Investigation of Property Sale Prices Using Machine Learning in Philadelphia

Team 11

Shusaku Asai

Congjun Huang

Jingjing Wang

1. Abstract

Many machine learning applications model home sales price. We extended this work by applying various sales price models to Single-Family, Multi-Family, Mixed-Use, Industrial, Commercial, and Vacant Land property types in Philadelphia. Within each property type, we built and tuned a best performing model and tested its generalizability on future unseen data. We found that the Single-Family property type model had the best performing model on the test data and the Commercial property type had the worst performing model as defined by RMSE. Across all property types, XGBoost and Random Forest models showed superior performance. Future home buyers in Philadelphia may find our model useful in predicting sales prices of homes. Government and businesses should use the Commercial, Industrial, Mixed-Use and Vacant Land models with caution due to their poorer generalizability.

2. Introduction and Motivation

2.1 Introduction

Property in the United States plays an integral role in the economy, community, and social development. Properties are often categorized by government, housing, and community entities and have socio-economic impacts.

For example, housing prices in the US are rapidly growing (1) in urban areas (2). Further, the degree of urbanization in the US has been steadily increasing from 1970 to 2020 (3). Research has also shown that owning a home is associated with financial (4) and social-psychological benefits (5). Property types such as Industrial and Commercial are also impactful, as commercial development supported 8 million jobs and generated $338.1 billion in salaries in 2020 (6). Mixed-Use property types have been suggested to reduce automobile dependency in urban areas (7). In these ways, property types have been shown to impact the economy, individuals, and communities.

2.2 Motivation

Given the increasing demand and prices of homes, and the socioeconomic impacts of construction of differing property types, there exists a need to investigate the sales price of properties from specific property types in urban areas. We investigate six property types: Single-Family, Multi-Family, Commercial, Industrial, Mixed-Use and Vacant Land. Our investigation implemented machine learning algorithms that predict the sales price of a property within each property type. Our research has utility for home buyers to predict future prices given the characteristics of a home, and for government officials and businesses to forecast sales prices of properties for budgeting and taxation purposes.

3. Background

There exists rich literature that models housing prices. Analysis of other property types is less prevalent. Table 1 provides a literature review on methodologies from various papers pertaining to our analysis.

Table 1 Literature Overview

Type	Take-away	Author
Commercial	Renovation status and zip code were vital predictors of commercial loan defaults.	(8)
Commercial	A decision tree has lowest RMSE on Malaysian commercial property price prediction.	(9)
Single & Multi-Family	A random forest and GBM work best to predict housing sales prices in Hong Kong per RMSE.	(10)
Single & Multi-Family	House amenities, zip codes, and days on market were used to predict home prices in Virginia.	(11)
Single & Multi-Family	Ensemble methods are competitive to predict home sales price in the Ames and Boston.	(12)
Single & Multi-Family	Area, bedrooms, flooring type, and condition were used to construct an aggregated model to predict Mumbai home sales prices.	(13)
Single & Multi-Family	Geospatial features are used to fit Network Embedded regression models.	(14)
Commercial	Commercial property is strongly dependent on spatial features and econometrics in Malaysia.	(15)

Geospatial information is required for all property types. Regression trees, random forests, boosting models, and neural networks tend to predict sales price best. Papers that utilize sales prices outcome use Mean Squared Error (MSE) or Root Mean Squared Error (RMSE) as evaluation metrics.

4. Data

4.1 Data Source

We utilized the publicly available Philadelphia Properties and Assessment History dataset (16). Table 2 provides a dataset description.

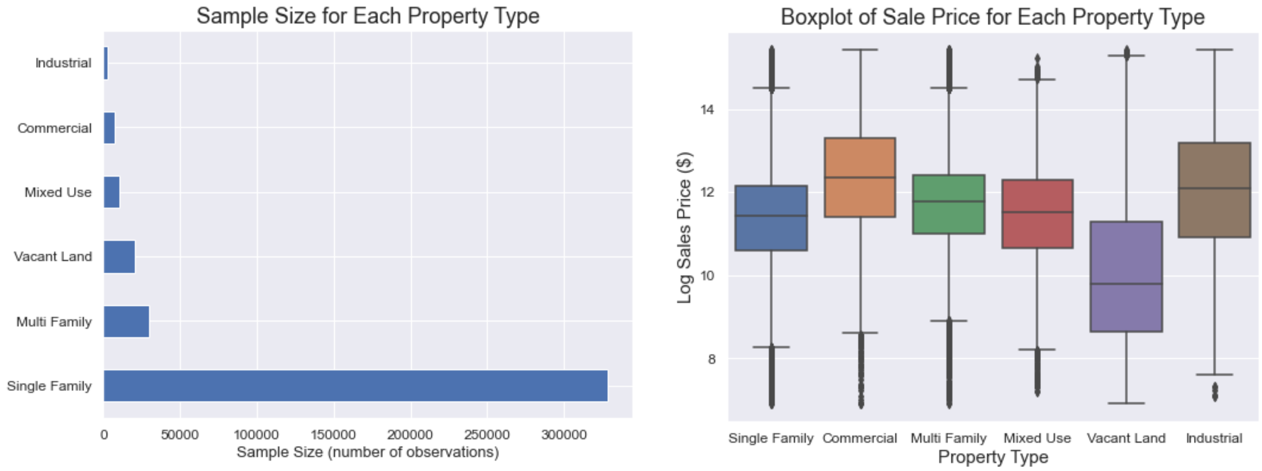
Table 2: Dataset Relevant Descriptions

Observations	Numeric features	Categorical features	Property types
581,191	32	46	Single-Family Multi Family Commercial Industrial Mixed-Use Vacant Land

Each observation in the dataset is categorized as Single-Family, Multi-Family, Commercial, Industrial Mixed-Use or Vacant Land. A comprehensive data dictionary is found online (17).

4.2 Exploratory Data Analysis (EDA)

Figure 1 displays the sample size of each property type prior to processing. Single-Family observations are most populous. Figure 2 outputs the distribution of log sales price by property type. Commercial and Industrial property types have the highest median sales prices, while Single-Family, Multi-Family, and Mixed-Use property types have similar medians.



Figures 1 and 2: Barplot for Number of Available Data on Each Year (L)
Boxplot of Sale Price Distribution for Each Property Type (R)

Two maps are used to explore sales price and property types by latitude and longitude.

