greatlearning
Learning for Life

PROJECT INTERIM REPORT

Batch details	PGP-DSE-Feb’22-Chennai
Team members	1. Shreedhar Velmurugan 2. Bhuvanesh Chellapandian 3. Shanjanaa J 4. Sudharsan K 5. Aditya Ram B B
Domain of Project	Real Estate
Proposed project title	Brooklyn Housing Price Prediction
Group Number	Group 3
Team Leader	Aditya Ram B B
Mentor Name	Mrs. Vidhya K

Date: 27-07-2022

Signature of the Mentor

Signature of the Team Leader

TABLE OF CONTENT

PROJECT DETAILS :

OVERVIEW

Overpricing a property can become a real problem for both the buyer and the seller. Most sellers want to maximize profit from the sale of their home and some are also monetarily and emotionally invested in their property. Agents who help in selling/buying any property cannot be completely trusted as their primary goal is to maximize their commission from the property sale deal. Buyers on the other hand should not fall for overpriced properties.

The objective of this project is to help Brooklyn property buyers in property valuation and predict the "near future" price using the Regression model and extrapolation of the fitted model.

Business problem statement (GOALS)

1. BUSINESS UNDERSTANDING:

The real estate market growth in Brooklyn is driven by many factors. Buyers are not sure of how to properly evaluate a property and make a smart purchase. It is a well-known fact that buyers will be doing their research and homework before they decide to buy a property. The real problem faced by many buyers is,

a) On what basis should a property be evaluated.
b) Are the claims about any price driving factor true or not?
c) what will be the future value of a property?

2. BUSINESS OBJECTIVE:

The primary objective is to be able to explain and point out various significant features that drive the price of properties in different neighborhoods of Brooklyn.

The ultimate objective is to come up with a model which will be able to successfully extrapolate the "near future" price of a property.

APPROACH:

- Data Understanding
- Data Pre-Processing
- Exploratory Data Analysis
- Feature Engineering
- Model Building
- Model Evaluation
- Model Optimization

CONCLUSION:

Implementation of this model will help Brooklyn property buyers to evaluate the property they are planning to buy and learn about its predicted future valuation to maximum degree of accuracy that could be obtained by a regression machine learning model.

NOTE: accuracy refers to the ML model's R2 value. (Capturing maximum possible variation with the given set of features)

2. DATA UNDERSTANDING:

2.1 DATA DICTIONARY:

- **Borough**

New York City is divided into five boroughs: Brooklyn, Manhattan, Queens, the Bronx, and Staten Island. Borough number '3' refers to Brooklyn'

- **PLUTO**

PLUTO: Extensive land use and geographic data at the tax lot level in comma–separated values (CSV) file format. The PLUTO files contain more than seventy fields derived from data maintained by city agencies.'''

- **'MAPPLUTO_F' ,'PLUTOMapID'**

A code indicating whether the tax lot is in the PLUTO file, the MapPLUTO file with water areas included, and/or the MapPLUTO file that is clipped to the shoreline.

- **'APPBBL' (APPORTIONMENT BOROUGH, BLOCK, and LOT)**

The originating BBL (borough, block, and lot) from the apportionment before the merge, split, or property conversion to a condominium.

- **TaxMap**

A tax map is a special purpose map, accurately drawn to scale showing all the real property parcels within a city, town, or village. These maps are used to locate parcels and obtain other information required in assessment work. As changes take place in ownership, size, or shape of the parcels, the tax map system must be updated.

- **Sanborn**

The Sanborn Map number is associated with the tax block and lot.

- **Tract2010**

The 2010 census tract in which the tax lot is located.property tax related column

- **BBL**

A concatenation of the borough code, tax block, and tax lot.

- **BoroCode**

contains borough code.

- **CT2010**

The 2010 census tract in which the tax lot is located.

- **CB2010**

The 2010 census block in which the tax lot is located.

- **CD**

The community district (CD) or joint interest area (JIA) for the tax lot. The city is divided into 59 community, districts and 12 joint interest areas, which are large parks or airports that are not considered part of any community district.

- **FireComp**

The fire company that services the tax lot.

- **HealthArea**

The health area in which the tax lot is located.

- **HealthCent**

The health center district in which the tax lot is located. Thirty health center districts were created by the City in 1930 to conduct neighborhood-focused health interventions.

- **ProxCode**

If there are multiple buildings on the lot, CAMA data for building number 1 is used.

Value Description

0 Not available

1 Detached

2 Semi-attached

3 Attached'

- **IrrLotCode**

A code indicating whether the tax lot is irregularly shaped or not

- **LotType**

Value Description

1 Block assemblage – a tax lot that encompasses an entire block

2 Waterfront – a tax lot bordering on a body of water. Waterfront lots may contain a small amount of submerged land.

3 Corner – a tax lot bordering two intersecting streets

4 Through – a tax lot connecting two streets, with frontage on both streets. Note that a lot with two frontages is not necessarily a through a lot. For example, an L-shaped lot with two frontages is considered an inside lot (5).

5 Inside – a tax lot with frontage on only one street. This value comes from CAMA, but is only assigned in PLUTO if CAMA has no other lot types for the tax lot.

6 Interior lot – a tax lot that has no street frontage

7 Island lot – a tax lot that is surrounded by water

8 Alley lot – a tax lot that is too narrow to accommodate a building. The lot is usually 12 feet or less in width.

9 Submerged land lot – a tax lot that is totally or almost completely submerged'"

- **YCoord**

The Y coordinate of the XY coordinate pair depicts the approximate location of the lot.

- **XCoord**

The X coordinate of the XY coordinate pair depicts the approximate location of the lot.

- **AssessLand**

The assessed land value for the tax lot. The Department of Finance calculates the assessed value by multiplying the tax lot's estimated full market land value, determined as if vacant and unimproved, by a uniform percentage for the property's tax class

- **AssessTot**

The Department of Finance calculates the assessed value by multiplying the tax lot's estimated full market value by a uniform percentage for the property's tax class.

- **YearAlter1,YearAlter2**

If a building has only been altered once, YEAR ALTERED 1 is the date that alteration began.

If a building has been altered more than once, YEAR ALTERED 1 is the year of the

The Department of Finance defines alterations as modifications to the structure that,

according to the assessor, change the value of the real property.

The date comes from the Department of Buildings permits and may either be the actual date or an estimate.

- **ExemptTot**

The exempt total value, which is determined differently for each exemption program,

is the dollar amount related to that portion of the tax lot that has received an

exemption.

- **Owner Name**

Contains the owner's name

- **PolicePrct**

The police precinct in which the tax lot is located. This field contains a three-digit police precinct number which is preceded with leading zeros if the precinct number has less than three digits.

- **building_class**

Building class during construction

- **SanitDistr**

The sanitation district that services the tax lot.

- **SanitSub**

The subsection of the sanitation district that services the tax lot

- **LandUse**

A code for the tax lot's land use category.

VALUE DESCRIPTION

01 One & Two Family Buildings

02 Multi-Family Walk-Up Buildings

03 Multi-Family Elevator Buildings

04 Mixed Residential & Commercial Buildings

05 Commercial & Office Buildings

06 Industrial & Manufacturing

07 Transportation & Utility

08 Public Facilities & Institutions

09 Open Space & Outdoor Recreation

10 Parking Facilities

11 Vacant Land

- **sale_date**

The actual date on which the sale took happened

- **ExemptTot, ExemptLand**

The exempt total value, which is determined differently for each exemption program, is the dollar amount related to that portion of the tax lot that has received an exemption.

- **Easements**

An easement is a nonpossessory right to use and/or enter onto the real property of another without possessing it.

2.2 <u>VARIABLE CATEGORIZATION :</u>

Independent variables:

Numerical column: 44

Categorical column: 65

Null columns: 0

Target variable:

Numerical column - 1

Total columns: 111

2.3 <u>DISTRIBUTION OF VARIABLES:</u>

There are 44 numerical variables and 65 categorical variables inclusive of the target variable in the dataset. It is observed some of the numerical variables are not normally distributed and most of the categorical features are having heavy data imbalance. The target variable is also skewed.

2.4 REDUNDANT FEATURES :

1. All records belong to Brooklyn, so both 'Borough' and 'Borough' columns can be dropped.
2. 'Unnamed: 0' is a unique row identifier. It can be dropped.
3. 'Version' refers to the PLUTOmap version. It can be dropped.
4. 'PLUTOMapID' refers to the map ID. It can be dropped
5. 'MAPPLUTO_F' refers to the PLUTO file. It can be dropped.
6. 'APPBBL' is an insignificant feature. It can be dropped.
7. 'TaxMap' can be dropped - it is a categorical feature with 200 subcategories.
8. 'Sanborn' feature is insignificant, it can be dropped.
9. 'Tract2010' feature is insignificant, it can be dropped.
10. 'BBL' feature is insignificant, it can be dropped.
11. 'BoroCode' feature is insignificant, it can be dropped.
12. 'CT2010' feature is insignificant, it can be dropped.
13. 'CB2010' feature is insignificant, it can be dropped.
14. 'FireComp' feature is insignificant, it can be dropped.
15. 'HealthArea' feature is insignificant, it can be dropped.
16. 'HealthCent' feature is insignificant, it can be dropped.
17. 'YCoord', 'XCoord' : we do not need specific plot locations. Both features can be dropped
18. 'OwnerName'. This column can be dropped as it does not hold any significance
19. 'PolicePrct' feature is insignificant, it can be dropped. (it has got nothing to do with property price)
20. 'building_class' can be dropped because building class at the sale is more important.
21. 'SanitDistr' feature is insignificant, it can be dropped.
22. 'SanitSub' feature is insignificant, it can be dropped.
23. 'ZoneDist1': since other Zones are dropped due to a high percentage of null values, this can also be dropped
24. 'sale_date' feature is too specific.It can be dropped.
25. 'ExemptTot', 'ExemptLand'. These two features do not influence the price of a building.
26. government decides the exemption and revises it twice a year
27. 'Easements'. All the values are zero in this feature. It can be dropped.

