House Price Modelling

EC.48W

Group Project: Final Report

Atalay İlhanlı	2014300063
Aydın Bayraktar	2013300174
Emre Boran	2013300075
Gül Evin Yılmaz	2012300204
Orhan Yaman	2011300531

Abstract

In this project, on the basis of actual data we forecasted the values of house prices in Melbourne City, Australia. In order to select a prediction method, we explored three different machine learning methods, Linear Regression, Decision Tree Learning and Random Forest Regressor. According to results we could decide which method is better to generate predictions for housing prices in Melbourne. Experiment results show that Linear Regression is superior to Decision Tree Regression and Random Forest Regression in terms of accuracy.

1. Introduction

Buying a house is a difficult decision for almost every consumer. It is an important question for consumers whether the price of the house they want to buy is fair or too expensive compared to market conditions. On the other hand, housing market carries important signals for the national economy's well-being. So, generating predictions about housing values is important for both consumers and economic professionals. There are several features which may affect the price of a house: location, number of rooms, age, area, etc.

In this study we used a data set obtained from Domain.com.au. The data set includes Address, Type of Real estate, Suburb, Method of Selling, Rooms, Price, Real Estate Agent, Date of Sale and Distance. Our main purpose is to decide which machine learning method is better to forecast housing prices according to different variables.

2. Methods Review and Background

Methods

In this study we will implement three different machine learning methods:

1. Linear Regression
2. Decision Tree Regression
3. Random Forest Regression

Before forecasting the value of house prices, we tried to determine which model fits better to our data.

Background

In the past 25 years, the median housing prices in Australia increased approximately by an average of 6.8% per annum. Furthermore, the highest increase in the housing prices is experienced in Melbourne, where values increased with the rate of %8.1. If this trend continues the value of a median house in Melbourne is expected to increase from $825.000 to $5.8 million by 2043. This trend suggests that investing real estate may bring high profits in the long run.

Besides from this trend, the price of a house in Melbourne may depend on a lot of variable like location, size, age, etc. So, our aim is to model the relationship between the price of houses in Melbourne and the independent variables.

3. Data Cleaning and Reconstruction

Target for grouping variables: Categorical variables, numerical variables and date object.
'Suburb', 'Address', 'Type', 'Method', 'SellerG', 'CouncilArea', 'Regionname' and ‘PostCode’ will be our categorical variables.
Date will be our date object and the rest will be numeric variables either in form of integer or float. Necessary steps have been taken to group our variables according to our target.

Missing values

Unfortunately, in some of the columns, more than half of the observations are missing. In dealing with the missing values, we used “.dropna()” function to remove any rows which have at least one missing value. At the end of the process number of observations significantly reduced to 6196 from almost 20 thousand.

