LA Housing Price Prediction Report

Data Collection and Understanding

The initial dataset provided contained information about location, type, size, use and value of properties. It was further enriched with demographic data using the US Census 2010 Decennial Dataset. The demographic data was at block level and was a macro variable for locality of property. Out of the 10,000 properties, 46 had incomplete/null demographic information in database. Such entries were dropped.
The different macro features, which could influence a properties price were assessed. The following were included in final data collection-

- Total Housing Units
- Total Population
- Occupied Houses
- Avg. Household Size
- Avg. Household Size(Occupied)
- Avg. Household Size(Rented)
- Total Household Size
- Urban Housing
- Population above 18
- Median Age
- Urban Pop

Data Description

The dataset initially contained 53 columns (whose description is provided in problem statement) and 11 columns having demographic info were added.
Out of these the information on following won't be available while making predictions. Hence these were dropped-

- LandValue
- LandBaseYear
- ImprovementValue
- ImpBaseYear
- TotalLandImpValue
- HomeownersExemption
- RealEstateExemption
- FixtureValue
- FixtureExemption
- PersonalPropertyValue
- PersonalPropertyExemption
- TotalExemption
- netTaxableValue

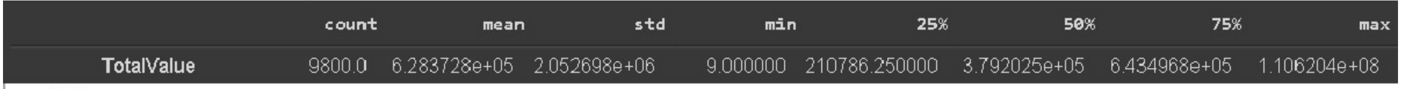
Other columns containing information of repetitive nature/ having lot's of missing information were also dropped.

- ZIPcode- Repeated column (ZIPcode5)
- AIN,AssesorID – Identifiers for Property assessors(Categorical feature), all rows having unique values hence dropped
- TaxRateArea_CITY- Meta variable for TaxRateArea i.e TaxRateArea column contains all info
- PropertyLocation – Already used with coordinates to obtain demographic location. Coordinates represent spatial info in a better way
- GeneralUseType – has only one value throughout- residential
- SpecificUseType,SpecificUseDetail1,SpecificUseDetail2 – These when combined with PropertyUseCode have the same no. of unique values as PropertyUseCode, essentially meaning that PropertiesUseCode codifies info on these
- isParcelTaxable? – has 2 values 'y'/'n' .For 'n' TotalValue=0, those rows were already dropped. Hence only 1 value remains and column could be dropped
- SpecialParcelClassification – Has lots of null values
- ParcelBoundaryDescription- Written by assessor- format highly variable- mostly contains tract and block info – which can be leveraged using coordinates and geoencoding
- HouseNo,HouseFraction,StreetDirection,StreetName,UnitNo,City – Indicate PropertyLocation, which was already used up, StreetDirection and UnitNo have lot of null values
- RowID – Unique identifier formed using AIN,Rollyear, irrelevant to model
- Location 1 – Geo coordinates of place. CENTER_LAT and CENTER_LON already represent them.

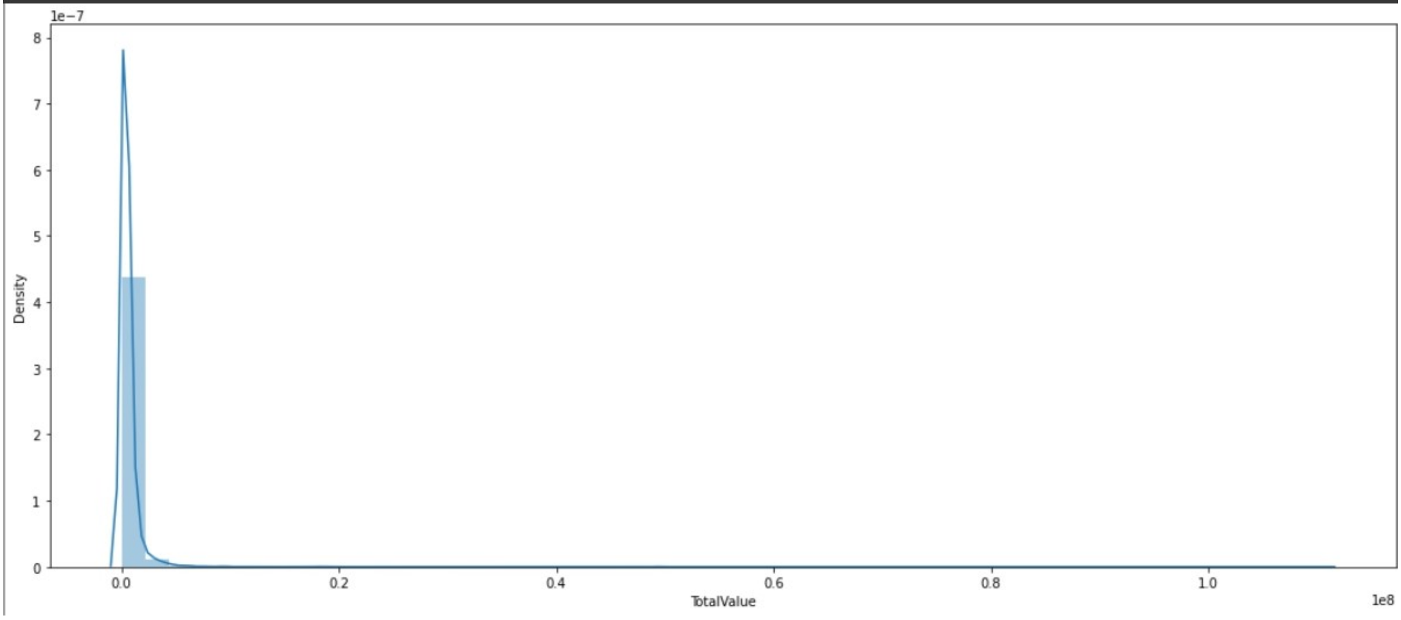
RecordingDate was converted into a new column based on difference of RecordingDate with date of valuation I.e. 1st Jan 2019 AdministrativeReigon and Cluster – These two were combined into a new categorical feature - AdministrativeRegion_Cluster PropertyUseCode had around 40 categories. These were combined into 7 broader categories-
0100 Single Family
Residence 0101 Single
Family Residence Pool
010C Single Family Residence
Condominium 02xx- 2 units
01xx- 1 units
other 05xx- 5 or
more units
03xx- 3
04xx- 4 units
Hence only 26 variables will be used in modelling.

Exploratory Data Analysis

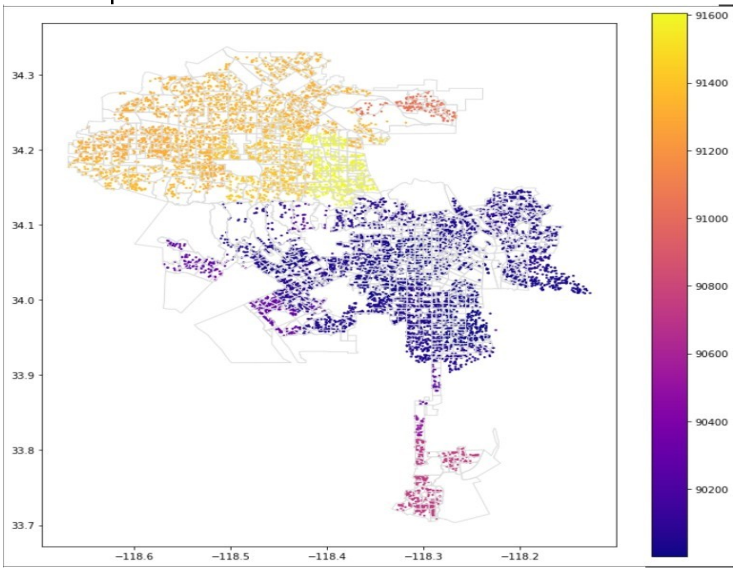
in EDA scatterplots and boxplots were used to recognize outliers for numerical featrures. However non-linear models will be used to map these.
Given below is the distplot for TotalValue, most of the houses have TotalValue in lower than $643k



