Abstract

Buying or selling a home is among the most significant transactions an individual conducts in their lifetime. In this project I utilized a data mining lens to predict how much people are willing to pay for a particular property in the region of Melbourne, Australia based on historical property auction data. The goal is to be able to support buyer and seller decision-making in the “real time” dynamic of the auction experience. I employed python libraries for exploratory data analysis and machine learning modeling. The best two models are compared and analyzed.

1. Introduction

Melbourne has consistently led Australia in terms of population growth (Melbourne Demographics. Wikipedia, 2021). Approximately 80% of transactions for the purchase and sale of residential properties in Melbourne are completed through public property auctions.

The auction method so commonly used in Australia for property sales differs from the private sale/treaty method used in many other jurisdictions. The chief differences between the two methods of sale are that in the private sale method the seller states an expected sale price. In an auction, while the seller may establish a reserve price known only to the seller and the auctioneer, no expectation of a selling price is tabled for potential buyers. While the auction method is more transparent for buyers it can result in higher prices than the private sale/treaty alternative. (Frino et al., 2010). Certainly the “real time” dynamic of the auction experience increases the competitive and emotional context of the sale.

2. Analytical objectives

In Melbourne there is a need for buyers and sellers to be knowledgeable about auction market pricing that includes emotional component. Sellers by auction need to strike the right reserve price to encourage bidders. Buyers need to take care not to overbid in the heat of the moment. The need for pricing information in the Australian based context is increased as most Australian buyers do not use agents, while sellers typically do. (Just Landed, 2021).

This project attempts to predict the property auction sale prices, or in other words, people's willingness to pay a certain price for a particular property, based on data from auction property sales in 2016-2017. It makes use of several libraries in Python: pandas for data cleaning, analysis, and feature engineering, matplotlib and seaborn for data visualization, and scikit-learn and xgboost for machine learning. The goal is to train several different models (including linear, tree-based, and ensemble methods) to come up with the best model that can predict auction selling price within one standard deviation of the mean.

3. Overview of the Data

Work on the project commences with analysis of the dataset (21 columns and 13580 rows) provided for use: "Melbourn Housing Snapshot" created by Tony Pino and made publicly available at: https://www.kaggle.com/dansbecker/melbourne-housing-snapshot?select=melb_data.csv.

Notes on Specific Variables:

Rooms: Number of rooms
Price: Price in dollars
Address: address
Method: S - property sold; SP - property sold prior; PI - property passed in; VB - vendor bid; SA - sold after auction; SS - sold after auction price not disclosed. S, SA, and SP are a complete selling and will be used to filter the dataset.
Type: h - house, cottage, villa, semi, terrace; u - unit, duplex; t - townhouse
Date: Date sold
Distance: Distance from CBD (city business centre)
Regionname: General Region of the city (West, North West, North, North east …etc)
Propertycount: Number of properties that exist in the suburb.
Bedroom2: Scraped # of Bedrooms (from different source)
Bathroom: Number of Bathrooms
Car: Number of car spots
Landsize: Land Size
BuildingArea: Building Size
YearBuild: The year when the property was built

