GreenScape – Data Quality Report

This file contains data quality report of Team GreenScape. Our research questions are:

1. ***What are the factors, that might influence green score in Breda?***
2. ***Based on these factors, can we predict the green score?***

To answer these research questions, we had to perform data collection process, which had led us to Exploratory Data Analysis. Based on this analysis of each theme, that might be related to green index score, we performed an analysis of each topic.

Introduction: Each team member has contributed their skills and efforts to gather valuable information for this project. We have explored various aspects of Breda's city data, ranging from green index and liveability scores to the total number of houses and population in different neighbourhoods. Our data collection primarily relied on public access sources, such as Gemeente's website called Breda in Cijfers and the renowned data.politie.nl website. In addition, we utilized BUasVPN to access specific datasets. By combining and pre-processing these datasets, we aimed to conduct comprehensive exploratory data analysis (EDA) and evaluate the quality of the data for this project.

Datasets' Structure: Our team, GreenScape, has thoroughly examined multiple datasets for our project, focusing on key variables for performing exploratory data analysis (EDA). The datasets are primarily in CSV format and consist of the following variables: green index (green index score), liveability (liveability score), income recipients, workforce (working population per neighbourhood), neighbourhoods (districts and regions), public safety and nuisances in different regions. To assess the structure and quality of these datasets, we developed a universal Python function that provides detailed information about each dataset. The function outputs include the number of rows and columns, data types, the number and percentage of missing values per column, the total count of missing values, and the overall percentage of missing values. In the images provided, we have combined the outputs

applied to all the datasets we analysed. This comprehensive overview allows us to identify any inconsistencies or issues within the datasets. By addressing missing values through techniques like backfilling or replacing with zero, we aim to ensure the integrity and reliability of our research outcomes.

Fig. Structure of Green Index, Houses per Neighbourhood, Population per Neighbourhood, Population in general, grid keys data

```
                                Number of Rows  Number of Columns  \
Soort misdrijf                             748                  4
Perioden                                   748                  4
Wijken en buurten                          748                  4
Geregistreerde misdrijven (aantal)         748                  4

                                Column Data Types  \
Soort misdrijf                             object
Perioden                                    int64
Wijken en buurten                          object
Geregistreerde misdrijven (aantal)          int64

                                Number of Missing Values  \
Soort misdrijf                                         0
Perioden                                               0
Wijken en buurten                                      0
Geregistreerde misdrijven (aantal)                     0

                                Percentage of Missing Values  \
Soort misdrijf                                           0.0
Perioden                                                 0.0
Wijken en buurten                                        0.0
Geregistreerde misdrijven (aantal)                       0.0

                                Total Missing Values  \
...
Soort misdrijf                                   0.0
Perioden                                         0.0
Wijken en buurten                                0.0
Geregistreerde misdrijven (aantal)               0.0
```

Fig. Structure of Public Safety Data

```
                               Number of Rows  Number of Columns  \
Overlast                                  612                  4
Perioden                                  612                  4
Wijken en buurten                         612                  4
Geregistreerde overlast (aantal)          612                  4

                               Column Data Types  Number of Missing Values  \
Overlast                                  object                         0
Perioden                                   int64                         0
Wijken en buurten                         object                         0
Geregistreerde overlast (aantal)           int64                         0

                               Percentage of Missing Values  \
Overlast                                                0.0
Perioden                                                0.0
Wijken en buurten                                       0.0
Geregistreerde overlast (aantal)                        0.0

                               Total Missing Values  \
Overlast                                          0
Perioden                                          0
Wijken en buurten                                 0
Geregistreerde overlast (aantal)                  0

                               Total Percentage of Missing Values
Overlast                                                      0.0
Perioden                                                      0.0
Wijken en buurten                                             0.0
Geregistreerde overlast (aantal)                              0.0
```