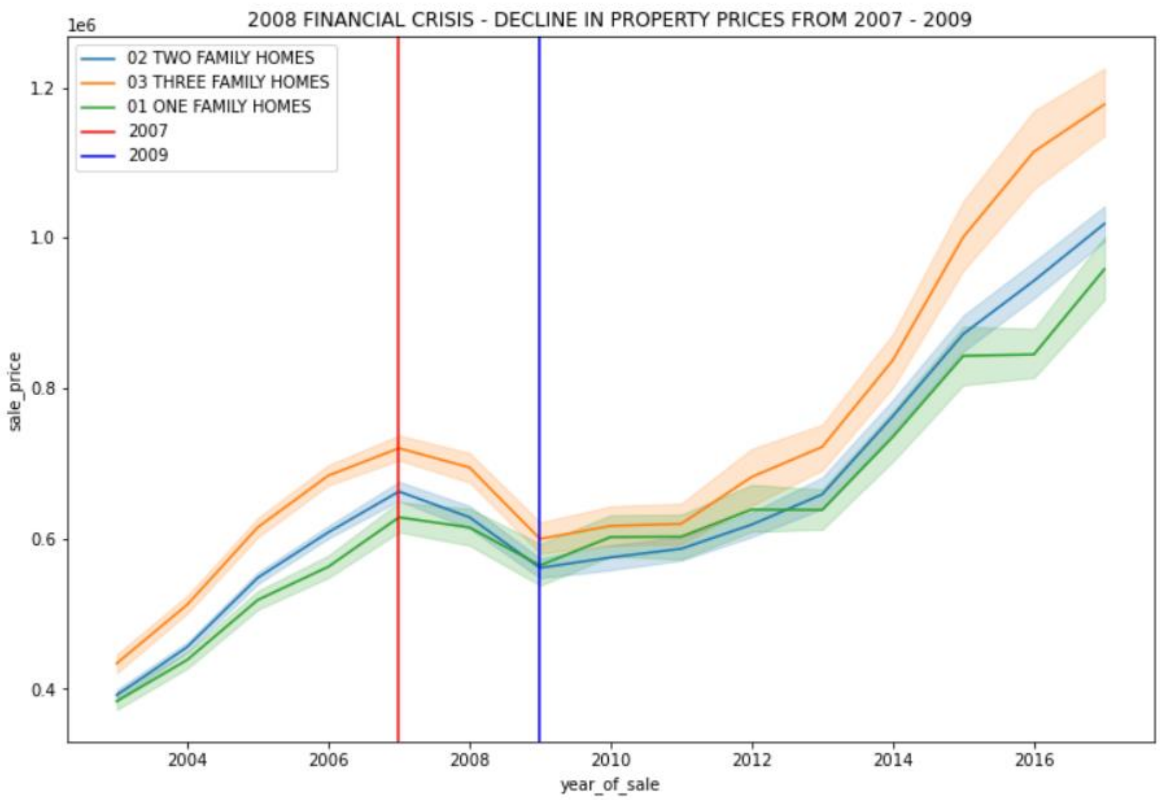
2008 Financial crisis

The financial crisis of 2008 created the biggest disruption to the U.S. housing market since the Great Depression

We can divide the real estate market trend into two categories. Before 2008 and after 2008

If we consider the whole data from 2004 to 2016 for training (2017 as a test set), our model performance will be greatly affected because we are trying to capture two different trends. Things have drastically changed after the 2008 financial crisis.

To solve this, we will be considering the sales that took place after 2008. This will lead to the concept of trying to build a model that will capture the new trend that was formed after 2008.



It can be seen that the price of properties has declined during the period 2007 - 2009.

3. DATA PREPROCESSING:

3.1 NULL VALUE TREATMENT:

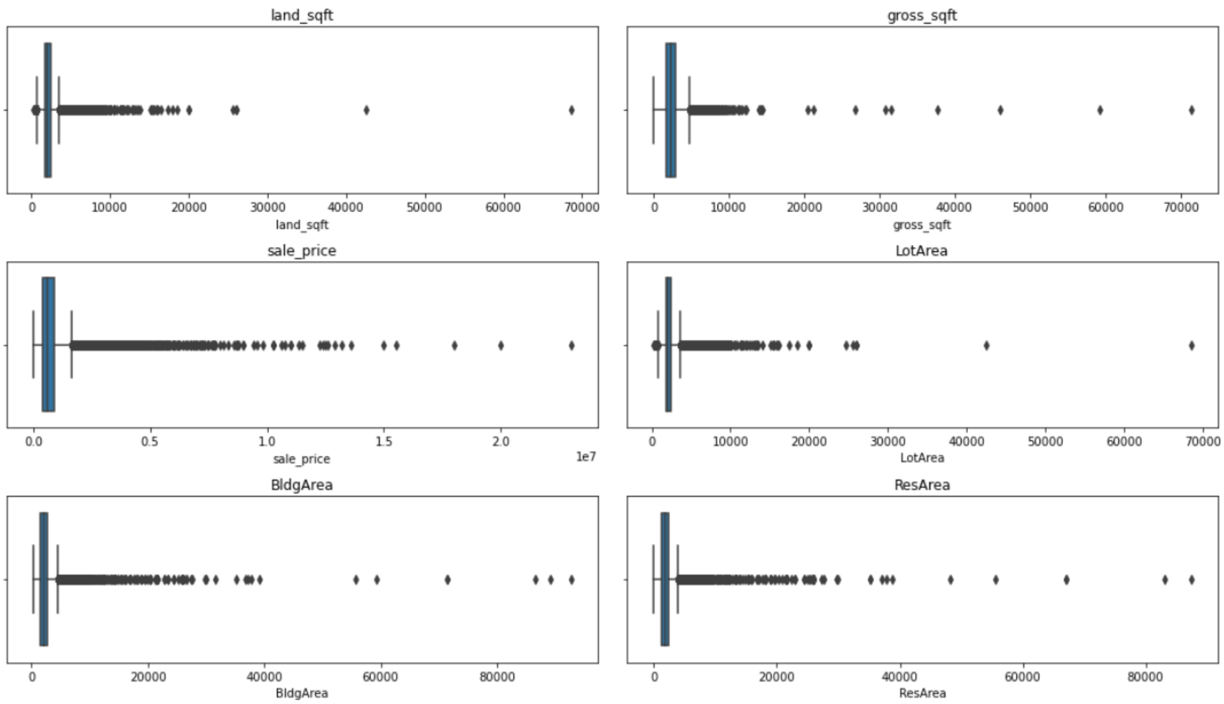
Null value treatment is essential to building most of the commonly used machine learning models such as linear regression, decision tree, KNN, and others. In that context, the percentage of missing values in the dataset was calculated and the same is shown below.

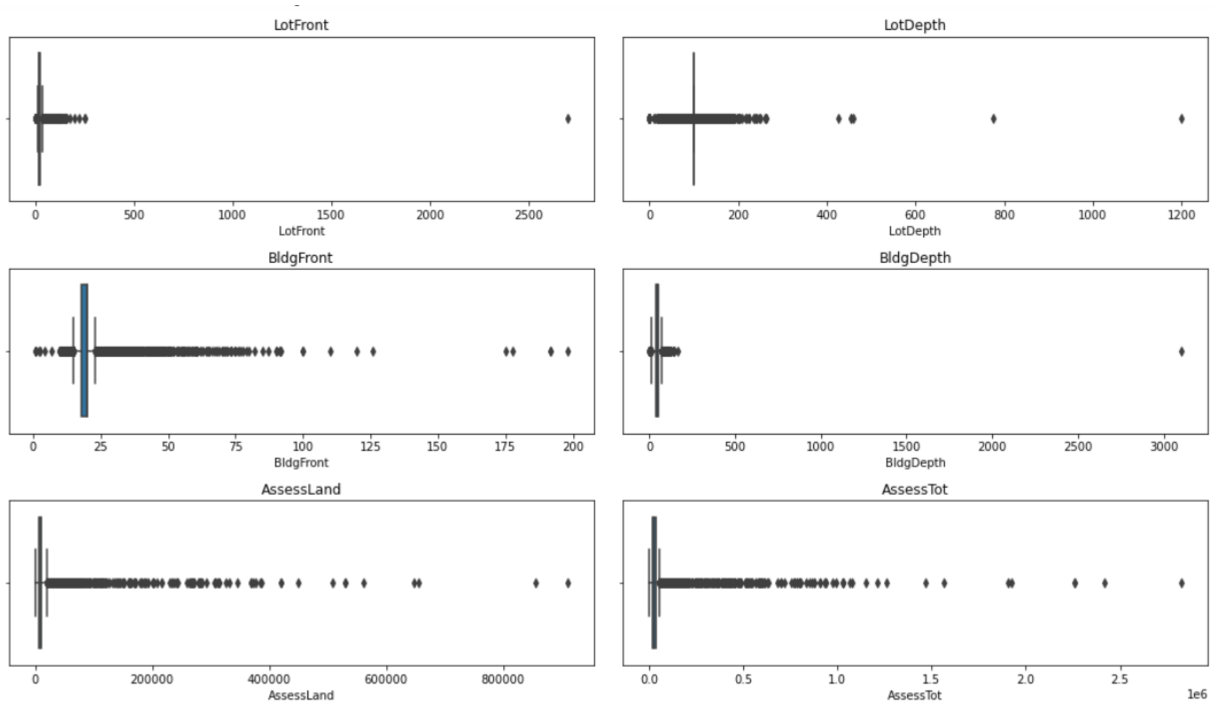
	columns	Null count	Percentage of Null values
0	Unnamed: 0	0	0.000000
1	borough	0	0.000000
2	neighborhood	0	0.000000
3	building_class_category	83	0.021234
4	tax_class	6934	1.773932
5	block	0	0.000000
6	lot	0	0.000000
7	easement	390883	100.000000
8	building_class	6934	1.773932
9	address	1	0.000256
10	apartment_number	305267	78.096771
11	zip_code	0	0.000000
12	residential_units	0	0.000000
13	commercial_units	0	0.000000
14	total_units	0	0.000000
15	land_sqft	0	0.000000
16	gross_sqft	0	0.000000
17	year_built	0	0.000000
18	tax_class_at_sale	0	0.000000
19	building_class_at_sale	0	0.000000
20	sale_price	0	0.000000
21	sale_date	0	0.000000
22	year_of_sale	0	0.000000
23	Borough	87155	22.296953
24	CD	87155	22.296953
25	CT2010	87447	22.371656
26	CB2010	88368	22.607276
27	SchoolDist	87195	22.307187
28	Council	87155	22.296953
29	ZipCode	87155	22.296953
30	FireComp	87403	22.360399
31	PolicePrct	87155	22.296953
32	HealthCent	87155	22.296953
33	HealthArea	87155	22.296953
34	SanitBoro	87645	22.422311
35	SanitDistr	87645	22.422311
36	SanitSub	87931	22.495478
37	Address	87178	22.302837
38	ZoneDist1	87169	22.300535
39	ZoneDist2	375768	96.133114
40	ZoneDist3	390697	99.952415
41	ZoneDist4	390880	99.999233
42	Overlay1	348962	89.275307
43	Overlay2	390835	99.987720
44	SPDist1	355383	90.917998
45	SPDist2	390859	99.993860
46	SPDist3	390879	99.998977
47	LtdHeight	385762	98.689889
48	SplitZone	87177	22.302582
49	BldgClass	87177	22.302582
50	LandUse	88172	22.557133
51	Easements	87155	22.296953
52	OwnerType	337389	86.314575
53	OwnerName	87259	22.323560
54	LotArea	87155	22.296953

