Influence of Educational Resources on Employment Outcomes across Different Socioeconomic Backgrounds

Group 10: Laurynas Jagutis, Guangyu Li, Xiaohan Wang

November 25, 2024

This datasheet was prepared following the “Datasheets for Datasets” template of Gebru et al. [9], extended where needed following the frameworks of Pushkarna et al. [14] and Paullada et al. [13].

Research Question Explore the influence of educational resources on employment outcomes across different socioeconomic backgrounds.

Client Context The client is the government that wants to research whether investing in educational resources will improve economic mobility in the country, by comparing progress in other countries.

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?)

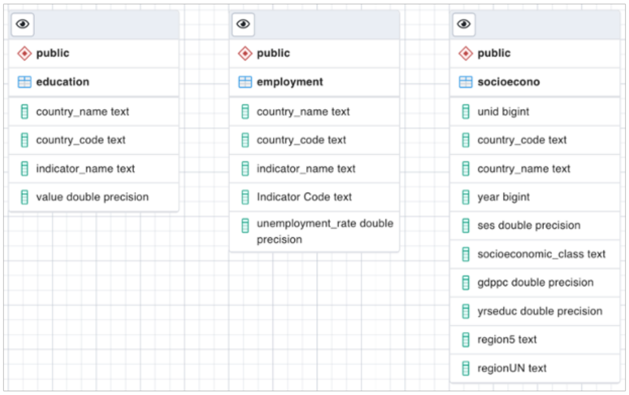
The knowledge graph (KG) of the European countries, their educational, employment indicators and socioeconomic data is created specifically for the **task** of our client to discover the possible relationship between educational resources of different countries in different socioeconomic backgrounds, and their respective employment rate.

With respect to the gap to be filled, by employing the method of knowledge graph, we become able to query the specific relationships between countries’ educational resources and employment rate, devise more complex queries that integrate country-country relationships, and discover implicit knowledge. Knowledge graphs excel at representing entities of our interest (in our case, the different countries and indicators) and their relationships. They have the potentiality for organizing the countries, relations between countries and countries, countries and indicators, indicators and indicators, and their own properties into a well-represented topological structure, on which we could apply graph theory methodology to explore, search and query to achieve our analytical goals.

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

Thanks to the extensibility of the knowledge graph, it is also beneficial for conducting **alternative** future researches on, for example, the impact of the proximity of a certain clique of countries and their employment outcome, for which it is convenient to construct and build geographical ties between countries into the KGs.

It is also worth noticing that the indicators involved in our study could be reorganized into a more hierarchical structure. Owing to the flexibility of the KG and its nature to capture structural information in graph queries, we think this dataset/knowledge graph is also suitable for studying the impact of different categories of indicators on employment rates of countries, assisting the government to put more effort on important social agendas.



public
education
country_name text
country_code text
indicator_name text
value double precision

public
employment
country_name text
country_code text
indicator_name text
Indicator Code text
unemployment_rate double precision

public
socioecono
unid bigint
country_code text
country_name text
year bigint
ses double precision
socioeconomic_class text
gdppc double precision
yrseduc double precision
region5 text
regionUN text

Figure 1: Enter Caption

Any other comments?

Though the approach of KGs are powerful to some extent, we also reflected **alternative** methodology to the problem, such as traditional data merging, regression analysis.

On the one hand, the nature of the research question is also suitable for a more traditional regression analysis with an assumption on potential correlational relationships.

On the other hand, our relational data source is characteristic of sharing `country_name` as the primary key, while lacks foreign keys. The data schema of the data sources makes it easy to join over three main tables involved, without creating many duplicates. This prevents its expressiveness after converted into a knowledge graph, since not so many complex queries could be made by utilizing the graph structure, and the dataset, in this regard, is hence more ideal for traditional relational data analysis.

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 main instances that comprise our datasets represent unique countries.

However, since we are investigating the employment statuses, socioeconomic statuses and education statistics in those countries, each of these three statistics can be considered being instances.

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

We will treat employment statuses, socioeconomic statuses, and education statistics in our countries as instances. Therefore, we have four types of instances in total.

