# BUSA3020 Group Assignment - Predicting Airbnb Listing Prices in Sydney

---

**Due Date:** Friday, 3 June 2022 @ 11.59pm (Week 13)

**Overview:**

- In the group assignment you will form a team of up to 3 students (minimum 2) and participate in a forecasting competition on Kaggle
- The goal is to predict listed property prices of Airbnb stays based on various Airbnb characteristics and regression models
- You will:
    - Write a problem statement and perform Exploratory Data Analysis
    - Clean up data, deal with categorical features and missing observations, and create new variables (feature engineering)
    - Construct and tune forecasting models, produce forecasts and submit your predictions to Kaggle
    - Each member of the team will record a video presentation of their work
    - Marks will be awarded producing a prediction in the top 3 positions of their unit as well as for reaching the highest ranking on Kaggle amongst all teams.

**Instructions:**

- Form a team of 3 students (minimum 2 students)
- Each team member needs to join https://www.kaggle.com
- Choose a team leader and form a team in the competition https://www.kaggle.com/t/caad5fd1f5134d86a15ab13d37d98d19
    - Team leader to click on `team` and join and invite other team members to join
    - There are two MQBS BUSA units competing in this competition
    - Your **team's name must start** with your unit code, for instance you could have a team called BUSA3020_PR3D1CT0RS
- All team members should work on all the tasks listed below however
    - **Choose a team member who will be responsible for one of each of the 3 tasks listed below**

**Marks**:

- Total Marks: 40
- Your mark will consist of:
    - 50% x overall assignment mark + 45% x mark for the task that you are responsible for + 5% x mark received from your teammates for your effort in group work
- 7 marks will be deducted from each Task for which there is no video presentation

**Competition Marks:**

- 5 marks: Ranking in the top 3 places of your unit on Kaggle (make sure you name your

team as instructed above)
- 2 marks: Reaching the first place in your unit (make sure you name your team as instructed above)

**Submissions:**

1. On Kaggle: submit your team's forecast in order to be ranked by Kaggle
   - Can do this as many times as necessary while building their model
2. On iLearn **only team leader to submit** this Jupyter notebook re-named `Group_Assignment_MQ_ID.ipynb` where MQ_ID is team leader's MQ ID number
   - The Jupyter notebook must contain team members names/ID numbers, and team name in the competition
   - Provide answers to the 3 Tasks below in the allocated cells including all codes/outputs/writeups
   - One 15 minute video recording of your work
     - Each team member to provide a 5 minute presentation of the Task that they led (it is best to jointly record your video using Zoom)
     - When recording your video make sure your face is visible, that you share your Jupyter Notebook and explain everything you've done in the submitted Jupyter notebook on screen
     - 7 marks will be deducted from each Task for which there is no video presentation or if you don't follow the above instructions
3. On iLearn each student needs to submit a file with their teammates' names, SID and a mark for their group effort (out of 100%)

---

**Fill out the following information**

For each team member provide name, Student ID number and which task is performed below

- Team Name on Kaggle: `(insert here)`
- Team Leader and Team Member 1: `(insert here)`
- Team Member 2: `(insert here)`
- Team Member 3: `(insert here)`

---

# Task 1: Problem Description and Initial Data Analysis

1. Read the Competition Overview on Kaggle https://www.kaggle.com/t/caad5fd1f5134d86a15ab13d37d98d19
2. Referring to Competition Overview and the data provided on Kaggle write about a 500 words **Problem Description** focusing on key points that will need to be addressed as first steps in Tasks 2 and 3 below, using the following headings:
   - Forecasting Problem
   - Evaluation Criteria
   - Types of Variables/Features

- Data summary and main data characteristics
- Missing Values (only explain what you found at this stage)

Total Marks: 11

In [9]:
```
#Task 1 code here, insert more cells if required
```

`(Task 1, Text Here - insert more cells as required)`

---

# Task 2: Data Cleaning, Missing Observations and Feature Engineering

- In this task you will follow a set of instructions/questions listed below.
- Make sure you **explain** each step you do both in Markdown text and on your video.
  - Do not just read out your commands without exaplaining what they do and why you used them

Total Marks: 11

**Task 2, Question 1**: Clean **all** numerical features and the target variable `price` so that they can be used in training algorithms. For instance, `host_response_rate` feature is in object format containing both numerical values and text. Extract numerical values (or equivalently eliminate the text) so that the numerical values can be used as a regular feature.
(2 marks)

In [11]:
```
## Task 2, Question 1 Code Here
```

`(Task 2, Question 1 Text Here - insert more cells as required)`

**Task 2, Question 2** Create at least 4 new features from existing features which contain multiple items of information, e.g. creating `email`, `phone`, `reviews`, `jumio`, etc. from feature `host_verifications`.
(2 marks)

In [12]:
```
## Task 2, Question 2 Code Here
```

`(Task 2, Question 2 Text Here - insert more cells as required)`

**Task 2, Question 3**: Impute missing values for all features in both training and test datasets.
(2 marks)

In [13]:
```
## Task 2, Question 3 Code Here
```

`(Task 2, Question 3 Text Here - insert more cells as required)`

**Task 2, Question 4**: Encode all categorical variables appropriately as discussed in class.

Where a categorical feature contains more than 5 unique values, map the features into 5 most frequent values + 'other' and then encode appropriately. For instance, you could group then map `property_type` into 5 basic types + 'other': [entire rental unit, private room,

entire room, entire towehouse, shared room, other] and then encode.
(2 marks)

In [14]:
```
## Task 2, Question 4 Code Here
```

```
(Task 2, Question 4 Text Here - insert more cells as required)
```

**Task 2, Question 5**: Perform any other actions you think need to be done on the data before constructing predictive models, and clearly explain what you have done.
(1 marks)

In [15]:
```
## Task 2, Question 5 Code Here
```

```
(Task 2, Question 5 Text Here - insert more cells as required)
```

**Task 2, Question 6**: Perform exploratory data analysis to measure the relationship between the features and the target and write up your findings. (2 marks)

In [16]:
```
## Task 2, Question 6 Code Here
```

```
(Task 2, Question 6 Text Here - insert more cells as required)
```

---

# Task 3: Fit and tune a forecasting model/Submit predictions/Report score and ranking

Make sure you **clearly explain each step** you do, both in text and on the recoded video.

1. Build a machine learning (ML) regression model taking into account the outcomes of Tasks 1 & 2
2. Fit the model and tune hyperparameters via cross-validation: make sure you comment and explain each step clearly
3. Create predictions using the test dataset and submit your predictions on Kaggle's competition page
4. Provide Kaggle ranking and **score** (screenshot your best submission) and comment
5. Make sure your Python code works, so that a marker that can replicate your all of your results and obtain the same MSE from Kaggle

- Hint: to perform well you will need to iterate Task 3, building and tuning various models in order to find the best one.

Total Marks: 11

In [10]:
```
#Task 3 code here
```

```
(Task 3 - insert more cells as required)
```