Yuliia Bobrovytska: Public safety data + Nuisances data

***Introduction:*** In terms of my role in gathering and examining data for this particular project, I made the deliberate decision to focus on uncovering a plethora of information pertaining to the vital aspect of public safety within the city of Breda. As a result of my research efforts, I successfully managed to procure 2 datasets specifically obtained from the renowned data.politie.nl website.

***Datasets’ structure:*** To check how structured are the dataset, I’ve used 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 images below, I put every output of the function applied to every dataset mentioned before.

*Figure 1. Structure of Public Safety Data*

A picture containing text, screenshot, menu

Description automatically generated

*Figure 2. Structure of Nuisances Data*

A picture containing text, screenshot, menu

Description automatically generated

***Data accuracy:*** Data accuracy is a pivotal concept that involves the assessment of the usefulness of the data at hand in relation to a proposed business scenario. When considering the present project and its corresponding business case, which revolves around enhancing the municipality of Breda through the implementation of data analysis and predictive techniques, and taking into account the initial research questions posed, it is my assertion that the data I have managed to gather possesses a high degree of accuracy. The primary aim we have set forth is to construct a predictive model of the green score, drawing upon various factors that are likely to influence its value, such as public safety and nuisances across different regions within Breda. Therefore, I concur that the data I have procured is indeed accurate and aptly aligned with the requirements of this particular business case.

***Data Completeness:*** Within this particular section, our task is to address a rather straightforward inquiry, which pertains to the extent of comprehensiveness in relation to our data. It is essential for a dataset to be considered whole and comprehensive that it possesses a lack of any missing values. As we observe the contents presented in figures 1 and 2, it becomes evident that these datasets exhibit an absence of any missing values. Consequently, I am able to confidently assert that the data I have come across stands as a true testament to its complete nature, attaining a flawless completion rate of 100%.

***Data Relevance:*** In relation to the significance of data, this particular term essentially elucidates the degree of coherence existing between the substance of the data and the specific domain of focus. Following an extensive exploratory data analysis (EDA) performed on the datasets, it becomes apparent that both of my datasets exhibit a considerable degree of relevance to this particular business scenario.

***Data Consistency:*** Regarding the issue of data consistency in the datasets, after conducting thorough research on this subject, I can confidently state that the data utilized for this project exhibits a remarkably high level of The Public Safety Dataset encompasses an extensive range of available data spanning from the year 2012 to 2022, providing a comprehensive historical perspective. Conversely, the Nuisances Dataset solely offers available data exclusively for the year 2022.

***Data Accessibility:*** The datasets are conveniently available for download and can be readily accessed through the website data.politie.nl. In terms of the accessibility of the data among team members and product owners, multiple individuals can effortlessly access the data without any difficulty.

***Data Timeliness:*** Data timeliness refers to the swiftness with which data is obtained, processed, and presented to us. Upon examining various datasets, I observed that the majority of the data is compiled on an annual basis, meaning that we receive information only once per year. Regrettably, I was unable to locate a dataset that offers monthly representation, underscoring the lack of timeliness in my data. It would undoubtedly be more timely and up-to-date if it were updated on a monthly basis, with the data itself being organized and represented at the monthly level.