Vanika Hans

Milestone 2: House Price Prediction Project

Agenda Style

01

Data Pre-Processing

Data cleaning to prepare for modelling in next milestone

02

Contents Additional Necessary Exploratory Data Analysis

Explanation of Additional Data Analysis

03

Analytical Approach

Discussion of analytical approach for the next milestone, Milestone 3: Modelling

Data Pre-Processing

1. Removal of Unwanted Variables
2. Missing Value Identification and Treatment
3. Variable Transformation and Creation of New Variables
4. Outlier Detection and Treatment
5. Encoding categorical variables

1: Removal of Unwanted Variables

Index	Attribute	Description
0	cid	a notation for a house
1	dayhours	Date house was sold
2	price	Price is prediction target (in $)
3	room_bed	Number of Bedrooms per house
4	room_bath	Number of bathrooms per bedrooms
5	living_measure	square footage of the home
6	lot_measure	square footage of the lot
7	ceil	Total floors (levels) in house
8	coast	House which has a view to a waterfront (0 - No, 1 - Yes)
9	sight	Has been viewed
10	condition	How good the condition is (Overall out of 5)
11	quality	Grade given to the housing unit, based on grading system
12	ceil_measure	square footage of house apart from basement
13	basement	square footage of the basement
14	yr_built	Built Year
15	yr_renovated	Year when house was renovated
16	zipcode	zip code
17	lat	Latitude coordinate
18	long	Longitude coordinate
19	living_measure15	Living room area in 2015 (implies-- some renovations) This might or might not have affected the lot size area
20	lot_measure15	lotSize area in 2015 (implies-- some renovations)
21	furnished	Based on the quality of room (0 - No, 1 - Yes)
22	total_area	Measure of both living and lot

- Removed **cid** immediately – did not have any impact on the price specifically
- Continued/analyzed in Variable Transformation section:
  - **Dayhours** (cleaned in previous milestone)
  - **Living_measure**
  - **Living-measure15**
  - **Lot_measure**
  - **Lot_measure15**
  - **Yr_built**
  - **Yr_renovated**

2a: Missing Value Identification

Issue 1: The following columns/variables had missing values (NaN):

Variable Name	Number of Missing Values
living_measure15	166
room_bed	108
room_bath	108
condition	57
sight	57
lot_measure	42
ceil	42
furnished	29
total_area	29
lot_measure15	29
living_measure	17
ceil_measure	1
quality	1
yr_built	1
coast	1
basement	1

For example:

	price	room_bed	room_bath	living_measure
131	649000	NaN	NaN	1530.000000
150	1200000	5.000000	2.750000	3650.000000
269	620000	4.000000	3.000000	2130.000000
300	635000	NaN	NaN	1210.000000
1451	465000	NaN	NaN	NaN

Issue 2:
The following columns/variables had rows with $ as the value:

Variable Name
ceil
coast
condition
yr_built
long
total_area

For example:

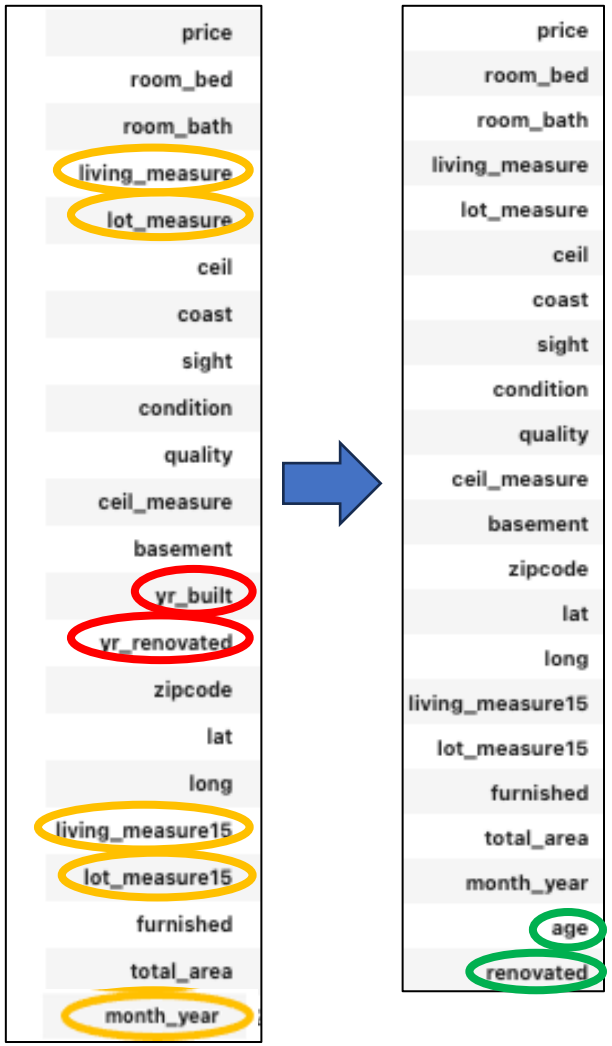
condition	quality	ceil_measure	basement	yr_built	yr_renovated	z
3	9.000000	1530.000000	0.000000	$	0	

