**Team Project Proposal**

**(Group 12)**

**Data Mining the House sales prices**

**Objective:**

We will be using the King county, USA dataset for predictive analytics using Enterprise SAS Miner. King County is a [county](https://en.wikipedia.org/wiki/County_(United_States)) located in the [state](https://en.wikipedia.org/wiki/U.S._state) of Washington. King is the most populous county in Washington, and the 13th-most populous in the United States. As of the [2010 United States Census](https://en.wikipedia.org/wiki/2010_United_States_Census), there were 851,261 housing units at an average density of 402.4 per square mile (155.4/km2).[[22]](https://en.wikipedia.org/wiki/King_County,_Washington#cite_note-census-density-22) The goal of our project is to enable us to study and evaluate the variation of price of a house in King County based on the different house features.

**Business Objective:**

Results from this study can be used to predict the future pricing of houses based on the current trends.

**Data Set:**

Our project uses second-hand data retrieved from[Kaggle](https://www.kaggle.com/harlfoxem/housesalesprediction)**.** This dataset encapsulates a varied set of variables which play a major role in defining the actual price of the house. This dataset contains house sale prices for homes sold between May 2014 and May 2015 in King County, USA. In all, it contains 19 house features plus the price and the id columns, along with 21613 observations.

**Data Mining Technique:**

In our dataset, initially data cleaning will be performed to replace/remove missing values; moving further with understanding the importance of missing values in some of the variables. Besides that, we would like to perform Market Basket analysis and Cluster Segmentation on the dataset. Furthermore, we would classify the target variable using Logistic Regression, decision trees and gradient boosting to determine the category of output variable. We would also like to use techniques such as Neural networks and Random forests to build an ensemble model.

**List of Variables:**



**High Level Data Summary:**

* Data observations in the dataset : 21613
* Binary variables : 1
* Nominal variables : 5
* Interval variables : 15