**Introduction**

The project was carried out on behalf of the local home builder's association, and it involved analysing the state of the local housing market, which had suffered during the recent economic crisis. Two datasets were provided: the Pre-Crisis data contains information on 1,978 single-family homes sold in the one-year period preceding the burst of the "housing bubble," and the Post-Crisis data contains information on 1,657 single-family homes sold in the one-year period following the burst of the housing bubble. In addition, there is a collection of 2,000 observations classified as a test set, which corresponds to residences that are currently for sale.Because the residences in the test set had not yet been sold, the selling price in each of these 2,000 observations was set to zero. The variables in each data set are presented in the table below.

| **Variable** | **Description** |
| --- | --- |
| LandValue | assessed value for the land ($) |
| BuildingValue | assessed value for the building structure ($) |
| Acres | size of lot home sits on (acres) |
| AboveSpace | above ground living space (sq ft) |
| Basement | below ground finished living space (sq ft) |
| Deck | total deck space (sq ft) |
| Baths | number of full- or 3/4-baths |
| Toilets | number of 1/2-baths |
| Fireplaces | number of fireplaces |
| Beds | number of bedrooms |
| Rooms | number of rooms, which are not bedrooms |
| AC | 1 if home has air conditioning for at least 1/2 of living space, 0 if not |
| Age | age of home at time of sale |
| Car | total space for parking cars in covered structures (attached garage + carport) (sq ft) |
| PoorCondition | 1 if overall condition of home is below average, 0 if not |
| GoodCondition | 1 if overall condition of home is above average, 0 if not |
| Price | price the home was sold for (seasonally adjusted $) |

**Problem Definition**

The project had the following objectives:

1. Predict the sale price with the help of k-nearest neighbours using Price as the target or response) variable and all other values as input variables except AC, PoorCondition, and GoodCondition.
2. To account for the varying magnitudes of the variables, standardise the input variables.
3. To determine the value of k that minimises the RMSE on a static validation set or using a 10-fold cross-validation technique.
4. To determine the RMSE
5. To forecast selling prices of properties in the test set, utilise k-nearest neighbours with the value of k that minimises RMSE on the validation set.
6. The above objective were both for pre-crisis data and post-crisis data
7. Calculate the percentage difference in expected price between pre-crisis and post-crisis models, where/pre-crisis predicted price
8. Calculate the average percentage difference in expected price between the pre- and post-crisis models.

**Exploratory Data Analysis**

Exploratory data analysis (EDA) is used by data scientists to explore and investigate data sets and define their key attributes, typically employing data visualisation tools. It helps data scientists determine how to best modify data sources to get the answers they need, making it simpler for them to spot patterns, test hypotheses, and check assumptions.

EDA is generally used to investigate what data might reveal outside of formal modelling or hypothesis testing activities, as well as to better understand data set variables and their relationships. It may also help you determine whether the statistical procedures you're considering for data analysis are appropriate.

The basic goal of EDA is to aid in data analysis prior to making any assumptions. It may help spot obvious mistakes, as well as get a better understanding of data trends, find outliers or unexpected events, and uncover intriguing links between variables.

Python was utilised in this project for exploratory data analysis, together with the Matplotlib and Seaborn packages.

The stakeholders benefited from the exploratory by identifying important performance measures. Through the use of plots and dashboards, this stage allowed stakeholders to see the real estate performance and features of the properties under examination at a glance.

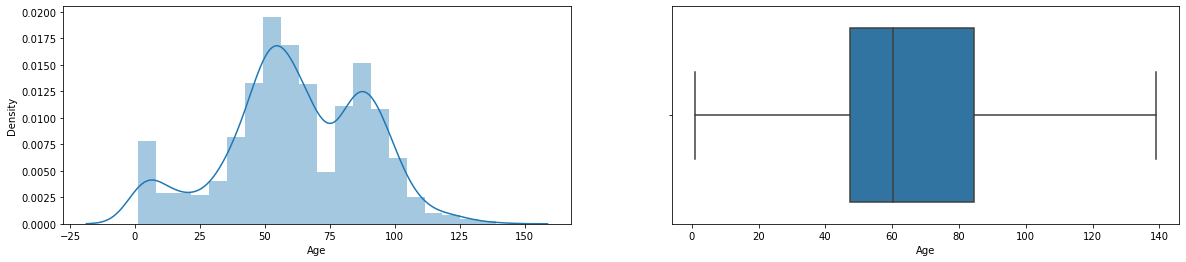
Three different kinds of plots were used in this project:

1. Distribution Plot
2. Boxplots and
3. Count Plots

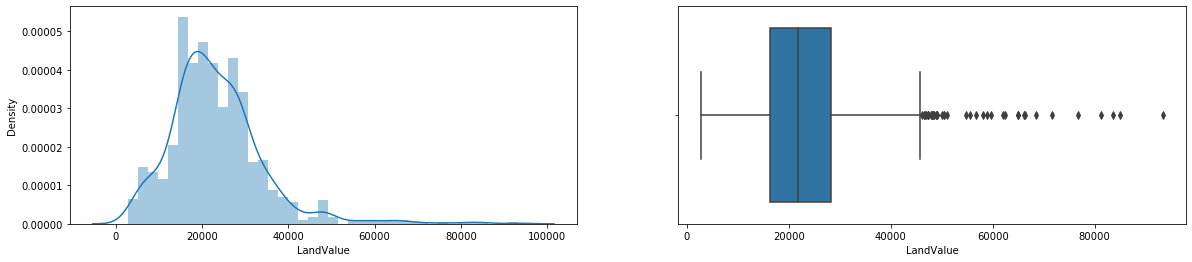
Each feature was examined and insights from each other:

1. 

