

Purwadhika

JCDS 1202

Final Project

Matplotlib Team



Hello!

We are Matplotlib Team

In this final project, we were assigned to make a regression machine learning model using DC Properties dataset.

Table of Contents

- BACKGROUND & PROBLEM IDENTIFICATION
- DATA UNDERSTANDING & DATA EXPLORATION
- MODELING & EVALUATION
- DEPLOYMENT
- FUTURE WORKS

1. Background & Problem Identification

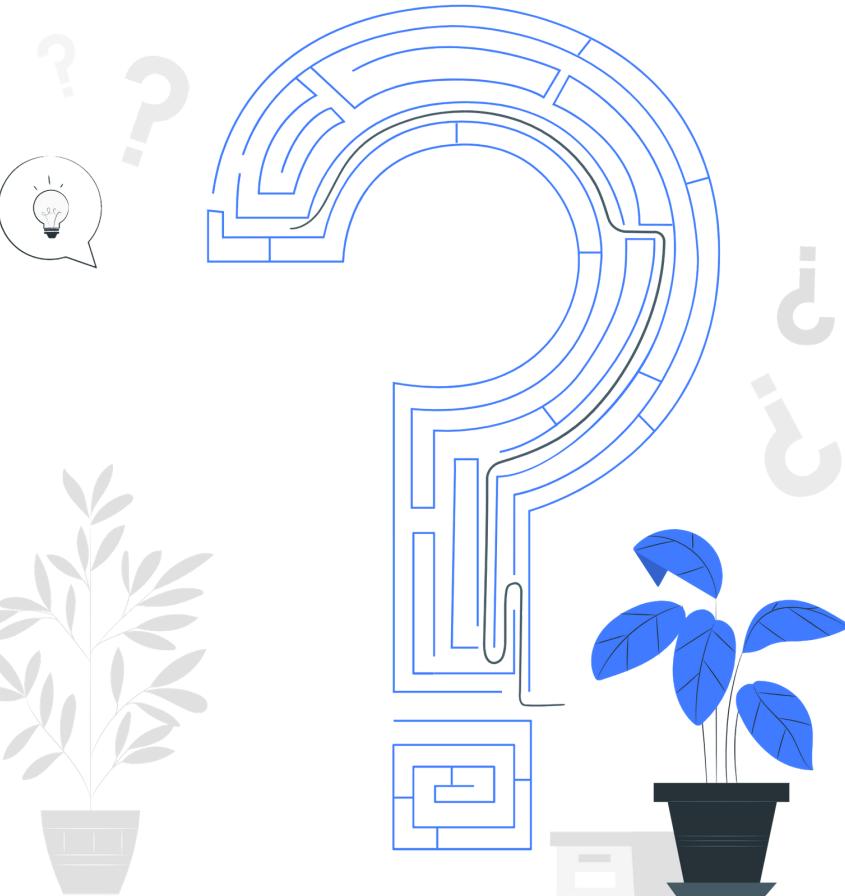


Background

We position ourselves as a Data Scientist Team working at MPL Bank located in Washington DC, USA.

We were assigned to work on a project to develop a Machine Learning (ML) solution for Underwriter Team of MPL Bank. We will help the Underwriter Team to make an improvement in their process of underwriting, specifically in the process of property appraisal and valuation.

Problem Identification



Problem
Definition

Business
Objective

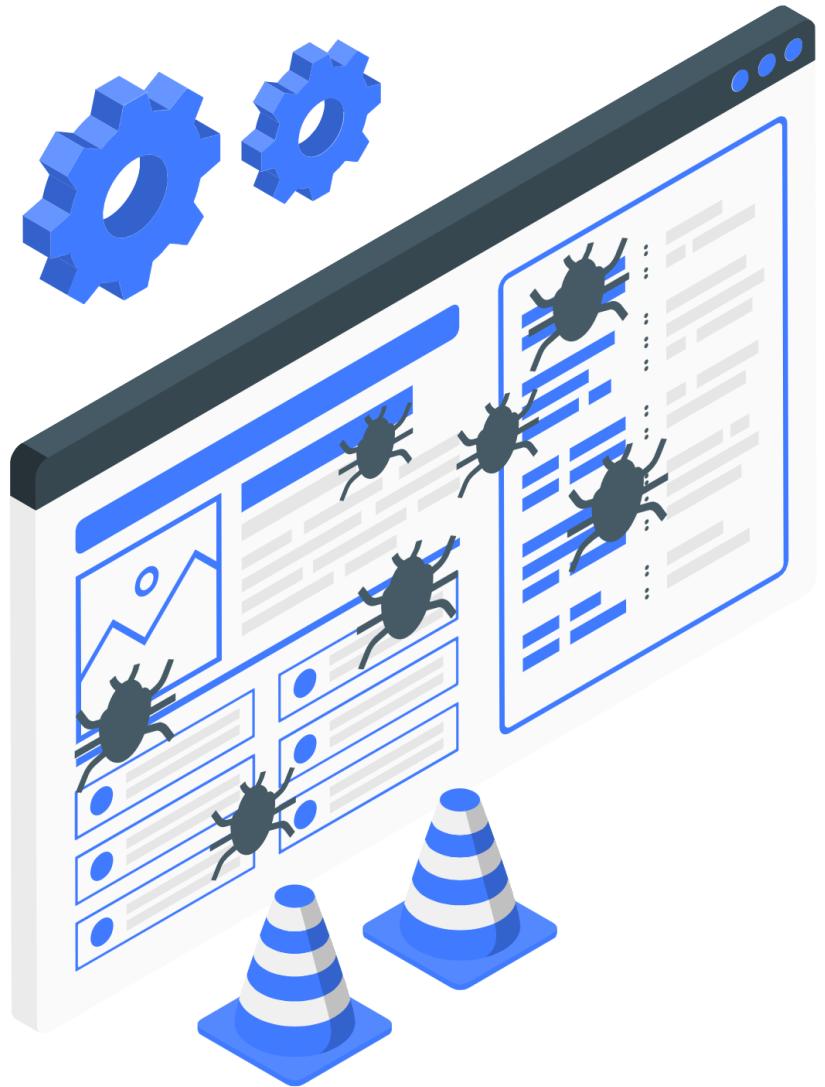
Data
Requirements

Analytic
Approach

Action

Value

Problem Definition



**Risk of Fraud & Erroneous
Appraisal by AMC**

**Difference between the
Agreed Offer and the Actual
Property Valuation**

**Improving Accuracy in
Appraisal Evaluation Process**

Why?



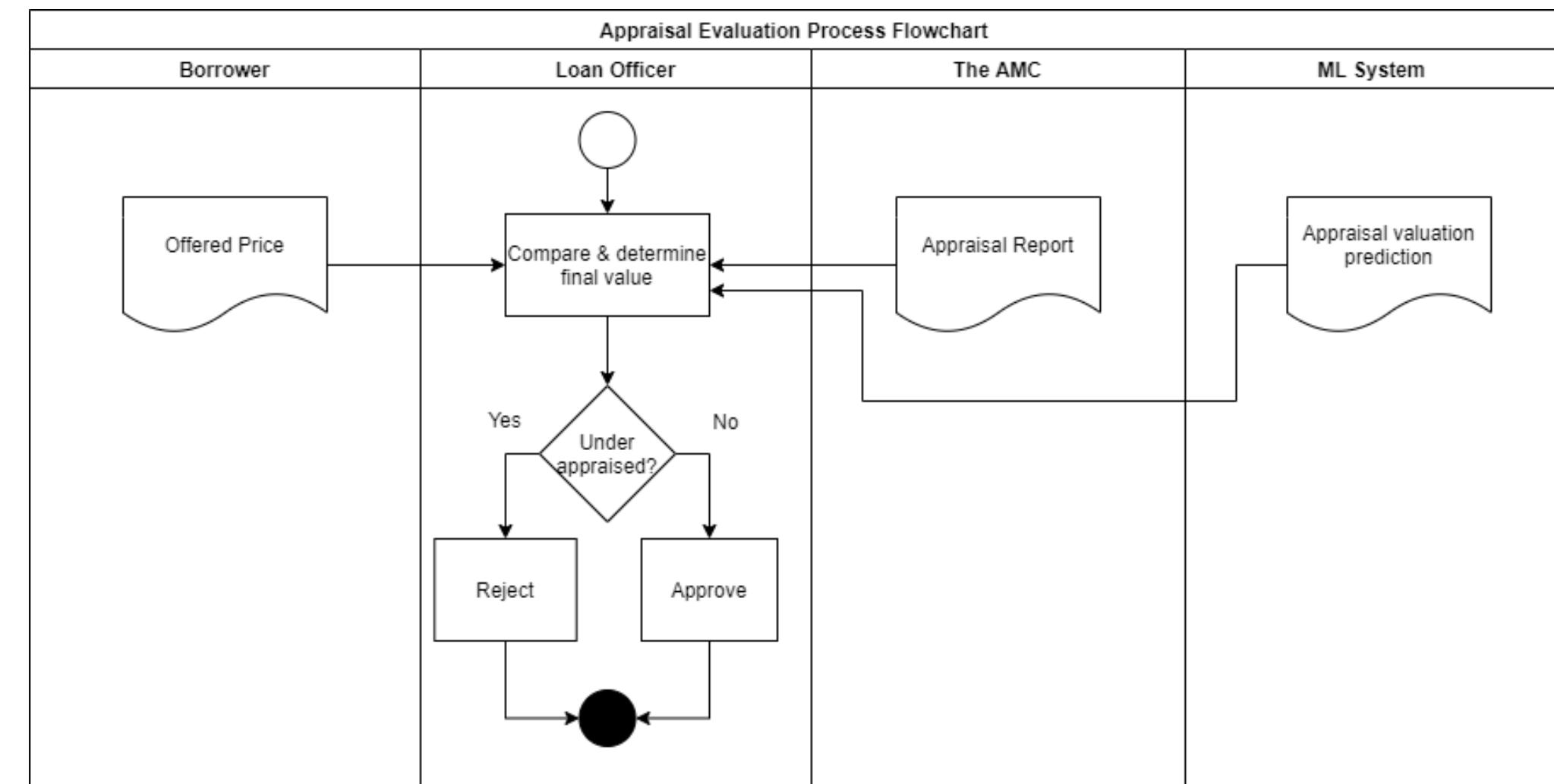
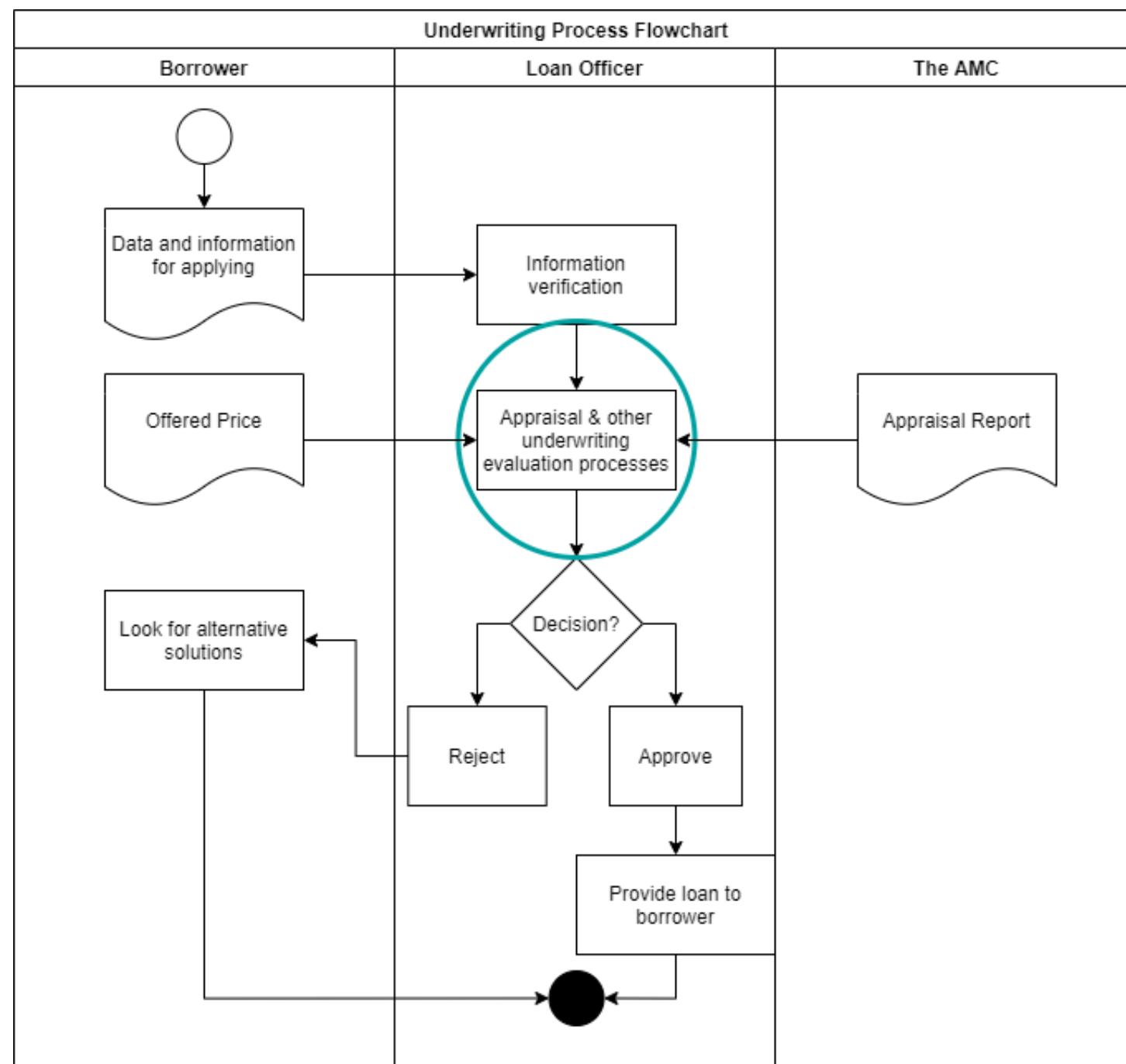
Appraisal process directly affects company's revenue.