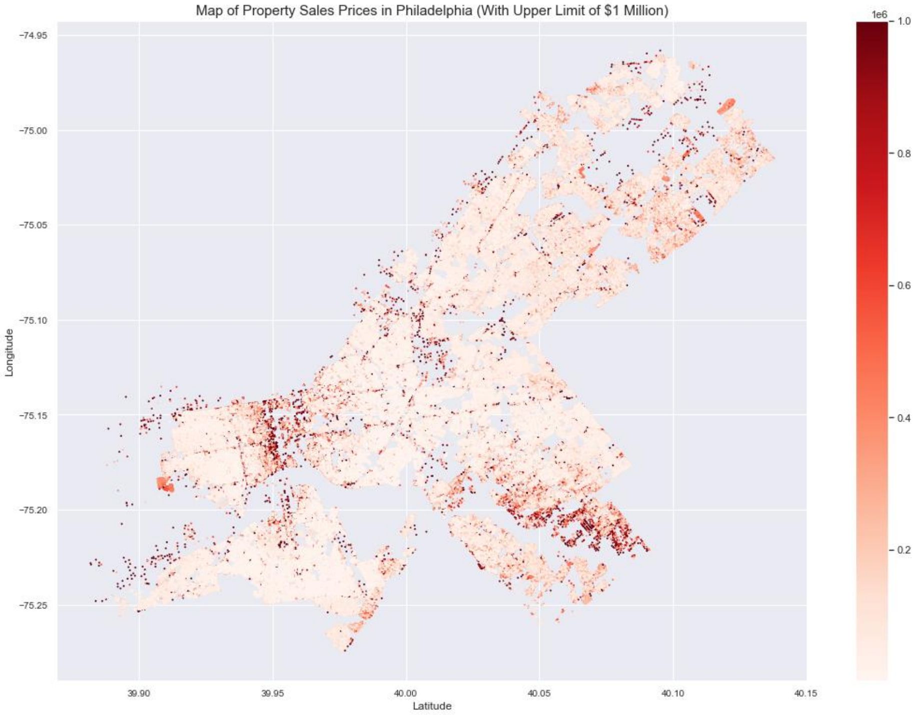


Figure 3: Map of Sales Price with Upper Limit of $1,000,000

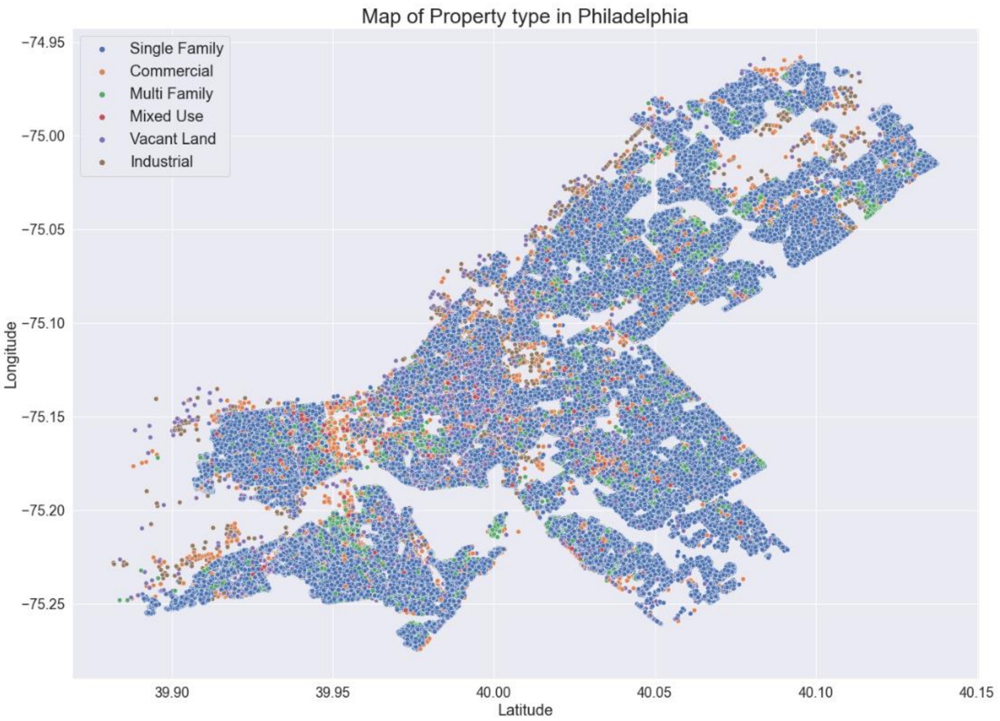


Figure 4: Map of Property Types

Higher sales prices naturally group together, and Industrial types are grouped along the coasts. A grouped boxplot by year of the log sales price is plotted in Figure 5. Commercial and Industrial types are most expensive across years, and there exists a general increase in log price across all types.

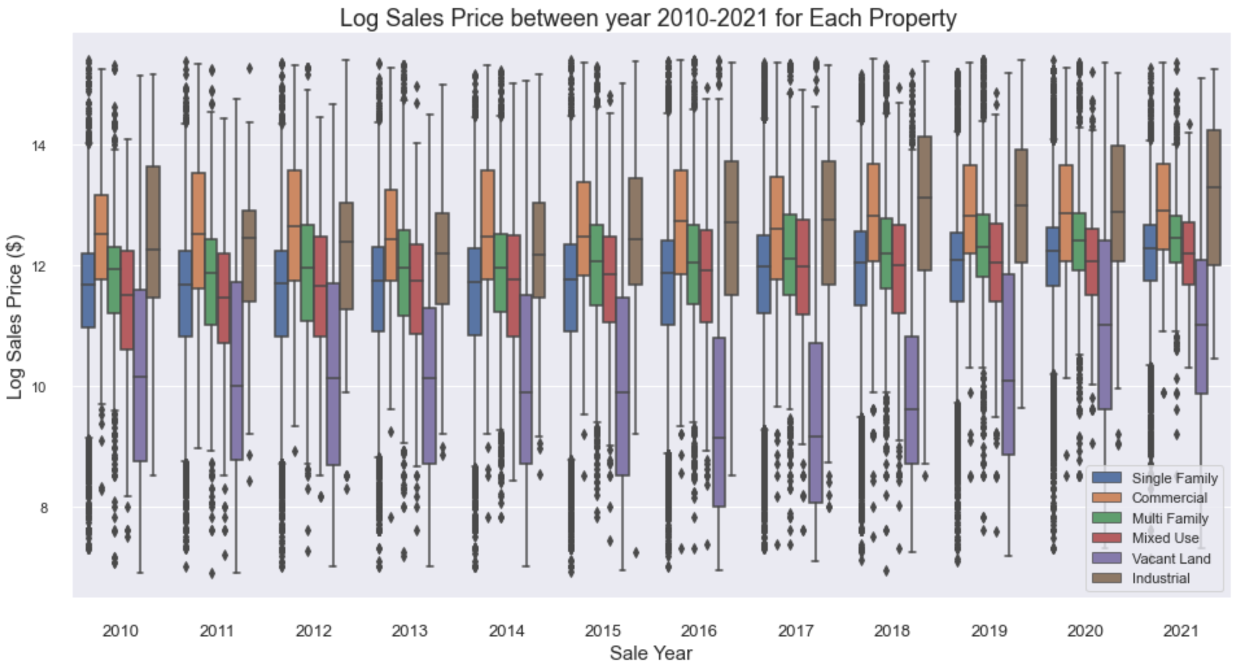


Figure 5: Boxplot of Sale Price (Year 2010-2021) for Each Property Types

5. Methods and Experiments

Our experiment was the prediction of "future" sales prices within each property type. We sub-set outcomes to prices of greater than $1000 and less than $5 million to remove outliers and to specify the analysis to reasonable prices. We modeled the outcome in thousands of dollars to reduce extreme variability. Within each property type dataset, we prepared training + validation (2010 – 2019) and test (2020 – 2021) datasets. The detailed workflow is seen in Table 3.

