Data report

Introduction

The purpose of this report is to assess the data used in the project and provide an overview of its quality, reliability, and accuracy. It focuses on the various datasets related to housing, demographics, neighbourhood features and moving trends in the city of Breda. The data quality has a significant impact on the effectiveness and reliability of any insights gained from analysis and modelling for the decision-making process.

Data Collection

The following datasets were collected for the analysis:

- CBS grid data (2018): This dataset contains the geometry of 100 by 100-meter grids their corresponding codes and the grid features (housing, income, demographics, point of interest (POI), leefbaarometer)
- Neighbourhood data (Most recent): This dataset contains the data on the neighbourhood level and the neighbourhood features (Type of employment contracts, income, education, leefbaarometer, government benefits)
- Move house data (2013 to 2020): This dataset contains the number of people moving and the type of movements linked to the CBS grid codes. It was provided to us by the municipality of Breda.
- Housing data (2020): This dataset contains information about the housing stock in a neighbourhood. It was pulled from a PDF using an API.
- Boom per buurt (Unknown): This dataset contains information about the type of development of the city. The data was provided by the Program at Buas.

The CBS grid data and neighbourhood data are more sensitive to expiration because they contain information that is collected every few years and is subject to minor changes. The CBS grid data has the shortest expiration period, typically around 1 year due to more frequent updates for small-scale analysis. The neighbourhood data has a relatively longer duration, estimated at approximately 2.5 years unless significant changes take place at a higher level in Breda. Housing stock data has the longest decay period, typically around 5 years, as it represents a more stable aspect of the housing market.

On the other hand, the move data set is historical data and has no specific expiration date. It captures past movements and serves as a reference for analysing historical patterns and trends.

As for the POI data, it is less likely to change significantly over time as it shows the average distance to certain POIs from the front doors of homes and the number of POIs within a certain number of kilometres. Therefore, the expiration date can be considered long, and the data can be assumed to remain relatively stable unless major changes occur in the surrounding areas.

Given these expiration periods and the stability of the datasets, it is essential to periodically review and update the data used in the project to ensure the most up-to-date and accurate analysis.

Data Description

See Data overview.xlsx

Data Cleaning & Preprocessing

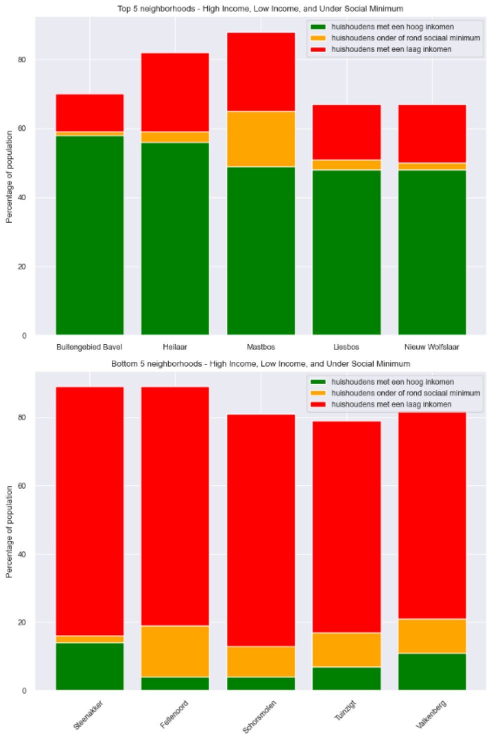
The team checked for missing values inside the datasets. Besides some NaN values, we found large negative numbers in the CBS grid data which represents values ranging from 0 to 4 or values that are kept confidential because of the CBS privacy policy.
These negative values required us to drop certain columns, which consisted of more than half of these negative values. The negative values in the columns that had enough data points were replaced by the mean of that column, based on neighbourhood to try to reach maximum accuracy.

The different datasets had multiple spelling and naming inconsistencies which required us to compare them. As well as the columns and string values.