The result would determine whether to give loan to a borrower.

A property's value is appraised by AMC and checked by MPL Bank's internal appraisal team.

This is where we come in to help improving the process of property appraisal evaluation.

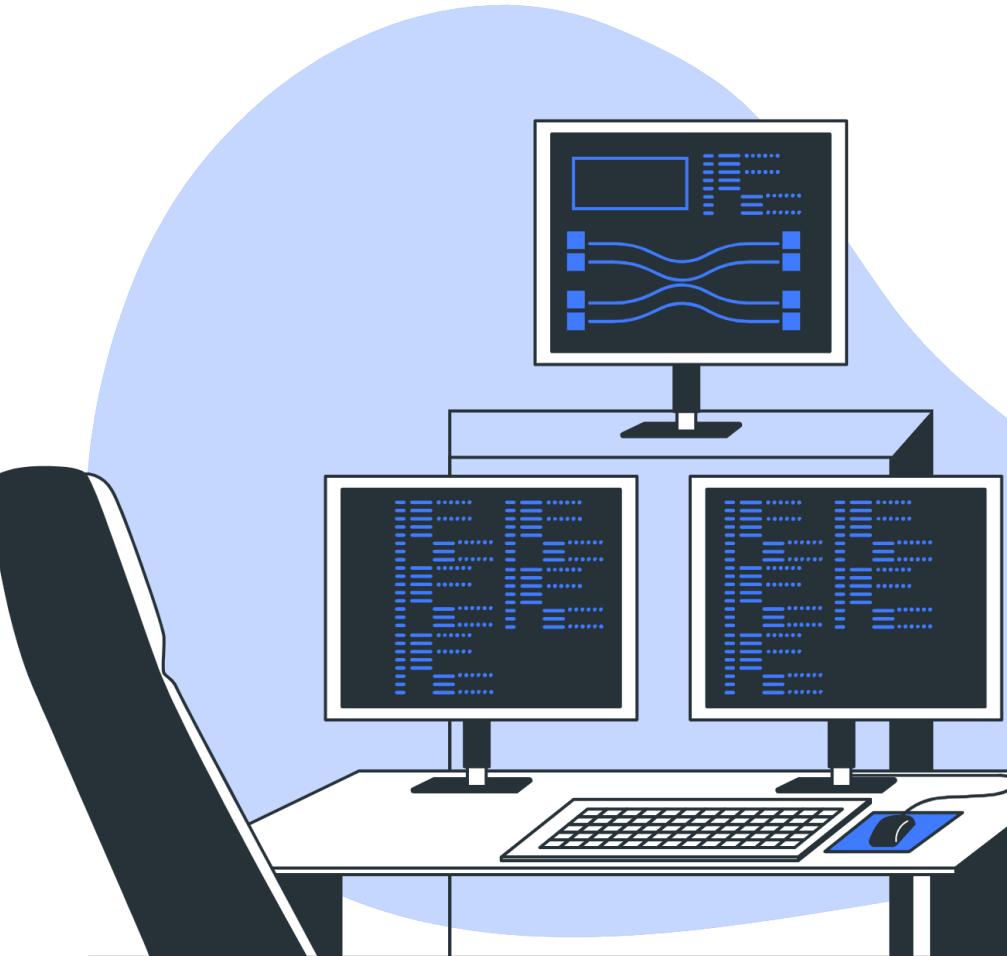
Appraisal Evaluation Process



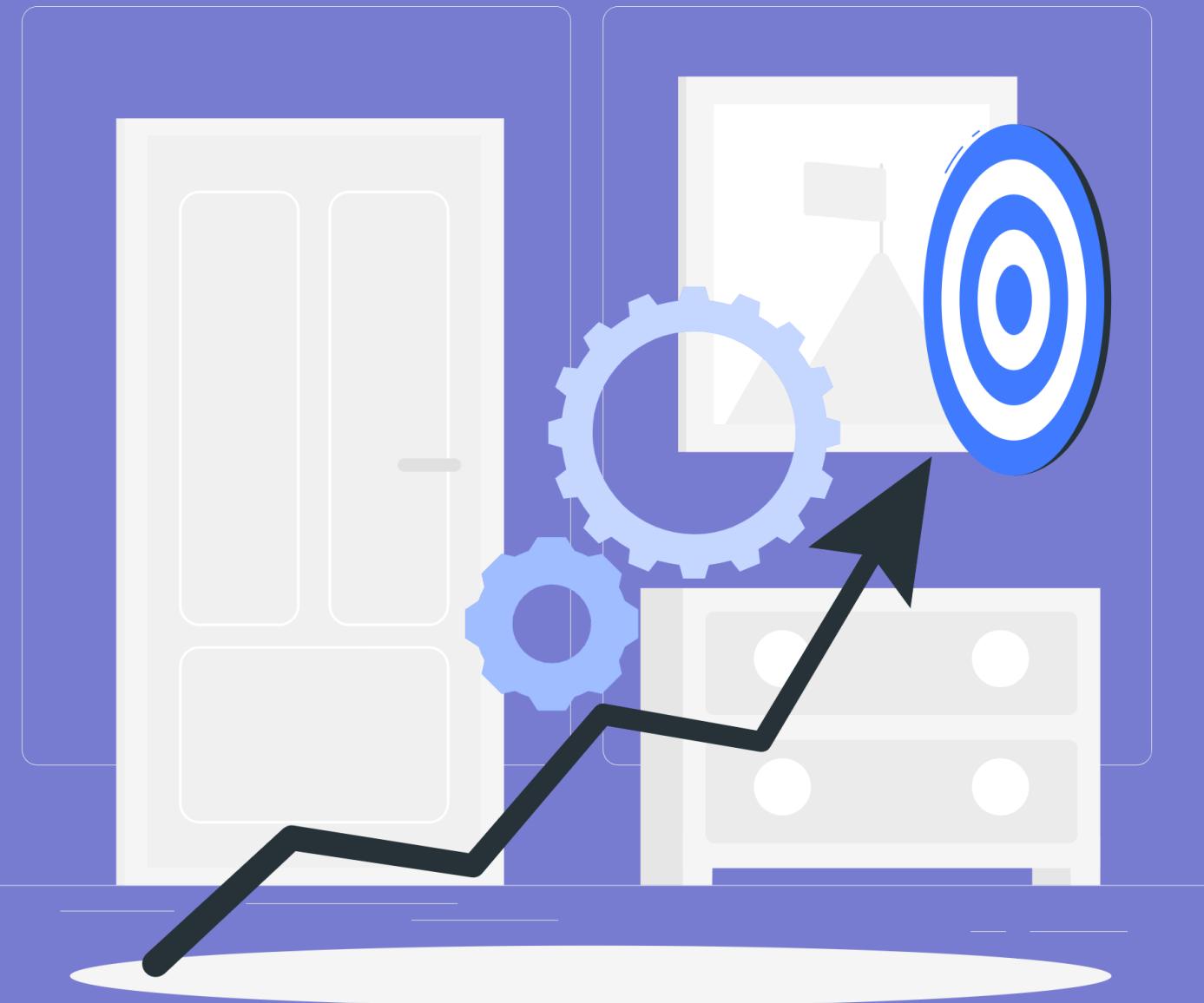
Expected Output

A system that can make an estimation of an accurate and reasonable value (price) for a property based on the aspects of the property by using Machine Learning.

Notes : Due to the limitation of our time budget, we limit the capability of our model in this project to predict an output only for properties with grade lower than Exceptional, since Exceptional properties have a price range that is very different than the rest of properties with other grades.



Business Objectives



Maximize Profit

by helping underwriter team to make the right decision whether to give a loan or not, with an optimal amount.

Minimize Loss

originated from fraud and erroneous valuation.

Notes: In this project we were asked to reach the Mean Absolute Error (MAE) metrics at most 12% to the median of property price.

Data Requirements

- The features of the property (e.g., gross building area, the number of rooms, the number of bedrooms, etc)
- The condition of the property
- The location of the property



Analytic Approach



Machine Learning Technique

Target : Continuous Value

Technique : Supervised Learning - Regression

Type : Model-based Learning

Risk

- The actual value of the property > prediction value
→ Reject giving loan and resulting in loss of potential borrower.
- The actual value of the property < prediction value (the model gives an under appraised value)
→ Suffer loss when the borrower is unable to pay back.

Performance Measures

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- The Coefficient of Determination (R^2)
- With MAE as the vital metrics as agreed with Underwriter Team.

ACTION

The business user can utilize the prediction result by comparing it with the appraisal value given by the AMC to determine a reasonable property value.

VALUE

To improve the underwriting process by providing a good appraisal prediction model and thus helps business user to make the right business decision, resulting in maximized profit and minimized loss.

2. Data Understanding & Data Exploration



DATA ACCESS & PRIVACY

The data was downloaded from [Kaggle](#).

All data is available at [Open Data D.C.](#).

The residential and address point data is managed by the [Office of the Chief Technology Officer](#).

Distribution Liability: [Data Terms and Conditions](#).

RESIDENTIAL DATA SHAPE

- 106,696 Rows
- 49 Columns

Data
Collection

Data Description

NUMERICAL

- BATHRM : Number of full bathroom
- HF_BATHRM : Number of half bathroom
- NUM_UNITS : Number of units
- ROOMS : Number of rooms
- BEDRM : Number of bedrooms
- GBA : Gross building area (in sqft)
- KITCHENS : Number of kitchens
- FIREPLACES : Number of fireplaces
- LANDAREA : Land area (in sqft)
- LATITUDE : Latitude
- LONGITUDE : Longitude
- AYB (ayb_age) : The earliest time the main portion of the building was built
- EYB (eyb_age) : The year an improvement was built more recent than actual year built
- PRICE : Price (in USD)

Data Description

CATEGORICAL - NOMINAL

- HEAT : Heating system
- AC : AC availability
- QUALIFIED : Government's criteria
on whether the sale is
representative of market value
- USECODE : Use code
- STYLE : Style
- ASSESSMENT_NBHD :
Neighborhood ID
- WARD : Ward
- EXTWALL : Exterior wall

- INTWALL : Interior wall
- QUADRANT : City quadrant (NE, SE,
SW, NW)

CATEGORICAL - ORDINAL

- GRADE : Grade
- CNDTN : Condition
- STRUCT : Structure
- ROOF : Roof type
- SALEDATE : Date of most recent
sale

Data Pre-Processing

Drop Unused Features

Unnamed: 0, SALE_NUM, CMPLX_NUM, LIVING_GBA, X, Y, ASSESSMENT_SUBNBHD, SOURCE, CITY, STATE, NATIONALGRID, GIS_LAST_MOD_DTTM, CENSUS_BLOCK, YR_RMDL, STORIES, FULL_ADDRESS

Fill Missing Values

AYB, QUADRANT, AC, NUM_UNITS

Drop Rows with Missing Values

SALEDATE, KITCHENS, ROOMS, GRADE, HEAT, BATHRM

Create New Features

- ayb_age = Present Year - AYB
- eyb_age = Present Year - EYB

Data Pre-Processing

Drop Unusual/ Erroneous Data

- BEDRM > ROOMS
- SALEYEAR <= 1991
- EYB < AYB

Merging Values with Very Few Occurrence

- CNDTN (Default → Good)
- GRADE (Low Quality → Fair Quality)