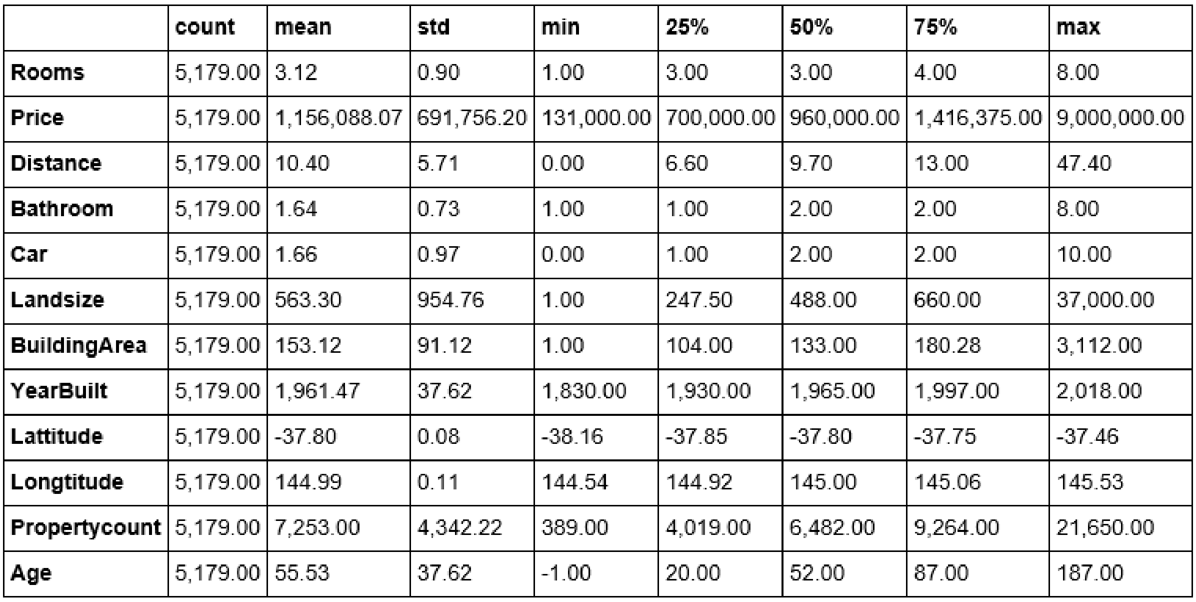
Removing Outliers

While we look at our data to check for the outliers, we observe there are some illogical observations like a house whose building area and land size is zero. Since it is hard to imagine a house that covers zero space, we dropped that rows. Furthermore, there is a house whose age is 800, we dropped that rows.

Rather than using the variable “Yearbuilt”, we prefer adding a new variable named “Age”.
(Age= 2017(date of releasing of data) – Yearbuilt)
We classify houses into two, historic and contemporary. (If age is more than 50 it is a historic house) Age is classified as a categorical variable rather than numerical.
While checking the data we realized there are duplicate variables that has the same observations. In our case Room and Bedroom2 are the same. By looking at the difference between the two variables which is zero, it would be logical to drop one of these.
After these steps, we have done with the data cleaning and reconstruction.