5.1 Preprocessing and Feature Engineering

Table 3: Preprocessing and Feature Engineering Step Description

Step	Task	Description
1	Variable Removal	Removed clearly unrelated features. Dropped 10 columns with more than 50% missing values.
2	Outcome Removal and Sub-setting	Removed samples with missing 'sale_date' or 'sale_price'. Sub-set the data to properties with a sales price of greater than $1000 and less than $5 million. Modified the outcome, sales price, to thousands of dollars.
3	Recode Categorical Features	Checked and replaced erroneous categorical variables and combined sparse categories into one.
4	Feature Engineering	a) Created 'sale_year', 'sale_month', 'sale_week', 'sale_day' and 'sale_date_of_week' b) Created 'property_age' c) Calculated 'bedrooms_per_room' and 'livable_area_ratio' d) Removed highly correlated numeric features
5	Split Datasets	a) Split datasets into each property type (6 resulting datasets) b) Within each dataset, split into training, validation, and test sets. Training + validation dataset contains property sold during 2010 – 2019. Test set contains property sold during 2020 - 2021.
6	Pipeline Application	Within each property type dataset: a) Filled missing values with median and standardized numeric features b) Filled missing values with mode and one-hot encoded categorical features c) Processed validation and test sets with same training pipeline

Step 4 included the removal of highly correlated features. Figure 6 shows the heatmap for the entire dataset that was used to drop correlated features.