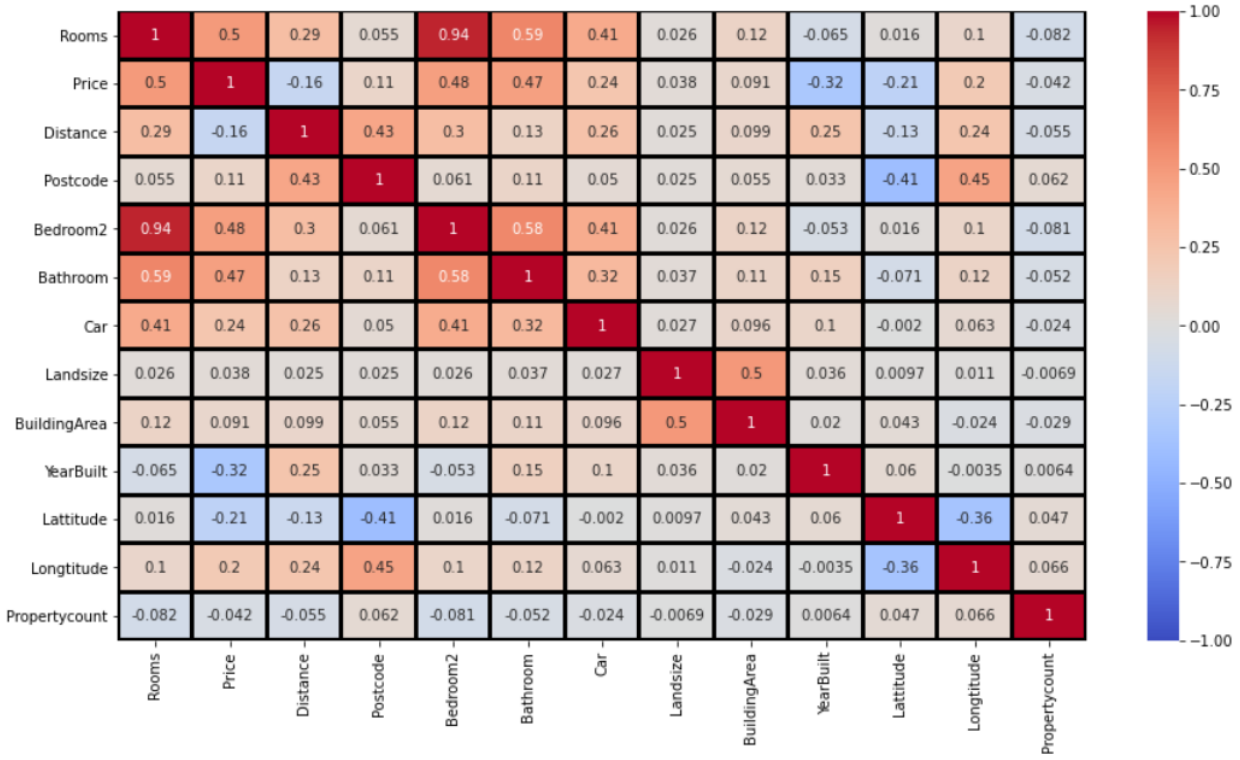
4. Exploratory Data Analysis

The following table displays general summary statistics for the continuous variables:

	count	mean	std	min	25%	50%	75%	max
Rooms	13580.0	2.937997e+00	0.955748	1.00000	2.000000	3.000000	3.000000e+00	1.000000e+01
Price	13580.0	1.075684e+06	639310.724296	85000.00000	650000.000000	903000.000000	1.330000e+06	9.000000e+06
Distance	13580.0	1.013778e+01	5.868725	0.00000	6.100000	9.200000	1.300000e+01	4.810000e+01
Postcode	13580.0	3.105302e+03	90.676964	3000.00000	3044.000000	3084.000000	3.148000e+03	3.977000e+03
Bedroom2	13580.0	2.914728e+00	0.965921	0.00000	2.000000	3.000000	3.000000e+00	2.000000e+01
Bathroom	13580.0	1.534242e+00	0.691712	0.00000	1.000000	1.000000	2.000000e+00	8.000000e+00
Car	13518.0	1.610075e+00	0.962634	0.00000	1.000000	2.000000	2.000000e+00	1.000000e+01
Landsize	13580.0	5.584161e+02	3990.669241	0.00000	177.000000	440.000000	6.510000e+02	4.330140e+05
BuildingArea	7130.0	1.519676e+02	541.014538	0.00000	93.000000	126.000000	1.740000e+02	4.451500e+04
YearBuilt	8205.0	1.964684e+03	37.273762	1196.00000	1940.000000	1970.000000	1.999000e+03	2.018000e+03
Lattitude	13580.0	-3.780920e+01	0.079260	-38.18255	-37.856822	-37.802355	-3.775640e+01	-3.740853e+01
Longtitude	13580.0	1.449952e+02	0.103916	144.43181	144.929600	145.000100	1.450583e+02	1.455264e+02
Propertycount	13580.0	7.454417e+03	4378.581772	249.00000	4380.000000	6555.000000	1.033100e+04	2.165000e+04

The first data quality conclusions:
- The extreme difference on the Price (target) variable between the max and quartile values indicates some high outliers likely skewing other statistics.
- BuildingArea and YearBuilt have a significant number of missing values, but most of the attributes are well populated.
- There are variables with "0" values (Distance, Bathroom, Bedroom2, Car, Landsize, BuildingArea).
- There is at least one awkward YearBuild value (1196) that is most likely written by mistake.

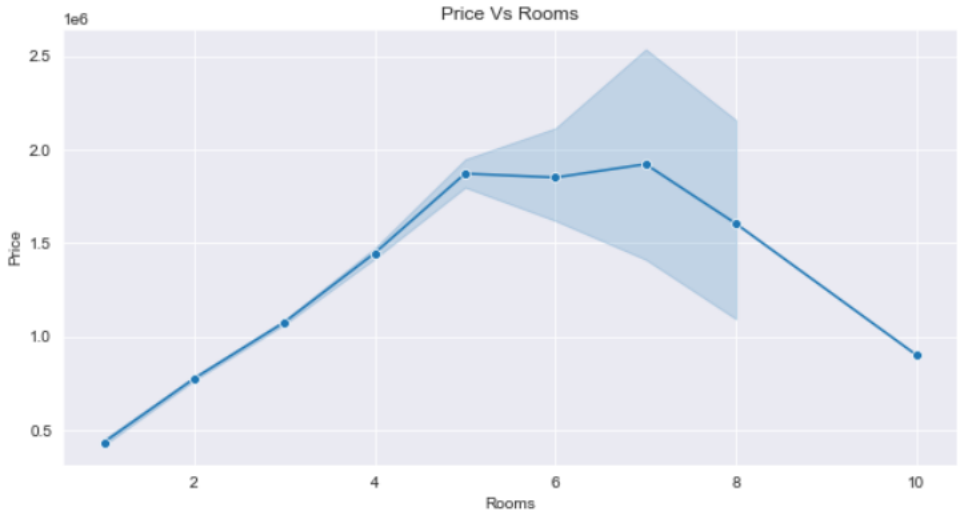
Next, we look at the linear correlation between each of the continuous variables in the dataset and check for collinearity in a heatmap. Values closer to 0 indicate weak or no correlation, positive values indicate positive correlation, and negative values indicate negative correlation.