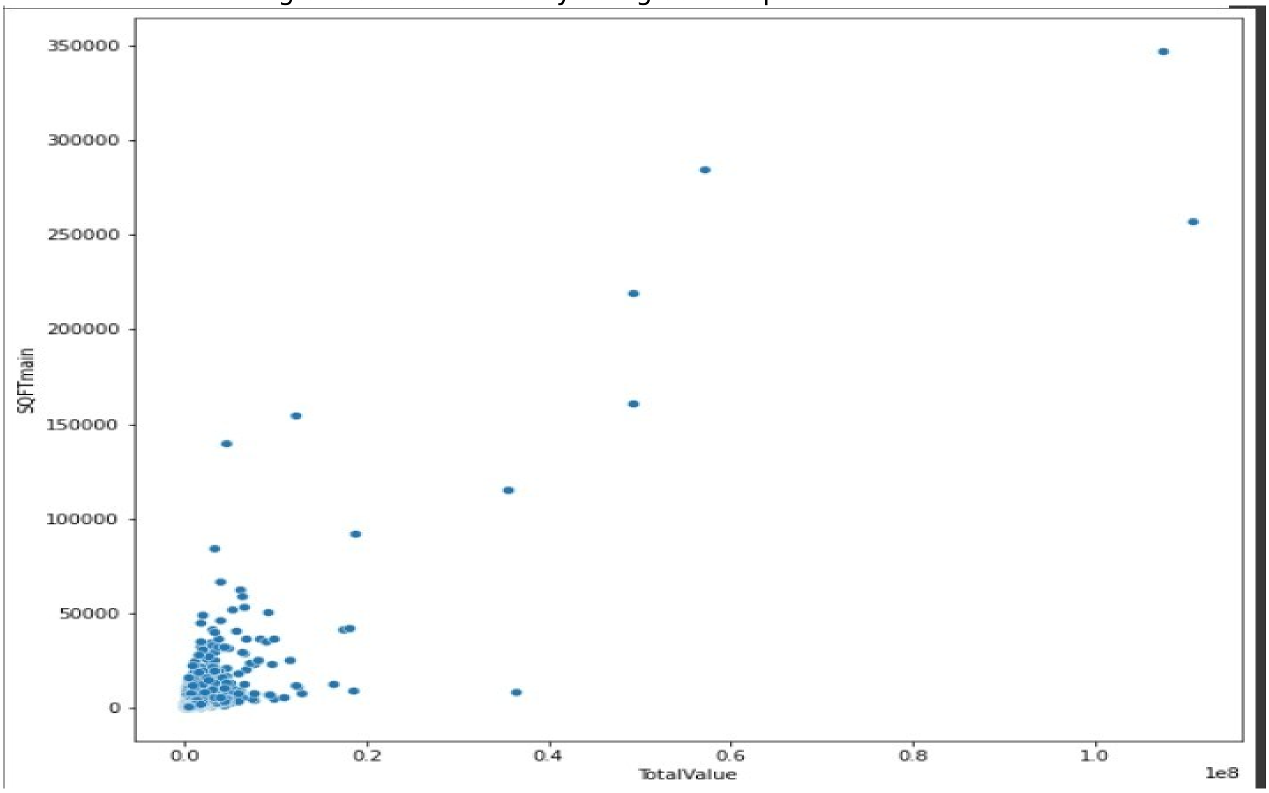
	count	mean	std	min	25%	50%	75%	max
TotalValue	9800.0	6.283728e+05	2.052698e+06	9.000000	210786.250000	3.792025e+05	6.434968e+05	1.106204e+08



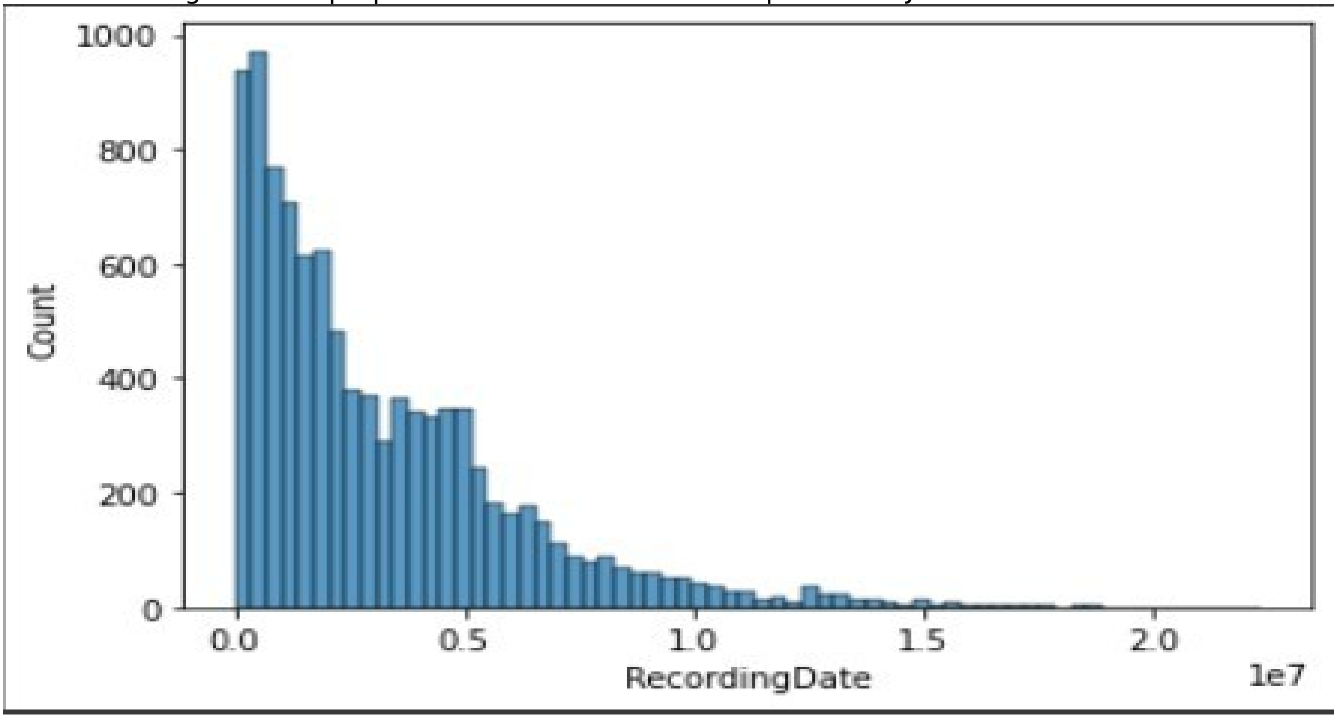
Plotting categorical features on map of LA showed how ZipCode5 and AdministrativeRegion_Cluster are highly dependent on coordinates – which was as expected
Heatmap for ZIPcode5 number