4. Statistical Analysis and Data Exploration

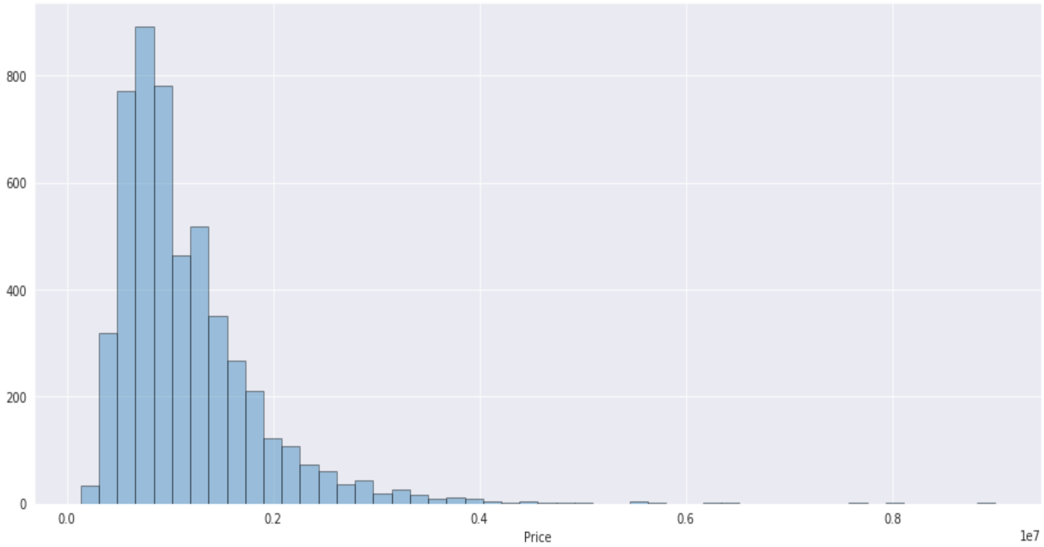
Description of data



	count	mean	std	min	25%	50%	75%	max
Rooms	5,179.00	3.12	0.90	1.00	3.00	3.00	4.00	8.00
Price	5,179.00	1,156,088.07	691,756.20	131,000.00	700,000.00	960,000.00	1,416,375.00	9,000,000.00
Distance	5,179.00	10.40	5.71	0.00	6.60	9.70	13.00	47.40
Bathroom	5,179.00	1.64	0.73	1.00	1.00	2.00	2.00	8.00
Car	5,179.00	1.66	0.97	0.00	1.00	2.00	2.00	10.00
Landsize	5,179.00	563.30	954.76	1.00	247.50	488.00	660.00	37,000.00
BuildingArea	5,179.00	153.12	91.12	1.00	104.00	133.00	180.28	3,112.00
YearBuilt	5,179.00	1,961.47	37.62	1,830.00	1,930.00	1,965.00	1,997.00	2,018.00
Lattitude	5,179.00	-37.80	0.08	-38.16	-37.85	-37.80	-37.75	-37.46
Longtitude	5,179.00	144.99	0.11	144.54	144.92	145.00	145.06	145.53
Propertycount	5,179.00	7,253.00	4,342.22	389.00	4,019.00	6,482.00	9,264.00	21,650.00
Age	5,179.00	55.53	37.62	-1.00	20.00	52.00	87.00	187.00

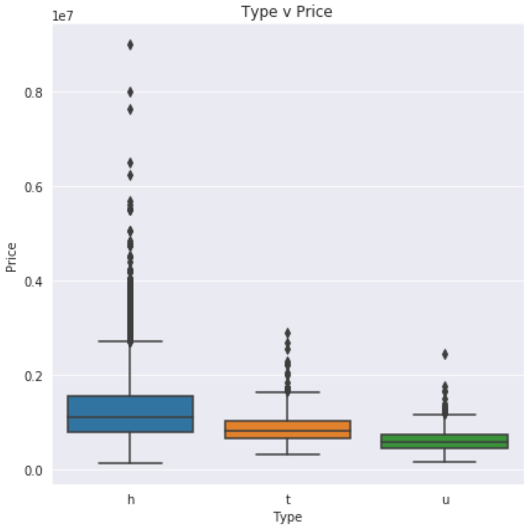
Figure above shows statistical values of relevant variables. (Ex:The houses have on average 3.12 room and have a price over 1 million Australian dollars.)

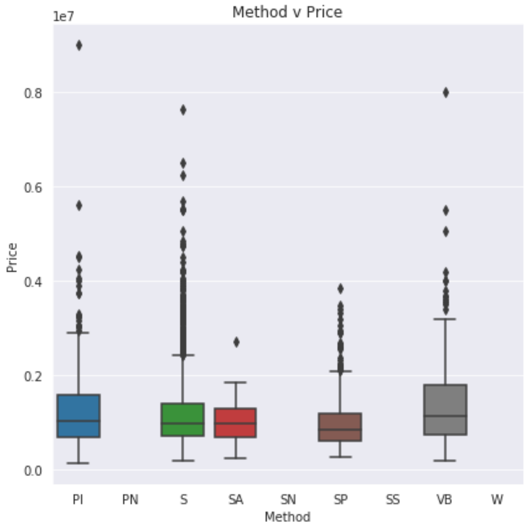
Distribution of price



The figure above shows most of the houses are around 1 million dollar. There are some exceptional prices which we may take as outliers.

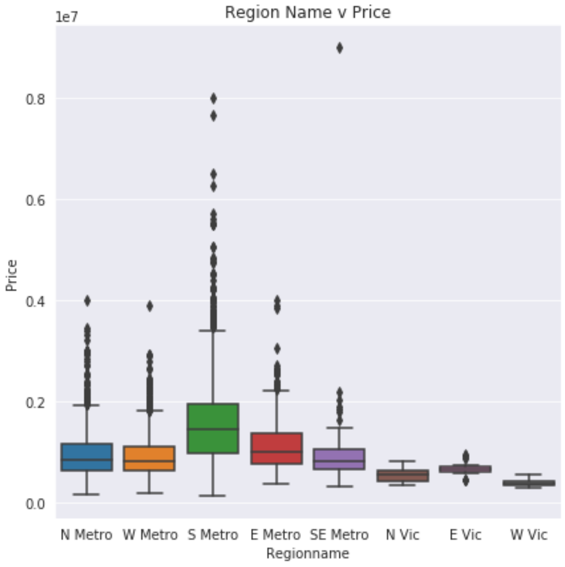
Price Distributions according to categorical features

