Data quality report

Isa van der Mierde, Max Meiners, Marwa Rouah, Michal Dziechciarz

Table of contents

A. Introduction..3

I. Scope and objectives of the report...3

B. Background information on the data...3

I.Used data sets:..3

C. Data Frameworks...5

D. Evaluating the datasets.. 6

I.Evaluating the completeness of the data:...6

II.Identifying inconsistencies and anomalies.. 8

E. Bias.. 9

F. Evaluating the quality of the sources...10

G. Bibliography..11

A. Introduction

I. Scope and objectives of the report

This data quality report focuses on evaluating diverse datasets utilized to address two research questions: “What are the most important factors that the municipality of Breda needs to take into account to effectively allocate resources in order to prevent public nuisance?” and “Can the development of an algorithm aid in predicting the risk of public nuisance in specific areas within the municipality?”

The main objective of this report is to provide a comprehensive understanding of the datasets used, highlighting our efforts to ensure data integrity by assessing its quality, identifying biases, and evaluating the reliability of data sources. By conducting a thorough examination, stakeholders will gain valuable insights into the accuracy and dependability of the datasets.

In addition to assessing data quality, this report aims to promote data governance, offer recommendations, and support decision-making processes. It serves as a communication tool to foster transparency, responsibility, and collaboration among stakeholders. Ultimately, the report acts as a catalyst for improving data quality, enabling informed decision-making, and advancing the organization's data management practices.

B. Background information on the data

I. Used data sets:

Registered nuisance:
The "registered nuisance" collection includes information on reported instances of annoyance and public disturbance, in particular, Breda locations. In the city of Breda, this directory lists information such as complaints and reports concerning noise, pollution, public health risks, and other types of disruptive activities.

personen met uitkering wijk en buurt:
Information on the number of people in Breda who get social or welfare help is available in the "Personen_met_uitkering_wijk_buurt" data directory. This collection of data offers insights into the socioeconomic standing of various regions, assisting decision-makers and social service organizations in comprehending the distribution of welfare beneficiaries, identifying regions with greater needs, and efficiently allocating resources.

Boas:
The dataset "BOA" consists of categories recorded by the "Special Investigation Officers" in Breda. These categories comprise multiple subcategories. One specific example is the category of cycling nuisance, which encompasses issues such as improperly parked bicycles and instances where cyclists ride in prohibited areas.

population:
The "population" dataset details Breda's population density and dispersion. Planning public services, performing social and economic research, and comprehending population dynamics all depend on accurate demographic data analysis.

Heat stress index:
The heat stress directory includes information about conditions associated with heat stress or heat in a particular area of Breda. When relative humidity and air temperature are combined, the heat index represents how hot it feels to a human body. This data may be used to analyze pinpoint hotspots, gauge the effect on public health, and create plans for reducing heat-related dangers.

Green index:
The data directory for the "green index" offers information for measuring and assessing green areas in a certain area or region, in this case, Breda. By assessing the quantity and quality of parks, gardens, woods, and other natural places, this data set sheds light on the neighborhood's environmental sustainability and livability.

Registered crime:
The registered crime dataset includes information on the total number of crimes that have been recorded in Breda's various neighborhoods during the last 10 years. Law enforcement agencies, decision-makers, and researchers can detect crime patterns, allocate resources for crime prevention, and assess the effectiveness of crime control measures by analyzing recorded crime data.

C. Data Frameworks

DEDA (Data Ethics Decision Aid) is a dynamic and interactive framework designed to support ethical inquiry in AI and data initiatives. It offers a structured framework to reflect upon and justify their ethical judgments related to data. By employing DEDA, it allows us to navigate complex ethical considerations and make informed decisions regarding data usage.

Additionally, the Assessment List for Trustworthy AI (ALTAI) serves as a comprehensive tool for evaluating the compliance of AI systems with seven key requirements for Trustworthy AI. ALTAI ensures that AI systems developed, deployed, procured, or used meet specific criteria, including Human Agency and Oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Diversity, Non-discrimination, and Fairness.

