House Price Prediction Project

03.04.2022

—

Vinay Santosh

PGP DSBA

March 2021 - B

Contents

List of Figures

List of Tables

Introduction

- Predicting the price of a house can be challenging. Many guesses on the price can be made based on the history of a house such as its location, the age of the property, the size and condition at which it is sold.
- However these guesses can only be limited to a basic understanding of the housing market and human intuition.
- In order to make predictions more accurate, eliminate errors in judgment and have consistent results based on data, we implement various analytical as well as machine learning tools to help come up with the best house price prediction model.
- The following report provides a step by step process involved in building the most optimal model for this prediction problem.
- All of the analysis is done using the tools and concepts that were taught during the course of the Data Science and Business Analytics program. The appropriate code used for the analysis is shared along with this report in a python notebook file.

Exploratory Data Analysis

- The given dataset contains around 21,600 entries of data on houses sold over a period of 1 year.
- Each entry contains about 22 variables about the house such as its measurements, condition, location etc.
- All the properties are located in the state of Washington, around the city of Seattle and its suburbs in the United States.+
- We can observe that the dataset contains a mixture of numeric, categorical and date data types.
- The target variable for this problem statement is the **'price'** variable which is in dollars.
- In this section, we will understand the relationship of the independent variables with the target variable and each other.
- Post removal of the ID variable which will be insignificant in the analysis, we have a total of 21 variables to work with.
- We observe that the dataset contains a total of 689 null values present in it.
- Upon further inspection, we also find anomalies in the data that will be treated in the following sections.

	count	mean	std	min	25%	50%	75%	max
price	21613.0	540182.2	367362.2	75000.0	321950.0	450000.0	645000.0	7700000.0
total_beds	21505.0	3.4	0.9	0.0	3.0	3.0	4.0	33.0
total_baths	21505.0	2.1	0.8	0.0	1.8	2.2	2.5	8.0
living_measure	21596.0	2079.9	918.5	290.0	1429.2	1910.0	2550.0	13540.0
lot_measure	21571.0	15104.6	41423.6	520.0	5040.0	7618.0	10684.5	1651359.0
ceil_measure	21612.0	1788.4	828.1	290.0	1190.0	1560.0	2210.0	9410.0
basement	21612.0	291.5	442.6	0.0	0.0	0.0	560.0	4820.0
living_measure15	21447.0	1987.1	685.5	399.0	1490.0	1840.0	2360.0	6210.0
lot_measure15	21584.0	12766.5	27287.0	651.0	5100.0	7620.0	10087.0	871200.0
total_area	21613.0	17164.8	41550.6	1423.0	7041.0	9578.0	12960.0	1652659.0

Table I

- The average price of the houses in the dataset is close to 500K starting from 75K and goes upto 7.7M.
- The majority of the houses are priced between 250K - 500K.
- **93.3%** of houses are priced under 1M and only **1%** of the houses are over 2M.

Fig I

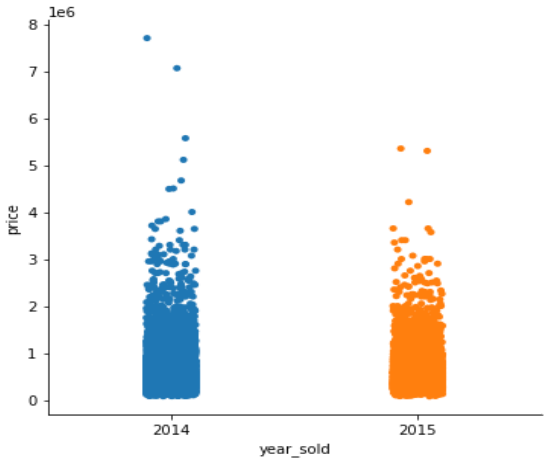

Fig II

- The **'dayhours'** variable indicates the date on which the property was sold. However, it needs to be formatted to an appropriate data type (date) for further analysis.
- We observe that the properties sold are from May of 2014 to May of 2015.
- The distribution of this variable against the target variable is shown in the plot below. We see that the most expensive properties are sold in the year 2014.
- There seems to be an equal distribution of average prices per month.
- We separate the data based on its datatype to perform univariate and bi-vatiate analysis.
- The data types of certain variables like **'total_area','quality'** etc were changed accordingly based on whether or not they are aggregatable.
- The renaming of certain column names was also done.

