DS110 Final Project

Exploring the Connection Between Open Spaces and Crime Rates in Boston Neighborhoods

Anmar Abdi & Owen Chappellet

Introduction

Our main goal of this project was to find out if there is any relationship between the percent of open space and daily crime rates in urban neighborhoods. We had an initial hypothesis that neighborhoods with more open spaces would have less crime, so proving this correlation could prove to be useful for urban planners and policymakers to make better decisions. This in turn would hopefully reduce the crime rate. It is important to note that the crime rate is impacted by a huge amount of variables and the solution is much bigger than just adding more open spaces to neighborhoods.

Methodology

To perform the investigation, we first had to prepare the data. We used four publicly available data sets:

- BPD Crime Incident Reports - https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system
- Boston Open Spaces - https://data.boston.gov/dataset/open-space
- GEOJson file of BPD Districts - https://data.boston.gov/dataset/police-districts
- BPD Offense Codes - https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system/resource/3aeccf51-a231-4555-ba21-74572b4c33d6

As well as an interactive map of Boston neighborhood populations from https://statisticalatlas.com/place/Massachusetts/Boston/Population

We started by cleaning the crime data sets by merging all the separate years into one combined crime data set (2015-2023). We then dropped all unnecessary columns then matched the neighborhood names with the District code from the BPD Districts file. We then calculated the area of each neighborhood using the GEOJson file and hand calculated the population for each neighborhood using the interactive map from *statisticalatlas*. Once we merged those two data sets we had to filter out the irrelevant crimes. We went through the Offense Codes data set and removed all codes related to crimes that wouldn't have a direct impact on public safety. After filtering out the irrelevant crimes we then found the total count of crimes in each neighborhood and calculated the crime rate per capita to normalize the crime rates. We then merged it with the count and area of open spaces in each neighborhood then found the percentage of open

spaces relative to the area of each neighborhood. We then calculated the amount of crimes reported per day for each neighborhood and found the daily crime rate by dividing the amount of crimes per day by the population. After this, we split the set into training and testing data, 2015-2022 and 2023 respectively. We tried out several machine learning models including Decision Trees, Random Forest Regression, and K-Nearest Neighbors, and found that the Random Forest model was the most accurate when tested.

```
[ ] # Good Random Forest Model!!!
    from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
    X = train_df[['Population', 'Population_Density', 'Open_Space_Percentage']]
    y = train_df['Crime_Rate']

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
    rf_model.fit(X_train, y_train)

    y_pred_rf = rf_model.predict(X_test)

    mae_rf = mean_absolute_error(y_test, y_pred_rf)
    mse_rf = mean_squared_error(y_test, y_pred_rf)
    rmse_rf = np.sqrt(mse_rf)
    r2_rf = r2_score(y_test, y_pred_rf)

    print("Random Forest Model (Test Data: 2015-2022):")
    print(f"Mean Aabsolute Error: {mae_rf}")
    print(f"Mean Sq. Error: {mse_rf}")
    print(f"Root MSE: {rmse_rf}")
    print(f"R2 score: {r2_rf}")
```

```
Random Forest Model (Test Data: 2015-2022):
Mean Aabsolute Error: 0.00023708926000062789
Mean Sq. Error: 1.6783742373668885e-07
Root MSE: 0.0004096796590110965
R2 score: 0.8051928863768013
```

```
[ ] X_test_2023 = test_df[['Population', 'Population_Density', 'Open_Space_Percentage']]
    y_test_2023 = test_df['Crime_Rate']

    y_pred_rf_2023 = rf_model.predict(X_test_2023)

    mae_rf_2023 = mean_absolute_error(y_test_2023, y_pred_rf_2023)
    mse_rf_2023 = mean_squared_error(y_test_2023, y_pred_rf_2023)
    rmse_rf_2023 = np.sqrt(mse_rf_2023)
    r2_rf_2023 = r2_score(y_test_2023, y_pred_rf_2023)

    print("Random Forest Model (2023 data):")
    print(f"Mean Aabsolute Error: {mae_rf_2023}")
    print(f"Mean Sq. Error: {mse_rf_2023}")
    print(f"Root MSE: {rmse_rf_2023}")
    print(f"R2 score: {r2_rf_2023}")
```

```
Random Forest Model (2023 data):
Mean Aabsolute Error: 0.00023378971962957095
Mean Sq. Error: 1.4370444128302654e-07
Root MSE: 0.0003790836863847171
R2 score: 0.7032684970787804
```