- All continuous variables have a correlation to the 'Price' with the highest value for 'Rooms' and 'Bathroom' and can be predictor variables.
- The correlation between Rooms and Bedroom2 is higher than 0.7, so they are collinear and cannot independently predict the value of the Price. Given Bedroom2 is from different source, it will be dropped.

It is important to note that heatmaps only show linear one-to-one correlation, so it is possible that some variables are correlated to the target in tandem with each other or in non-linear ways.

Let's look at Price vs Rooms:

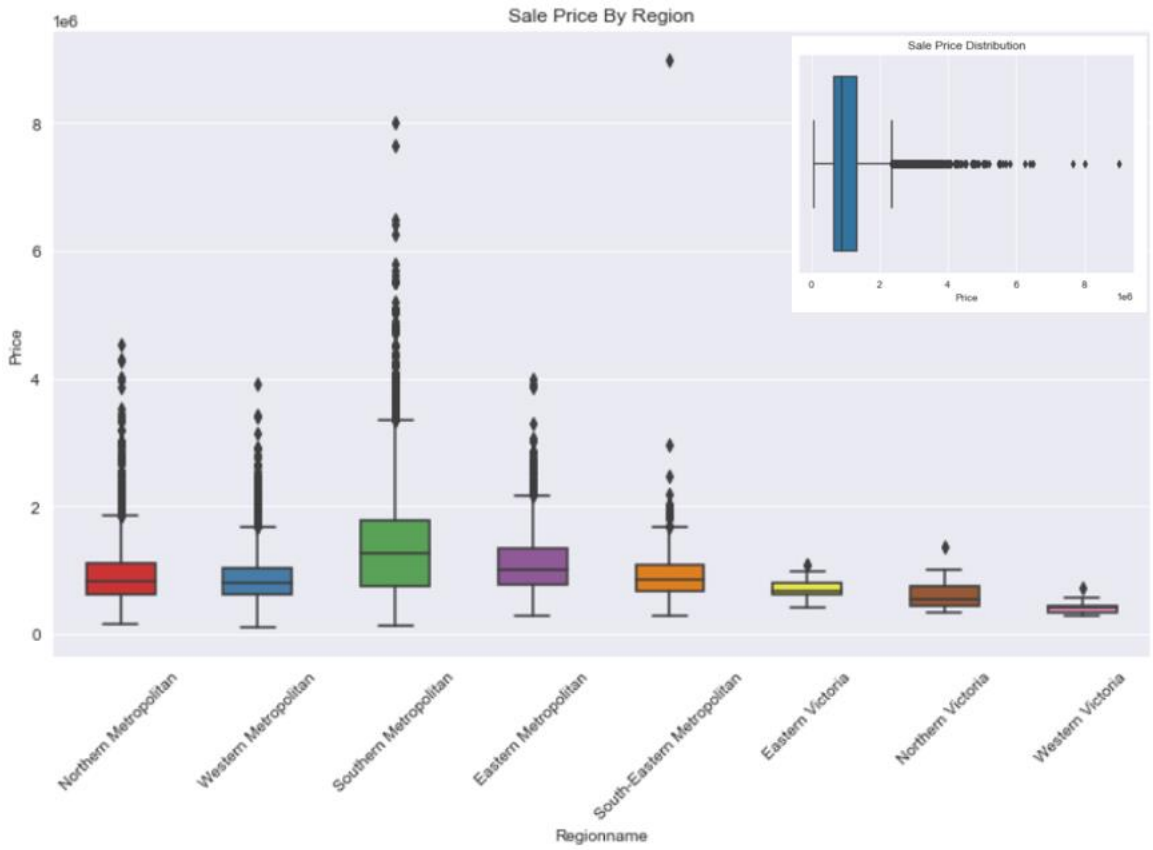


The Price is indeed higher for properties with more rooms, which makes sense since the number of rooms reflects the size of the property. However, it drops for properties with 7 rooms or more. Possible reason is that larger houses with more than 7 rooms may be in more rural regions where in general prices are lower and fewer people participate in auctions.

Thus, the first assumption is that the Region and Distance to the business center may influence the price.