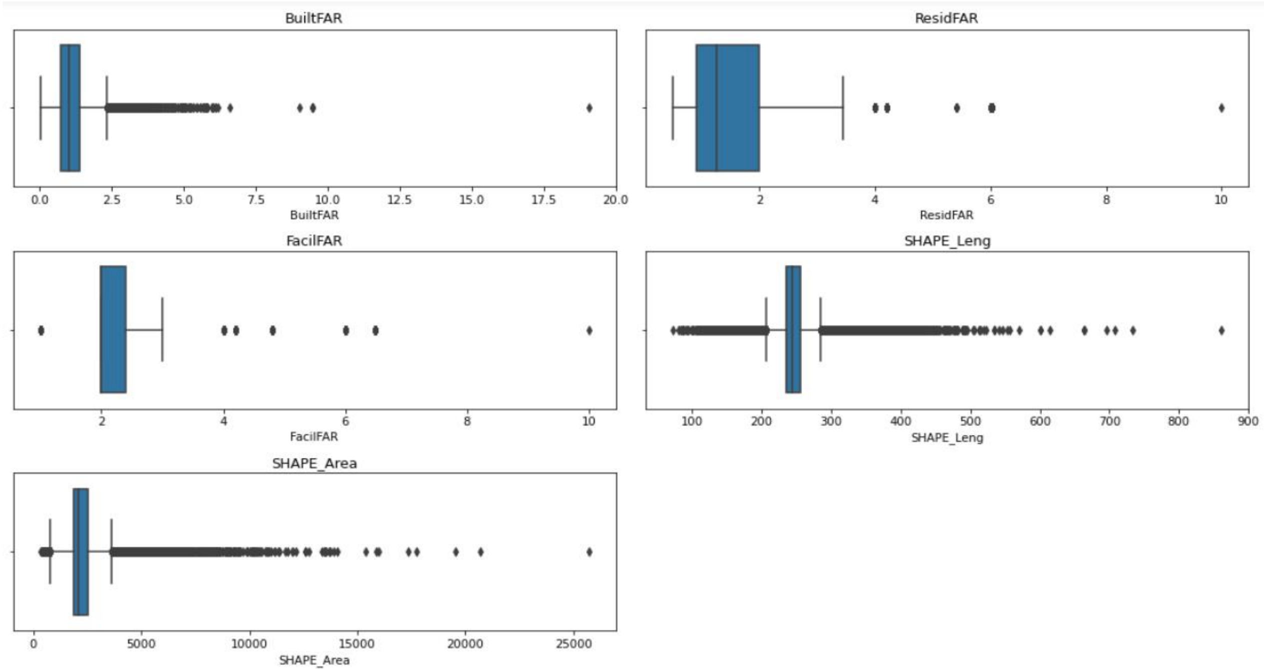
55	BldgArea	87155	22.296953
56	ComArea	87155	22.296953
57	ResArea	87155	22.296953
58	OfficeArea	87155	22.296953
59	RetailArea	87155	22.296953
60	GarageArea	87155	22.296953
61	StrgeArea	87155	22.296953
62	FactryArea	87155	22.296953
63	OtherArea	87155	22.296953
64	AreaSource	87155	22.296953
65	NumBldgs	87155	22.296953
66	NumFloors	87155	22.296953
67	UnitsRes	87155	22.296953
68	UnitsTotal	87155	22.296953
69	LotFront	87155	22.296953
70	LotDepth	87155	22.296953
71	BldgFront	87155	22.296953
72	BldgDepth	87155	22.296953
73	Ext	324750	83.081127
74	ProxCode	87177	22.302582
75	IrrLotCode	87177	22.302582
76	LotType	87177	22.302582
77	BsmtCode	87177	22.302582
78	AssessLand	87155	22.296953
79	AssessTot	87155	22.296953
80	ExemptLand	87155	22.296953
81	ExemptTot	87155	22.296953
82	YearBuilt	87155	22.296953
83	YearAlter1	87155	22.296953
84	YearAlter2	87155	22.296953
85	HistDist	371209	94.966780
86	Landmark	390757	99.967765
87	BuiltFAR	87155	22.296953
88	ResidFAR	87155	22.296953
89	CommFAR	87155	22.296953
90	FacilFAR	87155	22.296953
91	BoroCode	87155	22.296953
92	BBL	87155	22.296953
93	CondoNo	87155	22.296953
94	Tract2010	87155	22.296953
95	XCoord	87155	22.296953
96	YCoord	87155	22.296953
97	ZoneMap	87155	22.296953
98	ZMCode	384771	98.436361
99	Sanborn	87173	22.301558
100	TaxMap	87173	22.301558
101	EDesigNum	387329	99.090777
102	APPBBL	87155	22.296953
103	APPDate	371624	95.072950
104	PLUTOMapID	87155	22.296953
105	FIRM07_FLA	382230	97.786294
106	PFIRM15_FL	363110	92.894805
107	Version	87155	22.296953
108	MAPPLUTO_F	87155	22.296953
109	SHAPE_Leng	87155	22.296953
110	SHAPE_Area	87155	22.296953

From the above figure, it is evident that in 20 features a maximum percentage of missing value is above **70%**. This means these features can be removed from the data frame.

PRESENCE OF OUTLIERS AND TREATMENT: (17 numerical features)







'land_sqft': There are 5 values above 23000.

'gross_sqft': There are 6 values above 30000.

'sale_price': There are 5 values above 15000000.

'LotArea': There are 6 values above 21000.

'BldgArea': There are 7 values above 50000.

'ResArea': There are 6 values above 40000.

'LotFront': There are 4 values above 200.

'LotDepth': There are 6 values above 270.

'BldgFront': There are 5 values above 126.

'BldgDepth': There is 1 value above 165.

'AssessLand': There are 8 values above 509000.

'AssessTot': There are 6 values above 1600000.

'BuiltFAR': There are 6 values above 7.

'ResidFAR': There is 1 value above 6.2.

'FacilFAR': There is 1 value above 6.5

'SHAPE_Leng': There is 1 value above 800.

'SHAPE_Area': There are 8 values above 15000.

SKEW TABLE

	numerical_feature	original_skew	percentage_loss (0.99)	new skewness 0.99	sqaure_root_skewness	Log skew
0	land_sqft	9.62	0.6933	1.83	9.55	0.73
1	gross_sqft	12.04	0.9750	0.52	4.58	-0.38
2	sale_price	5.71	1.0003	1.94	3.71	-1.61
3	LotArea	9.60	0.7204	1.83	4.62	0.73
4	BldgArea	21.73	0.9931	0.70	3.52	0.12
5	ResArea	21.59	1.0003	0.86	3.52	0.33
6	LotFront	105.87	0.8540	2.14	3.29	1.64
7	LotDepth	6.44	0.9858	-1.88	4.04	-2.86
8	BldgFront	8.59	0.9768	1.27	2.24	0.94
9	BldgDepth	106.64	0.8450	0.27	0.82	-0.53
10	AssessLand	26.08	1.0003	1.19	5.61	-0.08
11	AssessTot	27.63	1.0003	1.18	3.66	0.03
12	BuiltFAR	2.56	1.0003	0.75	1.11	-0.16
13	ResidFAR	0.99	0.3214	0.67	-0.29	-0.12
14	FacilFAR	1.07	0.2401	1.05	0.26	0.31
15	SHAPE_Leng	0.90	0.9985	-0.49	3.92	-0.92
16	SHAPE_Area	3.55	1.0003	1.70	4.51	0.75

It is very evident from the above table that outliers are affecting the skew and have to be removed. Even square root transformation is not enough to reduce the skewness.

Even after removing outliers and doing square root transformation, we are not able to reduce the skewness of all features to under 1.5.

From the above table, we can conclude that log transformation is good for all features except 'LotDepth'.

FEATURE ENGINEERING:

List of new features that were created:

- Two features 'YearAlter1' and 'YearAlter2' were converted into a single feature.
- Using 'year_built' and 'year_of_sale' features, the age feature was created.
- All years in the year category were replaced with numbers in ascending order.
- All data points in the 'SchoolDist' column were replaced with alphabets.
- All data points in the 'LandUse' column were replaced with their respective details.
- All data points in the 'ProxCode' column were replaced with their respective details.
- All data points in the 'LotType' column were replaced with their respective details.
- All data points in the 'BsmtCode' column were replaced with their respective details.

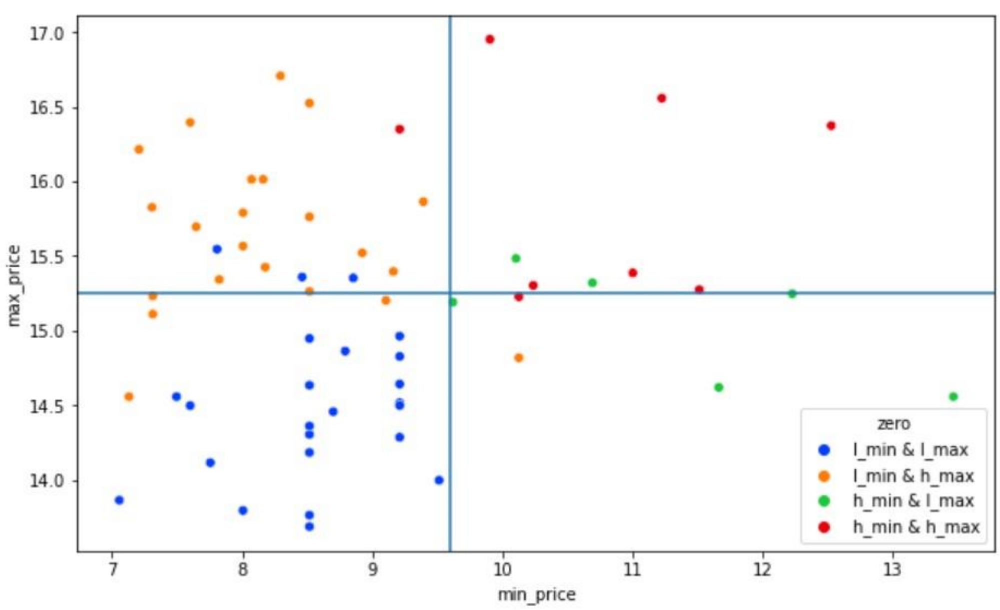
List of new features created using clustering Algorithm:

To reduce the number of columns after encoding, we have to take care of the subcategories in features that have a large number of subcategories. A single numerical feature was used to create 5 numerical features and clustering algorithm was applied on them.

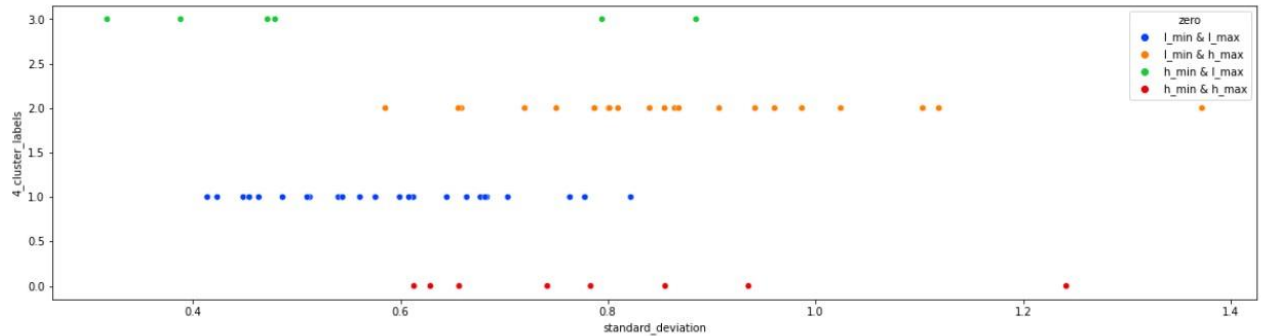
- 'neighbor_clusters'
- 'bclass_clusters'
- 'landuse_group'
- 'school'
- 'Lot'

These features contain cluster groups of their respective original features.