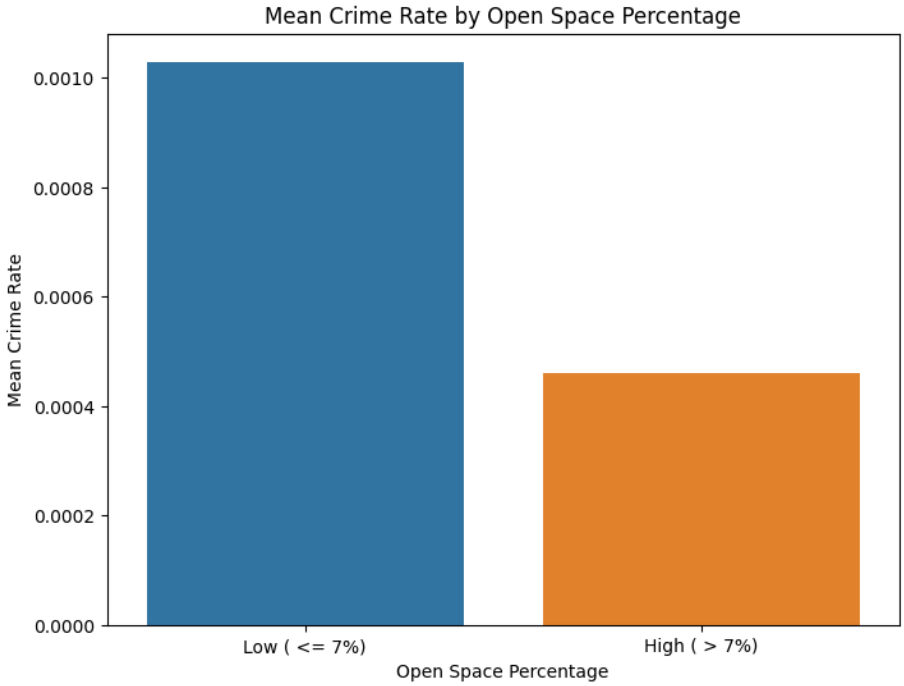
This model tested between population, population density, and open space percentage and found that the open space percentage had the most influence on models predictions.
Now that we know that the open space percentage has an impact on crime we can test it further by answering our main question:

Do neighborhoods with a high percentage of open space have a lower daily crime rate than those with low percentages?

Results

We found that neighborhoods with more than 7% of open space have lower crime rates than those with open space less than or equal to 7%

	Neighborhood	Open_Space_Percentage
1	Charlestown	3.832395
2	Dorchester	4.208172
4	Eastie	4.237573
0	Allston-Brighton	4.353470
9	South Boston	4.705957
7	Mattapan	4.849526
10	South End	5.109626
3	Downtown	6.531396
11	West Roxbury	8.860555
5	Hyde Park	10.083669
6	Jamaica Plain	10.542415
8	Roxbury	11.887019



On the left in blue, we see the districts which contained less than 7% open spaces with over double the amount of crimes than the districts on the right in orange, with higher than 7% open spaces.

Conclusion

In conclusion, our results support the initial hypothesis that there is a discernible relationship between the percent of open space in urban cities and their crime rates. This study could be used as a foundation for further exploration into urban planning and crime reduction strategies, emphasizing the potential impact of open spaces on community safety.

Finally, a key takeaway from this project was the sheer amount of data cleaning we had to do before conducting the analysis. We realized how messy real data is and learned how crucial this step is to gaining quality insights from an analysis.