

# 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



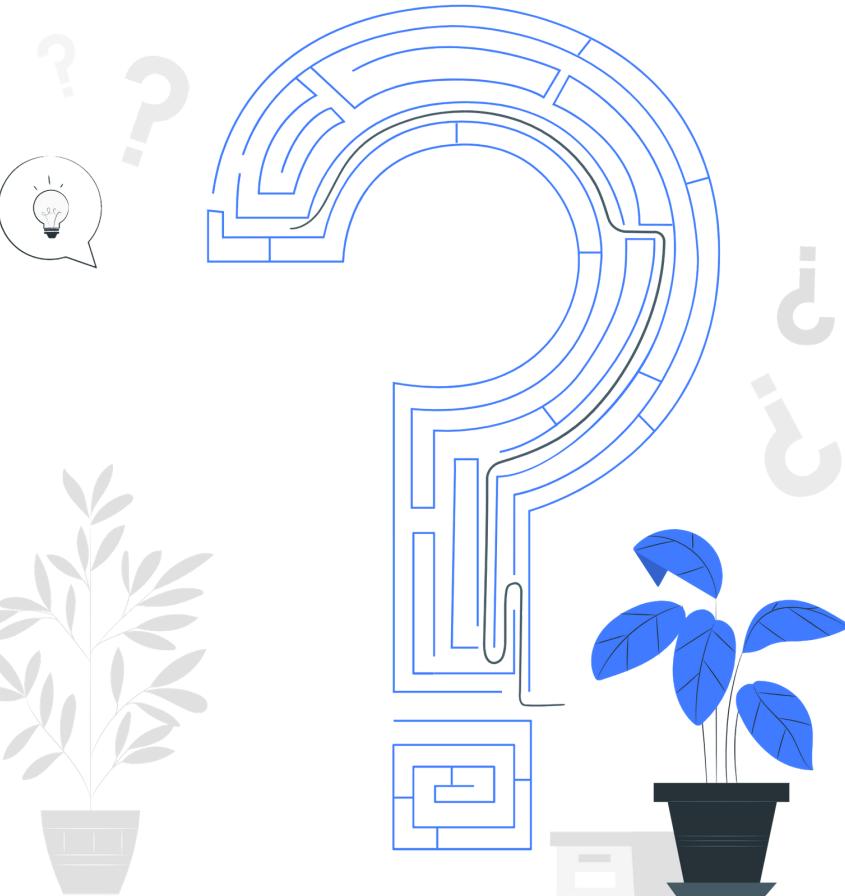
# Background

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

Difference between the  
Agreed Offer and the Actual  
Property Valuation

Improving Accuracy in  
Appraisal Evaluation Process

# Why?



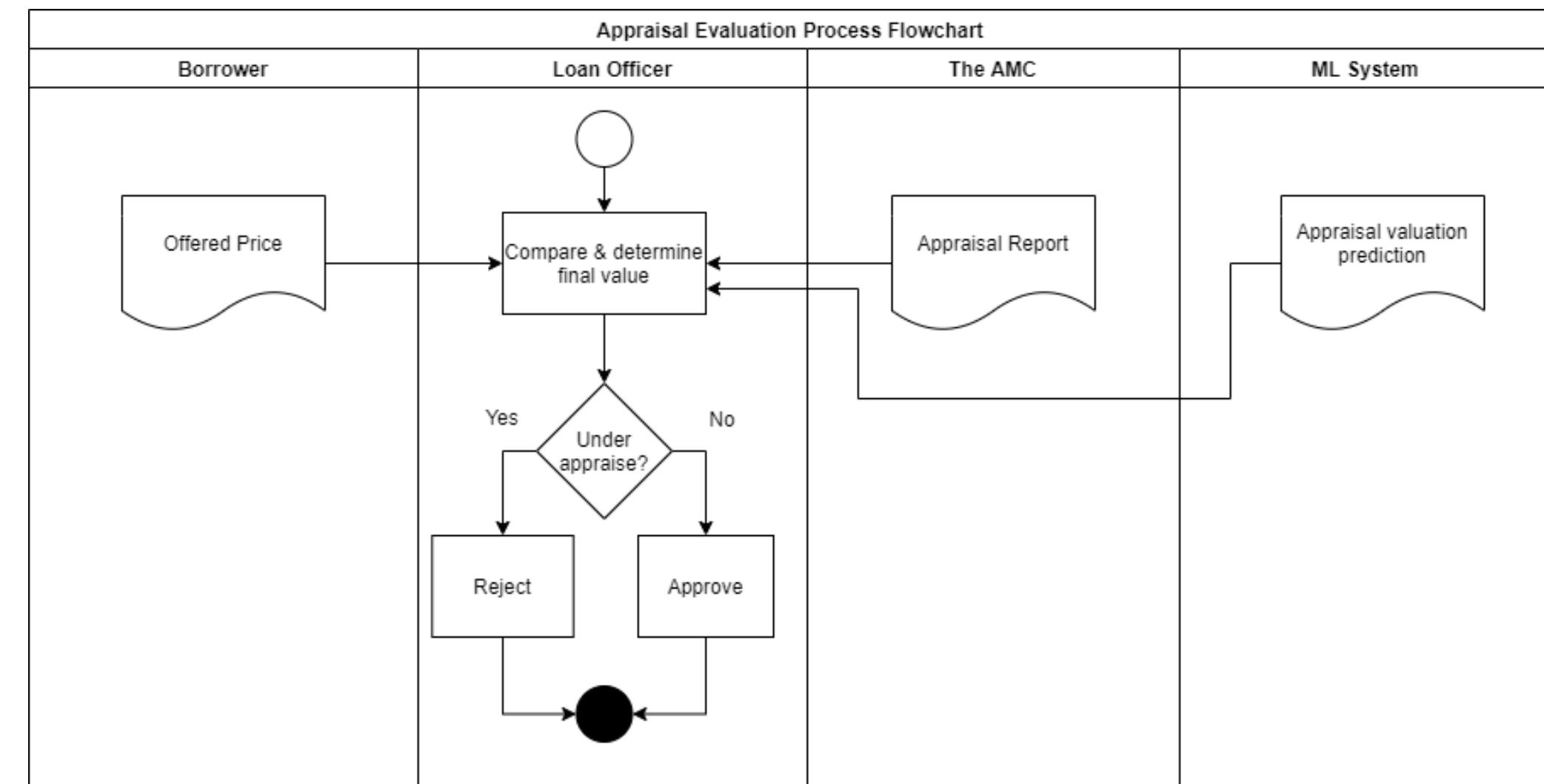
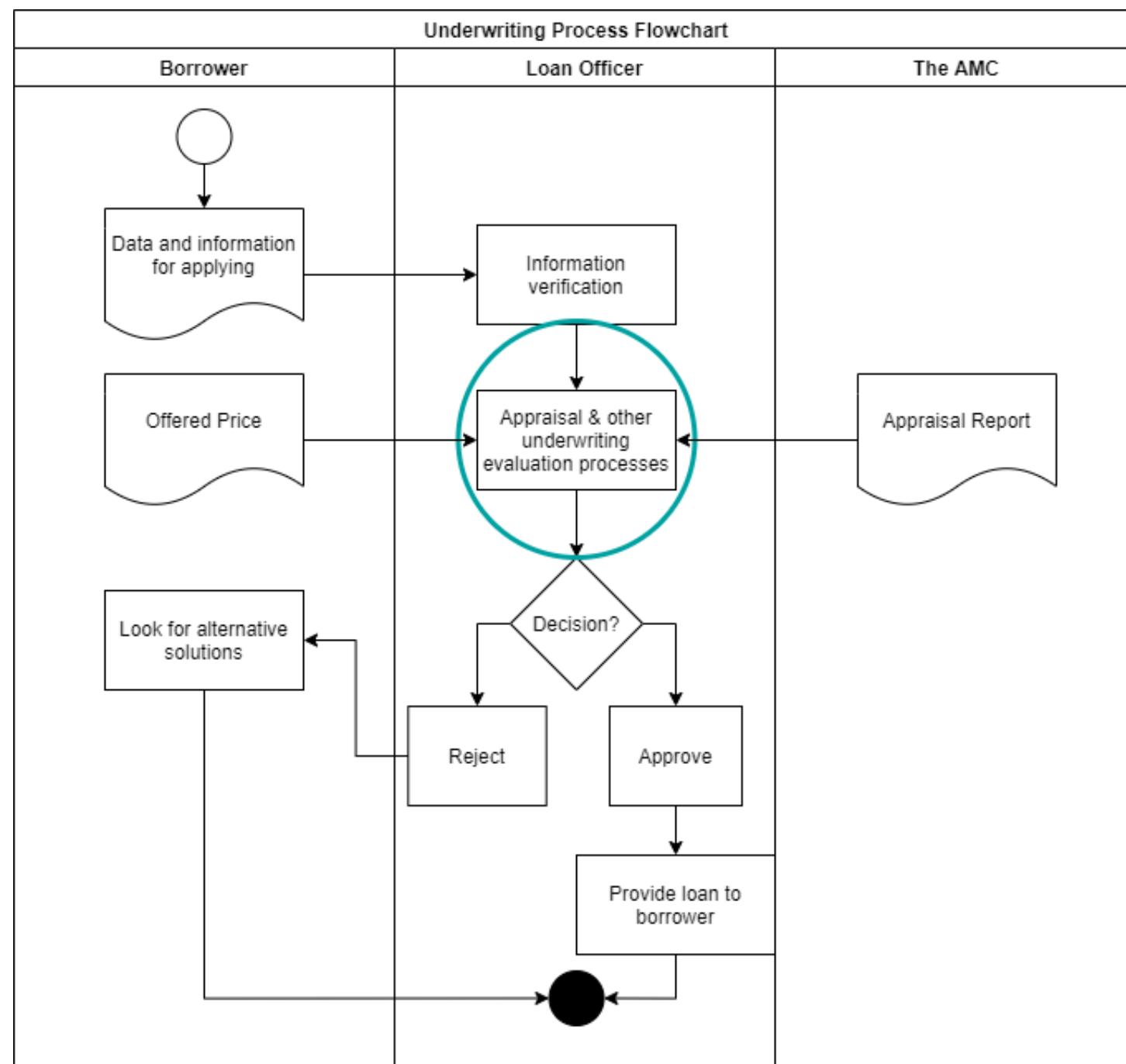
**Appraisal process directly affects company's revenue.**

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.

