# NIT6160 Final Project: Data Warehousing and Mining

## *Final project*

This project is worth 40% of the total marks of this unit.

### Introduction

Airbnb has successfully disrupted the traditional hospitality industry as more and more travellers decide to use Airbnb as one of the primary accommodation providers. From the Barwon South West data set provided in inside Airbnb, we hope something valuable to potential investors and hosts could be found out with data mining and machine learning.

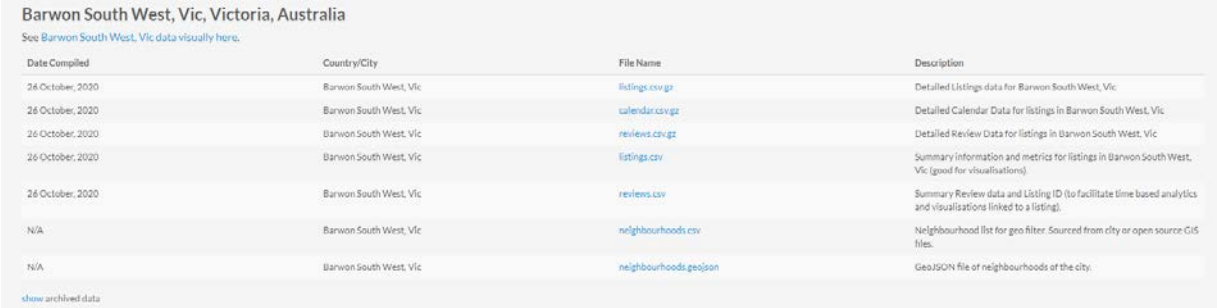
### Dataset

Inside Airbnb - Barwon South West, Vic, Victoria, Australia

The dataset could be downloaded from the link below:
http://insideairbnb.com/get-the-data.html

### Files:

There are two files that will be mainly used for quantitative analysis with python data mining.
(1): **listings.csv**: Detailed listings for Barwon South West
(2): **reviews.csv:** Detailed Review Data for Barwon South West



Barwon South West, Vic, Victoria, Australia

See Barwon South West, Vic data visually here.

| Date Compiled | Country/City | File Name | Description |
|---|---|---|---|
| 26 October, 2020 | Barwon South West, Vic | listings.csv.gz | Detailed Listings data for Barwon South West, Vic |
| 26 October, 2020 | Barwon South West, Vic | calendar.csv.gz | Detailed Calendar Data for listings in Barwon South West, Vic |
| 26 October, 2020 | Barwon South West, Vic | reviews.csv.gz | Detailed Review Data for listings in Barwon South West, Vic |
| 26 October, 2020 | Barwon South West, Vic | listings.csv | Summary information and metrics for listings in Barwon South West, Vic (good for visualisations) |
| 26 October, 2020 | Barwon South West, Vic | reviews.csv | Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing). |
| N/A | Barwon South West, Vic | neighbourhoods.csv | Neighbourhood list for geo filter. Sourced from city or open source GIS files. |
| N/A | Barwon South West, Vic | neighbourhoods.geojson | GeoJSON file of neighbourhoods of the city. |

show archived data

### Task 1: Data Pre-processing

Pre-processing is designed to select the proper columns data to work with and clean the dataset like removing the Nan values and dealing with the data format.

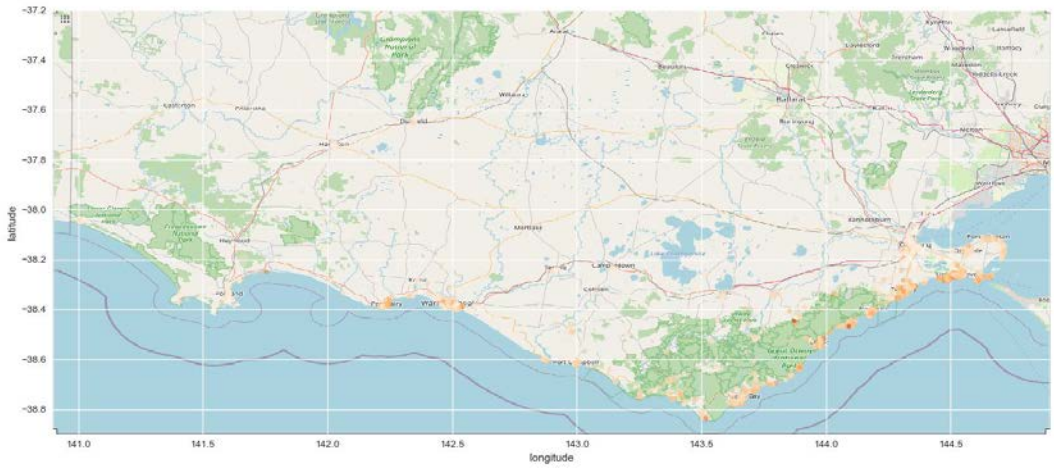
1) Deciding which columns to work with
   We want to keep the information from the dataset as much as possible while removing those irrelevant columns. Removing the irrelevant information could effectively reduce the unnecessary information and avoid the curse of dimensionality, thus to increase the model's performance.

2) Cleaning prices and dealing with missing values
   Operations to change the currency to float values and drop the rows with Nan values

## Task 2: Exploratory Data Analysis (EDA) with Data Visualization

As we are focusing on predicting the prices for accommodations and finding out features that contribute to high prices. We first see the price distribution through boxplot and real street map.

3) Price column visualization
   Vilsulize the price distribution of accommodations with boxplot
4) Accommodation distribution on maps
   You could use the opensource leaflet (python interface), google maps or any other tools to visualize the accommodation distribution based on the latitude and longitude of each accommodation.




5) Summarize the number of accommodations in each market/ each region

6) Summarize the mean price of accommodations in each market/ each region

## Task 3: Building the Accommodation Prediction Model

Now, we are trying to build a model to predict the price. The samples are divided into a training set (80% data samples ) and a testing set (20% data samples).

7) Choose a supervised model such as Xgboost, ANNs and other models to implement your price predictor
8) Perform an analysis to discuss what kinds features are most related to the accommodation price

## Task 4 （Advanced tasks）: Sentiment analysis

9) Perform the sentiment analysis for the review comments
10) Analysis of the reasons why people like and dislike the accommodations (open question)

The analysis could be performed in the following aspects: The hotel view, the location, staff attitude and

### Prepare a report

Your report should contain the following:

- **Introduction**
- **The methods applied for solving each task and why you choose it**
- **Results:** Include results and screenshots of the above experimentations.
- **Discussion and error analysis:** Try to interpret the results of your model. Discuss intuitions or hypothesis that can be obtained by visual inspections of the resulting classes or clusters. Mention about assumptions if any, discuss issues that might have affected the model's performance.
- **Challenge and problem during project**
- **References:** If you are using information from other sources apart from R manual and official website, you should cite them.