Neighborhood cluster:



Neighborhood price deviation:



Inference from the above clusters

'''Low minimum price and low maximum price'''

'l_min & l_max': They were low priced in the past and still they are at a low price range (low deviation)

'''Low minimum price and high maximum price'''

'l_min & h_max': They were low priced in the past but later they moved to a premium price range (good deviation) (good growth)

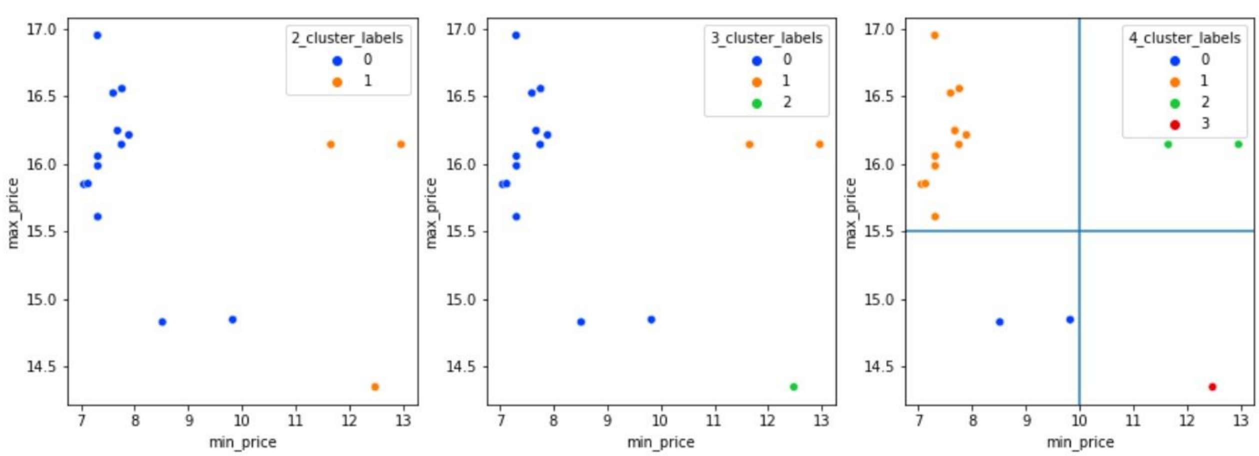
'''High minimum price and low maximum price '''

'h_min & l_max': They were premium priced in the past and but they did not experience any relative growth (low deviation)

'''High minimum price and high maximum price'''

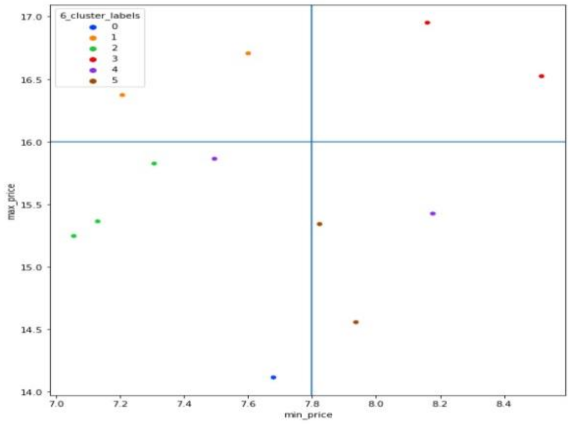
'h_min & h_max': They were premium priced in the past and still they are maintaining their premium level (good deviation)

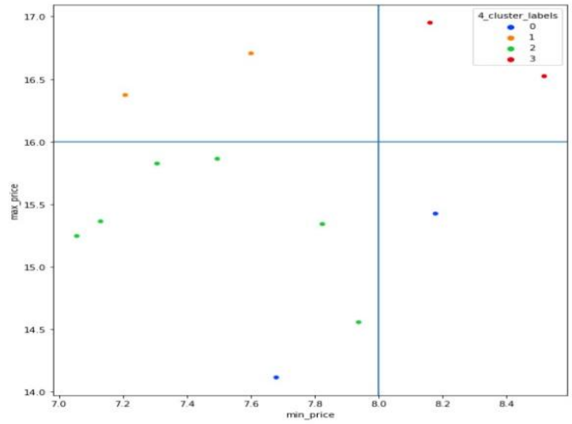
Business_class during sale cluster:



Using 4 clusters gives good data separation and data grouping for the Business_class_during_sale feature.

SchoolDist cluster:



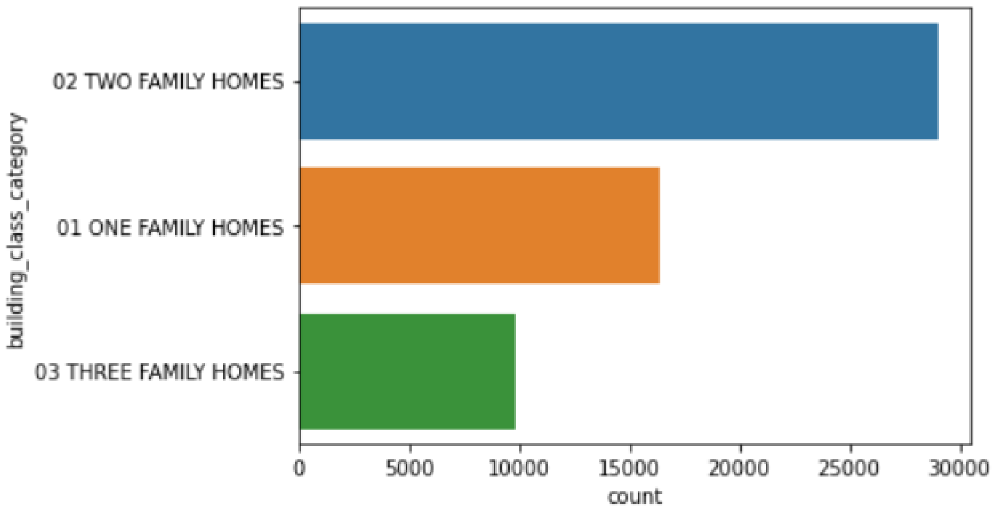


For the SchoolDist column using 4 clusters gives good data separation and data grouping.

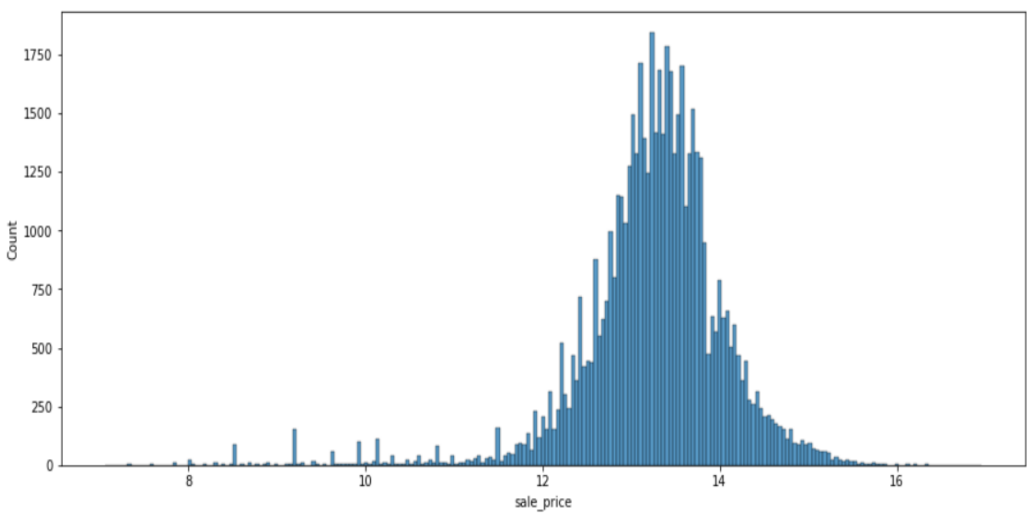
	SchoolDist	standard_deviation	mean	min_price	max_price	skewness	2_cluster_labels	3_cluster_labels	4_cluster_labels	6_cluster_labels
0	a	1.103454	13.667628	7.207860	16.372738	-1.602857	1	2	1	1
1	b	1.005050	13.633908	7.600902	16.705882	-2.142633	1	2	1	1
2	c	0.852184	13.887611	8.160518	16.951005	-1.425340	1	0	3	3
3	d	0.940668	13.042936	7.130899	15.363073	-2.083999	0	1	2	2
4	e	0.971854	13.176709	7.307202	15.825537	-2.254260	0	1	2	2
5	f	0.639898	12.801647	7.679714	14.115615	-3.368829	0	1	0	0
6	g	0.773303	12.672582	7.056175	15.246220	-2.718886	0	1	2	2
7	h	0.564925	13.547395	8.177516	15.424948	-2.774532	0	1	0	4
8	i	0.640679	13.453727	8.517193	16.523561	-1.092003	1	0	3	3
9	j	0.620781	13.195562	7.495542	15.863203	-1.963140	0	1	2	4
10	k	0.871614	12.749243	7.937375	14.557448	-2.241842	0	1	2	5
11	l	0.858048	13.038240	7.824046	15.341567	-2.146495	0	1	2	5

	building_class_at_sale	standard_deviation	mean	min_price	max_price	skewness	2_cluster_labels	3_cluster_labels	4_cluster_labels
0	A0	0.531143	13.161520	12.476100	14.346139	1.678087	1	2	3
1	A1	0.756700	13.375124	7.679714	16.245609	-1.187997	0	0	1
2	A2	0.741987	12.874582	8.517193	14.827111	-1.236087	0	0	0
3	A3	0.644920	14.198084	11.652687	16.142788	0.331994	1	1	2
4	A4	1.083471	13.733353	7.892826	16.213406	-1.269867	0	0	1
5	A5	0.665585	13.079633	7.755339	16.142788	-1.221634	0	0	1
6	A7	0.995355	14.836194	12.959844	16.143763	-0.916539	1	1	2
7	A9	0.690531	13.098808	7.763021	16.556351	-1.050371	0	0	1
8	B1	0.783142	13.225271	7.056175	15.847598	-1.819171	0	0	1
9	B2	0.828595	13.142549	7.313220	15.607270	-2.189418	0	0	1
10	B3	0.879839	13.251154	7.130899	15.852175	-1.590324	0	0	1
11	B9	0.919411	13.359174	7.313220	16.056220	-2.086756	0	0	1
12	C0	0.927948	13.326311	7.600902	16.523561	-1.854213	0	0	1
13	S0	1.018425	13.264751	9.825526	14.845130	-1.862462	0	0	0
14	S1	0.836369	13.262758	7.313220	15.984564	-1.647810	0	0	1
15	S2	0.966424	13.413737	7.307202	16.951005	-1.816031	0	0	1

