Compiling Trends in Enrollment rates, Government Spending and Impoverished Children in the Netherlands

Group 12

Thijs Blom 2038447

Jan-Willem Koopman 1472070

Cristóbal Sendín 2025825

Darshan Sudheer Amadalli 1961306

June 20, 2024

This datasheet was prepared following the "Datasheets for Datasets" template of Gebru et al. [2], extended where needed following the frameworks of Pushkarna et al. [6] and Paullada et al. [5].

1 Motivation For Data/Knowledge Creation

Why was the dataset/knowledge graph created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

The dataset we have gathered primarily aims to address the research question posed by our client.

How does the distribution of educational resources across different regions and socioeconomic groups in the Netherlands influence enrollment rates over time?

With the following subquestions:

1. What trends have been observed in the distribution of educational resources (e.g., teachers, funding, etc.) across different regions and socioeconomic groups in the Netherlands?
2. How have enrollment rates in primary, secondary, and tertiary education evolved in relation to these resource distributions over different time periods?

What (other) tasks could the dataset be used for?

The dataset is highly focused on addressing our client's main question and subquestions. These questions are quite specific, targeting different aspects such as socio-economic factors and enrollment rates. Therefore, apart from our client's concern regarding identifying the relationship between educational resources, enrollment rates, and socio-economic factors, there are few other ways to interpret the data present in this dataset other than using each metric individually.

Any other comments?

None.

2 Dataset and Knowledge Graph Composition

What do the instances that comprise the dataset represent?(e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

The datasets we have used comprise different instances, mainly the following categories:

1. **Municipality expenses**: include a wide range of expenses and different sources of income. We are interested in the ones related to education, namely everything under post 4 of the Dutch system: education.

2. **Enrolment rates**: number of students at the three levels of education. For secondary and tertiary education, these are subdivided by **VMBO**, **HAVO** and **VWO** for secondary, and **MBO**, **HBO** and **WO** for tertiary, respectively.

3. **Children poverty rates**: the instance includes number of children living under diverse income levels.

4. **Municipality code**: includes number associated to the municipality.

How many instances are there in total (of each type, if appropriate)?

In the final dataset we have gathered the following instances:

- Number of students enrolled in the following categories:
  - Primary Education
  - Secondary Education
    - VMBO
    - HAVO
    - VWO
  - Tertiary Education
    - MBO
    - HBO
    - WO
- Number of children living in poverty
- Municipality code
- Municipality expenses on education categories
  - Public primary education
  - Educational housing
  - Education policy and student affairs

All of this data is gathered per year.

What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

In the processed dataset, each instance consists of the number of students enrolled in the different dutch education level and their respective subdivisions, the number of children living in poverty, and the different municipality expenses. The data is not related to people, but to municipalities. Thus, the data is categorized by the municipality code.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

The time frames were slightly different, hence data does not overlap mostly at the beginning of the period (from 2000 to 2010). Also, while tertiary education is divided in its different types in the CBS dataset, secondary types are aggregated as a ”secondary” category. Therefore, we needed to use the DUO dataset for that specific data.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

The instances are linked through the included municipalities codes, or mapped using the municipality common name.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set (bootstrapping)? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all instances. However, when presenting the data to the client, we selected major Dutch cities —specifically Amsterdam, Utrecht, and Eindhoven— as a basis for drawing conclusions for the research question, as smaller municipalities would not provide statistically strong enough data.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

As the data contains only aggregated data per municipality and per year, there are no recommended data splits.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Within the categories of primary education data (regular (BO) or special needs (SBO, SO, VSO)) there are many -1 values of students in each category. This indicates data that was anonymized as to not reveal the true number of students, as there are less than 4. Nevertheless, it was mostly in the special education category, which to the overall sum of primary education students is a minor percentage. Additionally, this division in primary schools does not offer any more insights in our client's question.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset we created is self-contained. Given that the datasets used to create it belong to CBS and DUO, and these organizations offer public and transparent data, we can safely assume that access to the datasets will not be interrupted in the near future. However, as municipality borders and budget guidelines change, so does the schema of the dataset making the process of adding new data as it comes in more difficult.