	price	total_beds	total_baths	living_measure	lot_measure	ceil_measure	basement	living_measure15	lot_measure15	total_area
0	600000	4.0	1.75	3050.0	9440.0	1800.0	1250.0	2020.0	8660.0	12490.0
1	190000	2.0	1.00	670.0	3101.0	670.0	0.0	1660.0	4100.0	3771.0
2	735000	4.0	2.75	3040.0	2415.0	3040.0	0.0	2620.0	2433.0	5455.0
3	257000	3.0	2.50	1740.0	3721.0	1740.0	0.0	2030.0	3794.0	5461.0
4	450000	2.0	1.00	1120.0	4590.0	1120.0	0.0	1120.0	5100.0	5710.0

Table II

- The above two tables show the data set (Table 1.1 - numeric 1.2 - categorical) separated for ease of analysis.

	sold_on	floors	coast	sight	condition	quality	yr_built	yr_renovated	zipcode	lat	long	furnished
0	2015-04-27	1	0	0	3	8	1966	0	98034	47.7228	-122.183	0
1	2015-03-17	1	0	0	4	6	1948	0	98118	47.5546	-122.274	0
2	2014-08-20	2	1	4	3	8	1966	0	98118	47.5188	-122.256	0
3	2014-10-10	2	0	0	3	8	2009	0	98002	47.3363	-122.213	0
4	2015-02-18	1	0	0	3	7	1924	0	98118	47.5663	-122.285	0

Table III

- We can analyze the distribution of variables by plotting a distribution plot to visualize the spread of data recorded.