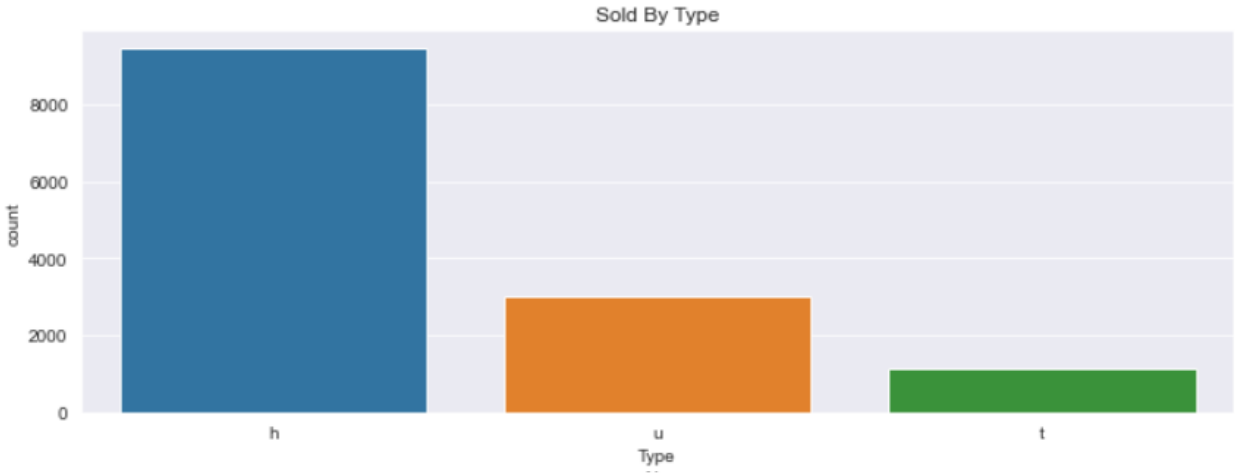
- The first coutplot shows that Southern Metropolitan has the highest house sales followed by Northern and Western Metropolitan.
- The next line chart shows that Southern Metropolitan has also the highest average sale prices whereas Western Victoria has the lowest.
- Northern, Western, Southern, and Eastern Metropolitan regions are the sale leaders, which is logical, because it appears they are located within 20 kms of city business centre.

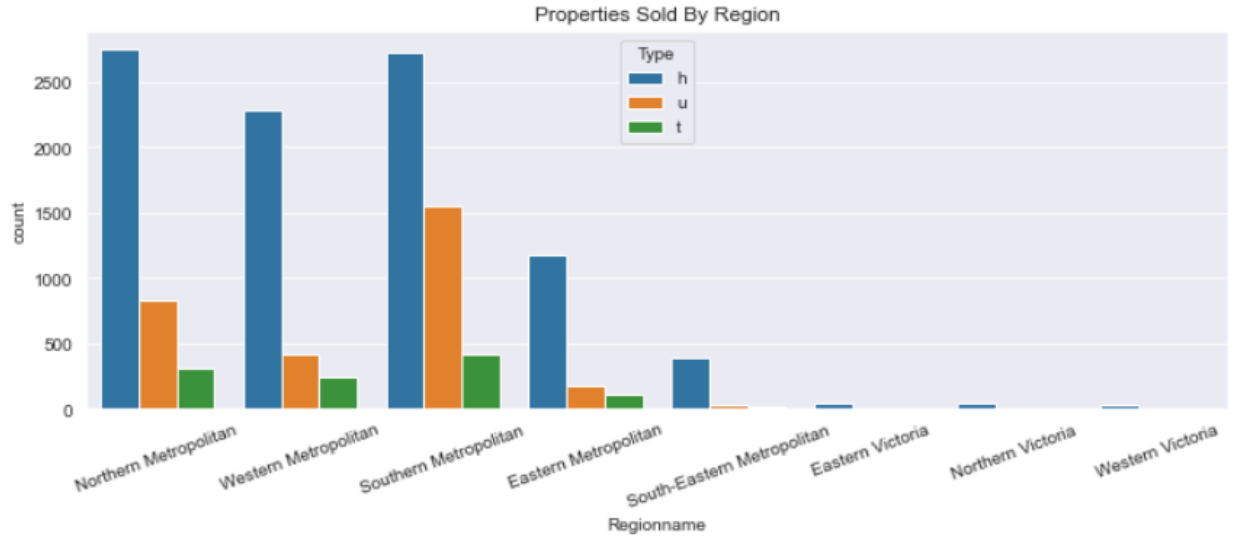
The region seems to influence the price, so property prices from all regions should be sufficiently represented in the final dataset. Let's look at the range of property prices by region and across the dataset:



Price distribution within each region is different with price outliers starting from 2M to 3M. The most houses are sold within 2,25M range. At the same time, estimating auction prices in this range is more difficult than in the luxury segment, where prices are more distinct. So, anything beyond this range can be considered as luxury (outliers) and will not be included into the final dataset.