Excluding Outliers

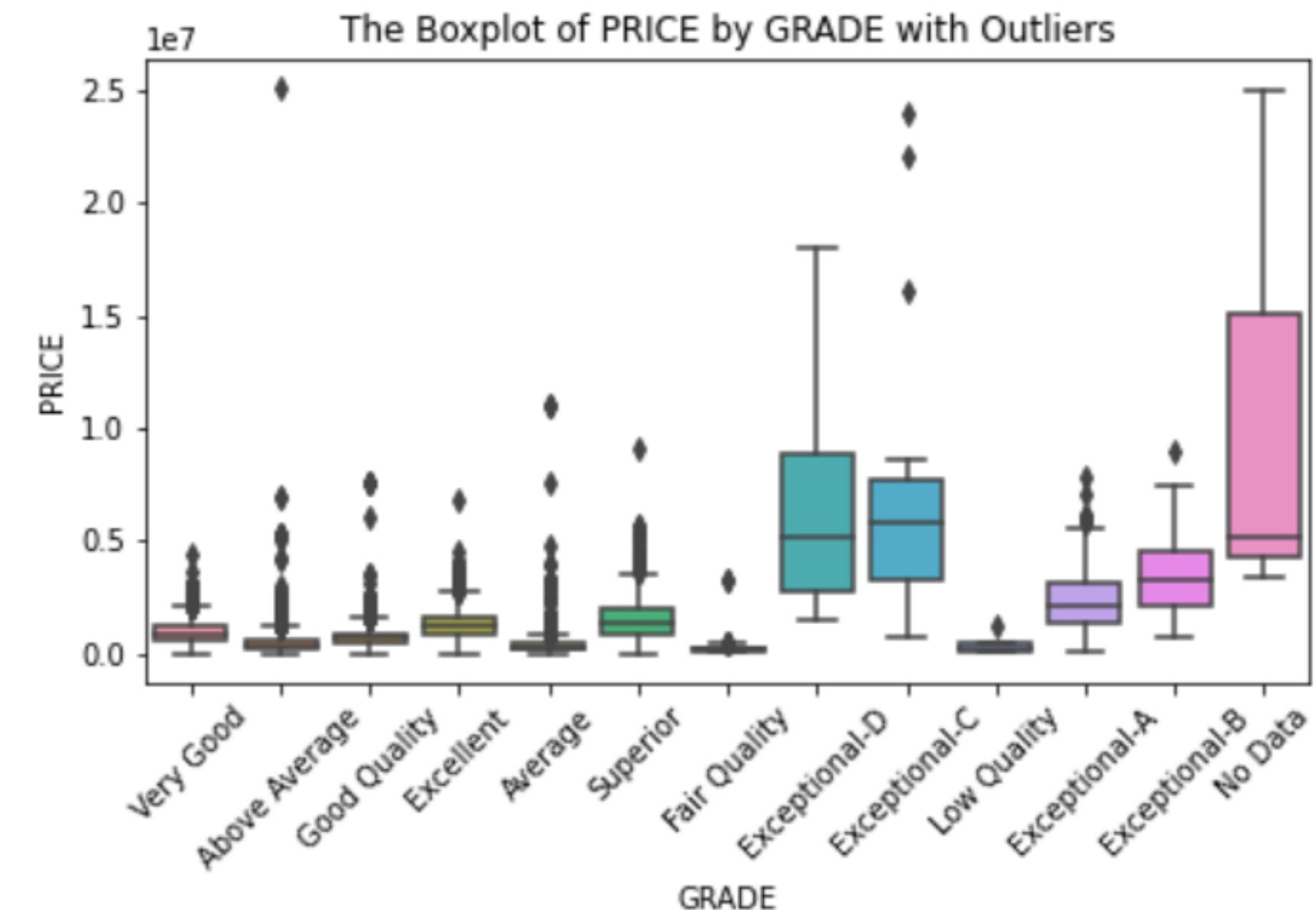
KITCHENS, PRICE, Price based on CNDTN & GRADE

Re-classify Features

CNDTN, STRUCT, ROOF

Exceptional Properties Exclusion

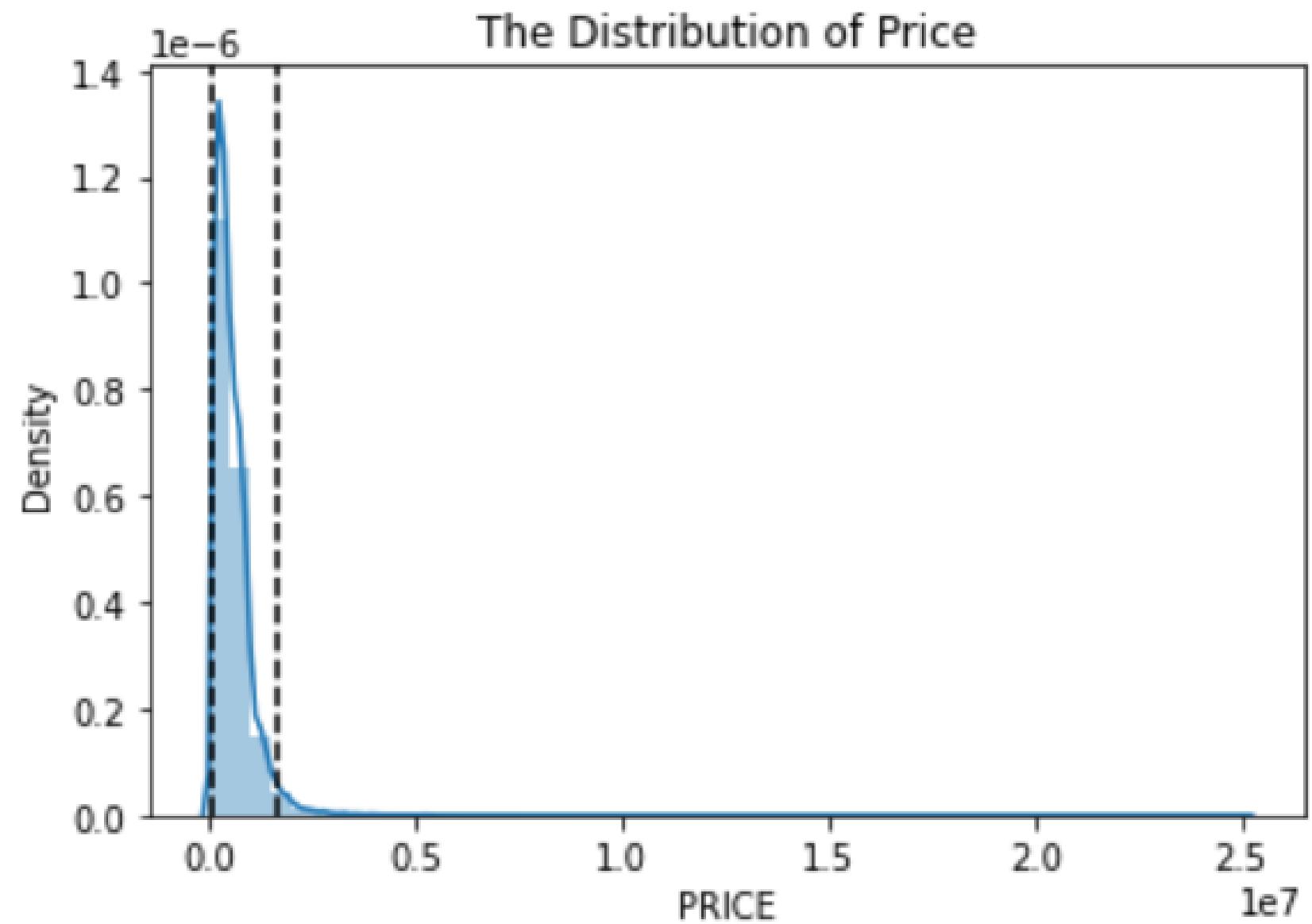
We limit the capability of our model in this project to predict an output only for properties with grade lower than Exceptional since Exceptional properties have a price range that is very different than the rest of other grades and another model needs to be built specifically for them.



Identifying Outliers in Price

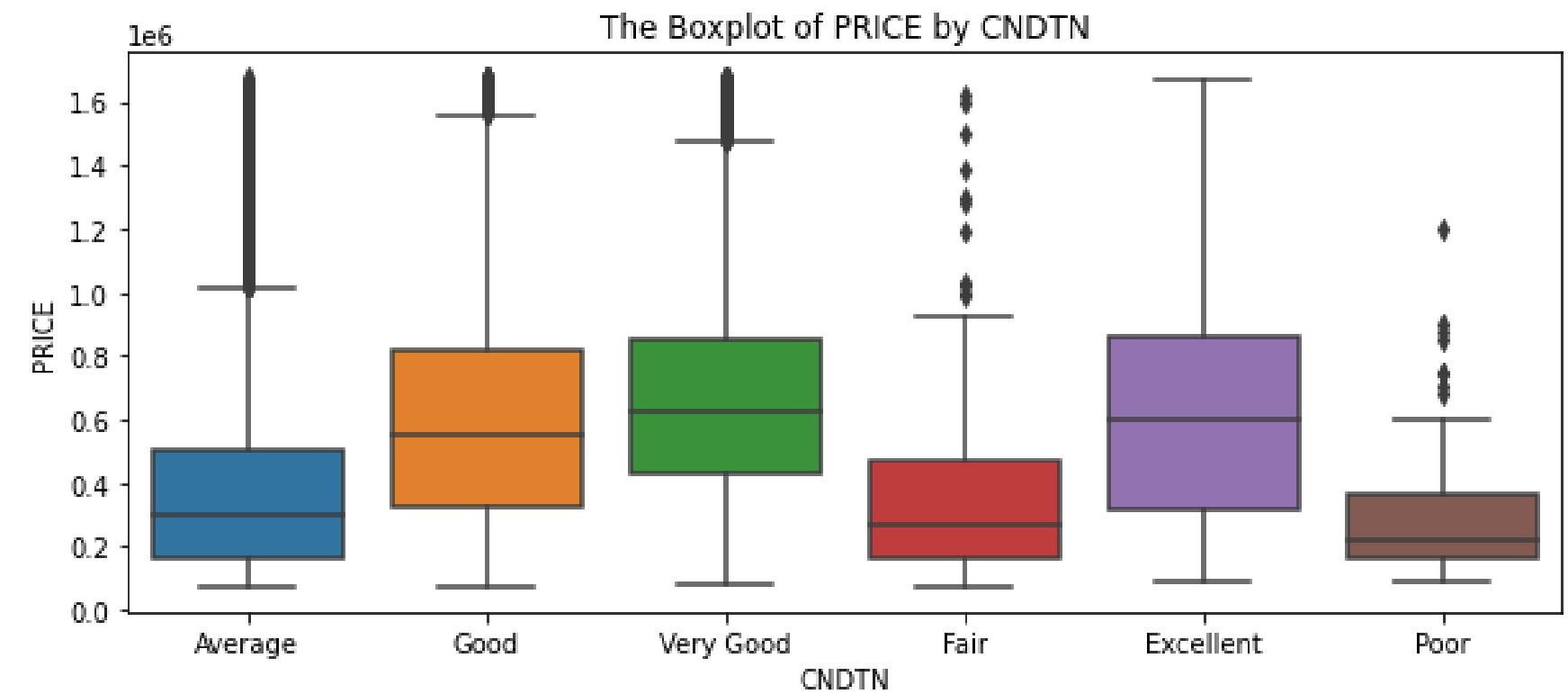
- Erroneous data with very low price (e.g., \$1.0) or extremely expensive price.
- Kaggle : "Property prices are very high in Washington DC, averaging \$647,000 in 2017."
- Drop data with $\text{PRICE} \leq 25\text{th percentile}$ or $\geq 97.5\text{th percentile}$ to assume 95% confidence interval.

Please note that we did not perform z-test due to the power of the distribution. Our PRICE distribution is extremely right-skewed as such if we use z-test, data that will be considered as outliers are only those with high price while data with very low price (e.g., \$1.0) will still be considered as non-outliers.



Re-classify CNDTN

There are some overlap between categories, but between 'Fair, Poor, Average' vs 'Good, Very Good, Excellent' they differ quite significantly; as such CNDTN still could potentially be a good predictor of price.



Median Price: \$435,000

We re-classify the categories in CNDTN into 2 groups :

- 0 : Poor (Under median)
- 1 : Good (Above median)

Re-classify STRUCT & ROOF

Median Price : \$435,000

We re-classify the categories
in STRUCT and ROOF into 2
groups :

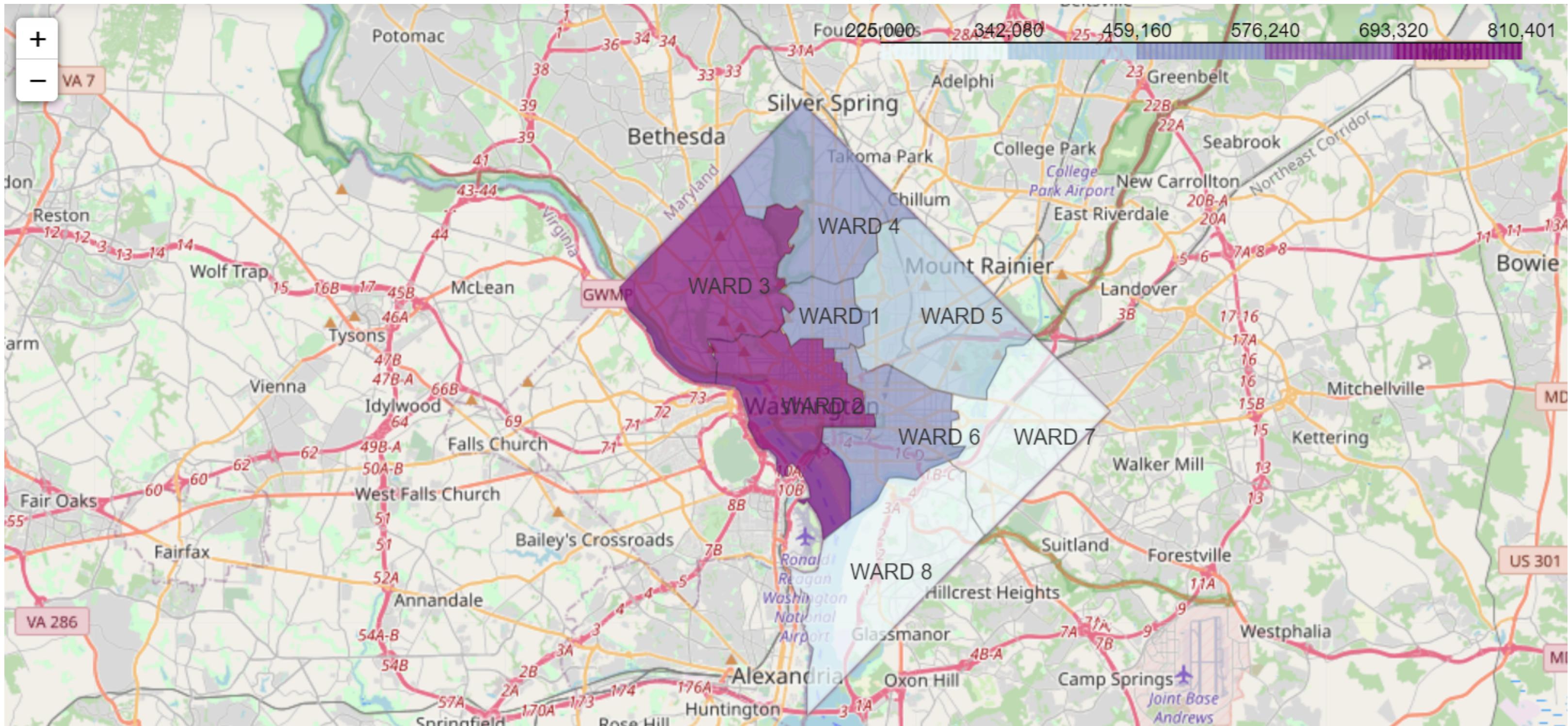
- 0 : Under median
- 1 : Above median

| | STRUCT | median_PRICE | total |
|---|---------------|--------------|-------|
| 0 | Semi-Detached | 275000.0 | 6107 |
| 1 | Multi | 310500.0 | 2082 |
| 2 | Town Inside | 349312.5 | 146 |
| 3 | Town End | 383730.0 | 61 |
| 4 | Row End | 432090.0 | 5145 |
| 5 | Single | 489000.0 | 11266 |
| 6 | Row Inside | 490000.0 | 17565 |
| 7 | Default | 625000.0 | 3 |