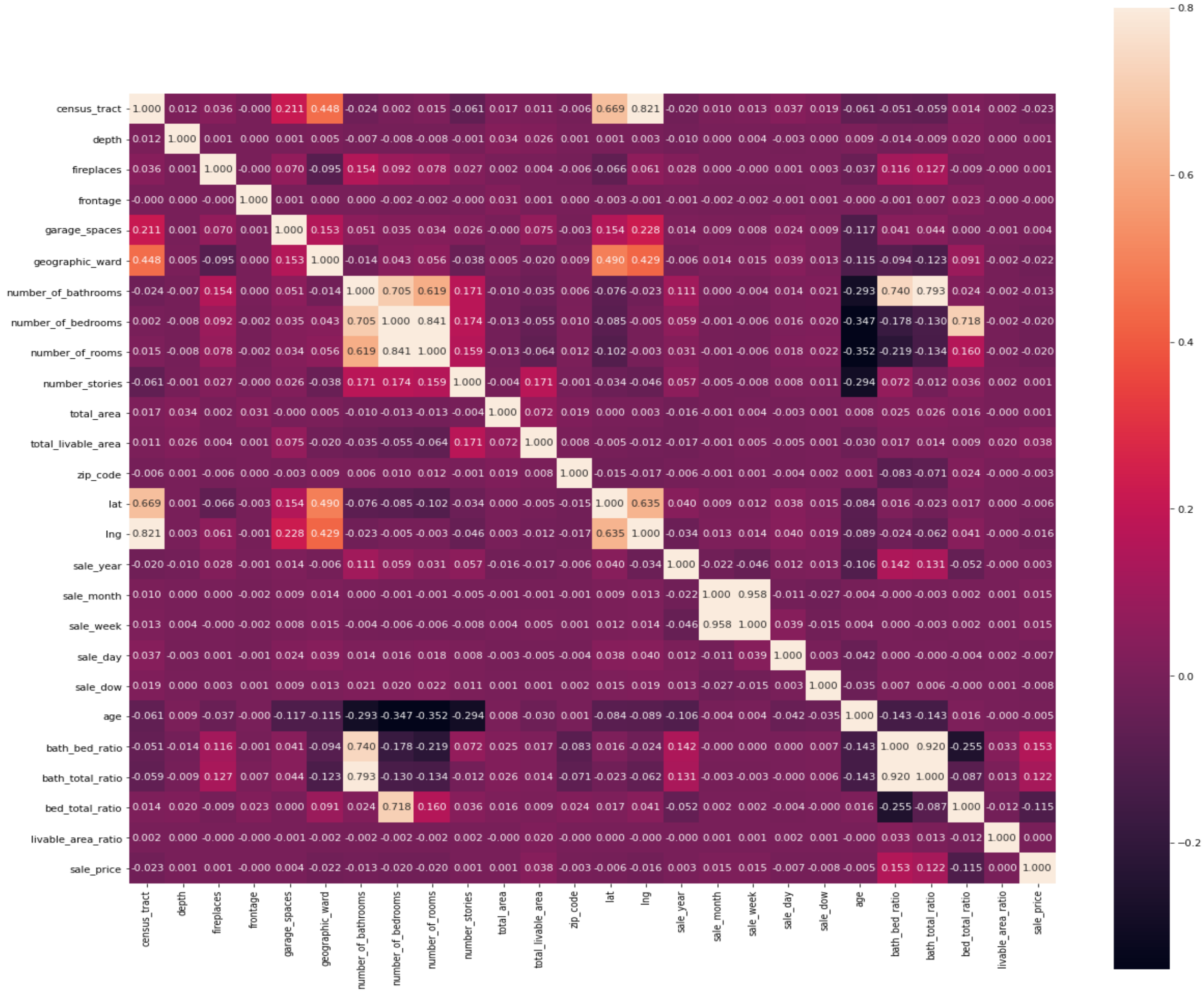


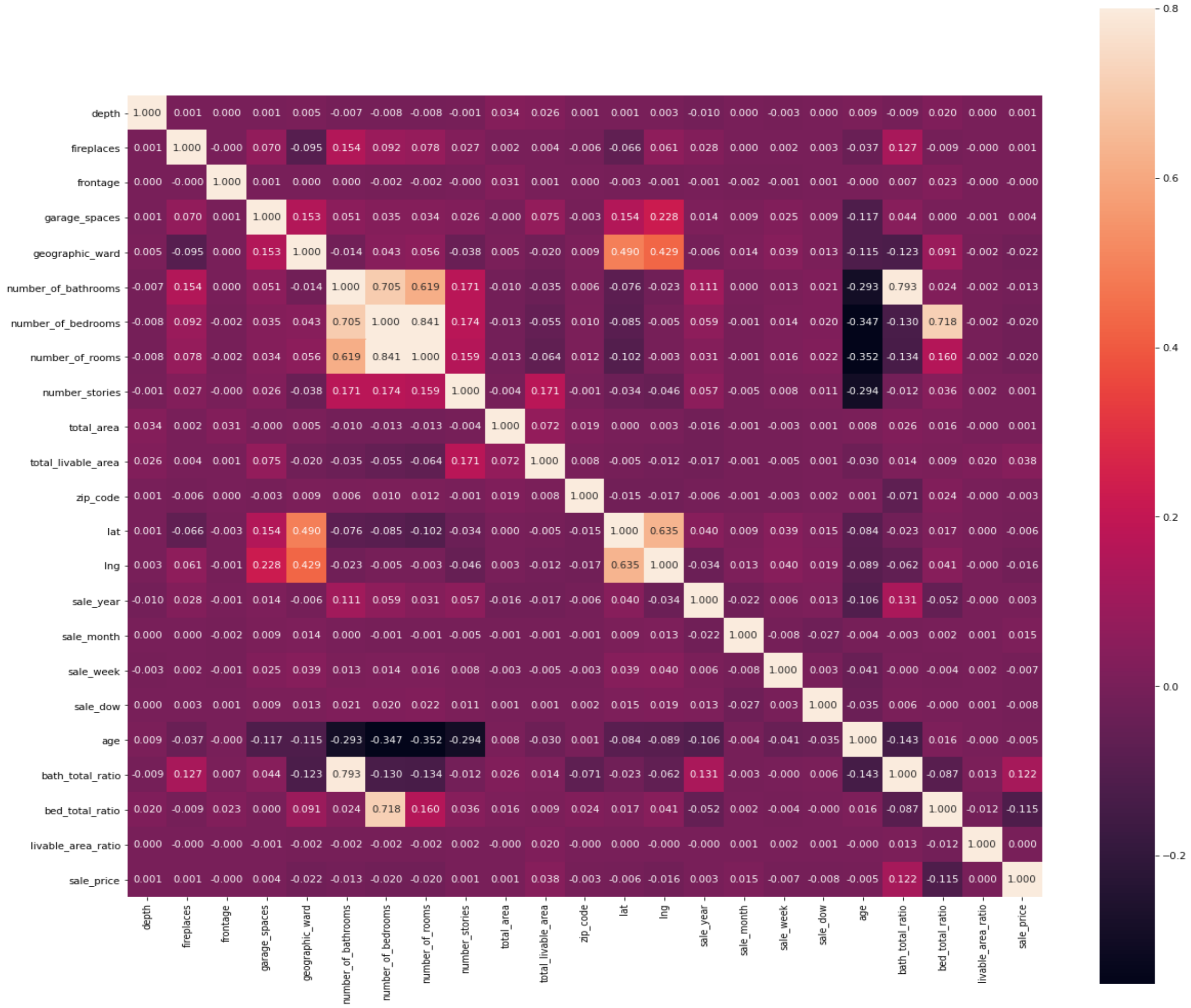
Figure 6: Correlation Heatmap

Highly correlated features were defined as pairs with a value ≥ 0.8. Table 4 describes how these correlated variables were handled.

Table 4: Methods Taken to Address High Correlation Problems

Correlated Features	Solution
‘census_tract’, ‘lng’	Removed the less essential feature ‘census_tract’
‘sale_month’, ‘sale_week’	Changed ‘sale_week’ to the week number in a month instead of in a year
‘bath_bed_ratio’, ‘bath_total_ratio’	Removed ‘bath_bed_ratio’
‘number_of_bedrooms’, ‘number_of_rooms’	No change since these features are common in other research (18)

After these changes, another matrix was output to show residual correlations. Figure 7 shows that all resulting variables had a correlation of under 0.8.



