Impact of Proximity to Subway and Parks on Property Values in Los Angeles

- Wesley Hall

Springboard Data Science Capstone Project

September 27, 2024

Problem Statement

- Analyze how proximity to subway stations and parks impacts property values in L.A.

- Critical for real estate investors, city planners, and homeowners.

- Key questions addressed:

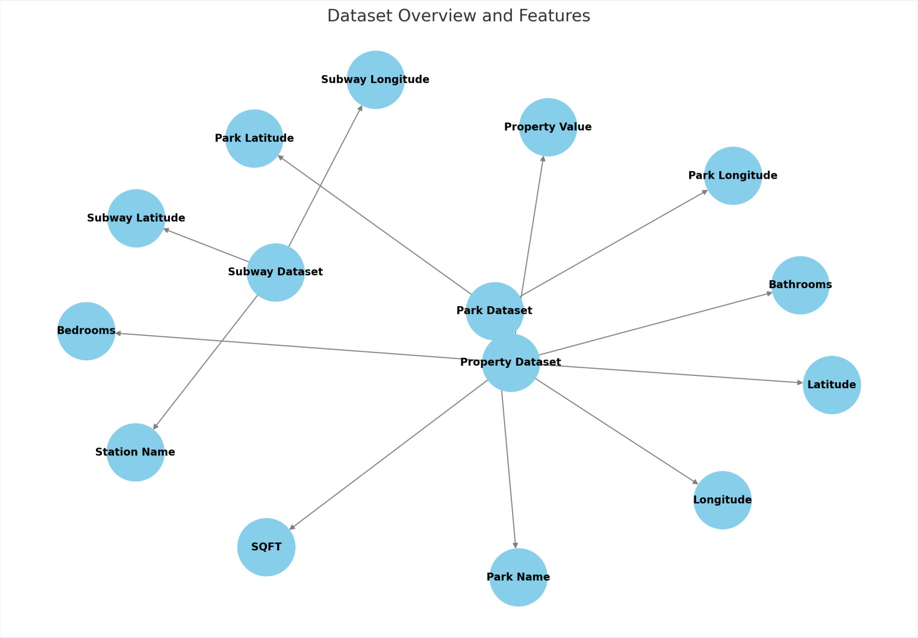
1. Does proximity to subways increase or decrease property values?

2. How do parks influence property prices in nearby areas?

Data Overview

This analysis relies on three main datasets:

- Property Dataset: Includes property details such as value, location (latitude/longitude), square footage, bedrooms, and bathrooms.
- Subway Dataset: Contains the latitude and longitude of subway stations across Los Angeles.
- Park Dataset: Lists the location of parks and green spaces within the city.



Data Preprocessing

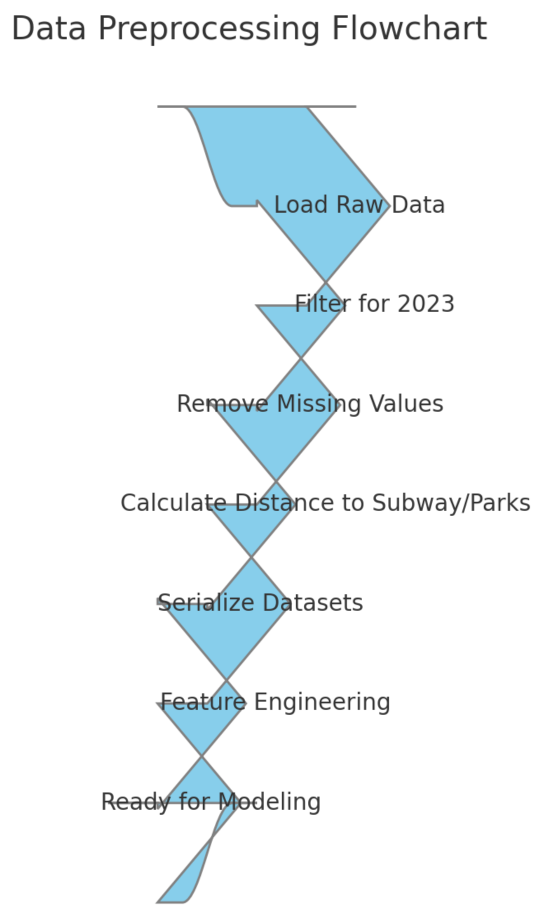
Key steps in preparing the data include:

1. Filtering the property dataset to focus on the year 2023, which provided the most recent property value data and pare our data for efficiency.