Fig. Structure of Nuisances Data

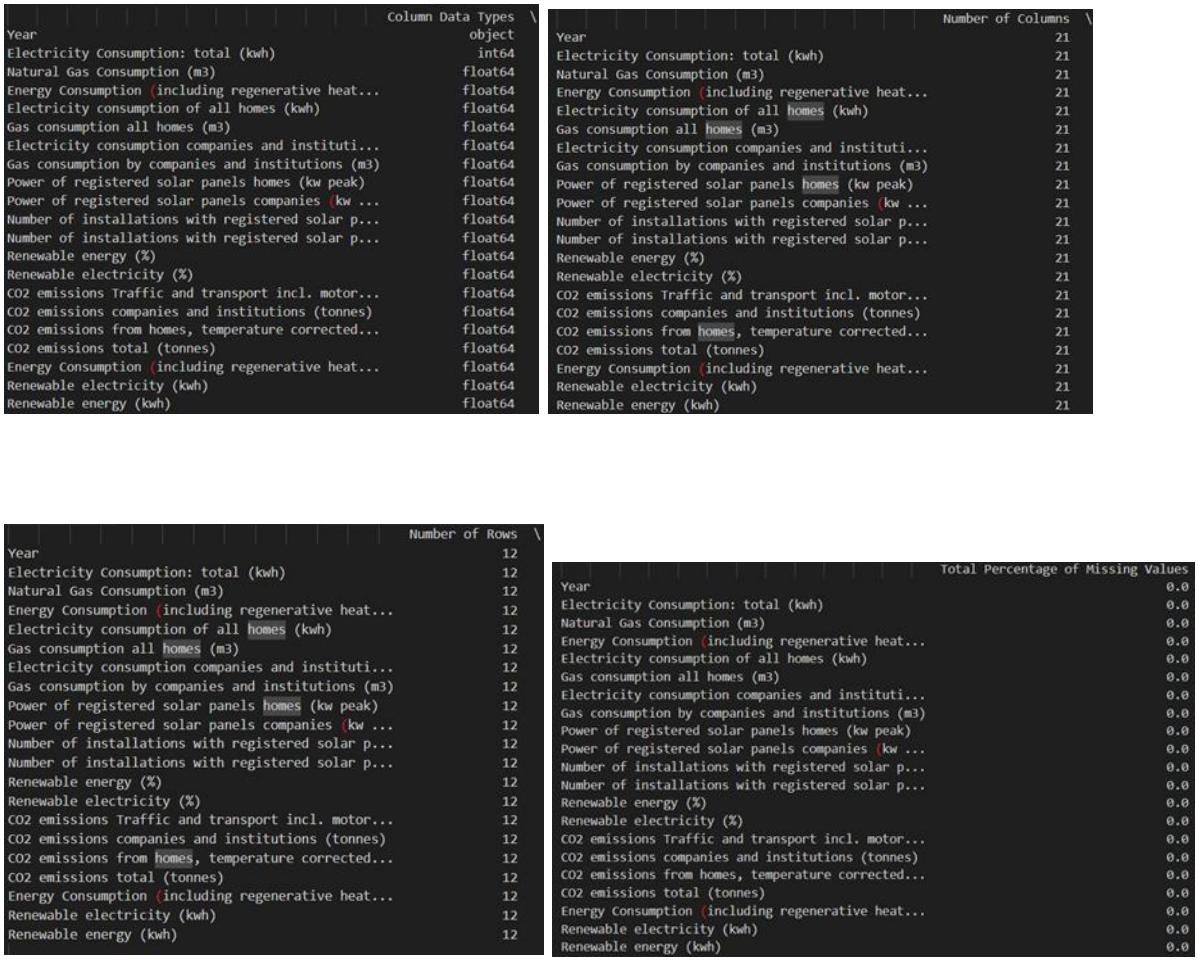
```
df_population.info()
✓ 0.2s
<class 'pandas.core.frame.DataFrame'>
Int64Index: 840 entries, 774 to 464
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   neighborhood  840 non-null    object
 1   regions       840 non-null    object
 2   year          840 non-null    object
 3   population    840 non-null    int64
dtypes: int64(1), object(3)
memory usage: 32.8+ KB
```

```
df_income.info()
✓ 0.2s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 448 entries, 0 to 447
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   neighborhood  448 non-null    object
 1   regions       448 non-null    object
 2   year          448 non-null    object
 3   income_score  448 non-null    object
dtypes: object(4)
memory usage: 14.1+ KB
```

```
df_working_pop.info()
✓ 0.0s
<class 'pandas.core.frame.DataFrame'>
Int64Index: 112 entries, 46 to 72
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   neighborhood        112 non-null    object
 1   regions             112 non-null    object
 2   year                112 non-null    object
 3   working_population  112 non-null    object
dtypes: object(4)
memory usage: 4.4+ KB
```

```
{'Number of Rows': 840, 'Number of Columns': 4}
{'Number of Rows': 448, 'Number of Columns': 4}
{'Number of Rows': 1171, 'Number of Columns': 7}
{'Number of Rows': 112, 'Number of Columns': 4}
```