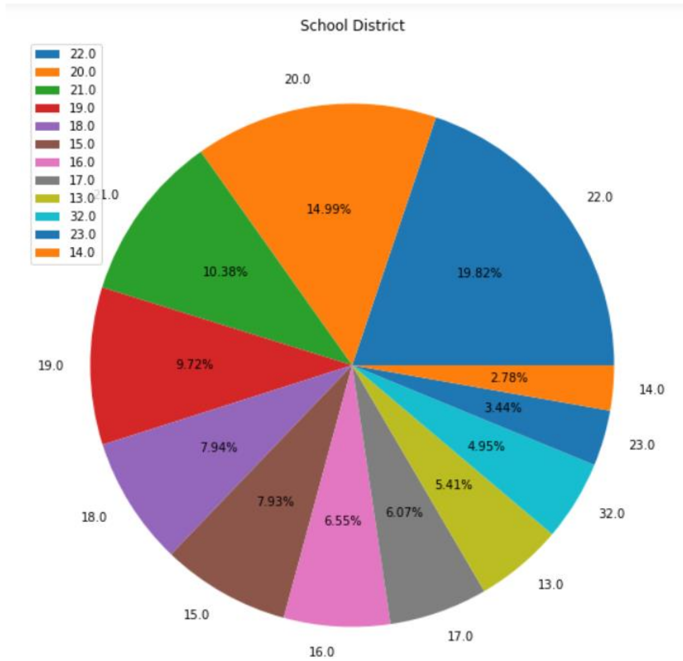
EXPLORATORY DATA ANALYSIS:



Distribution of three building class categories

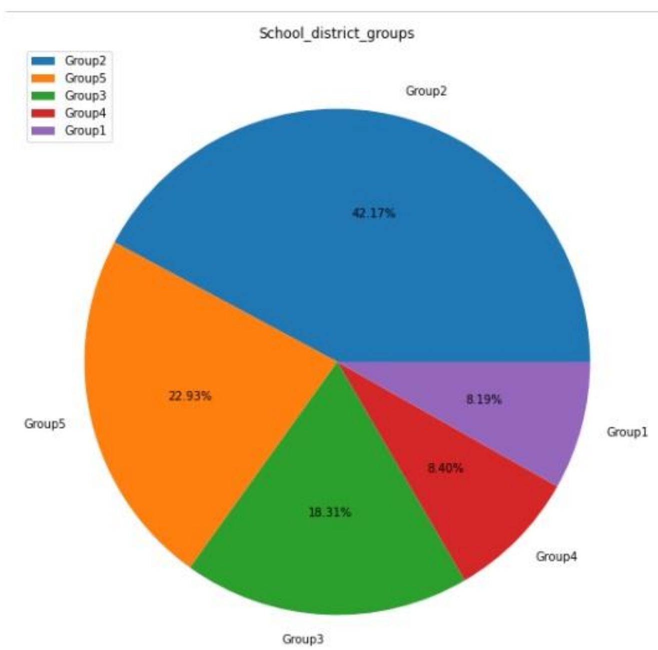


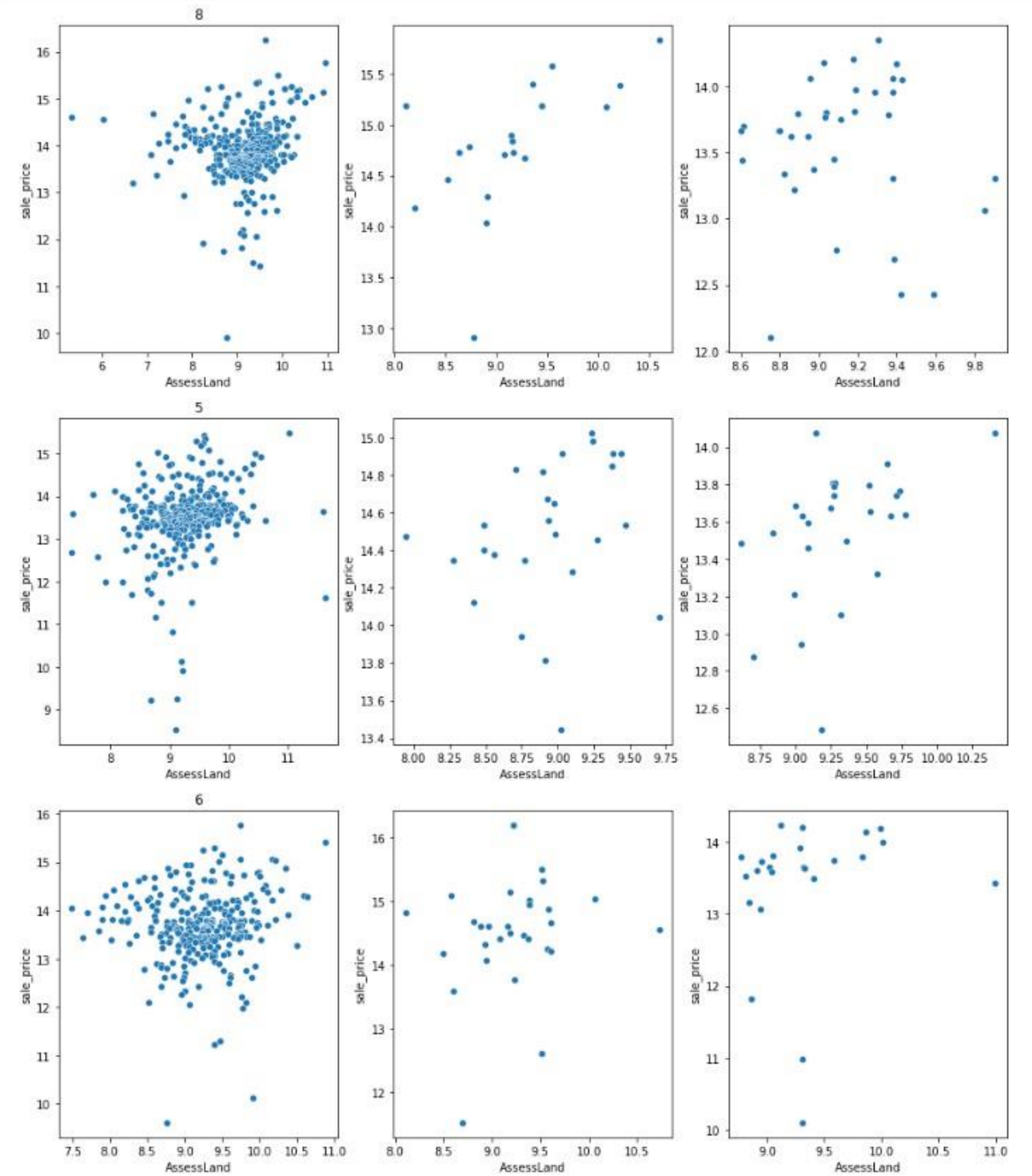
It can be seen that our target column is normally distributed.



School district distribution.

We can observe that the number of categories have been reduced.

