Project: Life in Breda

Data quality report and ML lifecycle

Team Tech Fusion 5

Andrea(220666), Borislav(220472), Johuar(223787), Neil(221270), Vincent(223377)

DISCOVER YOUR WORLD

Applied DS & AI/Game Department

Friday 23 June 2023

Breda University OF APPLIED SCIENCES

Life in Breda

Andrea Tosheva, Borislav Nachev, Neil Ross Daniel Manohar, Jouhar Birakdar, Vincent van der Beek

Breda University of Applied Sciences

Block D - Team-based Working

Irene van Blerck

June 23, 2023

Index

Contents

1 Introduction

1.1 Who are we working for?

The municipality of Breda has decided to work towards equal opportunities, no matter where you grow up in the city. Within one generation or by 2040 they want to achieve this and to create a valuable life for the citizens of Breda, with a focus on housing, education, and income. Numerous public services such as: schools, businesses, police, housing corporations, healthcare and welfare organizations, the municipality, and the national government work together towards collecting data to make their goal achievable.

1.2 Why do they need us?

With the exponential increase in terms of size and sources more and more data is collected and stored on numerous storage warehouses or even local hard drives. However, just the collection of data can be meaningless when there is no understanding of it. Nowadays, the task of the stakeholders, to take everything into account when taking a decision, gets harder and harder. After seeing the results of applying machine learning to different industries, the municipality of Breda approached us with the request to analyse and understand their data, so that they can have a better idea of how to help their citizens to improve their life in every aspect possible.

1.3 What is our goal?

Our goal is to see which are the factors that are contributing to the quality of life in the different neighbourhoods of Breda.

2 Datasets used

Full freedom was given in terms of data collection. From the data provided by the municipality of Breda we used the green index data as it can be influential to other key metrics such as air quality, parks which on the other hand have influence on the quality of life. To support our theory even more, we decided to include data from official government sources such as: their police website, the Central Bureau of Statistics, additional data provided by our stakeholders, open-source data for the POI and other statistical information per neighborhood. By using this kind of datasets we'll have a clear view of the most influential factors to the quality of life in each neighborhood.

Fig. 1 highlights some additional information about each dataset used in this project:

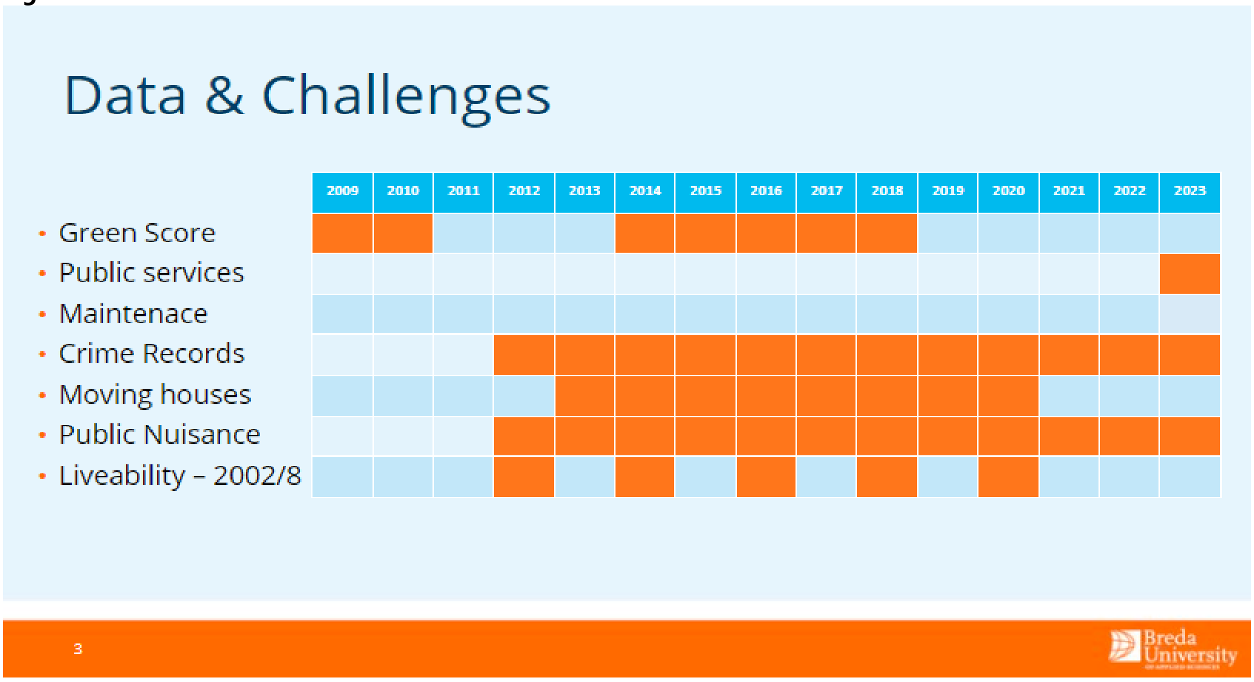
Figure 1

Datasets used summary

Dataset:	Publisher:	Format	Rows	Columns	Description:
Green Index	Municipality of Breda	.csv	298038	7	Data showing the green index score per neighborhood, per year.
Public Nuisance	Police	.csv	89964	7	Data showing the amount of recorded public nuisance incidents per neighborhood, per year
Crime data	Police	.csv	545632	5	Data showing all recorded crimes per neighborhood
Livability data	Ministry of the Interior and Kingdom relations	.csv	56	8	Data showing the livability index per neighborhood, per year
Move Houses	Municipality of Breda	.csv	39374	5	Data showing the frequency of the different types of house movement in and from the city.

3 Data quality - cleaning and preparation

Figure 2



Data & Challenges

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
• Green Score	🟧	🟧				🟧	🟧	🟧	🟧	🟧					
• Public services															🟧
• Maintenace															
• Crime Records				🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
• Moving houses					🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧			
• Public Nuisance				🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
• Liveability – 2002/8				🟧		🟧		🟧		🟧		🟧			

3

Since we took most of our data from different sources, every member of our team conducted their own individual research which resulted in a lot of missing values and inconsistencies in the data they've collected.
Based on figure 2 it was decided to use data only between the years 2014 and 2020. To prepare our data for our Machine Learning models it was decided to divide the preprocessing between Andrea and Borislav.

> Andrea preprocessed the datasets with the livability index and the recorded crimes. The 'recorded crimes' dataset had to be filtered out so that we are using information on neighborhood level and also, we had to ensure that we are working with the total number of records. Also, we needed remove all unnecessary years which don't fit in our time period. In the dataset connected with the livability index we encountered more issues. There were multiple missing years which had to be filled up by using data imputation. Also, the livability index had to be converted in number scale so that it can be used in a predictive model.

> Borislav preprocessed the green index, moving houses and public nuisance datasets. Except the mandatory preprocessing and filtering of the data to match the desired time period, the green index dataset contained a lot of missing data even in the years in which data was officially collected, therefore additional preprocessing was required. The problem was caused by the fact that the municipality of Breda didn't do complete record of the green index every year. Therefore, an algorithm was developed to calculate the most accurate result by taking the mean between the closest value on the left and right around the missing value. After fixing the missing data a predictive model was developed, which filled the missing values of the green index for the years of 2019 and 2020.

3.1 Data merging

> All datasets were merged on the present neighborhood columns, since we based the work on our model on neighborhood level.

4 Data exploration

The exploratory data analysis of the livability index dataset showed that there aren't big and significant movements of its values throughout the years. The least common index is "less than insufficient" ("ruim onvoldoende"), and the most common index is "more than sufficient" ("ruim voldoende") (**Fig. 3**). The overall index between the districts has gradually improved for a period of almost 20 years – the biggest percentage of neighborhoods had index "bad" ("zwak") and "more than sufficient" ("ruim voldoende") for the year of 2002(**Fig. 4**) and for 2020 is "very good" ("zeer goed") and "excellent" ("uitstekend") (**Fig. 5**).