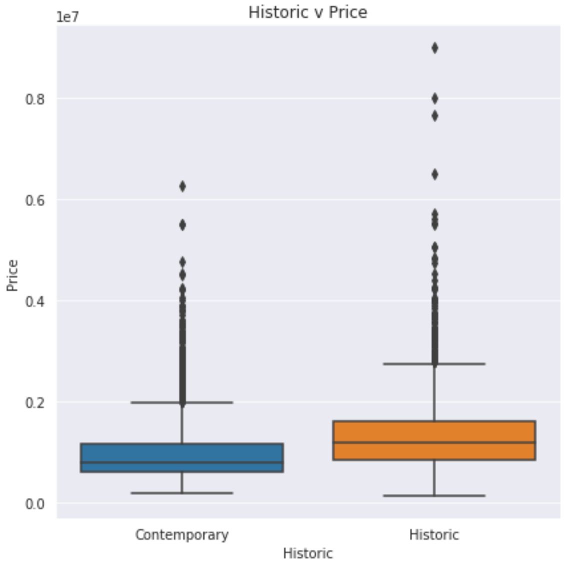




Type: (h=house, u= unit, t=townhouse) vs price. Median value of house is the most while median value of unit is the least.

Method: There is no significant difference between the different sale methods.

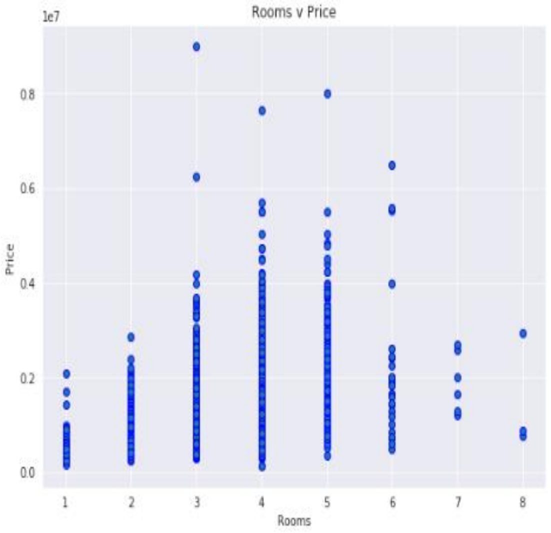


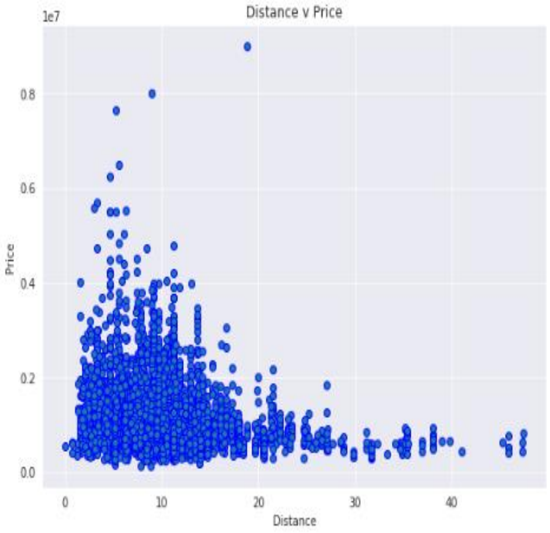


Region name: Houses located in Southern Metro Area have the highest median prices.

Historic vs price: Historic houses have higher median value also higher difference between the lowest price historic house and highest price historic house.