2b: Missing Value Imputations

Issue 1: The following columns/variables had missing values (NaN)
Issue 2: The following columns/variables had rows with $ as the value

- Issue 2: Replaced all the records with the '$' symbol with NaN to be handled with Issue 1
- Issue 1:
  - General rule of thumb: if the variable distributions are normal, we can add in mean values
  - If the variable distributions are skewed, then we can add median values
  - In our case, from Milestone 1, we replaced all the missing values with the median value since they were all skewed
- Finally – we ended up with no missing values!

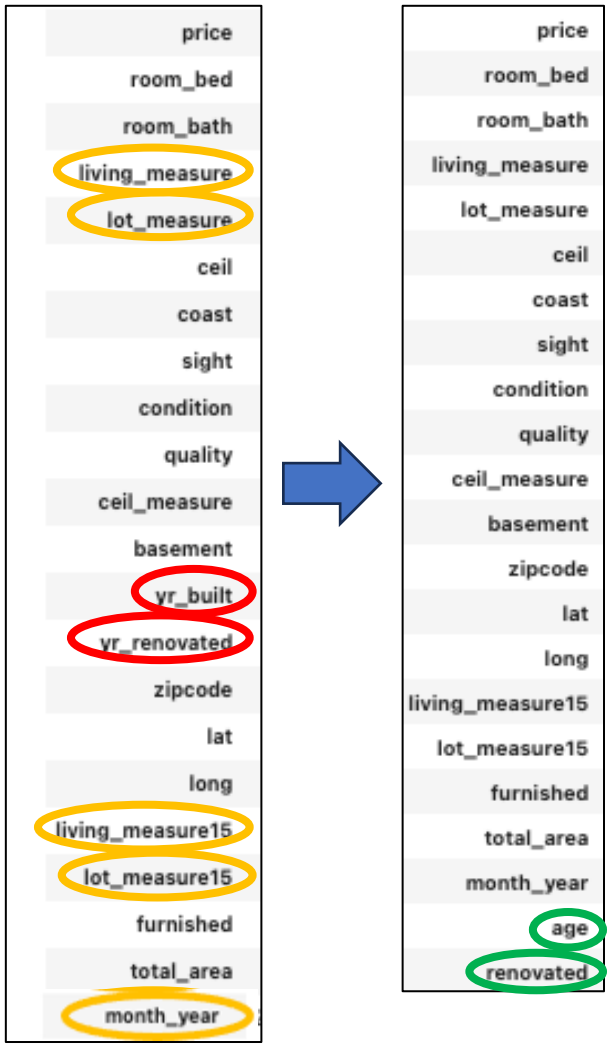
3: Variable Transformation and New Variables



- This is what the data snapshot looked like at this point in the analysis
  - cid was removed in Step 1
  - Note that dayhours was

