

Research Questions

1. To what extent does criminal activity influence the housing market in the Zealand and Capital Regions of Denmark?
2. Can we predict a house's price based on the local crime rate? What is/are the predictor/predictors?

This research consists of two notebooks:

1. Preprocessing housing data
2. Analysis

Preprocessing housing data

[preprocessing_housing_data.ipynb](#)

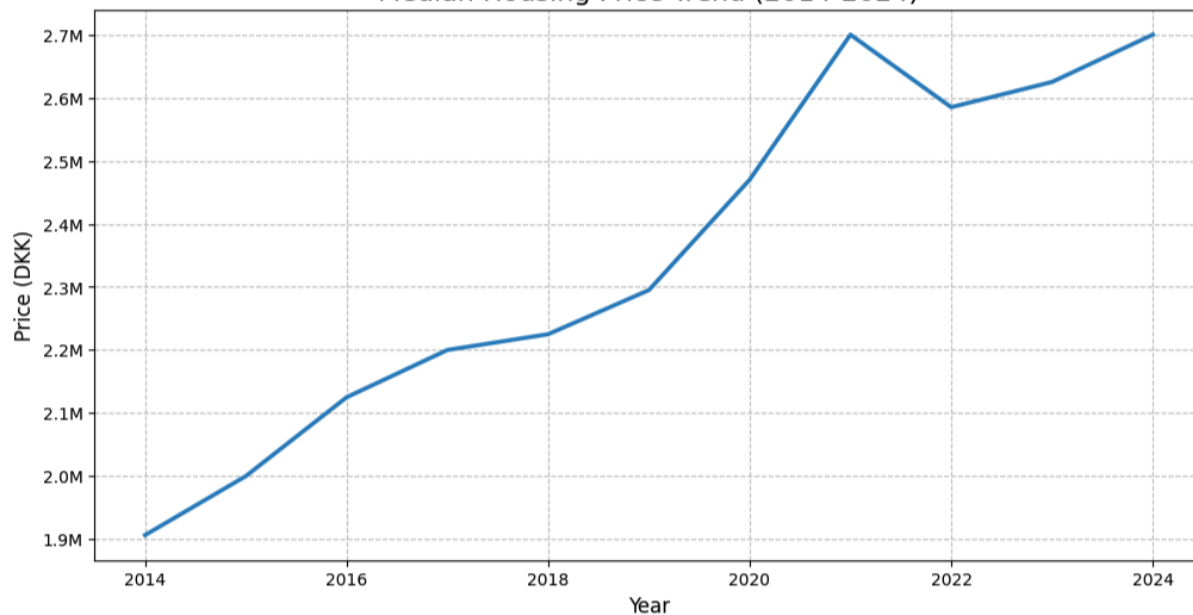
We used a public dataset “*residential house prices from 1992 to 2024*” containing records of property sales that happened in Denmark. The dataset was in a parquet file which was read using the appropriate pandas function. We explored the nature of the dataset by checking the types of the columns, the unique areas and regions and their frequencies, as well as if there are any null values. We found two null values in the columns `sqm`, `sqm_price` but those didn't affect our analysis. That is because we decided to only keep these columns: `date`, `year_build`, `purchase_price`, `city`, `region`. We came to this decision because we care about when the property was sold and at what price, in order to combine it with data from the crimes dataset. Also the city and region columns were useful in order to define the specific municipalities that these properties belong to.