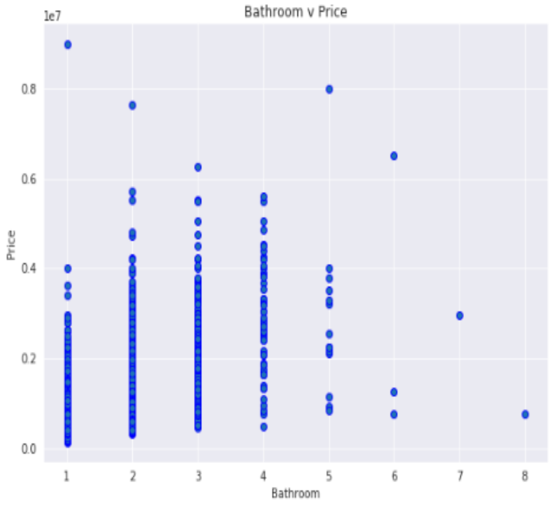
Distribution of numerical features vs price

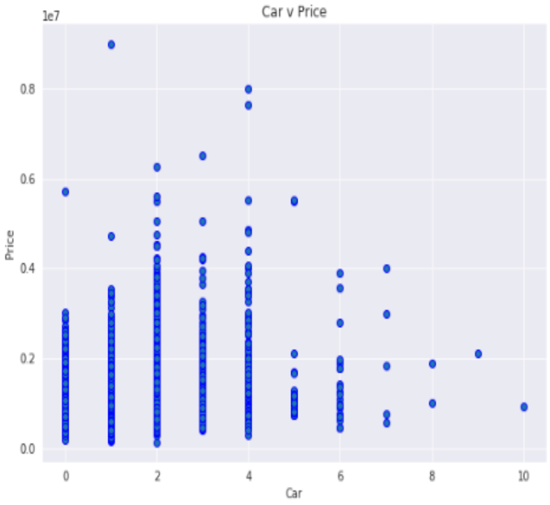


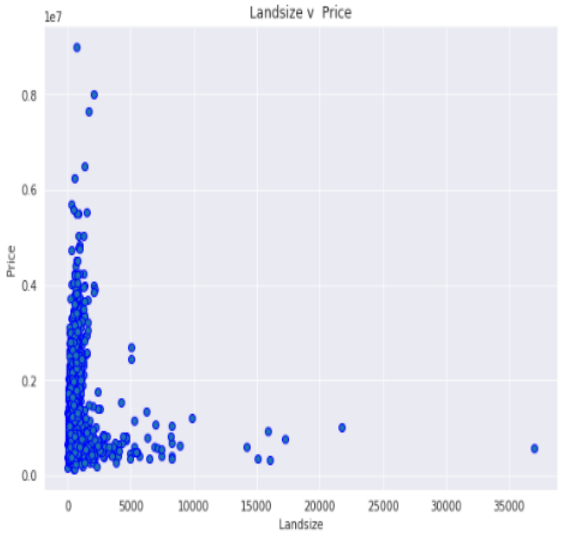


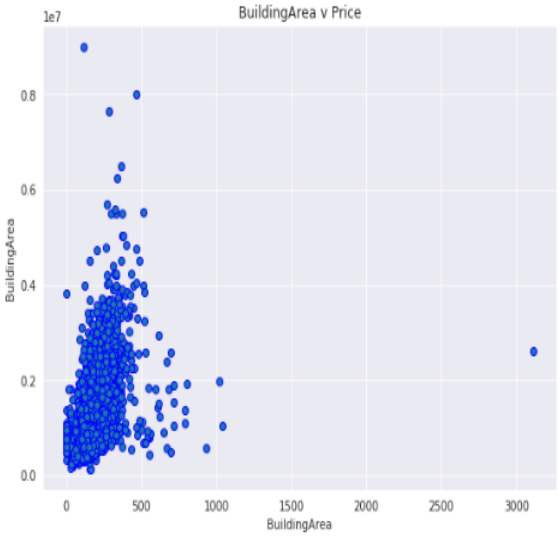
We see a trend that while number of rooms increases price is also increases until 4 rooms. After that there is no significance.

There is a negative relationship between distance and price.