By incorporating these frameworks into our project, we prioritize trust, openness, and responsible data practices. The utilization of DEDA and ALTAI fosters an environment that promotes data quality and ethical standards in AI and data-driven initiatives. It enables us to enhance data accuracy, integrity, privacy protection, inclusiveness, and responsible data handling, thus elevating overall data quality and adhering to ethical principles.

D. Evaluating the datasets

I.Evaluating the completeness of the data:

To assess the integrity and precision of the data, the Python library "Missingno" was employed, which offers a range of visualizations to examine the occurrence and distribution of missing data within a pandas data frame. These visualizations facilitate the identification of missing values' locations, the extent of their absence, and the potential presence of correlations among missing values. While missing values often lack informative value, a closer analysis may reveal underlying patterns or insights.

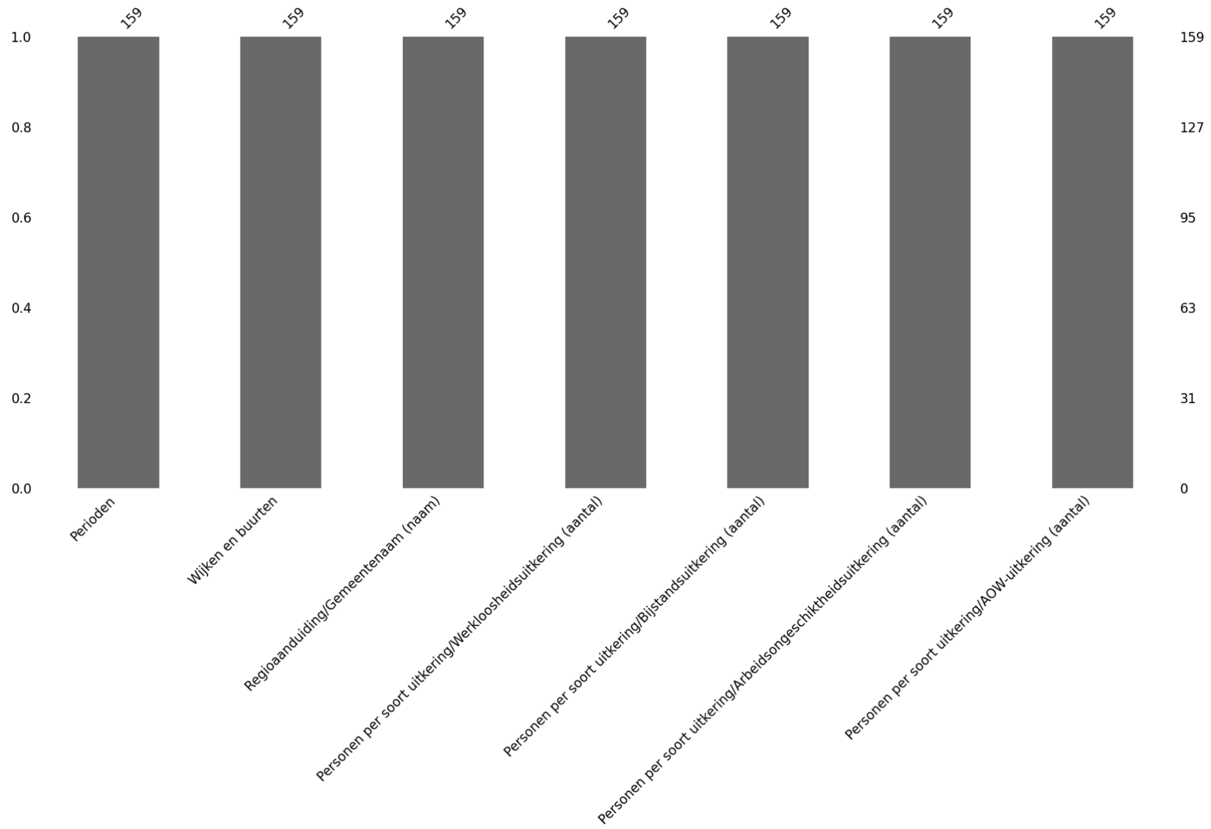
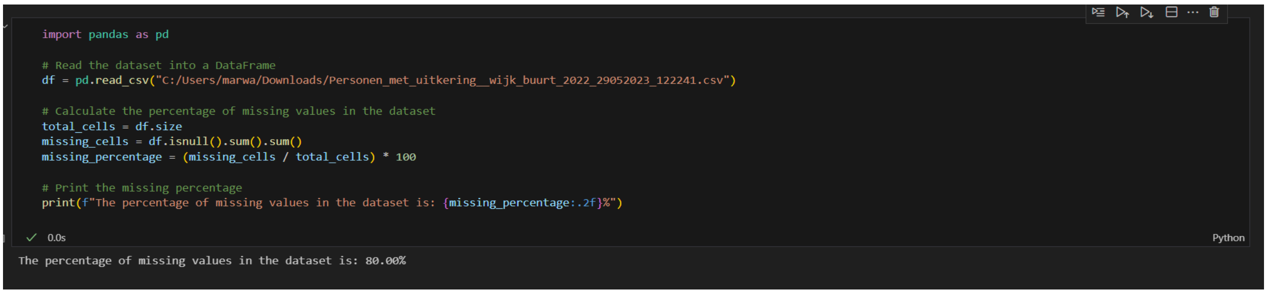