Cleaning

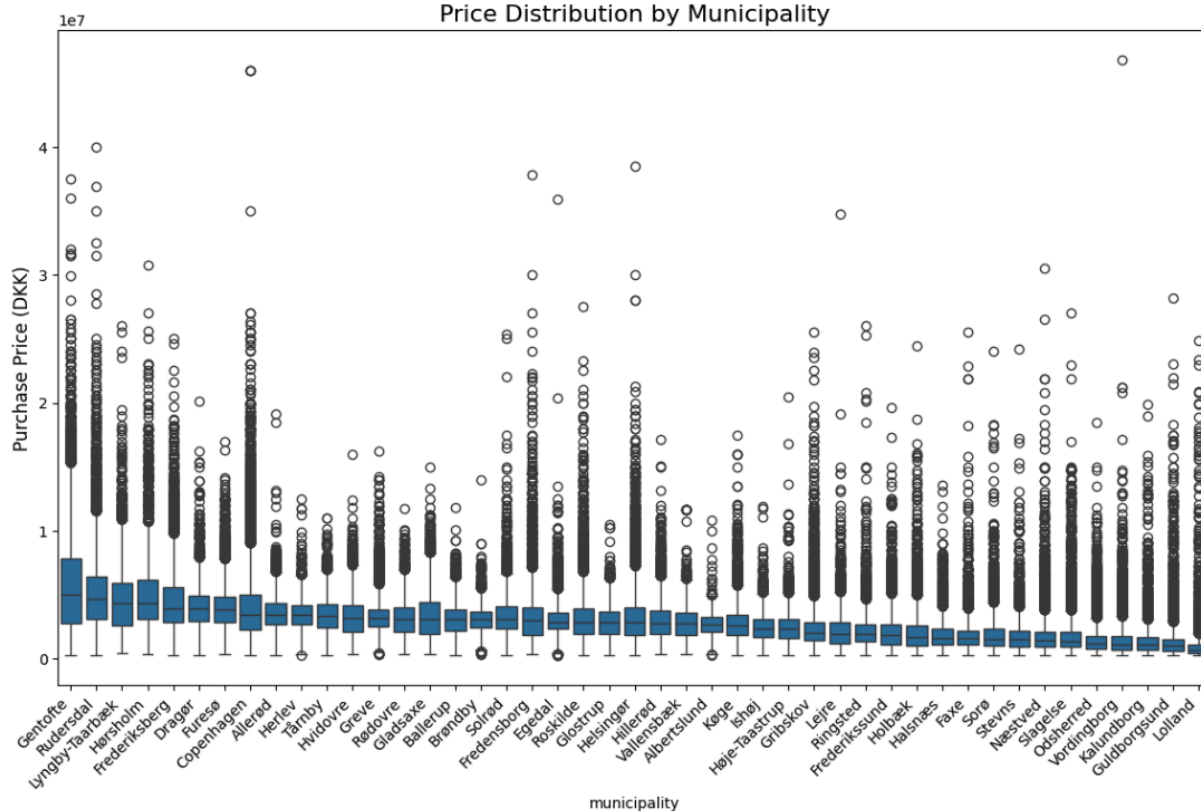
We decided to keep data from the last decade because we wanted to focus our research on the most recent records to better understand the most recent trends. Then we kept only records from Zealand, focusing on the most populated area of Denmark. Based on that we filtered the dataset and created a new one with the new data.

