

**UNIVERSITY OF NAIROBI**

**SCHOOL OF BUSINESS**

**TOPIC; MELBOURNE HOUSING MARKET DATA ANALYSIS WITH DATA MINING TOOLS AND CRISP-DM METHODOLOGY IN AUSTRALIA**

**BY**

**MUTUA PETER KINYWA**

**REG NO; D33/2883/2017**

**DIS 406: BUSINESS ANALYTICS AND DECISION SUPPORT**

**A REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF BACHELOR OF COMMERCE OF THE UNIVERSITY OF NAIROBI**

**MAY 2021**

Contents

[A BRIEF DESCRIPTION OF THE REAL ESTATE INDUSTRY IN KENYA 3](#_Toc73192631)

[Challenges facing the real estate industry; 3](#_Toc73192632)

[Why machine learning especially predictive analytics is important in addressing the challenges; 3](#_Toc73192633)

[CRISP-DM 4](#_Toc73192634)

[BUSINESS UNDERSTANDING 4](#_Toc73192635)

[Business objectives 4](#_Toc73192636)

[Data mining objectives 5](#_Toc73192637)

[DATA UNDERSTANDING 5](#_Toc73192638)

[Data set 5](#_Toc73192639)

[DATA PREPARATION 6](#_Toc73192640)

[Data Cleaning 6](#_Toc73192641)

[Exploratory Analysis and Data Visualization 7](#_Toc73192642)

[MODELLING 11](#_Toc73192643)

[EVALUATION 13](#_Toc73192644)

[Goodness of fit 13](#_Toc73192645)

[R-Squared 14](#_Toc73192646)

[Model Tuning 14](#_Toc73192647)

[Polynomial Regression With scikit-learn 14](#_Toc73192648)

[Gradient Boosting Regression 16](#_Toc73192649)

[IMPLEMENTATION 17](#_Toc73192650)

[CONCLUSION 17](#_Toc73192651)

[REFERENCES 19](#_Toc73192652)

[APPENDIX 19](#_Toc73192653)

# A BRIEF DESCRIPTION OF THE REAL ESTATE INDUSTRY IN KENYA

There has been a boom that started in the mid to the late 2000s because of the entire property market responding to the rise in demand over the years. To date, Kenya’s housing demand less supply is approximately at 2 Million units (Africa Housing Finance Yearbook, 2019).

To deal with the problem, the Jubilee government in collaboration with the private sector established an ambitious housing development initiative. In 2018, the NHC instituted the Affordable Housing Initiative, as part of the Jubilee government’s ‘Big 4’ agenda. The program plans to supply 500,000 of affordable houses by 2022.

For the past 2 decades, the Kenyan real estate market has grown exponentially evidenced by its contribution to the country’s GDP which increased from 10.5% (2000) to 12.6% (2012) and 13.8% in 2016.

## Challenges facing the real estate industry;

* Access to capital; no doubt that real estate can be an expensive venture and there is need for fast access to capital
* Financing risk
* Insufficient public markets; most real estate transactions are done in private markets
* Difficulty in transacting; real estate closings require a number of interested parties and a lot of paper work which makes the whole process take weeks

## Why machine learning especially predictive analytics is important in addressing the challenges;

Some of the predictive analytics tools and techniques include heat map analysis and projected revenue charts.

Machine learning has brought about a whole new dimension to data management by making it automated hence less paper work leading to enhanced data management in the industry. This entirely leads to a smooth and an easy transaction process.

# CRISP-DM

Cross Industry Standard Process for Data Mining is a freely available model that has become the leading methodology in data mining. This methodology groups all the activities and task involved in data mining into 6 consecutive phases;

* Business Understanding
* Data Understanding
* Data Preparation
* Modeling
* Evaluation
* Deployment

Due to data variety and other related issues as well as the business objectives we may require various degrees of flexibility when applying the CRISP-DM reference model. Biggest challenge that came up was the many features in the data set that were to be used and analyzed to come up with the model.