2. Calculating the distance from each property to the nearest subway station and park.

3. Removing properties with missing values, reducing the dataset to 2.4 million records.

4. Serializing filtered datasets to improve performance and avoid crashes during analysis.



Feature Engineering

To capture the impact of subway and park proximity, distance features were created. We used the column TaxRateAreas to further smooth property values.

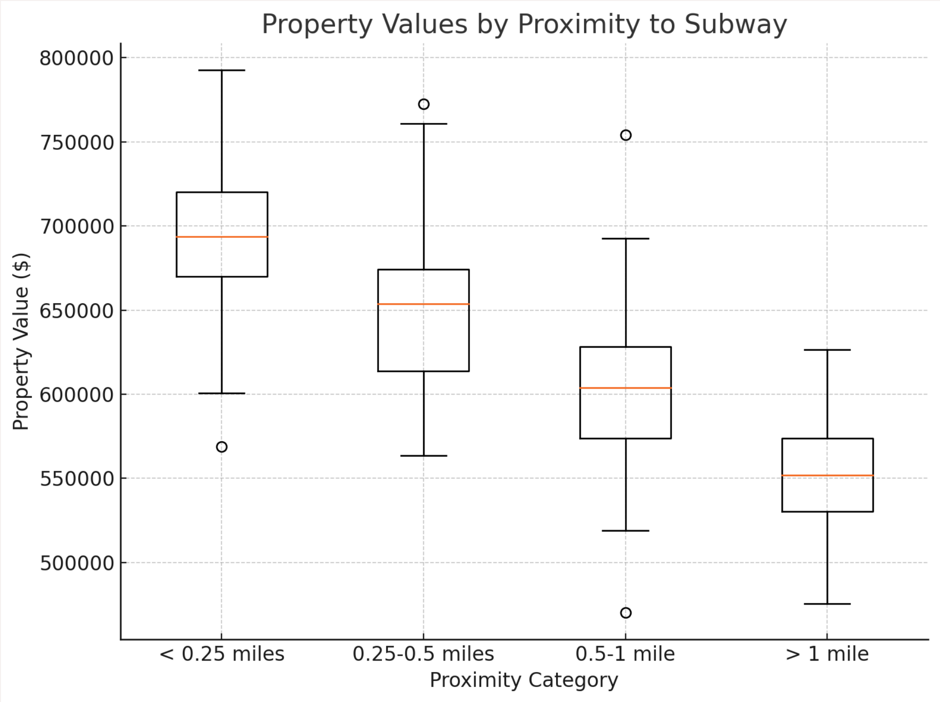
Distance intervals (e.g., properties within 0.25 miles, 0.5 miles, up to 2 miles) from subway stations and parks.

A control group for properties located more than 2 miles from either a subway station or park.

These features helped assess the influence of proximity on property values.