We have three different datasets, one for each statistical category. Since these statistical instances are based on countries, the number of statistical instances depends on the number of unique countries in the datasets. The first dataset of ours, holds 149 countries and their socioeconomic statuses throughout 13 decades. Thus, there would be 1,937 socioeconomic status instances and 149 country instances. The second dataset holds 266 country instances and one unemployment rate instance per country, therefore 266 unemployment instances in total. The final dataset, holds 242 country instances and 238,225 education statistics instances.

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?

- The employment statistic instance contained raw data of unemployment percentage per country in 2023. The list of all attributes for this instance are: country name, country code, indicator name, indicator code and the year 2023.
- The socioeconomic data instance contained the raw data of socioeconomic class, GDP per capita, socioeconomic status per country over 13 decades. The list of all attributes for this instance are: unique country id, wbid - which is the country code, country itself, year, socioeconomic status (SES), socioeconomic class, GDP per capita, population average of years of education received, region of the world, and region of the world according to UN.
- The education statistic instance contained raw data of almost 4000 education indicators, such as enrollment records, graduation rates, etc. per country over 60+ years. The list of all attributes for this instance are: country name, country code, (educational) indicator name, indicator code, and 60+ years.
- The country instance itself does not have any attributes, but it is what connects the statistical instances to one another.

There is no real target associated with the instances, however for a machine learning task one could select the unemployment percentage per country, the socioeconomic class per country or one of the educational indicators as a target. Furthermore, the instances are not related to people.

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.

A lot of information is missing from the educational statistics instance. Multiple educational indexes, such as adjusted net enrolment rate, lower secondary, female (%) or Barro-Lee: Average years of primary schooling, age 25+, have no values. The data was immediately unavailable to us, but we can observe a strong correlation between the lack of data availability and the economic status of countries. Specifically, third world countries or poorer countries tend to have significantly less data on education statistics.

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.

There can be some explicit relationships between instances observed, but also there are some implicit ones. Every statistical indicator instance has an explicit relationship with the country instance, since as previously mentioned the indicators describe mostly the economic situation in specific countries. On the other hand, implicit relationships can be observed between statistical indicator instances. Because each statistical instance is associated with a specific country, their relationships are inherently linked through their connection to that country.

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 for the employment statistic instance contains all possible instances, since there can be all countries and their unemployment rates found.
- The dataset for socioeconomic status instance could be called a sample, since it does not represent every possible country in the world. The larger set would have records of socioeconomic status for every country. Nonetheless, we believe that the sample is representative of the larger set, because it includes countries from all the regions of the world.
- The dataset for education statistic instance could not be called a full dataset. It has rows and columns for all the countries and all the education instances, however a lot of data-points in those columns are missing. If the data was presented then the dataset would definitely classify as a full dataset. At the moment this dataset is not representative, due to the fact that it lacks data, especially data for the poorer countries. This dataset was the reason why we and our clients agreed on focusing on the countries in Europe, instead of the whole world.

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.

There are no recommended data splits.

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

There are no errors, sources of noise, or redundancies in our datasets.

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.

All of our raw datasets are self-contained, they do not rely on any other external sources.

3 Collection Process

For each of the datasets you have collected and integrated to answer your client's knowledge request, please answer the following questions. You can give each dataset a name and refer to it by that name.

(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?

Educational Resource Dataset Originally from World Bank EdStats All Indicator Query, the educational resource dataset contains annual coverage of data on 4000 internationally comparable indications that describe education access, progression, completion literacy, teachers, populations, expenditures, and regional learning assessments (e.g. PISA, TIMSS, PIRLS). The link to the data is `https://datacatalog.worldbank.org/search/dataset/0038480/education-statistics`

Reliability Analysis This dataset, together with the following Employment Rate Dataset is provided by World Bank's Data Log website [1], which disseminates data through World Bank sites. The World Bank Group makes this data available using a combination of licenses based on the Access to Information policy. Thus, the data source is appropriate in the way that it is open for public research and meet up with general legal and ethic requirements. Moreover, since the World Bank is one of the most renowned UN financial institution aiming to provide loans for developing countries around the world, the data maintainer is also reputable for its reliability.