figure 1 : amount of values in each column of the "personen met uitkering wijk en buurt (2014)" dataset

Figure 1 showcases a bar plot generated using the "personen met uitkering wijk en buurt (2014)" dataset. The plot demonstrates the dataset's completeness, as all 159 values in each column are present, as indicated by each bar reaching the value of 1.0 on the y-axis. Employing this technique on each dataset enables us to verify the data's quality. The majority of the data proved to be complete and accurate, with the exception of the "personen met uitkering wijk en buurt (2022)" dataset.

Figure 2:Percentage of missing values within the dataset

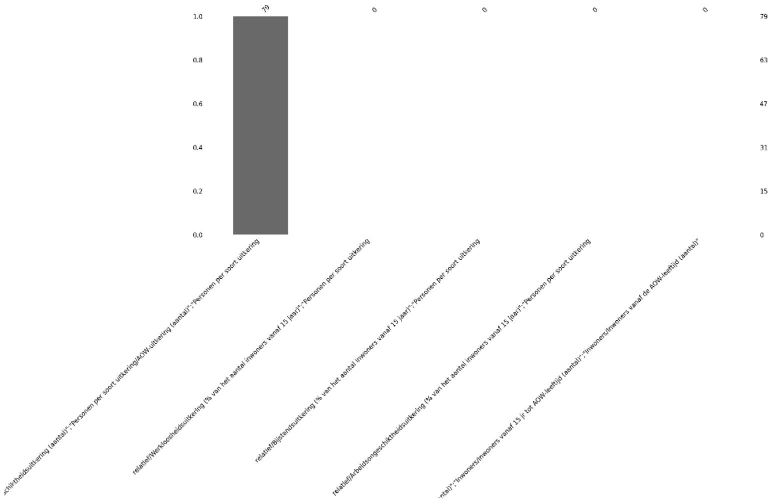


```
import pandas as pd

# Read the dataset into a DataFrame
df = pd.read_csv("C:/Users/marwa/Downloads/Personen_met_uitkering__wijk_buurt_2022_29052023_122241.csv")

# Calculate the percentage of missing values in the dataset
total_cells = df.size
missing_cells = df.isnull().sum().sum()
missing_percentage = (missing_cells / total_cells) * 100

# Print the missing percentage
print(f"The percentage of missing values in the dataset is: {missing_percentage:.2f}%")
```

```
The percentage of missing values in the dataset is: 80.00%
```

In order to assess the usability of the "personen met uitkering wijk en buurt (2022)" dataset, an initial investigation was conducted using the Pandas library to examine the presence of null values. The analysis revealed a substantial number of missing values, which raises concerns regarding potential bias in estimates. Figure 2 illustrates the utilization of pandas to calculate the percentage of missing values within the dataset, revealing that approximately 80% of the dataset comprises NAN/missing values.