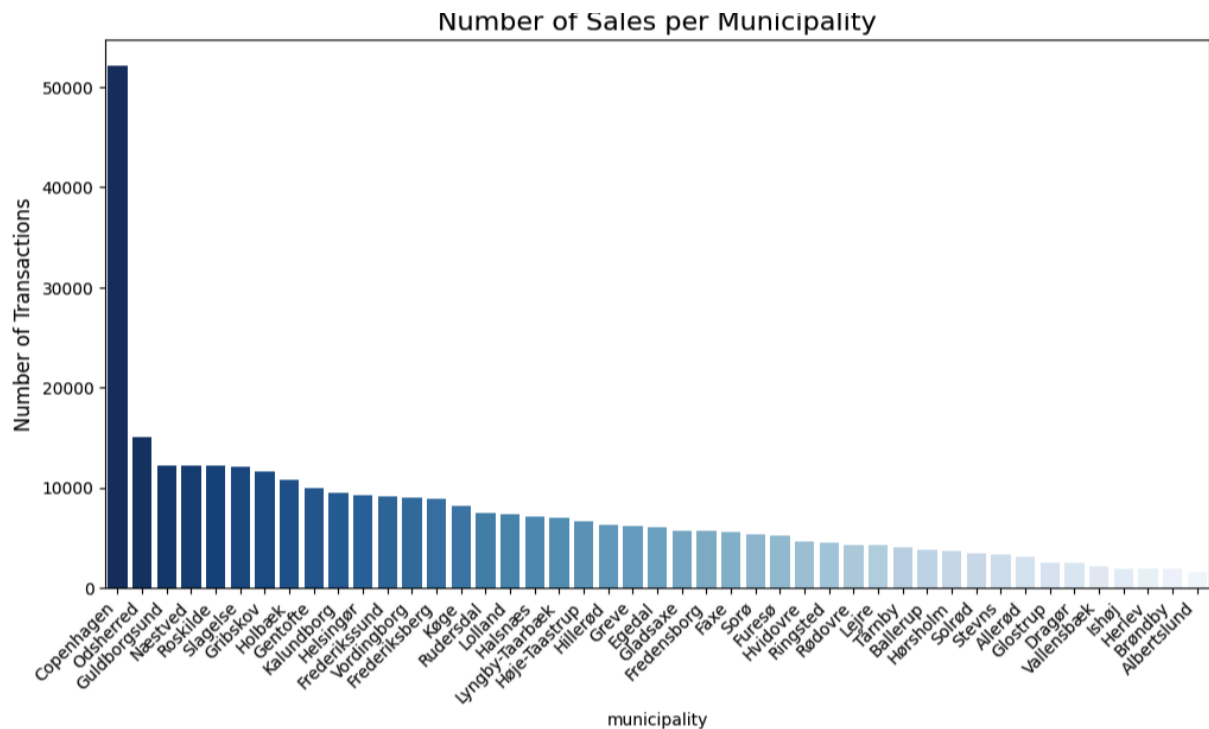
Then we had to match specific cities to municipalities. This had to be done because the other datasets that we used from statbank provide records only for municipalities and not for specific cities. For example, in the housing dataset we had records where the city could be “København K” or “Valby”, but these belong to the same municipality. We achieved this by creating a dictionary where the keys were the 46 municipalities found in Hovedstaden and Zealand, and the corresponding values were cities as written in the housing data. By doing that we created a new `municipality` column. Then we made some plots to better understand the data we had in our hands.

Median Housing Price Trend (2014-2024)



Price Distribution by Municipality





The plots reveal some valuable insights:

- The median housing prices have approximately a growth of 42% in the last 10 years.
- All the municipalities contained a lot of outliers.
- Non-surprisingly Copenhagen municipality had by far the most sales.

In the end we exported the dataset into a new csv file in order to use it in our analysis.

