Forecasting the Return on Investment (ROI) for a New House in Melbourne: A Five-year Sales Regression Analysis

BDA & DaViz Group Member:

2540135473 – Dora Kalifa Dharmawan – 30%

2501983105 – Pristian Budi Dharmawan – 40%

2540131310 – Pirelli Rahelya Piri – 30%

INTRODUCTION

A real estate company just got an investment with a total of $100 million. This investment is aimed at constructing several new home clusters with a target Return on Investment (ROI) of around $200 thousand for each unit sold. As we know, the cost of a property is increasing every year by around 7.9% per year in Melbourne (propertyupdate.com).

This company has difficulties developing home clusters for certain areas in Melbourne. Their target market is the new family or a family with a medium- to high-class monetary level. They wanted to open these home clusters with a low population density, a good environment, and the nearest to the centre of the city. They took a dataset from 2017 to 2018 to see which region has the highest ROI in Melbourne. However, there are several limitations to building a home cluster, such as the law, construction costs, land availability, etc.

With those criteria, the company assumes that they will be able to overcome their limitations in this project. To increase their ROI, they could build certain facilities in their area, bundle packages, sell furnished or unfurnished houses, etc. Other than that, to make their forecasting of ROI more valid, they wanted to use the dataset that had already been received.

However, they are unclear about the dataset they received. Thus, this company hired a data scientist to process the data and give them the best advice on which region they should construct to have the highest ROI within five years of analysis. The data scientist suggested that to build a forecasting model, he wanted to conduct regression analysis by leveraging several regression models, such as Multiple Linear Regression (MLR), Lasso Regression (LR), and Random Forest Regression (RFR).

METHODOLOGY

The dataset of this project can be accessed through the link below:

Melbourne Housing Market

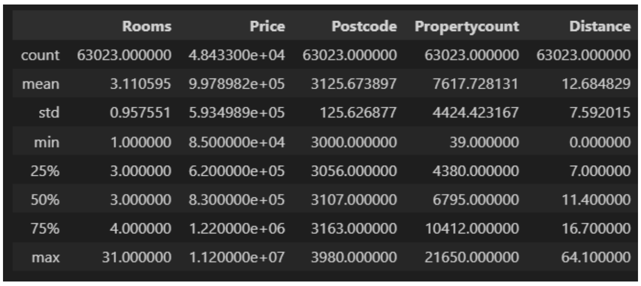
This dataset contains the following attributes:

- Suburb
- Address
- Rooms
- Price
- Method
- Type
- SellerG
- Date
- Distance
- Postcode
- Bedroom2
- Bathroom
- Car
- Landsize
- BuildingArea
- YearBuild
- CouncilArea
- Lattitude
- Longitude
- Regionname
- Propertycount

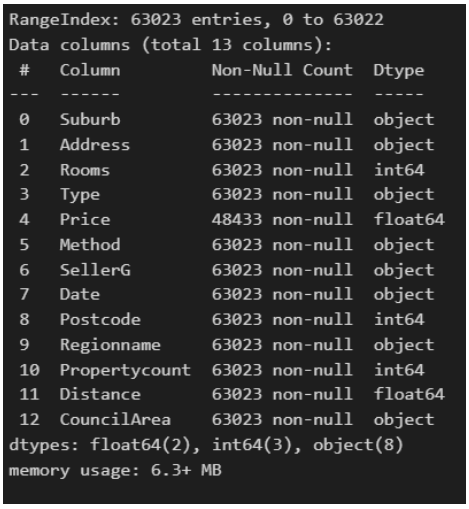
However, there are two datasets, which are:

1. MELBOURNE_HOUSE_PRICES_LESS.csv
2. Melbourne_housing_FULL.csv

The only difference between these two datasets is the number of attributes. Thus, we will forecast it using these two datasets and two results. Below are the summary statistics for dataset number 1:



	Rooms	Price	Postcode	Propertycount	Distance
count	63023.000000	4.843300e+04	63023.000000	63023.000000	63023.000000
mean	3.110595	9.978982e+05	3125.673897	7617.728131	12.684829
std	0.957551	5.934989e+05	125.626877	4424.423167	7.592015
min	1.000000	8.500000e+04	3000.000000	39.000000	0.000000
25%	3.000000	6.200000e+05	3056.000000	4380.000000	7.000000
50%	3.000000	8.300000e+05	3107.000000	6795.000000	11.400000
75%	4.000000	1.220000e+06	3163.000000	10412.000000	16.700000
max	31.000000	1.120000e+07	3980.000000	21650.000000	64.100000



```
RangeIndex: 63023 entries, 0 to 63022
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Suburb         63023 non-null  object 
 1   Address        63023 non-null  object 
 2   Rooms          63023 non-null  int64  
 3   Type           63023 non-null  object 
 4   Price          48433 non-null  float64
 5   Method         63023 non-null  object 
 6   SellerG        63023 non-null  object 
 7   Date           63023 non-null  object 
 8   Postcode       63023 non-null  int64  
 9   Regionname     63023 non-null  object 
 10  Propertycount  63023 non-null  int64  
 11  Distance       63023 non-null  float64
 12  CouncilArea    63023 non-null  object 
dtypes: float64(2), int64(3), object(8)
memory usage: 6.3+ MB
```