cluster include being affordable, with prices ranging from 90 to 140 Australian dollars, and being surrounded by bars, cafes, burger joints and Italian restaurants where people can get fast food, pizza, pasta, etc. Notably, there are quite a few private rooms in this category, which is in distinct contrast to the previous two clusters. In my opinion, these places are a very good choice for those with a relatively tight budget but who still want to enjoy the convenience of the downtown location. Moreover, people who travel alone can also choose these listings.

Remember that in the exploratory data analysis stage, I did a preliminary clustering based on price, and room type only. To a large extent, the results are consistent for the two clustering analyses in that they both show that private room are generally cheaper than entire homes, which is very intuitive.

Last but not least, I created a map visualization for the 252 listings, with blue, red, and green representing cluster 1 to 3 respectively.



Map 2: distribution of listings in downtown Sydney

We can see that quite a few of the red listings are located very close to river banks, which verifies my previous hypothesis. Furthermore, blue listings tend to cluster around city centre, whereas green listings are scattered in further distances from the city center, as reflected by their general price levels.

5. Conclusion

For this project, I have analysed the public Airbnb listings in Sydney to get a general idea of the characteristics of these renting properties. I have explored the listing availability in different months. On top, I have looked at which neighbourhoods have the largest number of listings and the average price in those areas. An exploratory clustering was performed on the entire dataset to discover the relationship between price, monthly reviews, and room types. Furthermore, a map visualization informs us where the most and least expensive listings are situated. Correspondingly, the property prices in those areas will be high and low respectively.

In the analysis of listings in downtown Sydney, I have identified three distinct groups and their individual features, based on which I offered my predictions and recommendations. I believe all these findings can be of great value to people who

either plan to become a local Airbnb host in Sydney, or who plan travel to Sydney and want to find suitable accommodation.

Reference

https://www.bloomberg.com/news/articles/2017-04-28/airbnb-goes-after-business-travelers-with-new-booking-tool, by Zaleski, O., viewed on 22 Dec. 2019
https://en.wikipedia.org/wiki/Choropleth_map, viewed on 23 Dec. 2019
https://travel.usnews.com/Sydney_Australia/When_To_Visit/, viewed on 24 Dec. 2019