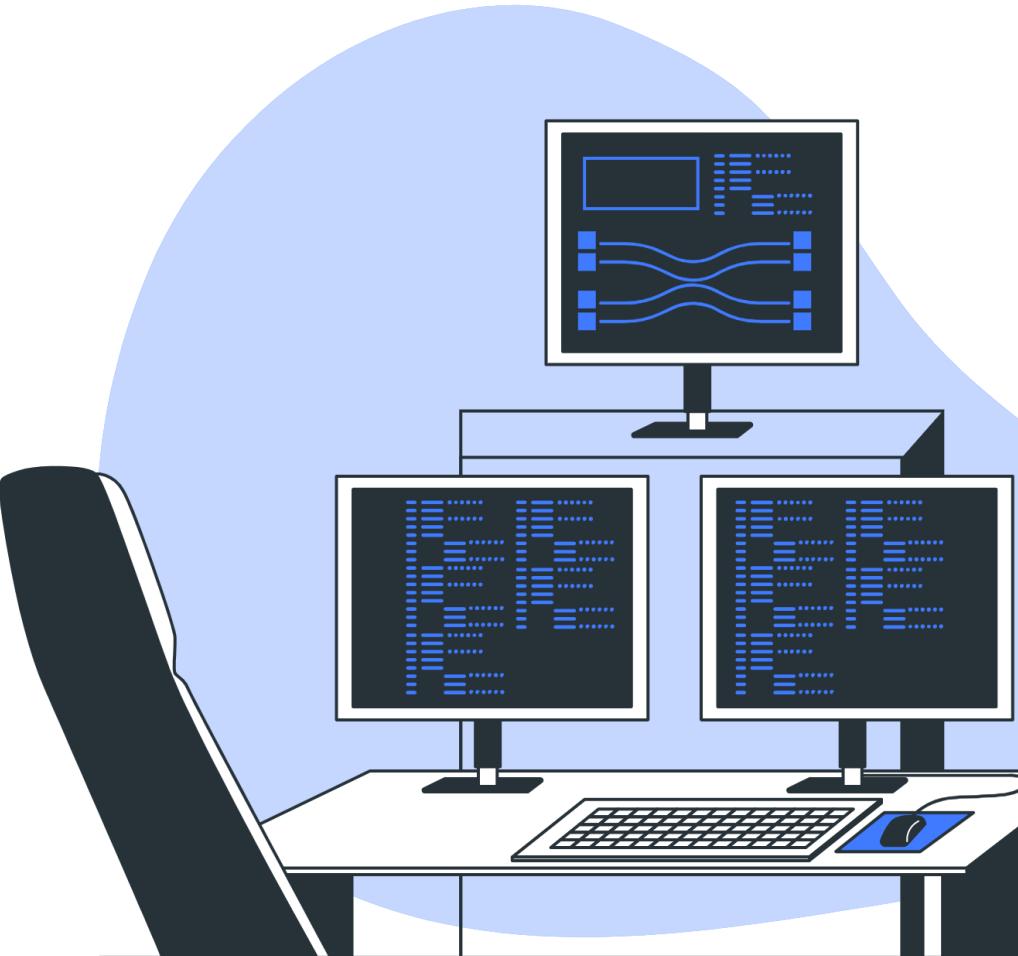
# Appraisal Evaluation Process



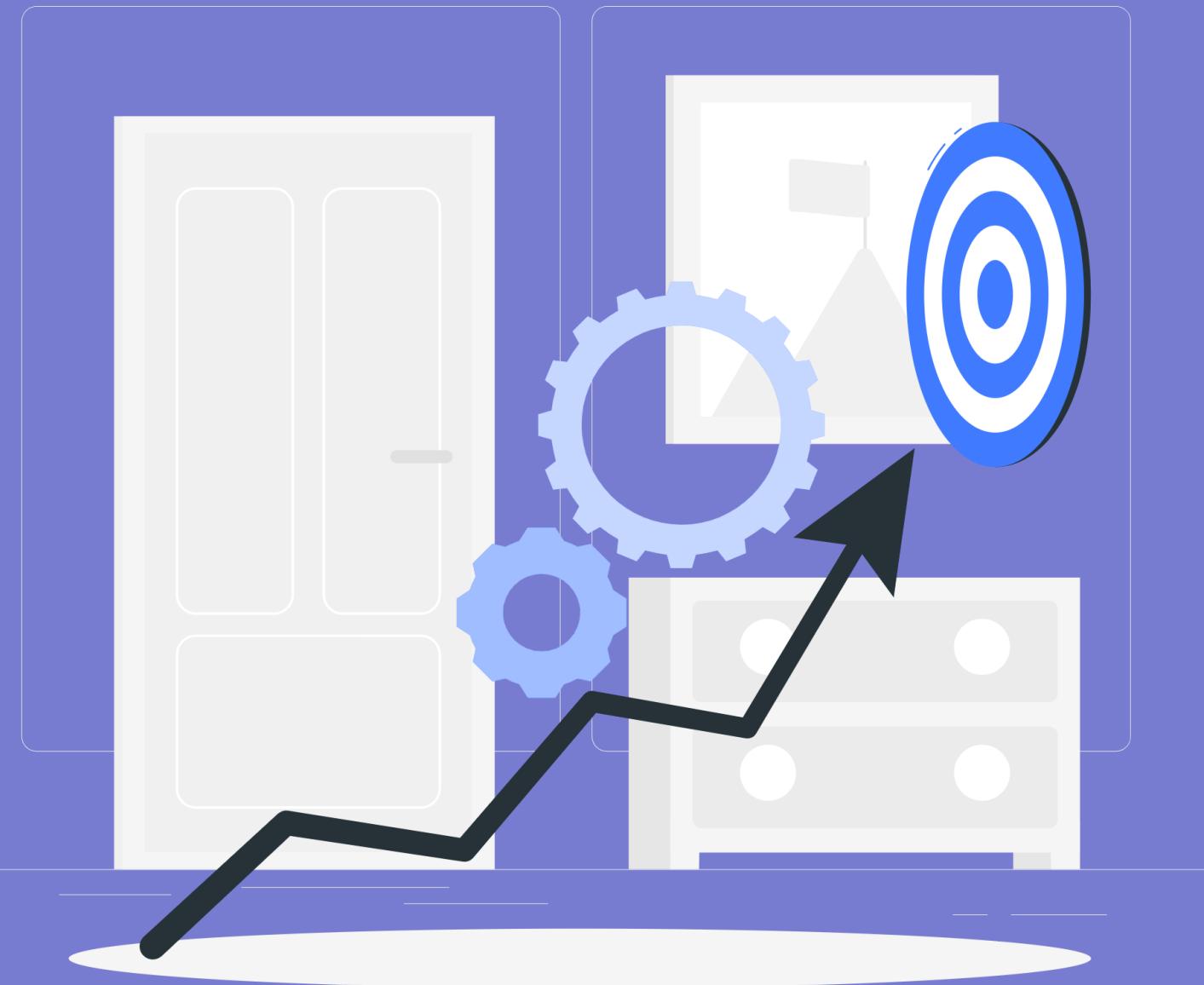
# Expected Output

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 13% 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 Techniques

Target : Continuous Value

Machine Learning : 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

<b>BATHRM</b> Number of Full Bathroom	<b>HF_BATHRM</b> Number of Half Bathroom	<b>HEAT</b> Heating System	<b>AC</b> AC Availability
<b>NUM_UNITS</b> Number of Units	<b>ROOMS</b> Number of Rooms	<b>BEDRM</b> Number of Bedrooms	<b>AYB</b> The earliest time the main portion of the building was built
<b>EYB</b> The year an improvement was built more recent than actual year built	<b>QUALIFIED</b> Government's Criteria on whether the sale is representative of market value	<b>GBA</b> Gross Building Area (in sqft)	<b>STYLE</b> Style

# Data Description

<b>STRUCT</b> Structure	<b>GRADE</b> Grade	<b>CNDTN</b> Condition	<b>EXTWALL</b> Exterior Wall
<b>ROOF</b> Roof Type	<b>INTWALL</b> Interior Wall	<b>KITCHENS</b> Number of Kitchens	<b>FIREPLACES</b> Number of Fireplaces
<b>LANDAREA</b> Land Area (in sqft)	<b>LATITUDE</b> Latitude	<b>LONGITUDE</b> Longitude	<b>ASSESSMENT_NBHD</b> Neighborhood ID

# Data Description

WARD	QUADRANT	SALEDATE	PRICE
Ward	City Quadrant (NE, SE, SW, NW)	Date of Most Recent Sale	Price

# Data Pre-Processing

## Drop Unused Columns

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

## Create New Columns

- ayb\_age = Present Year - AYB
- eyb\_age = Present Year - EYB

## Drop Unusual/ Erroneous Data

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

## Drop Rows with Missing Values

SALEDATE, KITCHENS, ROOMS, GRADE, HEAT, BATHRM

## Merging Values with Very Few Occurrence

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

## Excluding Outliers

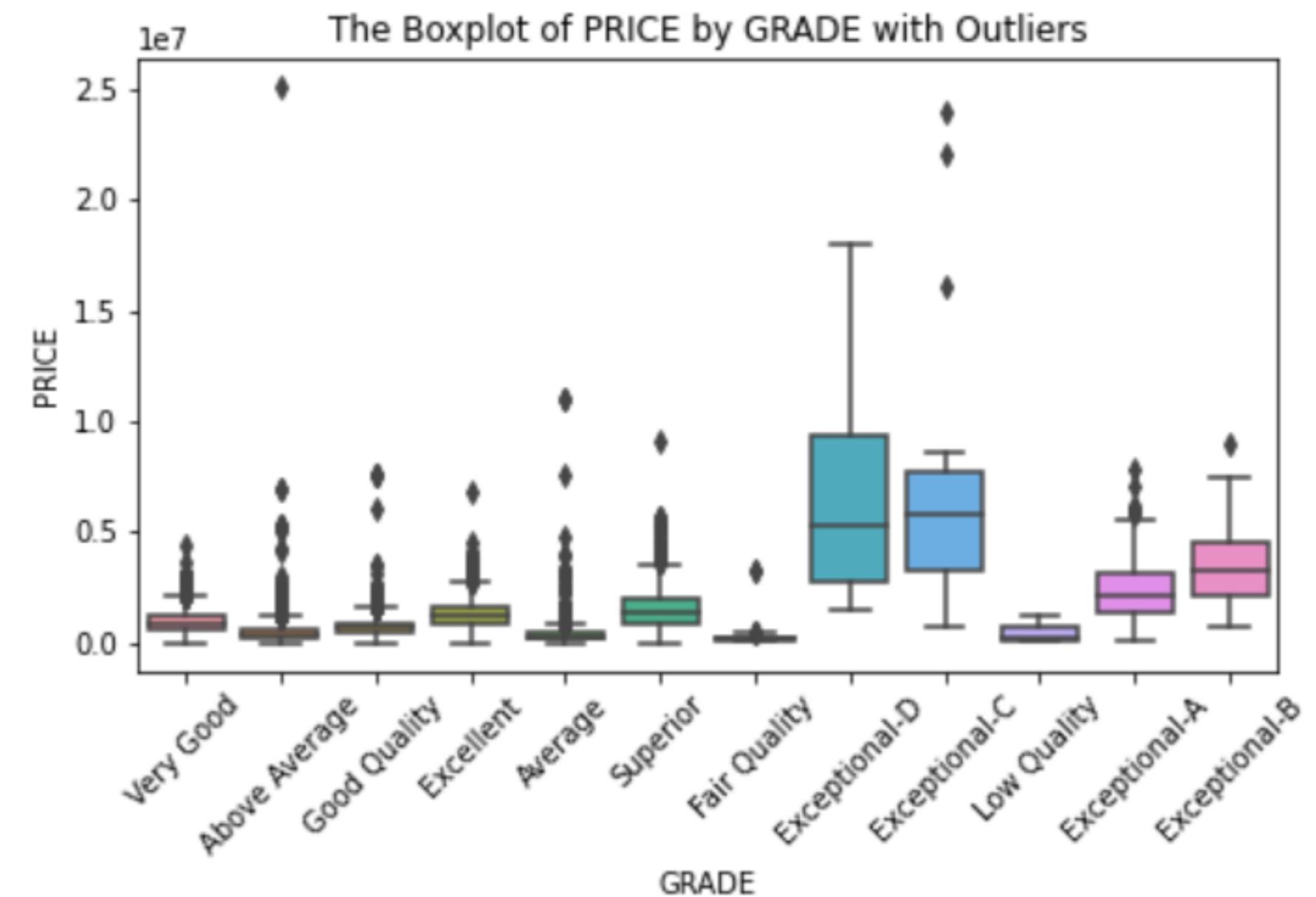
KITCHENS, PRICE, Price based on CNDTN

## Re-classify Columns

CNDTN, STRUCT, ROOF

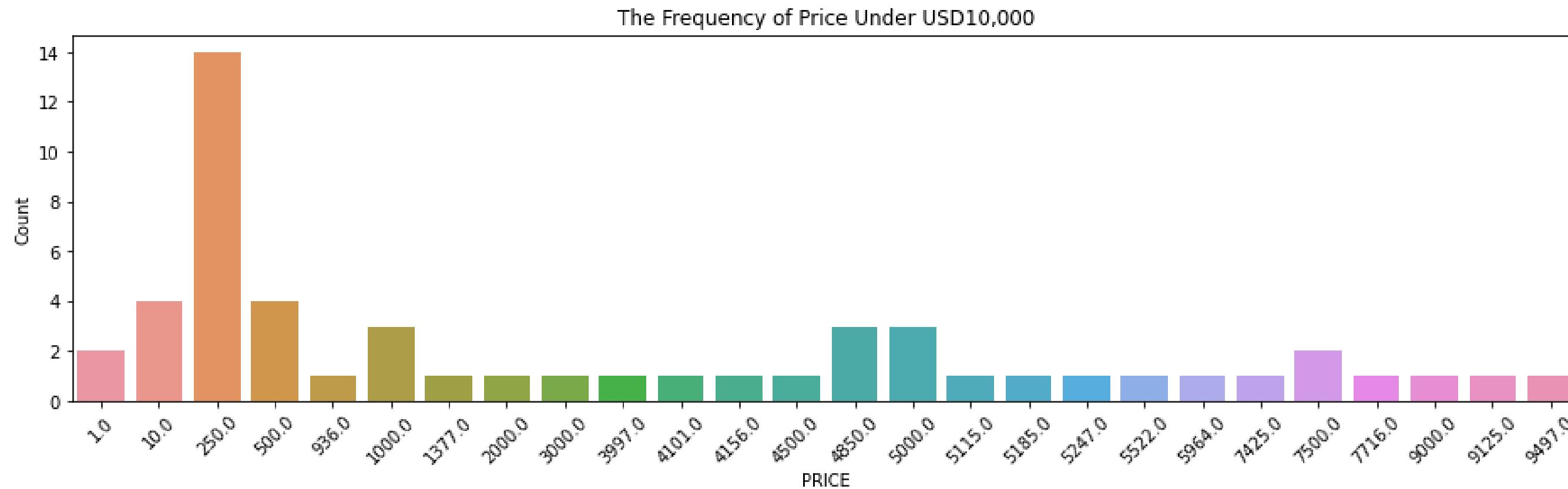
# Identifying Outliers in PRICE

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.



