Greenscape – Data quality report, personal contribution of Kacper

Kacper Janczyk: Green Index, Houses per Neighbourhood, Population per Neighbourhood, Population in general, grid keys data

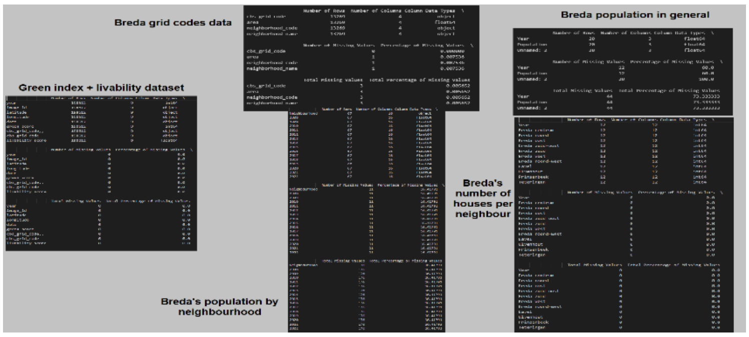
Introduction: When it comes to my part of the data collection and analysis for this project, I chose to search for data regarding green index (green score and liveability score), total number of houses in Breda per neighbourhood, total number of people living in Breda per neighbourhood and total number of people living in Breda. Majority of the datasets I found are from Gemeente's website containing public access data called Breda in Cijfers. The green index + liveability score dataset was downloaded by accessing Gemeente's database with usage of VPN. Simple query was conducted to combine green_index and livability_index tables.

Datasets' structure: To check how structured are the dataset, I've created a universal Python function, which prints out information about every dataset (data frame in Pandas), such as:

1. Number of rows
2. Number of columns
3. Data type
4. Number of missing values (per column)
5. Percentage of missing values (per column)
6. Total missing values
7. Total percentage of missing values

In the image below, I combined every output of the function applied to every dataset mentioned before.

Figure 1. Structures of datasets


Breda grid codes data
Breda population in general
Green index + livability dataset
Breda's number of houses per neighbour
Breda's population by neighbourhood

Data accuracy: Data accuracy is a term used for determining, whether the data we have is useful or not, regarding proposed business case. In case of this project and its' business case, which is how to improve the municipality of Breda using data analysis and predictions, and considering our research questions mentioned at the beginning, I would say the data I was able to collect is accurate. Our main objective is to make a prediction model of green score, based on different factors which might contribute to its' value, such as number of people in each of the neighbourhoods, number of houses in the neighbourhoods and liveability score. Therefore, I can agree, that the data I found is accurate for this business case.

Data Completeness: In this section, we need to answer a simple question, which is "Is my data comprehensive?". A complete dataset shouldn't have any missing values, duplicates or irrelevant values. In the figure 1, we can see, that some of the datasets found are indeed missing values (3/5 datasets found do have some number of missing values). The highest total percentage of missing values can be seen in data regarding total number of inhabitants in Breda. What's important to mention, is that the dataset was downloaded as xlsx file, which had to be formatted correctly to work with it, using pandas. Other datasets with missing values are population by neighbourhood and grid codes. Therefore, I cannot tell, that the data I found is 100% complete. Important preprocessing steps had to be performed, to have data, which can be called "complete" and can be used for further investigation and preparation for Machine Learning.

Data Relevance: In terms of data relevance, this term basically describes the level of consistency between the content of data and the area of interest. After conducting EDA on the datasets, I can tell, that 4/5 datasets are relevant for this business case. The only dataset that I later found not to be relevant is the total amount of population in Breda. This dataset could be skipped, since I found a dataset representing population of Breda on neighbourhood level, which is more relevant in terms of this project. What could have been argued are columns of latitude and longitude in green index dataset, but at the end I also created an interactive map, which represents green score points on the map of Breda, which I couldn't find on Gemeente's website. Therefore, geographical coordinates are relevant for this project as well.

Data Consistency: When it comes to data consistency of the datasets, based on research I did on this matter, the data used for this project is partially consistent. What I mean by that, is for example with total population in Breda. When we search for "population in Breda" on Google, the number is equal to January values for each year, that can be found on Breda in Cijfers website. It is not updated consistently according to each year. Also, the datasets have different time periods, for example there's a time gap when it comes to green score dataset (2009-2010 and then from 2014).

Data Accessibility: The datasets can be downloaded and easily accessed on Breda in Cijfers website. When it comes to accessibility of the data within team members and product owners, the data can be easily accessed by multiple people.

Data Timeliness: Data timeliness is basically about how quickly data is captured, processed and made available to us. When it comes to datasets I discovered, the data is mostly yearly. We only get an information per year. I couldn't find a dataset that would be represented by month for example, so I can say that my data is not so timeliness. It would be more timeliness, if it was updated monthly and if the data itself was represented by month.

Recommendations:

Data Completeness: I suggest addressing the issue of missing values in the datasets. It is crucial to investigate the reasons behind the missing data and take steps to fill in the gaps where possible. If certain data is consistently missing or unavailable, I recommend exploring alternative data sources or strategies to ensure data completeness.

Data Consistency: It would be beneficial to establish a more consistent and regular updating schedule for the datasets. This will help ensure that the data remains relevant and up to date. I believe that adopting a standardized approach to data collection and processing will contribute to maintaining consistency over time.

Data Timeliness: Considering the current data collection frequency, I propose collecting and updating data at shorter intervals, such as monthly or quarterly. This will provide a timelier view of the municipality's characteristics and enable better monitoring of changes and trends over time.

Data Integration: I suggest exploring opportunities to integrate data from multiple sources to enrich the existing datasets. Identifying additional relevant datasets that complement the current information will provide a more comprehensive understanding of the municipality. This could include incorporating data on environmental factors, infrastructure, economic indicators, or social demographics, depending on the project's objectives.

Data Quality Assurance: To ensure the accuracy and reliability of the data, I recommend implementing quality assurance processes. This may involve data validation, cross-referencing with external sources, and employing data cleansing techniques to identify and rectify inconsistencies or errors in the datasets.

Data Accessibility and Collaboration: Enhancing data accessibility for team members and stakeholders is crucial. I propose implementing a centralized data management system that facilitates secure and efficient sharing of data among authorized users. Implementing appropriate access controls and permissions will ensure data security and privacy while enabling collaboration and knowledge sharing.

Data Documentation: I believe it is important to develop comprehensive documentation for each dataset, outlining data sources, collection methods, and any preprocessing or transformations applied. This documentation will provide a clear understanding of the data's context and limitations, promoting transparency and reproducibility in analyses.

Sources:

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