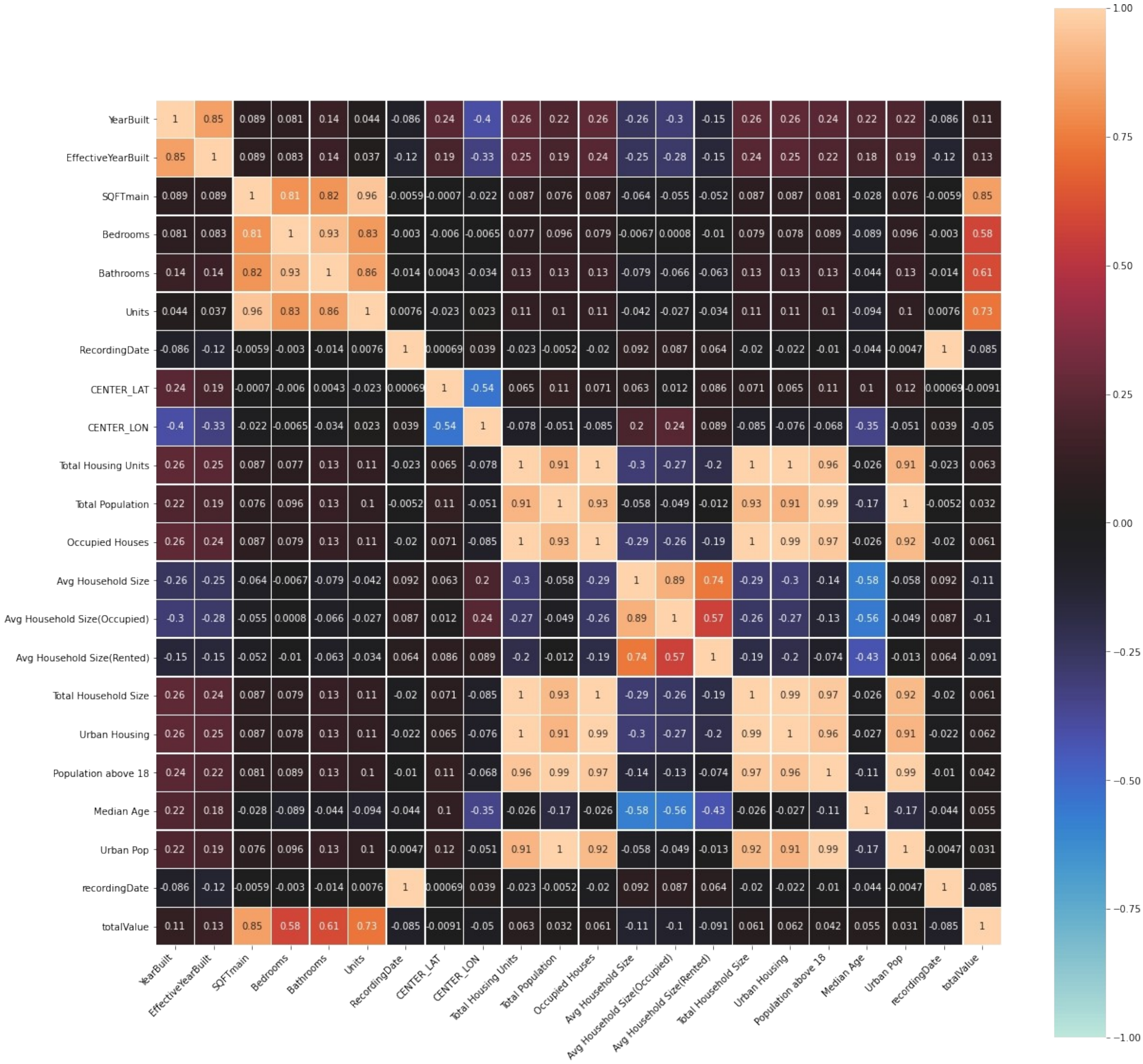
A positive correlation between TotalValue of property and SQFTmain, Bedrooms, Bathrooms and Living units Is observed by suing scatter plots



Most of the recording dates (difference between original recording date and 1st Jan 2019) have a small value meaning- most of properties were assesed in near past of 1st Jan 2019

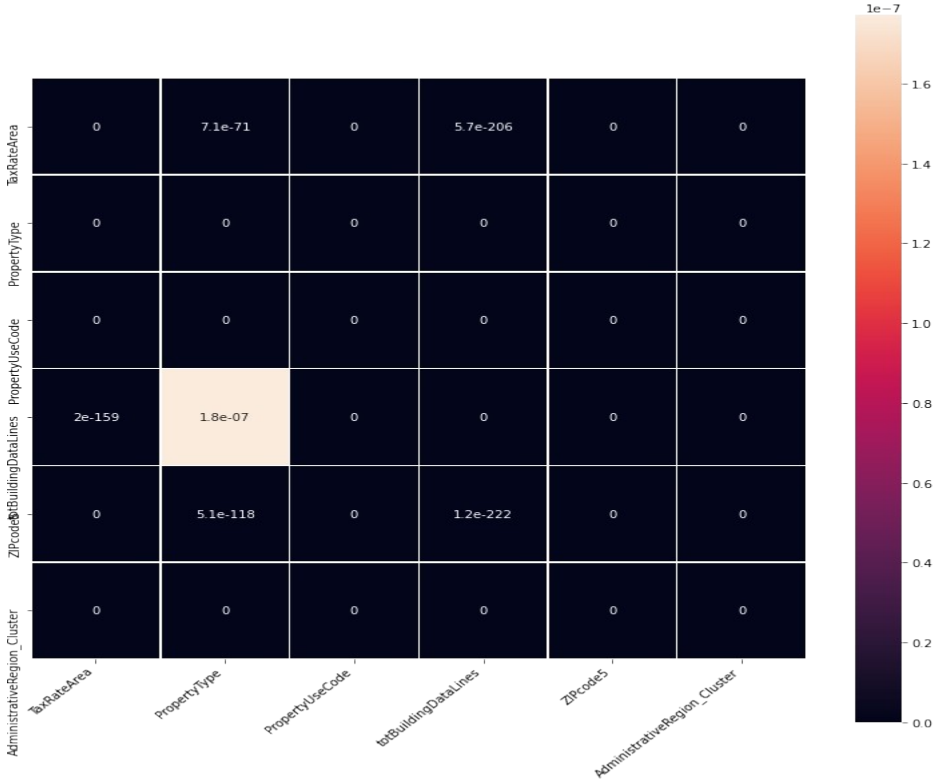


Correlations between different numerical variables were calculated.
Large Positive correlations were observed between value and size of property - TotalValue of property and SQFTmain, Bedrooms, Bathrooms and Living units as already expected.
Median Age has slight positive correlation with value – A trend which can be attributed to increasing spending capacity with age Avg Household Size has negative correlation with value meaning that poor people live in smaller houses
Yet and EffectiveYear builarBuilt also have positive correlation with value - Newer houses are more expensive.



