**PROPERTY PRICE FORECASTING**

Real Estate properties like houses are good investments, the owner can have it rented or developed to be sold at a higher price. Unlike cars, real estate property value appreciates over a long period of time, depending on its geographic location. This study will help businessmen or landlords identify the location with a high rental market, and type of house that are commonly rented or sold. And for tenants to find the perfect location and type house that suits their budget.

**GOALS:**

* Predict the house rental number and sales per location
* Predict the increase or decrease in rental and sales price per location
* Relationship between sales and rent and find which one is better for landlord .

**DATA SOURCE AND BACKGROUND**

The house rent and sales reports obtained are from mid-2017 to mid-2019, which are available in the Family and Community Services of New South Wales website. While the house rental and sales report from 2015 to early 2017 were requested from Family and Community Services. These reports are published quarterly, and are based on two sources. Data provided on the bond lodgement form that is registered with Rental Bonds (RB), NSW Fair Trading, Department of Finance, Services and Innovation, for the Rental Statistics. While the data provided on the ‘Notice of Sale or Transfer of Land’ form that is submitted with Land and Property Information (LPI), Department of Finance, Services and Innovation, for the Sales Statistics. The geography of the data is limited on the state of New South Wales, and reports information are based on the Australian Statistical Geography Standard (ASGS) and Australian Bureau of Statistics (ABS). The datasets are composed of information from the combined area of Greater Sydney, which are grouped into Inner, Middle and Outer rings, and the LGAs of Newcastle, Cessnock, Maitland, Port Stephens, Lake Macquarie, Wollongong, Shellharbour and Kiama is defined as the “Greater Metropolitan Region (GMR)”. Other parts of the state which are not mentioned are grouped as "Rest of NSW" for these datasets.

**DATASET FORMAT**

All files are in csv format, mid 2017 to 2019 House Rental Reports consist of 15 columns for LGA tab, and 12 columns for Postcode tab. While for House Sales Reports, each consist of 14 columns for the LGA tab and 11 columns for the Postcode tab. In the LGA tab, data are grouped by Local Government Areas (LGA), while in Postcode tab, the data are grouped by postcodes.

Requested 2015 to early 2017 House Rent and Sales Report from the Family and Community Services has different tabs and columns format. For the Rental Reports, each csv files consist of 6 tabs, as listed below:

* lga\_ad – contains summary of rental price data for both separate house and flat/units in the Greater Metropolitan Region. The tab is categorized by Local Government Area and Codes, and the number of bedrooms.
* lga\_sh – contains rental price data for separate houses in the Greater Metropolitan Region. The tab is categorized by Local Government Area and Codes, and the number of bedrooms.
* lga\_fu – contains rental price data for flats/units in the Greater Metropolitan Region. The tab is categorized by Local Government Area and Codes, and the number of bedrooms.
* lga\_bond – contains total number or rental data in the Greater Metropolitan Region. The tab is categorized by Local Government Area, dwelling type, New Bonds Lodged and Total Bonds Held.
* sa3\_rent – contain rental price data for all dwelling types in the Rural Statistical Areas. The tab is categorized by Rural Statistical Area and Code, and dwelling type.
* sa3\_bond – contains total number or rental data in the Rural Statistical Area. The tab is categorized by Rural Statistical Area, dwelling type, New Bonds Lodged and Total Bonds Held.

While the House Sales Reports for 2015 to 2017 are consist of 5 tabs each, as listed and described below:

* gmr\_a – contain Sales Price data for all dwellings in the Greater Metropolitan Region.
* gmr\_ns – contain Sales Price data for non-strata in the Greater Metropolitan Region.
* gmr\_s – contain Sales Price data for strata in the Greater Metropolitan Region.
* sa3\_a – contain Sales Price data for all dwellings in the Rural Statistical Area.
* sa3\_ns&s – contain Sales Price data for all strata and non-strata in the Rural Statistical Area.

There are 34 csv files in total, 17 for Rental Report and 17 for Sales Report. 15 files for mid-2017 to mid-2019 and 19 requested files for 2015 to early 2017.

**WORK NEEDED FOR THE PROJECT**

Data Consolidation and Data Alignment

Data are divided per quarter, and retrieved data from the Family and Community Services website has different data formatting from the ones that were requested. These needs to be consolidated and aligned to be generated in one DataFrame for easier data cleaning and data manipulation. In data analysis and prediction, data alignment is important to obtain accurate results.

Loading Data and Packages

When all the data has been consolidated and data alignment is done it is necessary to load all essential packages that we gonna use in this project and then we will load dataset.

Analysing the test variable

Since we contain dataset of sales and rent ,we can use rent or sale as our test variable and try to find various things about it.

Multivariate Analysis-

Here we gonna find numerical and categorical variable and try to find any outliers there,if we get any outliers we need to remove it and find correlation matrix so that we can get an idea about the relationship between different variables

Input Missing data and Cleaning Data

For both sales and rental datasets, there are records with null values and records with value "s". Value "s" is found on columns New Bonds Lodged No. and Total Bonds Lodged No. for rental reports, and sales no. for sales report, which indicates small sample. For these datasets all null values will be deleted, due to house types category with null values does not contain any quantitative record that can be used during the data manipulation. But records with value "s" will be replaced by small integers.

Identify the relationships between variables

To be able to perform data analysis, identifying the relationship between the variables is important. Knowing the dependent and independent variables will make it easier for the group to perform different manipulation with the data to obtain accurate predictions on the later part of the project.

During the data manipulation the group will answer the questions and identify the relation of variables listed below.

Relationship between:

* price
* geographic location
* number of sales/signed bonds

Answer questions like:

* Type of house people most people rent
* Type of house with the highest sales
* Comparison of number bonds and price per year
* Comparison of number sales and price per year
* Locations that may increase the rental and sales price in the future based on the data given
* Locations that may increase the rental bonds and sales in the future based on the data given

Feature Engineering or Transformation-

In this step we will try to transform some of our features using some techniques.

Modelling and Prediction-

In this step we will use different models and will try to find which model is better and will predict the main goal of our project.

**TECHNIQUES TO BE USED**

* Support Vector Machine
* Random Forest
* Neural Network
* Decision Tree

**PROJECT PLAN**

* Week 8 – Data Reading
* Week 9 - Data Exploration
* Week 10 – Data Visualization
* Week 11 – Pre Processing
* Week 12 - Model Implementation