# BUSINESS UNDERSTANDING

Price of a real estate property is one of the key determinants in the entire real estate market or industry. Price determines the purchasing power or whether investors or customers will purchase or rent a real estate property. Real estate agents, realtors or developers make independent decisions as to whether a customer ready to buy or purchase a property/house or there has to be an examination of various factors or features to determine the price before purchase

## Business objectives

The business goal of this project is to analyze the various features like the number of rooms, land size, the location, property count; that affects the price of a house or a real estate property and also examine the relationships between the price and the various features that affect the price in the market.

Some serious special attention is directed to the price as it is the key determinant in the whole real estate market.

## Data mining objectives

CRISP-DM is the data mining methodology applied in the whole project to achieve the business objectives mentioned above.

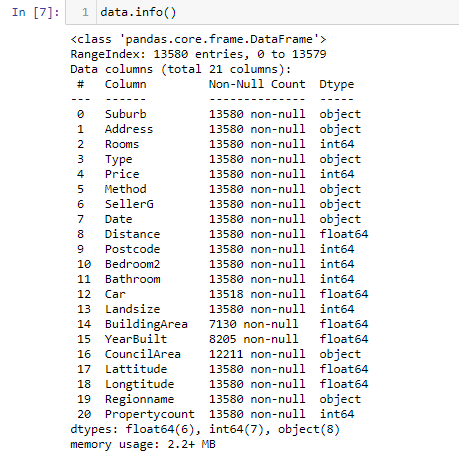
First an exploratory analysis of the different features in the data set is conducted then later there is application of data mining tools that are used to identify the underlying patterns and bring up a working model that can be used to determine the price of a house in the market.

# DATA UNDERSTANDING

## Data set

The data set used in this project is an extract from a real world data set. The data is for Melbourne Housing Market. The data was scraped from publicly available results posted every week from Domain.com.au and was last updated in 2018. The dataset includes Address, Type of Real estate, Suburb, Method of Selling, Rooms, Price, Real Estate Agent, Date of Sale and distance from C.B.D. The data set includes 13580 rows and 21 columns. The whole data set is a representative of the various features in the housing industry that have a relationship with the Price of a house and what is considered before a customer decides to buy a house in the market.

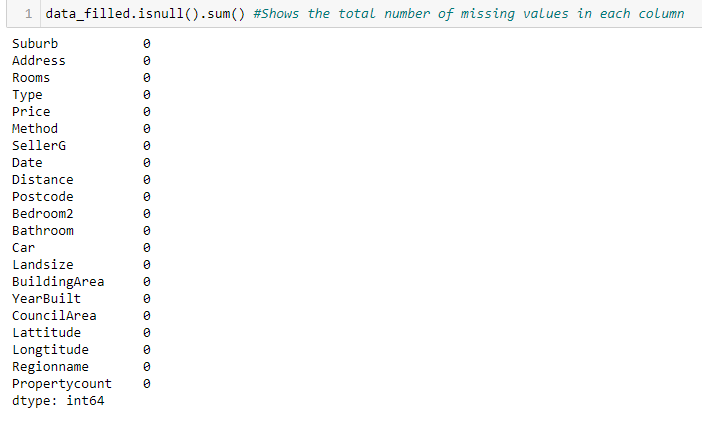
The data set has float, integer and object data types as shown below;



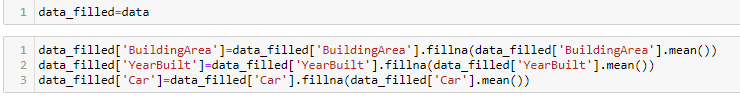
# DATA PREPARATION

## Data Cleaning

A detailed analysis resulted in identifying a large number of missing values. The total number of missing values in each column is shown below;