Fig. Structure of Green index score, Livability score, Income recipients, Workforce, Neighborhoods



	Column Data Types	Number of Columns	Number of Rows	Total Percentage of Missing Values
Year	object	21	12	0.0
Electricity Consumption: total (kwh)	int64	21	12	0.0
Natural Gas Consumption (m3)	float64	21	12	0.0
Energy Consumption (including regenerative heat...	float64	21	12	0.0
Electricity consumption of all homes (kwh)	float64	21	12	0.0
Gas consumption all homes (m3)	float64	21	12	0.0
Electricity consumption companies and instituti...	float64	21	12	0.0
Gas consumption by companies and institutions (m3)	float64	21	12	0.0
Power of registered solar panels homes (kw peak)	float64	21	12	0.0
Power of registered solar panels companies (kw ...	float64	21	12	0.0
Number of installations with registered solar p...	float64	21	12	0.0
Number of installations with registered solar p...	float64	21	12	0.0
Renewable energy (%)	float64	21	12	0.0
Renewable electricity (%)	float64	21	12	0.0
CO2 emissions Traffic and transport incl. motor...	float64	21	12	0.0
CO2 emissions companies and institutions (tonnes)	float64	21	12	0.0
CO2 emissions from homes, temperature corrected...	float64	21	12	0.0
CO2 emissions total (tonnes)	float64	21	12	0.0
Energy Consumption (including regenerative heat...	float64	21	12	0.0
Renewable electricity (kwh)	float64	21	12	0.0
Renewable energy (kwh)	float64	21	12	0.0

Fig. Structure of CO2 Emissions, Energy and Electricity Usage, Renewable Energy and Electricity production, and Solar Panel Installations

	Number of Rows	Number of Columns	Column Data Types	Number of Missing Values	Percentage of Missing Values	Total Missing Values	Total Percentage of Missing Values
Year	298038	7	int64	0	0.000000	54	0.002588
Image ID	298038	7	object	0	0.000000	54	0.002588
Latitude	298038	7	object	0	0.000000	54	0.002588
Longitude	298038	7	object	0	0.000000	54	0.002588
Date	298038	7	object	0	0.000000	54	0.002588
Green Score	298038	7	int64	0	0.000000	54	0.002588
CBS Grid Code	298038	7	object	54	0.018118	54	0.002588

Fig. Data quality assessment for Green Index dataset

	Number of Rows	Number of Columns	Column Data Types	\
Neighborhood	41	2	object	
Income Index	41	2	int64	

	Number of Missing Values	Percentage of Missing Values	\
Neighborhood	0	0.0	
Income Index	0	0.0	

	Total Missing Values	Total Percentage of Missing Values
Neighborhood	0	0.0
Income Index	0	0.0

Fig. Data quality assessment for Income dataset

Data Accuracy: As a team, GreenScape, we recognize the critical importance of data accuracy in relation to our proposed business case, which aims to leverage data analysis and predictions to improve the municipality of Breda. Considering our research questions and objectives, we have thoroughly assessed the data we collected for its accuracy. The datasets we obtained, including variables such as green index, liveability score, income recipients, workforce, and neighbourhoods, have been sourced from reliable and up-to-date platforms, such as Breda in Cijfers and the municipality of Breda's website. This ensures that the data aligns with our project's requirements and can effectively support the development of a prediction model for the green score. While discrepancies in specific data points have been observed, such as variations in reported CO2 emissions, we believe that these inconsistencies arise from different measures and sources of data. We trust the data provided to us as it reflects the unique insights and information, we require to make accurate predictions and drive positive changes in Breda. By carefully validating and cross-referencing the data and leveraging our domain knowledge, we are confident in the accuracy of the collected data for our business case.