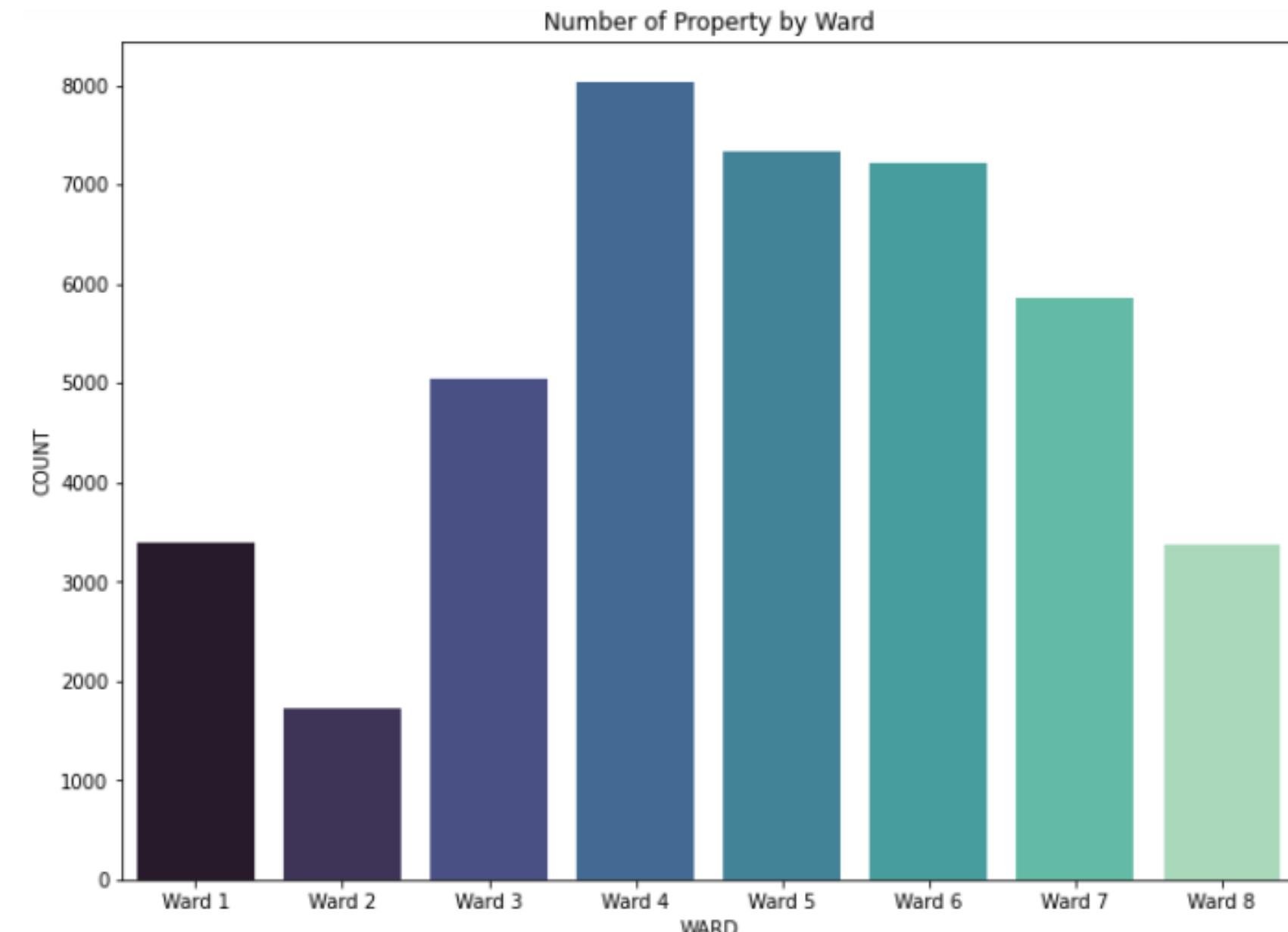
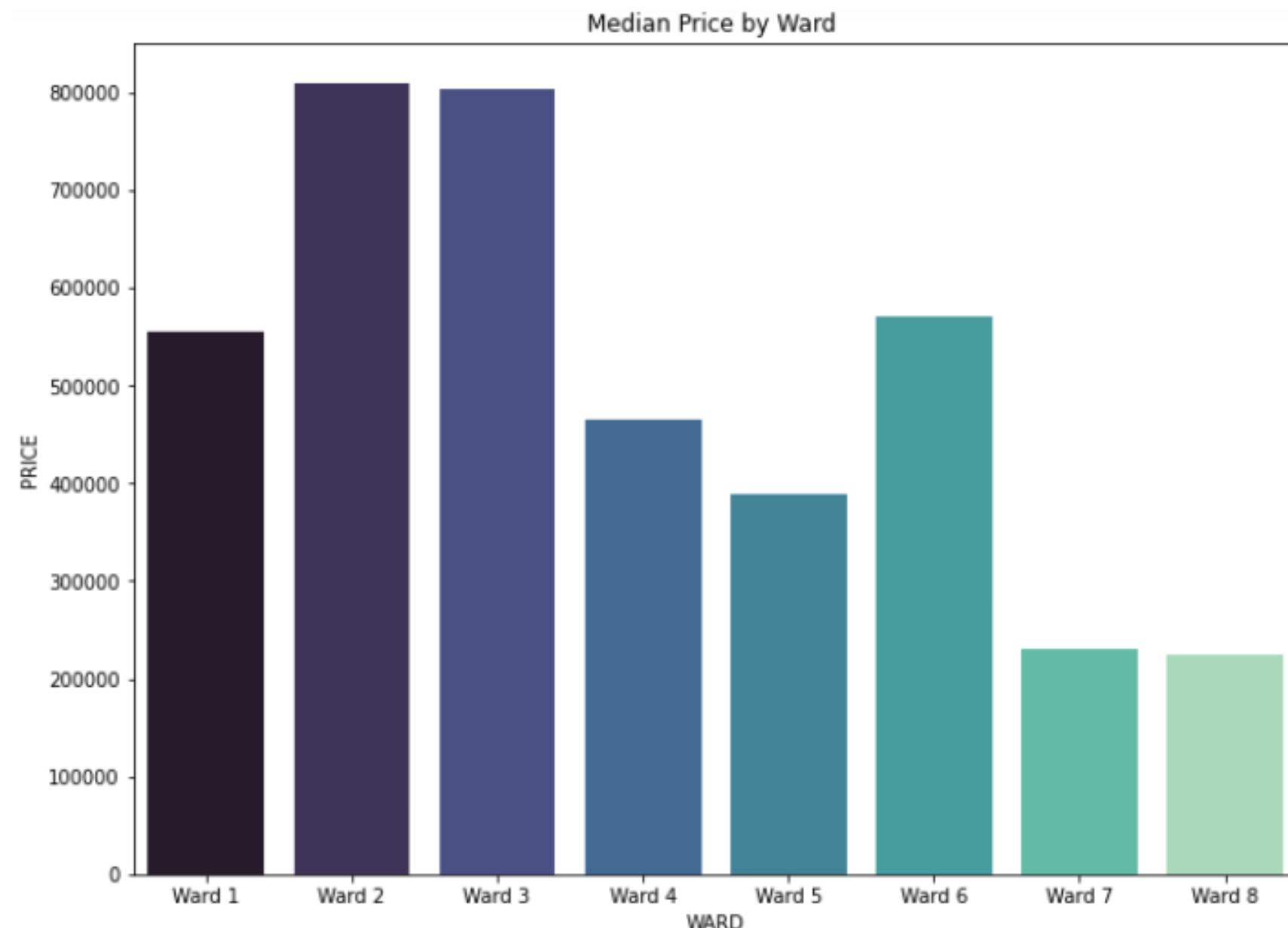
| | ROOF | median_PRICE | total |
|----|----------------|--------------|-------|
| 0 | Concrete | 299900.0 | 1 |
| 1 | Comp Shingle | 359000.0 | 12051 |
| 2 | Typical | 373750.0 | 66 |
| 3 | Built Up | 376000.0 | 13117 |
| 4 | Composition Ro | 410075.0 | 48 |
| 5 | Metal- Pre | 410749.5 | 96 |
| 6 | Shake | 480000.0 | 257 |
| 7 | Metal- Sms | 500000.0 | 12016 |
| 8 | Concrete Tile | 512500.0 | 2 |
| 9 | Water Proof | 594000.0 | 4 |
| 10 | Shingle | 649435.5 | 168 |
| 11 | Clay Tile | 649900.0 | 191 |
| 12 | Slate | 700000.0 | 3649 |
| 13 | Neopren | 724110.0 | 699 |
| 14 | Wood- FS | 762500.0 | 2 |
| 15 | Metal- Cpr | 764250.0 | 8 |