# Erroneous PRICE

We found few suspicious data where properties were being sold for a very low price. We obtain the information that the median of the whole USA residential price is around  $\pm$ USD50,000. Thus based on this information, we drop data points with an assumption that there are very few cheap properties with values as low as one-fifth to the median price, which is USD10,000.



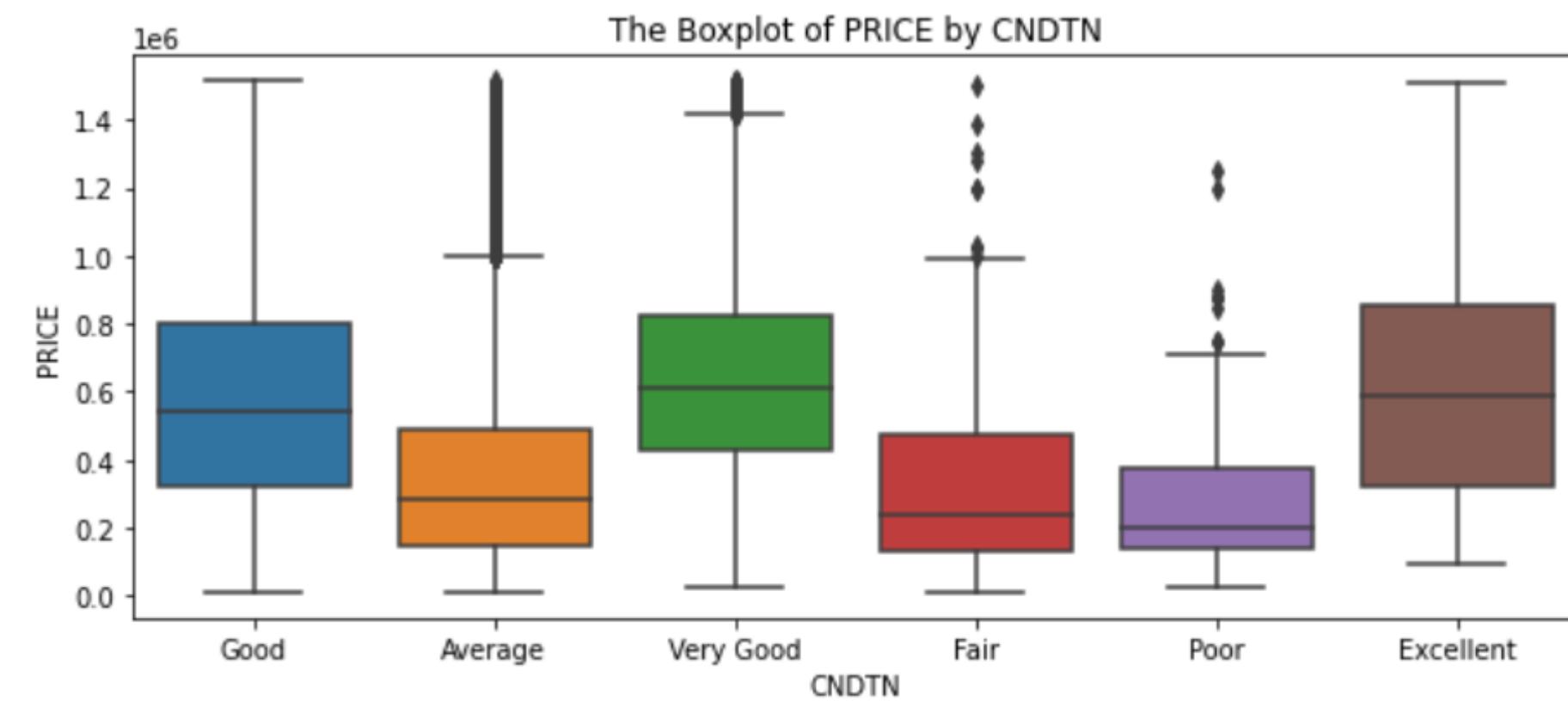
# Re-classify CNDTN

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.

Price Median : USD420,000

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

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



# Re-classify STRUCT & ROOF

Price Median : USD420,000

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

- 0 : Under median
- 1 : Above median

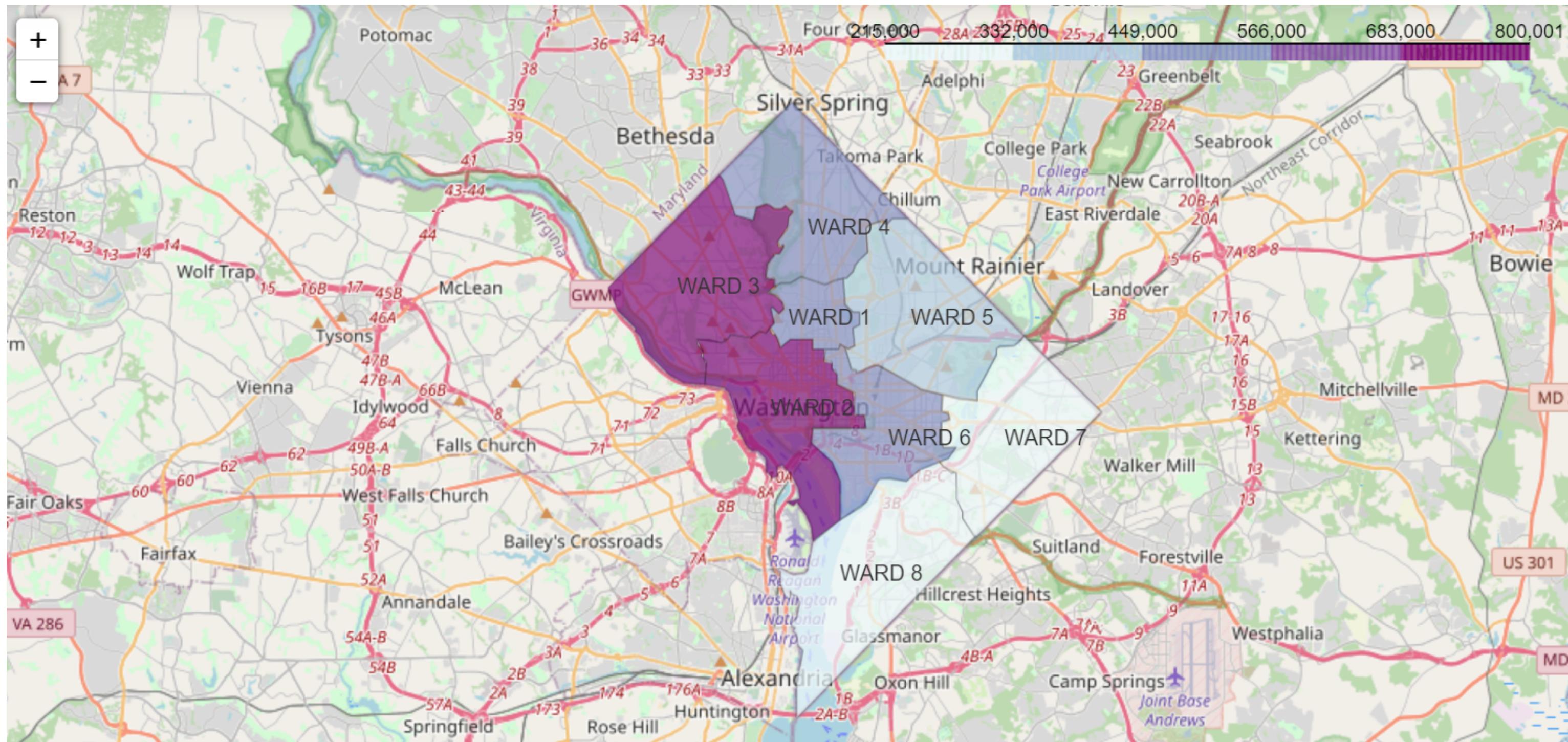
	STRUCT	median_PRICE	total
0	Semi-Detached	267000.0	7961
1	Multi	295000.0	2748
2	Town Inside	349500.0	185
3	Town End	388478.5	76
4	Row End	425000.0	6530
5	Single	471000.0	14267
6	Row Inside	475000.0	22444
7	Default	560000.0	2

ROOF	median_PRICE	total	
0	Concrete	299900.0	1
1	Comp Shingle	350000.0	15368
2	Built Up	362500.0	16971
3	Metal- Pre	366500.0	133
4	Typical	370000.0	84
5	Composition Ro	410150.0	61
6	Shake	455000.0	334
7	Metal- Sms	484000.0	15369
8	Shingle	493506.0	215
9	Concrete Tile	512500.0	2
10	Water Proof	594000.0	4
11	Clay Tile	650000.0	237
12	Slate	685000.0	4568
13	Neopren	711500.0	849
14	Wood- FS	847500.0	4
15	Metal- Cpr	853500.0	13