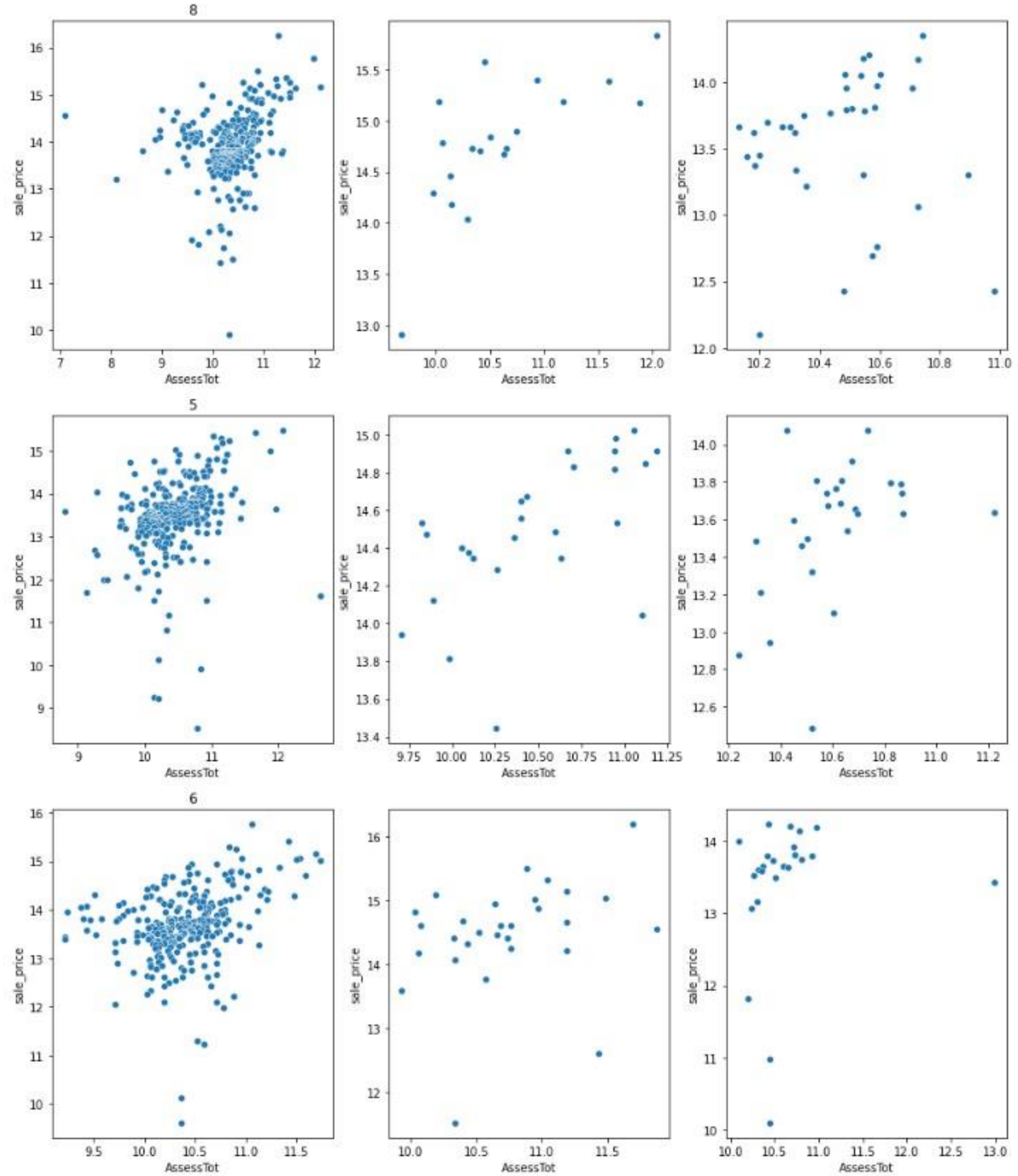




Column 1: 'low_min_high_max_neighborhood', School group 3 ,2 Basements

Column 2: 'high_min_high_max_neighborhood', School group 3, 2 Basements

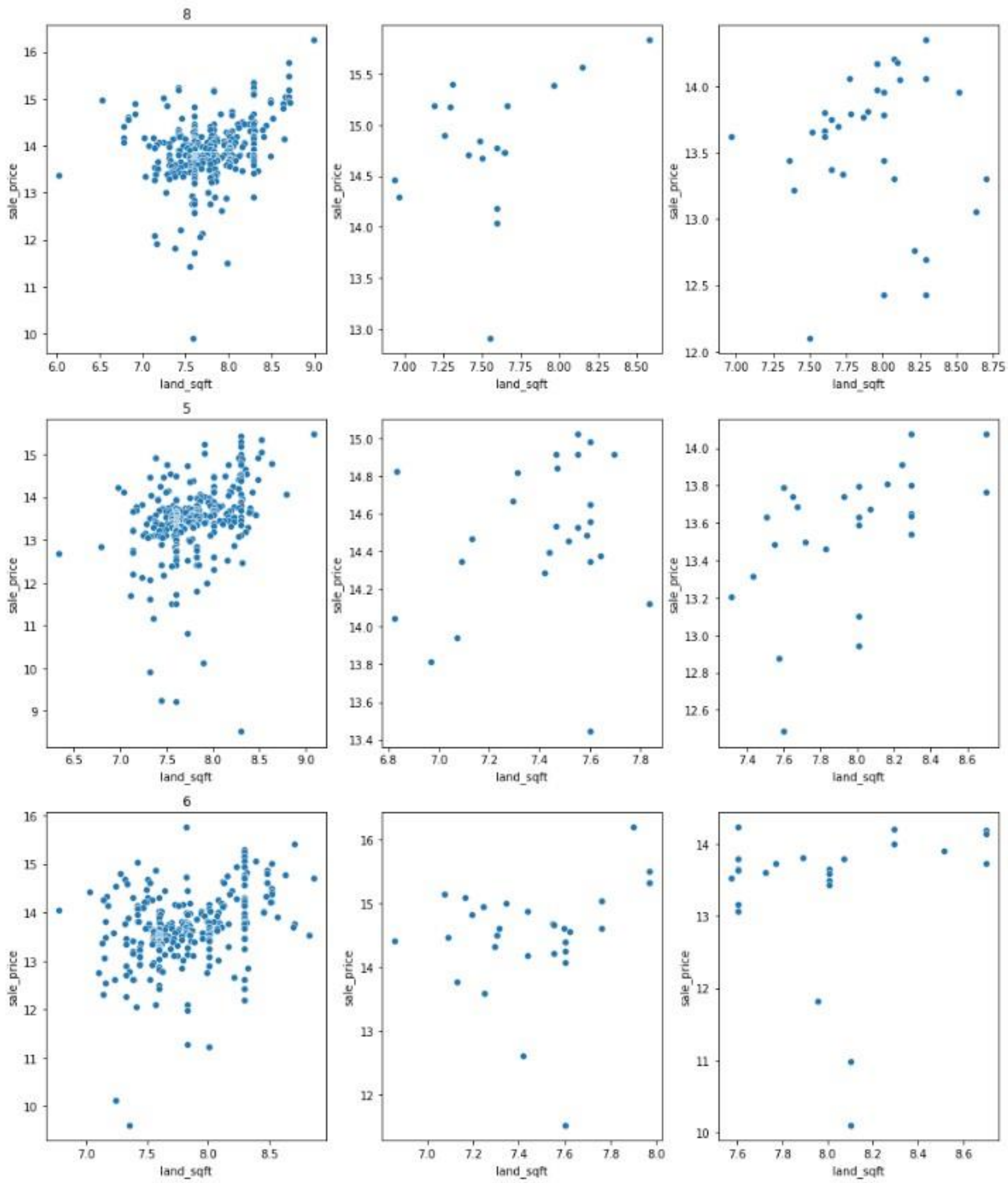
Column 3: 'high_min_low_max_neighborhood' School group 3, 2 Basements



The sale price in the first column is clustered around a certain region and the regions move every year.

The sale price in the second seems to be positively correlated to the AssessTot.

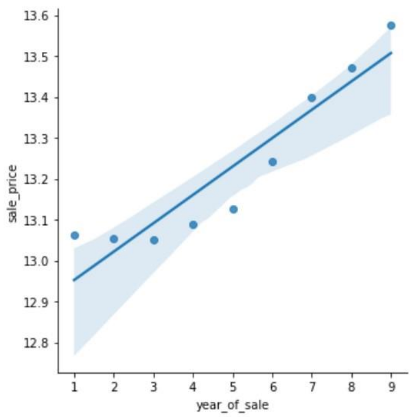
The sale price in the third column has a weak positive correlation to AssessTot.

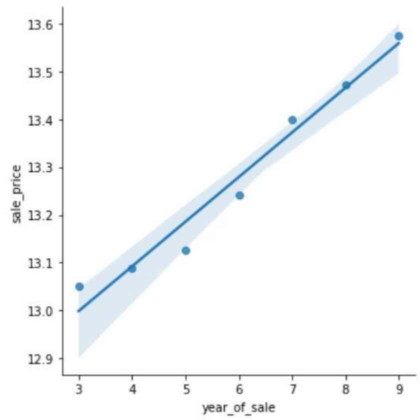


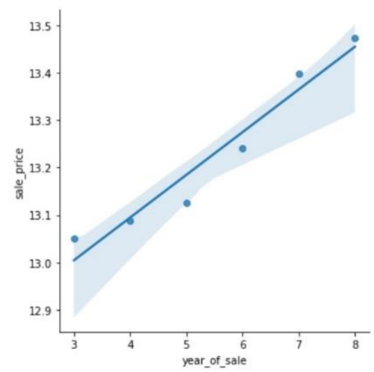
The sale price in the first column is clustered around a certain region and the regions move very year.

The sale price in the second seems to be positively correlated to the land_sqft.

The sale price in the third column has a weak positive correlation to land_sqft.







Inference:

Plot 1 – For the first three years, the price increase is negligible.

Plot 2 – From the 3rd year, there is a steady increase in average sale price.

Plot 3 – When we use an lm plot we can observe that the regression line is able to roughly estimate the 9th year's average price.

BASELINE MODEL BUILDING:

We can use the OLS method to build the base model:

- OLS regression

and the metrics that we use to validate our model is

- R2
- RMSE

NOTE: Before building this model, a primary OLS model was built and features having p-values greater than 0.05 were removed

	features	p_values
4	LotArea	0.284346
5	BldgArea	0.318451
9	LotDepth	0.710410
10	BldgFront	0.492356
11	BldgDepth	0.589462
14	BuiltFAR	0.262592
19	Total_alterations	0.060379
23	ProxCode_Detached	0.143007
24	ProxCode_Not available	0.963632
26	IrrLotCode_Y	0.468191
27	BsmtCode_four basements	0.102863
29	BsmtCode_one basement	0.413465
30	BsmtCode_three basement	0.461047
35	bclass_clusters_hila_bclass	0.932284
39	landuse_group_One & Two Family Buildings	0.134654
45	Lot_Inside	0.302241
46	Lot_other_lots	0.236842

OLS REGRESSION:

Dep. Variable:	sale_price	R-squared:	0.298
Model:	OLS	Adj. R-squared:	0.297
Method:	Least Squares	F-statistic:	662.6
Date:	Wed, 27 Jul 2022	Prob (F-statistic):	0.00
Time:	08:49:35	Log-Likelihood:	-51808.
No. Observations:	48491	AIC:	1.037e+05
Df Residuals:	48459	BIC:	1.040e+05
Df Model:	31		
Covariance Type:	nonrobust		

Omnibus:	32210.122	Durbin-Watson:	0.340
Prob(Omnibus):	0.000	Jarque-Bera (JB):	517870.818
Skew:	-3.007	Prob(JB):	0.00
Kurtosis:	17.837	Cond. No.	5.00e+03

The condition number is large, 5.00e+03. This might indicate that there are strong multicollinearity or other numerical problems.

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.5392	0.169	38.674	0.000	6.208	6.871
land_sqft	0.1617	0.039	4.111	0.000	0.085	0.239
gross_sqft	0.1777	0.016	11.357	0.000	0.147	0.208
year_of_sale	0.0707	0.001	49.116	0.000	0.068	0.074
ResArea	0.0239	0.018	1.303	0.193	-0.012	0.060
NumFloors	0.0413	0.009	4.746	0.000	0.024	0.058
AssessLand	-0.2498	0.010	-24.175	0.000	-0.270	-0.230
AssessTot	0.6600	0.013	51.041	0.000	0.635	0.685
ResidFAR	0.1629	0.015	10.527	0.000	0.133	0.193
FacilFAR	-0.2139	0.015	-14.071	0.000	-0.244	-0.184
SHAPE_Leng	-0.1606	0.045	-3.571	0.000	-0.249	-0.072
SHAPE_Area	0.1283	0.043	2.988	0.003	0.044	0.212
Total_alterations	-0.0337	0.009	-3.746	0.000	-0.051	-0.016
age	0.0026	0.000	19.510	0.000	0.002	0.003
building_class_category_02 TWO FAMILY HOMES	-0.0933	0.009	-10.496	0.000	-0.111	-0.076
building_class_category_03 THREE FAMILY HOMES	0.0666	0.027	2.430	0.015	0.013	0.120
ProxCode_Detached	-0.0373	0.013	-2.959	0.003	-0.062	-0.013
ProxCode_Semi-attached	-0.0383	0.008	-4.726	0.000	-0.054	-0.022
BsmtCode_four basements	-0.0755	0.045	-1.660	0.097	-0.165	0.014
BsmtCode_no basement	-0.1002	0.020	-4.979	0.000	-0.140	-0.061
BsmtCode_two basements	-0.0315	0.008	-3.963	0.000	-0.047	-0.016
neighbor_clusters_high_min_low_max_neighborhood	-0.5673	0.025	-22.328	0.000	-0.617	-0.518
neighbor_clusters_low_min_high_max_neighborhood	-0.6207	0.021	-28.905	0.000	-0.663	-0.579
neighbor_clusters_low_min_low_max_neighborhood	-0.7309	0.022	-32.508	0.000	-0.775	-0.687
bclass_clusters_liha_bclass	-0.3230	0.048	-6.678	0.000	-0.418	-0.228
bclass_clusters_lila_bclass	-0.4750	0.054	-8.735	0.000	-0.582	-0.368
landuse_group_Multi-Family Walk-Up Buildings	-0.2403	0.026	-9.160	0.000	-0.292	-0.189
landuse_group_other use	-0.8069	0.066	-12.204	0.000	-0.936	-0.677
school_Group2	-0.2512	0.014	-17.503	0.000	-0.279	-0.223
school_Group3	0.1087	0.015	7.386	0.000	0.080	0.138
school_Group4	-0.1835	0.019	-9.857	0.000	-0.220	-0.147
school_Group5	-0.0840	0.016	-5.298	0.000	-0.115	-0.053

Inference from the OLS model:

AssessTot, year of sale, low_min_low_max_neighborhood, low_min_high_max_neighborhood, AssessLand, high_min_low_max neighborhood, and age are features that contribute more to the sale price.

Let’s analyze the school group columns:

- When a property is present in School_Group3 it experiences a 10.87% price variation on average.
- When a property is present in School_Group2 it experiences a -25.12% price variation on average.
- When a property is present in School_Group4 it experiences a -18.35% price variation on average.
- When a property is present in School_Group4 it experiences a -8.40% price variation on average.
- In other words, we could say that with respect to School_Group1 School_Group3 is valued more while other School groups are valued lower.
- Due to n-1 encoding the value of the first school group is fused into the constant term and whenever a property falls in any neighborhood category the linear equation balances it by making the coefficient of other terms negative or positive based on its value.

Let’s analyze the neighborhood features.