Published Dataset Usage Due to the open publicity that the dataset enjoys, a number of researches already studied it, such as the work of Fiofanova and Toporkova [8] on international analysis of national databases of educational statistics for macro-regions.

Sponsors World Bank's Data Log is funded by various sources including trust-funded program, The Trust Fund for Statistical Capacity Building (TFSCB) [4], direct financing from the World Bank for data collection activities and funding from global partnerships like the Marrakech Action Plan for Statistics [3].

Employment Rate Dataset The dataset contains the latest employment rates per country. The data can be accessed at `https://tradingeconomics.com/country-list/employment-rate`

Reliability Analysis The reason why we deem this dataset to be from appropriate source is similar as the above one.

Published Dataset Usage We had a hard time finding studies with keywords that focuses solely on the employment rate, since this dataset is a part of a larger dataset, but we believe this data is referenced in other works that utilize employment rate to support their conclusions.

Sponsors Same as above.

Socioeconomic Status Dataset This dataset provides decadal socioeconomic status (SES) percentile rankings for 149 countries from 1880 to 2010, based on average income and education rankings, enabling long-term comparative analysis and integration with other global databases. If country A has an SES score of 55, for example, it indicates that 55 percent of the countries in this dataset have a lower average income and education ranking than country A.

The data is hosted currently on Kaggle at `https://www.kaggle.com/datasets/sdorius/countryses/data`

Reliability Analysis This is a data source provided by our clients, and we did integrated this data into our research as per their request. However, the data is not directly provided by a single reliable source. Rather, it is a compilation by an author Shawn Dorius (sdorius@iastate.edu) from a large number of data sources. In his sources, the data provider listed works of Barro and Lee [7], Maddison [10], Morrisson and Murtin [11], etc. As the author did not disclose his methods of producing the resulting dataset, we are skeptical about the dataset's reliability. We acted as per the client's request regardless to include it in our study.

However, since we are majorly only concerned about the socioeconomic status part of the data, which can be easily testified through other open data, the reliability issue with the dataset did not impose too much of a hardship.

Published Dataset Usage There is only one Kaggle notebook that demonstrates exploratory data analysis on the dataset with regard to GDP, Education and SES scores. The link to the notebook can be found at `https://www.kaggle.com/code/dongxu027/analysis-of-gdp-education-ses-scores`.

Sponsors The data provider did not disclose the funders, but one of his sources, the World Bank is funded by a number of sources which is discussed in the above dataset.

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.

Since our datasets include employment rates, education statistics, and socioeconomic status, these attributes are directly associated with each country. The education indicators are categorized into four groups: Government Expenditure, Gender Equality, Labor Force Education, and PISA Score, which measures performance in reading, mathematics, and science. By leveraging the hierarchical structure of education statistics indicators, we can directly link the data to their corresponding instances. Although the education statistics contain projections for future decades, we do not use this projection data in our knowledge graph, as we are not studying the potential growth trends of education indicators over time.

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

Instead of sampling, due to the significant number of missing values in the dataset, we focus on data filtering. We target European countries because they have more complete data compared to others. For sampling from the instances, we select countries equally based on their classification in the socioeconomic status dataset. For example, Austria represents a typical high (core) country, while Bulgaria represents a middle (semi-periphery) country.

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

contractors and research scientist. It does not influence on our project

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

The motivation for collecting these dataset is to study the employment rate, education background and socioeconomic status for different countries. In our project, we want to combine them to find relation between education and employment rate.

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

Countries with different socioeconomic backgrounds are the focus of our project, specifically within Europe. European countries have fewer missing data points, making it easier to extract certain statistical patterns. Additionally, countries in Europe have unique geographic features that relate to their socioeconomic backgrounds.

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

The government or research group. In our project, the client is also government which making the education policy based on these analysis.

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?

For the education indicator dataset, a possible research question might be to query a country's general education status across various dimensions. For the socioeconomic status dataset, the research question could involve comparing socioeconomic development within different countries across various regions of the world. For the employment rate dataset, the research question might focus on investigating employment rates in different countries. In our context, we will combine all these queries to investigate the relationship between education and employment rates within different socioeconomic backgrounds in Europe.

