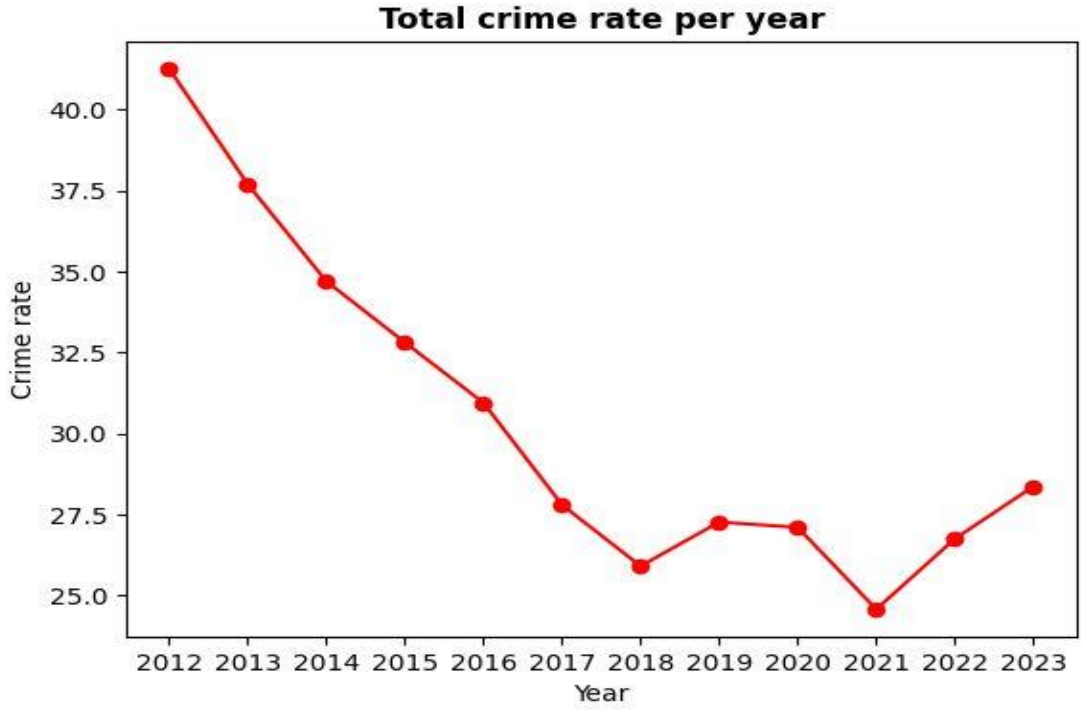
Data quality report

The panic room

The purpose of this report is to assess the quality of the collected data, which was used for achieving the goal of the project. The datasets used in this analysis include crime rate data provided by the Dutch police, educational level data, income data, quality of life data and population data from the municipality of Breda. The focus of this report is to evaluate the completeness, accuracy, consistency, timeliness, and validity of each dataset, providing valuable insights into the reliability and usability of the collected data.

The crime dataset from the Dutch police contains information about the number of different type of crimes in each neighbourhood in Breda throughout the years (2012 to 2023). This means there are 4 columns – Soort misdrijf (Type of crimes), Perioden (Period), Wijken en buurten (Regions and neighbourhoods) and Geregistreerde misdrijven (aantal) (Registered crimes (numbers)). Initially, it was checked the completeness of the data as “NaN”/null values can be exclusively noticed in the lastly mentioned column. Such values may indicate either the absence of certain crimes in Breda or their underreporting. Furthermore, there are not entirely duplicated columns (in a way that there are not rows, which repeat themselves) as each row provides unique information. Another thing, which must be considered is whether there are inconsistencies or anomalies in the data, which was checked by looking in the crimes throughout the years.



As it can be noticed in the visualization, the total crime rate in Breda has decreased throughout the years. This can be interpreted in two different ways – either way the crime is handled well in the past year, or it is not collected enough data. However, based on the other sources – the crime rate in Breda is increasing in the past 3 years, which can be noticed in the Dutch police data. Even though, it is not for the entire period from 2012 to 2023, it can be

assumed that probably there are not anomalies and that are the actual numbers. No identified biases were found in the data, since there is no personal information and it is opened for everyone. Lastly, it must be checked the accuracy of the data. It can be said that the data provided by Dutch police it is trustworthy based on few things – the data is collected with the help of Basisvoorziening Handhaving (BVH), which allows the police officers to report any type of incident, crimes, and activities, which they encounter. The info contains the location, period, and other relevant details. Also, the data is used by the Public Prosecution Service, so it can benefit the legal purposes.

The population, income level and educational level datasets will be described together, because they share a lot of similarities between them. The only difference they have is the categorical columns. For instance, the income level data has categorical column, which contains the unique values "Low income" and "High income", while the educational data contains "Laag" (low level of education), "Midden" (Middle level of education) and "Hoog" (High level of education). The similar columns are the "Buurten" (neighbourhoods), "Year" and "value" (number of people). The completeness of the data was checked and there are some similarities – for some of the data there are missing values for income level, population, liveability index and educational level for the same neighbourhoods in 2020 – Emer and Hazeldonk. However, in the income level, liveability index and educational level data all the rows for both neighbourhoods either contain "NaN" or "?". Of course, there also other neighbourhoods with missing data, but it was worth to be mentioned that. Regarding duplicates, there are none and each contains distinct information. Moreover, the datasets were checked for inconsistencies and anomalies as something unusual can be noticed only in the dataset for educational level. When the data is getting aggregated, it gives incredibly high numbers, which doesn't make sense. To make the data better for an analysis, it was made EDA only on one year – 2019. Furthermore, there is no bias in all three datasets since all the data is public and in the project is not used private data. The data provided from the municipality is collected from variety of sources such as CBS, Police, ABF Research etc. Furthermore, the data is up to date, which makes it a reliable resource.

The last dataset is for the Quality of life in Breda. Like the tables from mentioned in the previous paragraph, this one contains column for neighbourhoods and year with the only difference there is no categorical column and instead of having a column for number of people – there is for liveability index. Typical to the other data, there are missing values in the numerical column (which in this case is the "liveability score"). For the data about the income and education it was mentioned that there is not data for two specific neighbourhoods – Emer and Hazeldonk. This can be noticed in this dataset as well and is also not just for one specific year, but all of them. Also, there are no duplicated data, and each row shows the liveability score of a neighbourhood in a different year. There is no inconsistency in the data as the range of the liveability score is reasonable. It can be said that the part about the bias and the accuracy of data is the same as the previous 3 datasets.

There were more datasets, which EDA was done on, but those are the most important and the only ones, which are going to be used for creating the model for the project. Based on the analysis, all of them are suitable for work as some changes were made to make them more useful. For example, some of the neighbourhoods without any data were removed and the group decided to focus on specific year, so the educational data can be used.