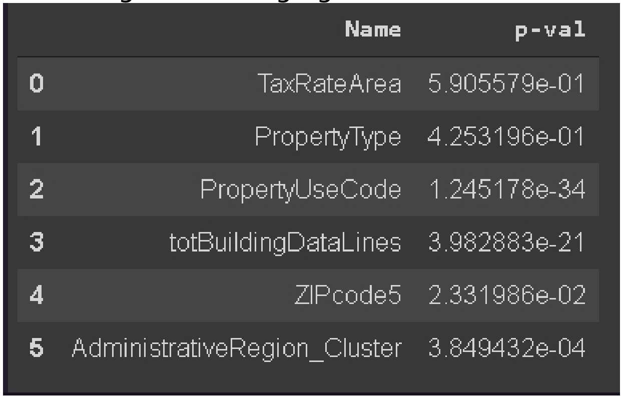
	YearBuilt	EffectiveYearBuilt	SQFTmain	Bedrooms	Bathrooms	Units	RecordingDate	CENTER_LAT	CENTER_LON	Total Housing Units	Total Population	Occupied Houses	Avg Household Size	Avg Household Size(Occupied)	Avg Household Size(Rented)	Total Household Size	Urban Housing	Population above 18	Median Age	Urban Pop	recordingDate	totalValue
YearBuilt	1	0.85	0.089	0.081	0.14	0.044	-0.086	0.24	-0.4	0.26	0.22	0.26	-0.26	-0.3	-0.15	0.26	0.26	0.24	0.22	0.22	-0.086	0.11
EffectiveYearBuilt	0.85	1	0.089	0.083	0.14	0.037	-0.12	0.19	-0.33	0.25	0.19	0.24	-0.25	-0.28	-0.15	0.24	0.25	0.22	0.18	0.19	-0.12	0.13
SQFTmain	0.089	0.089	1	0.81	0.82	0.96	-0.0059	-0.0007	-0.022	0.087	0.076	0.087	-0.064	-0.055	-0.052	0.087	0.087	0.081	-0.028	0.076	-0.0059	0.85
Bedrooms	0.081	0.083	0.81	1	0.93	0.83	-0.003	-0.006	-0.0065	0.077	0.096	0.079	-0.0067	0.0008	-0.01	0.079	0.078	0.089	-0.089	0.096	-0.003	0.58
Bathrooms	0.14	0.14	0.82	0.93	1	0.86	-0.014	0.0043	-0.034	0.13	0.13	0.13	-0.079	-0.066	-0.063	0.13	0.13	0.13	-0.044	0.13	-0.014	0.61
Units	0.044	0.037	0.96	0.83	0.86	1	0.0076	-0.023	0.023	0.11	0.1	0.11	-0.042	-0.027	-0.034	0.11	0.11	0.1	-0.094	0.1	0.0076	0.73
RecordingDate	-0.086	-0.12	-0.0059	-0.003	-0.014	0.0076	1	0.00069	0.039	-0.023	-0.0052	-0.02	0.092	0.087	0.064	-0.02	-0.022	-0.01	-0.044	-0.0047	1	-0.085
CENTER_LAT	0.24	0.19	-0.0007	-0.006	0.0043	-0.023	0.00069	1	-0.54	0.065	0.11	0.071	0.063	0.012	0.086	0.071	0.065	0.11	0.1	0.12	0.00069	-0.0091
CENTER_LON	-0.4	-0.33	-0.022	-0.0065	-0.034	0.023	0.039	-0.54	1	-0.078	-0.051	-0.085	0.2	0.24	0.089	-0.085	-0.076	-0.068	-0.35	-0.051	0.039	-0.05
Total Housing Units	0.26	0.25	0.087	0.077	0.13	0.11	-0.023	0.065	-0.078	1	0.91	1	-0.3	-0.27	-0.2	1	1	0.96	-0.026	0.91	-0.023	0.063
Total Population	0.22	0.19	0.076	0.096	0.13	0.1	-0.0052	0.11	-0.051	0.91	1	0.93	-0.058	-0.049	-0.012	0.93	0.91	0.99	-0.17	1	-0.0052	0.032
Occupied Houses	0.26	0.24	0.087	0.079	0.13	0.11	-0.02	0.071	-0.085	1	0.93	1	-0.29	-0.26	-0.19	1	0.99	0.97	-0.026	0.92	-0.02	0.061
Avg Household Size	-0.26	-0.25	-0.064	-0.0067	-0.079	-0.042	0.092	0.063	0.2	-0.3	-0.058	-0.29	1	0.89	0.74	-0.29	-0.3	-0.14	-0.58	-0.058	0.092	-0.11
Avg Household Size(Occupied)	-0.3	-0.28	-0.055	0.0008	-0.066	-0.027	0.087	0.012	0.24	-0.27	-0.049	-0.26	0.89	1	0.57	-0.26	-0.27	-0.13	-0.56	-0.049	0.087	-0.1
Avg Household Size(Rented)	-0.15	-0.15	-0.052	-0.01	-0.063	-0.034	0.064	0.086	0.089	-0.2	-0.012	-0.19	0.74	0.57	1	-0.19	-0.2	-0.074	-0.43	-0.013	0.064	-0.091
Total Household Size	0.26	0.24	0.087	0.079	0.13	0.11	-0.02	0.071	-0.085	1	0.93	1	-0.29	-0.26	-0.19	1	0.99	0.97	-0.026	0.92	-0.02	0.061
Urban Housing	0.26	0.25	0.087	0.078	0.13	0.11	-0.022	0.065	-0.076	1	0.91	0.99	-0.3	-0.27	-0.2	0.99	1	0.96	-0.027	0.91	-0.022	0.062
Population above 18	0.24	0.22	0.081	0.089	0.13	0.1	-0.01	0.11	-0.068	0.96	0.99	0.97	-0.14	-0.13	-0.074	0.97	0.96	1	-0.11	0.99	-0.01	0.042
Median Age	0.22	0.18	-0.028	-0.089	-0.044	-0.094	-0.044	0.1	-0.35	-0.026	-0.17	-0.026	-0.58	-0.56	-0.43	-0.026	-0.027	-0.11	1	-0.17	-0.044	0.055
Urban Pop	0.22	0.19	0.076	0.096	0.13	0.1	-0.0047	0.12	-0.051	0.91	1	0.92	-0.058	-0.049	-0.013	0.92	0.91	0.99	-0.17	1	-0.0047	0.031
recordingDate	-0.086	-0.12	-0.0059	-0.003	-0.014	0.0076	1	0.00069	0.039	-0.023	-0.0052	-0.02	0.092	0.087	0.064	-0.02	-0.022	-0.01	-0.044	-0.0047	1	-0.085
totalValue	0.11	0.13	0.85	0.58	0.61	0.73	-0.085	-0.0091	-0.05	0.063	0.032	0.061	-0.11	-0.1	-0.091	0.061	0.062	0.042	0.055	0.031	-0.085	1

Chi2 test between categorical variables showed that variables were not similar. All combinations had p-value smaller than 0.05. Null hypothesis stands void.



	TaxRateArea	PropertyType	PropertyUseCode	totBuildingDataLines	ZIPcode5	AdministrativeRegion_Cluster
TaxRateArea	0	7.1e-71	0	5.7e-206	0	0
PropertyType	0	0	0	0	0	0
PropertyUseCode	0	0	0	0	0	0
totBuildingDataLines	2e-159	1.8e-07	0	0	0	0
ZIPcode5	0	5.1e-118	0	1.2e-222	0	0
AdministrativeRegion_Cluster	0	0	0	0	0	0

Similarly ANOVA test was conducted. The finding showed that categories in columns TaxRateArea and PropertyCode didn’t have considerable effect on TargetValue . These columns were dropped for modelling later. Bringing down thew no. Of columns used in modelling to 24.



	Name	p-val
0	TaxRateArea	5.905579e-01
1	PropertyType	4.253196e-01
2	PropertyUseCode	1.245178e-34
3	totBuildingDataLines	3.982883e-21
4	ZIPcode5	2.331986e-02
5	AdministrativeRegion_Cluster	3.849432e-04

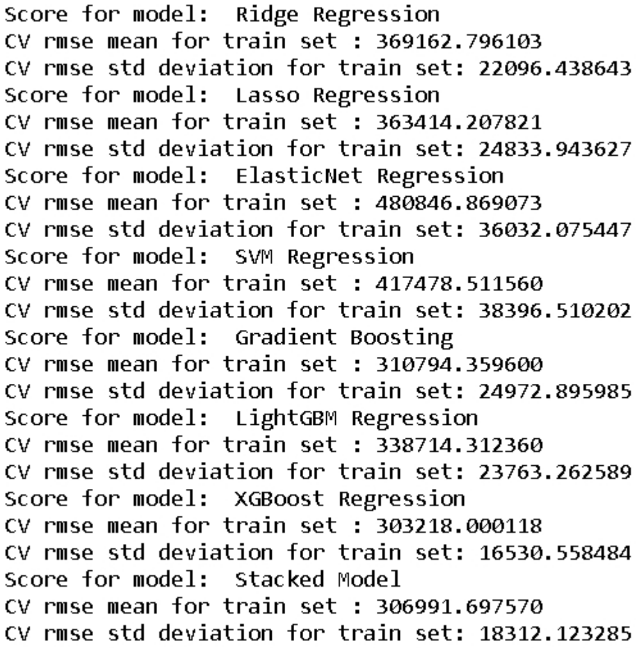
Data Preparation

For data preparation, rows having invalid entries - 0 value, 0 area, 0 Effective Year Built, 0 Year Built were dropped. These combined with entries dropped on basis of incomplete demographic data amounted to 2% of dataset.
AdministrativeRegion_Cluster, RecordingDate and PropertyUseCode were transformed as described previously. Following columns remained -
Categorical type 1 cols - ['ZIPcode5']
Categorical type 2 cols -
['PropertyUseCode','totBuildingDataLines'] Numerical cols -
['RecordingDate', 'YearBuilt', 'EffectiveYearBuilt', 'SQFTmain',
'Bedrooms', 'Bathrooms', 'Units', 'CENTER_LAT','CENTER_LON', 'Total Housing Units', 'Total Population', 'Occupied Houses', 'Avg Household Size', 'Avg Household Size(Occupied)',’Avg Household Size(Rented)', 'Total Household Size', 'Urban Housing', 'Population above 18', 'Median Age', 'Urban Pop','AdministrativeRegion_Cluster']

Categorical type 1 had large number of categories in them - These were encoded to numerical data using CatBoostEncoder. Categorical type 2 had large number of categories in them - These were encoded to numerical data using OneHotEncoder. Numerical cols were standardized.