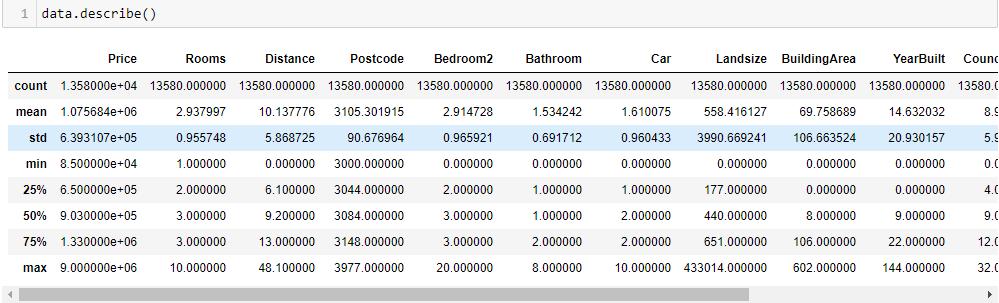


A number of missing values from these columns; Car, Building Area, YearBuilt and CouncilArea, resulted after running an analysis on the data set. Therefore, it was decided that the data must be cleaned and the missing values filled through imputation using the mean of the values in each of the columns where there is missing data. This worked by computing the mean of the non-missing values in a column and then substituting the missing values within each column separately and independently from the other values. The code used in this step is shown below;



## Exploratory Analysis and Data Visualization

The Describe() function in Pandas comes in handy in getting various summary statistics form the data set as shown below; (refer to the Jupyter Notebook)



Key observations form the above output;

* Notably large difference in 75th percentile and max values of predictors Landsize, YearBuilt and BuildingArea
* The mean value is more than median value of each column represented by 50th percentile in the index column

To make the analysis more informative, it was performed with the use of geospatial coordinates to visualize the geographical distribution. Scatterplot was used to check the most common type of house in the area and also the most common prices of the houses in the market in that geographical location.

* Longtitude and latitude plotted against the type of house showed that type h is the most common type of house in that geographical location (fig 1)
* Longtitude to Latitude against price and the Landsize showed that the central area in that geographical location is more developed and the price of a house ranges from sh 1.5M to sh 3.0M (fig 2)

Fig 1;