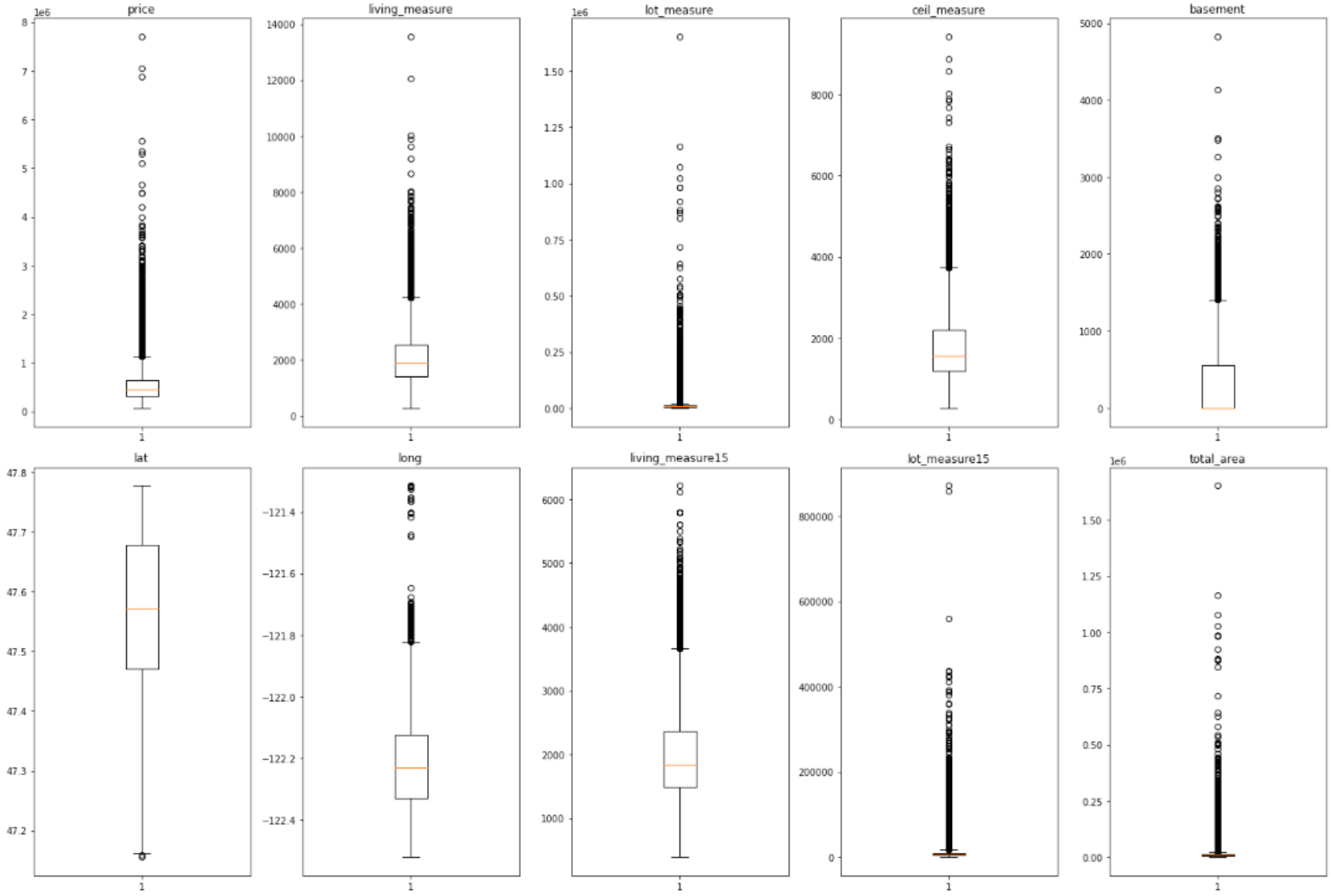
- Continued/analyzed in Variable Transformation section:
  - **Dayhours**
    - Converted from YYYYMMDDT000000 to simply MMYYYY in previous milestone, renamed as "**month_year**"
  - **Yr_built** and **month_year**
    - Combined the year the property was built and extracted the year from month_year to create a new variable: **Age**
    - Also dropped 12 rows where the Age was showing up as negative (implied issue with the provided **yr_built** or **month_year** value)
    - Kept month_year variable in case it deems useful in the model