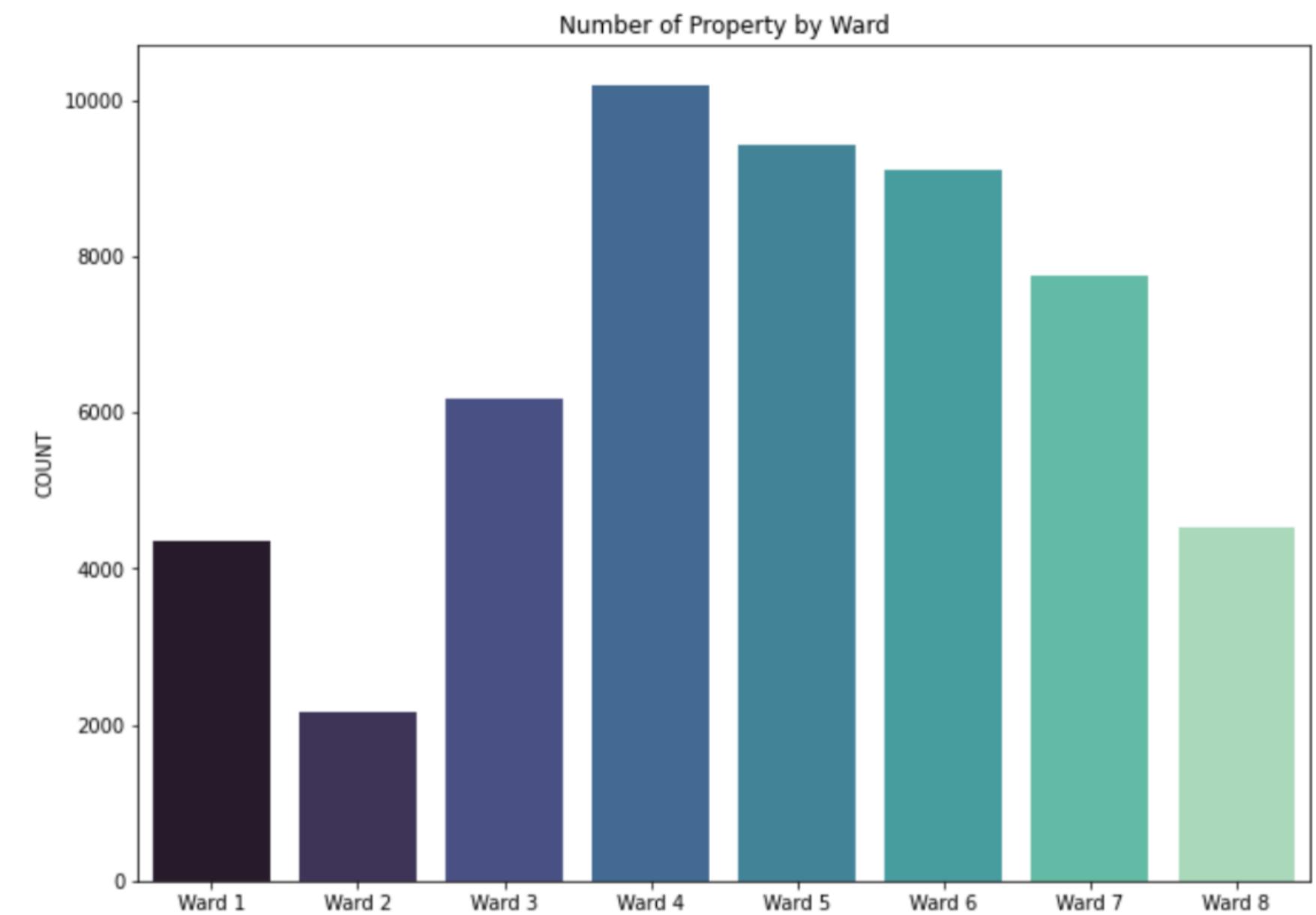
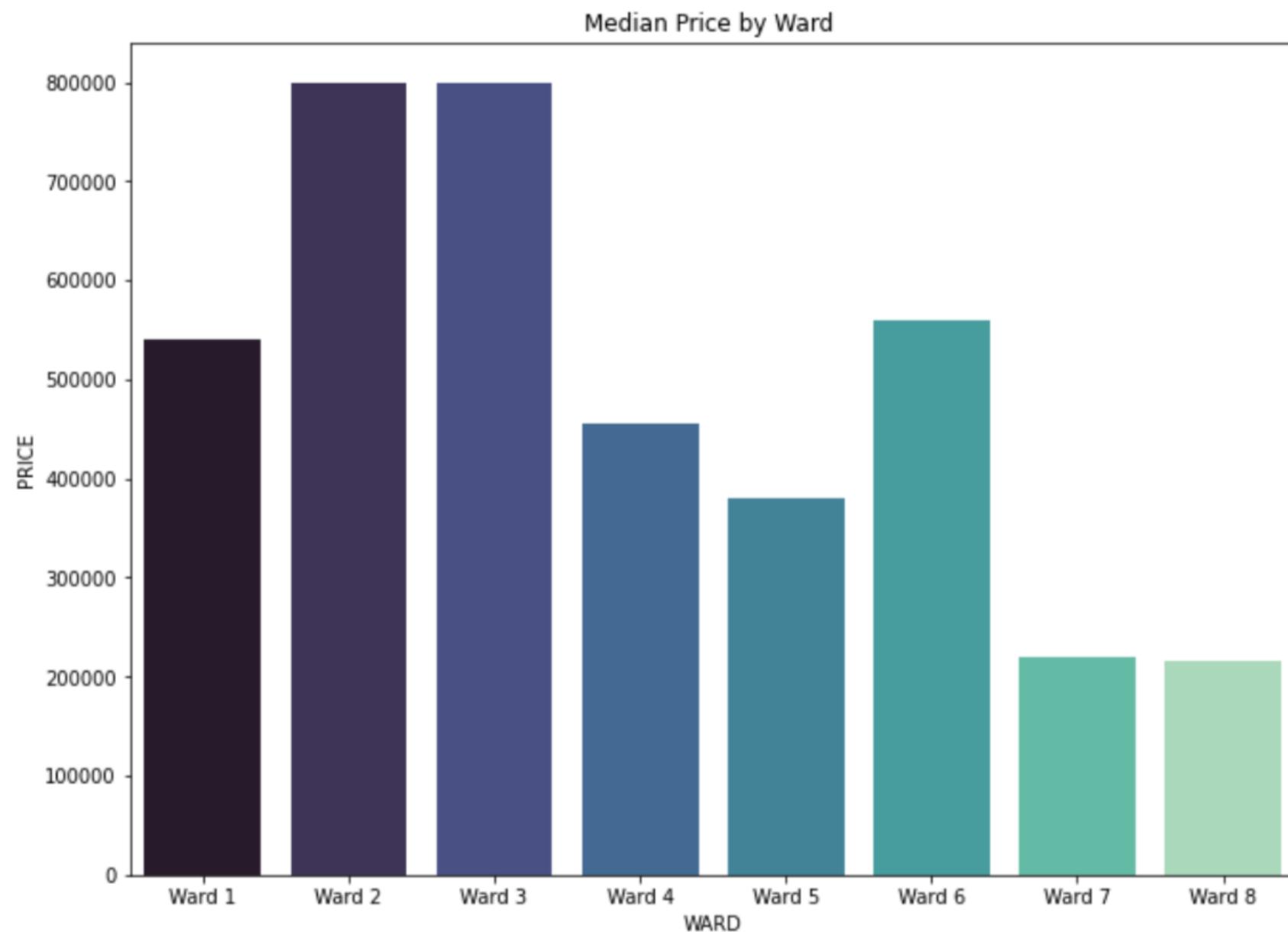
# Insights