Any other comments?

None.

3 Collection Process

(If you did not record the data yourself) Where did you download the data from? Please elaborate on why this is an appropriate source. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)? Who funded the creation dataset?

The datasets where obtained from CBS (*Centraal Bureau voor de Statistiek*) and DUO (*Dienst Uitvoering Onderwijs*). Both sources are government-funded and authoritative data sources in the Netherlands. The first one is the national statistical office of the country, whereas the latter is an agency of the Dutch Ministry of Education, Culture, and Science. In both cases, the data is collected using well-explained methodologies and is widely trusted.

As the different datasets on their own can be used in many scientific areas, we found that the data had been used and study for diverse purposes. As an example, in [3], the authors focused on the enrolment rates of 2012 to check the effect of funding arrangements on dropout rates.

Additionally, in [1] the municipality expenses data from CBS is used for the period 2005-2007 to study the efficiency of dutch municipalities managing resources for social assistance.

(If you did record the data yourself) What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

N/A

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data associated with each instance was obtained directly from the CBS and DUO datasets. This data has been gathered by professionals in these organizations and it is directly observable.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

No dataset is a sample from a larger set. Even though we just presented the data on the poster for three cities (Amsterdam, Utrecht, and Eindhoven), we had data for most dutch municipalities.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? What does this context mean for your project?

The data was collected by the aforementioned organizations, which possess teams of professionals that gather the data in a reliable and high-quality manner. As these employees are educated for their job and have presumably no ulterior motives, it can be assumed that this does not impact this work in any meaningful way.

What was the motive behind capturing the data? What does this context mean for your project?

For the enrolment rates data, the motive was to have a historical record of the different educational options popularity in the Netherlands over time. Regarding expenses, the data is collected to offer a detailed and transparent record of how the different municipalities are expending the public resources. The children living in poverty serves the purpose of identifying and addressing socio-economic disparities. Finally, the mapping between codes and municipalities offers a systematic approach to catalogue data related to municipalities, in order to ensure consistency. In the context of our project, this means we can systematically obtained the data for each municipality and examine how the enrolment trends relates to expending and poverty levels.

Where was the data recorded? (geographic location, time, sociological context and possible others.) What does this context mean for your project?

The data was recorded in the Netherlands' municipalities. This means that the data is relevant for the question proposed by the client, since they were interested in the result of dutch policies.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

TODO

All the source datasets had different timeframes, causing gaps in data when constructing the compiled dataset. Although this could limit the client's ability to find trends over time, the final dataset was created using the largest possible timeframe available. Most data available is in the time period 2010-2020. There are also discrepancies between larger and smaller municipalities, with smaller municipalities often having less data available.

Who was the data created for? What does this context mean for your project?

The data was designed for many stakeholders, like researchers, policy makers, teachers, students and journalists. In the case of our project, it means that the data collection we have done is appropriate to our client demand and objective.

Is there a research question associated with the original dataset, and if yes, what is it? If no, what could be the reason for sharing the data? What does this context mean for your project?

The datasets made public by the CBS and DUO are provided based on open government laws, they were not created with a certain research goal in mind. This is beneficial to our work, as it can be assumed there is no conflict of interest and the data is reliable.

4 Data Transformation and Presentation

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The relevant data is gathered from sources such as the Dutch Bureau for Statistics[1] (CBS) and the Dutch Education Agency[2] (DUO).

The datasets on municipality spending were collected from CBS' StatLine, these datasets are split by year. The child poverty rates were extracted from this CBS data source.

Primary and secondary education enrollment rates are gathered from the DUO dataset portal. Tertiary education enrollment rates per municipality are retrieved from the CBS' OpenData Portal.