3: Variable Transformation and New Variables



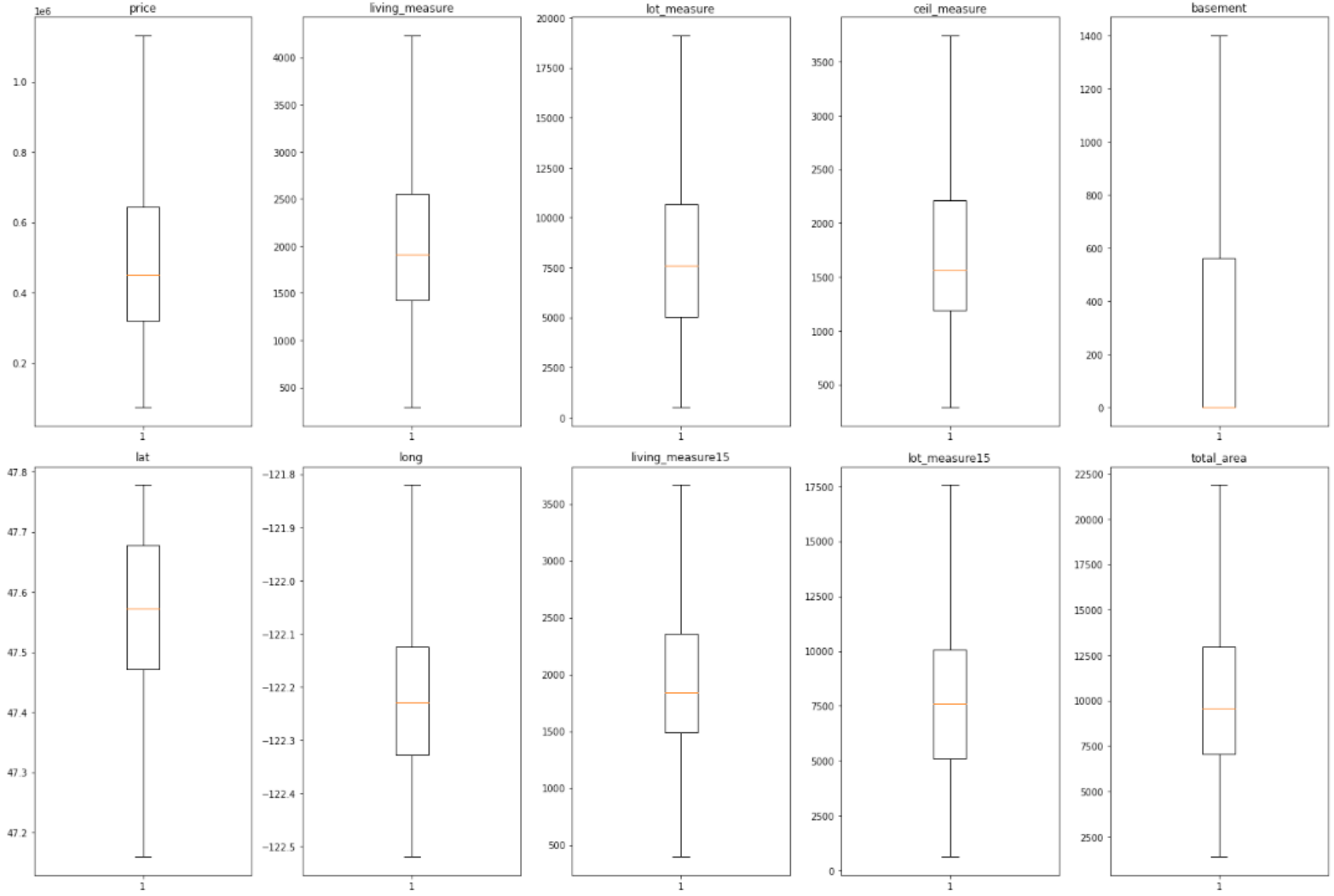
- **Living_measure, Living_measure15, Lot_measure, Lot_measure15**
  - Examined / Looked to verify the following from the variable description for living_measure15 and lot_measure15: "*Living room area in 2015 (implies-- some renovations)*" and "*lot Size area in 2015 (implies-- some renovations)*"
  - By this logic, if there were no renovations, then the area for both living_measure and lot_measure would be the same
  - Found that when there were no renovations, the values were actually not the same
  - Due to this reason, no data consolidation was done for this variable
- **Yr_renovated**
  - From the previous step, rather than focusing on what year the renovation happened, it became more important for whether or not there was a renovation
  - Transformed this variable into **renovated**

4a. Outlier Detection



Visually, summarized outliers in numerical variables. Note that categorical ‘outliers’ were not considered outliers in this project, but more of data that the model should be aware of.