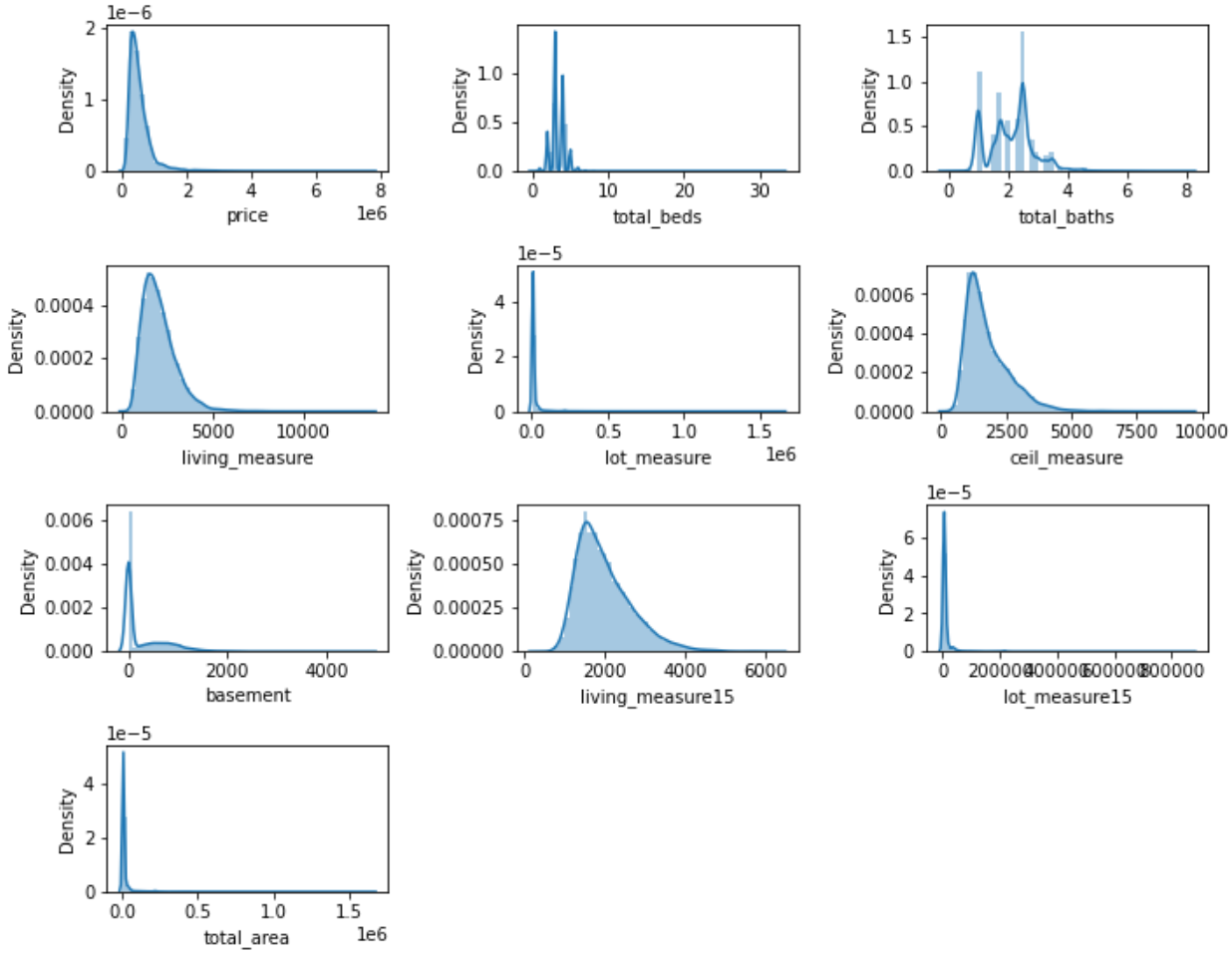

Fig III

- We notice that the majority of the variables are skewed to the right.
- The kernel density estimation shows the density of the distribution for each of the variables.