An utility dataset used for mapping municipality codes to their names was retrieved from OpenDataSoft.

[1] https://www.cbs.nl/

[2] https://duo.nl/particulier/

Which degree of interaction with the data was needed to prepare the data? (Discovery, Capture, Curation, Design, Creation)

- Discovery - This phase involved exploring a wide range of potential data sources such as CBS, DUO and the Dutch Ministry of Education. For each of the sources, the relevance and the quality of the dataset had to be evaluated to ensure the information aligns with our objectives.
- Capture - Relevant sources once identified and downloaded into CSV formats as available. We have used data extraction to extract specific attributes from the dataset that would help with our research objectives.
- Curation - The data obtained was free of inconsistency to a large extent, so basic data cleaning procedures of filling missing values, standardization/normalization was not required.
- Design - The basic structure of the data is to group the number of students enrolled every year (in a decade) by the region and understand how much funding a region gets and its influence.

Did you find contradictions in your data? If yes, how did you deal with them?

No contradictions have been found in any of the data sources. All values behave as one would expect.

Did you find conflicts in your data? If yes, how did you deal with them?

During the data exchange process, the main conflict rose from the discrepancies between municipality names between the different data sources. For example, the municipality of Utrecht would sometimes be called ”Utrecht (gemeente)” whereas other datasets simply contained ”Utrecht”. The same is true for municipalities with multiple common names such as ”’S Hertogenbosch” versus ”Den Bosch”.

How did you organize your data? Describe the categorization and/or classification of your data set. Feel free to include diagrams or other imagery.

By merging all the separate datasets, one full dataset was created with the year and municipality as the index and every property as a column. Total columns were added where appropriate. A example table of the schema can be found in Table 1.

Year	Municipality	Total education expenses	Impoverished children (relative)	Total enrollment	Primary	VMBO	HAVO	...
2021	Amsterdam							
	Eindhoven							
	Utrecht (gemeente)							
	...							
2022	Amsterdam							
	Eindhoven							
	Utrecht (gemeente)							
	...							

Table 1: Schema of the complete preprocessed dataset

How did you integrate the different datasets? What was the process for deciding which data to map to which data? If you wrote code for this, please link to it here.

The datasets where specifically chosen because of them containing the data per year and per municipality. Thus, after every dataset has received its own specific preprocessing treatment, all datasets could be merged on this shared index.

All relevant code is published on the author’s GitHub repository, shown in appendix A.

If you use data exchange in your project, describe your process and share your code.

The data exchange process involved transforming data from various sources such as CBS and DUO into a uniform format. This could include details such as year, number of students, region, funding (country-wide and per municipality), etc.

All relevant code is published on the author's GitHub repository, shown in appendix A.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? (i.e., to what extent does this dataset achieve answering the knowledge analytics requests of the client?) If so, how? If not, what are the limitations?

The limitations originate from dataset availability, rather than their contents. Some datasets only contain data for a limited year range, while others may have data missing for smaller municipalities.

Overall, the requested conclusion can only be substantiated for a very limited range of years, as this is the only range all datasets contain data for. Furthermore, some datasets are compiled by CBS, but provided by municipalities themselves. As CBS does not verify their correctness, some conclusions from the data must be made with caution.

Any other comments?

None.

5 Dataset Distribution

How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

Each dataset used from CBS and DUO are hosted on their respective website and can be accessed through relevant APIs. They do not have a DOI but the organization can be cited as a source when used.

The compiled data that is prepared for the client is available on the author's GitHub repository, detailed in appendix A.

An interactive dashboard for visualizing the findings is available in appendix B.

When will the dataset be released/first distributed? What license (if any) is it distributed under?

The data sources provided by DUO and OpenDataSoft are covered under the Creative Commons (CC) 0 1.0 license, with the CBS data being covered by the CC 4.0 license.

The compiled dataset for the client will be provided in the public domain using the CC0 1.0 license.