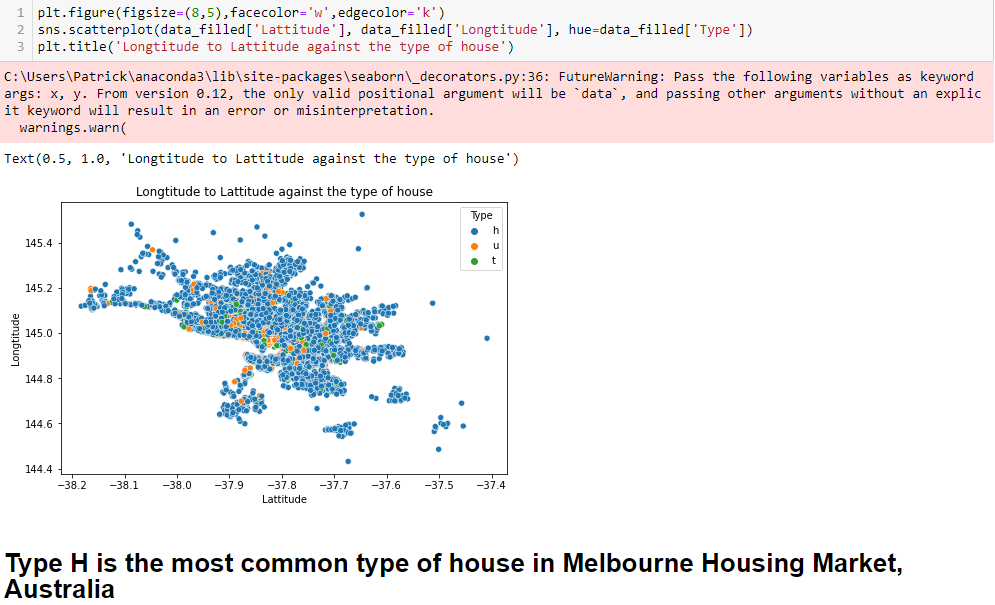
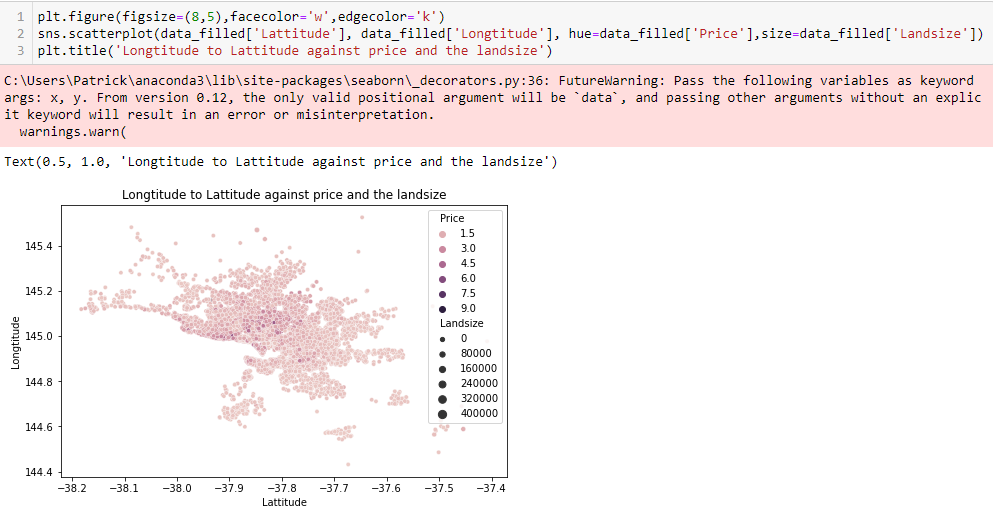
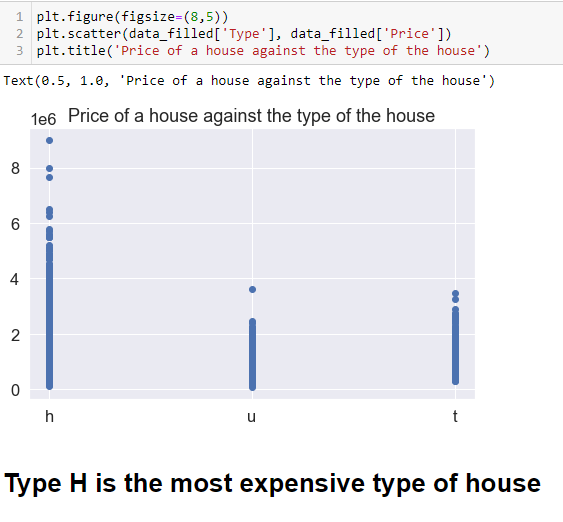


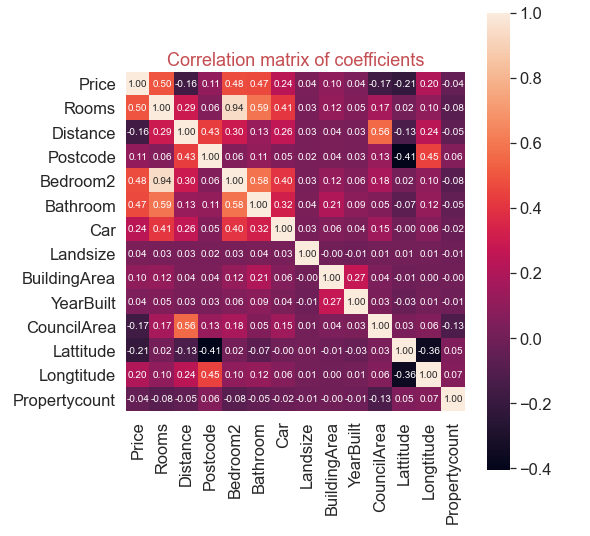
Fig 2;