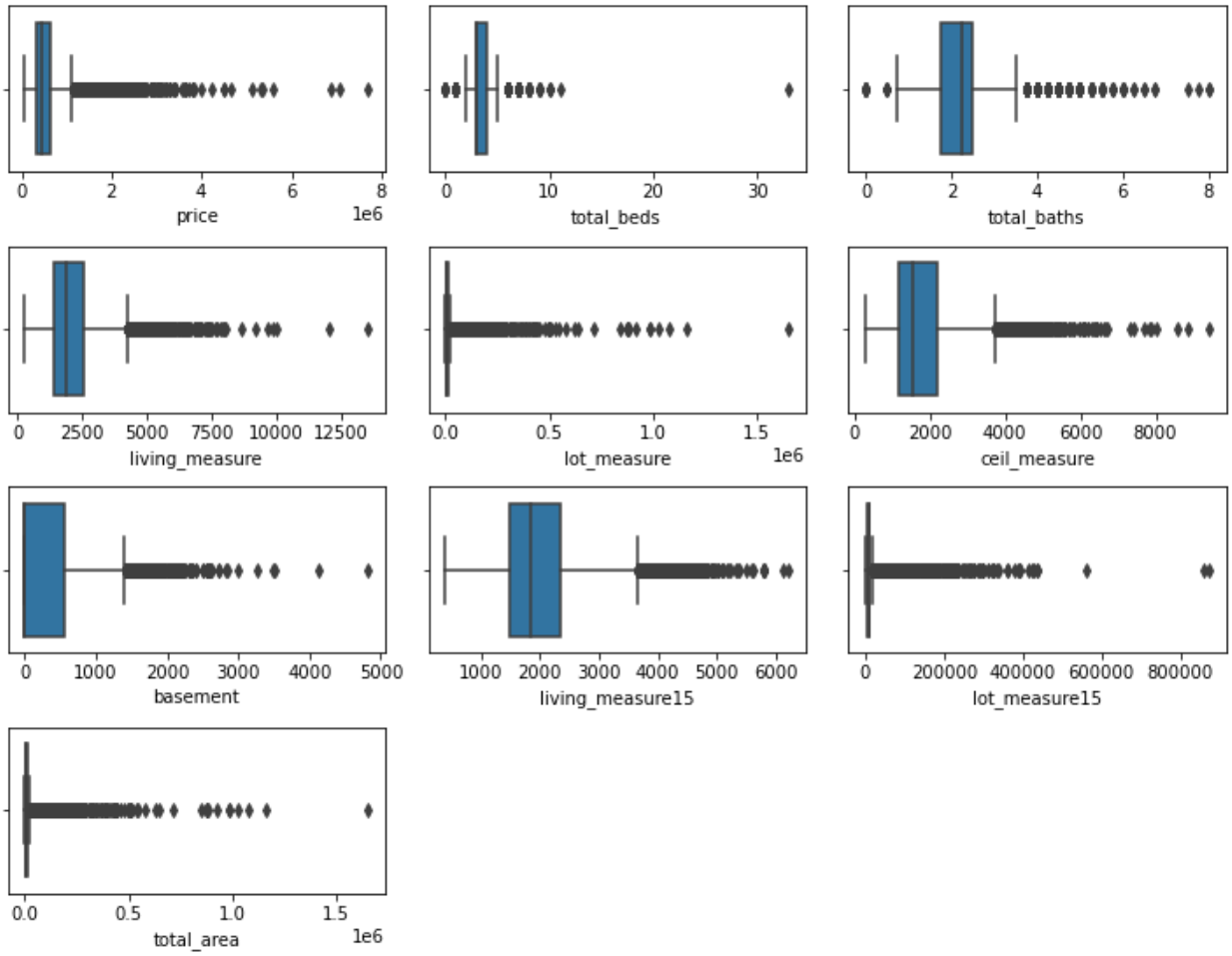

Fig IV

- The box plot above shows the presence of outliers in each of the variables.
- This indicates that there are extreme values present in variables such as **'total_beds'**, **'lot_measure'**.
- The median for the total number of bathrooms and bedrooms per house is 2 and 3 respectively.