	depth	fireplaces	frontage	garage_spaces	geographic_ward	number_of_bathrooms	number_of_bedrooms	number_of_rooms	number_stories	total_area	total_livable_area	zip_code	lat	lng	sale_year	sale_month	sale_week	sale_dow	age	bath_total_ratio	bed_total_ratio	livable_area_ratio	sale_price
depth	1.000	0.001	0.000	0.001	0.005	-0.007	-0.008	-0.008	-0.001	0.034	0.026	0.001	0.001	0.003	-0.010	0.000	-0.003	0.000	0.009	-0.009	0.020	0.000	0.001
fireplaces	0.001	1.000	-0.000	0.070	-0.095	0.154	0.092	0.078	0.027	0.002	0.004	-0.006	-0.066	0.061	0.028	0.000	0.002	0.003	-0.037	0.127	-0.009	-0.000	0.001
frontage	0.000	-0.000	1.000	0.001	0.000	0.000	-0.002	-0.002	-0.000	0.031	0.001	0.000	-0.003	-0.001	-0.001	-0.002	-0.001	0.001	-0.000	0.007	0.023	-0.000	-0.000
garage_spaces	0.001	0.070	0.001	1.000	0.153	0.051	0.035	0.034	0.026	-0.000	0.075	-0.003	0.154	0.228	0.014	0.009	0.025	0.009	-0.117	0.044	0.000	-0.001	0.004
geographic_ward	0.005	-0.095	0.000	0.153	1.000	-0.014	0.043	0.056	-0.038	0.005	-0.020	0.009	0.490	0.429	-0.006	0.014	0.039	0.013	-0.115	-0.123	0.091	-0.002	-0.022
number_of_bathrooms	-0.007	0.154	0.000	0.051	-0.014	1.000	0.705	0.619	0.171	-0.010	-0.035	0.006	-0.076	-0.023	0.111	0.000	0.013	0.021	-0.293	0.793	0.024	-0.002	-0.013
number_of_bedrooms	-0.008	0.092	-0.002	0.035	0.043	0.705	1.000	0.841	0.174	-0.013	-0.055	0.010	-0.085	-0.005	0.059	-0.001	0.014	0.020	-0.347	-0.130	0.718	-0.002	-0.020
number_of_rooms	-0.008	0.078	-0.002	0.034	0.056	0.619	0.841	1.000	0.159	-0.013	-0.064	0.012	-0.102	-0.003	0.031	-0.001	0.016	0.022	-0.352	-0.134	0.160	-0.002	-0.020
number_stories	-0.001	0.027	-0.000	0.026	-0.038	0.171	0.174	0.159	1.000	-0.004	0.171	-0.001	-0.034	-0.046	0.057	-0.005	0.008	0.011	-0.294	-0.012	0.036	0.002	0.001
total_area	0.034	0.002	0.031	-0.000	0.005	-0.010	-0.013	-0.013	-0.004	1.000	0.072	0.019	0.000	0.003	-0.016	-0.001	-0.003	0.001	0.008	0.026	0.016	-0.000	0.001
total_livable_area	0.026	0.004	0.001	0.075	-0.020	-0.035	-0.055	-0.064	0.171	0.072	1.000	0.008	-0.005	-0.012	-0.017	-0.001	-0.005	0.001	-0.030	0.014	0.009	0.020	0.038
zip_code	0.001	-0.006	0.000	-0.003	0.009	0.006	0.010	0.012	-0.001	0.019	0.008	1.000	-0.015	-0.017	-0.006	-0.001	-0.003	0.002	0.001	-0.071	0.024	-0.000	-0.003
lat	0.001	-0.066	-0.003	0.154	0.490	-0.076	-0.085	-0.102	-0.034	0.000	-0.005	-0.015	1.000	0.635	0.040	0.009	0.039	0.015	-0.084	-0.023	0.017	0.000	-0.006
lng	0.003	0.061	-0.001	0.228	0.429	-0.023	-0.005	-0.003	-0.046	0.003	-0.012	-0.017	0.635	1.000	-0.034	0.013	0.040	0.019	-0.089	-0.062	0.041	-0.000	-0.016
sale_year	-0.010	0.028	-0.001	0.014	-0.006	0.111	0.059	0.031	0.057	-0.016	-0.017	-0.006	0.040	-0.034	1.000	-0.022	0.006	0.013	-0.106	0.131	-0.052	-0.000	0.003
sale_month	0.000	0.000	-0.002	0.009	0.014	0.000	-0.001	-0.001	-0.005	-0.001	-0.001	-0.001	0.009	0.013	-0.022	1.000	-0.008	-0.027	-0.004	-0.003	0.002	0.001	0.015
sale_week	-0.003	0.002	-0.001	0.025	0.039	0.013	0.014	0.016	0.008	-0.003	-0.005	-0.003	0.039	0.040	0.006	-0.008	1.000	0.003	-0.041	-0.000	-0.004	0.002	-0.007
sale_dow	0.000	0.003	0.001	0.009	0.013	0.021	0.020	0.022	0.011	0.001	0.001	0.002	0.015	0.019	0.013	-0.027	0.003	1.000	-0.035	0.006	-0.000	0.001	-0.008
age	0.009	-0.037	-0.000	-0.117	-0.115	-0.293	-0.347	-0.352	-0.294	0.008	-0.030	0.001	-0.084	-0.089	-0.106	-0.004	-0.041	-0.035	1.000	-0.143	0.016	-0.000	-0.005
bath_total_ratio	-0.009	0.127	0.007	0.044	-0.123	0.793	-0.130	-0.134	-0.012	0.026	0.014	-0.071	-0.023	-0.062	0.131	-0.003	-0.000	0.006	-0.143	1.000	-0.087	0.013	0.122
bed_total_ratio	0.020	-0.009	0.023	0.000	0.091	0.024	0.718	0.160	0.036	0.016	0.009	0.024	0.017	0.041	-0.052	0.002	-0.004	-0.000	0.016	-0.087	1.000	-0.012	-0.115
livable_area_ratio	0.000	-0.000	-0.000	-0.001	-0.002	-0.002	-0.002	-0.002	0.002	-0.000	0.020	-0.000	0.000	-0.000	-0.000	0.001	0.002	0.001	-0.000	0.013	-0.012	1.000	0.000
sale_price	0.001	0.001	-0.000	0.004	-0.022	-0.013	-0.020	-0.020	0.001	0.001	0.038	-0.003	-0.006	-0.016	0.003	0.015	-0.007	-0.008	-0.005	0.122	-0.115	0.000	1.000

Figure 7: Residual Correlation Heatmap

After pre-processing and feature engineering, a total of 12 datasets existed: six training + validation sets and six test sets. All datasets have 84 independent variables. The dimensions of each dataset are found in Table 5.

Table 5: Dataset Dimensions

	Single-Family	Multi -Family	Commercial	Industrial	Mixed Use	Vacant Land
Train + Val (%)	137,072 (83.32%)	13,497 (85.51%)	2,996 (89.38%)	1,253 (90.73%)	4,581 (87.09%)	11,856 (84.57%)
Test (%)	27,436 (16.67%)	2,288 (14.49%)	356 (10.62%)	128 (9.27%)	679 (12.91%)	2,163 (15.43%)

5.2 Machine Learning Modeling

We modeled each property type’s sales price with models seen in Table 6 strictly on the training + validation datasets. Hyperparameters were tuned with the validation set with random search. The model with best performance for each property type was selected to apply on the test (“future”). Model performance was evaluated by RMSE and Mean Absolute Error (MAE). We used RMSE, MAE, and R^2 as evaluation metrics to mirror previous research (9, 10, 12, 14).

Table 6: Models Used

Model Type	Justification	Hyperparameters Explored	
Linear Regression	(9, 10, 12)	Linear	'fit_intercept' 'normalize'
		Lasso	'alpha' 'fit_intercept’ 'normalize'
		Ridge	'alpha’ 'fit_intercept' 'normalize'
KNN Regressor	(11)	'n_neighbors' 'weights' 'leaf_size'	
Random Forest Regressor	(19)	'n_estimators' 'max_depth' 'min_samples_split' 'min_samples_leaf' 'max_features'	
Multi-Layer Perceptron (MLP) Regressor	(20)	'hidden_layer_sizes' 'learning_rate_init' 'alpha' 'batch_size'	
XGBoost Regressor	(21)	'n_estimators' 'max_depth' 'learning_rate_init' 'min_child_weight'	

5.3 Experimental Design

The experimental workflow is visualized in Figure 8.