Data Completeness: When assessing the completeness of the datasets we have gathered, it is evident that some datasets contain missing values, indicating that the data is not 100% complete. Out of the five datasets found, three of them exhibit missing values. The dataset with the highest total percentage of missing values is the one concerning the total number of inhabitants in Breda. It is worth noting that this dataset was downloaded as an xlsx file and required proper formatting using pandas to work with it effectively. Additionally, the datasets related to population by neighbourhood and grid codes also exhibit missing values. To

ensure the data is comprehensive and suitable for further investigation and preparation for machine learning, significant preprocessing steps were performed. These steps involved properly formatting the xlsx file, handling missing values, and eliminating irrelevant columns. By performing these preprocessing steps, we aim to achieve a more complete dataset that can be utilized for data analysis and modelling. It is crucial to acknowledge that while the datasets may not be 100% complete initially, the preprocessing steps undertaken help address these issues and enhance the data's completeness and usability for our project.

Data Relevance: As a team, GreenScape, we have carefully assessed the relevance of the datasets for our business case. After conducting exploratory data analysis (EDA), it is evident that most of the datasets are highly relevant to our project objectives. These relevant datasets provide valuable insights into various factors that contribute to improving the municipality of Breda, such as the green index score, liveability score, income recipients, workforce, and neighbourhood-level data. By analysing these datasets, we can gain a comprehensive understanding of the factors that influence the green score and explore ways to enhance the city's green initiatives. While one dataset related to the total population of Breda was initially included, it was later determined to be less relevant due to the availability of a more specific and informative dataset representing population at the neighbourhood level. This decision ensures that our analysis focuses on the most relevant and granular data for our business case. Additionally, the inclusion of geographical coordinates (latitude and longitude) in the green index dataset proves to be relevant as it enables the creation of an interactive map showcasing the distribution of green score points across the city of Breda.

Data Consistency: We have carefully examined the consistency of the datasets used in our project. Based on our research and analysis, we have observed that the datasets exhibit partial consistency. One aspect of inconsistency is evident in the dataset representing the total population in Breda. When comparing this data to external sources, such as search engine results, we found that the population numbers provided in the dataset do not consistently reflect the most up-to-date information. It appears that the dataset is not

consistently updated year by year, but rather provides population values for specific points in time, such as January values for each year. Furthermore, there are variations in the time periods covered by the datasets. For example, the green score dataset contains data for the years 2009-2010 and then a gap until 2014. This discrepancy in time coverage can impact the analysis and interpretation of the data, requiring careful consideration and handling during our project. While these inconsistencies exist, they can be addressed by implementing appropriate data manipulation techniques and taking into account the limitations of the datasets. By acknowledging these inconsistencies and conducting thorough data analysis, we can still derive valuable insights and make informed decisions to improve the municipality of Breda.

Data Accessibility: The datasets used in our GreenScape project are readily accessible and can be obtained from reliable sources. The data is conveniently available for download on the Breda in Cijfers and data.politie.nl websites and can be accessed by team members and product owners without significant obstacles. With regards to the accessibility of the data among team members, individuals can easily retrieve and work with the datasets. The data is openly accessible, ensuring efficient collaboration and seamless sharing of information within the team. While there might have been occasional technical difficulties or remote connection issues experienced by some individuals, overall, the datasets were accessible to the team. The use of VPN and other necessary measures facilitated the retrieval of data from Gemeente's database. Additionally, the availability of the datasets on open platforms, such as data.politie.nl and Breda in Cijfers, further enhances the accessibility and ease of obtaining the required data. Therefore, considering the accessibility of the datasets from both the source platforms and within the team, we can affirm that the data used in our GreenScape project is accessible and can be efficiently utilized for analysis and decision-making processes.