Modelling

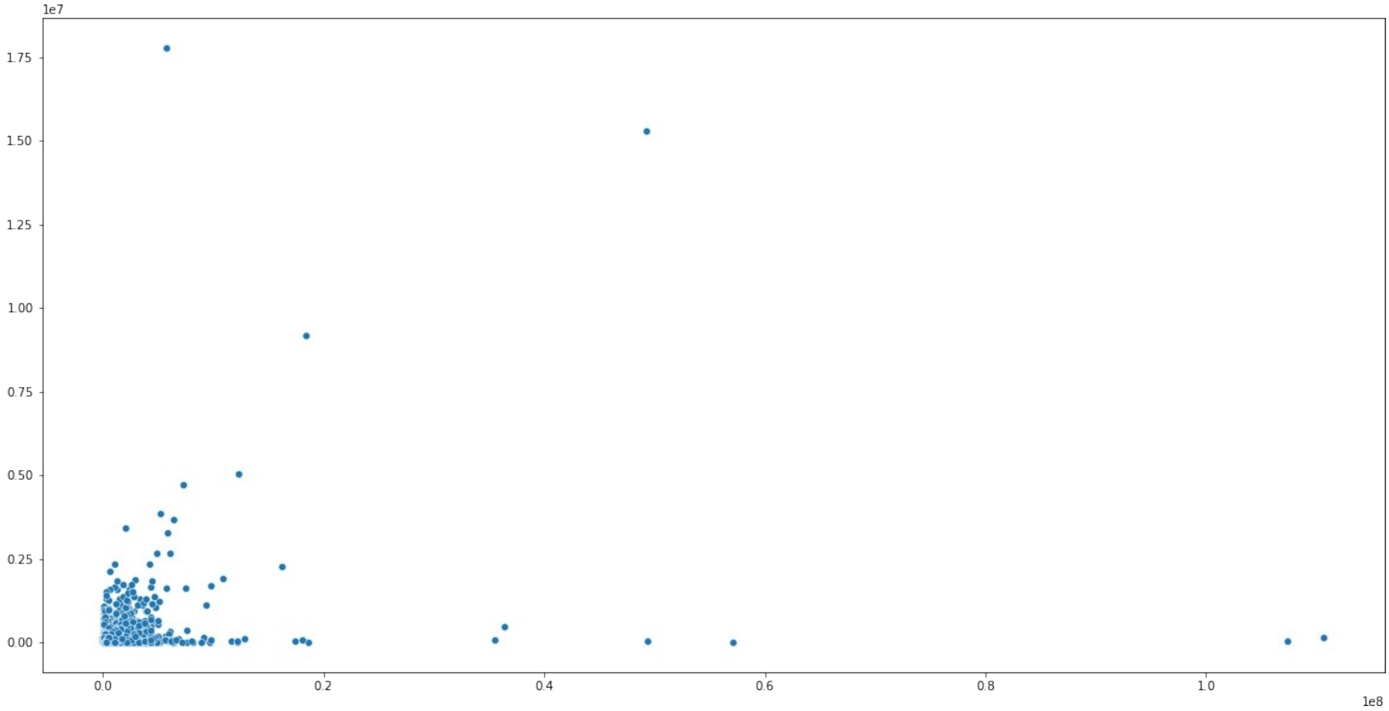
Data was split into training and tests set.
Different regression models were tried on this transformed data. XGBoost regressor produced the best results among the base models.



```
Score for model:  Ridge Regression
CV rmse mean for train set : 369162.796103
CV rmse std deviation for train set: 22096.438643
Score for model:  Lasso Regression
CV rmse mean for train set : 363414.207821
CV rmse std deviation for train set: 24833.943627
Score for model:  ElasticNet Regression
CV rmse mean for train set : 480846.869073
CV rmse std deviation for train set: 36032.075447
Score for model:  SVM Regression
CV rmse mean for train set : 417478.511560
CV rmse std deviation for train set: 38396.510202
Score for model:  Gradient Boosting
CV rmse mean for train set : 310794.359600
CV rmse std deviation for train set: 24972.895985
Score for model:  LightGBM Regression
CV rmse mean for train set : 338714.312360
CV rmse std deviation for train set: 23763.262589
Score for model:  XGBoost Regression
CV rmse mean for train set : 303218.000118
CV rmse std deviation for train set: 16530.558484
Score for model:  Stacked Model
CV rmse mean for train set : 306991.697570
CV rmse std deviation for train set: 18312.123285
```

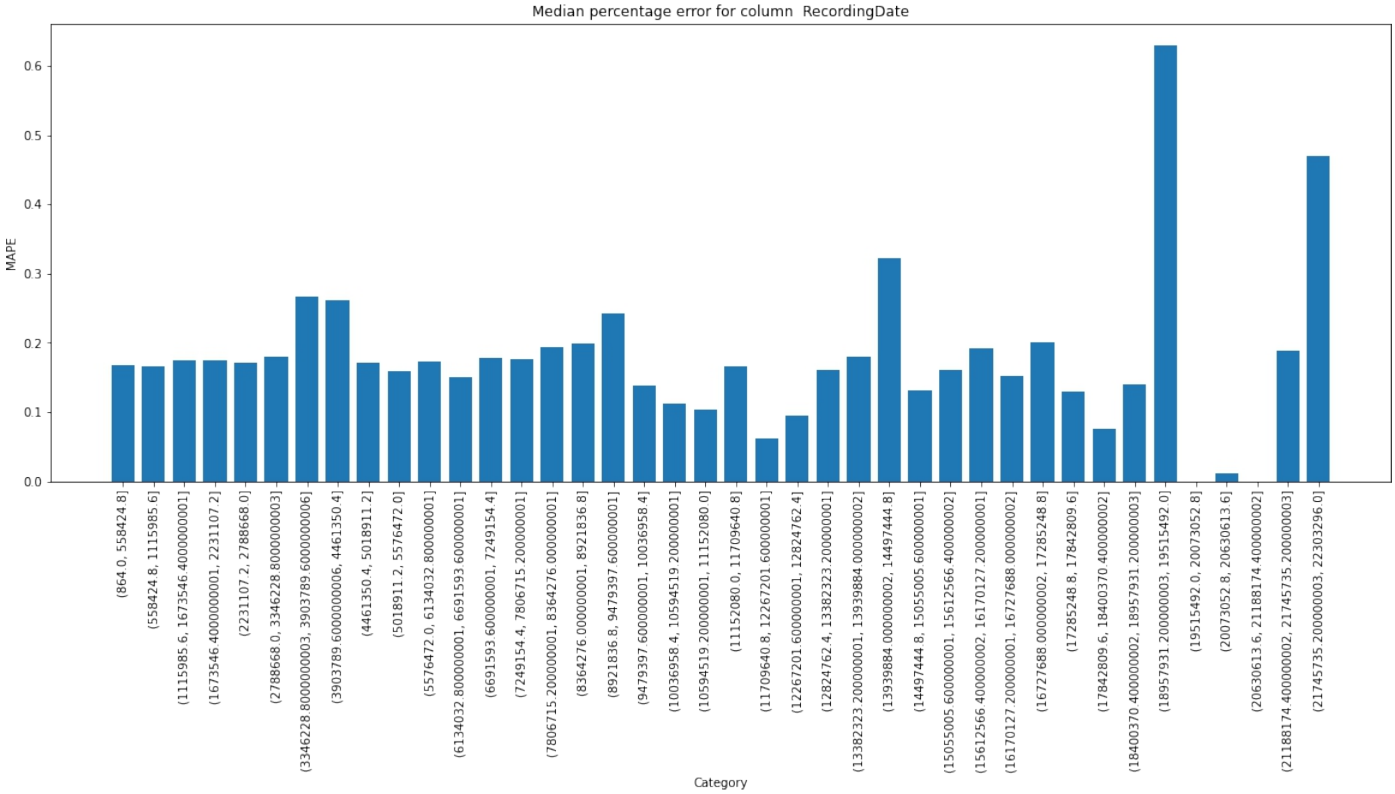
Xgboost was further trained on dataset and the analysis on nature of error was conducted. Individual error analysis -
Absolute Error for individual