Figure 3

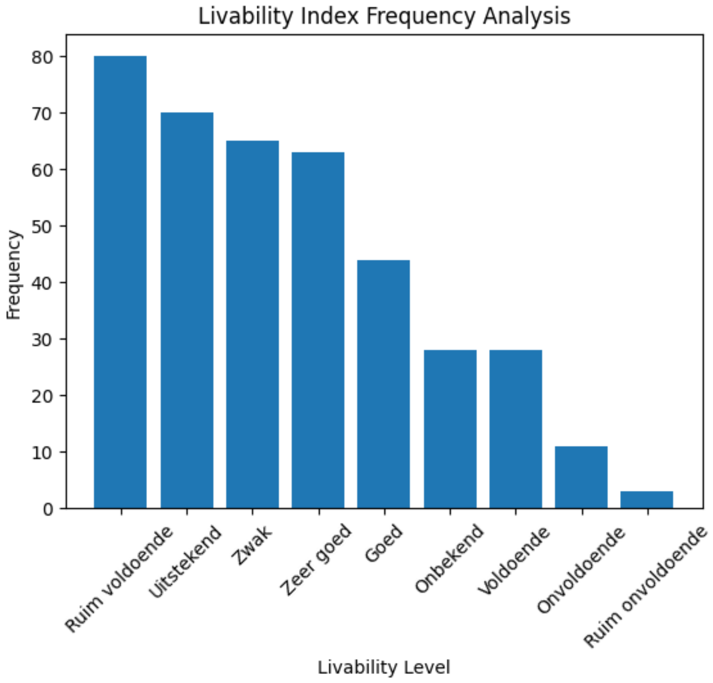


Figure 4

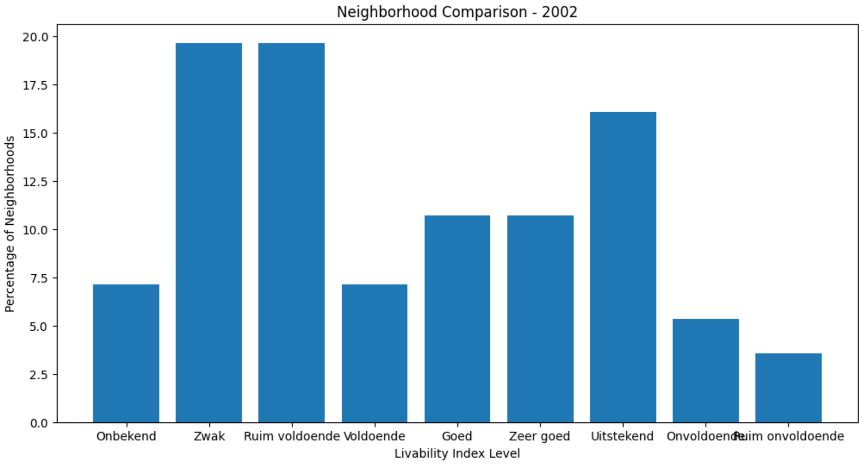