Plotting a scatter of Price of a house against the type of the house resulted to the key finding that type h is the most expensive type of house in the housing market as shown below;



Further trends and relationships between the features in the dataset were observed via data visualization with heatmap as shown in the output below;



Some of the key findings from the output (heatmap) above are;

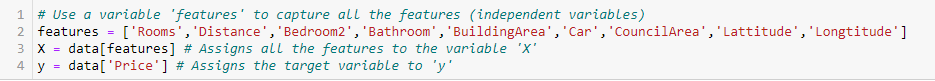
* Landsize, YearBuilt and Propertycount have almost no correlation with the price. Since correlation is almost zero, we can infer that there is no linear relationship between these three features with Price. However it is safe to drop these features when applying linear regression model so as to generate a model with good performance
* Propertycount has almost no correlation with all the other predictors
* Landsize and YearBuilt also have almost no correlation with all the other predictors
* We can also infer that Rooms has a strong positive correlation with Bedroom2, Bathroom and Car
* Bedroom2, Rooms and Bathroom have a strong positive correlation with the price
* Latitude has a strong negative correlation with Postcode and Longtitude

# MODELLING

As previously pointed, the data mining objective of this project was to use the application of various data mining tools to identify the underlying patterns between the various features and also come up with a business model with good performance. The variables were analyzed with correlation matrix to verify the relationship between the price and other variables or features in the dataset.

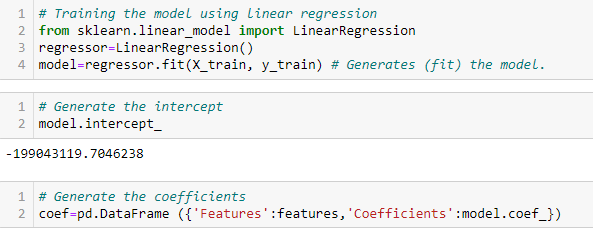
Variable ‘features’ is used to capture all the features, independent variables, and assigned to ‘X’.

The target variable ‘Price’ is assigned to ‘y’ as shown below;



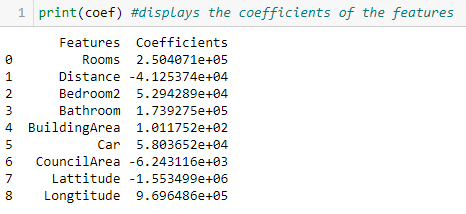
Splitting of the data is done randomly. The data is split into the training set and the validation set. The training set is set to 70 percent of the observations.

The model is trained using linear regression to generate (fit) the model, generate the intercept and also generate the coefficients of the features as shown below;



A display of the generated coefficients of the features which is much easier to interpret:

Code and output;



From the above output of the generated coefficients, we can interpret that;

* Rooms; holding all other variables constant, one unit increase in the Number of rooms will result in a 250,407.10 unit increase in the predicted value, Price.
* Distance; holding all other variables constant, one unit increase in the Distance from CBD in Kilometers will result in a -41,253.74 unit increase in the predicted value, Price.
* Bedroom2; holding all other variables constant, one unit increase in the number of bedrooms will result in a 52,942.89 unit increase in the predicted value, Price.
* Bathroom; holding all other variables constant, one unit increase in the Number of Bathrooms will result in a 173,927.50 unit increase in the predicted value, Price.
* BuildingArea; holding all other variables constant, one unit increase in the Building Size a will result in a 101.1752 unit increase in the predicted value, Price.
* Car; holding all other variables constant, one unit increase in the Number of car spots will result in a 58,036.52 unit increase in the predicted value, Price.
* CouncilArea; holding all other variables constant, one unit increase in the Governing council for the area will result in a -6,243.116 unit increase in the predicted value, Price.