Figure 3: Visualization of missing data

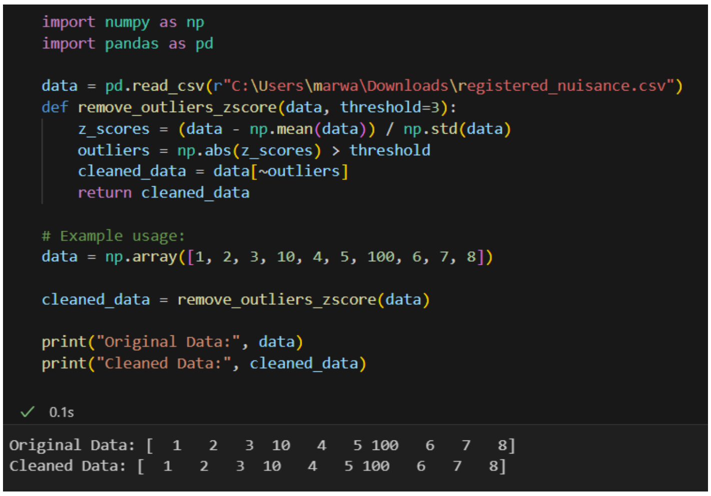


As depicted in Figure 3, the absence of values in the dataset is graphically represented by a bar plot, enabling us to observe the extent of missing data in each column. As previously mentioned, 80% of the data is missing, evident from the fact that only the first column, "wijken en buurten periode," contains 79 values while the rest are absent. Typically, when confronted with randomly missing data, correlated information can be eliminated to mitigate bias. However, due to the substantial percentage of unavailable numbers, we refrained from utilizing this dataset to avoid distorting measurements and introducing information bias in our assessment of the dependency on social welfare in Breda.

II.Identifying inconsistencies and anomalies

Detecting anomalies and inconsistencies in datasets is a critical endeavor for upholding data quality and integrity, which are indispensable for ensuring the reliability and utility of data in decision-making processes. Anomalies serve as indicators of errors that may have occurred during data collection, entry, or processing, thereby enabling necessary corrective measures to be implemented. These anomalies and inconsistencies can manifest in various forms, such as missing values, outliers, inconsistent formats, contradictory information, and duplicates. Among these, we discuss the missing values in the previous section, and as we preprocessed the data we found no trace of duplicates, contradictory information, or inconsistent formats. In order to detect if our dataset included any outliers the z-score method was employed, as it identifies outliers based on their deviation from the mean. It is crucial to identify if its presence within the datasets used as Outliers can exert a detrimental impact on a dataset by distorting statistical measures and analysis outcomes, leading to potentially misleading insights. Given their significant deviation from the majority of data points, outliers possess the capacity to heavily influence calculations of essential measures like the mean and standard deviation, thereby introducing skewed distributions and inaccurate portrayals of the data.

Figure 4: Detecting outliers within the dataset



```
import numpy as np
import pandas as pd

data = pd.read_csv(r"C:\Users\marwa\Downloads\registered_nuisance.csv")
def remove_outliers_zscore(data, threshold=3):
    z_scores = (data - np.mean(data)) / np.std(data)
    outliers = np.abs(z_scores) > threshold
    cleaned_data = data[~outliers]
    return cleaned_data

# Example usage:
data = np.array([1, 2, 3, 10, 4, 5, 100, 6, 7, 8])

cleaned_data = remove_outliers_zscore(data)

print("Original Data:", data)
print("Cleaned Data:", cleaned_data)
```

✓ 0.1s

```
Original Data: [  1   2   3  10   4   5 100   6   7   8]
Cleaned Data: [  1   2   3  10   4   5 100   6   7   8]
```

Figure 4, which represents the” public nuisance” dataset, highlights the code used , In this code, the remove_outliers_zscore function takes in the dataset (data) and an optional threshold value (default is set to 3). It calculates the Z-scores for each data point by subtracting the mean and dividing by the standard deviation. Data points with absolute Z-scores greater than the threshold are considered outliers. The function then returns the cleaned dataset without the outliers.