data points Error vs TotalValue

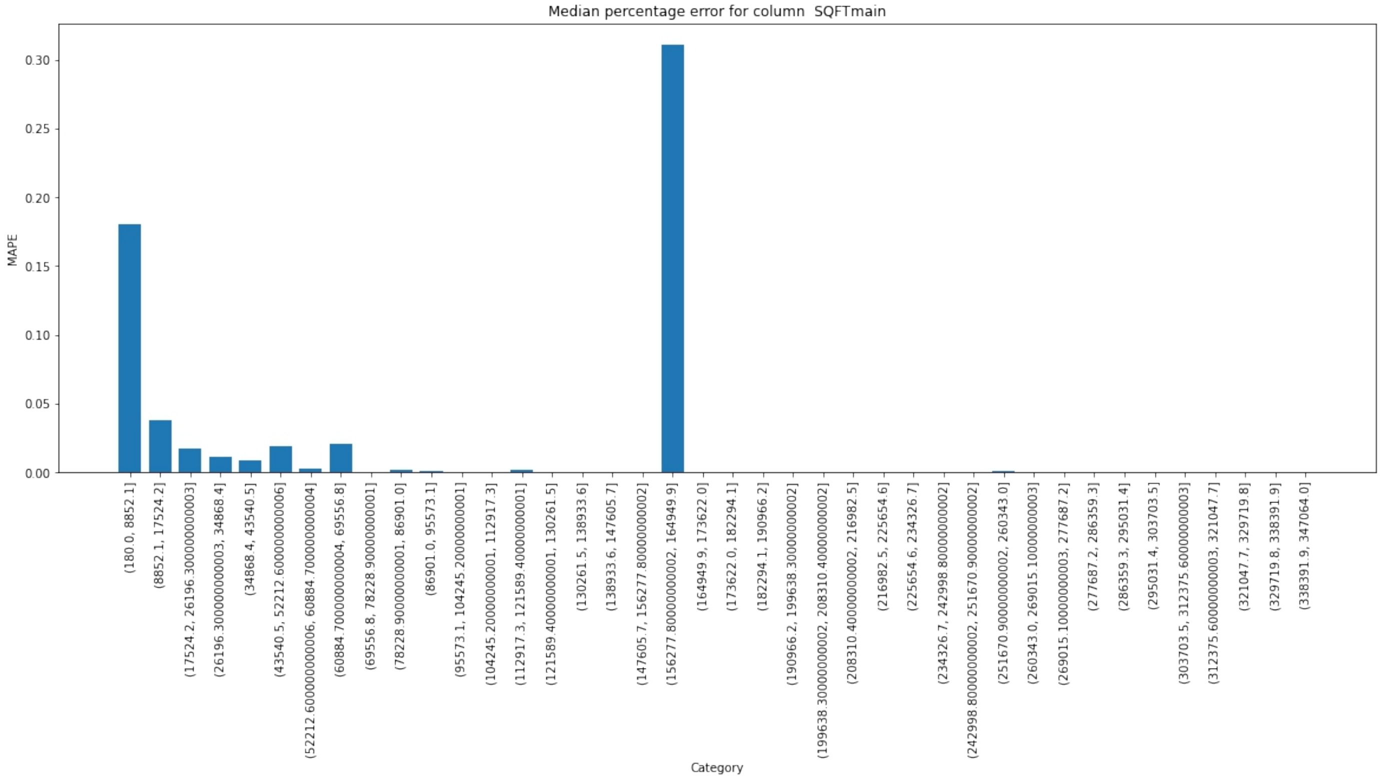


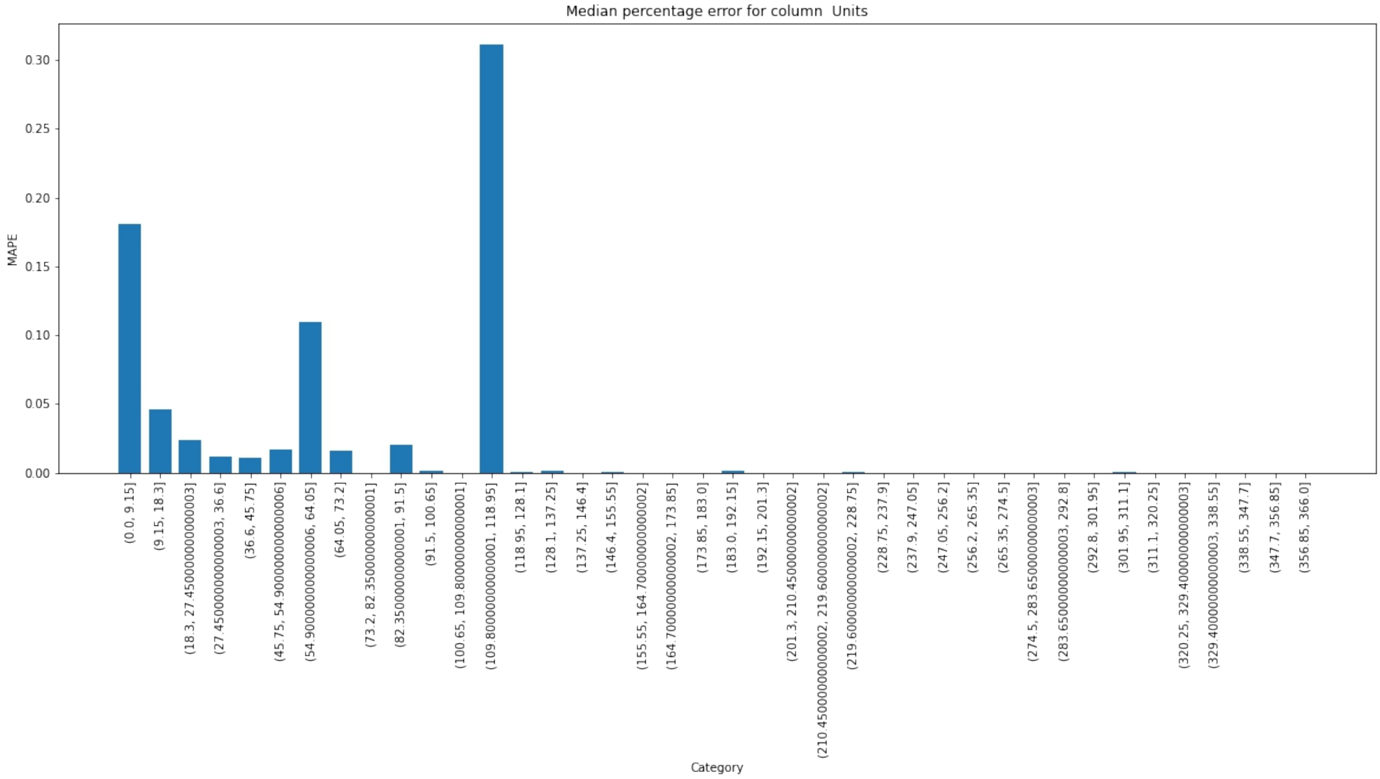
A very small no. of points are causing high error, these need to be

further analyzed. Group error analysis -
Median absolute percentage error was analyzed for different categories in categorical variables and bins in numerical variables Following bar graphs stood out

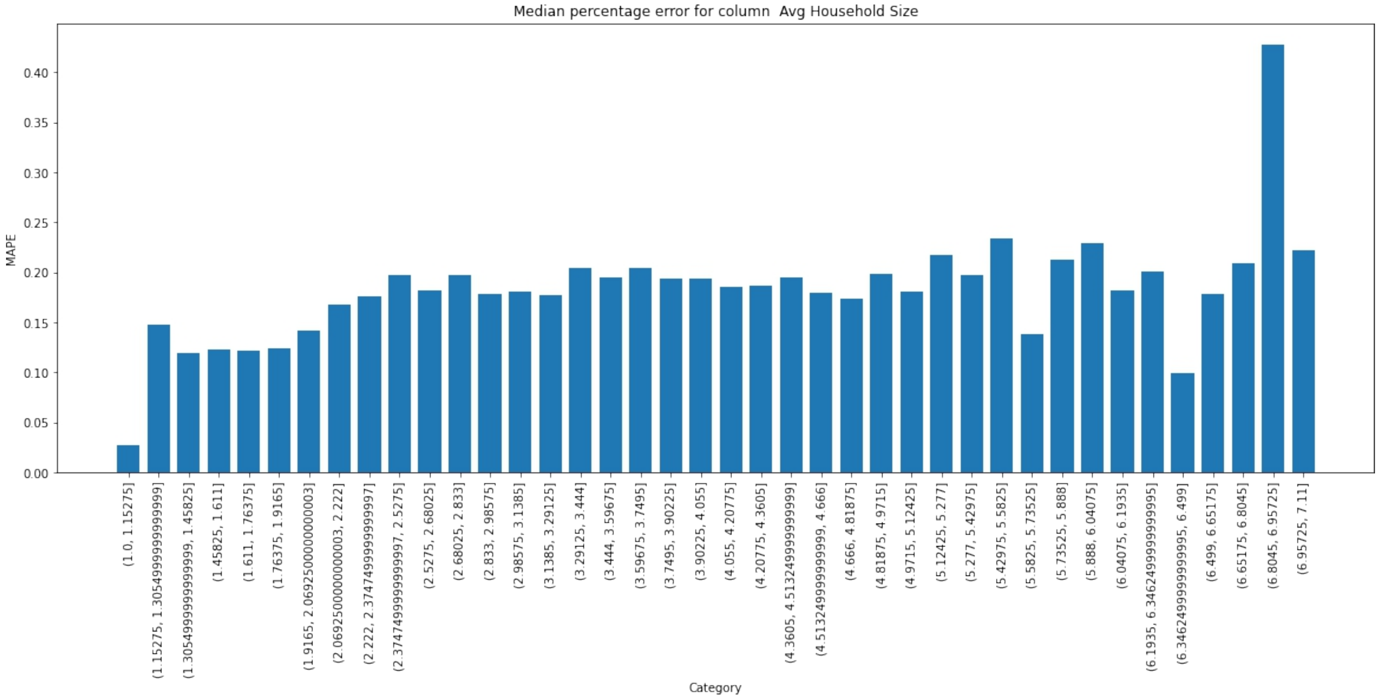


For recording date i.e. transformed difference between 1 Jan 2109 and recording date .It's observed that for bins with larger values median % err is high i.e model is performing bad for older recorded houses
Same trend is seen for houses which are very old (based on year build, effective year build)