Are there any copyrights on the data?

Copyrights exists for the socio-economic dataset (CBS - Copyright (c) Central Bureau of Statistics, The Hague/Heerlen) and the expenditure for each municipality (Copyright © Gemeenten.) On the contrary, the DUO and OpenDataSoft sources are public domain data and are not subjected to copyright restrictions when properly cited.

The compiled dataset constructed by the authors will not have a copyright attached.

Are there any fees or access/export restrictions?

The datasets exists on a public domain, hence there are no access or export restrictions on the data. However, it is requested that the data is properly cited wherever used.

The compiled dataset will be released in the public domain and will not have access and export restrictions.

Any other comments?

None.

6 Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

The raw datasets, containing the enrollment rates, socio-economic conditions and the city-expenditures are managed by DUO and the CBS.

The collected dataset as well as the fully compiled dataset will be hosted on the author's GitHub repository, detailed in appendix A. Changes and/or updates will be performed as per the requirements of the client.

Will the dataset be updated? If so, how often and by whom?

The datasets containing enrollment rates and socio-economic condition will be updated by the responsible organization every year either within the same dataset or a new separate dataset. For city expenditure, the CBS guideline towards the municipalities is to provide the new reports yearly. However, it cannot be said with certainty how often the data will be updated as smaller municipalities may decide not to send any data at all, or a municipality can simply miss the CBS' set deadline causing no data to be published.

For the client's compiled dataset, any updates that are required will be managed by the authors. Such changes will be visible through the author's GitHub Repository (appendix A). Special care must be taken to ensure that newly added data conforms to the schema of the datasets already in use.

How will updates be communicated? (e.g., mailing list, GitHub)

Dataset providers DUO and the CBS have no mode of automatically communicating an update to the existing data. However, updates are expected on a set date for which any interested parties can prepare.

For the compiled data for the client, any changes can be communicated to them via GitHub as well as emails containing release notes that would contain a detailed description of the changes that have been done along with an exact changelog.

If the dataset becomes obsolete how will this be communicated?

Dataset providers DUO and the CBS have no mode of automatically communicating if a published dataset becomes obsolete. For historical purposes, it is same to assume that these datasets will be preserved so long as the organizations are able to.

Any changes to the compiled dataset will be reported through GitHub.

Is there a repository to link to any/all papers/systems that use this dataset?

The compiled data created by the authors be found on the GitHub repository in appendix A.

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

So long as new data can be augmented to use the same index as the already compiled dataset (per year, per municipality) it can be integrated easily, as the code written is modular in nature. Contributions in the form of issues and pull-requests on the author's GitHub repository (appendix A) are welcomed.

7 Legal and Ethical Considerations

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

There is no information about any ethical review process conducted by our data sources - CBS, DUO and OpenDataSoft.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

The dataset does not contain data that might be considered as confidential. In case of the socio-economic data, The data is derived from the Government's own internal data based on tax registration but no further information are shared on how their internal data is being used with respect to privacy.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why

No, the dataset does not contain potentially harmful data.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the datasets containing statistics on impoverished children and education enrollment rate relate back to people, although only aggregate data is present in the datasets.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No, The data does not identify any specific subset of the population.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No, It is not possible to identify any individual(s).

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Although the dataset does not provide information of any specific individual, It provides information about the distribution of income and wealth of persons and households in the Netherlands. This information is used to understand how the poverty rates affect the enrollment of children into schools, as per the clients request.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

No data is collected from any individual. The data was already collected and aggregated by DUO and CBS and does not violate any individual's privacy.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No information is provided about this part of the data collection process by DUO or the CBS.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

No information is provided about this part of the data collection process by DUO or the CBS.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Since the collected datasets only contain aggregated data and not on individual persons or entities, this analysis has been deemed unnecessary. Whether DUO or the CBS have performed such analyses on their datasets is unknown.

Any other comments?

None.