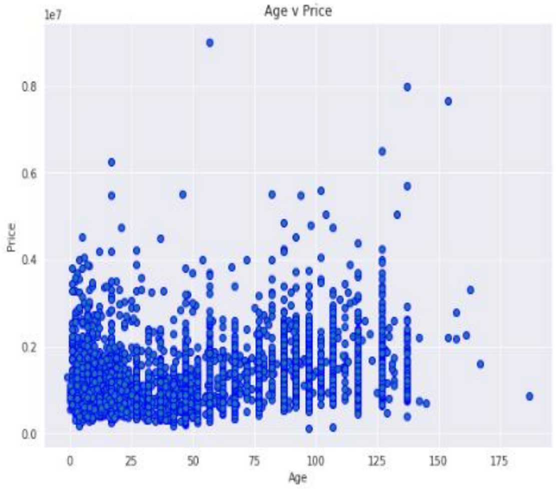


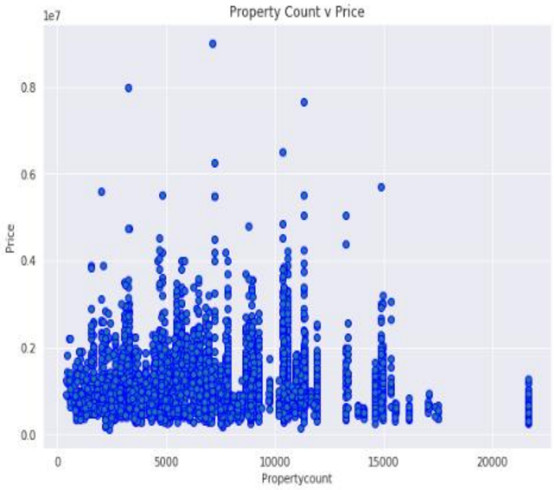




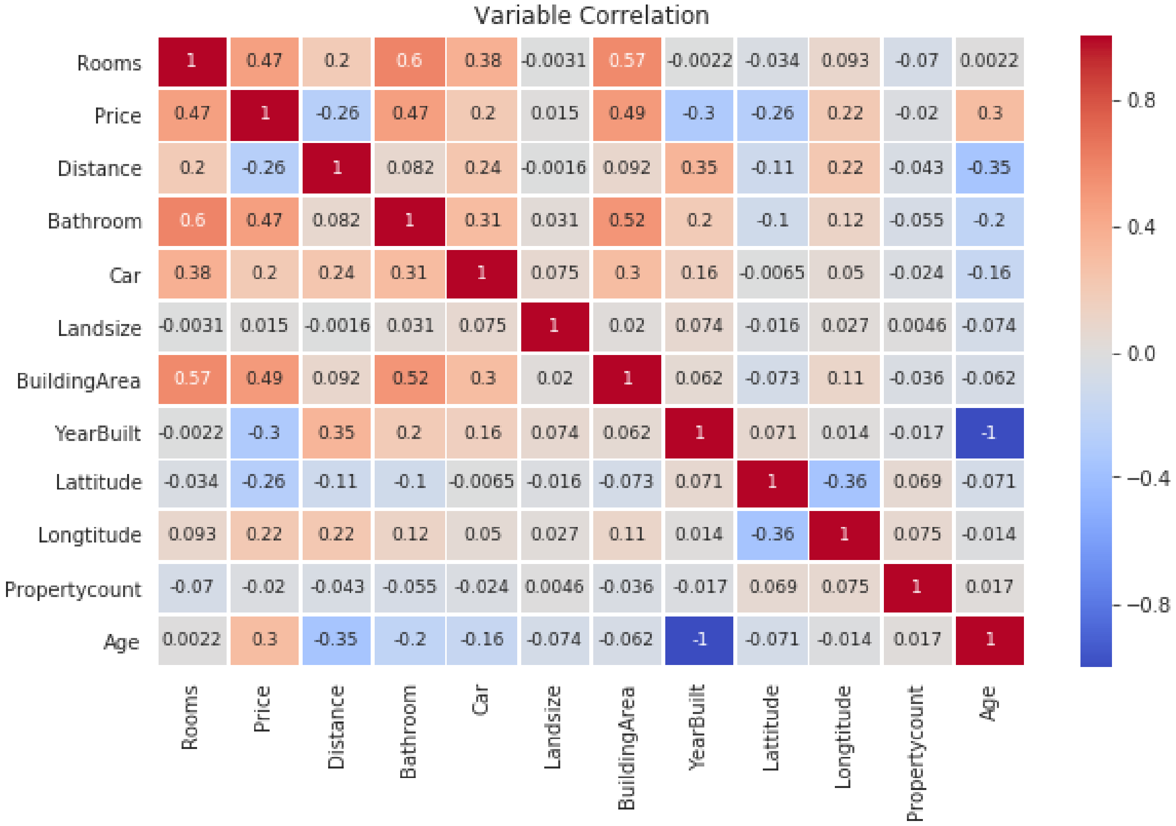


Higher the building area is associated with a higher price.