---

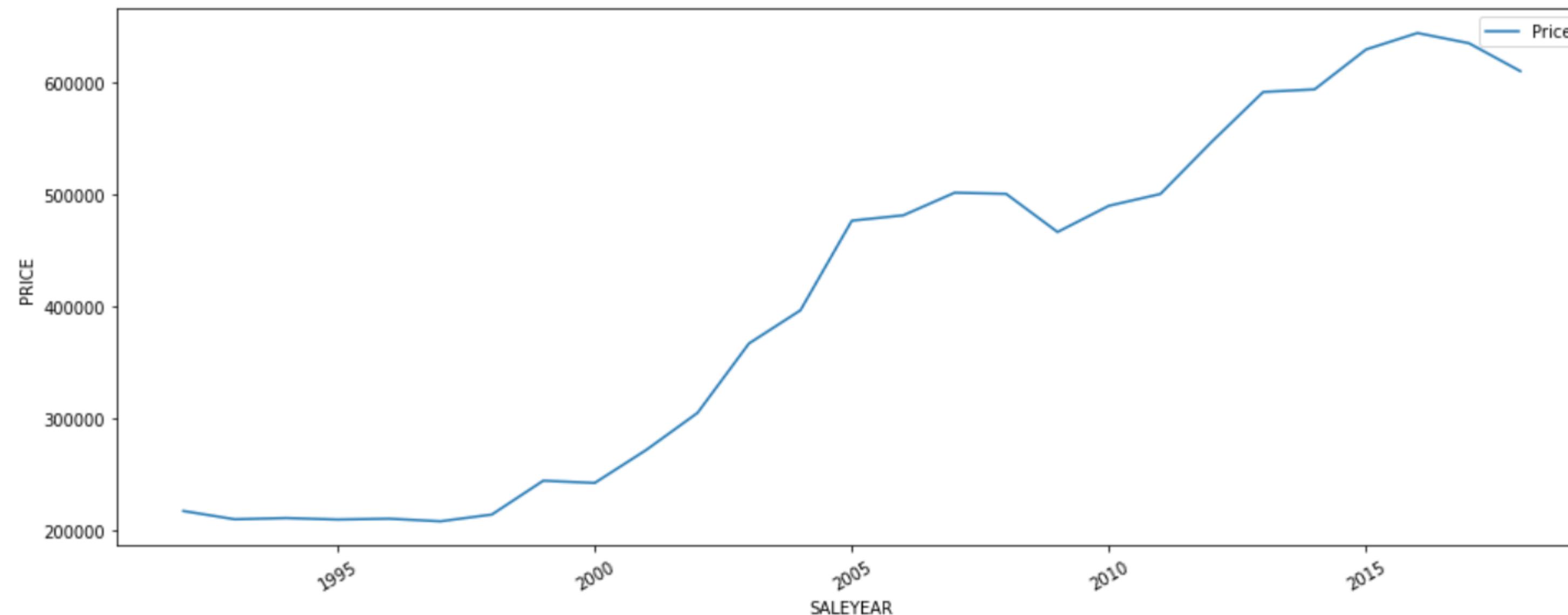
# Map by Ward



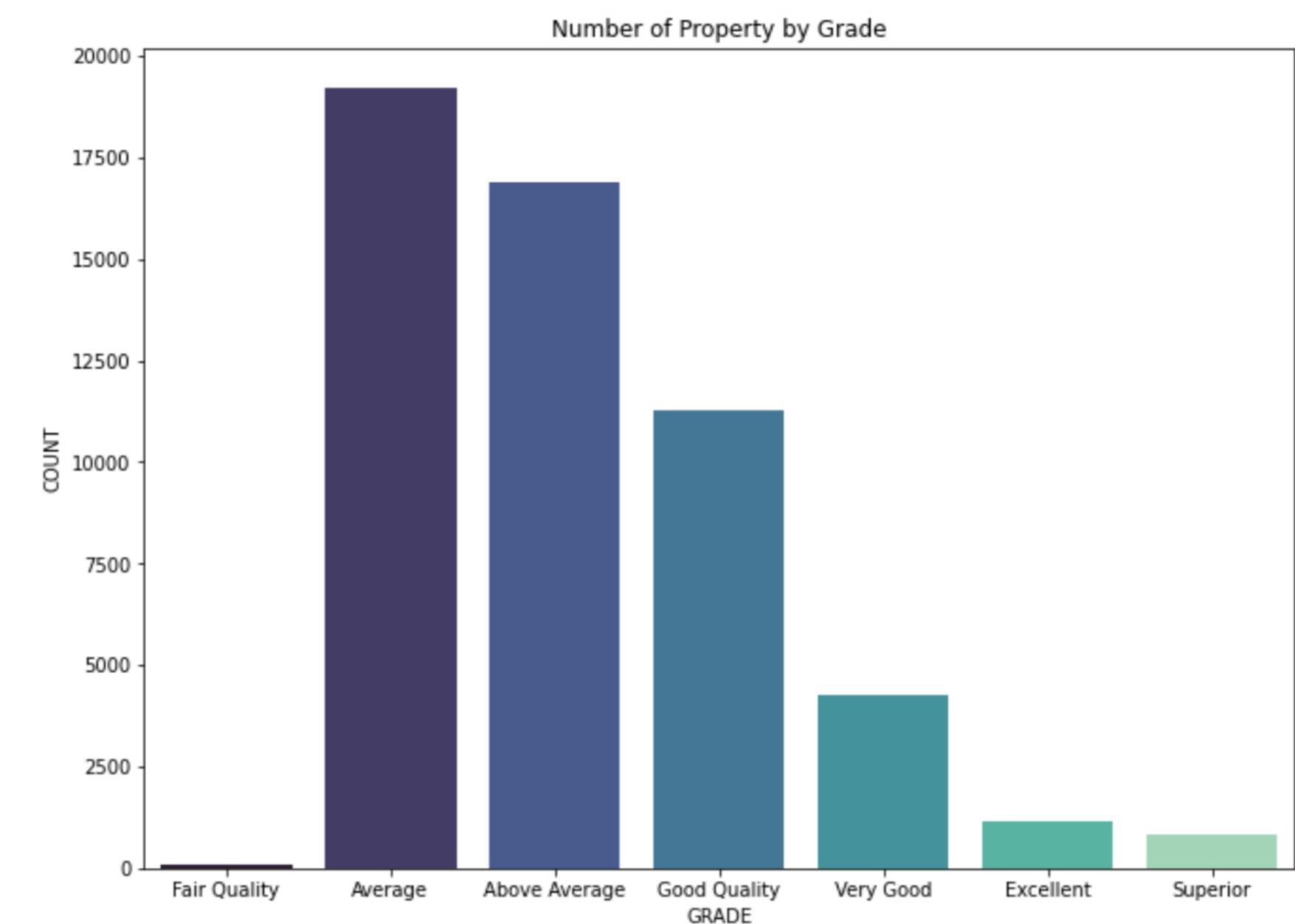
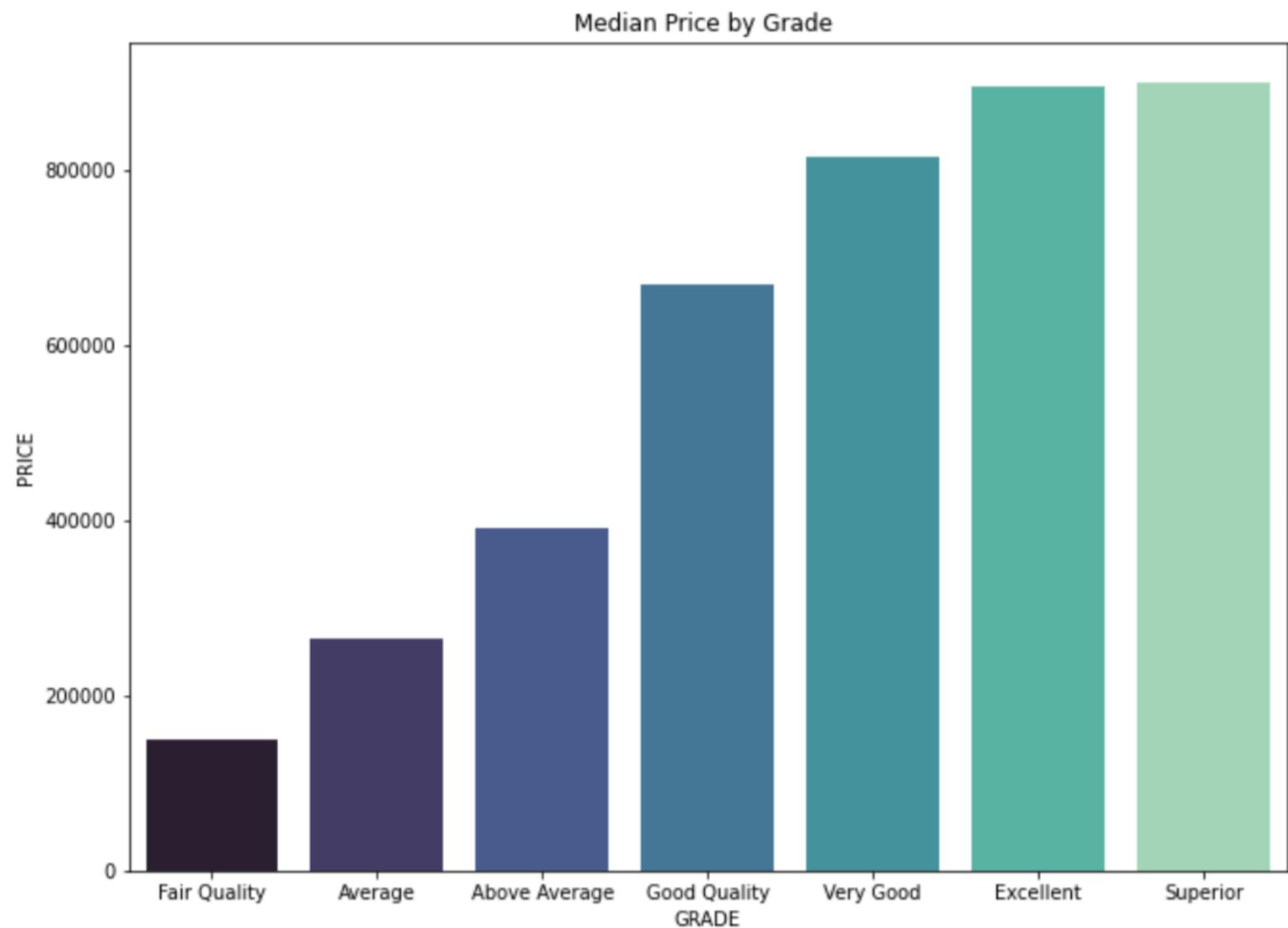
# Median Price & Number of Property by Ward



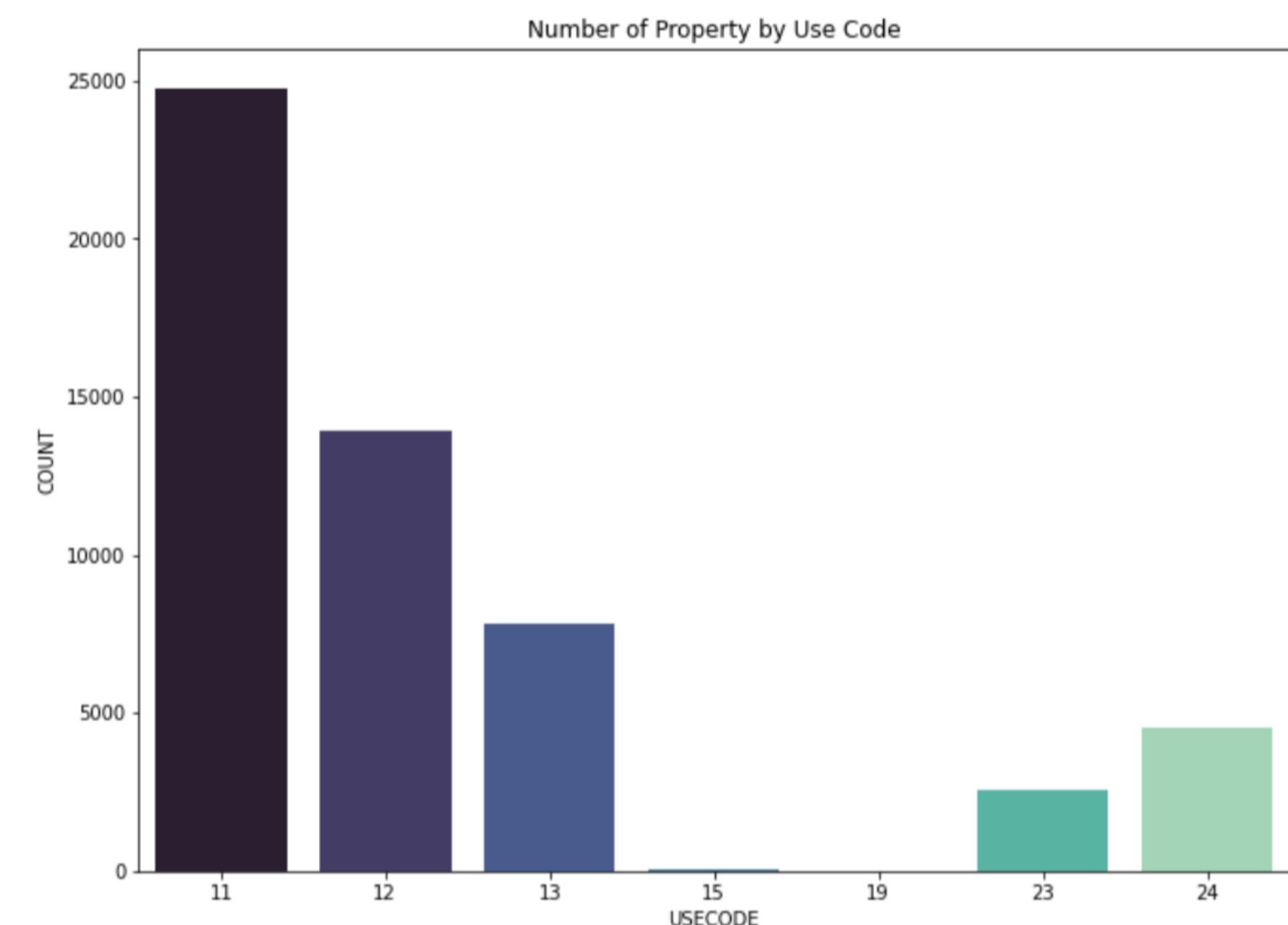
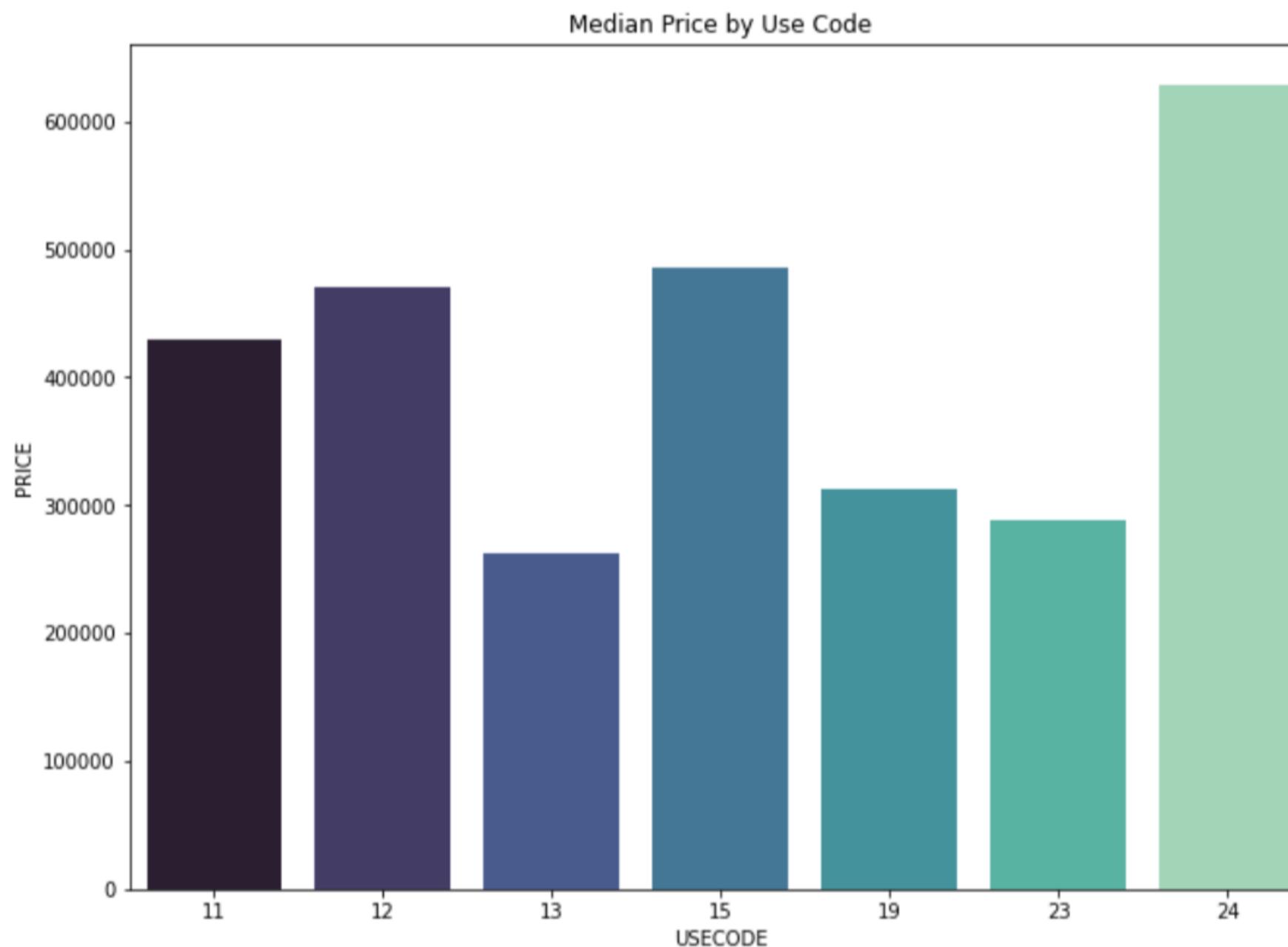
# Median Sale Price per Year



# 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 baseline models of regression model
- Hyperparameter tuning with GridSearchCV to get best parameters of the best baseline-model