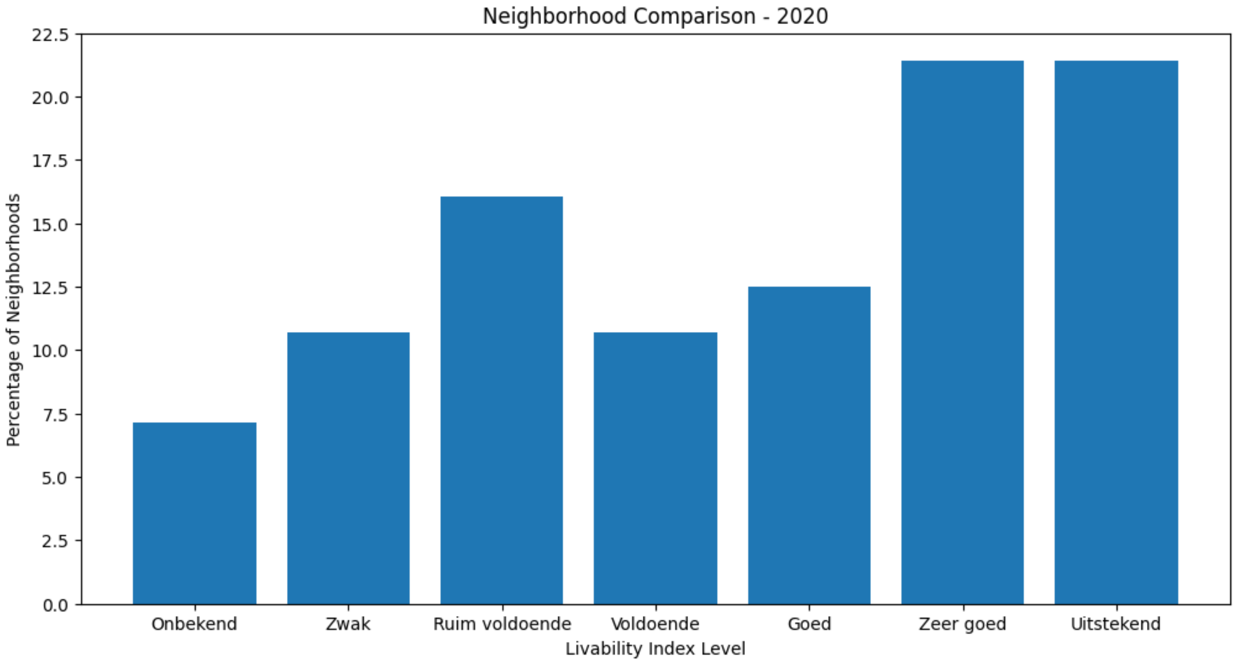
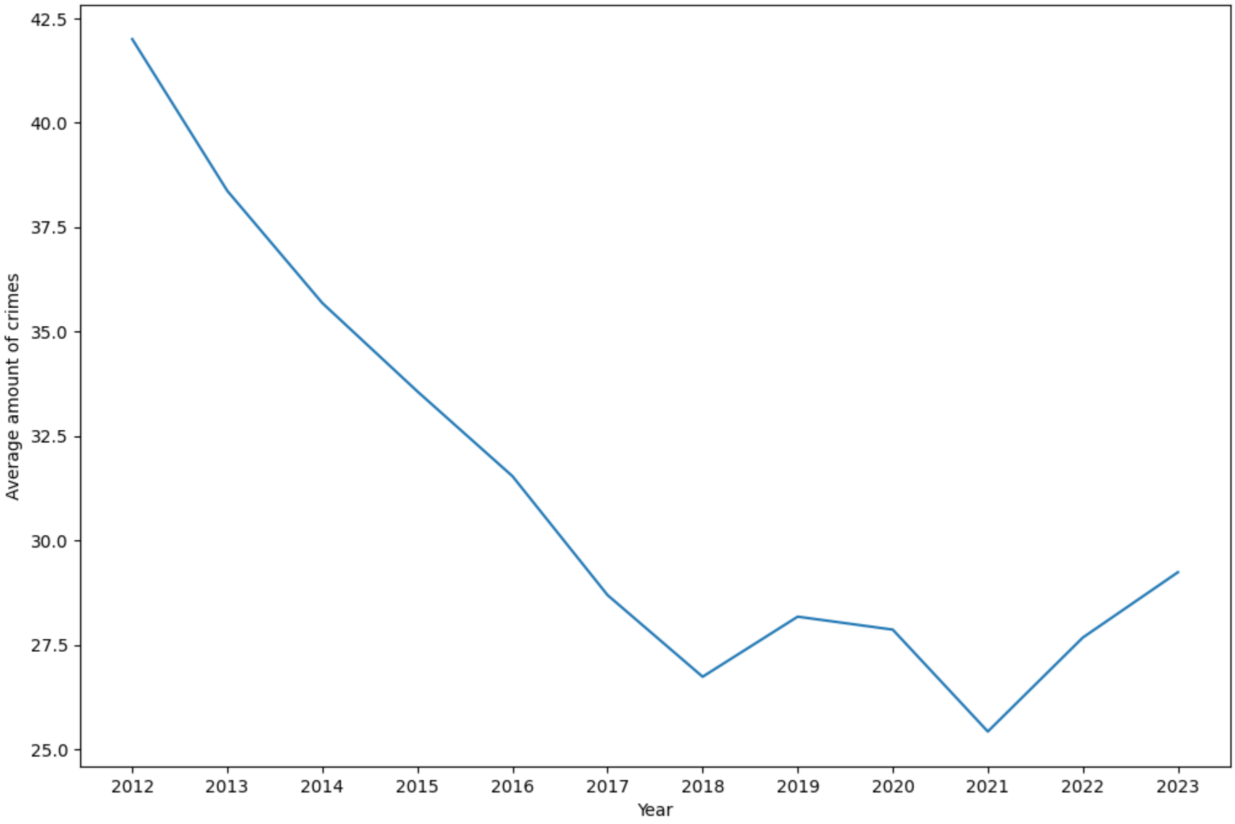