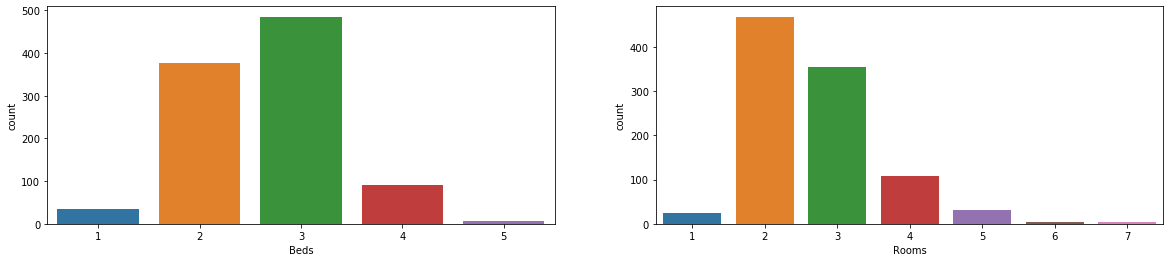


Examination of prices before and after the crisis their distribution did not change , though the average price of post crisis seemed to have been lower before the house prices bubble burst. It is important to note that there were outliers in terms of house prices’

The distribution of house owners did not change in the year after the crisis.There were no outliers also.

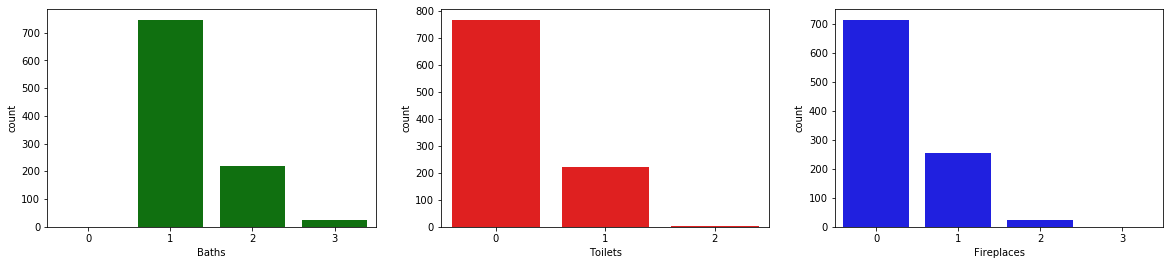


The land values increased between the two events, the average mean also increased in price. The building values also followed the same trend.



Most houses in this dataset had 3 bedrooms mostly and 3 to 4 other rooms.

These factors did not change after the crisis.



Most houses had one washroom, most toilets did not have toilets indoors and no fireplaces.

The features that were represented with cout plots did not change in a year, that is before and after the crisis.

**Modeling**

Before modelling a correlation plot was deployed to determine the relationship between the features. Using 80% correlation as the cap. There was no multicollinearity, indicating the model would perform better.

Due to the different scales of the various features of the data, scaling was used to bring uniformity to the dataset.

The model used in this project was k-Nearest Neighbours, whose main advantage is its reliance on the similarity of the various features of the dataset.

After training the model, the best performance was 76.06% when using the R-Squared as the metric and 4 neighbours.

The model had a Root Mean Squared Error (RMSE) of 19054.89 on the training set and 31271.02257758389 on the validation set. The results indicated the performance was not performing well.The above results were for the pre-crisis. For the post crisis, after training the model, the best performance was 71.76% when using the R-Squared as the metric and 7 neighbours. The model had a Root Mean Squared Error (RMSE) of 23741.7562 on the training set and 33901.52460681606 on the validation set. The model performed worse on the post crisis data, due to the smaller R-squared and the Root Mean Squared Error (RMSE).

On analysing the predicted values on both the pre-crisis and post crisis data, there was a 5.63% drop in prices after the burst of the house prices bubble.

**Conclusion**

From above we note that

1. Prices changed by 5% on the negative side. However we note that the land prices increased. However, this was expected since land almost always appreciates over time.
2. The model after scaling performed best on 4 neighbours on the pre-crisis data and 7 neighbours on 7 neighbours.
3. The models performed worse on the post crisis data than on the pre-crisis data, noted by the increase of the Root Mean Squared Error and drop in the value of the R-squared value.

31271.02257758389