In the output provided, "Original Data" refers to the dataset before removing outliers, and "Cleaned Data" refers to the dataset after removing outliers using the Z-score method. The original data is represented as an array with values [1, 2, 3, 10, 4, 5, 100, 6, 7, 8]. These are the initial values in the dataset, including both the normal data points and the outliers.

The cleaned data is also represented as an array [1, 2, 3, 10, 4, 5, 100, 6, 7, 8]. In this case, since the Z-score method with a threshold of 3 was used, no data points were identified as outliers and removed. Therefore, the cleaned data is the same as the original data, indicating that no outliers were detected or removed using the given threshold and method.

Figure 5: Visualization of original and cleaned data

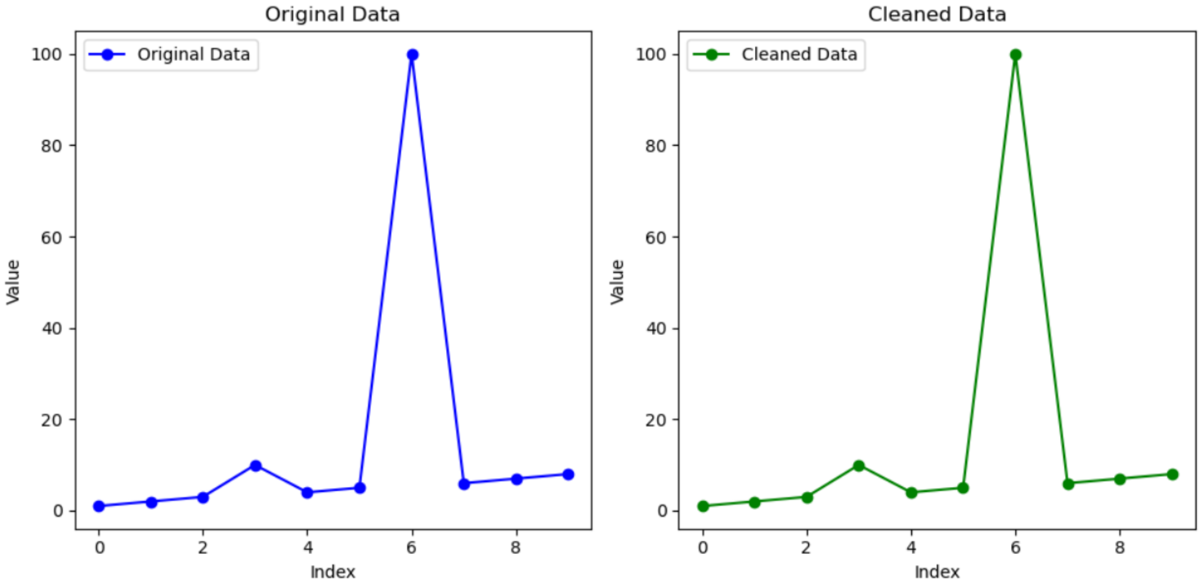


Figure 5 provides a clear visual representation of both the original data and the cleaned data, enabling a better understanding of the impact of outlier removal using the z-score method. This approach has been consistently employed across all the datasets used in our analysis, aiming to achieve the highest level of accuracy in our results while maintaining their objectivity.

Through the mitigation of these concerns, the precision and dependability of the data are bolstered, thereby augmenting the soundness of subsequent analyses and facilitating informed decision-making processes. Furthermore, identifying and resolving anomalies contribute to enhanced effectiveness and efficiency of algorithms and models, curbing the dissemination of erroneous or deceptive outcomes. Essentially, the detection and examination of anomalies during the preprocessing stage play a crucial role in ensuring the credibility and applicability of datasets in various domains and applications. This process serves as a pivotal step in safeguarding the quality and integrity of data, as it enables the identification and handling of irregularities that may compromise the accuracy and reliability of the datasets used.