4 Data Transformation and Presentation

For the intermediate status report, you need to finish this chapter. For those questions that you have not implemented or executed yet, please write out a plan for how you'd like to approach the data transformation.

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 "raw" data files were saved in our GitHub repository. All except one of them are stored there, since the size of that one was too large to store on GitHub.
Link to access them: https://github.com/Laurynas-Jagutis/Knowledge-engineering-10

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

In terms of the degree of interaction with the data, our practice of preparation of the final dataset will fall under the view of **Design** of data to make the three datasets from different sources more tractable and analyzable. More specifically, we will start scrubbing the data before constructing a knowledge graph from it. First, the dataset of educational resources involves considerable amount of missing data, where data imputation and selection will be conducted to align the educational background information with each country's employment rate and socioeconomic status. In the sense that we are not only passively ”given” the exact amount of data to tackle the research question of the influence factors on employment outcomes of different socioeconomic backgrounds, but we are handcrafting a joined dataset particularly for our research needs, we are engaging in **designing** our data.

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

The contradiction relates several dimensions. First, the difference of time slot across different dataset may produce different of socioeconomic status. For example the time slot of the SES dataset is decades, in which omit the influence of some major event among some countries such as WWII. Due to the rapid recovery of economics, it shows a smooth increase from 1920 to 1950 for some European countries. In that case, we will choose the dataset having shorter time slot. In the other hand, in general, small countries will be more vulnerable to some external factors, which will cause the contradiction of the underlying relation between education and socioeconomic. So, we will weight the data based on the population among countries to reduce the instability.

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

For the socioeconomic status data and the education level data, there are different sources which will cause data inconsistency on some indicators. So, we will set the priority queue for the choice of data. In some cases, when the data of some attributes missing, the relevant country and UNESCO Institute for Statistics (UIS) will provide their estimations, then we deem the data from UIS is more reliable than the country estimation. In the same way, When there is a discrepancy between UIS, World Bank between the relevant country, we also choose the data of neutral institutions as they have no intention to fabricate data.

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

The research question focus on the relation between socioeconomic status with education level, in which many indications can be observed. Empirically the education level is closely related to employee rate. The most significant factors can be conclude from the analysis of data of socioeconomic status and employee rate among the countries in the world. So the dataset for the research question focus on the employee rate, socioeconomic status and education indications among countries over years.

For the socioeconomic status (SES) data, it uses a composite measure including economic, education level of a country. The score of SES of a country represents to the income and education ranking in the world.

For the education indication data, it includes the education, the literacy, the basic scientific knowledge, the network attendance rate among different age groups, from which we can possibly correlate to the SES scores.

For the employment rate data, it includes the employee rate among different categories of works.

To investigate the correlation and interaction between socioeconomic status with the education level, we will align the time and countries between the two datasets and find correlation in the normalized data over decades.

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.

For our research question the key aspect of the data is the socio-economic status. The dataset for the socio-economic status also holds information of that status in different countries during different decades. It also holds information on the location of the countries in the world (i.e Southeast Asia, East Europe, etc.). Furthermore, the employment and education datasets both include a ”country” attribute, thus we will use this attribute to connect the datasets.

The final data would have the countries as an attribute, the socioe-conomic status in those countries in 2010 since this was the most recent data in the dataset that was provided to us. The final dataset would also have the unemployment percentage in the country and adult literacy rate (%) of both sexes over 15 years old if there are not too many empty values of this attribute, since the given dataset is scarce. We think that these attributes are the main ones, required to answer the research question.

Additionally, we will examine other attributes to delve deeper into the research question. Mainly, we will concentrate on various educational statistics. Such as:

- Adjusted net enrolment rate, lower secondary, both sexes (%)
- Adjusted net enrolment rate, primary, both sexes (%)
- Adjusted net enrolment rate, upper secondary, both sexes (%)
- Adjusted net intake rate to Grade 1 of primary education, both sexes (%)

We will also analyze data on continents and regions (e.g., Southeast Asia, Eastern Europe) to determine if the impact of education on employment outcomes varies across different socio-economic backgrounds in various parts of the world.