Analysis

`analysis.ipynb`

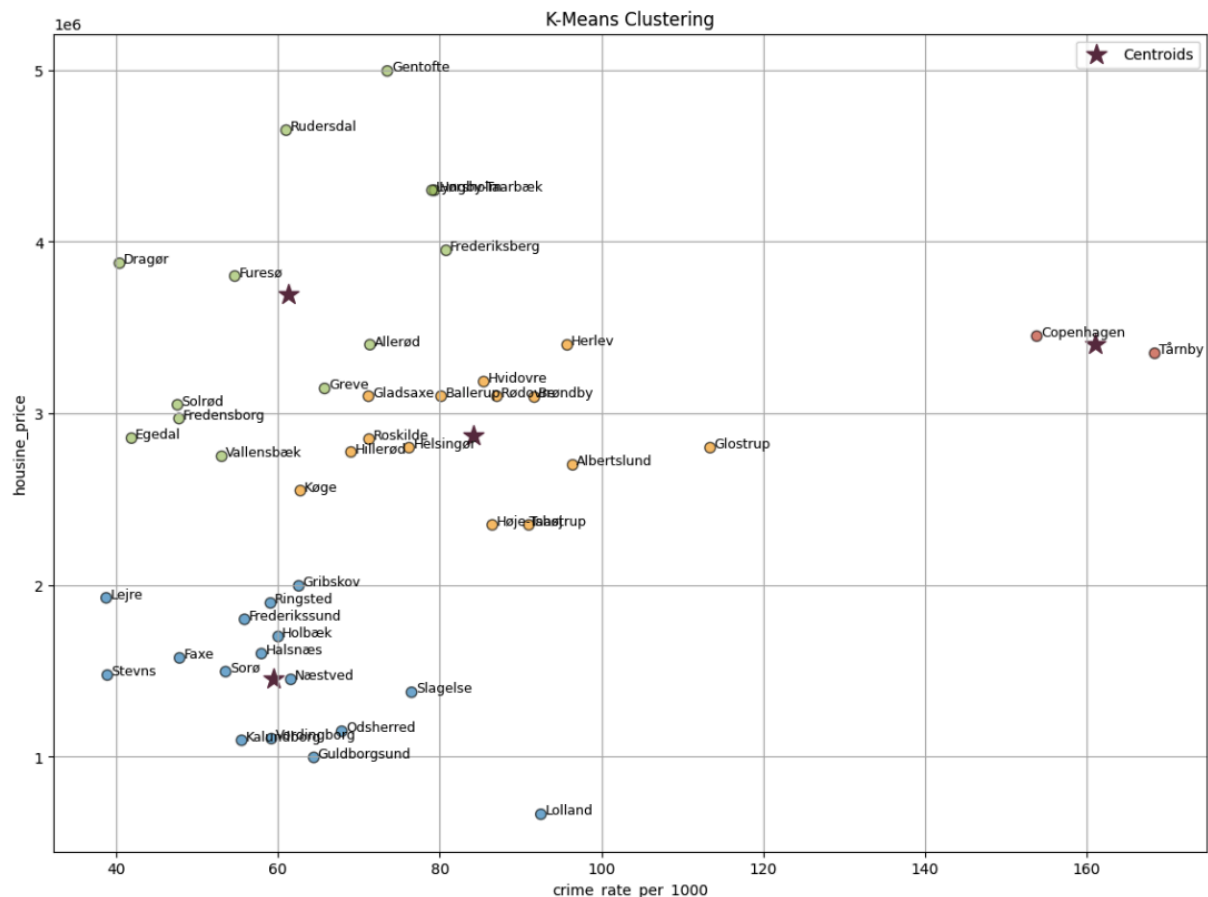
For the analysis we used the crimes dataset that was created from the other research question, because it contains the crime rate per 1000 people and also the percentage of non-danish residents in municipalities, which presented an interesting opportunity to be used in our regressions, alongside with other features.

We checked if both of the datasets have the same number of municipalities. We found only one missing from the housing dataset which was due to the lack of sales records.

We then merged the dataframes on municipality and date/year with an inner joining which created records where for each year we had data about these features: year of build, purchase price, city, municipality, non-danish population, danish population, crime count, total population, non danish ratio, crime rate per 1000.

To what extent does criminal activity influence the housing market in the Zealand and Capital Regions of Denmark?

In order to research this question we decided to use the clustering method and more specifically the K-Means. First we aggregated our data by grouping by the records on the municipality feature and calculating the median purchase price and crime rate of the past 10 years. Then we normalized the data. We used the elbow method to calculate the best number of clusters which was 4.



Cluster 0 (Bottom Left)

- Characteristics: Low Prices (~1m - 2m DKK) and Low-to-Medium Crime.
- Examples: Lolland, Guldborgsund, Kalundborg.
- Insight: These areas are safe, but economically less active. Low crime alone does not drive high prices here.

Cluster 1 (Top Left)

- Characteristics: High Prices (~3m - 5m DKK) and Low Crime.
- Examples: Gentofte, Rudersdal, Frederiksberg, Dragør.
- Insight: This is the "Ideal" market where high safety correlates with high value. Gentofte is the extreme outlier here—the most expensive municipality by far.