Model tends to perform worse for houses with small SQFTmain i.e main area
A large error is also observed for a narrow set of properties having SQFTmain around 16000 sqft This trend is also seen for other signifiers of property size - Bedrooms,Bathrooms,Units
This ties in individual error analysis performed before this step and EDA SQFTmain,Bedrooms,Bathrooms,Units were highly correlated to TotalValue
Therefore large error is seen for properties with low TotalValue and for a narrow set in between Some sort of outlier seems to be present for this small set whose value can't be predicted by the model/or insufficient data to fully describe the property
Possible reasons can be a large property which is very old/ needs a lot of repairing and upkeep or on flip side a property with unique features/location

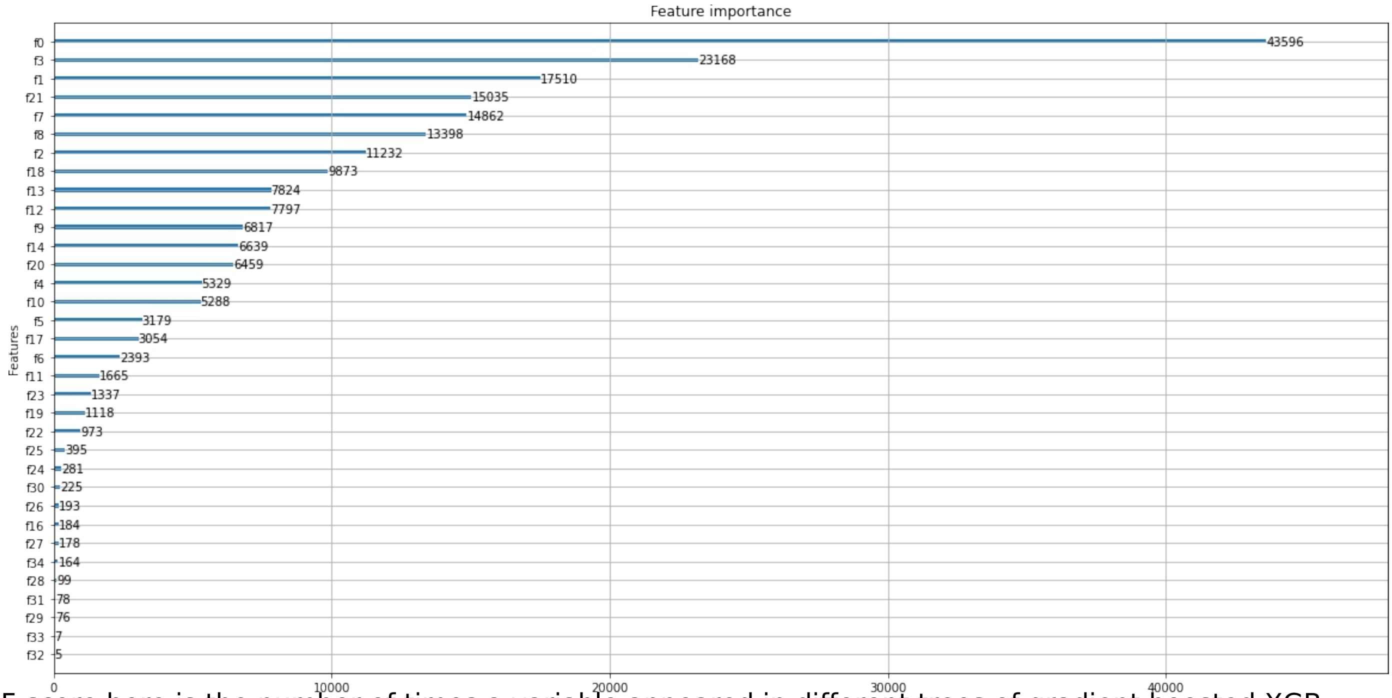


As average size of properties in a block increases model's error % also increases i.e. Properties in areas with larger properties are difficult to asses by model

Large error% for narrow set of Avg. Household size (both occupied and rented), Urban pop seems to confirm presence of some outliers

These outliers could be same as those observed with SQFTmain

To have better insights into the working of model, feature importance was plotted

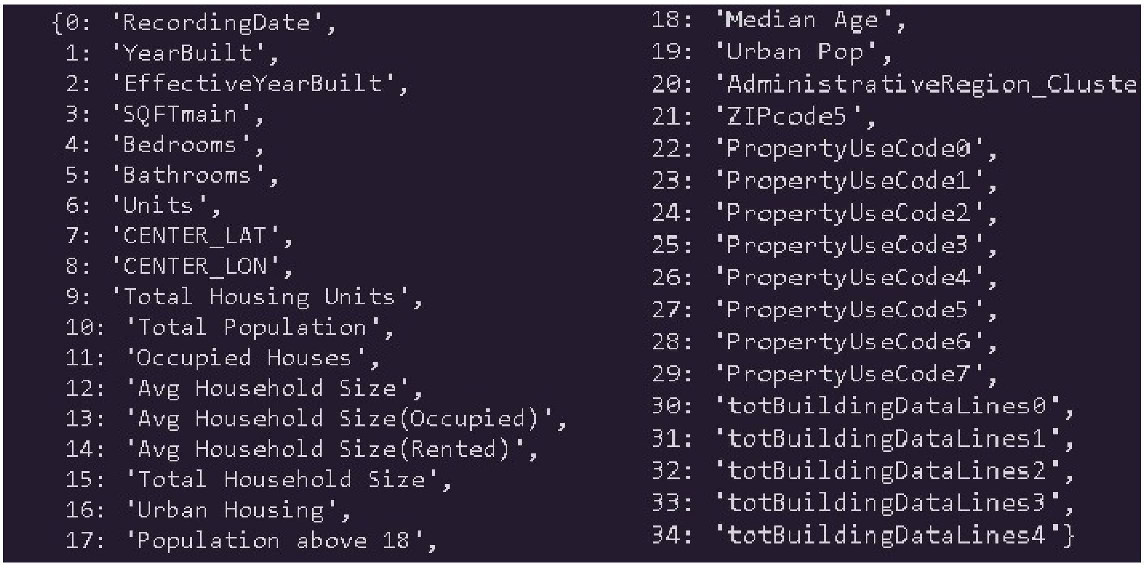


F-score here is the number of times a variable appeared in different trees of gradient boosted XGB

regressor after splitting Note that categorical features are the least important espacially

totBuildingLines
zip code has almost same importance as latitude and
longitude RecordingDate, SQFTmain, EffectiveYearBuilt are
most important features

