Data Quality Report

1. Introduction (Thomas)

In the following report, we will discuss some highlighted points for data usage in our project to ensure the data we use is valuable. The first step will be to evaluate the completeness of our data, essentially examining if there is any missing or duplicated data. Next is to identify inconsistencies or anomalies, this could include looking for outliers or unusual patterns in the data. Following we check for bias in our data, this step is especially important because we want to ensure that the data is accurate. The next step is evaluating the quality of the data which links to the same importance of accurate and usable data.

2. Evaluation of the completeness of our data (Thomas)

Evaluating the completeness of our data was quite simple since we have done the EDA process in previous blocks and once again, checking for completeness is part of EDA thus it comes quite automatically in the steps. In this step, as mentioned before we first go through the data and see if there are any quick and visible fixes, after this we can use functions that will delete or edit any empty data rows or duplicated data. This process will either remove rows or change the values to 0 or NaN following to this we can then delete the NaN values as well. After we finish evaluating the completeness of the data we have a clean dataset that we can work on without any concern for missing or unusable data. By having a complete dataset we can move to the next step which would be Identifying any inconsistencies or anomalies.

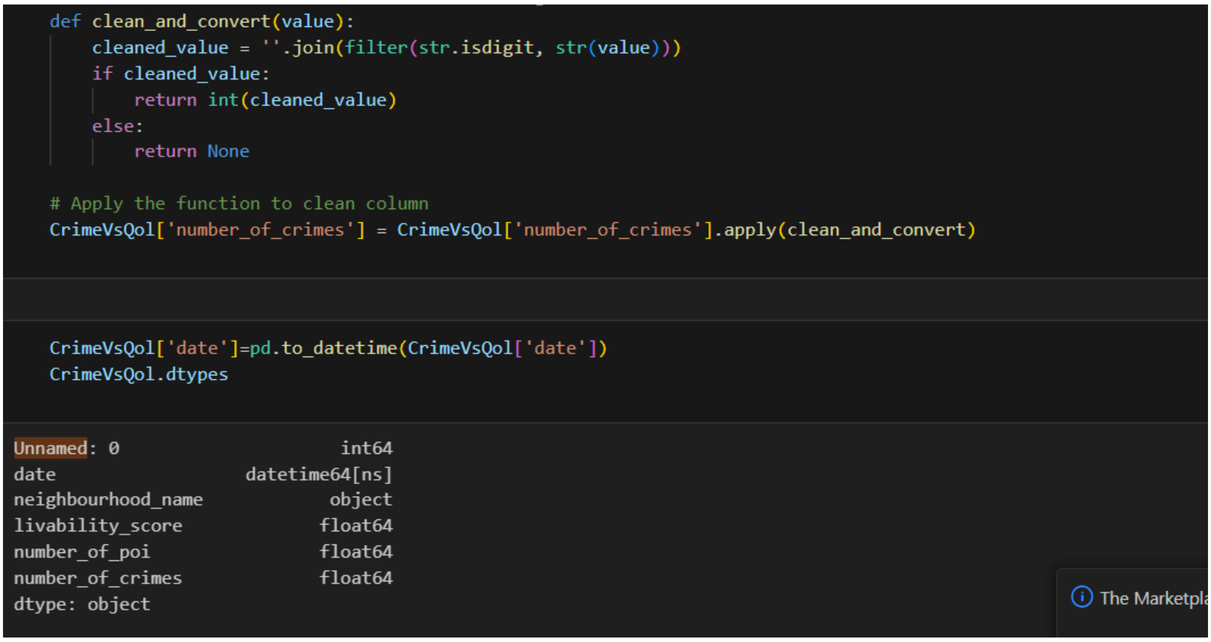
3. Identifying any inconsistencies or anomalies (Amyr)

Now we will focus on identifying inconsistencies or anomalies within the dataset. High-quality data is essential for obtaining clear insights and enables efficient visualization and predictive modeling. By addressing inconsistencies and anomalies, we can improve the accuracy and reliability of our data-driven analyses.

To begin our analysis, we will first load the dataset and examine it using a tabular representation. During this process, we will look for potential issues such as duplicates, missing values (NaN), and data type discrepancies. These issues can be indicators of inconsistencies or anomalies within the dataset, and by identifying and resolving them, we can ensure the integrity of our data.
we made use of the data that was provided
There are many types of inconsistencies or anomalies:
updated anomalies, deleted anomalies, inserted anomalies.