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

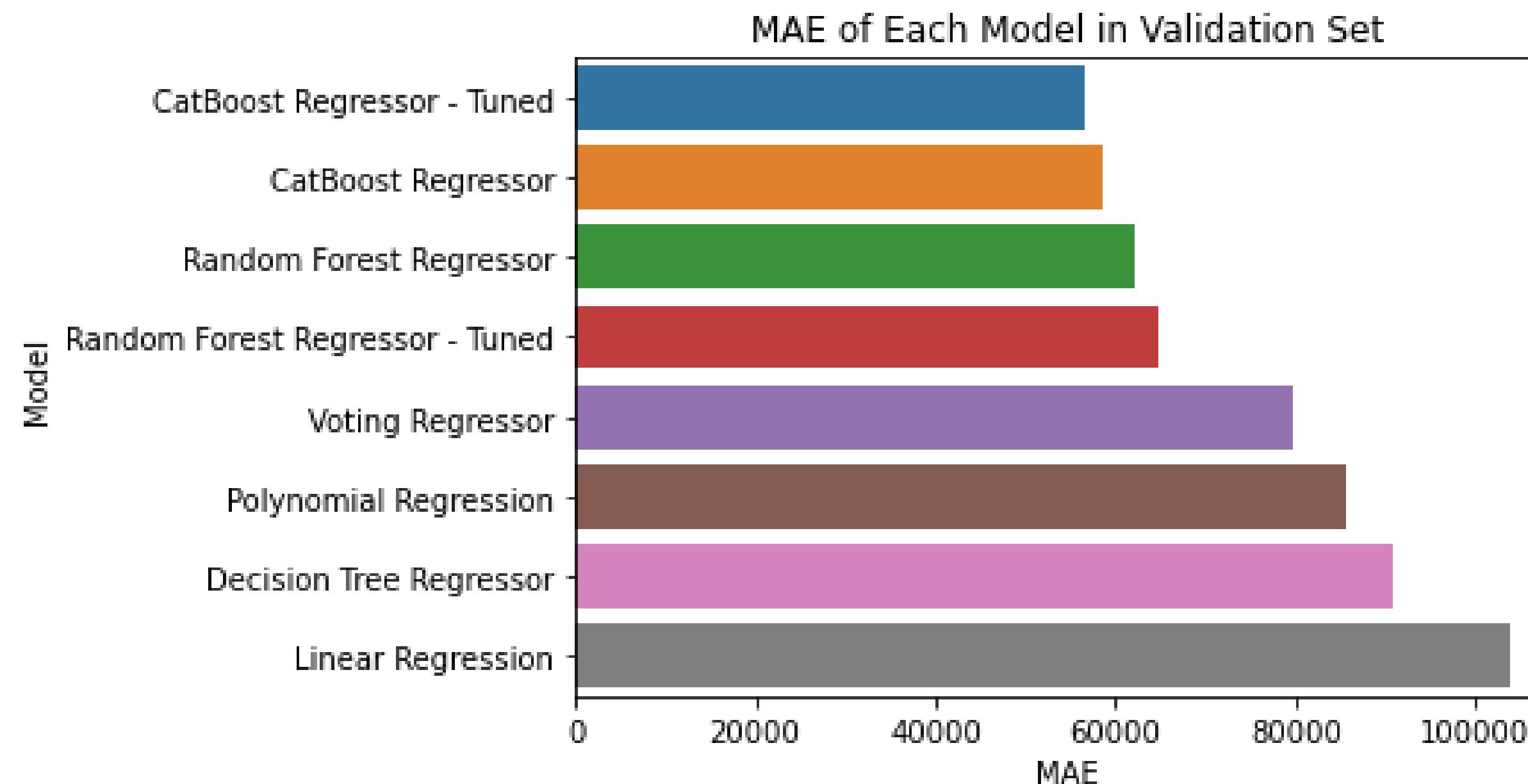
# Model Assessment

CatBoost Regressor - Tuned has the best R<sup>2</sup> score and the lowest MAE in the validation set.

	Model	Set	MSE	RMSE	MAE	R2
1	Linear Regression	Validation	1.903111e+10	137953.273042	103854.302812	0.805568
3	Polynomial Regression	Validation	1.385206e+10	117694.756695	85578.223564	0.858480
5	Decision Tree Regressor	Validation	1.946054e+10	139501.048949	90685.029856	0.801180
7	Random Forest Regressor	Validation	9.184056e+09	95833.478785	62224.824740	0.906171
9	Random Forest Regressor - Tuned	Validation	9.609771e+09	98029.440223	64903.556443	0.901821
11	Voting Regressor	Validation	1.237256e+10	111231.996331	79811.893996	0.873595
13	CatBoost Regressor	Validation	7.999273e+09	89438.655723	58547.042234	0.918275
15	CatBoost Regressor - Tuned	Validation	7.897812e+09	88869.633101	56658.213149	0.919312

# Model Assessment

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



# Final Model Evaluation

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

## Train Set

MSE: 6079714453.015866  
RMSE: 77972.52370557122  
MAE: 47183.05421572207  
R-squared: 0.9380661353784469

## Test Set

MSE: 7307725567.58184  
RMSE: 85485.23596260257  
MAE: 55157.87421335401  
R-squared: 0.9249561938597425

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 : 0.9157272844086888  
Std : 0.003736664180184315

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 13% of the median property price.

Price median : 410000.0

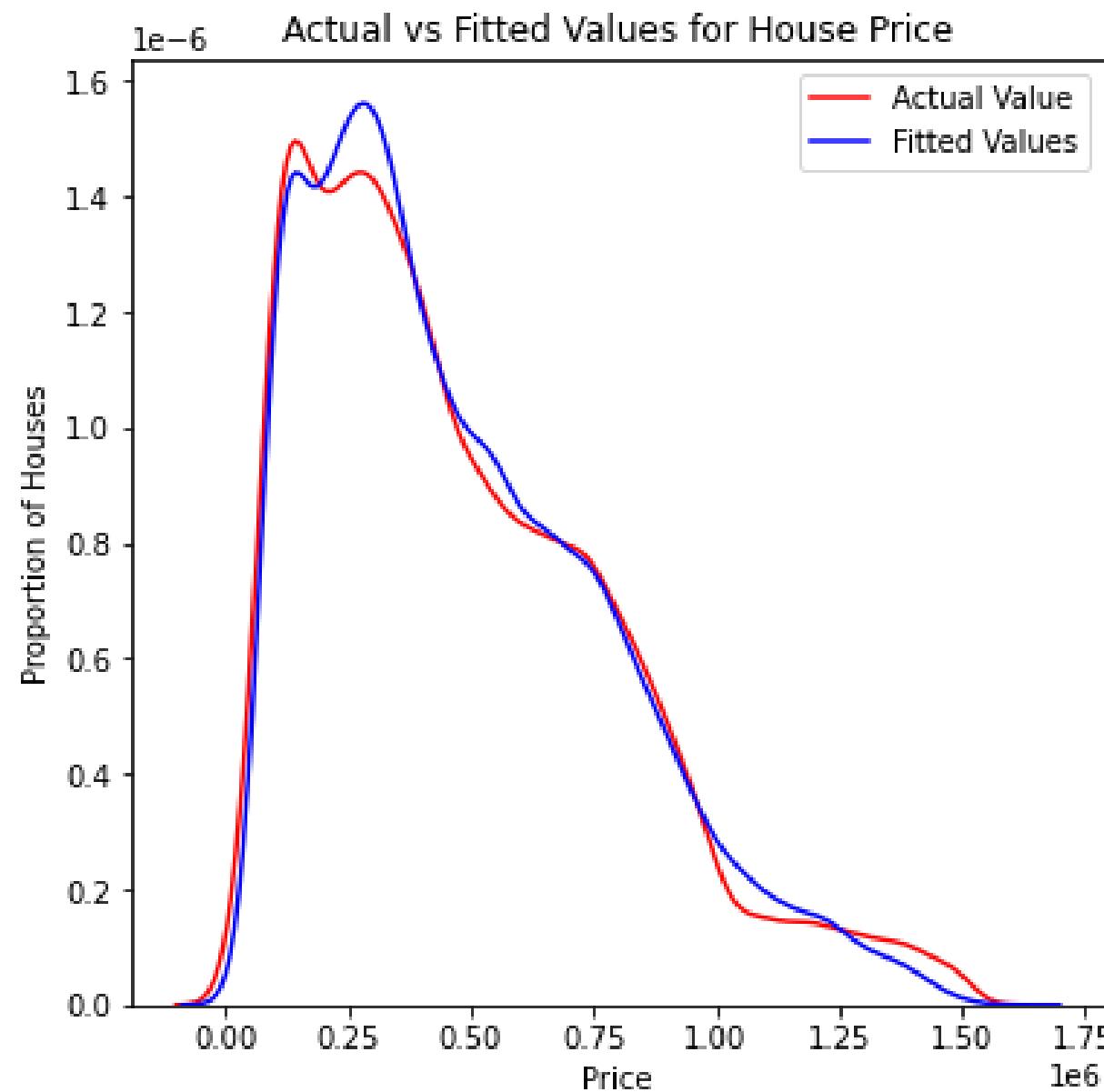
Desired MAE (13% \* median) : 53300.0

Achieved MAE : 47183.05421572207

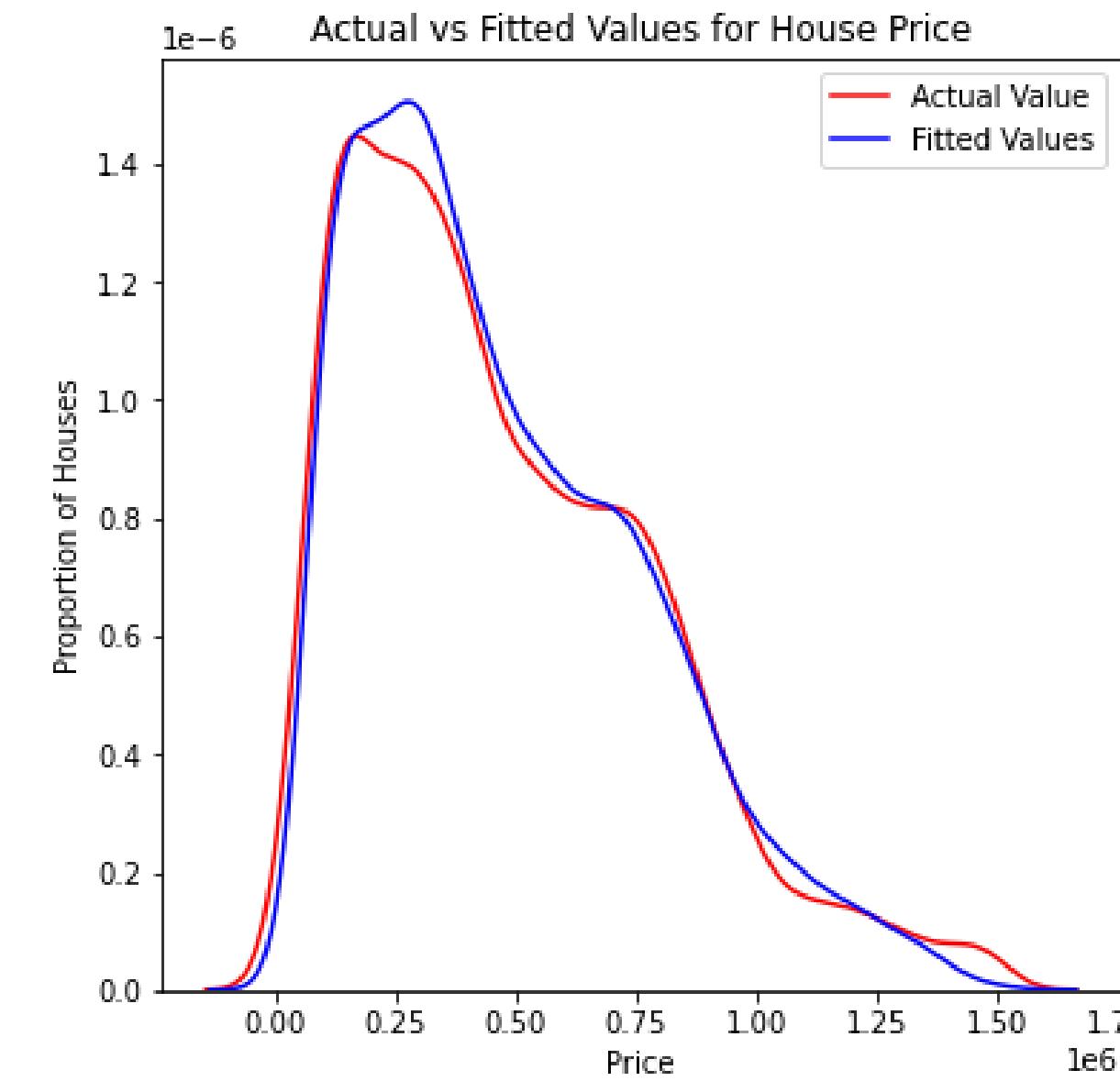
From the result shown above, we have achieved the desired MAE value (under USD53,300).