Whenever data was merged the number of values, the unique values and the NaN values were checked and compared from before the merge.
We translated every column name and value from Dutch to English to make our data more user-friendly.
values and the NaN values were checked and compared from before the merge.

EDA

We did our EDA on the neighbourhood data and the move house data. For the neighbourhood data we looked at the income and the correlation between income and bought or rented properties as well as the correlation analysis between the type of moving houses and the leefbaarometer on the CBS grid level. All types of moving show similar correlations:

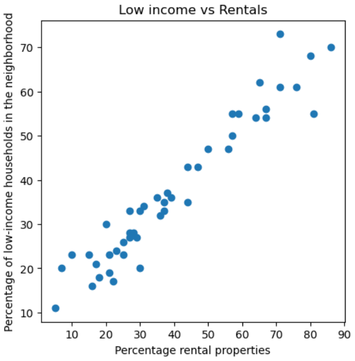


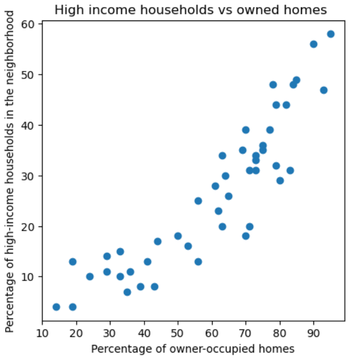
Top 5 neighborhoods - High Income, Low Income, and Under Social Minimum

Bottom 5 neighborhoods - High Income, Low Income, and Under Social Minimum

We found that there is a serious discrepancy within the income between neighbourhoods in Breda.

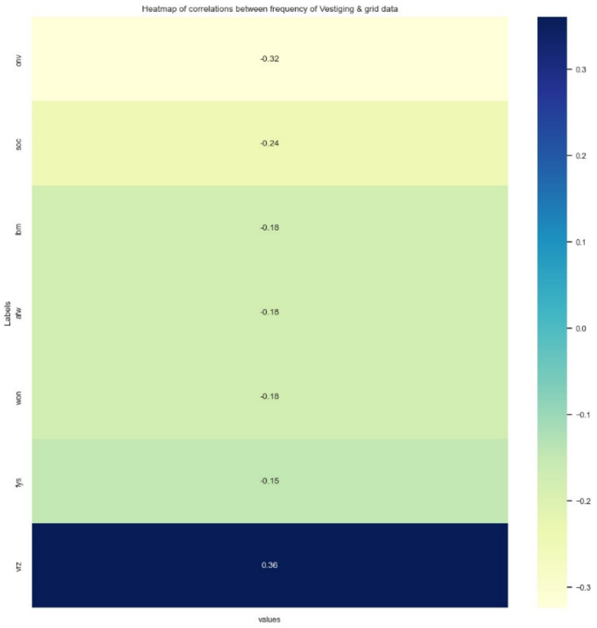
- Stacked bar plot of levels of income per household sorted on the top 5 and bottom 5.



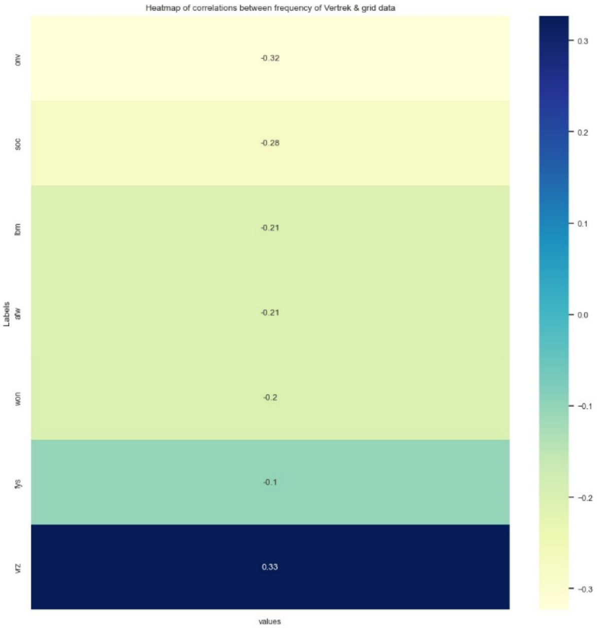


Then we looked at the correlation between level of income and compared them to rental & bought properties, where we found a strong correlation.