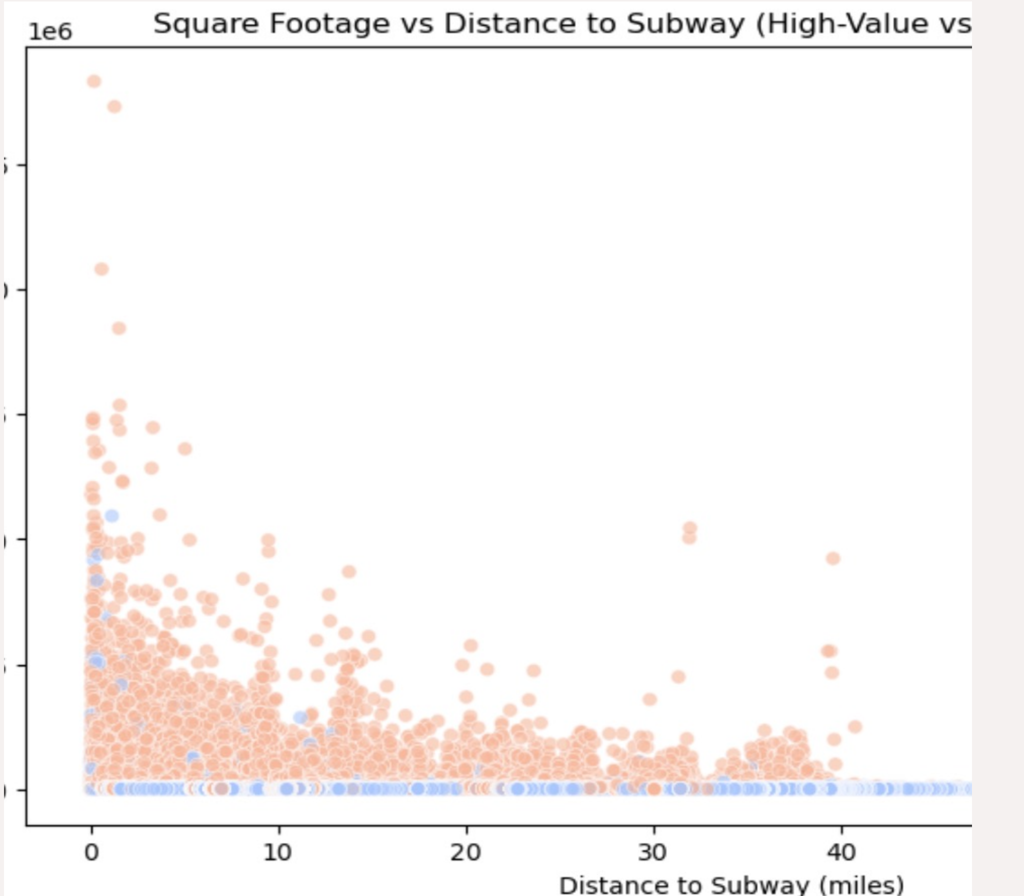
Exploratory Data Analysis (EDA)

- Initial findings from EDA showed:
- Properties closer to subway stations (within 0.25 miles) generally have higher values.
- Parks also positively impact nearby property values, though the effect was weaker compared to subways.
- Outliers were present, particularly in areas with high land values but fewer amenities.