Cluster 2 (Right Side)

- Characteristics: Moderate Crime and Moderate Prices.
 - Examples: Hvidovre, Brøndby.
 - Insight: These are municipalities where the prices are close to the median and the crime rate is mostly less than the median.
- This cluster neighbours very closely to Cluster 1

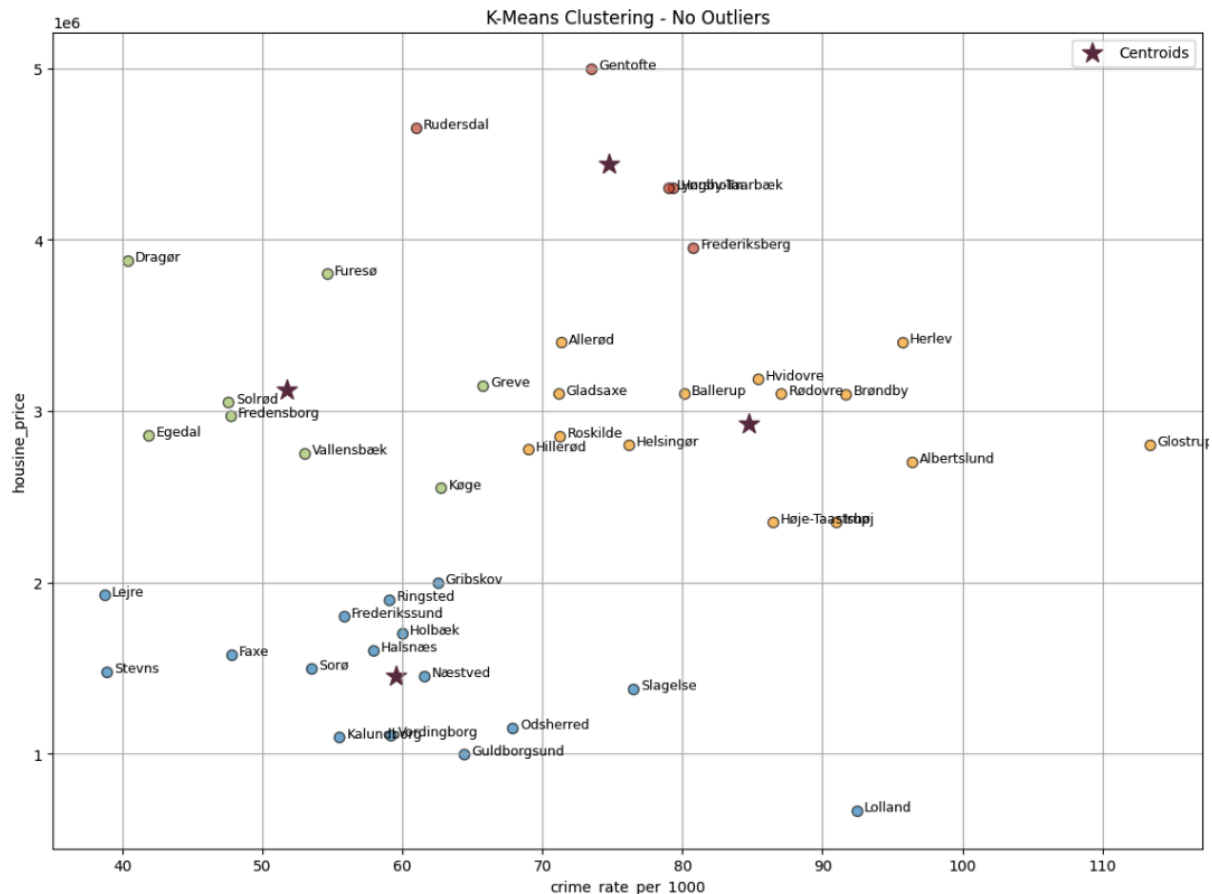
Cluster 3 (The outliers)

- Characteristics: High Crime and High prices.
- Examples: Copenhagen, Tårnby.
- Insight: Copenhagen has the most population so it makes sense to have the most crimes. Also it is the capital and it makes sense to have high prices, although clearly not the highest. Tårnby became an outlier based on the crime rate which most probably is affected by the fact the airport operates in Tårnby, so all the crimes that are reported in the airport belong to the Tårnby municipality.