The dictionary for variable names is provided below

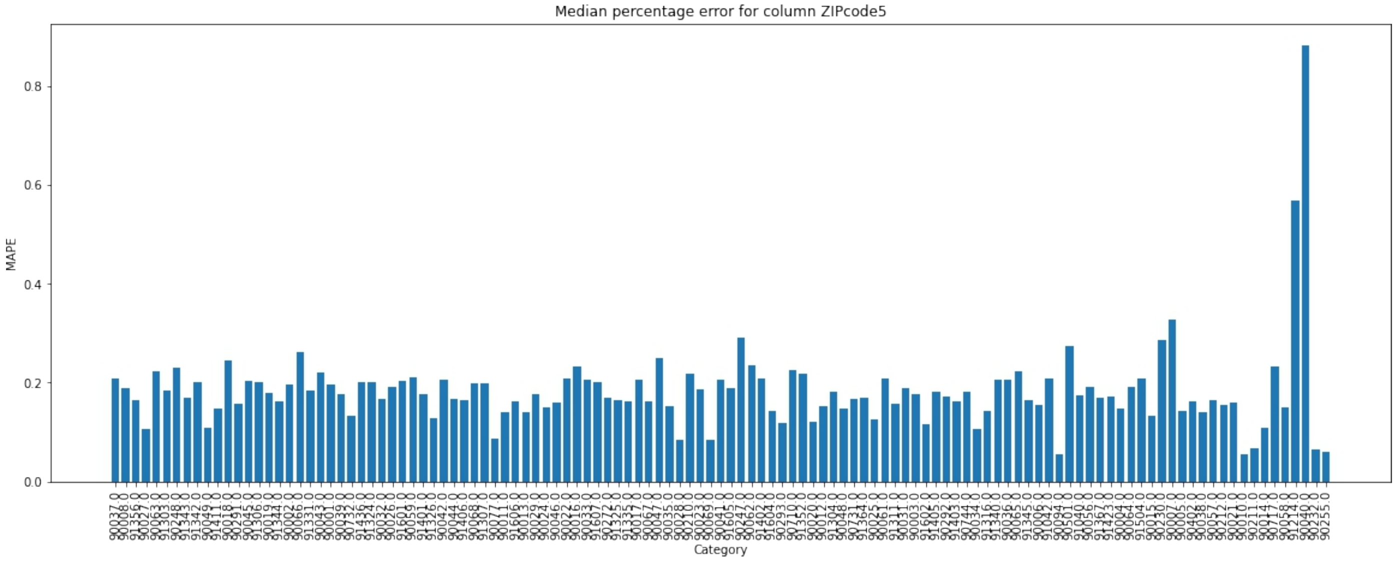


```
{0: 'RecordingDate',                  18: 'Median Age',
 1: 'YearBuilt',                      19: 'Urban Pop',
 2: 'EffectiveYearBuilt',             20: 'AdministrativeRegion_Cluste
 3: 'SQFTmain',                       21: 'ZIPcode5',
 4: 'Bedrooms',                       22: 'PropertyUseCode0',
 5: 'Bathrooms',                      23: 'PropertyUseCode1',
 6: 'Units',                          24: 'PropertyUseCode2',
 7: 'CENTER_LAT',                     25: 'PropertyUseCode3',
 8: 'CENTER_LON',                     26: 'PropertyUseCode4',
 9: 'Total Housing Units',            27: 'PropertyUseCode5',
 10: 'Total Population',              28: 'PropertyUseCode6',
 11: 'Occupied Houses',               29: 'PropertyUseCode7',
 12: 'Avg Household Size',            30: 'totBuildingDataLines0',
 13: 'Avg Household Size(Occupied)',  31: 'totBuildingDataLines1',
 14: 'Avg Household Size(Rented)',    32: 'totBuildingDataLines2',
 15: 'Total Household Size',          33: 'totBuildingDataLines3',
 16: 'Urban Housing',                 34: 'totBuildingDataLines4'}
 17: 'Population above 18',
```

30-34 represent values
[1,2,3,4,5] in order 22-29
represent (in order) -
0100 Single Family
Residence 0101 Single
Family Residence Pool
010C Single Family Residence
Condominium 02xx- 2 units
01xx- 1 units
other 05xx- 5 or
more units
03xx- 3 ur
04xx- 4 units

A set containing outliers was formed. And these outliers were dropped from dataset. The outliers constituted about 0.5% of total observations.

The column ZIPCode5 after some consideration was also dropped from the dataset. The reasoning for it is

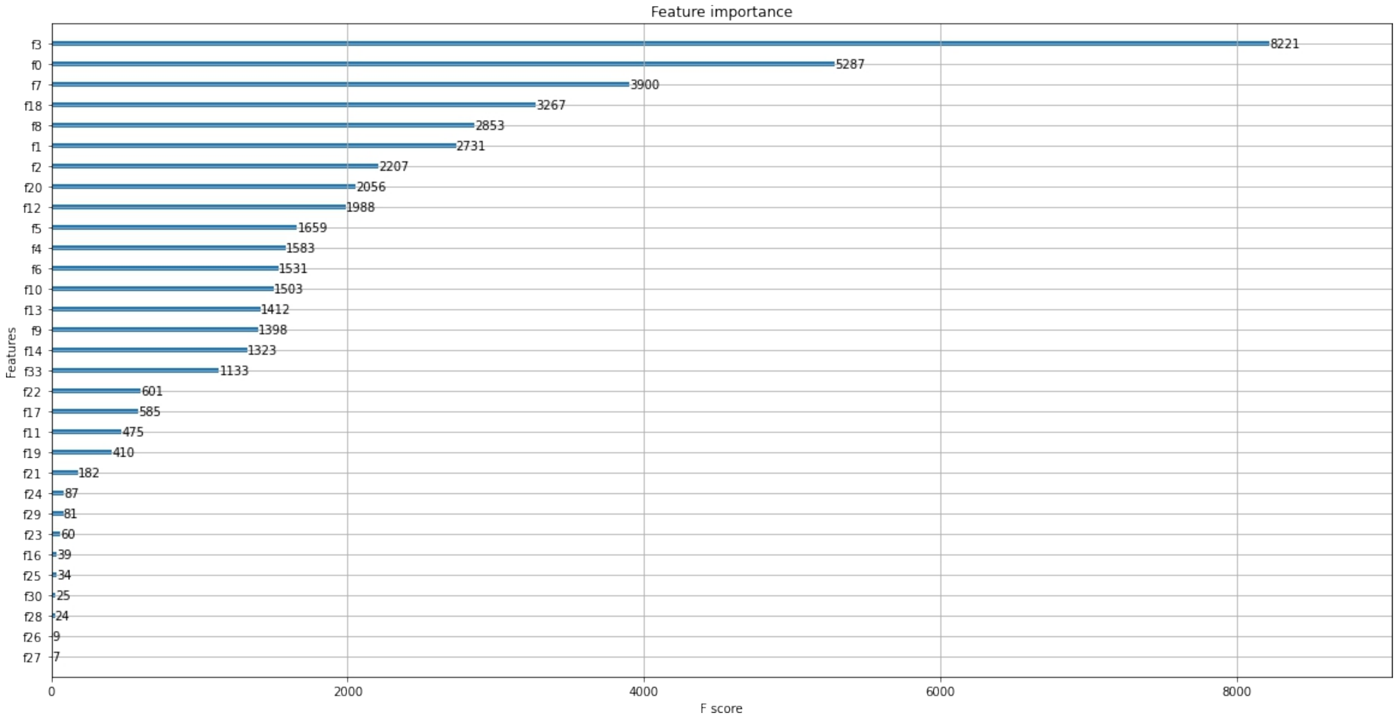


as follows-

- ZIPcode5 has lots of categories and there's no info on grouping these categories together
- Zip code has almost same feature importance as latitude and longitude
- And since we are using forest based models, spatial apace can be split into relevant partitions based on co-ordinate itself

The model was constrained to work on a smaller space and was retrained on complete dataset with K-fold cross validation to produce better results. This time hyper parameter tuning was also performed.

In the final model order of feature importance is changed
SQFTmain is the most important feature now follwed by RecordingDate



```
R2 score for final model: 0.9828916783945469
```

The model’s sensibilities as mentioned in above error analysis still remain. However the results did improve. The median absolute error for the model is around $109k and mean absolute error around $166k

Scope for Improvement

Because the project was conducted in a small time frame, a large scope for improvement of results remain. The following are the key steps that can be taken -

- Dataset can be further enriched using data about income of people and temporal data on property valuation .
- Data about key locations and buildings around the concerned property can be added using geospatial analysis.Examples being- No. Of malls within 1Km radius, Distance to nearest police station, nearest hospital
- More entries and enrichment can also open up the possibility of using Artificial Neural Networks for modelling
- Better EDA can be performed, K means can be used to reduce dimensionality of dataset.
- Missing Values can be better managed - Instead of dropping such values, imputation can be used.
- Ensemble models can be used to make better and more robust predictions.
- Different encoders along with more hyperparameter tuning can be tried to make pipeline better.