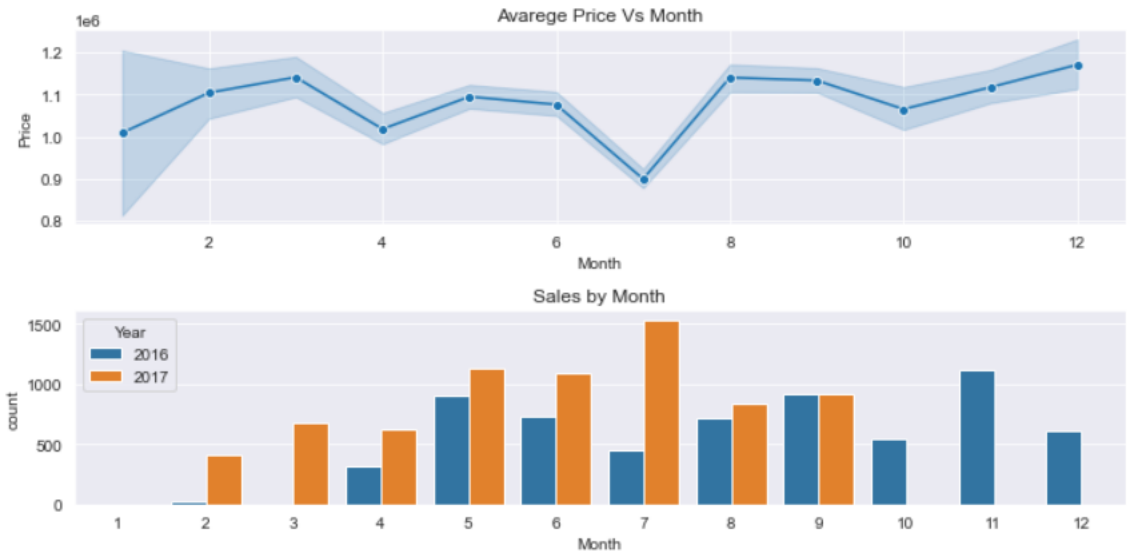
The next assumption is that the Type of the property may influence the Price.





- The dataset contains 75% houses and 25% units&townhouses.
- The h-type has the highest average price, the u-type has the lowest, and the t-type is in between, reflecting the usual distribution of prices across property types.
- Four sale leaders contain all three property types, with the first top seller having an h/u&t ratio of almost 50/50.

Another assumption is that season may influence the price. Let's look at average price and sales by month:



- We have almost 2 years of data, starting from Jan 2016 to Sep 2017, so we can compare months' sales from April to September (March, October-December data is only for one year).
- The number of sales increases from May to September although it's wintertime in Australia.
- The fall in average price in July is probably due to the large number of properties offers at lower prices in 2017.

There is not enough data for summertime (December-March), so we cannot assess the correlation between price and the season and use it for price prediction.

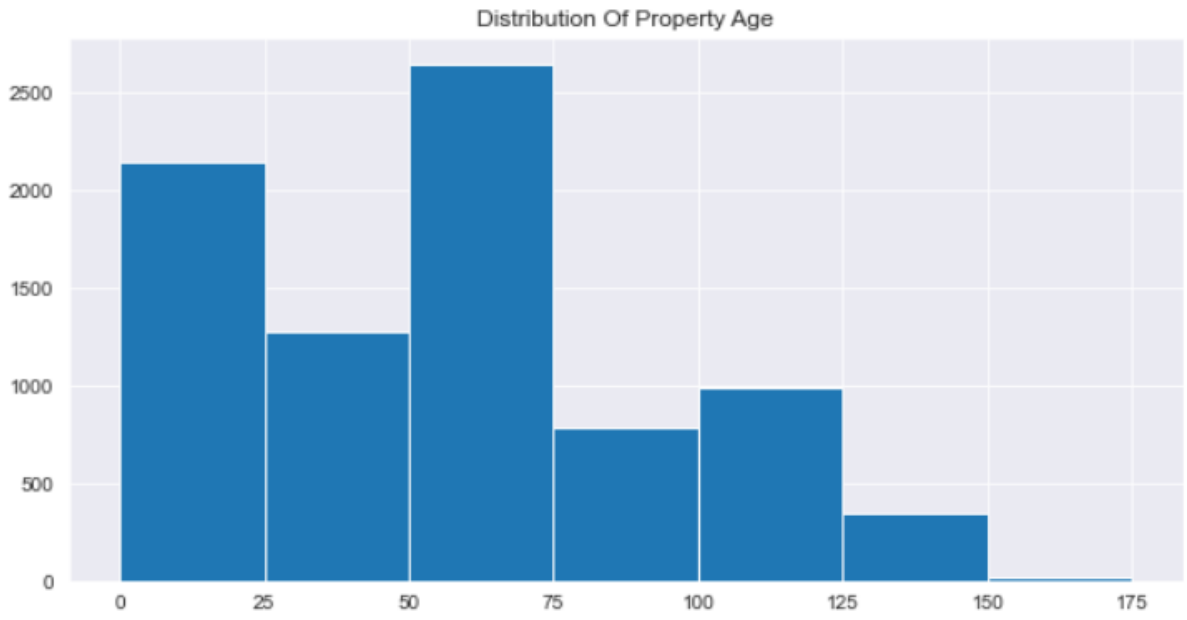
5. Missing and Null values

Missing values

"BuildingArea" has almost 50% missing data. NaN and Null values in this column will be filled with mean building area by property type.