Limitations:

- K-Means clustering struggles with non-spherical shapes, varying densities/sizes and outliers.
- Our outliers may have created distortion to the clusters.

We decided to use Z-Score to detect outliers and re-create the clusters to observe if we have any significant improvements. Firstly, we found that the outliers are Copenhagen and Tårnby. Then we used K-Means clustering again.



This time we can argue that the clusters are more distinct clusters.

- We have a cluster with low prices and a low crime rate (with the exception of Lolland).
- We have a cluster with median prices and a low crime rate.
- We have the expensive cluster, where prices are high and crime rates are close to the median.
- And lastly we have a cluster with median prices and crime rates starting from the median all the way to the highest values.

Findings

We can confidently argue that high crime rates do not necessarily mean low housing prices.

Can we predict a house's price based on the local crime rate?

The primary goal of this research was to determine if local safety (specifically the crime rate) could be used to predict house prices. To test this, we designed a series of Multiple Linear Regression experiments using various feature combinations, ranging from simple univariate models (Crime Rate only) to complex multivariate models including Municipality, Year Built, and Demographics (Non-Danish population ratio). We made a comparative analysis by running every experiment twice: once on the Full Dataset and once on a "Clean" Dataset where statistical outliers (Copenhagen and Tårnby) were removed.

The Impact of Outliers

The comparative results confirm that removing outliers significantly improved the model's reliability. In almost every scenario, the "No Outliers" models achieved lower Mean Absolute Error (MAE) and higher accuracy than their "With Outliers" counterparts. For example, in the complex demographic model, removing the outliers improved the R² score from 0.293 to 0.325. This validates our hypothesis that Copenhagen and Tårnby act as statistical anomalies; their unique "high crime / high price" dynamic contradicts the standard market rule (where crime usually lowers value), and removing them allowed the model to more accurately capture the trends affecting the majority of Danish municipalities.

Feature Performance & Best Results

The results conclusively show that Location (Municipality) is the dominant driver of house prices, while Crime Rate alone is a poor predictor.

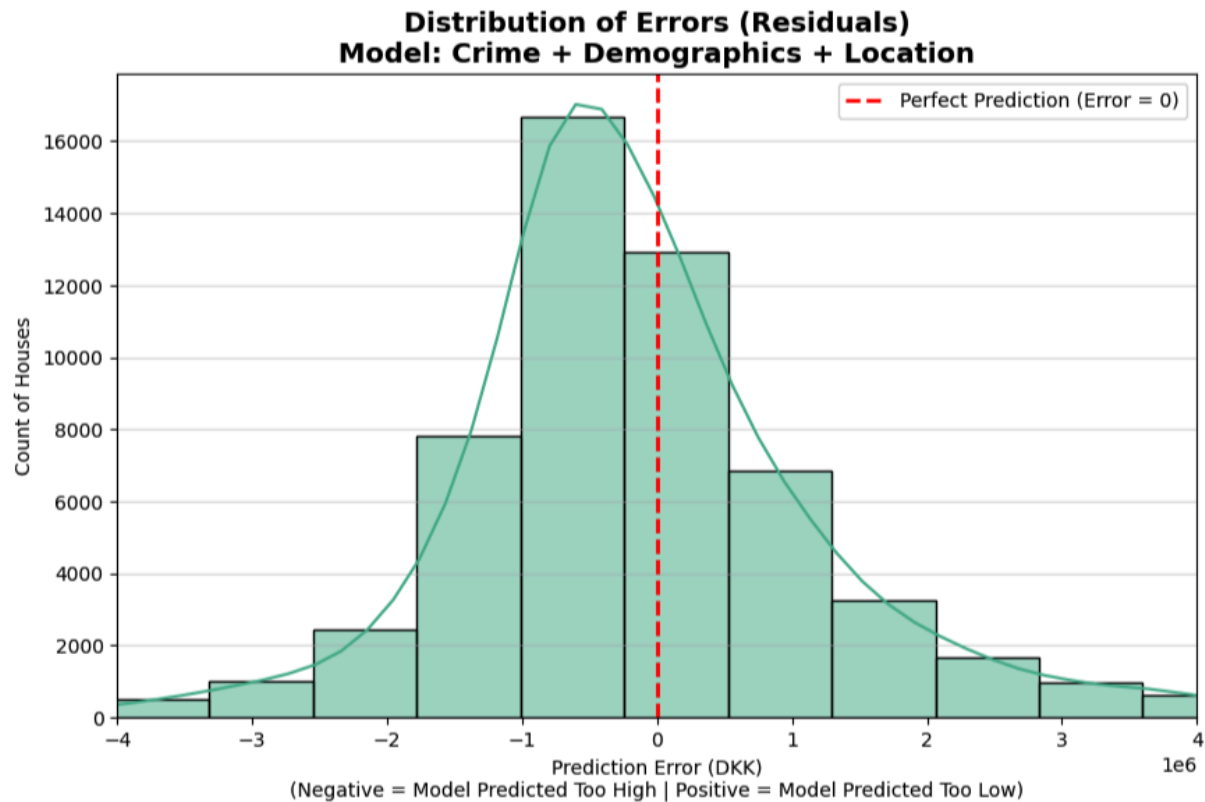
Crime Rate Only: This model failed completely, having an R² of 0.000 on the clean data, proving that without context, safety data cannot predict value.