Correlation between the variables:



	Rooms	Price	Distance	Bathroom	Car	Landsize	BuildingArea	YearBuilt	Lattitude	Longtitude	Propertycount	Age
Rooms	1	0.47	0.2	0.6	0.38	-0.0031	0.57	-0.0022	-0.034	0.093	-0.07	0.0022
Price	0.47	1	-0.26	0.47	0.2	0.015	0.49	-0.3	-0.26	0.22	-0.02	0.3
Distance	0.2	-0.26	1	0.082	0.24	-0.0016	0.092	0.35	-0.11	0.22	-0.043	-0.35
Bathroom	0.6	0.47	0.082	1	0.31	0.031	0.52	0.2	-0.1	0.12	-0.055	-0.2
Car	0.38	0.2	0.24	0.31	1	0.075	0.3	0.16	-0.0065	0.05	-0.024	-0.16
Landsize	-0.0031	0.015	-0.0016	0.031	0.075	1	0.02	0.074	-0.016	0.027	0.0046	-0.074
BuildingArea	0.57	0.49	0.092	0.52	0.3	0.02	1	0.062	-0.073	0.11	-0.036	-0.062
YearBuilt	-0.0022	-0.3	0.35	0.2	0.16	0.074	0.062	1	0.071	0.014	-0.017	-1
Lattitude	-0.034	-0.26	-0.11	-0.1	-0.0065	-0.016	-0.073	0.071	1	-0.36	0.069	-0.071
Longtitude	0.093	0.22	0.22	0.12	0.05	0.027	0.11	0.014	-0.36	1	0.075	-0.014
Propertycount	-0.07	-0.02	-0.043	-0.055	-0.024	0.0046	-0.036	-0.017	0.069	0.075	1	0.017
Age	0.0022	0.3	-0.35	-0.2	-0.16	-0.074	-0.062	-1	-0.071	-0.014	0.017	1

There is high positive correlation between room, bathroom and building area. One reason can be they all related to size of a house. Moreover, price is highly correlated with these three variables. Other than these variables, correlation is somehow ignorable.

5. Methods and Tools of Use

Using supervised learning, our aim is to employ a method that forecasts the outputs using independent variables when we present the example inputs and outputs to the system. Regression is a supervised learning method, which gives a continuous array of outputs. Because we are trying to predict values that can fall in a continuous range, our study can be classified as a regression problem. We tried to analyze relationship between independent variables we selected (Rooms, Distance, Bathrooms, Car, Landsize, Longitude, Latitude, Age, BuildingArea and Type) and the price of a house.

Independent Variables	Dependent Variable
Rooms, Distance, Bathrooms, Car, Landsize, Longitude, Latitude, Age, BuildingArea and Type	Price of property

Table: Variables used in our prediction models

1. Linear Regression

Linear regression allows us to model a relationship between a dependent variable and independent (explanatory) variables. Using Linear regression, we can answer:

1) Whether our independent variables are good at predicting an outcome?

2) Which independent variables are significant predictors and how they impact the outcome?

2. Decision Tree Regression

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target.

3. Random Forest Regression

Random forest regressor combines different decision trees to determine final output instead of depending on only one decision tree. This algorithm makes it robust to overfitting. When we aggregate imperfect decision trees, at the average imperfection gets minimized.