Figure 5



The EDA of the recorded crimes dataset showed that the mean number of crimes was in its peak in 2012 and that most recorded crimes happen in the city center of Breda. Overall the number of the recorded crimes gradually decreases for a period of 6 years, with lowest point in 2021, followed by small increase in the past 2 years. (**Fig. 6**)

Figure 6



In the public nuisance dataset, it was shown that most of the nuisance accidents are recorded in the city center, the same way like with the recorded crimes. Interestingly enough, the biggest number of recorded incidents fall into the category of the incidents that don't fit in any other category. Second after them come the incidents caused by mentally challenged people. (**Fig. 7**)

Figure 7

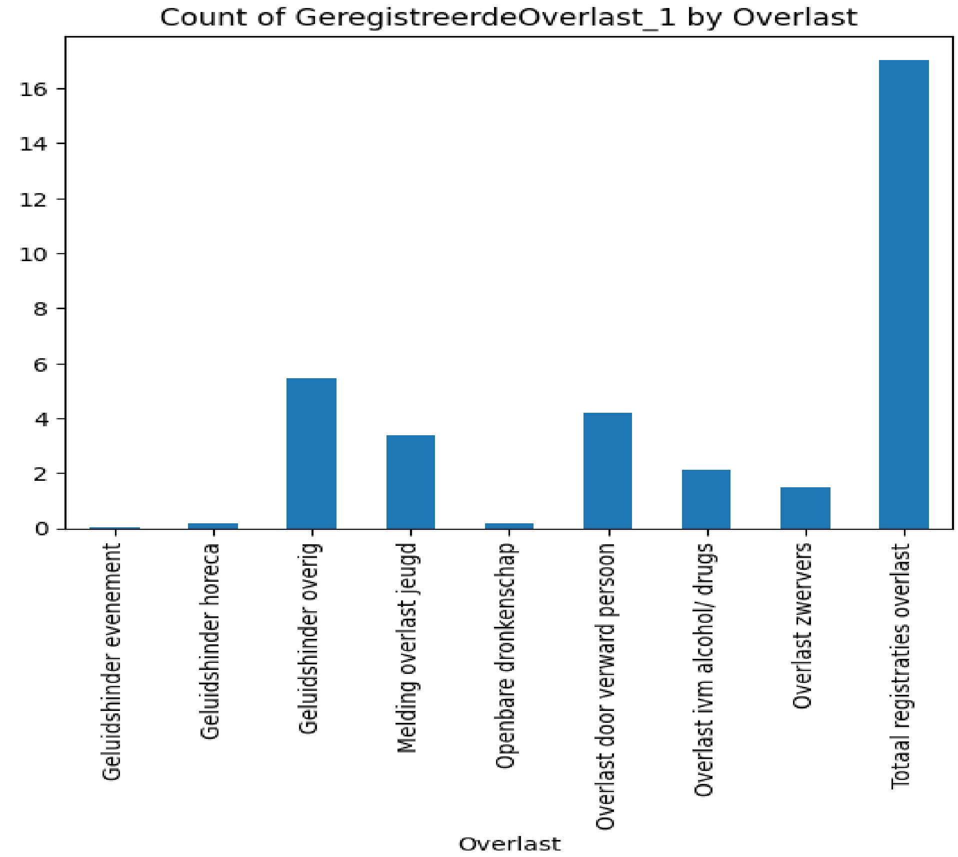
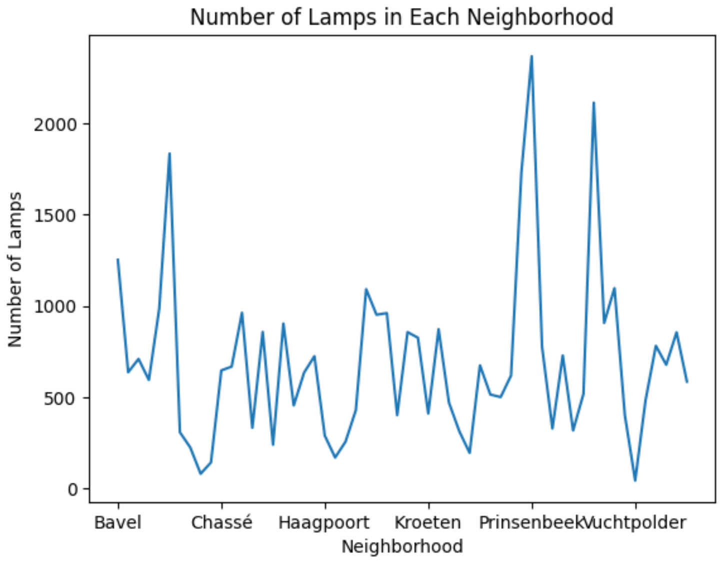