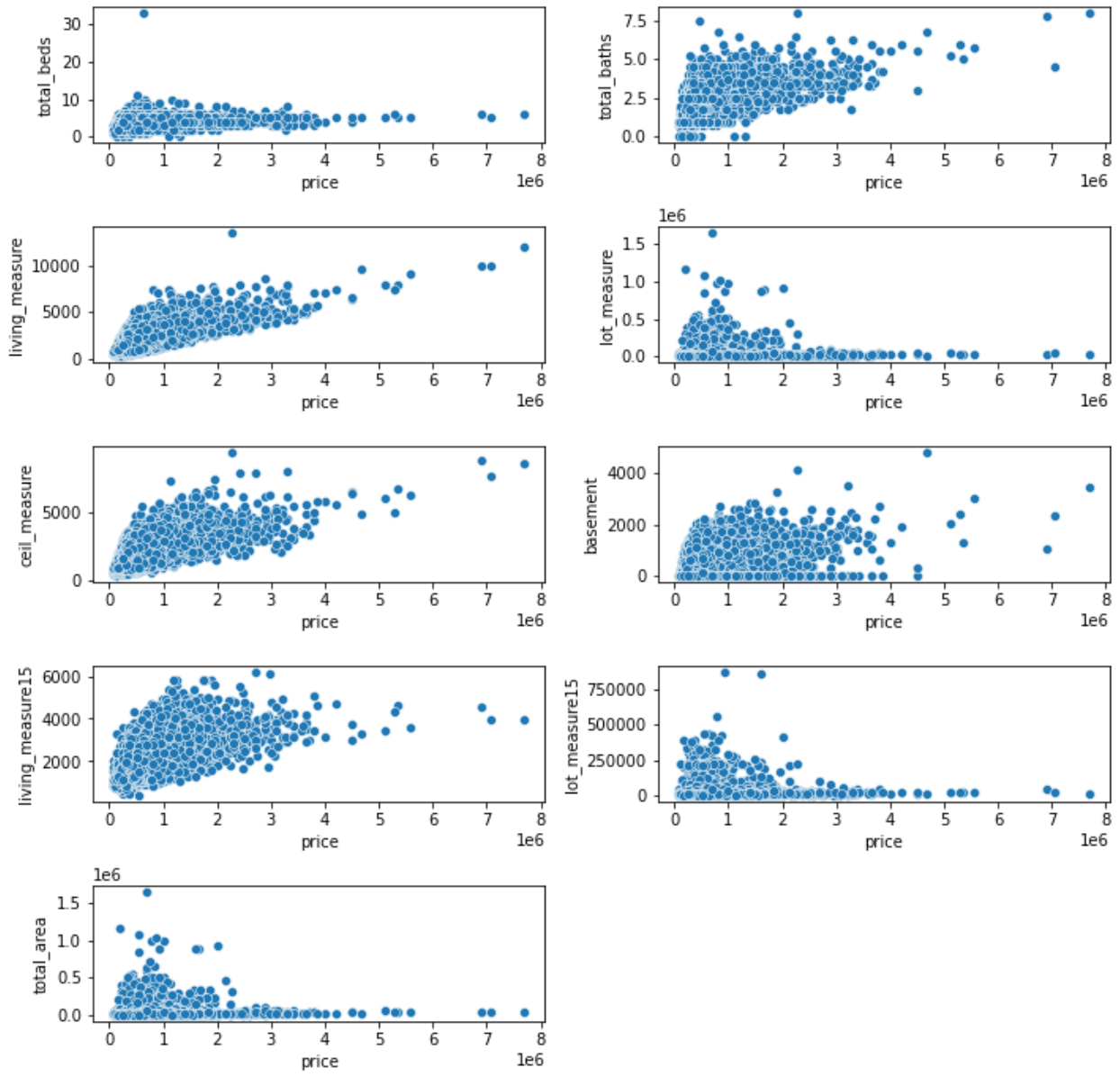

Fig V

- By plotting a scatter-plot for the numeric variables with the target variable, we find that there is a linear relationship between them.
- Further, we can deduce that on an average the price of a property is higher when it has a high condition rating and has more than 2 floors.