8 Knowledge graph description and solution

Describe the knowledge graph structure, and the arguments behind them, in detail. Provide a diagram of the graph structure as well as an excerpt of the graph with nodes and values. Provide queries that the client can use to find the answer to their question(s).

For our client, we have created a knowledge graph in Neo4j, a popular graph database management system. Access to the database is given in appendix B. Neo4j uses a specific query language called Cypher [4], which allows querying and updating the graph in a declarative way. A picture of the knowledge graph is provided in Figure 1, and a high-level diagram is shown in Figure 2. The knowledge graph is structured with three different types of nodes: City, Year, and Data. The term "Data" is generalized, in detail, it is replaced by the actual name of the data it represents (e.g., VWO, HBO, MBO, PRIMARY, etc.). Relationships between these nodes typically follow the format HAS_{DATA}, as illustrated in the diagram.

This allows us to easily identify the types of data available for each city and the specific years for which data is available. By grouping all data for a city, this structure facilitates comparisons between cities, trend analysis within or across cities, and detailed inspection of a specific city. The exact database information is summarized in the table at the end of this section.

Our client can use the following queries to get the data that they need to answer their research questions.

For the first subquestion: What trends have been observed in the distribution of educational resources across different regions and socioeconomic groups in the Netherlands?

Trends in Educational Expenses by City and Year:

```
MATCH (c:City)-[:HAS_DATA_FOR]->(y:Year)-[:HAS_EDUCATION_EXPENSES]->(d:Data)
RETURN c.name AS City, y.year AS Year, d.value AS EducationExpenses
ORDER BY c.name, y.year;
```

Trends in Impoverished Children by City and Year:

```
MATCH (c:City)-[:HAS_DATA_FOR]->(y:Year)-[:HAS_IMPOVERISHED_CHILDREN]->(d:Data)
RETURN c.name AS City, y.year AS Year, d.value AS ImpoverishedChildren
ORDER BY c.name, y.year;
```

Comparison of Resource Distribution Across Cities:

```
MATCH (c:City)-[:HAS_DATA_FOR]->(y:Year)-[:HAS_EDUCATION_EXPENSES]->(d:Data)
RETURN c.name AS City, AVG(d.value) AS AvgEducationExpenses
ORDER BY AvgEducationExpenses DESC;
```

For the second question: How have the enrollment rates in primary, secondary, and tertiary education evolved in relation to these resource distributions over different time periods?

Enrollment Rates in Primary, Secondary, and Tertiary Education by Year:

```
MATCH (y:Year)
OPTIONAL MATCH (y)-[:HAS_PRIMARY]->(p:Data)
OPTIONAL MATCH (y)-[:HAS_SECONDARY]->(s:Data)
OPTIONAL MATCH (y)-[:HAS_HBO]->(h:Data)
OPTIONAL MATCH (y)-[:HAS_WO]->(w:Data)
OPTIONAL MATCH (y)-[:HAS_MBO_TOTAL]->(m:Data)
RETURN y.year AS Year,
       SUM(p.value) AS TotalPrimary,
       SUM(s.value) AS TotalSecondary,
       SUM(h.value) AS TotalHBO,
       SUM(w.value) AS TotalWO,
       SUM(m.value) AS TotalMBO
ORDER BY y.year;
```

Enrollment Rates in Primary, Secondary, and Tertiary Education by City and Year:

```
MATCH (c:City)-[:HAS_DATA_FOR]->(y:Year)
OPTIONAL MATCH (y)-[:HAS_PRIMARY]->(p:Data)
OPTIONAL MATCH (y)-[:HAS_SECONDARY]->(s:Data)
OPTIONAL MATCH (y)-[:HAS_HBO]->(h:Data)
OPTIONAL MATCH (y)-[:HAS_WO]->(w:Data)
OPTIONAL MATCH (y)-[:HAS_MBO_TOTAL]->(m:Data)
RETURN c.name AS City,
       y.year AS Year,
       SUM(p.value) AS TotalPrimary,
       SUM(s.value) AS TotalSecondary,
       SUM(h.value) AS TotalHBO,
       SUM(w.value) AS TotalWO,
       SUM(m.value) AS TotalMBO
ORDER BY c.name, y.year;
```