# Distribution Plot of Actual vs Fitted Values

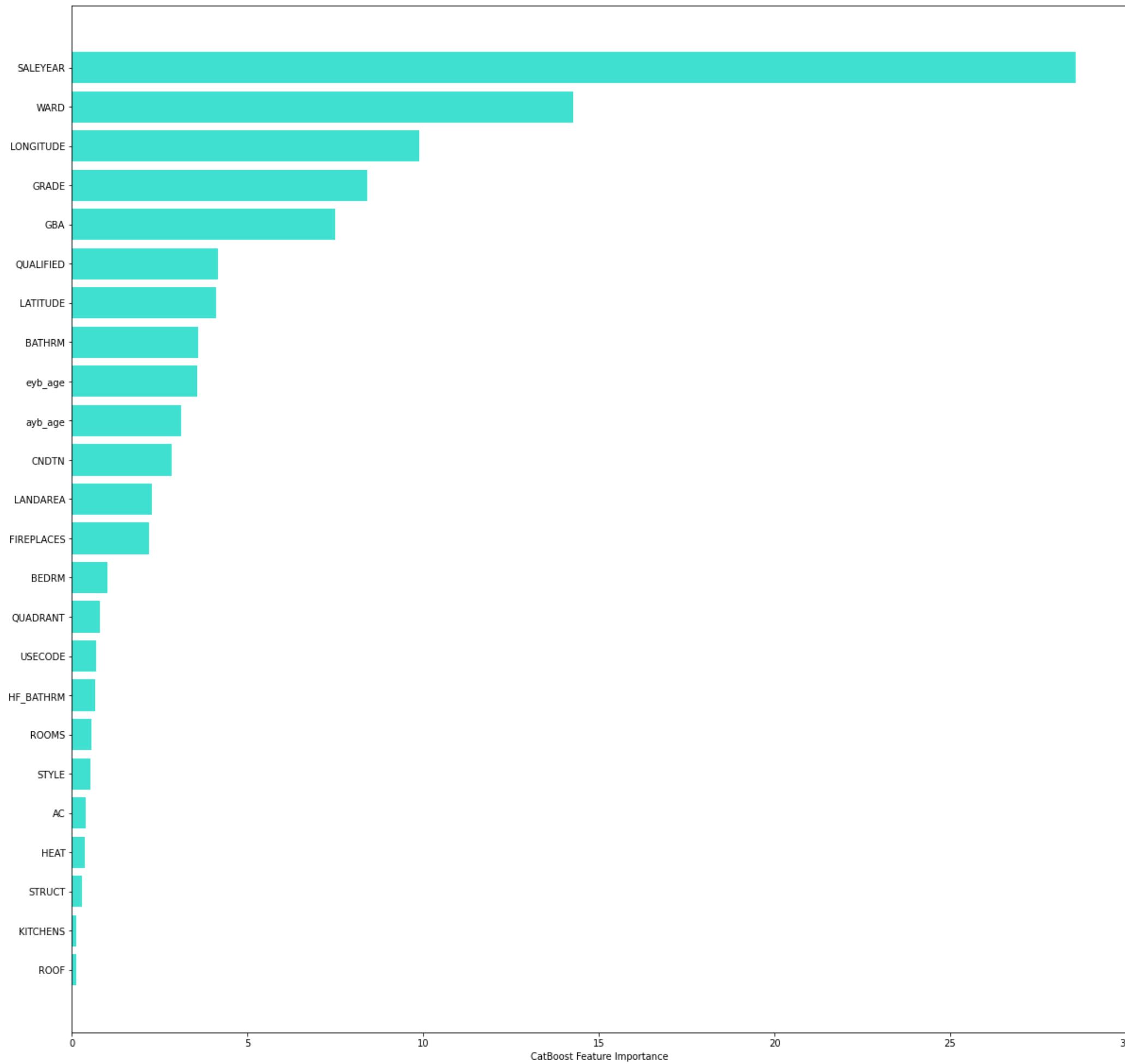
Train Set



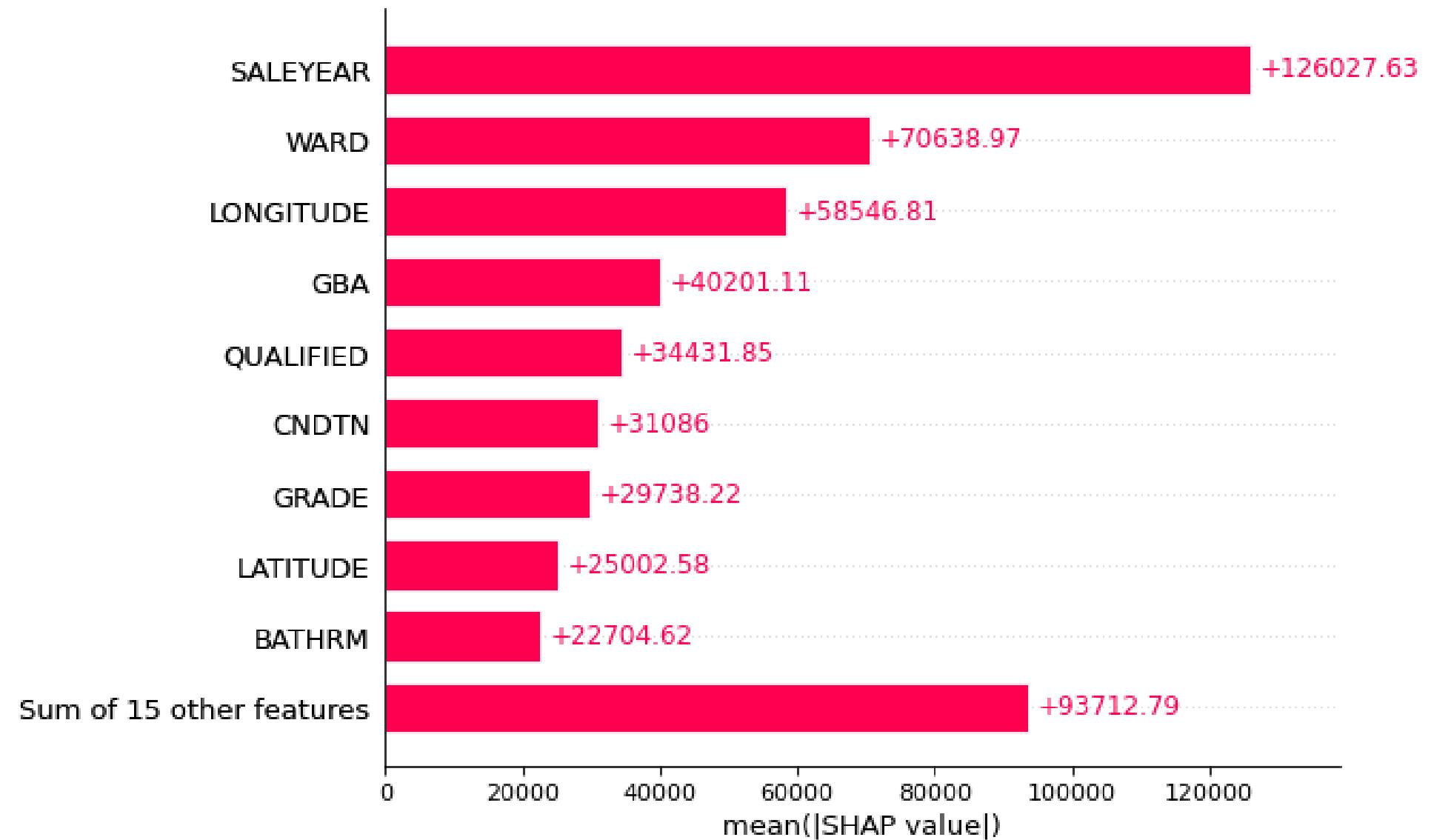
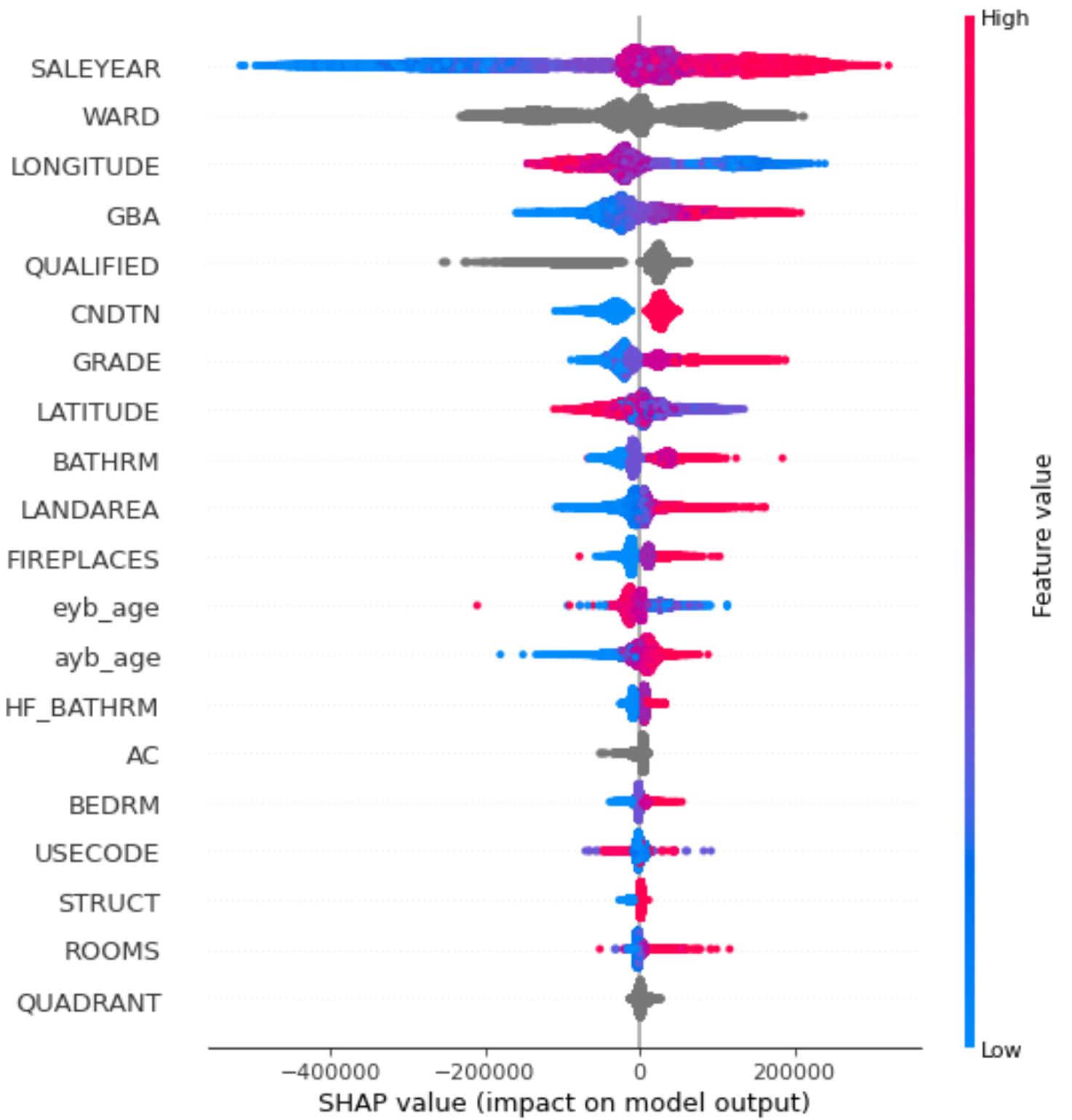
Test Set



# 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
- Risk
- Performance Measure
- Action
- Value

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.

The risk that may be caused by wrong prediction from the ML model is profit loss especially when the model gives underappraised value (price).

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 ( $R^2$ ).

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**

Washington DC, USA.