Therefore, all the attributes mentioned above will be included in the final dataset.

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

The data exchange part of our project involves direct mapping from the relational data table. We created entity-entity property, entity-relationship tuples, and used cypher scripts to import them into the neo4j database.

Since the countries and indicators could be identified as individual entities and could be canonicalized with unique identifiers, it is convenient to create a mapping from them to the nodes and edges in the neo4j database.

The code is readily available in our GitHub repository, the link of which can be found in Section 5.

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?

We believe that this dataset processing procedure will answer the knowledge analytics requests of the client, in a way that it encompasses the required indicator information.

The dataset is well-structured and includes relevant attributes to address the research question. It allows for a comprehensive analysis of the influence of educational resources on employment outcomes across different socio-economic backgrounds and regions. The integration of socio-economic, employment, and educational data provides a robust foundation for answering the client's knowledge analytics requests.

Any other comments?

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?)

Distribution of the data Our project is publicly available on GitHub at `https://github.com/gitkeniwo/knowledge-engineering-course-project`

We have requested an issue of DOI at Zenodo [5] for the repository at `https://doi.org/10.5281/zenodo.11984028`.

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

Release of the data It is already distributed on GitHub at `https://github.com/gitkeniwo/knowledge-engineering-course-project`.

We decide to release it under MIT license [6].

Are there any copyrights on the data?

There is no copyright concern from the data source where we obtained our data from. The distribution of the data and the corresponding copyrights shall follow the MIT license.

```
Copyright (c) 2024

Permission is hereby granted, free of charge,
to any person obtaining a copy of this software
and associated documentation files (the "Software"),
to deal in the Software without restriction,
including without limitation the rights to use,
copy, modify, merge, publish, distribute, sublicense,
and/or sell copies of the Software, and to permit
persons to whom the Software is furnished to do so,
subject to the following conditions:

The above copyright notice and this permission notice
shall be included in all copies or substantial portions
of the Software.
```

Are there any fees or access/export restrictions?

There is no fees or access/export restrictions attached to the dataset.

Any other comments?

6 Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

The members of our group are supporting/hosting/maintaining the dataset.

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

There is possibility for the dataset to be updated if new annual data is included in World Bank's data source. However, since our project does not include an automatic pipeline to fetch new data from the source, the update could not be scheduled at the moment.

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

The update and newer release will be published on GitHub. It is possible to follow the news of the repository by ”watching” the repository or through other means like RSS or GitHub webhook.

If the dataset becomes obsolete how will this be communicated?

Feel free to contact anyone of the group members through the issue forms of the GitHub repository.

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

As of now, there is no repository to link to any/all papers/systems that use this dataset.

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?

The extension of the dataset is possible through GitHub forks and pull requests. There is a document `README.md` file on how to initiate a neo4j file and import the data. We also plan to develop a docker pipeline by using DOCKERFILE configuration.

For tracking/assessing the quality of those contribution, new pull requests will be examined by the maintainer (our group members).

There is currently no procedure to examine the code. To achieve this, we plan to write unit tests and integration tests for the repository, and run the test automatically through GitHub Actions [2] triggered by new commits.

Communication will be kept open in the repository's pull requests and issue forms.

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.

No, the data did not go through any ethical reviews or screenings.

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.

No, the data is made publicly open to anyone by all of its sources.

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 data is safe with sensitive and inappropriate words.

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

No, the dataset describes macro characteristics in the level of countries.

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.

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.

(Question Skipped due to Irrelevance)

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.

(Question Skipped due to Irrelevance)

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

(Question Skipped due to Irrelevance)

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.

(Question Skipped due to Irrelevance)

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.

(Question Skipped due to Irrelevance)

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).

(Question Skipped due to Irrelevance)

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.

(Question Skipped due to Irrelevance)

Any other comments?

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).