Types of inconsistencies we found in the data is that some data that are supposed to be the data type int are object types, cause the data happens to have multiple spaces behind it (example:”14 .”).
The data frame registered the above-mentioned example as an object:

number_of_poi float64
number_of_crimes object



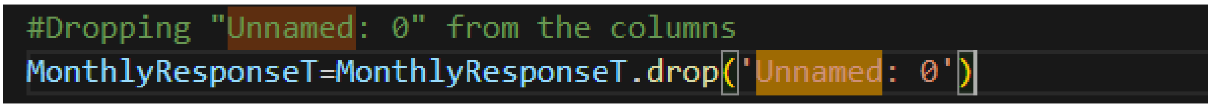
```
def clean_and_convert(value):
    cleaned_value = ''.join(filter(str.isdigit, str(value)))
    if cleaned_value:
        return int(cleaned_value)
    else:
        return None

# Apply the function to clean column
CrimeVsQol['number_of_crimes'] = CrimeVsQol['number_of_crimes'].apply(clean_and_convert)
```

```
CrimeVsQol['date']=pd.to_datetime(CrimeVsQol['date'])
CrimeVsQol.dtypes
```

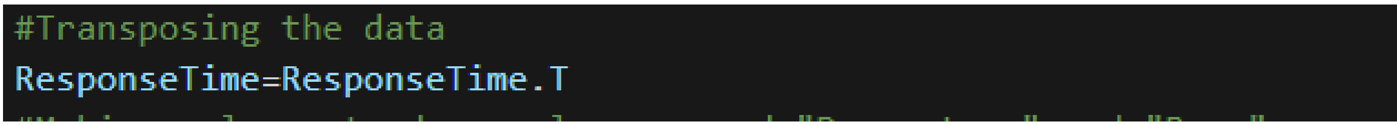
```
Unnamed: 0                    int64
date                 datetime64[ns]
neighbourhood_name           object
livability_score            float64
number_of_poi               float64
number_of_crimes            float64
dtype: object
```

We needed to erase the spaces behind the 14 to be able to transform it into an int or float data type.
There are some duplicated values in columns that hold different values in other columns so it’s hard to isolate them and work around it.
There are also times when the data has the wrong data type or cannot be transformed to a specific data type making it hard to work with.
There are times when we also get columns that are registered as “unnamed: 0” and we’ll have to see what value this column provides to drop it or rename it to visualize it.



```
#Dropping "Unnamed: 0" from the columns
MonthlyResponseT=MonthlyResponseT.drop('Unnamed: 0')
```

Most of the data started with unnecessary rows and we had to manually skip these to be able to work with the data properly, there were times when we had to transpose the data frame to be able to make a bit more sense of the data as well.



```
#Transposing the data
ResponseTime=ResponseTime.T
```

4. Biases (Dominik)

During our analysis of the datasets, we found that our dataset does not contain any sensitive information. Additionally, although we focused on data about specific neighborhoods, the presence of more crime in certain regions does not mean that the data is biased. Moreover, there are no neighborhoods without data about the number of crimes. While this suggests that our data may not contain obvious biases, it is important to be aware that we may still encounter biases that have emerged during the annotation process. However, it can be challenging, and in some cases, even impossible to precisely identify such biases.

5. Quality of data sources (Romina)

This data quality report evaluates the reliability and suitability of a merged dataset that combines information on livability scores, crime rates, and points of interest (POIs) in Breda. The dataset covers the period from May 2022 to April 2023 and aims to explore the variations across different neighborhoods in Breda.

The merged dataset is sourced from multiple data sources, including:

* Data source 1: Crime rates in Breda from May 2022 to April 2023
* Data source 2: Livability scores for different neighborhoods
* Data source 3: Number of POIs in each neighborhood in Breda

# Evaluation findings

## Data Accuracy

The accuracy of the data within the merged dataframe was carefully evaluated to ensure the correctness and precision of the reported value. To assess the data accuracy, several steps were taken:

* The reported crime rates, livability scores, and number of POIs were referenced with reliable sources, such as official crime statistics, reputable livability assessments, and verified POI databases from the police website of the Netherlands.
* Data cleaning techniques were applied to identify and correct any errors in it. This involved removing duplicate ratings, addressing missing or incomplete data, and applying appropriate data transformation methods.