MPL Bank Matplotlib Team Our Project

Lis Cory Kemal Isfan Rezki Fauziansyah

Home About Estimator Insights

© 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

**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

**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

**Matplotlib Team**

Washington DC, USA.

MPL Bank Matplotlib Team Our Project

Lis Cory Kemal Isfan Rezki Fauziansyah

Home About Estimator Insights

© Copyright 2021 Matplotlib Team. Designed by TemplateMo

**MPL Bank**

Home About Estimator Insights

**Matplotlib Team**

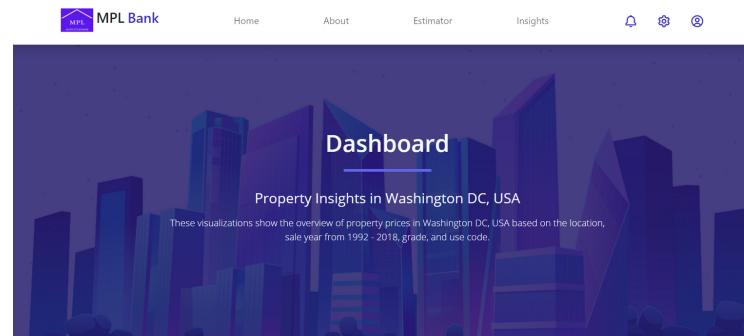
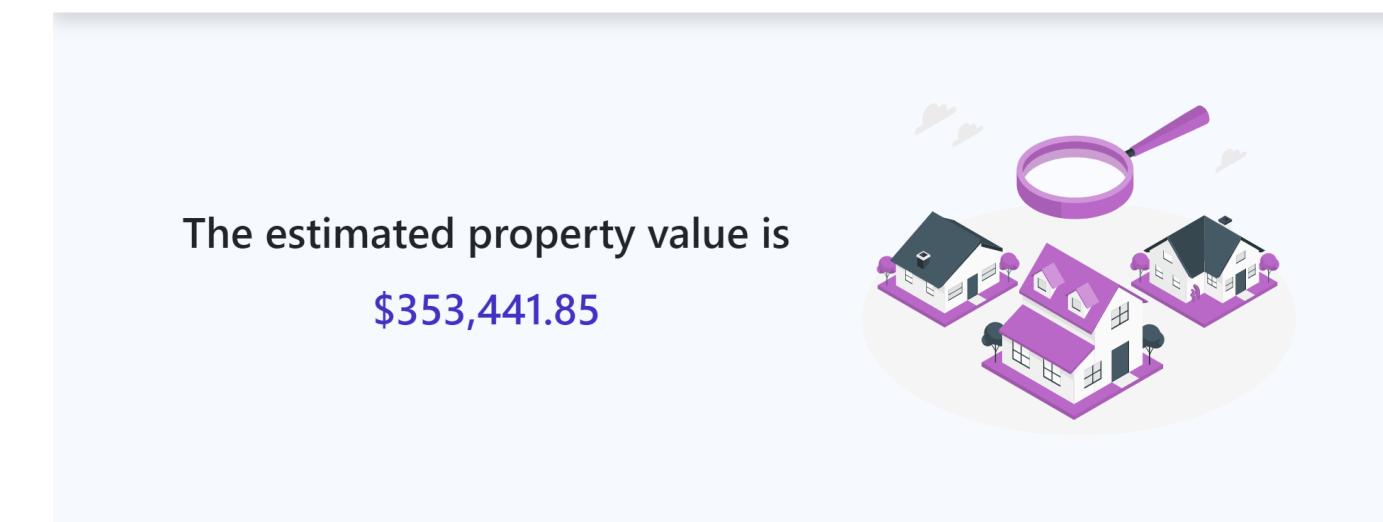
Washington DC, USA.

MPL Bank Matplotlib Team Our Project

Lis Cory Kemal Isfan Rezki Fauziansyah

Home About Estimator Insights

© Copyright 2021 Matplotlib Team. Designed by TemplateMo

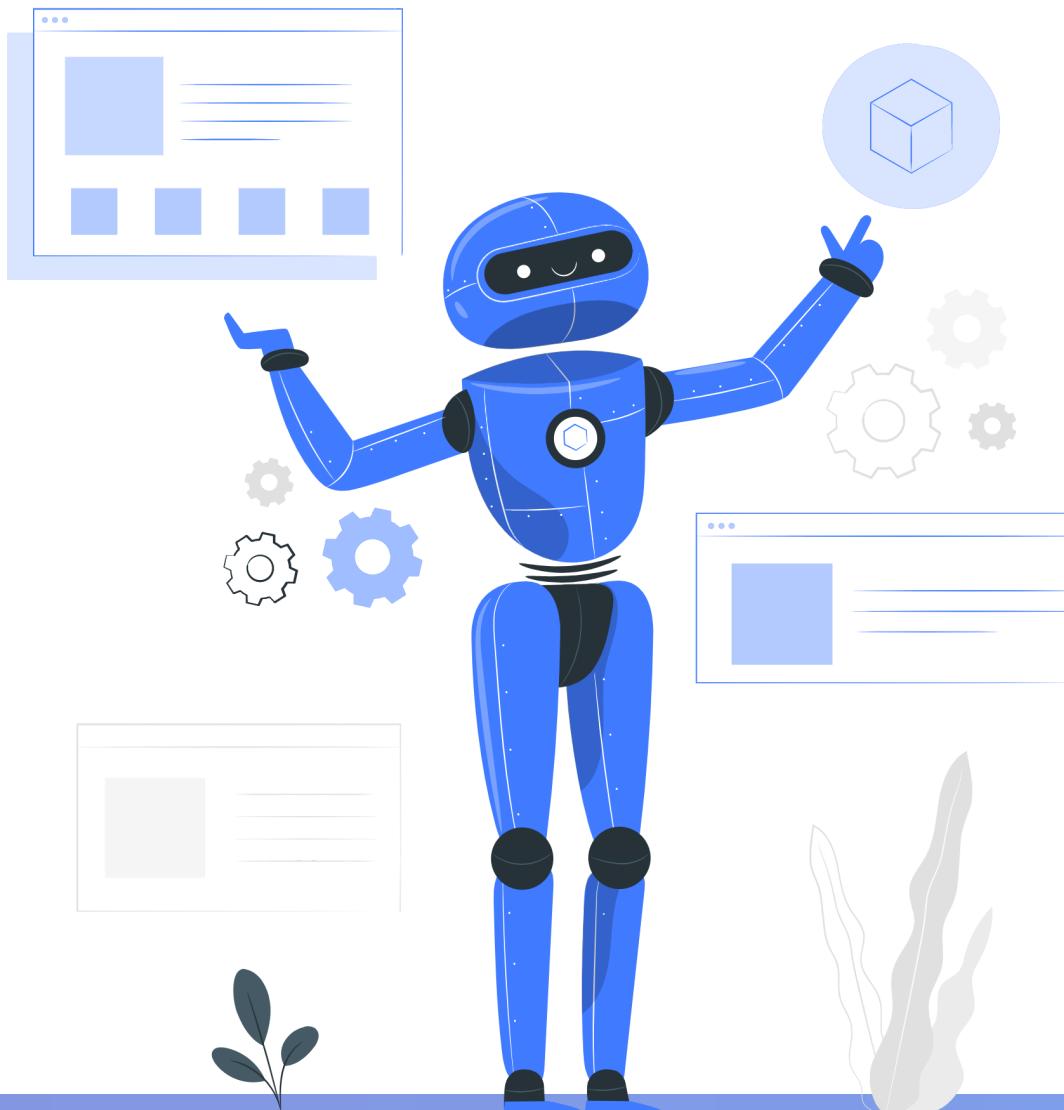


# 5. Future Works



# 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.

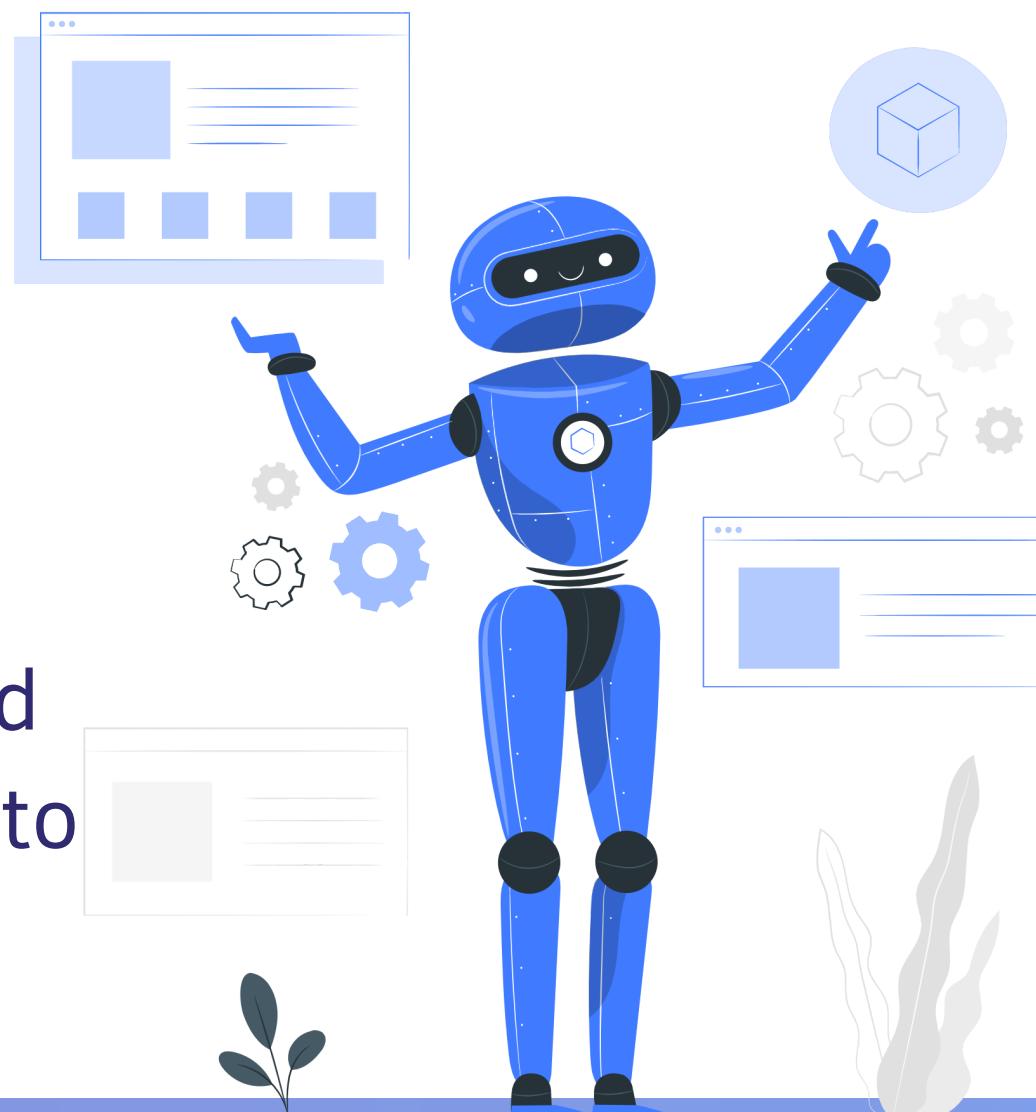


# Future Works

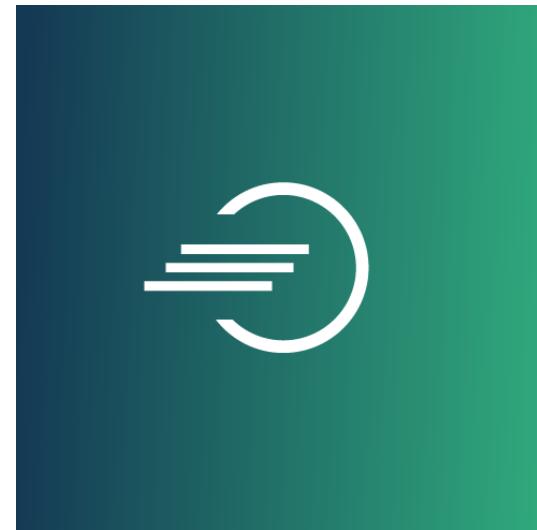
These are a few things that might help to improve model prediction result further :

- 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 .xlsxfile 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>