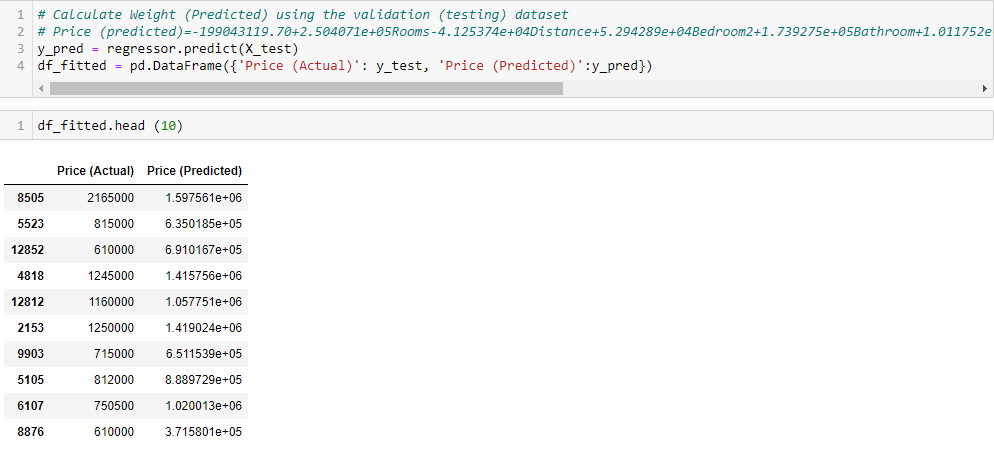
The large coefficients on the specific variables mean that the variables have a large impact on the value of the variable we are trying to predict, which in this case is Price.

Below, is the required linear regression model generated-

**Price (predicted)=-199043119.70+2.504071e+05Rooms-4.125374e+04Distance+5.294289e+04Bedroom2+1.739275e+05Bathroom+1.011752e+02BuildingArea+5.803652e+04Car6.243116e+03CouncilArea-1.553499e+06Lattitude+9.696486e+05Longtitude**

Making Predictions from the Model;

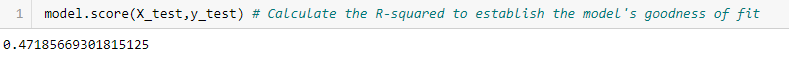
Code and output;



# EVALUATION

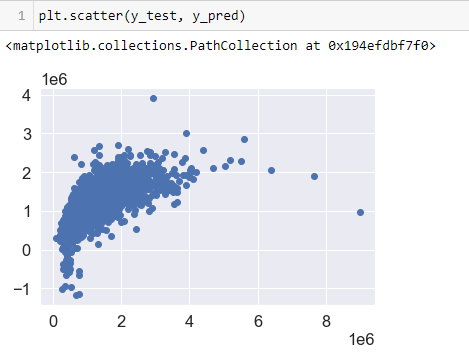
## Goodness of fit

This is a mathematical model that describes the differences between the actual/observed values and the expected/predicted values. This model shows how well the model fits a set of observations.



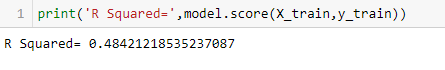
A model score of 0.4719 or 47.19%

**Predicted vs. Actual plot**



## R-Squared

This estimates the strength of the relationship between the linear regression model and the response variable, y, which in this case is the Price.