8.1 Knowledge Graph Structure and Schema

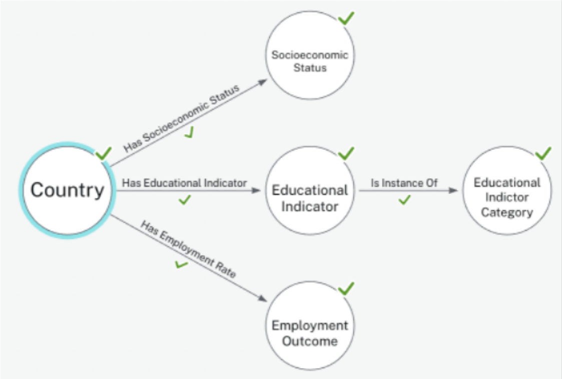


Figure 2: Graph Schema: Skeleton

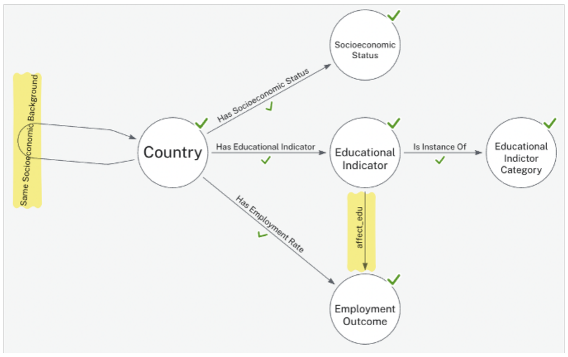


Figure 3: Graph Schema: Rounded off

The knowledge graph we have for the research question is built upon the three datasets we were given, namely educational resource, employment rate and socioeconomic status.

The Skeleton Given the essentially relationl datasets provided by our clients, and their sharing of the public keys and lack of foreign key, the structure of our knowledge graph remains simple in schema design. At the core of the knowledge graph are the Country nodes, each representing a country in our dataset. From the Country nodes branch out the indicator nodes of educational resource indicators, employment outcomes and socioeconomic status per country. To apply a categorization on the various educational indicators we covered, we grouped the indicators into four basic classes, *Government Expenditure, Gender Equality, Labor Force Education* and *PISA Score*, each of them measuring a different aspects of comprehensive assessment of a country. Upon the basic structure obtained from the relational dataset, we add an extra Educational Indicator Category node type that can be attached to educational indicator nodes. Up to now, the skeleton of the knowledge graph is presented, as is shown in 2, and it is characteristic of a number of scattered small connected components that center on the Country node at their cores. The connectivity of the knowledge graph is not satisfying to the extent that not so many insightful graph queries could be made regarding the client's request.

Properties The various property fields of the nodes of our graph determine that the graph is in nature heterogeneous. Country names are assigned to each country node, with country codes as the identifier of the nodes. For indicators, the properties are less diverse, and mainly used to keep trace of the exact numeric values of the indicators per se. For relationships, the properties are rare, but we did manage to add a special *correlates_with* relation that enjoys a property field of its exact numeric correlation coefficient. The details will be introduced later.

Constraints The graph is built with constraints that are aligned with the fundamental, low-level logic of the relational dataset we were offered. For example, country names are limited to string, and indicator values are limited to floats. Between countries and each indicator, there is only one relation to make sure each indicator is only associated with the country once. Similar constraints are also applied to relations between indicators and educational indicator categories. These constraints are of great importance to our knowledge as they maintain the **integrity** of the graph [12], making the graph more robust and less vulnerable to illegitimate CRUD operations.

Derivation of Implicit Knowledge Based on the fundamental graph structure we introduced above, it is possible to form new latent relations between the established entities in the graph, by derivation new information from the existing knowledge in the graph, as is shown in Figure 3. For example, we are now able to form **cliques** [15] of countries that share a same socioeconomic status grade with a *same_se_background* relation. Moreover, we could combine the educational indicators with the employment outcomes of countries, by adding a new *correlates_with* relation between them, populated with the exact correlation coefficient field, as is shown in Figure 8 and Figure 9. This kind of implicit knowledge inference is promising in the way that it could be utilized to make more advance queries. For example, if you further bin the correlation coefficients into categories, such as high-positive, low-negative, and then you are constructing a direct, qualitative correlational relationship between the educational and employment indicators, thus powering more graph queries that would benefit from it.