Insights

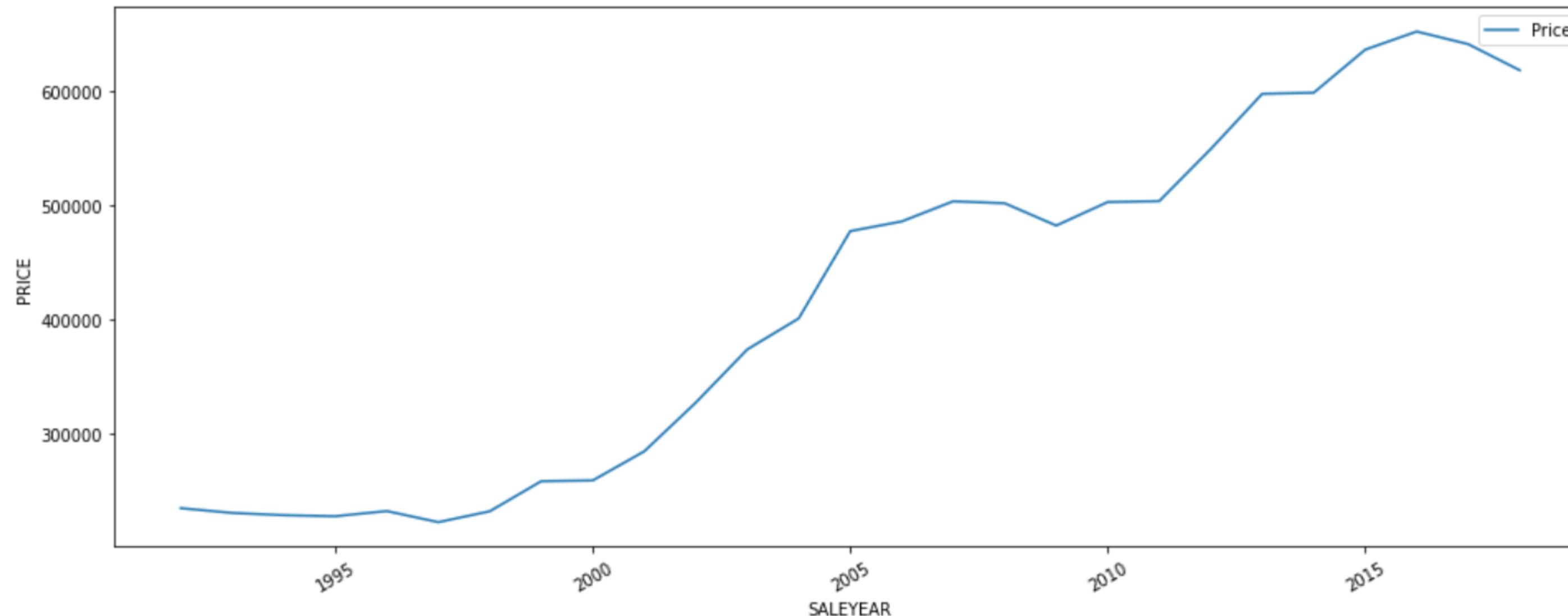
Map by Ward



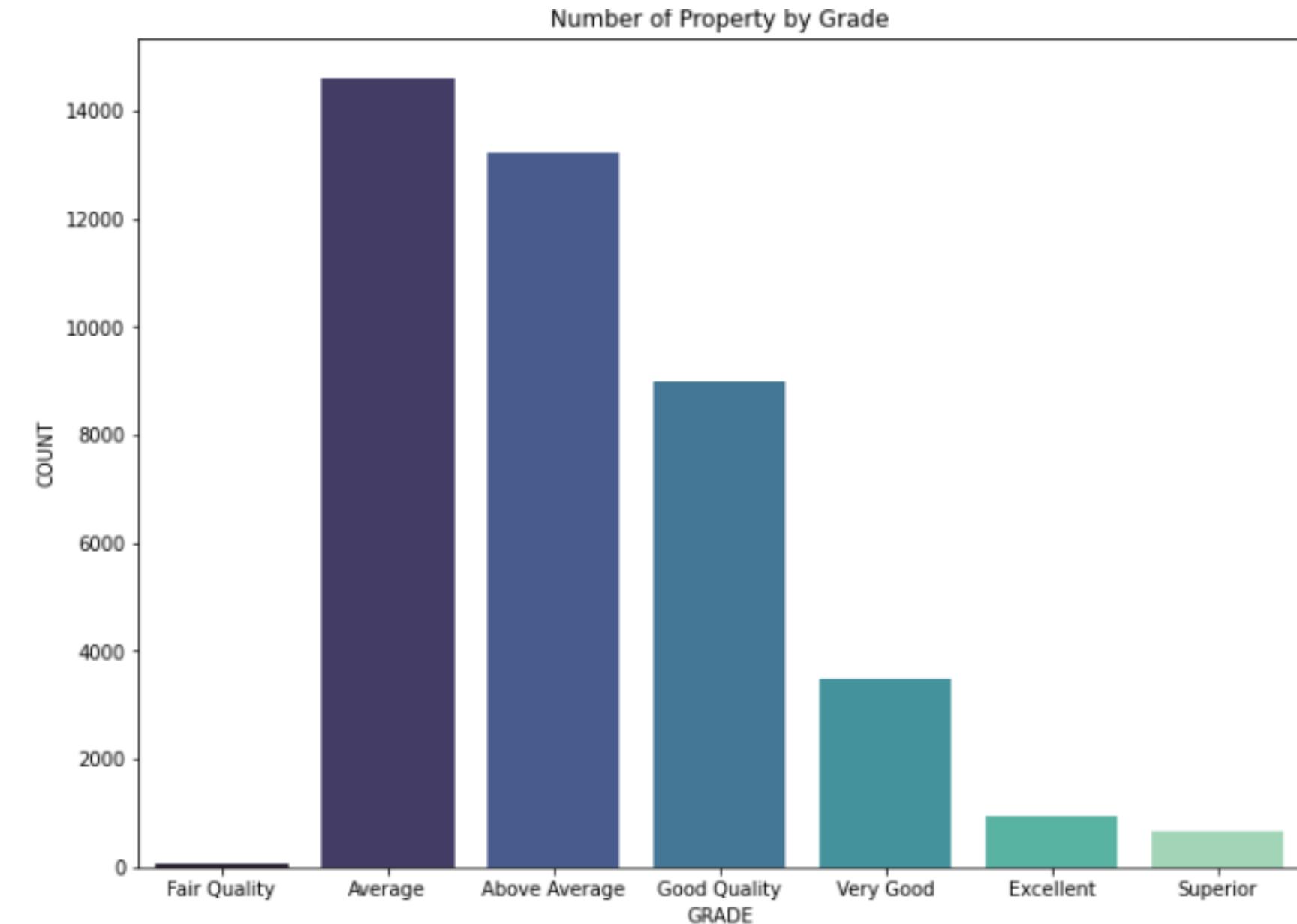
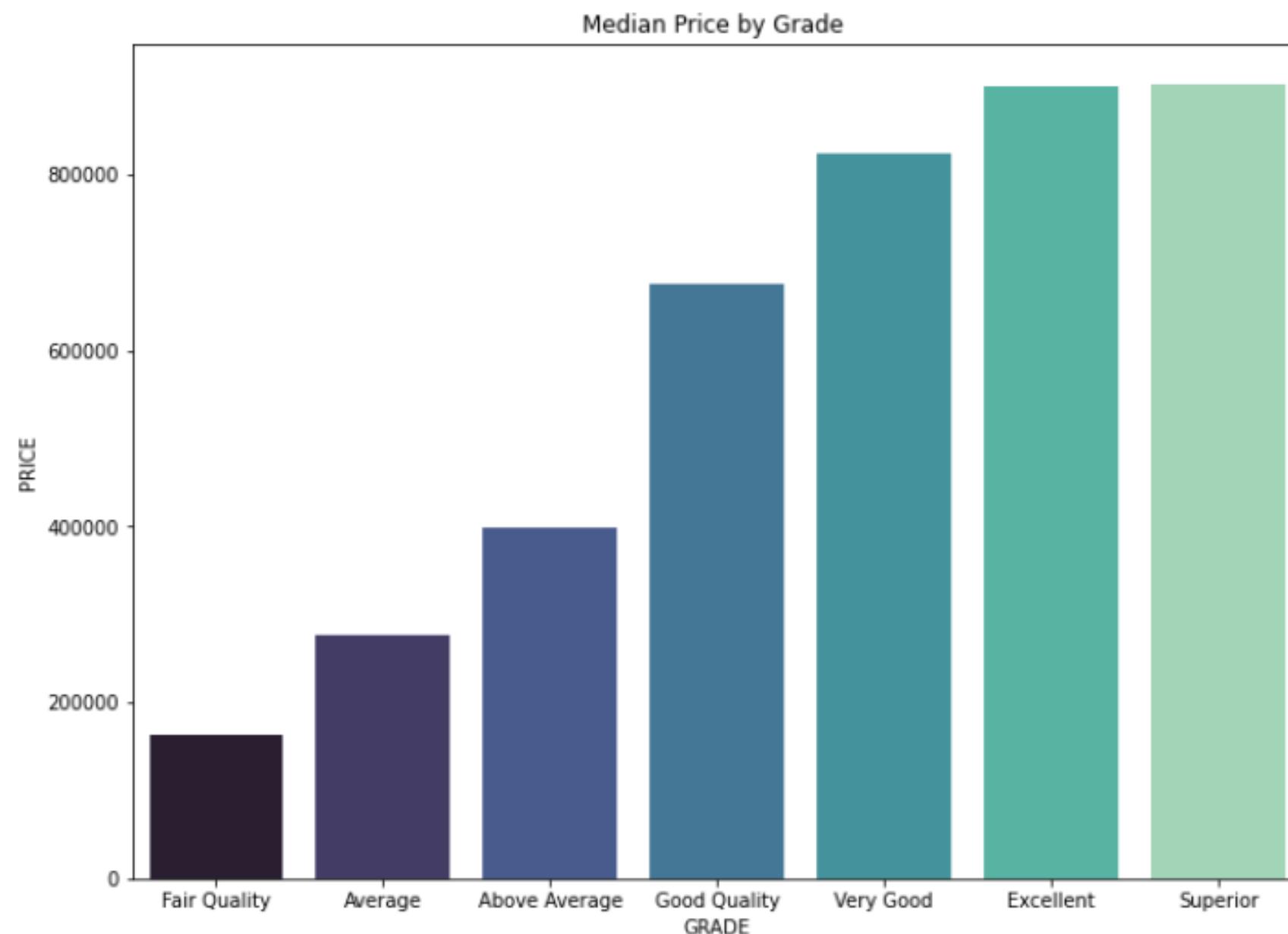
Median Price & Number of Property by Ward



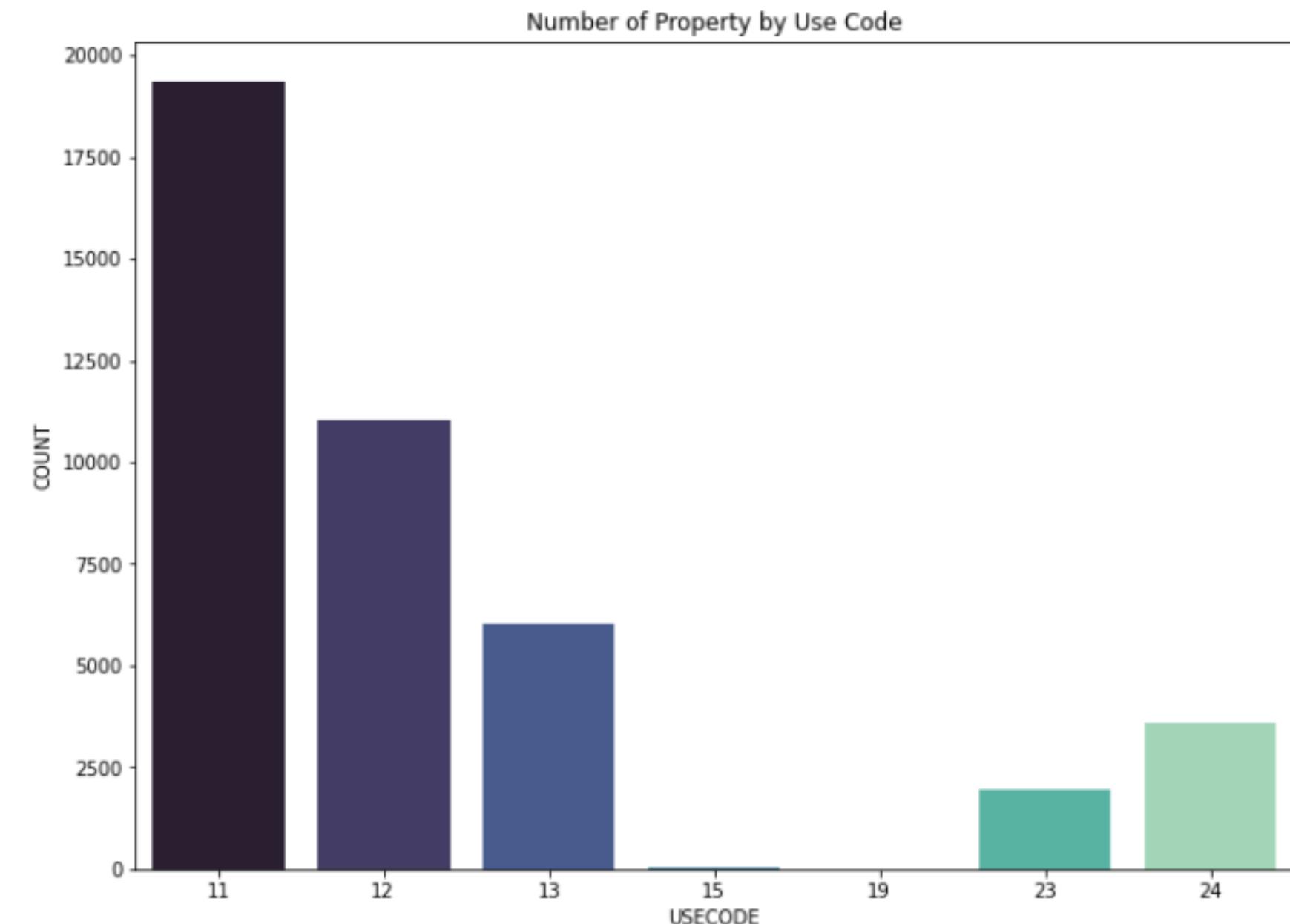
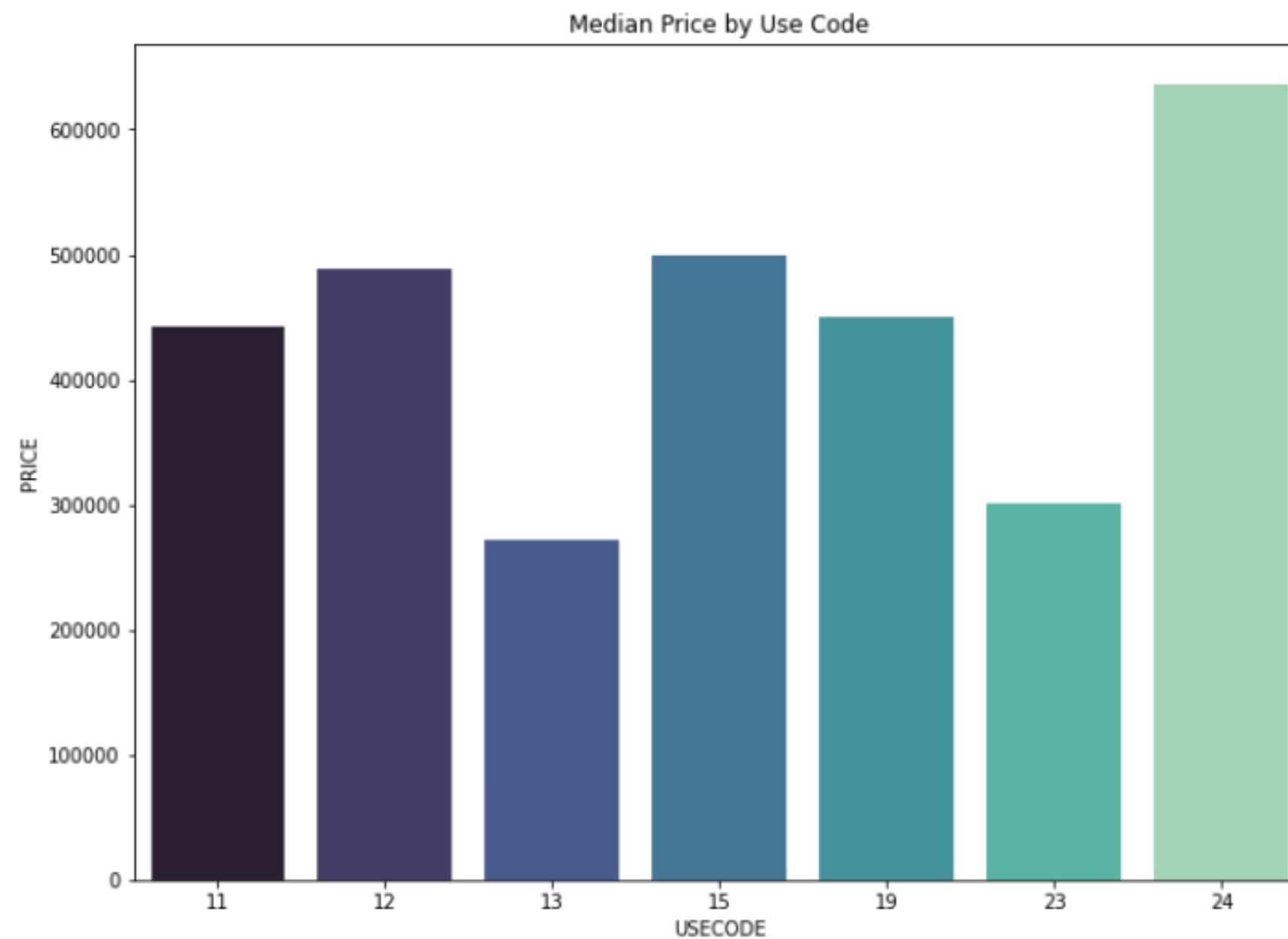
Median Sale Price per Year



Median Price & Number of Property by Grade



Median Price & Number of Property by Use Code



3. Modeling & Evaluation



Modeling Steps

- Encoding categorical features and scaling numerical features
- Training & evaluating different regression models

- Hyperparameter tuning with GridSearchCV to get best parameters of the best model

We choose the best model based on the R^2 score, Mean Absolute Error (MAE) and resource efficiency.