After that we looked at a few of our target variables and how they correlate with the leefbaarometer.



- Correlation heatmap between *Frequency of moving inside a grid cell* and the *Leefbaarometer on grid level.*



- Correlation heatmap between *Frequency of moving outside a grid cell* and the *Leefbaarometer on grid level.*

The number of people moving in and moving out is relatively the same, which means that the neighbourhood's population stagnates and doesn't grow. Our research objective was to find out what possibly attracts residents to move to certain neighbourhoods. But we found out that there are a few neighbourhoods with a lot more movement than others, which led us to investigate possible factors that play a role in this.

- Neighbourhoods' movement activity (Figure 3)

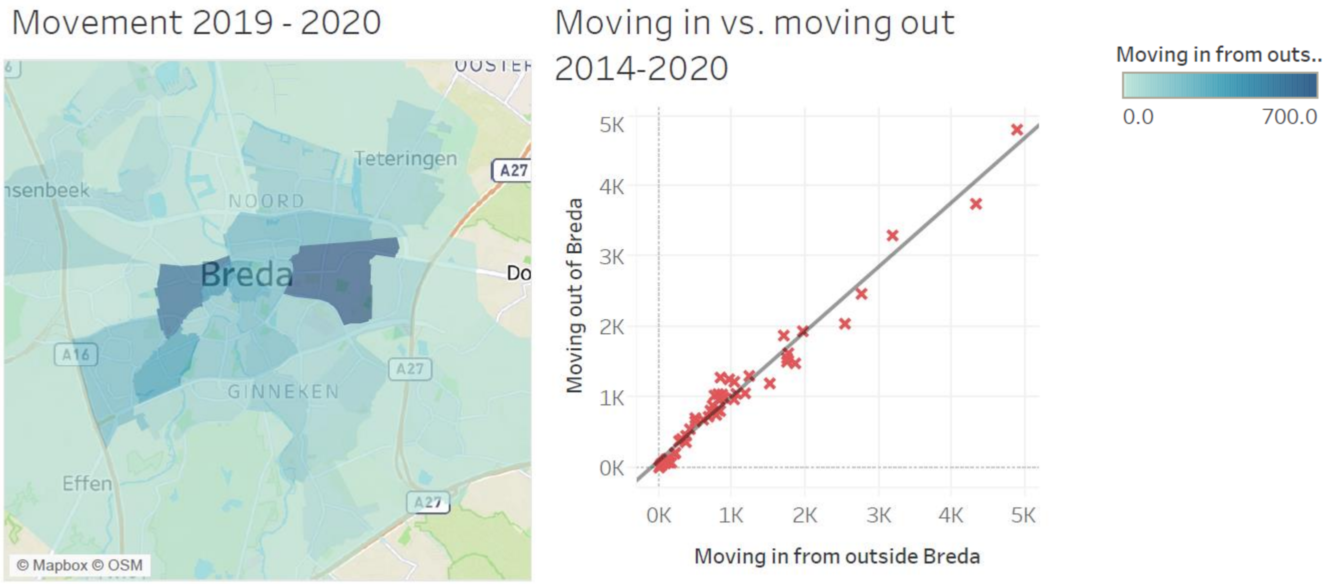


Figure 3

We investigated this further and found a moderate correlation between the safety & nuisance, the amenities, and the population. This leads us to believe that the city centre has a negative influence on the leefbaarometer.