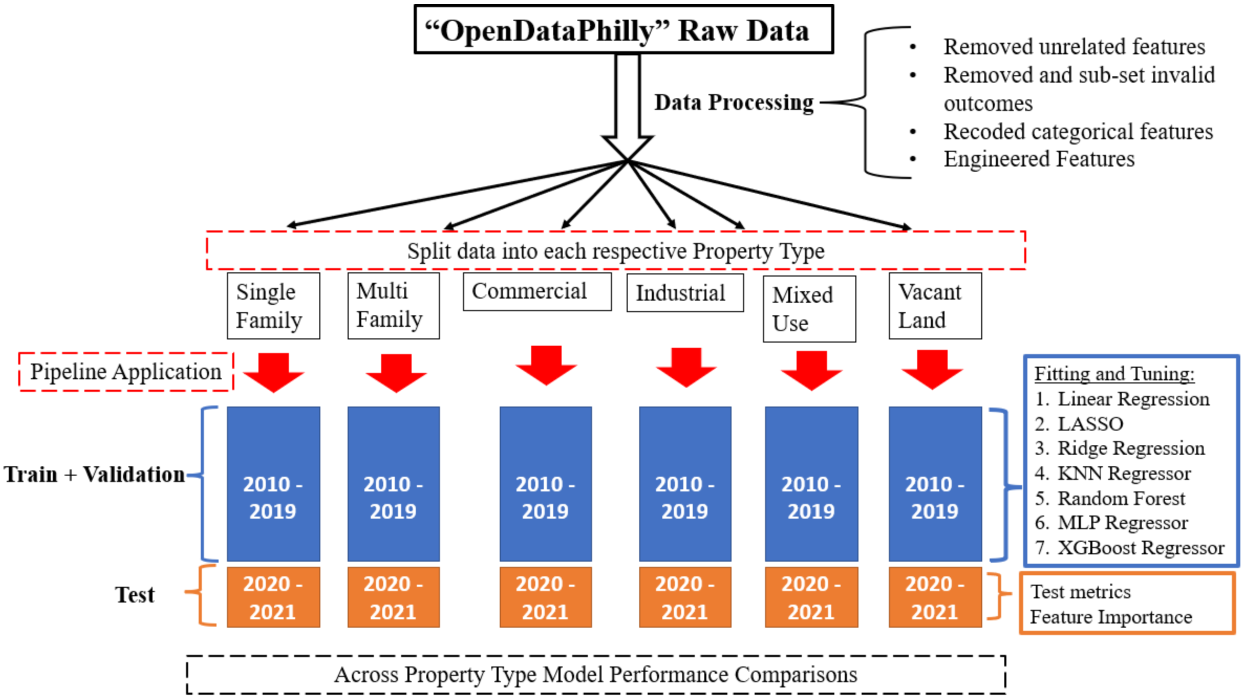


Figure 8: Experimental Design Flowchart

6. Results

6.1 Best Model Selection

Table 7 yields each model's performance on the validation sets. Green cells indicate that the model outperformed baseline and red indicates underperformance. The baseline model is defined as a prediction using the mean sale price of each category.

Table 7: Metrics for Seven Models on 6 Property Types

Model	Metrics	Single Family	Multi Family	Commercial	Industrial	Mixed Use	Vacant Land
Baseline	R^2	-2.10e-5	-3.21e-4	-6.00e-3	-2.63e-4	-8.19e-9	-3.97e-4
	RMSE	290.51	587.12	846.63	898.63	329.71	366.69
	MAE	155.75	308.05	587.72	620.17	205.83	166.43
Linear Regression	R^2	0.34	0.20	0.22	0.25	0.37	0.10
	RMSE	2.35e10	3.65e12	766.43	1.36e14	1.01e13	1.01e13
	MAE	106.66	266.99	489.94	450.98	166.26	156.22
Lasso	R^2	0.34	-0.01	0.22	0.28	0.37	0.10
	RMSE	237.02	654.70	763.83	715.86	261.78	331.11
	MAE	106.50	309.15	485.47	470.55	166.24	155.90
Ridge	R^2	0.34	0.14	0.22	0.28	0.37	0.05
	RMSE	236.94	1655.61	763.60	715.60	261.77	329.40
	MAE	106.58	264.41	485.93	474.56	165.30	160.61
KNN	R^2	0.99	0.99	0.99	0.99	0.99	0.99
	RMSE	170.73	410.46	693.86	713.67	240.56	267.69
	MAE	7.6e-2	2.77e-5	3.07e-5	2.86e-6	3.21e-6	1.46e-5
Random Forest	R^2	0.89	0.82	0.77	0.74	0.80	0.65
	RMSE	136.85	366.33	607.29	639.92	207.72	253.88
	MAE	34.79	103.60	221.43	251.95	73.97	68.28
MLP	R^2	0.62	0.24	0.27	0.39	0.51	0.24
	RMSE	178.03	587.13	693.10	692.12	236.55	317.71
	MAE	73.15	239.81	449.61	444.43	136.43	128.59
XGBoost	R^2	0.96	0.89	0.97	0.99	0.95	0.90
	RMSE	131.89	339.83	602.65	651.90	213.95	253.56
	MAE	31.24	44.78	112.68	1.7e-3	4.27	10.91

The Random Forest, XGBoost, and MLP Regressors had generally better performance. The best model for each property type from Table 6 as defined by smallest RMSE is output in Table 8.

Table 8: Best Model for 6 Property Types

Best Model	Single Family	Multi Family	Commercial
Model type	XGBoost Regressor	XGBoost Regressor	XGBoost Regressor
Hyperparameters	n_estimators: 100 max_depth: 11 learning_rate_init: 0.16 min_child_weight: 3	'n_estimators': 1000 'max_depth': 4 'learning_rate_init': 2.16 'min_child_weight': 3	n_estimators: 500 max_depth: 3 learning_rate_init: 1.47 min_child_weight: 1
RMSE	131.89	339.83	602.65
Best Model	Industrial	Mixed Use	Vacant Land
Model Type	Random Forest Regressor	Random Forest Regressor	XGBoost Regressor
Hyperparameters	n_estimators: 800 min_samples_split: 2 Min_sample_leaf: 5 Max_features: auto Max_depth: None	n_estimators: 800 min_samples_split: 2 Min_sample_leaf: 5 Max_features: auto Max_depth: None	n_estimators: 1000 max_depth: 5 learning_rate_init: 1.51 min_child_weight: 1
RMSE	639.92	207.72	253.56

The models shown in Table 8 were then used to predict on “future” property prices (test set: 2020 – 2021). Results in the test set are seen in Table 9.

Table 9: Model Generalization Performance on Test Data

	Single Family	Multi Family	Commercial	Industrial	Mixed Use	Vacant Land
Baseline R^2	-0.03	-3.04e-3	-3.78e-3	-4.24e-2	-9.27e-3	-3.90e-2
Baseline RMSE	312.63	501.71	916.28	1026.39	329.25	388.41
Baseline MAE	150.88	243.96	596.43	713.97	177.99	188.23
Test R^2	0.85	0.75	0.65	0.58	0.59	0.68
Test RMSE	266.54	360.50	692.85	649.69	210.38	295.21
Test MAE	95.60	167.07	416.46	404.58	101.60	139.06
Test R^2 – Train R^2	-0.11	-0.14	-0.32	-0.16	-0.21	-0.22
Test RMSE – Train RMSE	134.65	20.88	90.20	9.77	2.66	41.65
Test MAE – Train MAE	64.36	122.29	303.78	152.63	27.63	128.15

The Single-Family and Mixed-Use models showed smallest Test RMSE, whereas Commercial showed greatest test RMSE. Vacant Land also showed a low generalizability compared to other types.