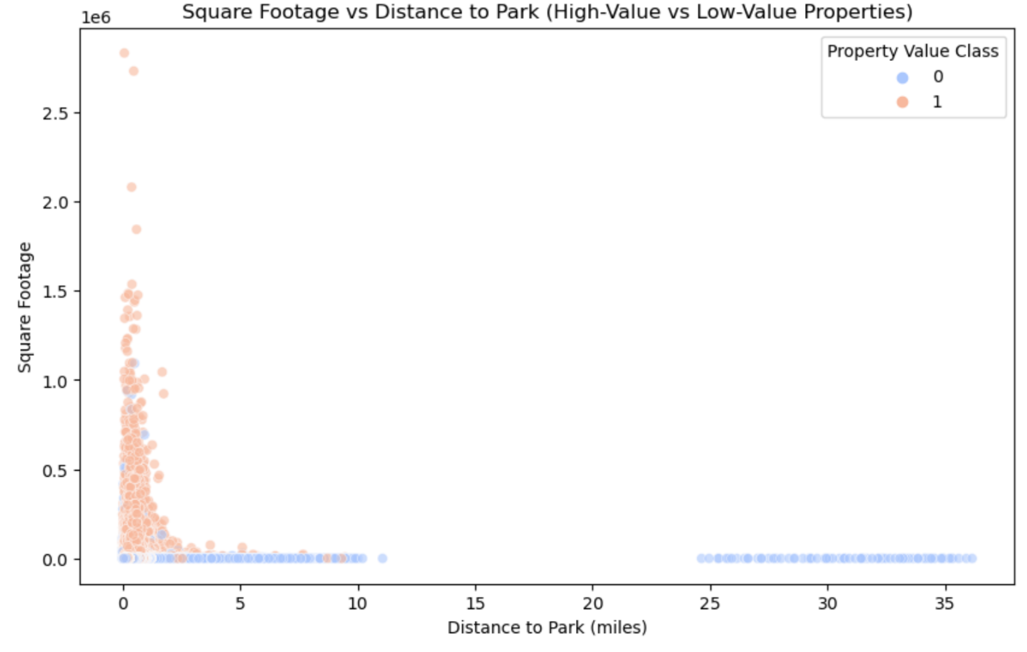
Square Footage vs. Subway Proximity

- Proximity to subways is a factor in valuations.
- Square footage has a much more direct impact on property value.
- High-value properties are often located relatively close to subways and have significantly larger square footage.
- Properties far from subways tend to be lower value, but there are exceptions where large square footage compensates for the distance.



Square Footage vs. Park Proximity

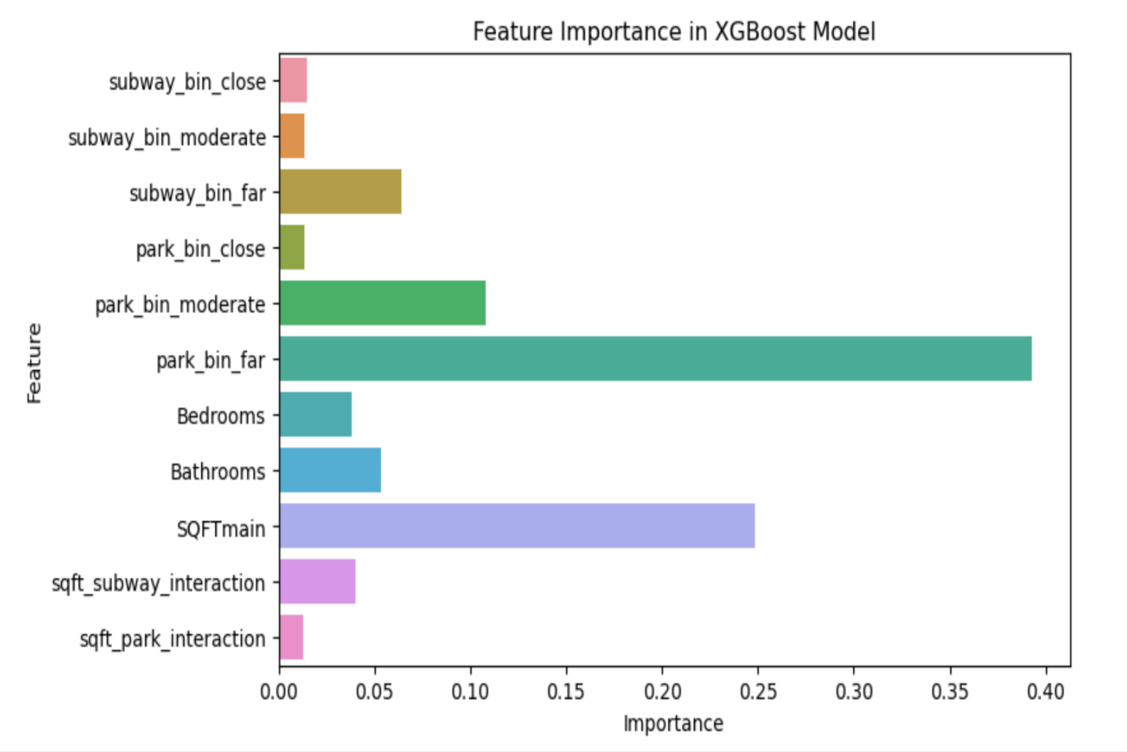
- Proximity to parks appears to have a stronger correlation with high-value properties compared to proximity to subways, where the high-value properties were more spread out.
- Square footage remains a key determinant, with larger properties being more valuable, especially when located closer to park
- Properties with more square footage and closer to parks are likely to have higher values.