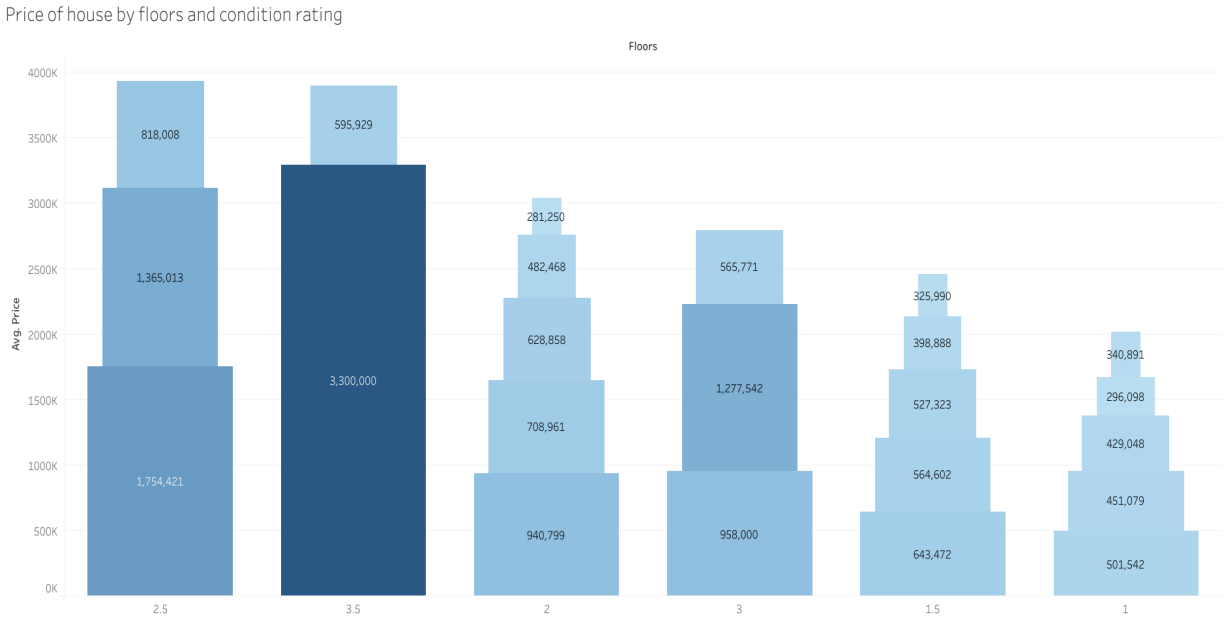

Fig VI

- The average price of a property with 3.5 floors and a condition rating of 5 is 3.3K which is the highest.

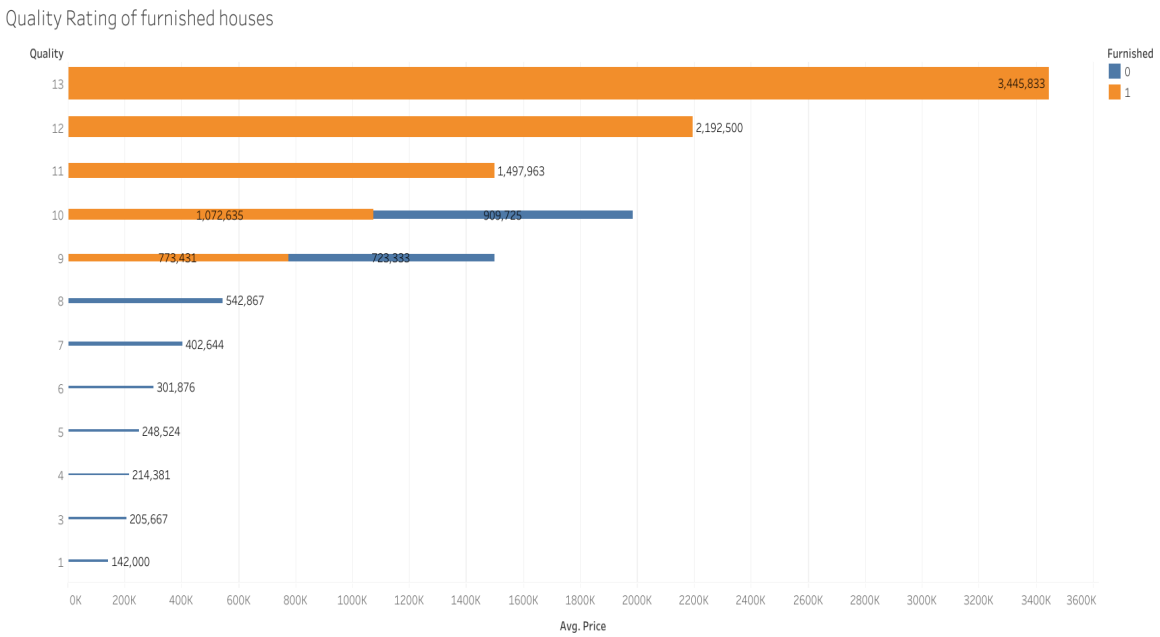

Fig VII

- Also, the average price of a house increases with an increase in quality rating.

- Furnished houses are sold at a higher price and have a minimum of 9 rating for quality.