Enrollment Rates and Education Expenses Correlation:

```
MATCH (c:City)-[:HAS_DATA_FOR]->(y:Year)
OPTIONAL MATCH (y)-[:HAS_PRIMARY]->(p:Data)
OPTIONAL MATCH (y)-[:HAS_SECONDARY]->(s:Data)
OPTIONAL MATCH (y)-[:HAS_HBO]->(h:Data)
OPTIONAL MATCH (y)-[:HAS_WO]->(w:Data)
OPTIONAL MATCH (y)-[:HAS_MBO_TOTAL]->(m:Data)
OPTIONAL MATCH (y)-[:HAS_EDUCATION_EXPENSES]->(e:Data)
RETURN c.name AS City,
       y.year AS Year,
       SUM(p.value) AS TotalPrimary,
       SUM(s.value) AS TotalSecondary,
       SUM(h.value) AS TotalHBO,
       SUM(w.value) AS TotalWO,
       SUM(m.value) AS TotalMBO,
       e.value AS EducationExpenses
ORDER BY c.name, y.year;
```

Impact of Socio-Economic Status on Enrollment Rates:

```
MATCH (c:City)-[:HAS_DATA_FOR]->(y:Year)
OPTIONAL MATCH (y)-[:HAS_PRIMARY]->(p:Data)
OPTIONAL MATCH (y)-[:HAS_IMPOVERISHED_CHILDREN]->(i:Data)
RETURN c.name AS City,
       y.year AS Year,
       SUM(p.value) AS TotalPrimary,
       i.value AS ImpoverishedChildren
ORDER BY c.name, y.year;
```

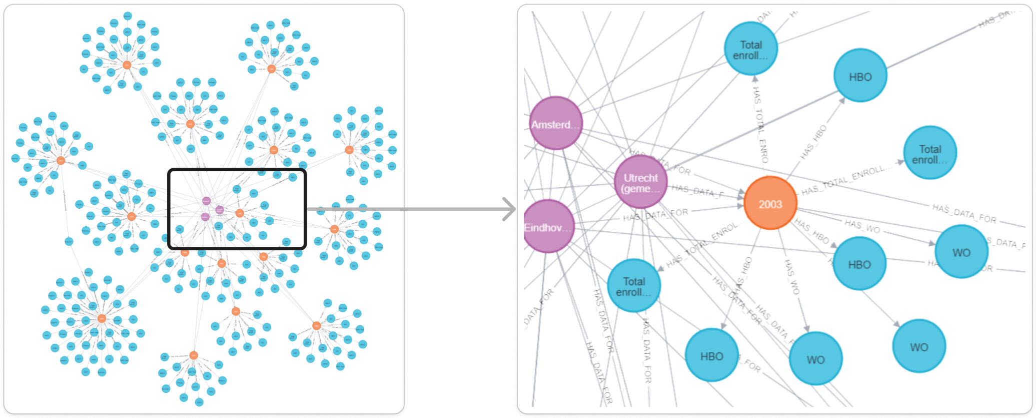

Figure 1: Knowledge Graph Overview

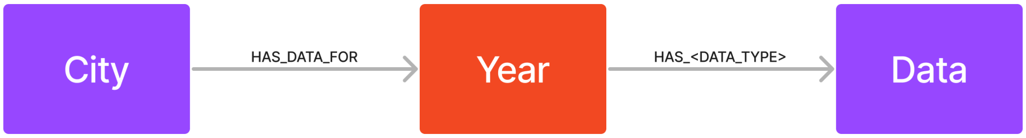

Figure 2: Knowledge Graph Diagram (High-Level)