- Heatmap of correlations between *nuisance & safety, amenities, and population* (Figure 4)

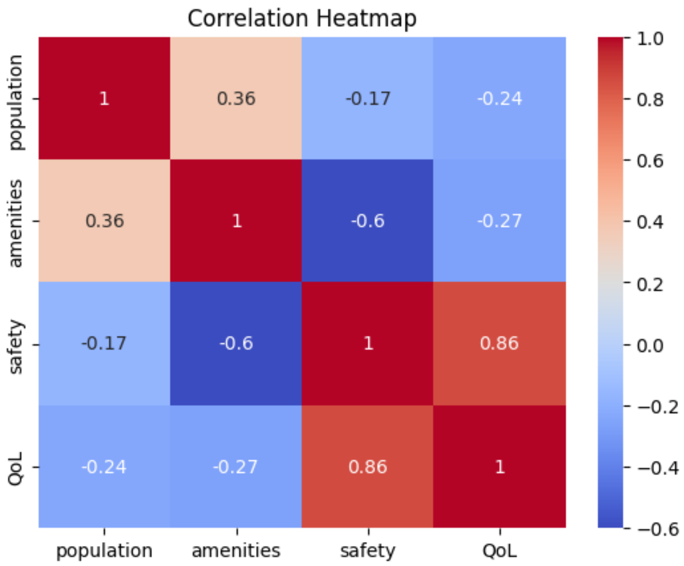
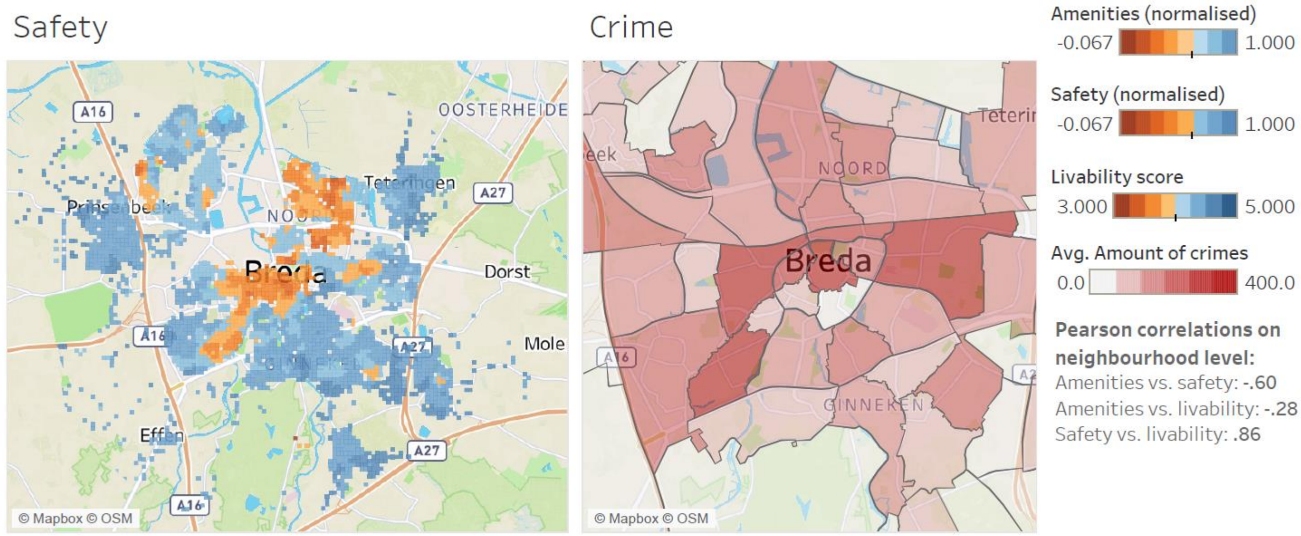


Figure 4

Because of the high correlations of liveability with amenities and safety we investigated how this can be connected to people moving. In the neighbourhoods with a lot of movement we discovered high levels of crime (after removing the outlier of the city centre) and unsafety. (Figure 5)



Pearson correlations on neighbourhood level:
Amenities vs. safety: -.60
Amenities vs. livability: -.28
Safety vs. livability: .86

Figure 5

To look further into possible factors resulting in these neighbourhoods scoring high in crime and unsafety we looked at socioeconomic indicators. As seen in the following figure, these neighbourhoods also have a high level of unemployment and low social cohesion.