Feature Engineering

Plot reveals the following insights from the XGBoost model:

- SQFTmain: size of the property is the primary factor affecting property value.
- Properties closer to parks tend to have higher values.
- Subway proximity is important, but less than nearness to parks.
- Properties with more square footage and closer to parks are likely to have higher values.



Modeling Approach

Train-test split (80% train, 20% test)

Models used: Random Forest, Extra Trees, XGBoost

Final model: XGBoost (best performing model)

Model 1: XGBoost (tuned)

Balanced precision, recall, and F1-scores across both classes

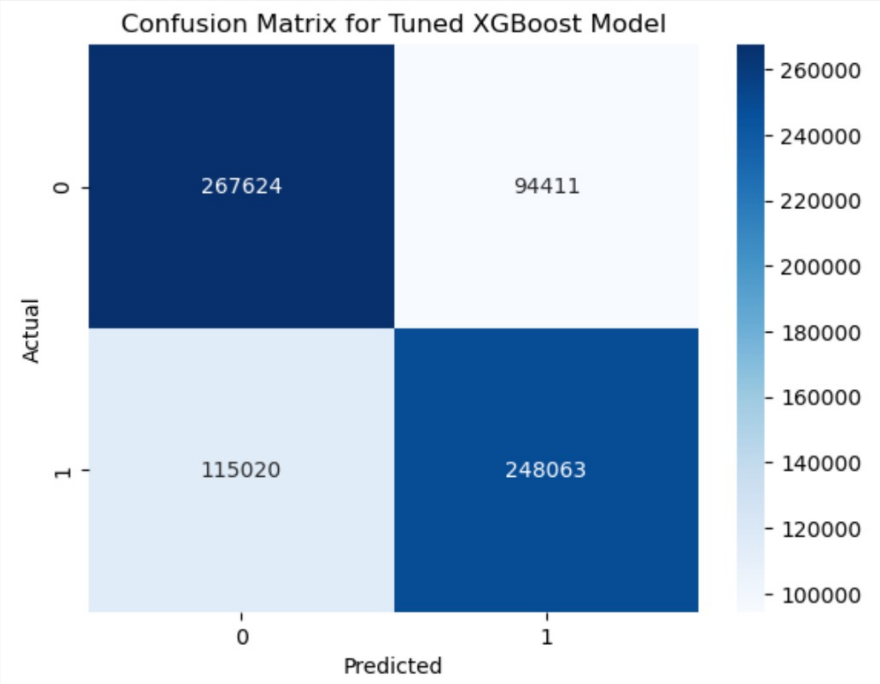
Best hyperparameters for the model are:
n_estimators: 300
max_depth: 10
learning_rate: 0.1
subsample: 1.0
colsample_bytree: 1.0

Tuned XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.70	0.74	0.72	362035
1	0.72	0.68	0.70	363083
accuracy			0.71	725118
macro avg	0.71	0.71	0.71	725118
weighted avg	0.71	0.71	0.71	725118