'Car' has 62 missing values. It is possible that after filtering data by method, price, and address duplicate there will be no 'Nan' left. Otherwise, they will be filled with median value.

"YearBuilt" has almost 40% missing data, and it seems impossible to replace all missing values with maintaining prediction accuracy. However, it is possible to create a new column "Build_Age" and calculate the current age of each property that has YearBuild value.



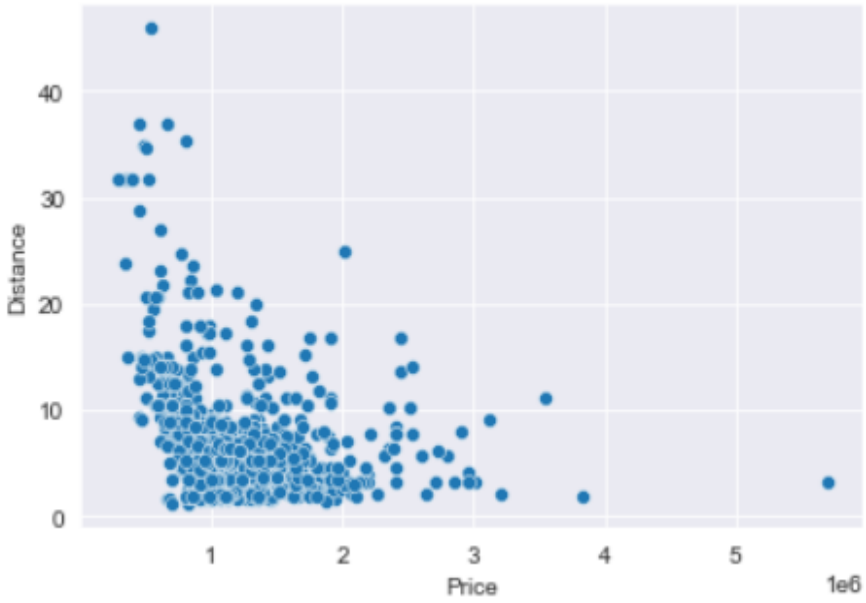
The mean and median building ages are very close, 57- and 52-year-old respectively, so either can be used. All NaN and incorrect values will be replaced with mean or median, and new column “Build_Age” will be used for price prediction instead of “YearBuilt".

Null values

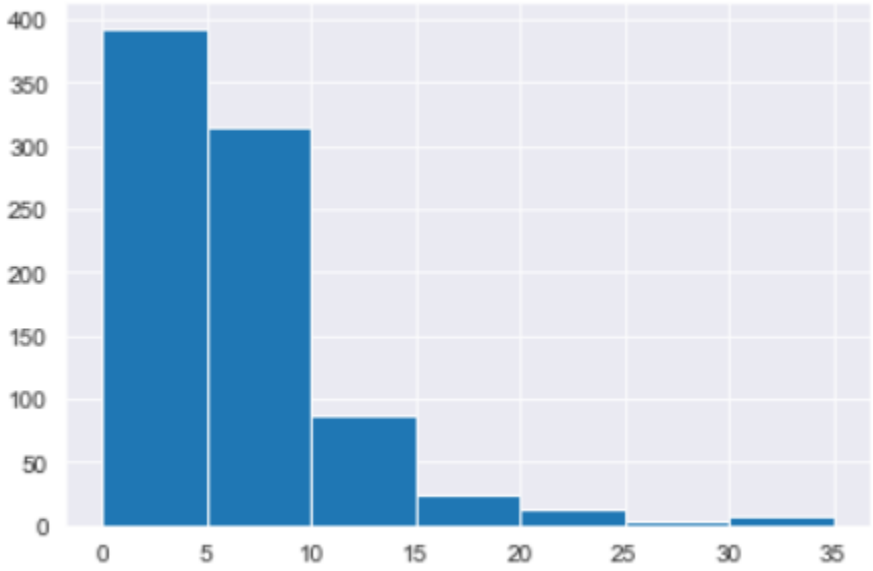
```
Number of zeros in "Landsize": 1939
Number of zeros in "Car": 1026
Number of zeros in "Bathroom": 34
Number of zeros in "Distance": 6
Number of zeros in "BuildingArea": 17
```

- “No bathroom” in property cannot be reasonably explained, and null values will be filled with median number of bathroom (1).
- The distance can be ‘0’ if the property is very close to a business center and can be left in the dataset.
- BuildingArea: Null values will be filled with mean building area by property type.
- "No car parking spots" can be true for units (apartments) and townhouses. Houses should have at least one car spot. However, in Melbourne, houses in the business center (old part of the city) may not have a place to park a car.

