## General Overview

This challenge is divided into the following 2 tasks: Exploratory Data Analysis (EDA) and a Modelling assessment. Task 2 should take approximately twice as long as Task 1 to complete. Each task is to be completed in a separate Python or R file and provided with your responses to the task questions.

Notes:

* The dataset and scenarios are fictitious
* Your work will be used solely for interview assessment purposes only and will not be used for commercial purposes
* You are free to use any format of your choosing (e.g. Jupyter Notebooks, Markdown, PowerPoint, Word, LaTeX, etc) in showing your thought process in answering the task questions
* Please include your code if it is written in a separate file from your answers.

Assessment Criteria:

* Code should be written in either Python or R, and should be well structured and easy to follow, adhere to coding and style best practices, and be well documented and clearly outline your thought process
* If you are unable to arrive at an answer, or if your confidence in an answer is low, please communicate this clearly in your response
* We are interested in more than just the results, your thought process and approach to the questions are equally important

## Data Details

The data dictionary for *dataset.csv* to be used in both tasks is specified below.

**property\_id**: the id of the property

**sale\_price**: the value that the property sold for; measured in dollars

**sale\_date**: the date of sale

**property\_type**: the type of property; either *House* or *Strata* (generally apartment, unit or townhouse)

**suburb**: the suburb of the property

**postcode**: the postcode of the property

**state**: the state that the property is in

**land**: the land area of the property; measured in squared metres

**floorplate**: the area of the base of the building; measured in squared metres

**bedrooms**: the number of bedrooms of the property

**bathrooms**: the number of bathrooms of the property

**garages**: the number of garages or car spaces of the property

**slope**: the slope of land the property is built on; measured in degrees where zero is flat, positive is above street level and negative is below street level

**max\_roof\_height**: maximum height of the roof; measured in metres

**year\_built**: the year that construction of the building completed

## Task 1 – EDA

You have been provided the dataset *dataset.csv* to explore.

Questions

1. What are your key findings when analysing the data? Please provide summary tables and/or visualisations for some of the more interesting variables.
2. Did you identify any anomalies or anything that looked suspicious in the dataset? If so, what were they and why?
3. What do you think the price difference between a similar property in NSW compared to QLD is? How did you arrive at your answer?

## Task 2 – Modelling

You have been tasked to build a predictive model that predicts property prices. Please use the results acquired from Task 1 (particularly Q3) and information provided in the below questions to build a predictive model using the given dataset (feel free to correct the data for any anomalies found in Task 1 Q2).

Questions

1. The dataset contains missing data. How does your model handle this during model fitting and prediction?
2. It is possible that not all variables included in the dataset are predictive. What variables did you include and exclude from model training? How did you determine this?
3. You received insight that property prices increase (due to inflation and other factors) at a flat rate of 5% each year (or part thereof). How did you incorporate this knowledge into your model?
4. You have reason to believe that the distribution of property prices between *House* and *Strata* is different. How does your model address this?
5. What are the assumptions of your model? Are there any concerns using this model to fit the dataset?
6. What does your model say the price increase is for each additional bedroom in a house (adjusting for all other applicable variables)?

Please follow the below specifications when implementing your model in either Python or R. Ensure you explain your thought process throughout your code.

Let *X* be a dataframe containing every column (including *property\_id*) except *sale\_price*, and let *Y* be a vector containing just the *sale\_price*. *X* and *Y* will be directly passed to the fit and predict functions of your model. Your model should handle all preprocessing including data cleansing, variable selection, feature engineering, etc inside the fit/predict functions (feel free to include any helper functions).

**For Python**

Create a *MyModel* class with *fit* and *predict* functions. The *fit* function should take both *X* and *Y* as arguments to fit the model. The predict function should take *X* as an argument and return a dataframe with columns [*property\_id*, *sale\_date*, *predicted\_price*].

Please ensure the following commands work when your code is completed:

*new\_model = MyModel()*

*new\_model.fit(X, Y)*

*new\_model.predict(X)*

**For R**

Create a *MyModel* function that takes *X* and *Y* as arguments, fits the model and then returns the model object of class *MyModel*. The code structure should roughly resemble

*MyModel <- function(X, Y) {*

*model <- structure(*

*list(model=…),*

*class="MyModel"*

*)*

*return(model)*

*}*

Now create the predict function with header *predict.MyModel <- function(obj, X) {…}*, where *obj* is an instance of *MyModel* and *X* is the input dataframe. The returned dataframe should have columns [*property\_id*, *sale\_date*, *predicted\_price*].

Please ensure the following commands work when your code is completed:

*new\_model <- MyModel(X, Y)*

*predict(new\_model, X)*