Model Assessment

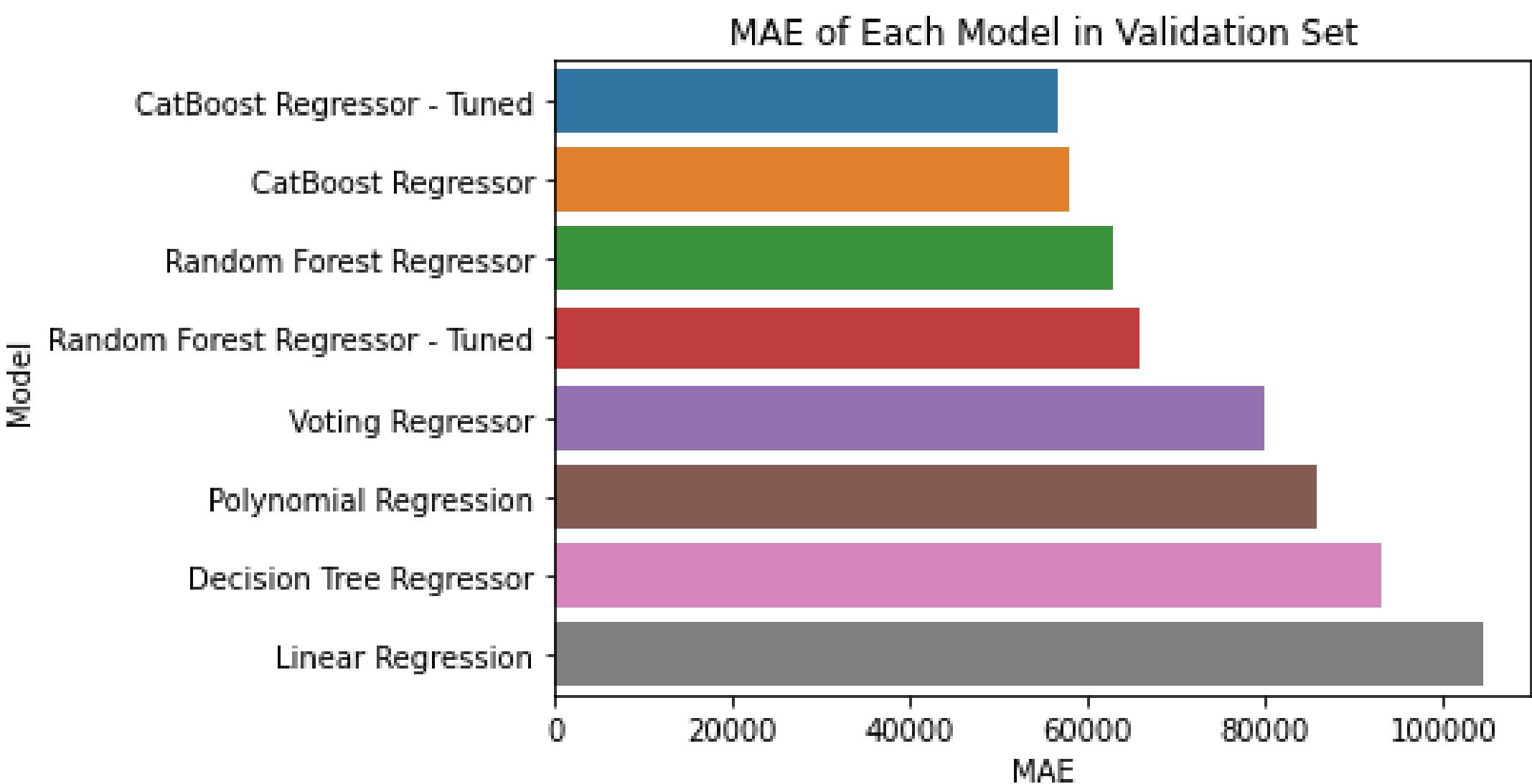
CatBoost Regressor - Tuned has the best R² score and the lowest MAE in the validation set.

| | Model | Set | MSE | RMSE | MAE | R2 |
|----|---------------------------------|------------|--------------|---------------|---------------|----------|
| 1 | Linear Regression | Validation | 1.902704e+10 | 137938.554633 | 104536.593792 | 0.803719 |
| 3 | Polynomial Regression | Validation | 1.397333e+10 | 118208.844948 | 85763.947553 | 0.855853 |
| 5 | Decision Tree Regressor | Validation | 2.010742e+10 | 141800.627564 | 92944.701611 | 0.792574 |
| 7 | Random Forest Regressor | Validation | 9.229200e+09 | 96068.722950 | 62796.015937 | 0.904793 |
| 9 | Random Forest Regressor - Tuned | Validation | 9.932636e+09 | 99662.611843 | 65918.276415 | 0.897536 |
| 11 | Voting Regressor | Validation | 1.242826e+10 | 111482.094846 | 79979.099662 | 0.871792 |
| 13 | CatBoost Regressor | Validation | 7.854494e+09 | 88625.585972 | 58064.522230 | 0.918974 |
| 15 | CatBoost Regressor - Tuned | Validation | 7.863497e+09 | 88676.363136 | 56813.042752 | 0.918881 |

Model Assessment

CatBoost Regressor - Tuned has the lowest MAE compared to other models.

We managed to improve the model performance by reducing the MAE by 46% from the first base model, Linear Regression (\$104,536.59), relative to the final model CatBoost Regressor - Tuned (\$56,813.04).



Final Model Evaluation

We chose CatBoost Regressor with tuned parameters as our prediction model.

Train Set

MSE: 5902490242.367122
RMSE: 76827.66586567057
MAE: 46975.76563974304
R-squared: 0.9392950310981597

Test Set

MSE: 10815149410.215195
RMSE: 103995.91054563249
MAE: 63782.41628337468
R-squared: 0.9073119043829377

Cross validation test for our best model to check how consistent the model and results are when measurement is repeated.

Training Cross Validation Scores
Mean : 0.9174836341473466
Std : 0.0011462151706577988

From the cross validation test, we still get a good result.

Final Model Evaluation

We were asked to reach the Mean Absolute Error (MAE) value below 12% of the median property price.

Price Median : 422970.0

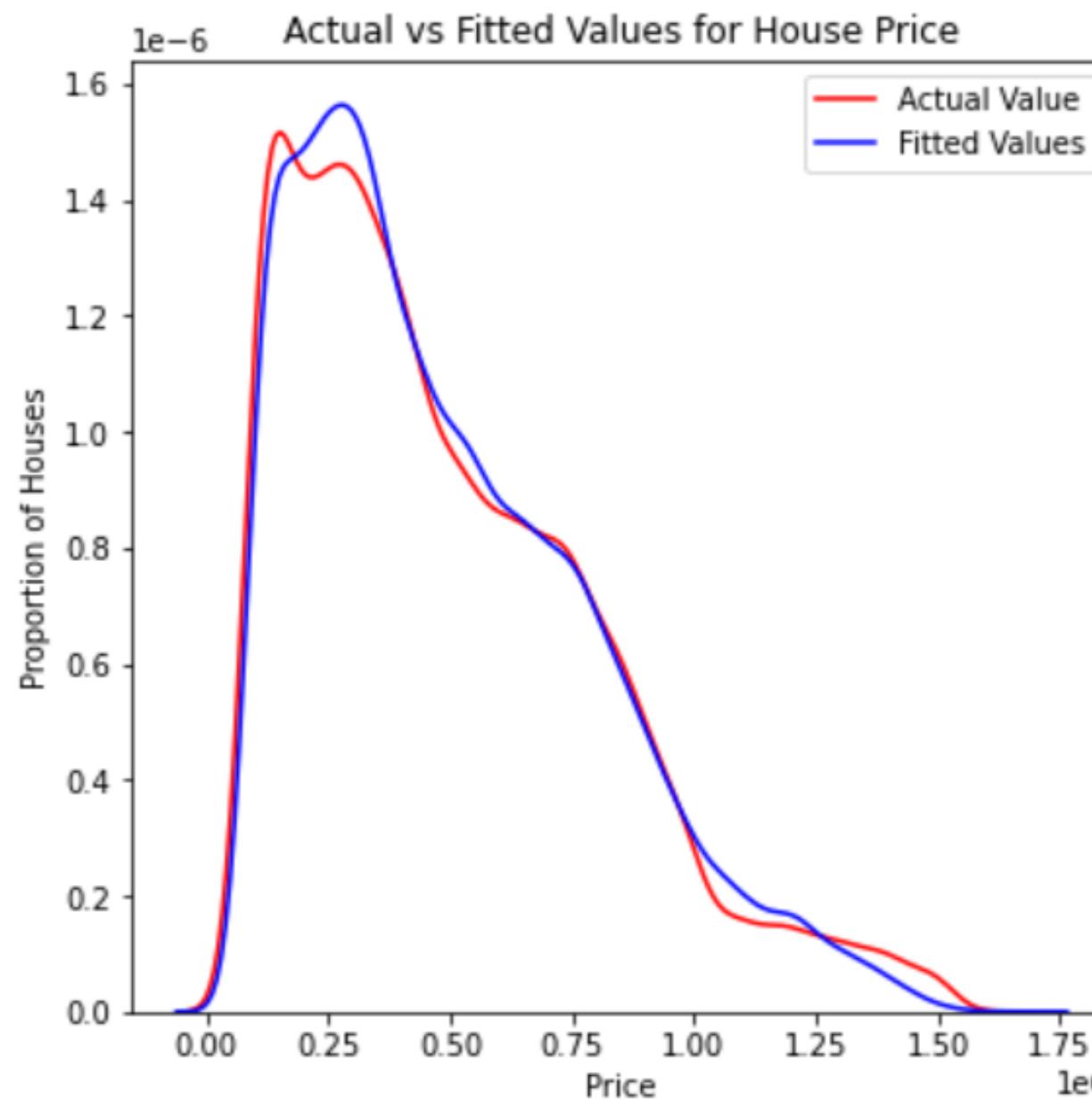
Desired MAE (12% x Median) : 50756.4

Achieved MAE : 46975.76563974304

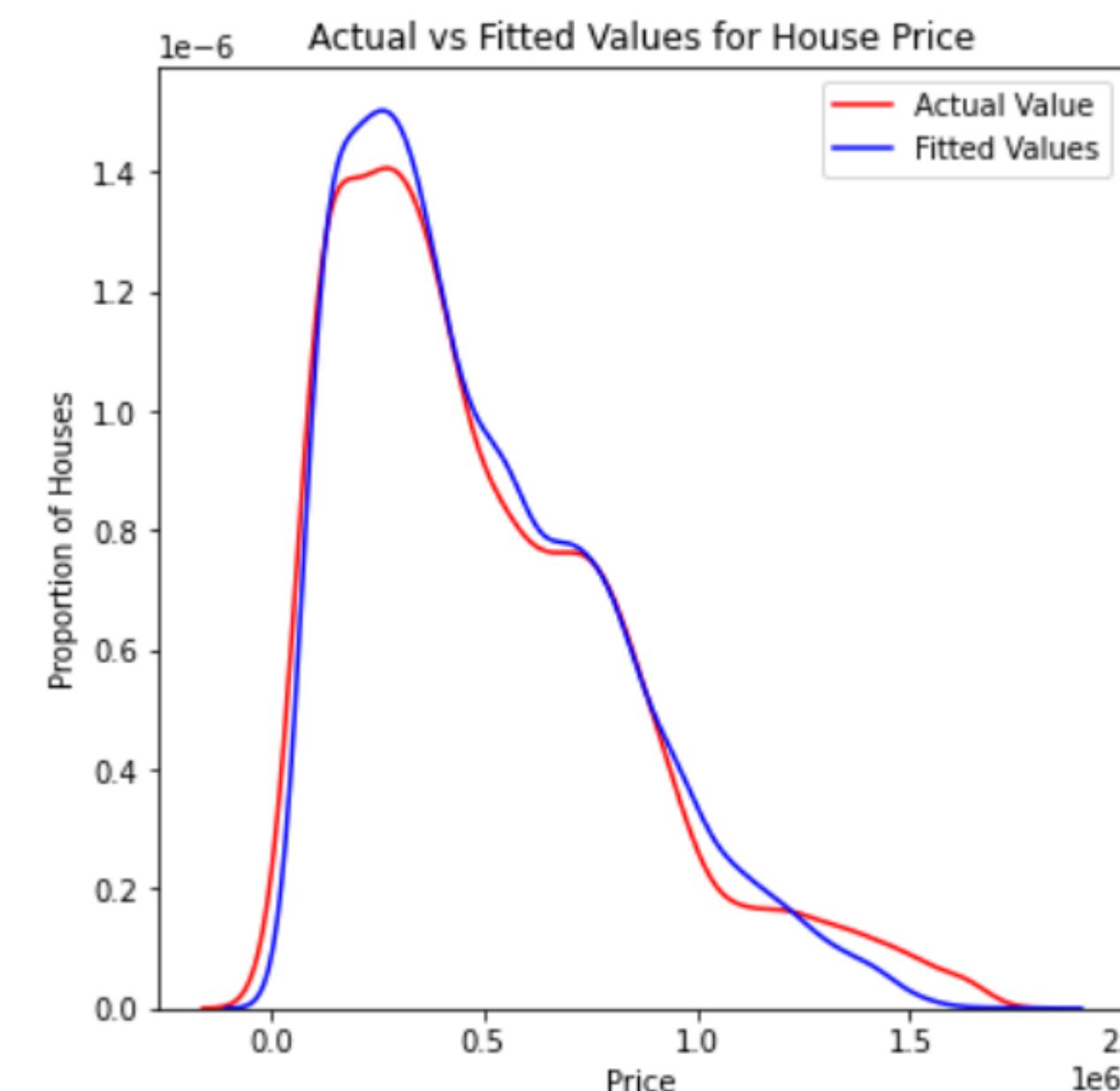
From the result shown above, we have achieved the desired MAE value (under \$50,756.4).