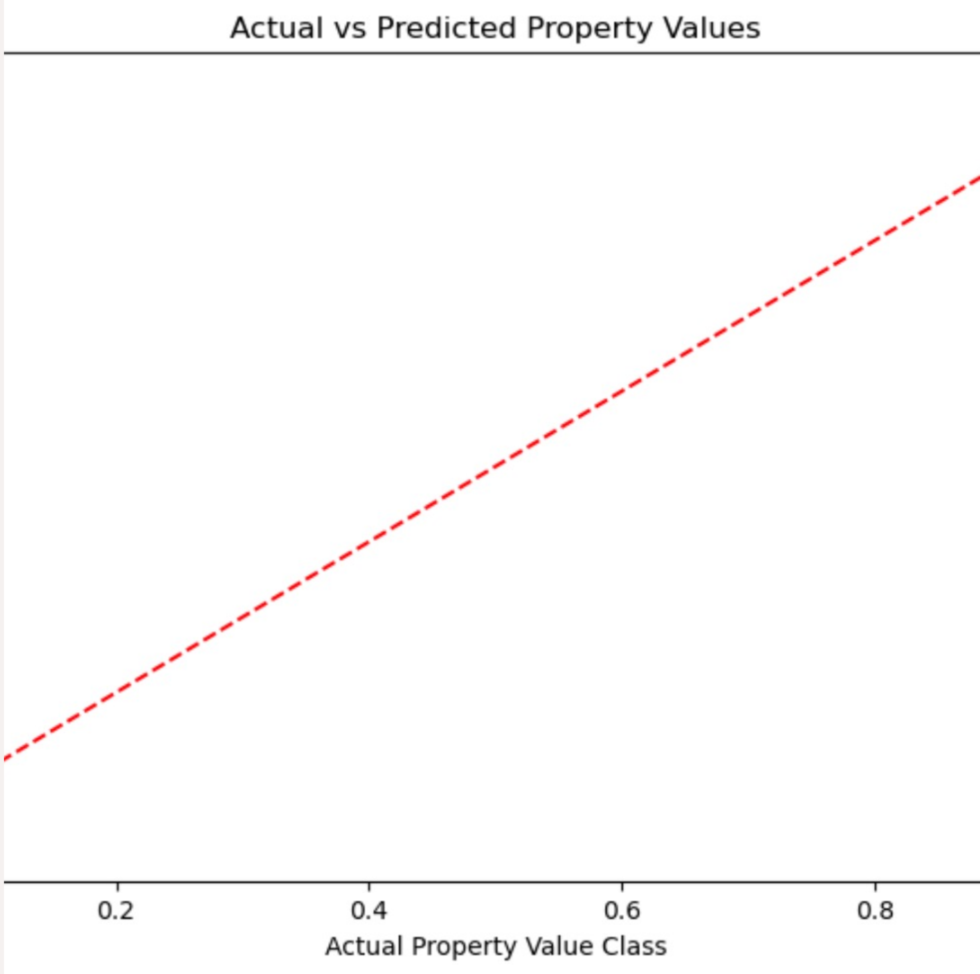
Confusion Matrix for Tuned SGBoost Model

- Our model does well overall.
- Moderate amount of false positives (predicting low-value properties as high-value) and false negatives (predicting high-value properties as low-value).
- This balanced confusion matrix is aligned with the overall accuracy and classification report.



Actual vs Predicted Values

- Near-perfect alignment of actual vs predicted property value classes along the 45-degree line.
- Shows that the XGBoost model is performing well in predicting the property value classes (high-value vs low-value).
- Model appears to accurately classify properties as either high-value or low-value with minimal errors.



Conclusion

Park Proximity:

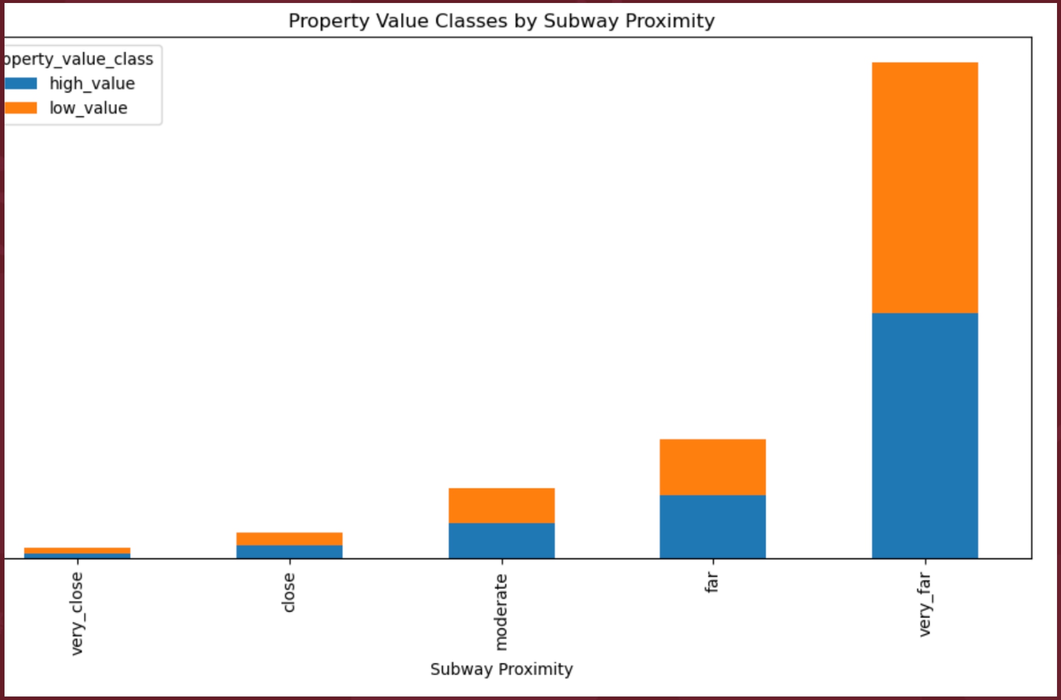
Properties located very close to parks tend to have higher values.