The distance to center and price for houses with no car parking:

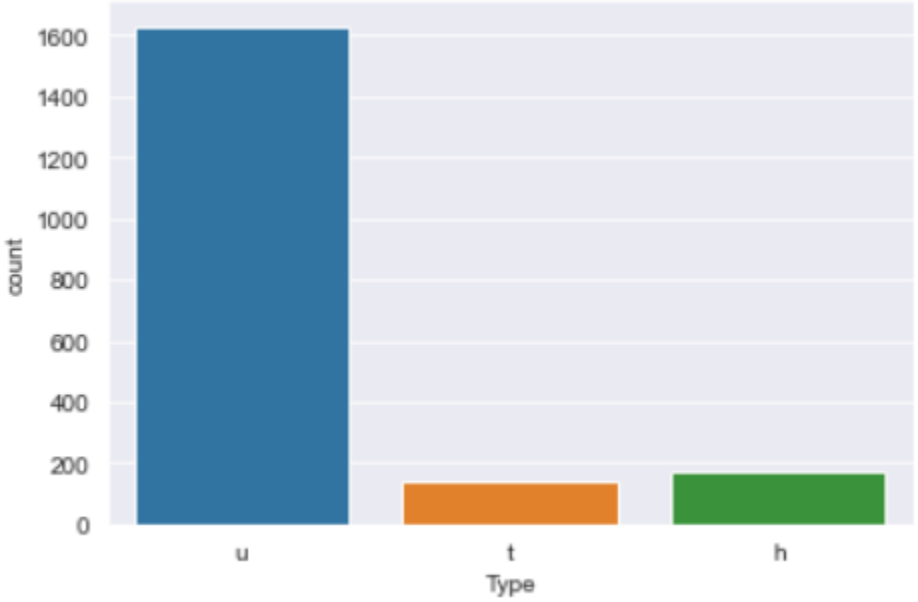


The distribution of distance to center for houses with no car parking:



No car spots for almost 400 houses can be explained by the localization within the business center (0-3 km), but null values for other houses will be replaced with median car spots (2).

- “Landsize” with null values may be associated with units (apartments) that do not own land.
  Property with no land by type:



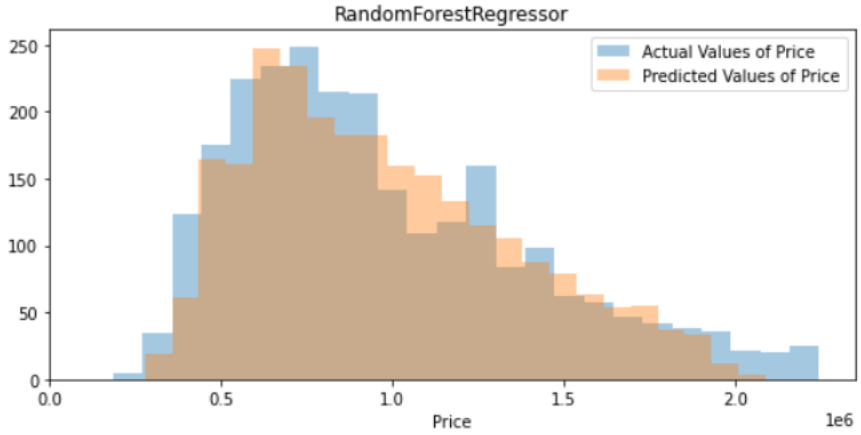
Indeed, the count plot shows that the main type of property without land is unit (apartment). Houses and townhouses can also have zero land depending on distance to the centre where land is expensive and type of residential development (row-houses, semi-detached). Null values will be left in this column.

6. Modeling

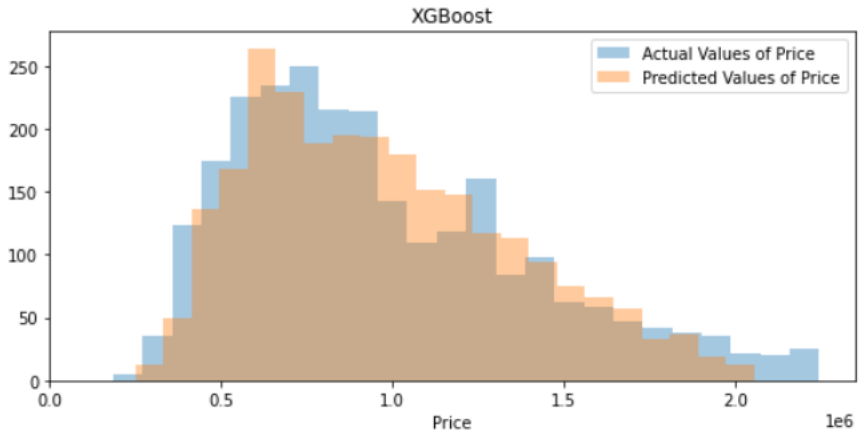
Before training the machine learning models, the data was pre-processed according to the decisions made during the exploration phase. The data also was got into a format that the models can understand and perform well on. Then, there were executed 11 models (linear, regression and ensemble methods) to find which ML model worked best for this dataset, and tune parameters for this model.