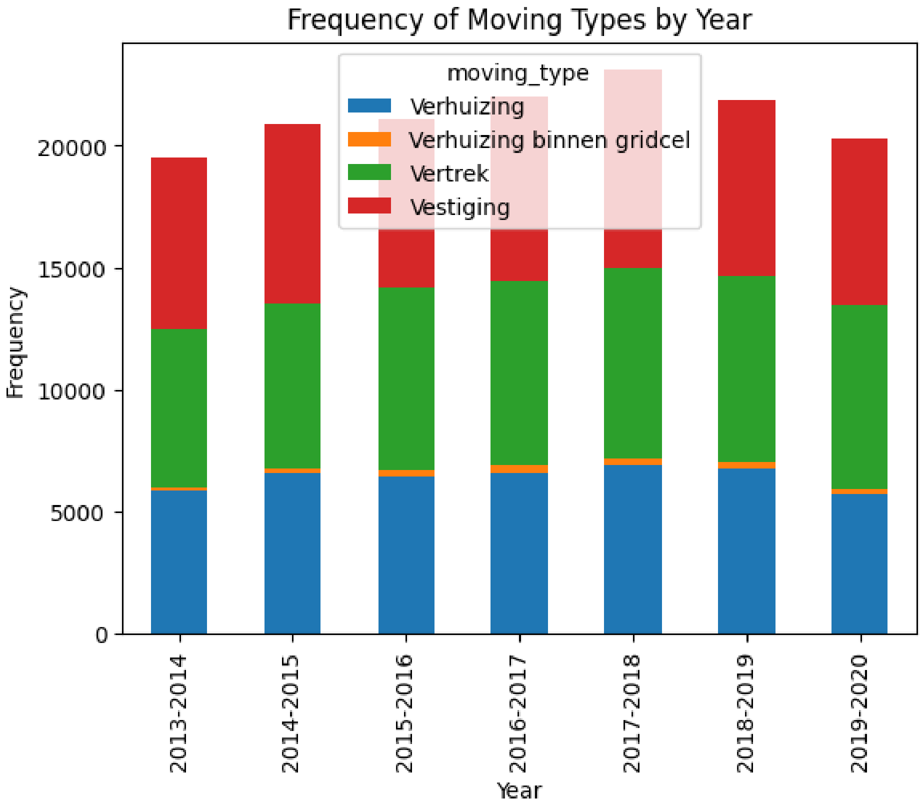
Figure 8



The EDA of the light dataset didn't show many interesting insights, besides the distribution of the amount of public lights across the main neighborhoods. (**Fig. 8**)