8.2 Queries for Solution of the Research Question

There are a variety of queries that could be made on the knowledge graph. Here are some examples:

- If you would like to query the related information with respect to a certain country, for example, Austria, you could query

  ```
  // Query on a single country
  MATCH (n:Country {country_name: ’Austria’}) -[r]-> (i:`Educational Indicator`)
    -[r2]-> (ic:`Educational Indictor Category`),
  (n:Country {country_name: ’Austria’})
    -[r3]-> (e:`Employment Outcome`),
  (n:Country {country_name: ’Austria’})
    -[r4]-> (s:`Socioeconomic Status`)
  return *
  ```

  And the result of the query is show in Figure 4

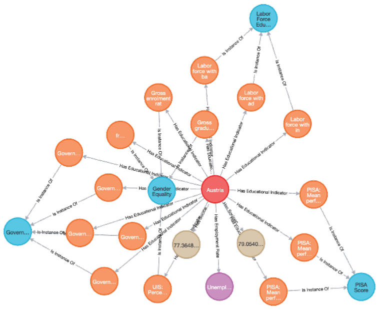


Figure 4: Query Results of Austria’s Relevant Indicators

- Another type of queries useful to make is to query on countries of similar socioeconomic backgrounds and their corresponding data. This could be done by

  ```
  // Query Countries that connect to a Socioeconomic Status
  // which has property "socioeconomic_class="Middle(semi-per)",
  // their employment outcomes and educational indicators that
  // is an istance of "Gender Equality"
  MATCH (n:Country) -[r]-> (s:`Socioeconomic Status`),
  (n:Country) -[r2]-> (e:`Employment Outcome`),
  ```

```
(n:Country) -[r3]-> (i:`Educational Indicator`)
-[r4]-> (ic:`Educational Indicator Category`
{category: "category: Gender Equality"})
WHERE s.socioeconomic_class = 'High(core)'
RETURN n, e, i, ic
```

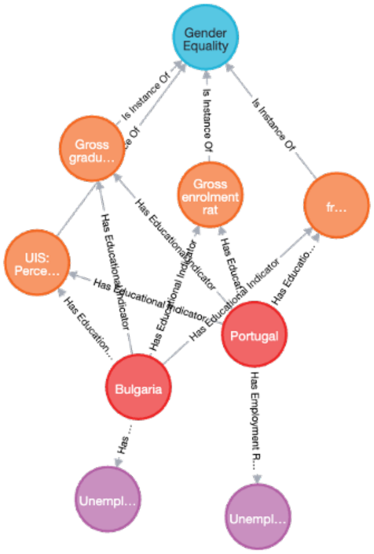


Figure 5: countries of middle socioeconomic background

Norway
Belgium
Spain
Netherlands
Austria
Denmark
Italy
Sweden
Finland
Ireland
France
United Kingd...
Greece
Germany

Figure 6: Countries of high socioeconomic background

- Also, it is possible to see all the indicator categories and their members, as is shown in Figure 7. Also, the correlation relation is displayed in Figuer 8.

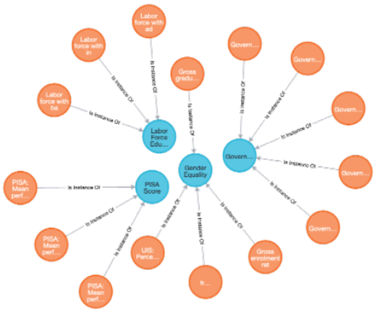


Figure 7: The indicator categories

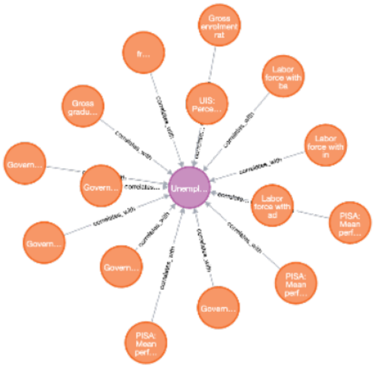
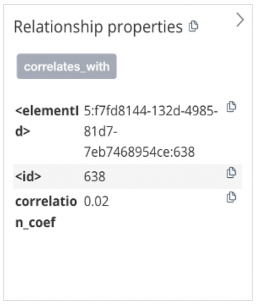


Figure 8: Correlation relations of the employment outcome with educational indicators



Relationship properties

correlates_with

<elementId>	5:f7fd8144-132d-4985-81d7-7eb7468954ce:638
<id>	638
correlation_coef	0.02

Figure 9: Example property fields of the correlation relations

- However, it is difficult to showcase the query that imputes such relations. To learn exactly how they work, please refer to our jupyter notebook at `https://github.com/gitkeniwo/knowledge-engineering-course-project/blob/main/notebooks/knowledge-graph-prepation.ipynb`.