Category	**Details**
Nodes (585)	
City	Represents different cities such as Amsterdam, Utrecht, Eindhoven
Data	Represents various data points (e.g., total enrollment, education expenses)
Year	Represents different years (e.g., 2020, 2021)
Relationships (635)	
HAS_DATA_FOR	Connects City to Year, indicating that a city has data for a specific year
HAS_EDUCATION_EXPENSES	Connects Year to Data for education expenses
HAS_HAVO	Connects Year to Data for HAVO

Category	Details
HAS_HAVO_VWO	Connects Year to Data for HAVO/VWO
HAS_HBO	Connects Year to Data for HBO
HAS_IMPOVERISHED_CHILDREN	Connects Year to Data for impoverished children (rate)
HAS_MBO_TOTAL	Connects Year to Data for total MBO
HAS_MBO1	Connects Year to Data for MBO1
HAS_MBO2	Connects Year to Data for MBO2
HAS_PRAKTIJK	Connects Year to Data for MBO Praktijk
HAS_PRIMARY	Connects Year to Data for primary education
HAS_SECONDARY	Connects Year to Data for secondary education
HAS_TOTAL_ENROLLMENT	Connects Year to Data for total enrollment
HAS_VMBO	Connects Year to Data for VMBO
HAS_VWO	Connects Year to Data for VWO
HAS_WO	Connects Year to Data for WO
Property Keys	
name	Name of the entity (e.g., city name)
type	Type of data node
value	Value associated with data points
year	Year associated with data

9 Other important aspects of the dataset and knowledge graph

Are there important aspects of the dataset/knowledge graph highlighted by the frameworks of Pushkarna et al. [6] and Paullada et al. [5], which are not covered in the sections above? If so, note them here.

None.

References

[1] Lourens Broersma, Arjen JE Edzes, and Jouke Van Dijk. Have dutch municipalities become more efficient in managing the costs of social assistance dependency? *Journal of Regional Science*, 53(2):274–291, 2013.

[2] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, 2021.

[3] Joyce Gubbels, Karien M Coppens, and Inge de Wolf. Inclusive education in the netherlands: How funding arrangements and demographic trends relate to dropout and participation rates. *International Journal of Inclusive Education*, 22(11):1137–1153, 2018.

[4] Inc. Neo4j. *Cypher Query Language*, 2024. Accessed: 2024-06-10.

[5] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

[6] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent documentation for responsible AI. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Seoul, South Korea, 2022.

Appendices

A Code repository

All code used, alongside a local copy of the exact datasets used, can be found on a public GitHub repository: github.com/Thais02/knowledge2024

B Interactive visualizations

To interactively explore the work, the authors provide EduViz, a dashboard visualizing the client's requested trends.

An interactive version of the query-able knowledge graph can be found on the Neo4j Browser, using the credentials:

- URI: REDACTED
- Username: REDACTED
- Password: REDACTED

C Client reflection on project outcomes

Dear Team, We want to express our appreciation for the thoroughness of the report. The detailed explanation of the data collection process, including the acknowledgment of dataset gaps due to availability issues, provides a comprehensive understanding of the project's scope. The diagrams and figures effectively illustrate the steps taken to transform raw data into analyzable formats, enhancing clarity and transparency. The structure of the knowledge graphs is intuitive and facilitates easy categorization, filtering, and access to essential data. These visual aids significantly enhance our ability to leverage the data for our specific research purposes. Additionally, the Cypher queries tailored to our research objectives have streamlined the retrieval of information crucial for analyzing educational resource allocation in the Netherlands. Furthermore, the interactive EduViz dashboard has proven to be an invaluable tool. Its user-friendly interface allows for seamless exploration of data trends and insights. The meticulous documentation and clear presentation of findings further underscore the reliability and applicability of the analysis to inform educational policies and decision-making processes. Thank you once again for your meticulous work and dedication to this project.