From the dataset with the information of the house movement within the area of Breda it became clear that the movement with the lowest frequency is the moving within the area of their current gridcel. Overall, the frequency distribution between the other three types of movement is pretty equal, with a slight tendency of increase of the movement inside of Breda and the movement outside of it. (**Fig. 9**)

Figure 9



In the dataset for the Green index of Breda its clearly shown that the highest average score is recorded in the year of 2010 and after that the score slowly decreases in the next couple of years. (**Fig. 10**) Also, the map in **Fig. 11** visualizes the distribution of greenness around the city area and is clearly visible which are the areas which are lacking "greenness".

Figure 11

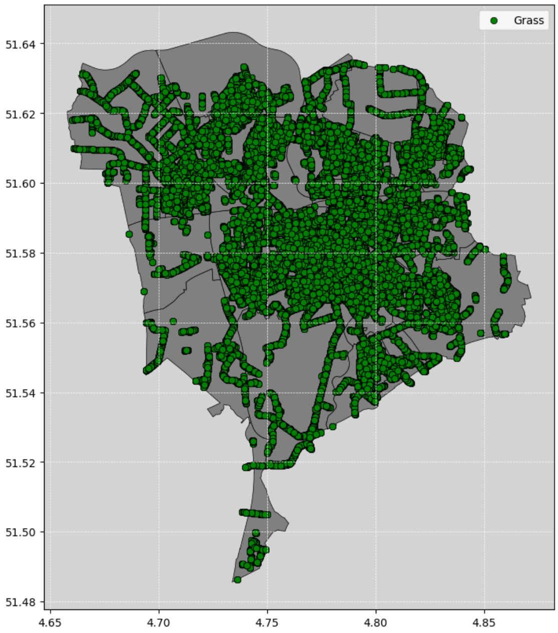
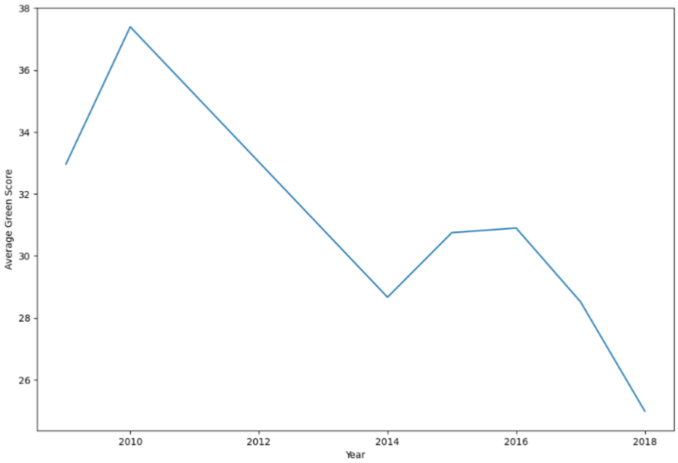


Figure 10



5 Machine Learning models used

Multiple ML model iterations, approaches and techniques were used to come up with the most efficient models:

- Random Forest Regression developed by Andrea.

The features used for the model are selected and engineered to improve their predictive power. The dataset for the model was grouped on neighborhood level, but also by years. Two new features, 'moving_out' and 'moving_in', were created. 'Moving_out' feature was assigned to the values from the 'Vertrek' column, and 'moving_in' was calculated as the average of 'Vestiging', 'Verhuizing binnen gridcel', and 'Verhuizing' columns. The unique neighborhoods in the dataset were iterated over and for each neighborhood, the following steps were executed:

- The neighborhood-specific dataframe was extracted using boolean indexing.
- The feature matrix 'X' was created using the columns 'green_score', 'GeregistreerdeOverlast_1', 'GeregistreerdeMisdrijven_1', 'moving_in', and 'moving_out'.
- The target variable 'y' was extracted from the 'Livability index' column.