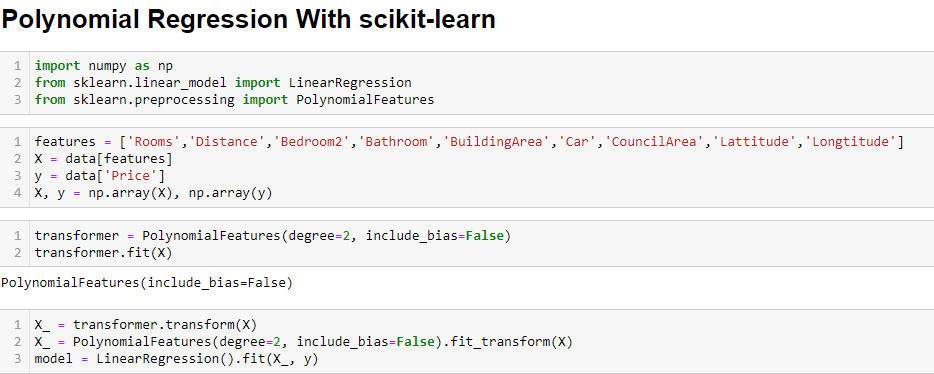
An R-squared 48.42% reveals that 48.42% of the data fit the linear regression model.

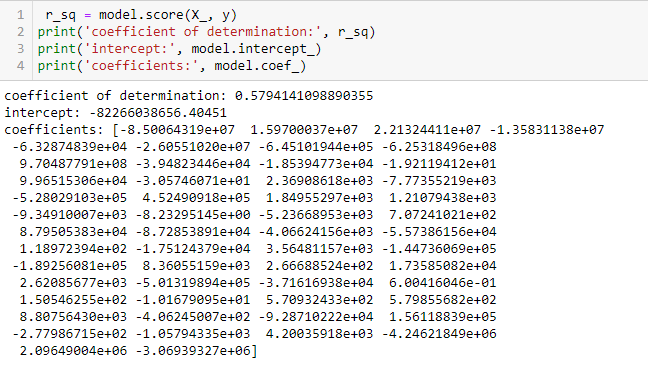
## Model Tuning

### Polynomial Regression With scikit-learn

This model provides the best approximation of the relationship between the dependent variables and independent variables.

Code and output;

****

****

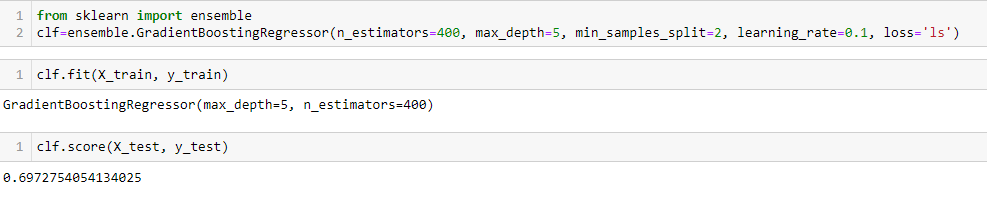
Coefficient of determination of 0.5794 shows that 57.94% percent of the variance in the dependent variable ‘y’ is predictable from the independent variables, represented by X

### Gradient Boosting Regression

Gradient boosting is a machine learning model for regression and classification problems, which results to a prediction model in the form of an ensemble of weak prediction models.

This technique reduces variance and bias.

Code and output;



This gradient boosting model gives us an accuracy of 0.6973 or 69.73% which is higher than all the other models used in this project. This makes it the best and the most applicable model in this project because of its high accuracy.

In conclusion, Gradient Boosting Regression is the most accurate model.

The number of rooms (Rooms), number of bathrooms (Bathroom), number of bedrooms (Bedroom2) and the number of car spots (Car) are the most important features when pricing a house in Melbourne, Australia.

# IMPLEMENTATION

In the real estate market predicting the price is a real business problem. Other than real estate firms and agents, also key players in the market such as financial institutions and mortgage lenders can also benefit from having accurate prediction models. Agents and firms can make better and informed decisions from the model and also analyze the risk factors associated with the market.

Financial institutions can use the price prediction model to provide an accurate real estate property or market appraisal. An accurate price prediction model, gradient boosting regression model, reduces the cost of real estate, housing, market analysis and also allows faster mortgage decisions by the mortgage lenders.