## Data completeness

We carefully checked the dataset to make sure it included all the necessary information for a thorough analysis.

We made sure that the data for all the neighbourhoods in Breda during the specified timeframe (May 2022 - April 2023) was present in the dataset. Below you can see the entire dataframe and the columns it contains.

Regarding the earlier mentioned heading about data accuracy and cleaning techniques, you may have noticed that the column "Unnamed: 0" consistently appears when loading a dataframe. It is not a problem to drop this column each time, as long as it is clear.

However, this ensures that our analysis covers all the neighborhoods in Breda and gives us a comprehensive understanding of the factors we're studying.

## Data consistency

We carefully looked at the livability scores, number of POIs, and crime rates for each neighborhood during different months. We wanted to see if there were any big changes or unusual patterns. Luckily, we didn't find any significant fluctuations or strange things happening. This means that the data is reliable and consistent, and we can trust the information it provides.

For example to prove it, let's say we examined the livability scores, number of points of interest (POIs), and crime ratings from the neighborhood "Valkenberg" in different months. We compared data from May, June, July, and so on.

After analyzing the information, we didn't notice any major ups and downs or unexpected trends. The livability scores remained relatively stable, without significant changes that could raise concerns. The number of POIs in the neighborhood also exhibited a consistent pattern, without drastic increases or decreases.

Additionally, the crime rates in Valkenberg remained relatively constant throughout the months, without any sudden spikes or unusual fluctuations. This consistency in the crime rates further confirms the reliability of the data.

Consistency in data fosters trust and reliability for analysis and decision-making.

## Data timeliness

The dataset includes recent months, allowing for an up-to-date analysis of trends in livability, crime rates, and POIs. This temporal relevance enhances the dataset's value in understanding the current dynamics in Breda.

## Data relevance

The information in the dataset is important and useful for studying how livable an area is, the level of crime, and the places of interest in different neighborhoods of Breda. It gives a complete picture of these factors, allowing us to compare and understand how they vary across different areas. This information is relevant and valuable for understanding the quality of life, safety, and attractions in Breda.

## Data documentation

Unfortunately, there is no detailed explanation about where the data comes from, how it was collected, or any limitations it may have. Having this kind of information is important because it helps us understand how reliable and accurate the data is. It also allows others to use the same methods and understand the data in the same way.

There could be one solution to conduct further research and investigation. This can involve reaching out to the entity or organization that provided the dataset and requesting additional information regarding its origin and collection methods. Additionally, reading and exploring documentation, such as research articles or reports, may provide insights into the data sources used.

## Data integration

This data integration process involved merging multiple data frames obtained from the official police website of the Netherlands. These data frames were carefully combined to create a comprehensive dataset for analysis.

By consolidating the information from various sources, we have created a unified dataset that provides a holistic view of crime rates, livability scores, and points of interest in different neighborhoods of Breda. This integration ensures that the data is complete, consistent, and ready for meaningful analysis and decision-making.

6. Data management strategy (Fedor)

While gathering, exploring, and working with the data provided by the municipality of Breda we encountered several issues regarding the management of the data. We would like to propose several improvements in the data management strategy which in our opinion will make it easier for upcoming projects to be valuable and provide new insights into our city.

a. While gathering the data we had to go through a significant amount of different web platforms with the data about Breda and download every dataset manually. These unnecessary steps are time-consuming and they lead to lower productivity. We suppose that creating an SQL server with all the data about Breda will significantly increase the productivity of Data Scientists as they will not have to spend time browsing the web looking for useful data, also SQL server is an easy way to keep track of the data that is available.

b. While working with several data sets we faced problems related to the lack of explanation of the data we are working with. We had to spend time figuring out how the data was gathered and what different observations in the datasets meant. Creating an overview with a basic explanation of the dataset will help Data Scientist to start working with the data right away.

c. We are living in an international community and it would be nice to have two versions of the datasets one in Dutch and one in English, so the data would be understandable for everyone right away without any preprocessing steps.

7. Conclusion (Fedor)

Throughout our project, we were lucky to work with relatively high-quality data. Of course, it was not perfect and we still had to deal with missing values, anomalies and other problems described in this report. We have thoroughly analysed and described the data we dealt with and proposed several changes for the data management strategy from which in our opinion municipality will only benefit.