The Random Forest model was instantiated by using the scikit-learn library. The feature importances of the trained model were calculated using the 'feature_importances_' attribute and the model was trained on the X and y variables. The Random Forest Regressor model was chosen as model prediction and the random state was set to 42. A loop was executed for each neighborhood in the 'neighborhoods' list. For each neighborhood, the previously stated steps were repeated. After that the dataset was split into training and testing sets using a split ratio of 5. The model was trained on the training set (X_train, y_train) and predictions were made on the testing set (X_test) using the trained model. The predicted values, actual values (y_test), and mean squared error (MSE) between the predicted and actual values were appended to the 'data' dictionary for the corresponding neighborhood.

- Linear Regression with Interaction terms developed by Borislav.
  - Based on the Simple Linear Regression model which tries the relationship between a scalar response and one or more explanatory variables, an interaction terms allows the model to make the relationship between one variable X1 and Y dependent on another variable X2.
  - The Linear Regression with interaction terms was the preferred model for our use case because we wanted to predict the livability score and at the same time keep the relationship that each neighborhood has on the livability score.

6 Iterations

1. Deep Learning Time Series Forecasting
   - We performed the Deep Learning Time Series Forecasting in the right way. The data was normalized and split into windows. However, after multiple discussions and meetings, it was decided not to use this model as a final one, because it's made for more complex time series problems than we have. The team also encountered a problem with a lack of data from which the Deep Neural Network can learn.
2. ARIMA Time Series Forecasting
   - ARIMA was used in the second iteration. We used multiple predictor parameters, and we could successfully predict the livability index. However, since the data that we had was too repetitive, the predictions weren't accurate, and we decided not to use this model. This model was also used during the preprocessing phase of our project where it predicted missing values from the downloaded datasets.
3. Vector Auto Regression Time Series Forecasting
   - We tried VAR Time Series Forecasting as a third iteration because our idea was to try to do multivariate time series. The problem here was that this didn't answer our initial research question since it doesn't say what influence each predictor has on the target variable and we decided not to use it as a final model.

4. Machine Learning Time Series Forecasting
   - Random Forest Regression
     - Since our project research question was “on a neighborhood” level. We trained 56 Random Forest Regressors (one per neighborhood) to maintain our “on neighborhood “level analysis. To be able to satisfy our research questions and look at which factors have influence on the livability score we took the feature importance of predictor for each model.
   - Linear Regression with Interaction terms
     - After our ‘BaseLine’ Lasso Linear Regression model, we decided to use Linear Regression with interaction terms because they allowed us to set interactions for each neighborhood and make our analysis more accurate with a single model.

7 Future Improvement

Data Management Recommendations:

- To store the datasets into a Relational Database which will make the datasets more secure. By setting constraints to each column type the values will be filtered automatically when loaded.
- When storing the data in a Relational Database data, the access time will be faster. Additionally, the preprocessing and merging processes will be faster as well.
- By using a Relational Database, we can add one more layer of security to the access of our data by setting different user roles to it.

Machine Learning Recommendations:

- The relationship between the chosen predictors can be further investigated by acquiring more data for each neighborhood in general on a neighborhood. Some known outliers can be described to the model as interactions to improve its accuracy even more.

8 References

Nvidia Linear Regression with Interactions **https://developer.nvidia.com/blog/a-comprehensive-guide-to-interaction-terms-in-linear-regression/**

Time Forecasting **https://www.youtube.com/@NachiketaHebbar**

Multivariate Time Forecasting **https://towardsdatascience.com/multivariate-time-series-forecasting-with-deep-learning-3e7b3e2d2b**

Games

Leisure & Events

Tourism

Media

Data Science & AI

Hotel

Logistics

Built Environment

Facility

Mgr. Hopmansstraat 2
4817 JS Breda

P.O. Box 3917
4800 DX Breda
The Netherlands

PHONE
+31 76 533 22 03
E-MAIL
communications@buas.nl
WEBSITE
www.BUas.nl

DISCOVER YOUR WORLD

Breda
University
OF APPLIED SCIENCES