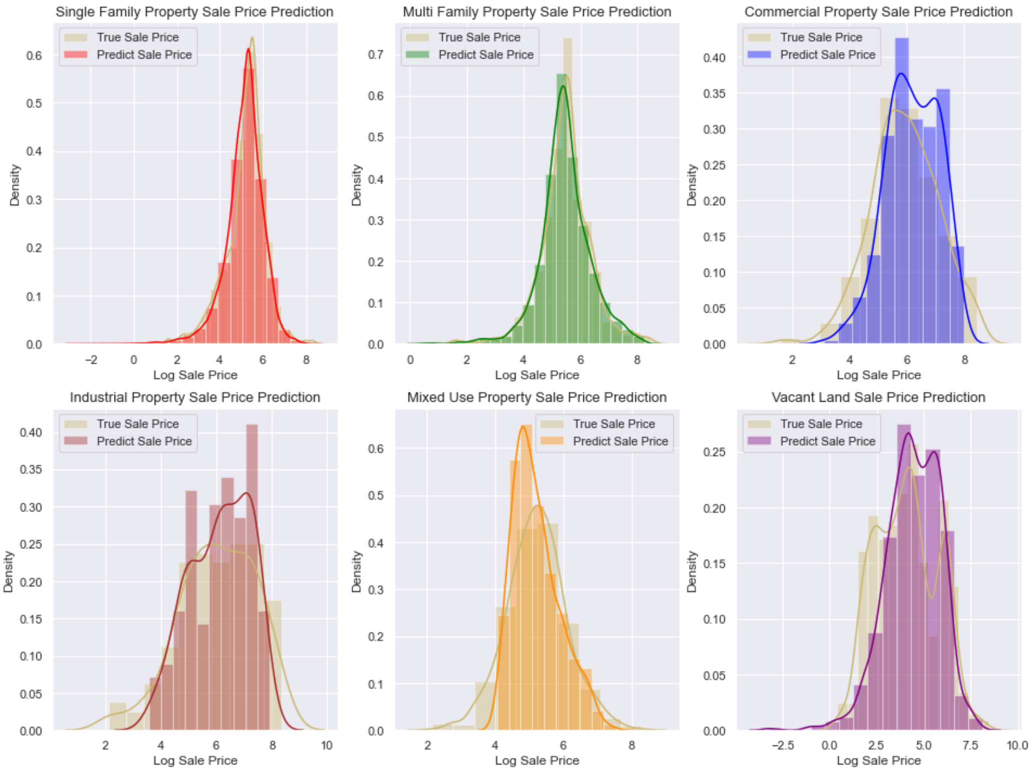
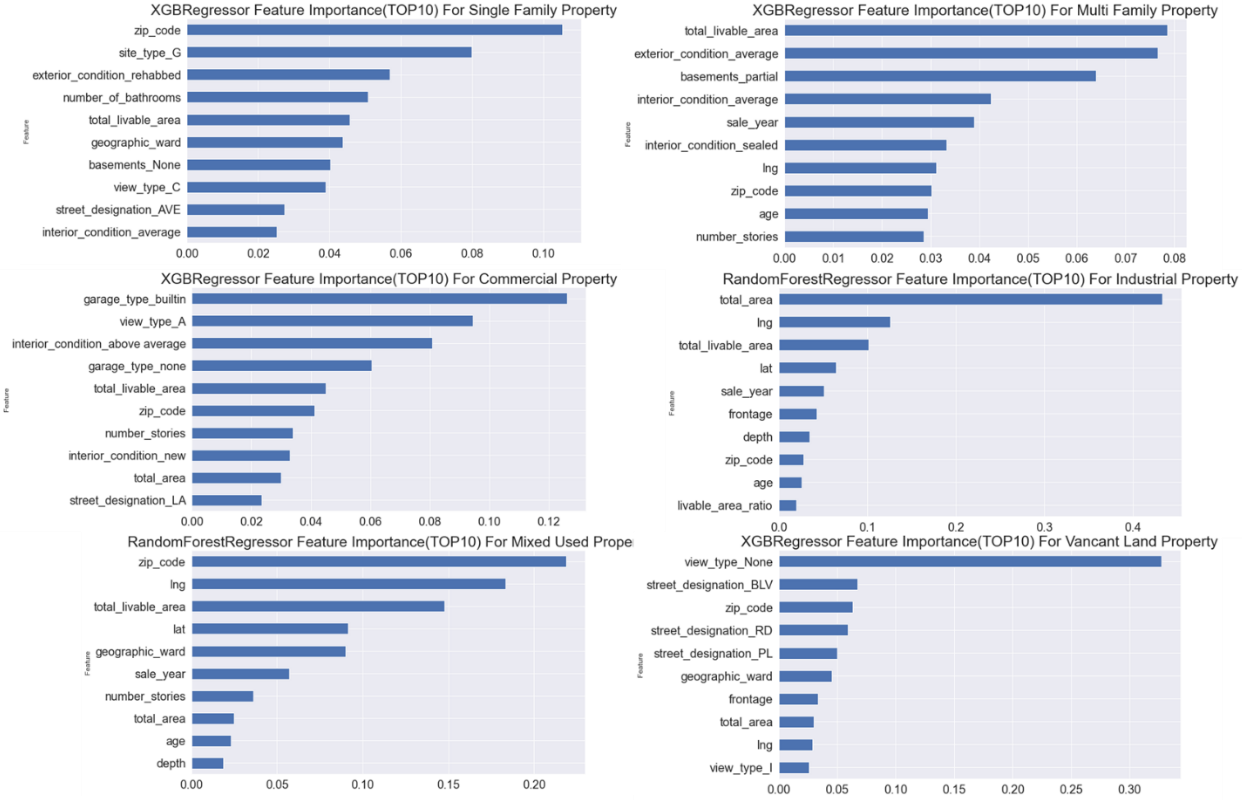


Figure 9: Distribution of True Sale Prices vs. Predictions for 6 Property Types

We plotted true and predicted log sale price distributions. Single and Multi-Family properties showed overlapping distributions, whereas Commercial and Vacant Land showed more visible differences.

We output model interpretation with feature importance measures. Figure 10 shows the top 10 most important features for the best model in each property type.