Distribution Plot of Actual vs Fitted Values

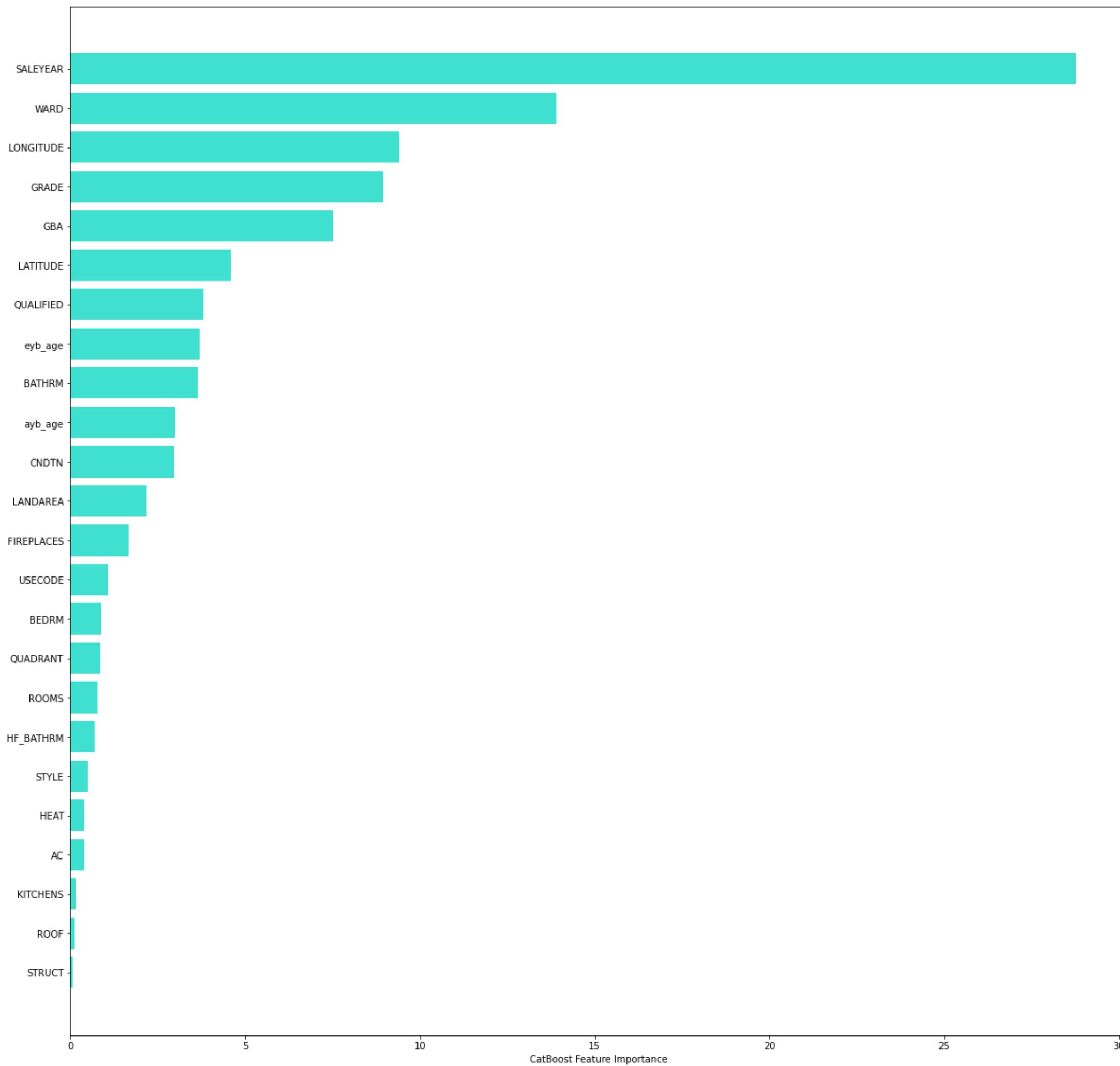
Train Set



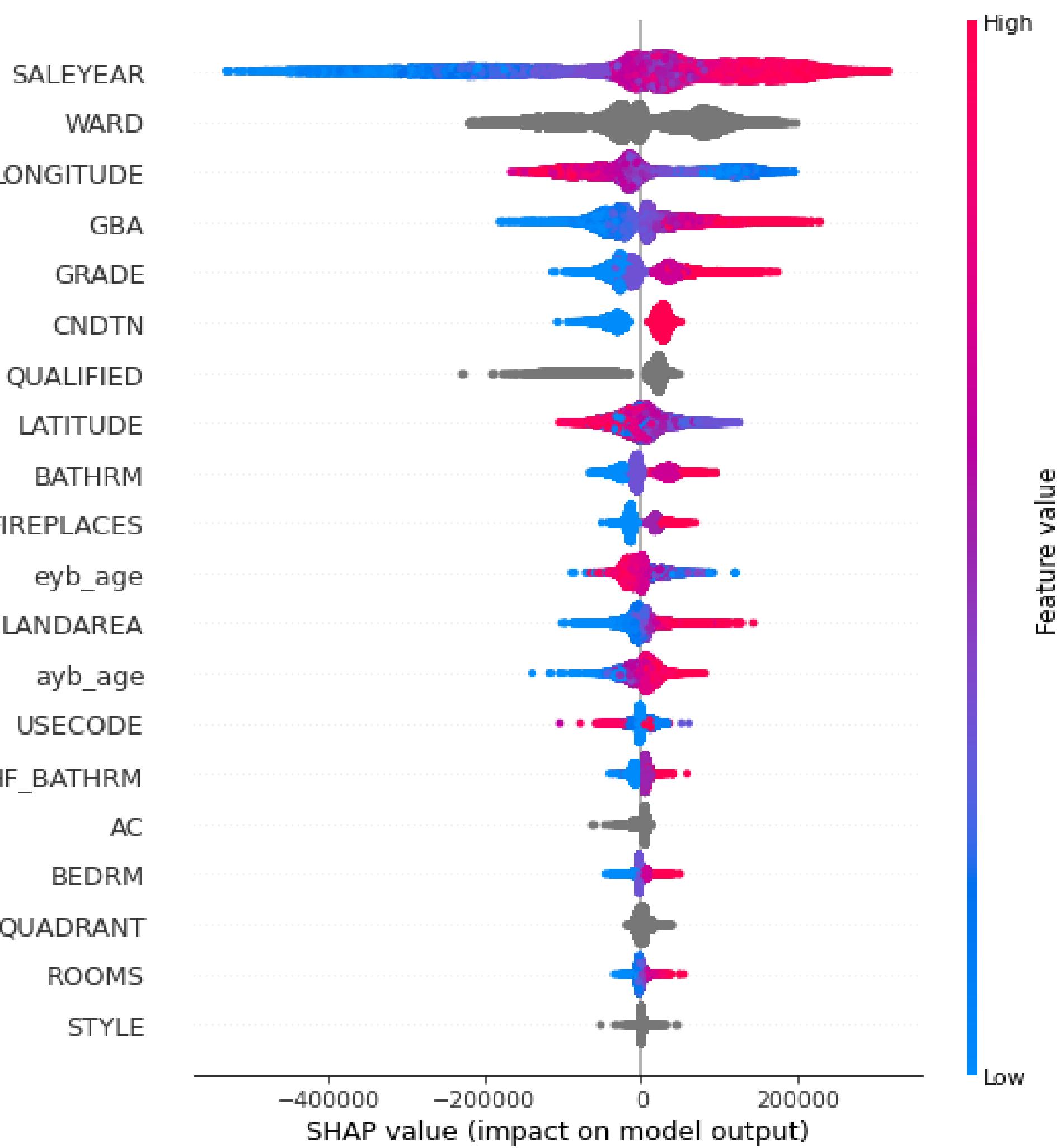
Test Set



Feature Importances



Feature Importances with SHAP



4. Deployment





Deployment Page

We deployed the model to a
webpage using FLASK.

MPL Bank

Home About Estimator Insights

Purwadhika JCDS Final Project - Matplotlib Team

In this project, we position ourselves as a part of the Data Scientist Team in a Financial Institution, MPL Bank, in Washington DC, USA. We are assigned to work on a project to develop a Machine Learning (ML) solution. The project owner is the Underwriter Team of MPL Bank. We will help the Underwriter Team to make an improvement in their process of underwriting, specifically in the process of property appraisal and valuation.

MPL orders the appraisal through a third party, an appraisal management company (AMC). In order to comply with the federal appraiser independence requirements, however, the appraisal process performed by an external party has a risk of fraud or producing erroneous results. Thus, the project owner wants to address these issues.

Problem Definition

Based on the elicitation process with the project owner, we found that they want to improve the accuracy of their underwriting process, specifically in the process of evaluating the appraisal. In the process of evaluating appraisals, there are some risks that the project owner wants to minimize, such as fraud and erroneous appraisal results given by the AMC. In addition, there is also a problem that often happens regarding the difference between the agreed offer made by a borrower and the property seller and the actual property valuation. Since lenders can't lend out money more than a property is worth, all of these risks may cause the project owner to determine wrong appraisal value and to make a wrong decision whether to give the loan to a borrower.

To address these risks and improve their business process, the project owner needs a reliable autonomous system that can provide an estimation value that can be used to compare the value given by the AMC.

The expected output of this project is a system that can make an estimation of an accurate and reasonable value (price) for a property based on the aspects of the property by using ML. However, due to the limitation of our time budget, we limit the capability of our model in this project to predict an output only for properties with grade lower than Exceptional, since Exceptional properties have a price range that is very different from the rest of properties with other grades.

Business Objectives

- Maximize profit by making the right decision to give a loan with an optimal amount.**
- Minimize loss and risks of fraud and erroneous valuation.**

Data Requirements

The value that we want to predict is the value (price) of a property. The required information needed to make a prediction are the features of the house (e.g., gross building area, the number of rooms, the number of bedrooms, etc), the condition of the house, the location, etc.

Analytic Approach

- ML Techniques**: Since the value (price) that we want to predict is a continuous value, this problem can be addressed with Supervised Learning, more specifically with Regression.
- Risk**: The risk that may be caused by wrong prediction from the ML model is profit loss especially when the model gives underappraised value (price).
- Performance Measure**: The performance measures to evaluate the ML model are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2).
- Action**: The business user can utilize the prediction result by comparing it with the appraisal value given by the AMC.
- Value**: The values created from the project are the improvement in the underwriting process and the maximized profit from giving the right appraisal and making the right decision in providing loan.

MPL Bank

Home About Estimator Insights

Washington DC, USA. Matplotlib Team Our Project

Lis Cory Kemal Isfan Rezki Fauziansyah

© Copyright 2021 Matplotlib Team. Designed by TemplateMo

MPL Bank

Home About Estimator Insights

Property Value Estimator

Washington DC, USA

The required data to estimate the value of a property are the location, the condition and the specification of the property.

Fill the form to estimate property value!

Location

| | |
|---------------------|-----------------------------|
| Ward Select Ward | Quadrant Select Quadrant |
| Longitude | Latitude |

Condition

| | |
|---|-------------------------------|
| Building Age (in years) | Renovation Years (in years) |
| Last Sale Year Select Last Sale Year | Condition Select Condition |
| Grade Select Grade | Qualified? Qualified? |

Specification

| | |
|-------------------------------|---|
| Gross Building Area (in sqft) | Land Area (in sqft) |
| Style Select Style | Use Code Select Use Code |
| Rooms | Bedrooms |
| Bathrooms | Half Bathrooms |
| Kitchens | Fireplaces |
| AC Has AC? | Heating System Select Heating System |
| Roof Select Roof | Structure Select Structure |

Estimate

MPL Bank

Home About Estimator Insights

Property Value Estimator

Washington DC, USA

The required data to estimate the value of a property are the location, the condition and the specification of the property.