- When a property is present in high_min_low_max_neighborhood it experiences a -56.73% price variation on average.
- When a property is present in low_min_high_max_neighborhood it experiences a -62.07% price variation on average.
- When a property is present in low_min_low_max_neighborhood it experiences a -73.09% price variation on average.
- In other words, we could say that with respect to high_min_high_max_neighborhood all other neighbourhoods are undervalued.
- Due to n-1 encoding the value of the high_min_high_max_neighborhood is fused into constant term and whenever a property falls in any other neighbourhood category the linear equation balances it by making the coefficient of other terms negative.

VARIANCE INFLATION FACTOR:

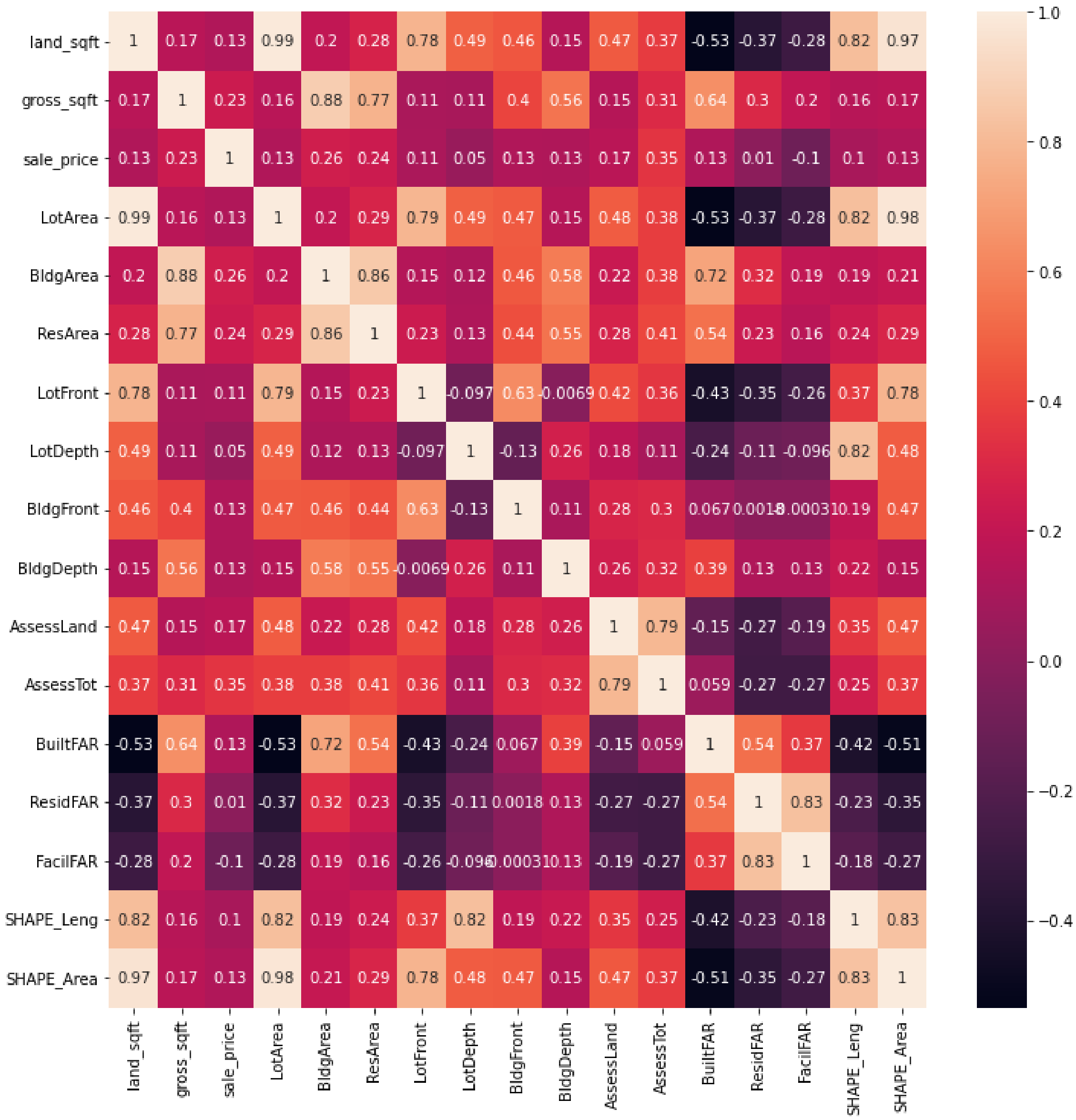
	VIF_Factor	Features
0	13104.515167	SHAPE_Area
1	9853.408315	land_sqft
2	6707.432211	SHAPE_Leng
3	1625.800318	ResArea
4	1224.366078	AssessTot
5	1074.285287	gross_sqft
6	767.935671	LotFront
7	726.783848	AssessLand
8	34.757881	NumFloors
9	14.961045	FacilFAR
10	12.015618	age
11	6.265201	ResidFAR
12	5.628373	year_of_sale

After multiple iterations, we end up with 4 columns from 13 columns that have a VIF value lesser than 10.

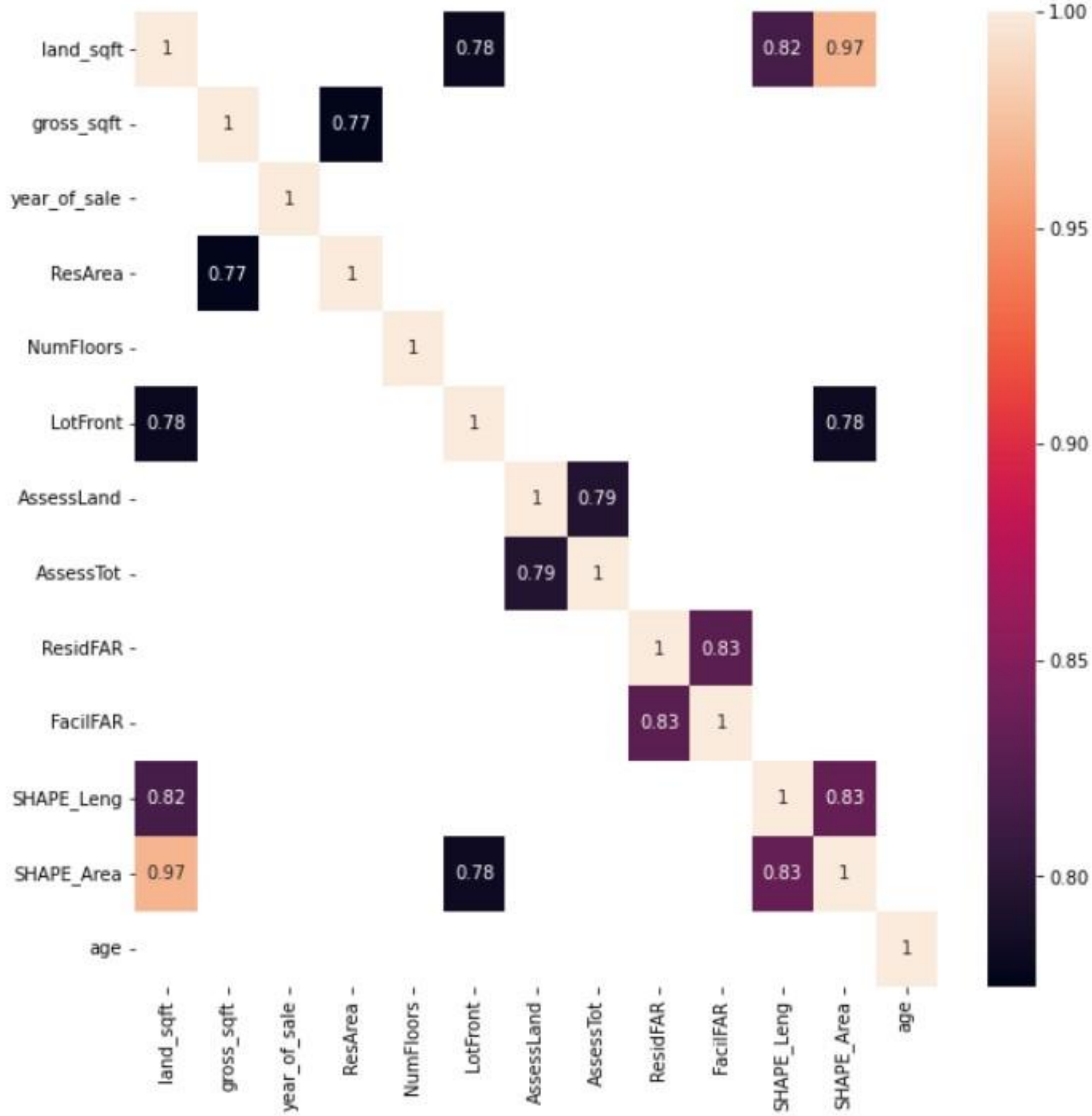
	VIF_Factor	Features
0	8.136861	NumFloors
1	7.609630	age
2	4.914631	year_of_sale
3	1.598188	ResidFAR

CHECKING FOR MULTI-COLLINEARITY AND TREATMENT:

When a pair of independent variables exhibit high correlation (that is when a pair of independent variables can explain one another with strong linear relation either positively or negatively) with each other it is termed a collinear effect. When more than one pair of independent variables exhibit a high correlation with each other it is termed a multi-collinear effect.



There is a threshold that is to be set to the correlation value to categorize it between the collinear effect and non-collinear effect. For this project, the threshold is set to be ± **0.7**. The dataset taken into consideration for this project has **more pairs** of independent variables which exhibit a multi-collinearity effect. This effect is visualized using a heatmap from the python seaborn library. The multi-collinearity effects were treated by dropping one of the columns in each pair of independent variables based on the domain knowledge.



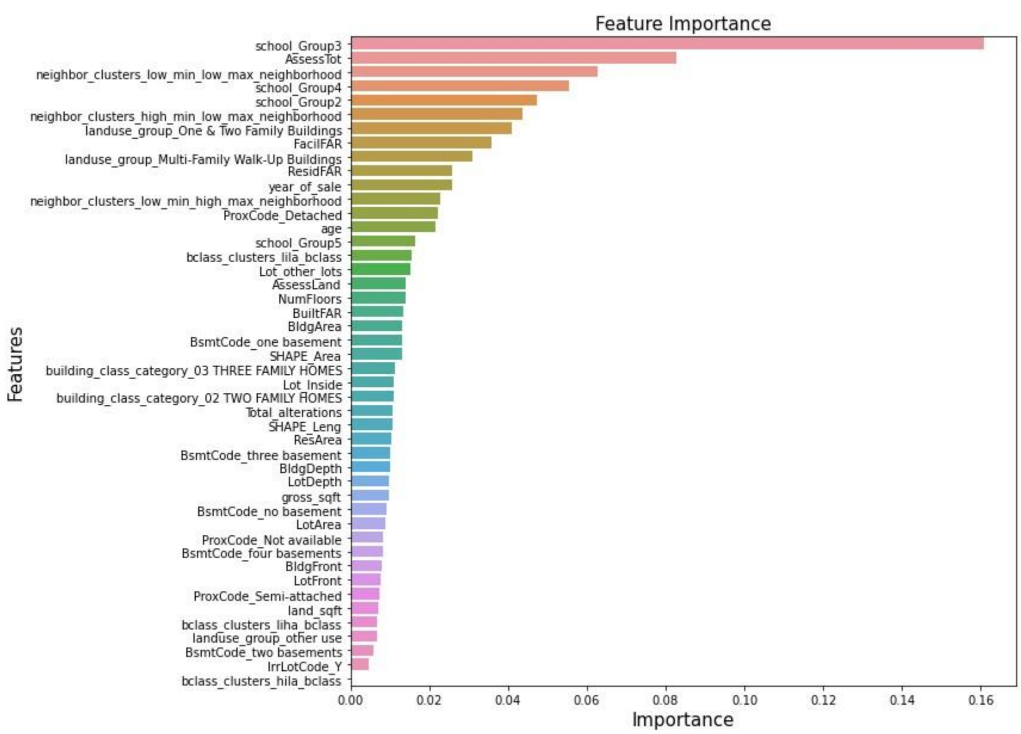
The above correlation plot contains numerical features that were selected from OLS model.