Two metrics from the scikit-learn library were used to measure performance. R2 (pronounced "R squared") evaluates the fit of the model, with 1.0 being a perfect fit. The Root Mean Squared Error (RMSE)- the square root of the average of the set of squared differences between each prediction and the real value. Since the standard deviation of auction sales price is around $423,000, a good model should have an RMSE well below that.

Two algorithms that performed the best on the dataset are called "ensemble methods," since they combine several models (called weak learners) into a one (called a strong learner). Random Forests work by creating multiple decision trees and then averaging their predictions. Specifically, each tree is allowed to use only a subset of rows, so that each tree will make slightly different predictions. Extreme Gradient Boosted Regressor builds decision trees one at a time, with each tree (called a base learner) learning from the mistakes of the previous tree.



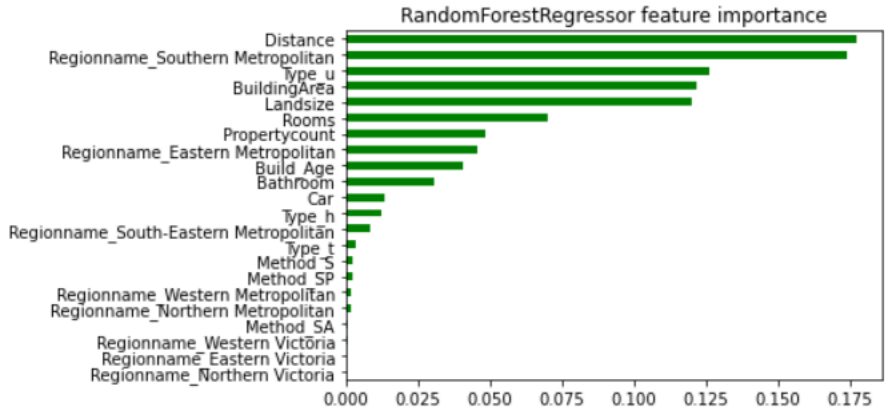

```
Rsquared = 0.8100968208308773
RMSE: 187896.9599204414
```



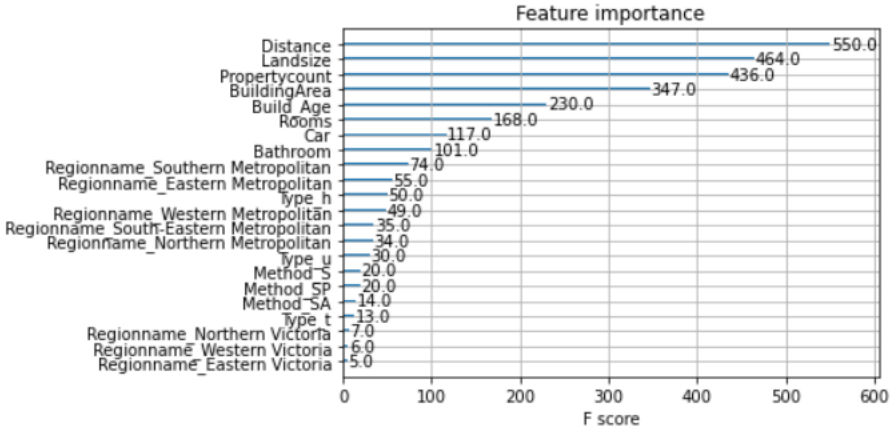

```
Rsquared = 0.82
RMSE: 184750.8362891691
```

Although Extreme Gradient Boosted Regressor is a bit better for the problem, it has the same prediction behaviour as Random Forest Regressor. Both models tend to underprice in the range of 250-900,000 and overprice in the range of 1-1.25 million.

It is interesting to compare features that are most important in explaining the auction price at each model level. For Random Forest Regressor the function feature_importances shows how much each feature contributes to decreasing the weighted impurity of the dataset and giving more information about the auction price. XGBoost evaluates the importance of features using the f(feature)score, which is the number of feature occurrences selected in the trees.



XGBoost



It seems that the top 4 most important features for both models are:

- Distance to the business center
- Landsize of the property
- Building area
- Number of rooms

7. Conclusion

The XGBoost Regressor performs the best on our dataset. Future work could tune this model further by adding new features. For example, we could add swimming pools, school district scores, crime count, floodplain, and average property tax by location as these can affect how willing people are to bid on a particular property.

Given the auction price is a result of ‘’voting’’, it is sensitive to unpredictable circumstances. Therefore, the model should be updated every quarter to keep it as accurate as possible while giving it time to accumulate new data, especially in our times of economic uncertainty.

References

Demographics of Melbourne (2021) In Wikipedia
https://en.wikipedia.org/wiki/Demographics_of_Melbourne

Frino, A., Lepone, A., Mollica, V., (2010). The Impact of Auctions on Residential Sale Prices: Australian Evidence. Australasian Accounting Business and Finance Journal 2010. Volume 4, Issue 3
https://ro.uow.edu.au/cgi/viewcontent.cgi?article=1092&context=aabfj