# Real Estate Industry Project
# Applied Data Science (MAST30034)

Authors: Akira Takihara Wang, Taha Bhatti, Liam Hodgkinson

## Industry Project Overview

### Predicting Rental Prices

This industry project aims to provide students with an opportunity to help predict rental prices for both residential properties and apartments throughout Victoria, Australia. The overall goal is for students to investigate the role of both internal and external variables on rental prices, and recommend where they are most likely to increase.

The overall aim is to determine the appropriate level of rent an online real estate company should be listing their properties, as well as which properties are most likely to increase in the next five years. Students can scrape property and internal attributes (number of bedrooms, land size, car park spots, etc) from a website of their choosing (e.g. www.domain.com.au).

Students are also expected to find external attributes such as population demographics (affluence and population growth of an SA2 district) and geospatial attributes (proximity to train stations, closest schools, parks, etc).

In other words, students should aim to answer the following questions **for Victoria**:

1. *What are the most important internal and external features in predicting rental prices? (This can be at the granularity of the groups' choosing)*
2. *What are the top 10 suburbs with the highest predicted growth rate?*
3. *What are the most liveable and affordable suburbs according to your chosen metrics?*

### Expected Skillset from Student Groups

We expect student groups to be proficient at writing Python in Jupyter Notebooks. Groups will also need to have at least one member who is familiar with web scraping data and building predictive models. If you need to revise Web Scraping, please review the `BeautifulSoup` component from COMP20008.

### Data Sources

- Property datasets scraped by students;
    * Some skeleton code will be provided that yields some basic features.
    * Students can use it (so long as it is attributed) and modify it to get more data where required.

- ABS Datasets (Public):
  - *Groups should determine which ABS Dataset(s) to use and find out where to get them.*
  - SA2 District Boundaries (ABS);
  - Total population by SA2 Districts;
  - Income by SA2 Districts;
  - Population Forecasts by SA2 Districts;
  - School locations;
- Public Transport Victoria data;
- and another other Geospatial API that can assist.

**Tips on Getting Started**

- For population growth, filter the dataset provided for `State=VIC, Sex=Persons, Years=202X or 203X` (`X` representing 1 or 2). Then find the percentage change in population from 202X to 203X for each SA2 district, for example, using:

  ```
  df.groupby('SA2_NAME')['Total'].apply(pd.Series.pct_change)
  ```

- To find if a coordinate falls within the area of a shapefile, students can search for the `geopandas` methods `.within()` and `.contains()`.
- API's can be used to determine the route distance from one property to a Point of Interest (e.g. Train stations, CBD). One example is Open Route Service, examples have been provided at the end of this document.

## Weekly Checkpoint Assessment

Groups that do not meet weekly checkpoints will lose the mark(s) allocated for that specific checkpoint. If your tutorial is earlier in the week, then you should show sufficient progress or plans to have it completed by the next week.

Whilst Sprints are usually 2-3 weeks long, we will keep them 1 week long for this subject. Checkpoints may be adjusted over the coming weeks depending on progress or updates from the authors.

- **By the End of Sprint 1 (Semester Week 6):**
  Start scraping your chosen real estate website for rental properties in Victoria. Find suitable attributes and external datasets which may assist with what features you can scrape. Additionally, groups should have at least one way of visualizing the geolocation of the properties either on a map and/or shapefile.
- **By the End of Sprint 2 (Semester Week 7):**
  Finish scraping all the required data (and saving it) for all rental properties. Groups are to determine *how many properties and features will be sufficient for their project.* At the minimum, we suggest at least a fair few thousand rental properties across several suburbs. Useful external datasets have been listed above and we expect students to find a way to use Statistical Areas Level 2 (SA2) to derive population forecast and affluence. We also recommend groups find any additional external dataset that may assist in the analysis to answer the 3 big questions.

- **By the End of Sprint 3 (Semester Week 8):**
  All scraping of data should have been completed. At the bare minimum, groups should aim to find the proximity to the closest train station or Melbourne CBD. Proximity/Distance should be calculated via routes (as travelled by car). This can be done by leveraging an API such as Open Route Service. Other features that are recommended also include information about nearby schools, parks, shopping centres, and entertainment districts. Features can be accessed by various public data sources, the most efficient being OpenStreetMap (either an export of data or via API). Once completed, groups should aim to start listing out the features that make a property more expensive or sought after.

- **By the End of Sprint 4 (Semester Week 9):**
  Begin forecasting the rental properties for the next 3-years by suburb or any other suitable granularity (we will let groups decide on the granularity). Whilst working on this, groups are expected to also present some analysis on what features are useful in making a certain suburb more expensive and sought after.

- **By the End of Sprint 5 (Mid-Semester Break):**
  Groups are to continue working on their predictions and analysis. Additionally, prepare a summary notebook (3-5 minutes max) and walk your Tutor through the current findings and any additional insights so far. This task should assist in helping groups formulate their answers to the 3 big questions.

- **By the End of Sprint 6 (Semester Week 10):**
  Summarise and output the 3-year predictions for your chosen granularity (i.e suburb or cluster of properties) and provide some form of an answer for the 3 big questions.

- **Presentations begin from Semester Week 11 onward**. Your slides must be uploaded by the Friday before Semester Week 11 by 2359 AEST.

- **By the End of Sprint 7 (Semester Week 11):**
  Deliverable Assessment Submissions Due. This final sprint should mainly be tidying and collating the repository for submission.

Sprints 1 & 2 will be marked together in Week 7. Since Sprint 5 is during the Mid-Semester Break, Checkpoints 4 and 5 will be marked together the following week. You may treat this as your Manager (tutor) being away on Annual Leave (holidays).

## Deliverable Assessment

### Summary of Outcomes and Repository - 10% of final mark

This project aims to give students the freedom to choose how to approach this as Industry work is quite open-ended when you are given full control over the Project methodology. For the summary of outcomes, we would like readable Jupyter Notebooks summarising your key findings which can be presented to clients and stakeholders.

- 5% for a readable Jupyter Notebook summarising the overall approach taken, issues that you may have run into, and the limitations/assumptions you made along the way. As this is open-ended, there is no right or wrong answer and we will assess you based on your overall approach.

- 5% for the readability and reproducibility of your group's GitHub repository/code quality. This will be personally assessed by the tutor team.

If you are unsure of how to write a readable Jupyter Notebook, your tutors have been permitted to help you along the way. For example, the first few tutorials on this subject consist of documents that are well documented using a combination of Markdown Cells, and appropriate Code Cells with comments.

Bonus marks will be awarded to groups who present their findings in an interactive, client friendly manner. This may include interactive visualisations or dashboards, so that clients are able to investigate and interact with some of the insights.

# References


[1] Relevant article: Towards Data Science Article

[2] OpenRouteService Documentation: Open Route Service API Docs

[3] OpenRouteService Basic Examples: Open Route Services Example Notebooks

[4] OpenRouteService Sample Project: Open Route Services Sample Project

[5] OpenRouteService Sample Project 2 (using `sjoin` and `geopandas`): Sample Project