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. [14] and Paullada et al. [13], which are not covered in the sections above? If so, note them here.

We do not have any comments for this section yet.

References

[1] Data Catalog | World Bank Group. https://datacatalog.worldbank.org/home.

[2] GitHub Actions documentation. https://docs.github.com/en/actions.

[3] Marrakech Action Plan for Statistics. https://www.worldbank.org/en/data/statistical-capacity-building/marrakech-action-plan-for-statistics.

[4] Trust Fund for Statistical Capacity Building (TFSCB). https://www.worldbank.org/en/data/statistical-capacity-building/trust-fund-for-statistical-capacity-building.

[5] Zenodo - Research. Shared. https://about.zenodo.org/.

[6] The MIT License. https://opensource.org/license/mit, October 2006.

[7] Robert J. Barro and Jong Wha Lee. A new data set of educational attainment in the world, 1950–2010. *Journal of development economics*, 104:184–198, 2013.

[8] lga leksandrovna Fiofanova and Ekaterina Sergeevna Toporkova. International Analysis of National Databases of Educational Statistics and Analysis of the Technologies' Educational Data in Countries of the World. *Journal of Advanced Pharmacy Education and Research*, 10(3-2020):90–101, 2020. ISSN 2249-3379.

[9] 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. doi: 10.1145/3458723. URL `https://doi.org/10.1145/3458723`.

[10] Angus Maddison. The west and the rest in the world economy: 1000-2030. *World Economics*, 9(4):75–99, 2008.

[11] Christian Morrisson and Fabrice Murtin. The Century of Education. *Journal of Human Capital*, 3(1):1–42, March 2009. ISSN 1932-8575, 1932-8664. doi: 10.1086/600102.

[12] Fernando Orejas, Hartmut Ehrig, and Ulrike Prange. A Logic of Graph Constraints. In José Luiz Fiadeiro and Paola Inverardi, editors, *Fundamental Approaches to Software Engineering*, pages 179–198, Berlin, Heidelberg, 2008. Springer. ISBN 978-3-540-78743-3. doi: 10.1007/978-3-540-78743-3_14.

[13] 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. doi: 10.1016/j.patter.2021.100336. URL `https://doi.org/10.1016/j.patter.2021.100336`.

[14] 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.

[15] Stephen B. Seidman and Brian L. Foster. A graph-theoretic generalization of the clique concept*. *The Journal of Mathematical Sociology*, 6(1):139–154, January 1978. ISSN 0022-250X. doi: 10.1080/0022250X.1978.9989883.

A Client reflection on project outcomes

From the client We are very pleased with the work done by the knowledge engineers. They stuck to the research question we gave them, which was about understanding how educational resources affect employment outcomes in different socio-economic backgrounds. They created a knowledge graph that includes data from European countries, focusing on education, employment, and socio-economic statuses. This setup has allowed us to look into the relationships between education and employment in detail, which was exactly what we needed.

One thing we really appreciated was how easy it was to communicate with the knowledge engineers. There was always room to discuss things with them, and they approached us quickly whenever they had questions. This made sure we were always on the same page and that any issues were resolved quickly.

The way they combined different datasets into a comprehensive knowledge graph was particularly impressive. It not only meets our current needs but also sets up a good foundation for future research.

However, one area that could be improved is the handling of missing data in the educational statistics. Although the focus on European countries was decided together to address this issue, a more detailed plan for filling these gaps or using data imputation techniques could enhance the reliability and completeness of the analysis.

Overall, we are very satisfied with the results and the process, and we look forward to continuing this productive collaboration.