**Predictive Analytics: Houses Data set**

# The Task

You will work on the **Houses Data set**. This is a dataset about residential property sales in the US, gathered from 2006 to 2010. The dataset consists of multiple variables measuring properties of the houses.

The assignment consists of applying models and model selection methodologies to arrive at models that predict the price of the sale of the houses, given some of the other variables measured.

## 1. Problem description

A primary goal of finding a model that is accurate in predicting the prices of the houses when they are sold. The accuracy of the predictions is measured in **Mean Absolute Error (MAE).**

A secondary goal is to get an understanding of which are the main factors that drive prices, according to the model, this would require that at least one of the models uses a few variables or that you can create a coherent explanation out of one of the models if all use many variables.

Select **three models, one from each model family** to predict the target variable 'SalePrice’. These model families are:

* a linear regression model,
* a kNN regression model,
* A third model. This model can be any model of your choice that is not linear regression nor kNN (might even be a model not covered in the QBUS2820 unit). This is to encourage you to self-explore and self-study, since the ability of self-study is critical in the field of machine learning which is evolving rapidly.

All the models need to be fine-tuned with hyperparameter search (when appropriate) and potentially variable selection. The methodology should maximize the predictive accuracy and the. When the three models have been tuned, you will compute an **accurate estimate of the prediction error** of these models and make a final decision among the three. In the conclusions, you also have to give an explanation of the driving factors of house prices, if the chosen model is not explainable, then use another (or several) and carefully justify the tradeoffs.

The model selection exercise:

* intro/business context/problem formulation
* exploratory data analysis
* The three models
* The conclusions section

Represents the main body of the report and makes 85% of the grade of the assignment.

In addition to the model selection above, the following short exercises. Create a section for each of the questions and remember to explain and discuss the methodology in the report as in the main body.

* Find the best predictive model that uses a single predictor (only one variable).
* Instead of optimizing for the *mean* absolute error, how would you change your methodology to optimize for the *median* error? This is a theoretical question, answer with a proposed methodology, no need to code it in the notebook.

Select 3 houses at random from the dataset and:

* Predict sale prices for those three houses for the year 2022. Reason your answer. You can use any of the three candidate models or use a new model.

Bonus question: Approximate bias and variance of the selected of the Expected Prediction Error of the more accurate model chosen in the main body of the report. The approximation is for a dataset of 70% size of the original dataset. Comment on the limitations of your solution. The bonus exercise is an extra 5% on top of the grade of the assignment, it can be used to counteract errors but cannot make the total grade for the assignment over 100%.

* The main marks come from the report, this is you can have a ‘perfect’ notebook but if there is no explanation if the report then it will not be given marks.
* Always give a reasoned answer, why do you chose a particular variable selection method and not other? Why did you choose a particular ‘third’ model? Why did you choose a particular method for estimating the errors? Etc.
* You do not need to ‘re-state’ the properties of the models, but need to critically justify what the models are adding to the analysis what are the benefits to those models and the drawbacks. The same goes for other decisions in the
* You might need to make ‘suboptimal’ decisions due to, for example, computation times, failure to meet assumptions of the models, etc. In this case remember to state the reason for the decision and the potential problems.

The grading of the assignment will be based on the methodology and justifications, removing points for methodological errors, incomplete sections, etc. There is no ‘minimum’ predictive accuracy to be reached, but you need to apply a good methodology.

The dataset is a popular one in data analysis, described in:

[**https://doi.org/10.1080/10691898.2011.11889627**](https://doi.org/10.1080/10691898.2011.11889627) or [**http://jse.amstat.org/v19n3/decock.pdf**](http://jse.amstat.org/v19n3/decock.pdf) It is used in a practice Kaggle competition: [**https://www.kaggle.com/c/house-prices-advanced-regression-techniques**](https://www.kaggle.com/c/house-prices-advanced-regression-techniques)