XGBRegressor Feature Importance(TOP10) For Single Family Property

XGBRegressor Feature Importance(TOP10) For Multi Family Property

XGBRegressor Feature Importance(TOP10) For Commercial Property

RandomForestRegressor Feature Importance(TOP10) For Industrial Property

RandomForestRegressor Feature Importance(TOP10) For Mixed Used Prope

XGBRegressor Feature Importance(TOP10) For Vancant Land Property

Figure 10: Top 10 Features for 6 Property Types

Geographic features were ranked within the top 10 for all property types. Interior condition features (interior condition, number of rooms, and livable area) were also ranked within top 10 for all property types except Vacant Land. Table 10 gives the most important non-geographic feature for each property type.

Table 10: Most Important Non-Geographic Feature for Each Property Type

	Single Family	Multi Family	Commercial	Industrial	Mixed-Use	Vacant Land
Most Important Non-Geographic Feature	Exterior Condition (Rehabbed)	Total Livable Area	Garage Type (Built in)	Total Area	Total Livable Area	View Type (None)

The most important non-geographic features for most property types match the subject matter need of each type's main use. For example, businesses who intent to buy Commercial property utilize garages for shipping and customer parking, and industrial properties must have enough total area to house factories and machinery.

7. Conclusions

The Single-Family property type model had the smallest test RMSE and thus had best generalizability for future predictions. Commercial property type model had the largest test RMSE, and thus had the worst generalizability. The XGBoost and Random Forest models were superior in both validations set and generalizability.

Our Single-Family and Multi-Family models show utility that future home buyers in Philadelphia may find useful. By training a model with all available data, a user may input a desired home's amenities and location to predict the sale price. Government organizations and private companies in Philadelphia may find our model for Mixed, Vacant, Commercial, and Industrial useful, but we recommend this be done with caution due to the larger test RMSE and poorer generalizability.

The poor generalizability of the Commercial and Industrial property types may be a result of the underlying variability of price and smaller data availability. The superior performance of the Single-Family type in contrast may be due to a larger and more robust sample size and tighter distribution in sale price. Future works that aim to predict Commercial and Industrial property types should ensure a generally non-missing feature space or attempt to aggregate datapoints from several cities to increase sample size.

8. Roles

Shusaku Asai:

- Coding: data processing, EDA, preliminary modeling, model selection and hyperparameter tuning (Multi-Family and Industrial properties), GitHub maintenance
- Report: Introduction, Motivation, Results, Conclusions writing
- Video: Transitions, result paraphrasing, introductory slides

Congjun Huang:

- Coding: Pipeline creation, model comparisons/selection, and hyperparameter tuning (Single-Family and Mixed-Use properties)
- Report: Result tabulation, table building, final editing
- Video: Results tabulations, figure adaptations, YouTube uploading

Jingjing Wang:

- Coding: EDA and exploration, pipeline application, model selection and hyperparameter tuning (Commercial and Vacant Land properties), feature importance analysis
- Report: EDA writing, figure editing/captions, proof-reading
- Video: Figure preparation, zoom preparations

Word Count excluding title page, figure captions, and References: 2469

9. References

Project Video: https://www.youtube.com/watch?v=Zv0nwZJKiHA&ab_channel=CongjunHuang
Project GitHub: https://github.com/delashu/Philadelphia-Housing

1. https://www.nar.realtor/newsroom/existing-home-sales-surge-6-7-in-january

2. https://www.prnewswire.com/news-releases/home-prices-in-cities-rise-16-surpassing-suburban-and-rural-price-growth-for-the-first-time-since-before-pandemic-301240469.html
3. https://www.statista.com/statistics/269967/urbanization-in-the-united-states/
4. Rohe, William M., and Mark Lindblad. "Reexamining the social benefits of homeownership after the housing crisis." *Boston: Joint Center for Housing Studies of Harvard University* (2013).
5. Rohe, William M., Roberto Quercia, and Shannon Van Zandt. "The social-psychological effects of affordable homeownership." *Chasing the American dream: New perspectives on affordable homeownership* (2007): 304-325.
6. Fuller, Stephen S. "Economic impacts of commercial real estate." *NAIOP Research Foundation* (2021).
7. McCormack, Edward, G. Scott Rutherford, and Martina G. Wilkinson. "Travel impacts of mixed land use neighborhoods in Seattle, Washington" Transportation Research Record 1780.1 (2001): 25-32.
8. Cowden, Chad, Frank J. Fabozzi, and Abdolreza Nazemi. "Default prediction of commercial real estate properties using machine learning techniques." *The Journal of Portfolio Management* 45.7 (2019): 55-67.
9. Mohd, Thuraiya, Syafiqah Jamil, and Suraya Masrom. "Machine learning building price prediction with green building determinant." *IAES International Journal of Artificial Intelligence* 9.3 (2020): 379.
10. Ho, Winky KO, Bo-Sin Tang, and Siu Wai Wong. "Predicting property prices with machine learning algorithms." *Journal of Property Research* 38.1 (2021): 48-70.
11. Park, Byeonghwa, and Jae Kwon Bae. "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data." *Expert systems with applications* 42.6 (2015): 2928-2934.
12. Shahhosseini, Mohsen, Guiping Hu, and Hieu Pham. "Optimizing ensemble weights for machine learning models: A case study for housing price prediction." *INFORMS international conference on service science*. Springer, Cham, 2019.
13. Varma, Ayush, et al. "House price prediction using machine learning and neural networks." *2018 second international conference on inventive communication and computational technologies (ICICCT)*. IEEE, 2018.
14. Das, Sarkar Snigdha Sarathi, et al. "Boosting house price predictions using geo-spatial network embedding." *Data Mining and Knowledge Discovery* 35.6 (2021): 2221-2250.
15. Usman, Hamza, Mohd Lizam, and Burhaida Burhan. "A REVIEW OF SPATIAL ECONOMETRICS IN EXPLICIT LOCATION MODELLING OF COMMERCIAL PROPERTY MARKET." *PLANNING MALAYSIA* 19 (2021).
16. https://www.opendataphilly.org/dataset/opa-property-assessments
17. https://metadata.phila.gov/#home/datasetdetails/5543865f20583086178c4ee5/representationdetails/55d624fdad35c7e854cb21a4/?view_287_page=1
18. Jengei Hong, Heeyoul Choi and Woo-Sung Kim. “A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea” International Journal of Strategic Property Management (2020) 24(3):1-13
19. Quang Truong, Minh Nguyen, Hy Dang and Bo Mei. “Housing price prediction via improved machine learning techniques” Procedia Computer Science 174 (2020), pp.433-442.

20. Lim, Wan Teng et al. “Housing price prediction using neural networks.” 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) (2016): 518-522.
21. Kokasih, Marco Febriadi, and Adi Suryaputra Paramita. "Property Rental Price Prediction Using the Extreme Gradient Boosting Algorithm." International Journal of Informatics and Information Systems 3.2 (2020) pp.54-59.