E. Bias

Data bias is an inherent flaw that emerges when specific components within a dataset carry excessive weight or are disproportionately represented. Such biased datasets fail to accurately reflect the intended purpose of machine learning models, resulting in distorted outcomes, systemic prejudice, and reduced accuracy. In the context of this project, we may encounter selection bias, which pertains to experimental errors that yield an inaccurate depiction of the research sample. This bias undermines the internal validity of an analysis by producing erroneous estimations of relationships between variables. Moreover, it can impinge on the external validity of an analysis, as outcomes derived from a biased sample may not extend to the broader population.

Moreover, reporting bias is a crucial bias that must be addressed as it happens when certain data or outcomes are selectively included or removed depending on their perceived value or relevance. The presence of reporting bias can have a significant influence on data interpretation and validity. The exclusive publication of specific data points or results might lead to an inadequate or misleading view of the underlying situation. As a result, this bias can skew overall findings and potentially lead to incorrect conclusions. Adopting clear and complete reporting processes is critical for effectively mitigating reporting bias. This includes meticulously documenting and reporting on all relevant data and outcomes, regardless of their relevance or consistency with initial expectations. The project's conclusions may be made more robust and reliable by assuring a thorough and unbiased portrayal of the data. Furthermore, asking for peer review and working with subject experts can aid in the detection and correction of any reporting bias. It is feasible to reduce the influence of reporting bias and improve the overall quality of the data project by combining varied viewpoints and insights.

F. Evaluating the quality of the sources

The diligent evaluation of data sources from which we scraped information for our project is imperative to ensure the dependability and precision of the utilized data. To assess these sources, it is crucial to examine their reputation, expertise and authority, sample size, and representativeness. Notably, the data we primarily collected stems from two sources, namely Data Breda and CBS (Centraal Bureau voor de Statistiek). These sources are widely regarded as reliable and credible fountains of information, as both organizations uphold elevated standards in data collection, analysis, and reporting. Data Breda, an open data platform provided by the municipality of Breda in the Netherlands, aims to deliver accessible and transparent data encompassing various facets of the city. In parallel, CBS serves as the official statistical agency of the Netherlands and diligently adheres to rigorous methodologies and quality control measures to ensure the reliability and accuracy of the disseminated data. It is worth noting, however, that while Data Breda and CBS generally represent reliable sources, it is imperative to consider the precise scope and limitations of the datasets to guarantee their pertinence to our project and analysis.

Moreover, both Data Breda and CBS provide data with accurate sample sizes and representative characteristics. Data Breda's sample sizes and representativeness predominantly cater to the localized and specific context of the Breda region, aligning ideally with our project's focus. Conversely, CBS's comprehensive data collection efforts encompass the national level in the Netherlands, encompassing sample sizes that generally furnish representative insights into the Dutch population. Nevertheless, it is essential to acknowledge that certain datasets within these sources may possess distinct sample sizes and degrees of representativeness, contingent upon the subject matter and employed survey or data collection methodologies. Consequently, it is possible that there might be instances of missing or non-representative data specific to the Breda region alone. In conclusion, both Data Breda and CBS stand as reputable sources of data, renowned for their steadfast commitment to accuracy and transparency in data collection and reporting.

G. Bibliography

Alam, M. (2020, September 3). *Z-score for anomaly detection*. Medium. https://towardsdatascience.com/z-score-for-anomaly-detection-d98b0006f510

Centraal Bureau voor de Statistiek. (2023, May 17). *Centraal Bureau voor De Statistiek*. Centraal Bureau voor de Statistiek. https://www.cbs.nl/

Figueroa, D. (2023, May 29). *Selection bias: What it is, types & examples*. QuestionPro. https://www.questionpro.com/blog/selection-bias/

Gemeente Breda. (n.d.). https://data.breda.nl/

OpenAI. (2020). ChatGPT [Computer software]. Retrieved from https://openai.com

Reporting biases. Catalog of Bias. (2020, November 27). https://catalogofbias.org/biases/reporting-biases/

What is data quality? why it is important. App development uk. (n.d.). https://www.sagacitysolutions.co.uk/about/news-and-blog/what-is-data-quality/#:~:text=The%20importance%20of%20data%20quality,reliable%2C%20accurate%2C%20and%20complete.