CAPSTONE PROJECT: BATTLE OF THE NEIGHBORHOODS

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.

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

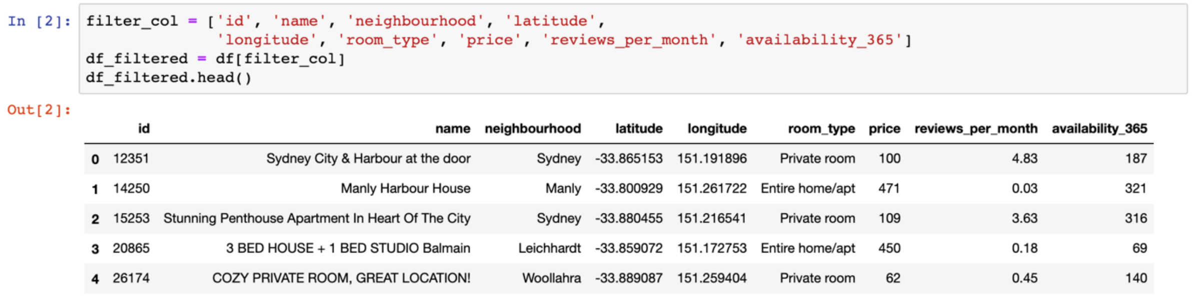
a) distribution of prices and number of listings in each neighbourhood
b) where the top and bottom 200 listings in terms of price are located
c) are there any similarities or patterns among the top and bottom 200 listings, and what's the difference between these two divergent categories

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 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 one of interest to me is the listings_summary_dec18 file, which contains all the essential information needed for this project, such as name of listing, neighbourhood, latitude and longitude, room type, price, reviews per month, and availability, shown in the following table:



```
In [2]: filter_col = ['id', 'name', 'neighbourhood', 'latitude',
                      'longitude', 'room_type', 'price', 'reviews_per_month', 'availability_365']
        df_filtered = df[filter_col]
        df_filtered.head()
```

Out[2]:

	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

Table 1: summary of listings

As of December 2018, there have been more than 30,000 listings in the city of Sydney. There won't be any issue performing exploratory analysis on this dataset. But obviously, the number will be too large for creating map visualizations or analysing surrounding facilities of each listing. Thus, when it comes to creating visualizations and making clusters, only the 200 most and least expensive listings will be taken into account, as defined in the problem section.

The following part is for week 5.

3. Methodology

- Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.
- Results section where you discuss the results.
- Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
- Conclusion section where you conclude the report.