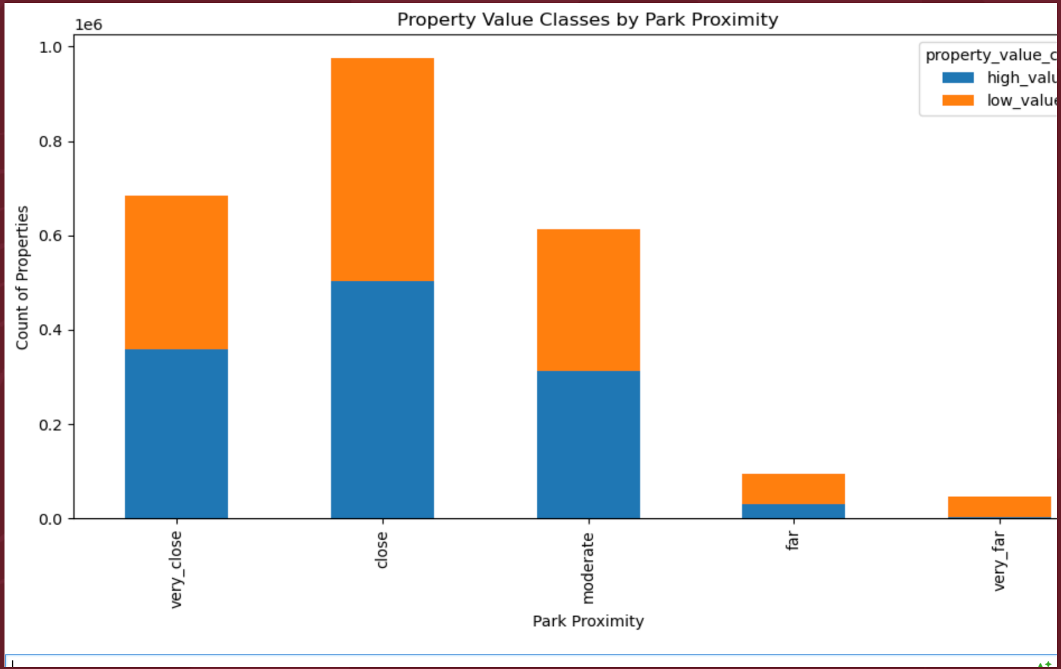
Nearness to green spaces is positively correlated with property value.

Subway Proximity:

Properties farther from subway stations tend to have higher values.

This suggests subways may not drive higher property values in this dataset, possibly reflecting different urban or economic factors.





Next Steps

Explore Additional Features: Analyze the impact of other amenities (schools, shops) and neighborhood factors (income, crime) on property values.

Advanced Feature Engineering: Test interaction terms and non-linear transformations for deeper insights.

Refine the Model: Experiment with new algorithms (SVM, Neural Networks) and use feature selection techniques to improve model accuracy.

Predict Future Trends: Conduct scenario analysis and make projections about future property values with upcoming infrastructure changes.

Deploy the Model: Propose deploying the model in a web-based dashboard for real estate professionals and urban planners.

Expand Geographic Scope: Apply the model to other regions or cities to assess the generality of our findings.