Data Timeliness: The datasets used in our GreenScape project primarily provide data on an annual basis, indicating that the information is updated and made available once per year.

However, there is a lack of datasets that offer more frequent updates, such as monthly or quarterly data. This restricts the timeliness of our data, as it does not provide real-time or granular insights into the variables of interest. While the data obtained is useful for estimating the green score index for future years and gaining a broad understanding of the city's performance over time, having more frequent measurements would enable us to capture seasonal variations and better identify factors influencing consumption patterns. In order to enhance the timeliness of our data, it would be beneficial to explore sources or datasets that offer more regular updates, such as monthly or quarterly data. This would provide us with more up-to-date and detailed information, facilitating the development of more effective strategies and interventions. Therefore, it is important to acknowledge the limitations of our current data timeliness and consider the potential benefits of incorporating more frequent and granular data sources into our analysis and decision-making processes.

Recommendations: Data Integration: Ensure that all relevant datasets are consolidated and integrated into a single cohesive dataset. This will facilitate easier analysis and modelling by reducing the need to work with multiple separate datasets.

Missing Value Handling: Develop a standardized approach for handling missing values in the datasets. Consider using techniques such as imputation (e.g., mean imputation, regression imputation) or advanced methods (e.g., multiple imputation) to fill in missing values, where appropriate. Document the missing value handling process to maintain transparency and reproducibility.

Data Validation and Cross-Referencing: Establish a robust process for validating and cross-referencing the collected data with external sources or benchmarks. This will help identify and address inconsistencies or discrepancies in the data, ensuring its accuracy and reliability.

Regular Data Updates: Explore options to obtain more frequent updates for the datasets, especially for variables that may change frequently or exhibit seasonal variations. Consider

collaborating with relevant data providers to establish a system for regular data updates, such as quarterly or monthly, to improve the timeliness of the data.

Documentation and Metadata: Maintain comprehensive documentation for each dataset, including details on data sources, collection methods, preprocessing steps, and any assumptions made during the data analysis process. This will enhance the transparency and reproducibility of the research findings and allow others to understand and validate the data management process.

Data Quality Metrics: Implement additional data quality metrics to assess the quality of the collected data. This could include metrics such as data consistency checks, outlier detection, or statistical validation against external benchmarks. These metrics will help identify potential data issues and ensure the reliability of the analysis and predictions.

Data Governance and Security: Establish clear data governance practices and protocols to ensure data security and compliance with relevant regulations and privacy policies. Implement appropriate access controls and data protection measures to safeguard sensitive information.

Collaboration and Communication: Foster collaboration and effective communication within the team to address data-related challenges and share knowledge. Regularly discuss and update the team on data management strategies, findings, and potential improvements to ensure everyone is aligned and informed.

.

Teamwork:

1. Kacper Janczyk: Green Index, Houses per Neighbourhood, Population per Neighbourhood, Population in general, grid keys data.
2. Yuliia Bobrovytska: Public safety and Nuisances data.

3. Simona Dimitrova (222667): Liveability, districts and regions in Breda, Population per Neighbourhood, Income recipients per Neighbourhood, Green Index, Working population per Neighbourhood.
4. Tatár Mátyás: CO2 Emissions, Energy and Electricity Usage, Renewable Energy and Electricity production, and Solar Panel Installations.
5. Collin Limoncelli-Buyskes: Green Index and Income.

All Contributors to this document have complied with Legal and Ethical guidelines by ensuring all steps of Data Selection and Pre-processing are available for inspection through the team Github Repository.

Sources:

IBM Documentation. (n.d.). https://www.ibm.com/docs/en/spss-modeler/18.2.0?topic=quality-writing-data-report

Monsanto, C. M. (2022, July 1). Data Quality: A Comprehensive Overview [+Examples]. Hubspot. https://blog.hubspot.com/website/comprehensive-overview-of-data-quality