Fill the form to estimate property value!

Location

| | |
|-------------------------|-------------------------|
| Ward Ward 5 | Quadrant NE |
| Longitude -76.994888 | Latitude 38.95709777 |

Condition

| | |
|-------------------------------|-----------------------------------|
| Building Age (in years) 67 | Renovation Years (in years) 48 |
| Last Sale Year 2014 | Condition Good |
| Grade Average | Qualified? Qualified |

Specification

| | |
|---|---|
| Gross Building Area (in sqft) 1088 | Land Area (in sqft) 2838 |
| Style 2 Story | Use Code 13 - Single family residential home with slight commercial/in |
| Rooms 6 | Bedrooms 3 |
| Bathrooms 1 | Half Bathrooms 1 |
| Kitchens 1 | Fireplaces 0 |
| AC Yes | Heating System Warm Cool |
| Roof Concrete / Comp Shingle / Built Up / Metal-Pre / Typical / Co | Structure Semi-Detached / Multi / Town Inside / Town End |

Estimate

MPL Bank

Home About Estimator Insights

Washington DC, USA. Matplotlib Team Our Project

Lis Cory Kemal Isfan Rezki Fauziansyah

© Copyright 2021 Matplotlib Team. Designed by TemplateMo

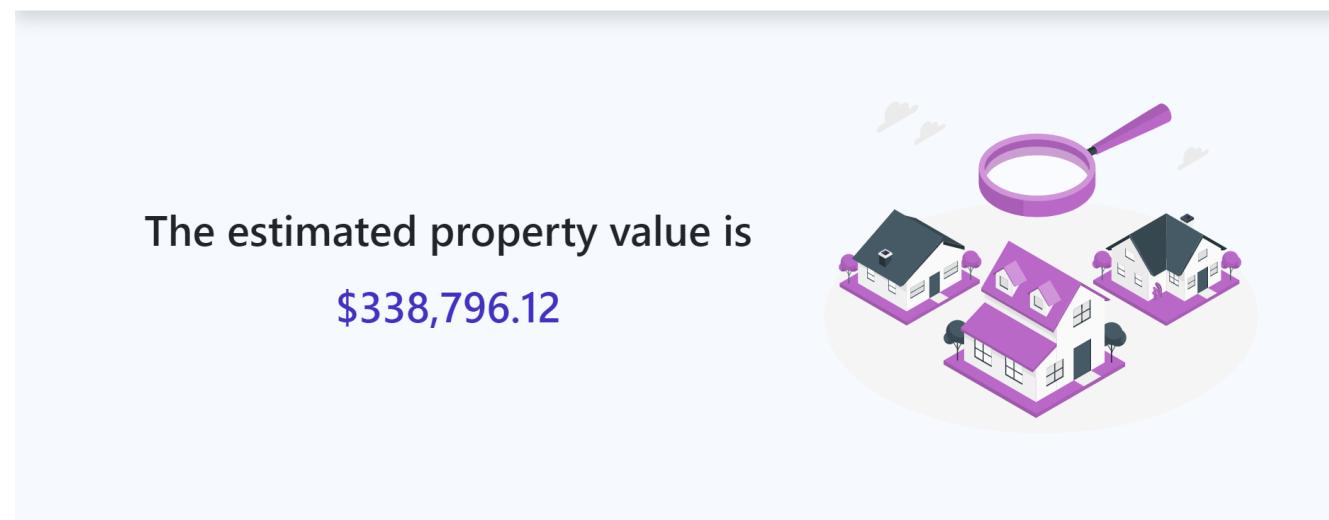
MPL Bank

Home About Estimator Insights

Washington DC, USA. Matplotlib Team Our Project

Lis Cory Kemal Isfan Rezki Fauziansyah

© Copyright 2021 Matplotlib Team. Designed by TemplateMo



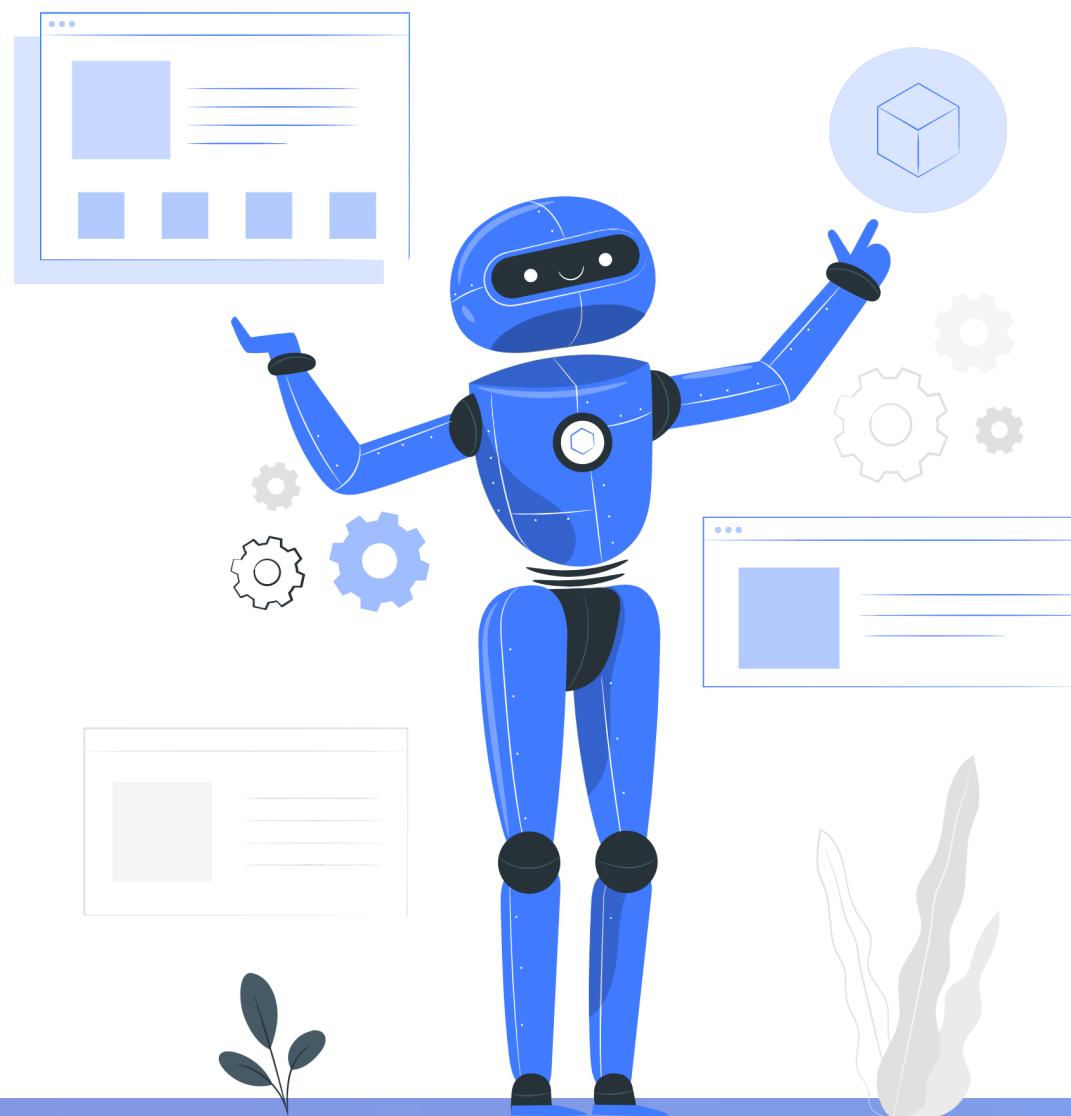
- Remarks**
- 11 - Single family residential home used as such
 - 12 - Single family residential home with non-economic 2nd unit
 - 13 - Single family residential home with slight commercial/ind
 - 15 - Townhouse - Planned Development
 - 19 - SFR - Manufactured Home (MH on permanent foundation)
 - 23 - Triplex, double or duplex with single family home
 - 24 - Four living units, e.g. fourplex or triplex w/SFR

5. Future Works



Future Works

We acknowledge that the result achieved is not perfect as there are more factors that could affect a property's price such as proximity to public services and facilities, tourism spots, purchasing power, area development prospect, etc.

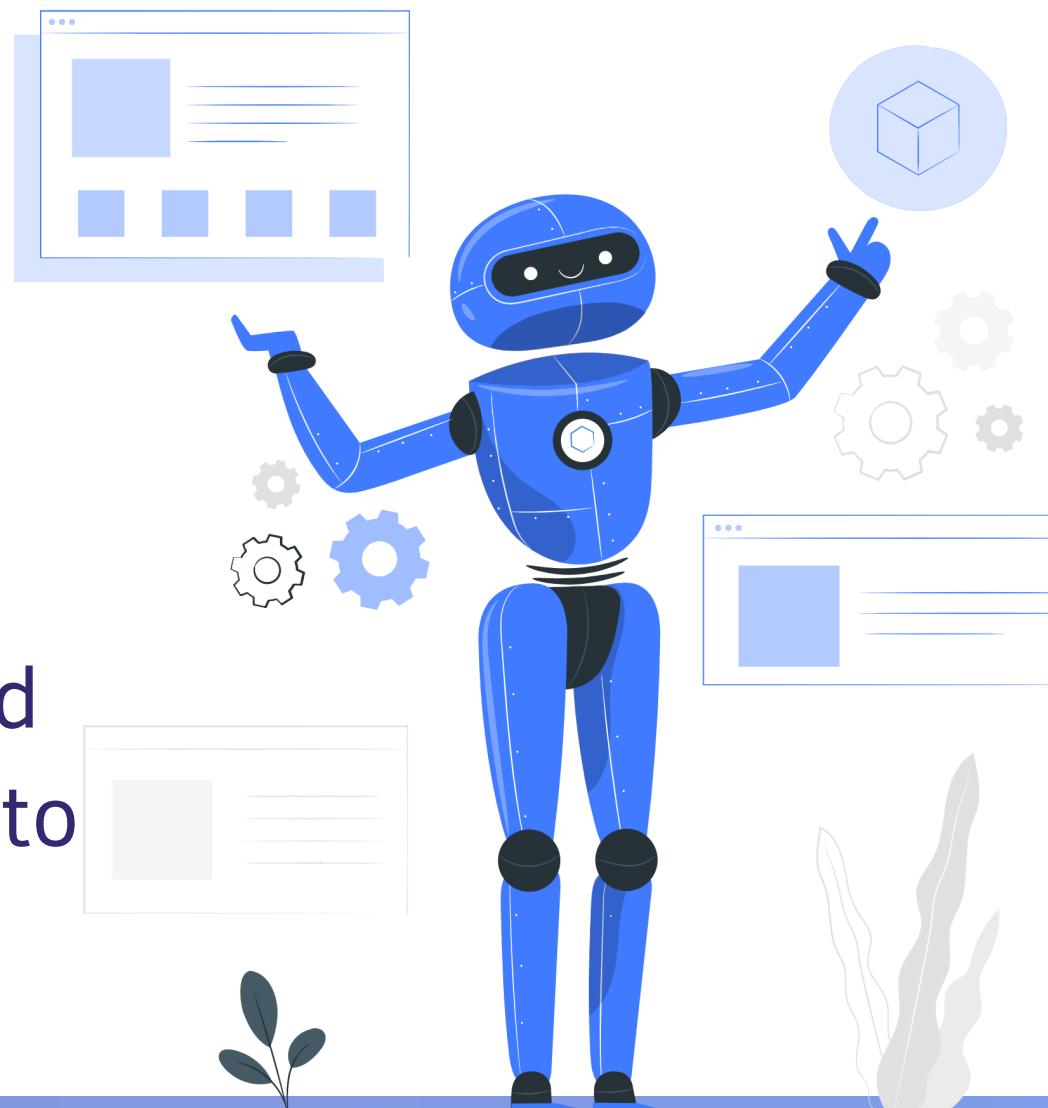


Future Works

These are a few things that might help to improve model prediction result further :

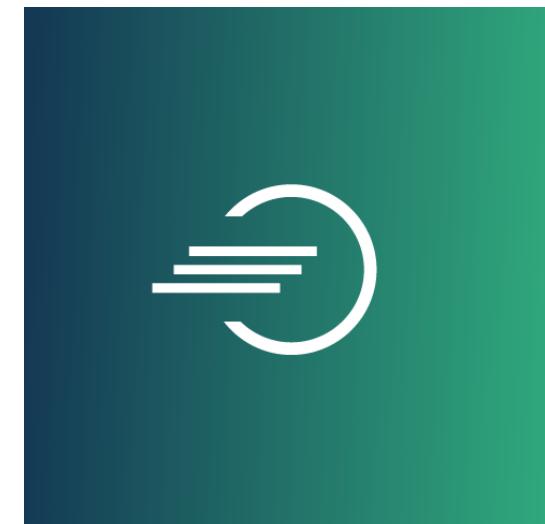
- Collect more property data in Washington DC.
- Get another relevant dataset such as DC Residents Demographic, Public Services and Facilities, etc.
- Try to experiment with more features.
- Use more parameters in grid search cv.

Additionally, for the ease of usage for the business users, we would recommend to add a feature in the application which allows users to input more data instantaneously in a .csv or .xlsx file format.



Thank You!

PURWADHIKA JCDS 1202
MATPLOTLIB TEAM



Lis Cory

Rezki Fauziansyah

Teuku Muhammad Kemal Isfan

GITHUB REPOSITORY

<https://github.com/ls-cy/Purwadhika-JCDS-Final-Project>

https://github.com/PurwadhikaDev/MatplotlibGroup_JC_DS_12_FinalProject