Problem Statement

Real estate investors benefit from house price prediction in Canberra, Australia by predicting price in the future to achieve the highest possible profit. The house price prediction can be achieved by visualising past data and observing market trend pattern.

Business Goal

To achieve the highest profit by predicting house price through visualising past data and observing market trend pattern in real estate.

Objectives

- To visualise data using Tableau visual analytics tool.
- To predict future house price in Canberra, Australia by observing market trend pattern in visualised data.

Scope

The past sales data of house in Canberra, Australia is from year 2006 to 2019, consist of 11 columns and 43,179 rows. The columns consist of sales date, price, location and specification of the house. Besides, Tableau visual analytics tool is employed for exploration and reporting to ensure effective data storytelling. Moreover, the predictive model is applied using SAS Enterprise Miner, the model will be validated and the outcome will be interpreted. However, before building the prediction model, exploratory data analysis (EDA), predictive modelling, segmentation, dimension reduction will be performed.

Methodology

Sampling, Exploring, Modifying, Modelling and Assessing (SEMMA) data mining methodology is applied to discover patterns in data for house price prediction. SEMMA data mining methodology approach is chosen because the methodology is developed with SAS Enterprise Miner analytics software.

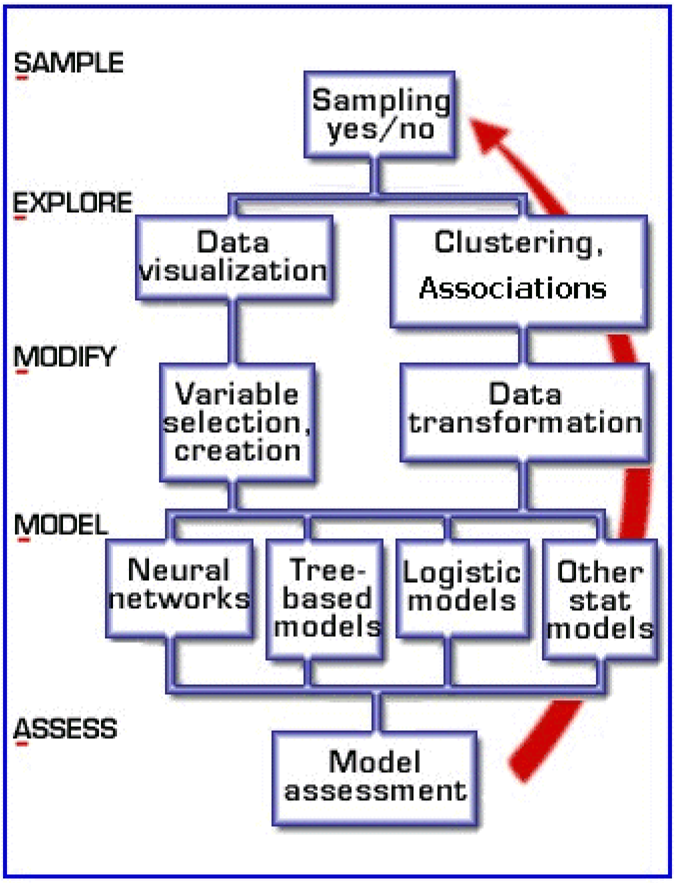


Figure 1: SEMMA Data Mining Process (SAS Institute, 2017)

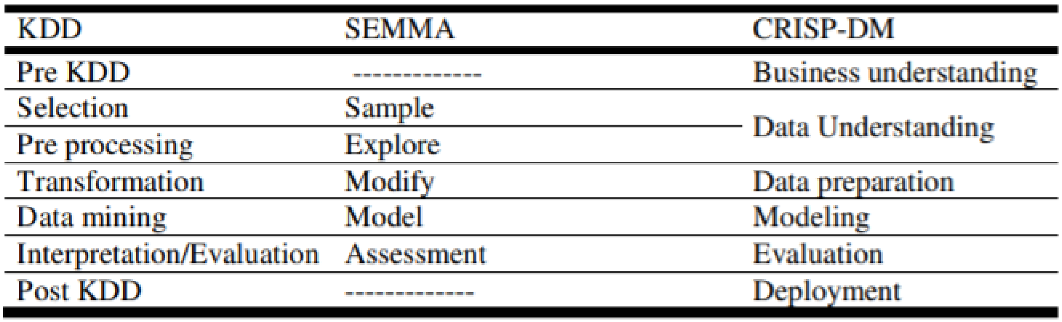
As shown in Figure 1, the methodology consists of 5 stages, prediction of house price in Canberra will follow these stages. First, data is sampled from Kaggle website which contains significant information and it has high usability. Second, the data is explored by searching relationships, trends and anomalies. Third, data is modified by features engineering for model selection process. Forth, in order to predict house price, regression model such as neural networks, tree-based models and linear regression model will be applied. Fifth, the data is assessed based on the usefulness and reliability of the outcomes from the data mining process.

Knowledge Discovery in Database (KDD) has same number of stages with SEMMA, KDD stages include: **selection**, **pre-processing**, **transformation**, **data mining** and **interpretation/evaluation**, it also can be translated to SEMMA as shown below:

- Selection → Sample
- Pre-processing → Explore
- Transformation → Modify
- Data Mining → Model
- Interpretation/Evaluation → Assess

As a result, KDD methodology can be implemented similarly to SEMMA if SAS Enterprise Miner is selected as the data mining tool.

Table 1: KDD, SEMMA, CRISP-DM Comparison (Azevedo & Santos, 2008)



KDD	SEMMA	CRISP-DM
Pre KDD	-------------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	-------------	Deployment

As shown in Table 1, the Cross-Industry Standard Process for Data Mining CRISP-DM is different from KDD and SEMMA, it consists of 6 stages. However, as shown in Table 1 CRISP-DM is comparable with KDD and SEMMA by excluding **business understanding** and **deployment** stages. Moreover, CRISP-DM is created by Daimler Chrysler, SPSS and NCR, therefore SPSS data mining software is required if using this methodology which is unsuitable for the requirement of the given task.

Reference

Azevedo, A. & Santos, M.F. (2008). KDD , SEMMA AND CRISP-DM : A PARALLEL OVERVIEW Ana Azevedo and M . F . Santos. *IADIS European Conference Data Mining*. p.pp. 182–185.