4b. Outlier Treatment



Utitlized IQR technique to remove outliers in data that were outside the scope 1.5 times the IQR value. Results are as shown.

5. Encoding Categorical Columns

The dataset consisted of ordinal categorical variables and non-ordinal variables.

Ordinal Categorical Variables
Room_bed
Room_bath
Ceil
Sight
Condition
Quality

Categorical Variables
Coast
Zipcode
Furnished
Month_year

In the case of the ordinal categorical variables, i.e. where the value indicated a clear ordering of the categories (for example: 5 bathrooms is greater than 2 bathrooms, a quality value of 1 is much less than a 5):

- These values were left as is

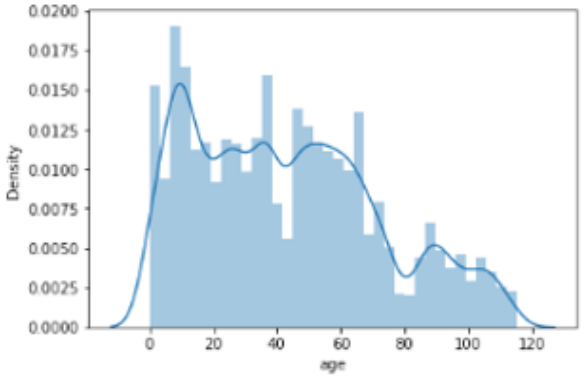
For the categorical variables:

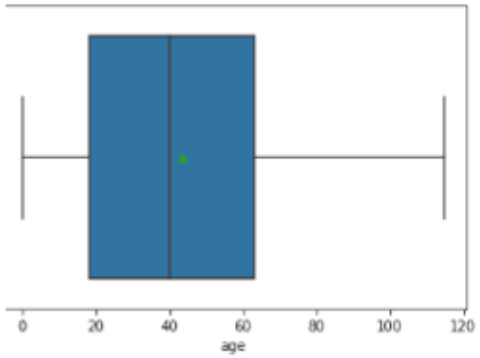
- All were already in Boolean format for the model besides zipcode and month_year
- So, I utilized one-hot encoding for these two variables in order to make the data ready for the model.

Additional Necessary Exploratory Data Analysis

- This step was done implicitly as part of previous steps
- For instance:
  - Examination of Living_measure, Living_measure15, Lot_measure, Lot_measure15 in order to determine if a pattern could be found to consolidate the variable
  - Boxplot visualization for outlier determination
    - Note: it was actually from this visualization that I determined that yr_renovated needed to be converted to a Boolean.
    - Since majority of the houses in the dataset were not renovated, the ones that were were seen as outliers and originally removed as part of the outlier removal. Changing to categorical and Boolean allowed this information to remain seen

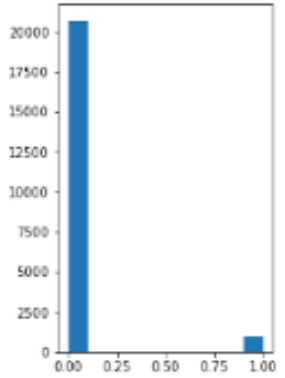
Creation of new variable Age and visualization/data cleaning:





There was an observation here of negative age and that was handled.

Creation of new variable renovated and visualization:



There were an overwhelming amount of houses not renovated, and all of the houses that were were now bucketized into one category.

Analytical Approach for Milestone 3: Modelling

- Data partitoning between training and test set
- After splitting the data, I will plan to approach this problem using the following classification algorithms:
  - Logistic Regression
  - Ridge Classifier
  - Lasso Regression
  - K-Nearest Neighbors (KNN)
  - Support Vector Machine (SVM)
  - Decision Tree
  - Random Forest
  - Gradient Boost
  - Xgboost
  - Adaboost
  - Bagging
- Will employ techniques such as pruning, feature importance, model tuning
- Model comparison using metrics
- After determining the best model: Business Insights and Recommendations