**C7081 Assessment Proposal**

**Can we predict house prices using features of each house and a supervised learning approach?**

The dataset I have chosen contains house prices along with a number of variables showing features of each house. The data is from houses in Melbourne and Sydney in 2014. House prices are an important part of the economy. This study could be useful for people wanting to sell their house so they can have an idea of the price it may sell for. The objective is to look at the association between types of houses and house price and see if the variables can be used to predict price. A second objective is to use different models in order to minimise the difference between predicted and actual prices. All the data contained in this data set is from one year, so I will not be looking at the effects over time on house prices.

The data is a collection of house prices along with information about the house. Original source was a dataset on [Kaggle](https://www.kaggle.com/shree1992/housedata) but I have removed 7 columns and changed two of the quantitative variables into binary variables. The raw data I will be using for my analysis is contained in my [GitHub repository](https://github.com/FlorenceGalliers/C7081-assessment.git).

Dimensions of data set = 4600 rows x 11 columns

Variables:

1. price = numerical, price (in dollars) for the house
2. bed = numerical, number of bedrooms in the house
3. bath = numerical, number of bathrooms in the house
4. sqft\_living = integer, area of living space in sq ft.
5. sqft\_total = integer, area of whole housing lot in sq ft.
6. floors = numerical, number of floors
7. condition = categorical, condition of house, 3, 4 or 5
8. basement : 0 if no basement, 1 if basement present
9. yr\_built : integer, year the house was built
10. renovated : 0 if no renovation, 1 if it has been renovated.
11. city : character, location of the house

**Analysis**

I am going to take a supervised learning approach to this dataset. I am going to carry out regression tasks. I may also split the price variable into chunks in order to make it categorical, e.g. low, medium and high price, in order to carry out some classification tasks.

I am going to use a range of methods in order to fit the best model to this data to be able to predict house prices using the other variables, or to see if this is possible. I will split my data into training and test data in order to fit the model. I will use different methods to do this such as k-fold cross validation or bootstrapping to see which method is most accurate. I will also bring in model selection methods to further improve accuracy of the models.

I would also like to use decision trees and random forests to see which variables are contributing most to the variations in house prices and which may have no or little impact.

**Comments** -

I am currently not sure how useful the city variable will be, I may think about turning it into a factor and see if there is a relationship between location and price.