SCORE CARD FOR DIFFERENT MODELS:

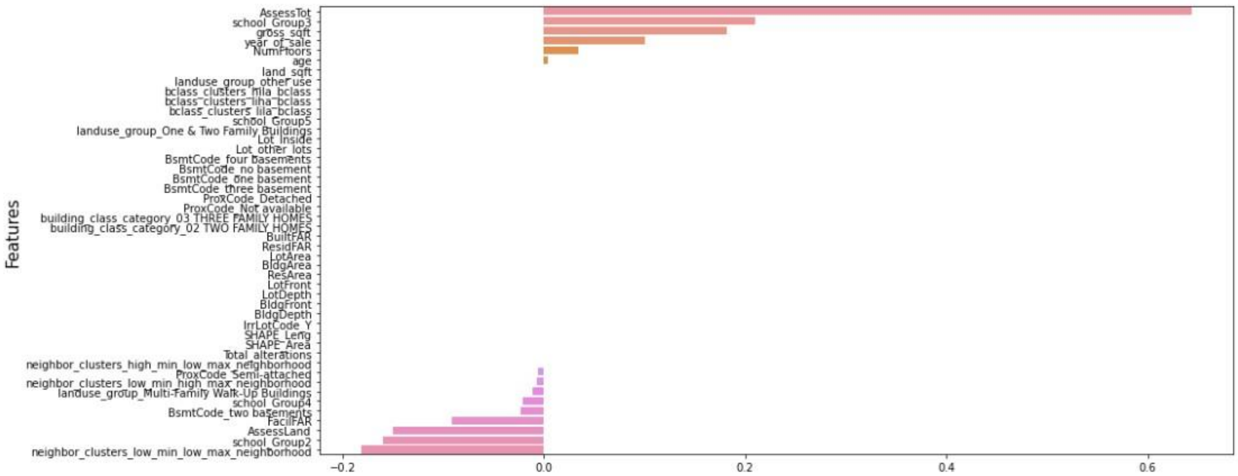
	Model_Name	Alpha (Wherever Required)	l1-ratio	R-Squared	Adj. R-Squared	Test_RMSE	Test_MAPE
0	Gradient boost regressor	-	-	0.374610	0.373864	0.650100	3.265836
1	Xtreme Gradient boost regressor	-	-	0.612411	0.611948	0.652200	3.211031
2	Multiple Linear Regression	-	-	0.315322	0.314505	0.655600	3.111154
3	Ridge regressor 1	1	-	0.315307	0.314489	0.655600	3.111485
4	Ridge regressor 2	2	-	0.315302	0.314485	0.655600	3.111407
5	Lasso regressor	0.01	-	0.278451	0.277590	0.670100	3.222914
6	Random forest regressor	-	-	0.884535	0.884397	0.671300	3.399994
7	Elastic Net	0.1	0.01	0.269290	0.268418	0.676200	3.290158
8	Decision tree regressor	-	-	0.960847	0.960800	1.008900	4.522860

Gradient boost regressor is giving a good RMSE compared to other algorithms. XGBoost is giving good R2 and adjusted R2 values compared to other algorithms.

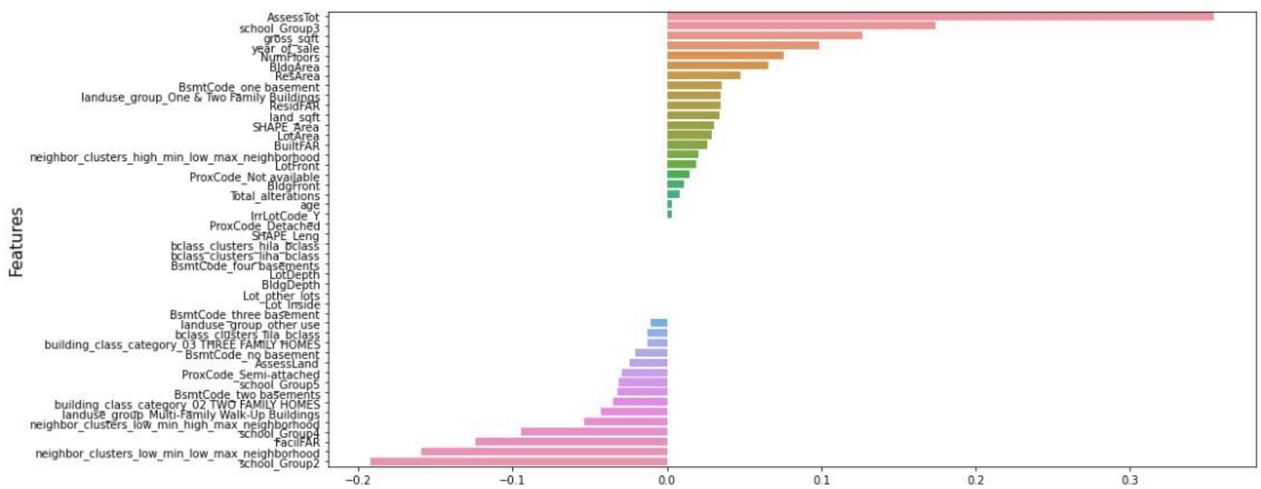
Xtreme Gradient Boosting Algorithm feature importance:



Lasso regularization Algorithm feature coefficients :

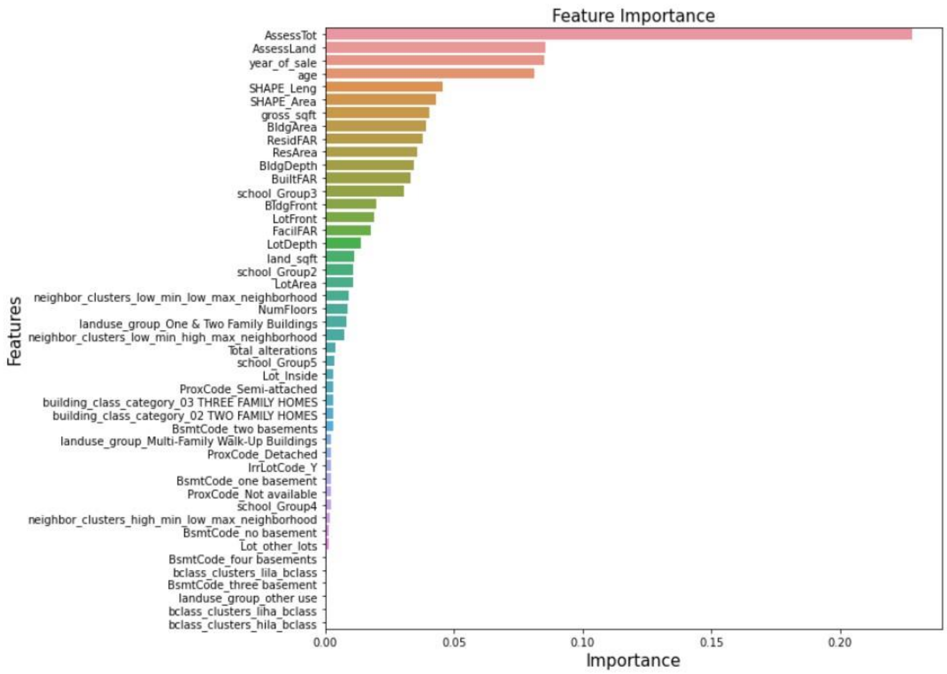


Elastic net regularization Algorithm feature coefficients :

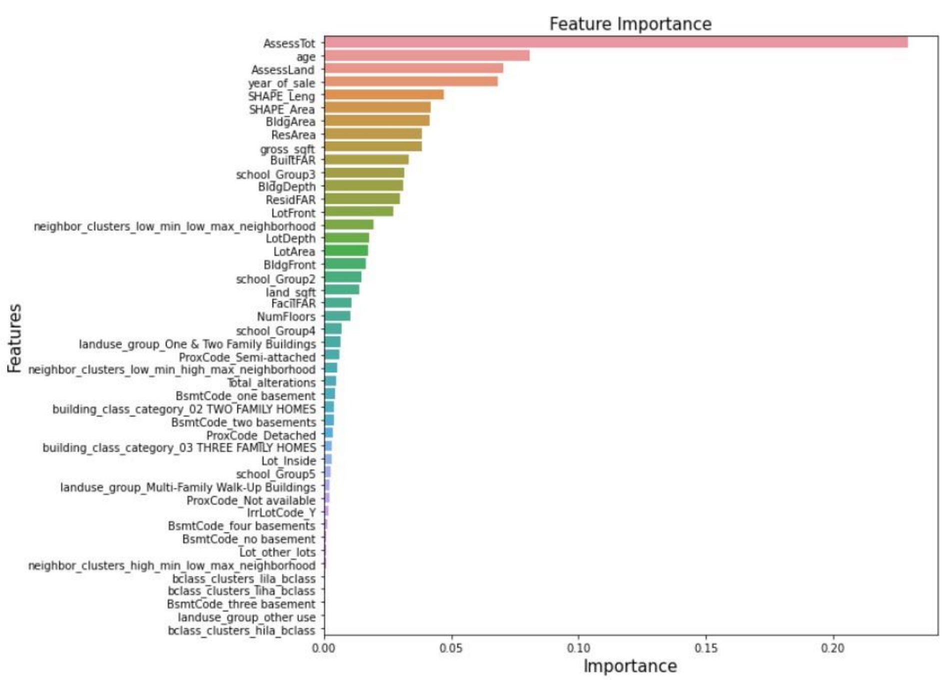


Inference: we can observe that many feature coefficients become 0 when we use Lasso regularization and when we use Elastic net fewer feature coefficients become 0.

Decision Tree Regressor Algorithm feature importance :



Random Forest Regressor Algorithm feature importance:



HYPERPARAMETER TUNING:

Carried out hyperparameter tuning with Random Forest Regressor and Xtreme Gradient Boost Regressor.

Did not achieve better performance (only marginal improvement) compared to the normal models.

Implications and Inference:

- From executing different algorithms, we can observe that a small group of features is given more weightage in all the above algorithms.
- It's also common knowledge to know that price of properties will increase with time.
- Apart from that, features like School group 3, age, assessed total value, assessed land value, number of floors in a building, gross square foot, and neighborhood clusters are very important and play a key role in determining the price of a property.

Closing Reflections:

We have learned that we have to separate the data again based on different regions and other features and multiple small models have to be built to increase the prediction accuracy.