Possible challenge from this model is that this model, gradient boosting regression model, will continue improving to minimize all errors in analysis which can overemphasize outliers and could cause over fitting.

For the model to perform accurately the data has to be sufficient, evenly distributed and substantial in that all data is available and has no unrealistic data and missing values in the data set.

# CONCLUSION

Due to the upward trend with decrease in inventories, increase in demand, low supply and naturally increase in prices made the key real estate market players and real estate market analyzers turn their attention and eyes to more precise prediction models to protect the economy from possible predictable threats.

In this project, the business problem (housing pricing) is analyzed using several machine learning techniques such as linear regression model, Advanced Linear Regression With statsmodels, Polynomial Regression With scikit-learn and Gradient Boosting Regression. A substantial computational experiment was performed on the business problem putting in place a set of features to develop a price prediction model with high R-Squared and low RMSE (Root Mean Squared Error). The solutions obtained by using the Gradient Boosting Regression model on the business problem outperforms in comparison with the model scores got from using the other prediction models hence Gradient Boosting Regression Model stands out.

Possible challenges in this project when addressing the business problem and modeling the project include lack of sufficient long time series of data, uneven distribution of data which leads to erroneous analysis results. For the model to perform well, the data available has to be dealt and analyzed by well experienced and trained real estate market analysts.

Analysts have recommended a variety of predictors of real estate pricing model beyond simple models. Some of the most successful forecasting variables recommended include; valuation ratios such as rent to price and income to price ratios; local economic variables such as the employment rate, income growth, construction costs; demographic trends such as population growth; local space market variables such as vacancy rates, transaction volume and housing starts.

The government should develop policies that ensure steady supply of land for housing investment and development, upgrading and development of the infrastructure such as access roads, electricity distribution and water distribution, increased availability and affordability of housing and also policies that ensure mobilization of funds for real estate development.

In conclusion, with the continued interest in the property market by both local and international players, continued investment in infrastructure and improvement of the legal environment, the real estate housing market sector is definitely poised for further growth in the long term.

# REFERENCES

1. <https://www.nairobibusinessmonthly.com/the-challenges-risks-and-rewards-of-real-estate/>
2. <https://blog.skyl.ai/machine-learning-in-real-estate/>
3. <https://www.globalpropertyguide.com/Africa/Kenya/Price-History>
4. Shearer, C, (2000), ‘The CRISP-DM Model: The New Blueprint for Data Mining’, journal of Data Warehousing, volume 5, Number, pp. 13-22
5. <https://moonbooks.org/Articles/How-to-implement-a-polynomial-linear-regression-using-scikit-learn-and-python-3-/>
6. <https://blog.mlreview.com/gradient-boosting-from-scratch-1e317ae4587d>
7. <https://en.wikipedia.org/wiki/Gradient_boosting>

# APPENDIX

**FEATURE DESCRIPTION**

**Rooms**  Number of rooms

**Price**  Price in Australian dollars

**Type**  br - bedroom(s); h - house, cottage, villa, semi, terrace; u - unit,

duplex; t - townhouse; dev site - development site; o res-other residential

**Method**  S - property sold; SP - property sold prior; PI - property passed in;

PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid;

VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available

**SellerG**  Real Estate Agent

**Date** Date sold

**Distance**  Distance from CBD in Kilometers

**Postcode**  Self explanatory

**Bedroom2**  Scraped number of Bedrooms (from different source)

**Bathroom**  Number of Bathrooms

**Car**  Number of car spots

**Landsize**  Land Size in Meters

**BuildingArea**  Building Size

**YearBuilt**  Year the house was built

**CouncilArea**  Governing council for the area

**Lattitude**  Self explanatory

**Longtitude** Self explanatory

**Regionname** General Region (West, North West, North, North east …etc.)

**Propertycount** Number of properties that exist in the suburb

**NHC**  National Housing Corporation