Data Pre-processing

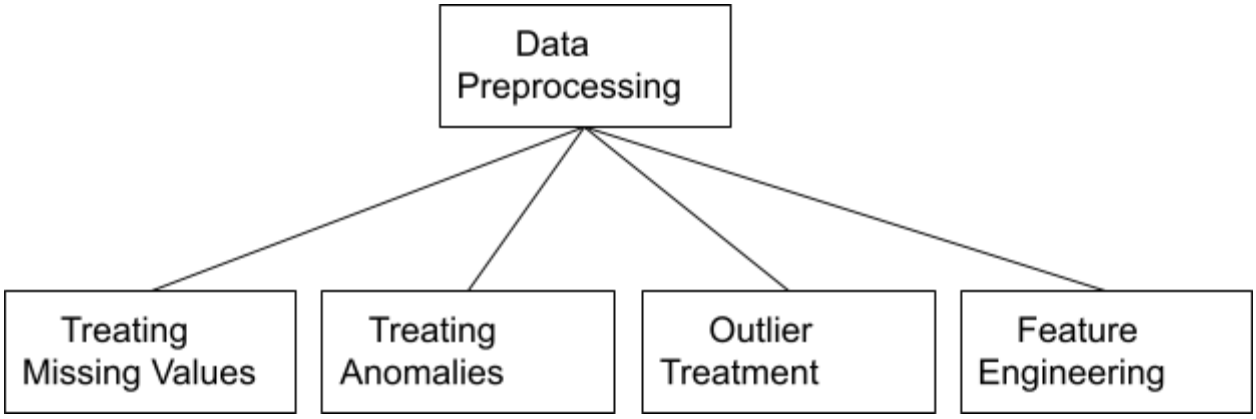

Fig VIII

- The above chart shows the various data preprocessing techniques involved in this prediction project.

Treating Missing Values:

- The given dataset has about *** missing or null values in total.
- The table below shows the total number of missing values in each variable/feature.
- To treat the missing values in the numeric data, we impute them with the median values of the respective feature.
- Since the number of datapoints is not large, we take care in handling the missing values rather than dropping them.
- For the variables like '**year_bulit**' and '**long**', we cannot impute the values based on the median and so we can go ahead and drop them.

Treating Anomalies:

- Upon analysis, we can observe that many of the categorical variables have anomalies such as “$” in them.
- These anomalies are treated based on the variable they are found in.
- For most cases, the anomalies are replaced by the most frequent or the median value.

Outlier Treatment:

- As shown in the EDA section, we notice that most variables have outliers present in them.
- Since we will be building linear regression models it is key to treat these outliers as these models are sensitive to them.
- We treat the outliers by capping the extreme values to the 75th percentile for higher extremes and 25th to the lower extremes.
- This will improve the performance of the regression model. However, treating outliers is not always mandatory.

Feature Engineering:

- With the given set of features, we can come up with new variables that will fetch us valuable insights.
- By extracting the year from the ‘**sold_on**’ variable, we can calculate the age of a house with the help of the **‘year_built’** variable.
- We can also segment the data based on the price range it lies in to understand the variation in prices of houses.
- To build a model that gives us the best result, we remove insignificant features.
- This needs to be done carefully as we do not want to eliminate data that is important for the prediction.

- Since the variable “coast” is unbalanced, i.e majority of the properties (99%) are non-coastal properties, we can make a decision to drop the column.
- Similarly, we will use the “**total_area**” instead of both the living_measure and lot_measure as these were identified as insignificant features from the previous Notes.

Model Building

- In order to address the regression problem, we can build linear models which are the most common and versatile of the regression models out there.
- The linear regression model finds the best fit line that is able to find a relationship between the independent variables in order to predict the dependent variable.
- It is usually represented by the equation : **y = mx + c** or **y = b0 + b1x1 +b2x2...bnxn** where y is the dependent variable, b0 is the intercept, and b1 to bn are the coefficients of independent variables.
- In the given problem statement, “**price**” is the dependent variable which will equate to a combination of coefficients of the respective independent variables and the y-intercept. These values can be derived from the stats.models library available in Python.
- Advantages of linear models are that it is easier to interpret and works well for linearly separable data.
- Handles overfitting by reducing dimensions regularization and cross-validation.
- More robust and gives consistent results.
- The disadvantages of a linear model is that they are prone to multicollinearity and outliers.