The Best Model: The most successful approach was Experiment #9 (Crime + Demographics + Location) run on the Clean Dataset. By combining location data with crime and demographic context, this model achieved the highest R² Score of 0.325.

Conclusion

The final R² score of 0.325 means that our best model is able to explain 32.5% of the variation in house prices based solely on neighborhood characteristics. While this establishes a strong baseline for evaluating "area value," the fact that ~67% of the price variation remains unexplained suggests that property-specific features are missing from the equation. We can conclude that while low crime adds value, it is the specific municipality that sets the price.

	Dataset	Experiment	Features Used	MAE (Error)	R ² Score
1	No Outliers	Crime Rate Only	crime_rate_per_1000	1,511,112 DKK	0.000
3	No Outliers	Year Build Only	year_build	1,514,913 DKK	0.001
2	With Outliers	Year Build Only	year_build	1,582,837 DKK	0.002
0	With Outliers	Crime Rate Only	crime_rate_per_1000	1,548,351 DKK	0.024
4	With Outliers	Municipality Only	municipality	1,245,927 DKK	0.276
6	With Outliers	Crime + Year Build + Location	crime_rate_per_1000, year_build, municipality	1,234,543 DKK	0.289
8	With Outliers	Crime + Demographics + Location	crime_rate_per_1000, non_danish_ratio, municipality	1,223,312 DKK	0.293
5	No Outliers	Municipality Only	municipality	1,162,290 DKK	0.308
7	No Outliers	Crime + Year Build + Location	crime_rate_per_1000, year_build, municipality	1,153,220 DKK	0.319
9	No Outliers	Crime + Demographics + Location	crime_rate_per_1000, non_danish_ratio, municipality	1,139,619 DKK	0.325



The histogram of residuals provides a visual analysis of the model's accuracy by plotting the frequency of prediction errors for the "Crime + Demographics + Location" experiment. The X-axis represents the Prediction Error in DKK (calculated as actual price - predicted price), where the red dashed line at zero marks a perfect prediction.

Values to the left (negative) indicate the model predicted a price higher than the actual sale.

Values to the right (positive) indicate the model predicted too low.

The teal bars and the overlying smooth curve reveal a classic normal distribution (Bell Curve), which is a critical validation that the model's errors are random rather than systematic, confirming that Linear Regression was the appropriate statistical tool. Despite the shape, the curve peaks slightly to the left of zero, indicating a minor tendency for the model to overestimate prices.