**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 the USA in 2014. House prices are an important part of the economy. This study could be useful for people looking to put their house on the market to get an idea of its value. The objective is to look at the association between types of houses and house price and see which of 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, 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 will carry out regression tasks, starting with simple linear and then expanding into multiple regressions. I have thought about splitting the price variable into chunks to make it categorical data, e.g. low, medium and high price, this will allow me to carry out classification based tasks on the data. For both regression and classification, I will use model selection methods to fit the best model and try to improve accuracy.

I will split my data into training and test data in order to fit the model. I will try multiple methods to do this including k-fold cross validation and bootstrapping, analysing the accuracy of each of these methods.

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.