- After dropping the variables that are insignificant, we can build a linear regression model.
- The above table represents the coefficients of the independent variables required to determine the target variable.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.759
Model:	OLS	Adj. R-squared:	0.759
Method:	Least Squares	F-statistic:	2644.
Date:	Fri, 01 Apr 2022	Prob (F-statistic):	0.00
Time:	16:37:35	Log-Likelihood:	-1.9834e+05
No. Observations:	15094	AIC:	3.967e+05
Df Residuals:	15075	BIC:	3.969e+05
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.587e+07	1.29e+06	-27.841	0.000	-3.84e+07	-3.33e+07
floors	9275.6378	1420.270	6.531	0.000	6491.736	1.21e+04
sight	4.08e+04	1462.731	27.892	0.000	3.79e+04	4.37e+04
condition	2.931e+04	1732.885	16.915	0.000	2.59e+04	3.27e+04
quality	6.491e+04	1849.899	35.090	0.000	6.13e+04	6.85e+04
yr_renovated	890.0489	99.584	8.938	0.000	694.852	1085.246
zipcode	-985.2877	65.743	-14.987	0.000	-1114.151	-856.424
lat	5.725e+05	7991.275	71.637	0.000	5.57e+05	5.88e+05
long	-6.838e+04	9864.214	-6.932	0.000	-8.77e+04	-4.9e+04
furnished	6.57e+04	4260.159	15.423	0.000	5.74e+04	7.41e+04
total_beds	-1.252e+04	1577.016	-7.937	0.000	-1.56e+04	-9426.069
total_baths	2.763e+04	2551.053	10.830	0.000	2.26e+04	3.26e+04
ceil_measure	94.0401	3.150	29.855	0.000	87.866	100.214
basement	83.2164	3.488	23.856	0.000	76.379	90.054
living_measure15	46.5693	2.715	17.150	0.000	41.247	51.892
lot_measure15	-3.5131	0.558	-6.300	0.000	-4.606	-2.420
total_area	0.2405	0.472	0.510	0.610	-0.685	1.166
year_sold	2.124e+04	2157.728	9.842	0.000	1.7e+04	2.55e+04
age_of_house	1882.4292	53.617	35.109	0.000	1777.334	1987.525

Fig IX

- The RMSE (Root Mean Square Error) is the measure of the spread of the residuals around the best fit line. We can see that both the train and test sets have a similar RMSE.
- The model score for both sets is close to each other which indicates that the model performs equally well on the test set and there is no overfitting in the model.

- The p-values indicate the significance of the independent variable in predicting the dependent variable. P-values higher than 0.05 are considered insignificant and the variable can be considered for removal or adjustments.

Model Validation

- In this section, we will try fine tuning techniques in order to improve its performance.
- The Ridge and Lasso are shrinkage models that improve the linear regression model.
- The Ridge regression model's objective is to find the best fit surface by adding a penalty factor to the SSE function.
- The penalty factor comes close to zero but does become zero.
- The Lasso regression shrinkage reduces the insignificant features to a zero coefficient.
- The below table represents RMSE scores and model scores of all the models that were built.
- We can say that the most optimal model is the Polynomial Linear Regression model which gives us a score of 0.80 i.e, 80% accuracy.

Model Name	Model Scores
Linear Regression	0.76
Ridge Regression	0.76
Lasso Regression	0.70
Polynomial Regression	0.80

Table IV

Final Interpretation / Recommendation

- The area of the property does not play a huge role in determining the price.
- Quality of the building, furnishing and number of floors are key factors in prediction.
- The age of the property does not influence the price of the house.
- Houses with higher views are more likely to be sold.
- Properties closer to the city are priced higher than those away from them.