For each type of regression, we split the data test and training sets to evaluate the performance of each model. We calculated the mean absolute percentage error:

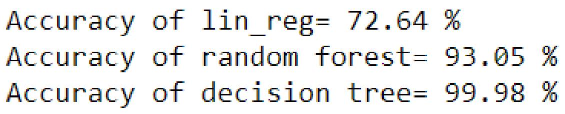
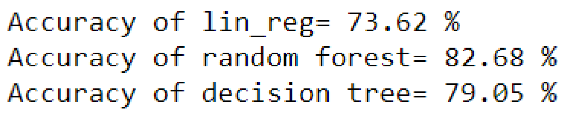
$$\mathrm{M} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|,$$

where *At* is the actual value and *Ft* is the forecast value. The difference between *At* and *Ft* is divided by the actual value *At* again. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points *n*. Multiplying by 100% makes it a percentage error.
We calculated accuracy for each model by subtracting the mean absolute percentage error from 1.

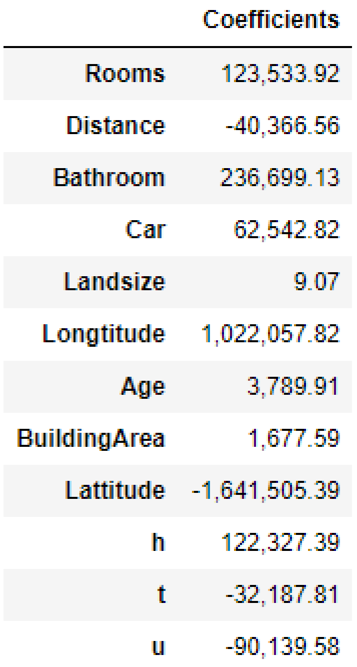
6. Results

In this study, we generated a forecast for the value of house prices in Melbourne City, Australia. Our data set includes 19741 houses and classified them to some categories. These categories: Suburb, Address, Rooms, Type, Price, Method, SellerG, Date, Distance, Postcode, Bed room2, Bathroom, Car, Land size, Building Area, YearBuilt, Council Area, Lattitude, Longtitude, Region name and Property count. Later, we calculated these values as count, mean, standard deviation, minimum number, maximum number, 25 percent, 50 percent and 75 percent.

We applied linear regression, random forest and decision tree to this data. We made an accuracy test and results are shown below.
Accuracy = 100 - mean absolute percentage error

Accuracy for training set	Accuracy for test set
 Accuracy of lin_reg= 72.64 % Accuracy of random forest= 93.05 % Accuracy of decision tree= 99.98 %	 Accuracy of lin_reg= 73.62 % Accuracy of random forest= 82.68 % Accuracy of decision tree= 79.05 %

By looking at the difference between training and test sets, we are concerned that there may be overfittings on the Random forest regression and Decision tree regression.
Due to that reasons, we pick linear regression model to estimate house prices.



	Coefficients
Rooms	123,533.92
Distance	-40,366.56
Bathroom	236,699.13
Car	62,542.82
Landsize	9.07
Longtitude	1,022,057.82
Age	3,789.91
BuildingArea	1,677.59
Lattitude	-1,641,505.39
h	122,327.39
t	-32,187.81
u	-90,139.58

Interpretation of coefficients

Every additional room increases the price 123,533 dollar while other variables unchanged.
Distance is negatively related with the price and every km away from the city center decreases the price almost 40 thousand dollars.
Every additional car rises price approximately 62 thousand dollars, but it is important to note that after a certain car spot, effects of an additional car spot on price might be lesser. (For ex. price differences between 1 car spot and 2 car spots should logically be more than 6 car spot and 7 car spots.)
Building area is also positively related with the price.

Further efforts could be on making a non-linear regression model or different machine learning methods that is more successful at predicting house prices without overfitting.

7. Resources and References

de Myttenaere, Arnaud, Boris Golden, Bénédicte Le Grand, and Fabrice Rossi. 2016. "Mean Absolute Percentage Error For Regression Models". *Neurocomputing* 192: 38-48. doi:10.1016/j.neucom.2015.12.114.

Harrington, Peter. 2012. *Machine Learning In Action*. Shelter Island, NY: Manning Publications.

Yardney, Michael. 2019. "Where Will House Prices Be 25 Years From Now?". *Property Update*. https://propertyupdate.com.au/where-will-house-prices-be-25-years-from-now-2/.