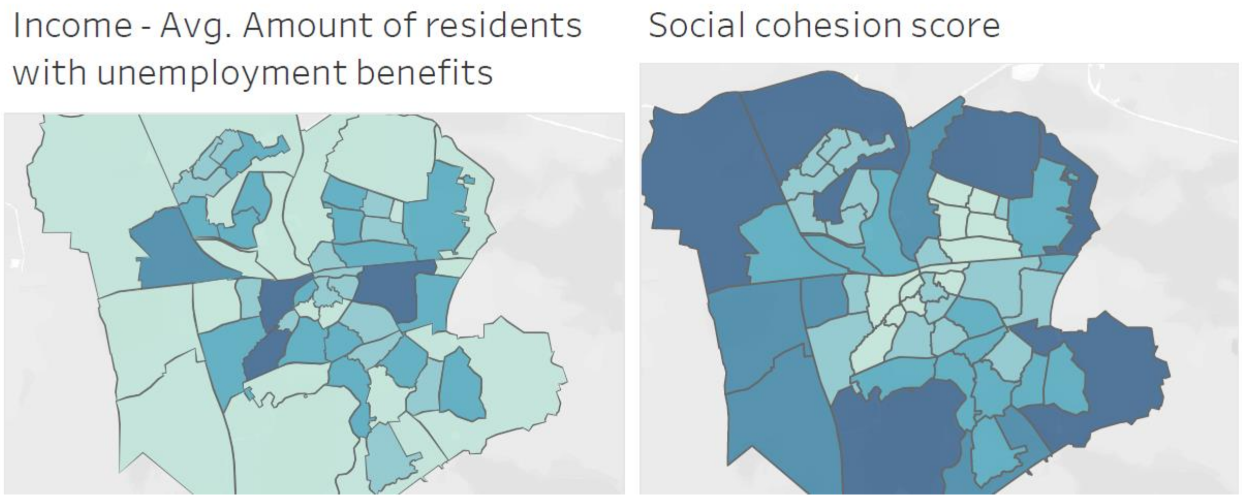

Income - Avg. Amount of residents with unemployment benefits
Social cohesion score

Figure 6

Pearson correlations on neighbourhood level:
Crime count vs. social cohesion: -.56
Crime vs moving out: .60
Amenities vs crime count: .62
Crime vs unemployment benefits: .74

Safety (normalised)
-0.067 1.000

Avg. Amount of crimes
0.0 400.0

Amount of residents with welfare benefit
0.0 400.0

Amount of residents with unemployment..
0.0 120.0

Social cohesion score
-0.0782 0.1476

Feature Selection

Before training our model on the dataset, we performed feature selecting using MultiTaskLassoCV with parameter tuning to increase the interpretability of our model and in most cases only slightly increase the performance metrics of the model.

Model Development

We have tried several different models in order to predict the number & type of people moving housing in relation to the CBS grid squares. Namely regression models and neural networks. Once we started working on Random Forest Regressor, we got good results which we could iterate over to produce a good prediction result. We used a dummy baseline in order to compare the various evaluation metrics. We compared these results after every change we made.
Below are all the iterations and the best performance metrics for each model:

	Linear Regression	Decision Tree Regressor	Random Forest Regressor	Neural Network
RMSE	2.497	2.875	2.187	5.316
MAE	1.405	1.331	1.061	2.571
R-squared	0.720	0.599	0.765	-0.321