London, March 2024

University of London

Template: Convert an existing Open Data resource into Linked Data and connect it to something else

Comprehending disparities in educational access and performance based on dropout data and the correlation with poverty levels across American States

Table of Figures

1. Preliminary Information

The source code, CSVs, RDFs files, and full Data Analysis of this project can be accessed on GitHub via this link.

2. Introduction

Across centuries the racial issue in the United States has been complex and has various layers that touch not only the educational system but several other systems of the American Society. Access to education and overall knowledge is considered for many thinkers one of the pillars that support a human being.

John Dewey (1859), an American educational reformer, psychologist, and philosopher, highlighted the importance of school as a tool for social progress. According to *Ho Lee (2016)*, Dewey believed that "*the school is an institution 'erected by society' which has the purpose of enhancing 'the welfare of society' and providing the society with a better future; the educational system without this 'ethical responsibility' is 'derelict and a defaulter'."*.

Poverty rates are one of the factors that have an impact on the educational level disparity, according to the *2022 Poverty Rate by Race/Ethnicity* table extracted from Kaiser Family Foundation (KFF), 9.5% of the White population in the US were living in the poverty line in 2022, against 21.4% of the Black population and 16.7% of the Hispanic:

Figure 01:

Location	White	Black	Hispanic
United States [1]	9.5%	21.4%	16.7%

Figure 1: 2022 Poverty Rate by Race/Ethnicity from KFF.

In addition, Miller-Cribbs *et al* (2007) state that compared to white students, black students in the South US are 3.8 times more susceptible to attending high-poverty schools than white students:

> "*Given these differences between children and classrooms in high– and low–ethnic minority and poverty schools, it is not surprising that, on average, test scores were lower in the high–ethnic minority and poverty schools. There is, as we would expect, a "gap" in achievement between these types of schools."* (Miller-Cribbs *at al,* 2007, p. 9)

In conclusion, it is undeniable the disparities in terms of poverty between races that persist across the United States. This economic barrier has a strong link and affects the educational level of the population.

Personal Motivation

This project serves as both exploratory and explanatory research, aiming to provide an understanding of the problem to make it explicit. This, in turn, enables the identification of factors that determine or contribute to the educational disparity.

It is clear to me that racial issues are complex, making them almost impossible to resolve simply. Nevertheless, I am also aware that is the government's responsibility to address this issue with governmental policies.

However, we must use our skills and privileged access to higher education to work towards contributing to and building an equitable society, creating ways and means by which this can be discussed and widespread.

Aims

The main aim of this project is to integrate and analyse three different data sets:

1. *State-Level Public School Dropout Data: 2002–2003 through 2004–05*, extracted from the NCES website;
2. *ACS DEMOGRAPHIC AND HOUSING ESTIMATES, 2010* - filtered by race and extracted from the United States Census Bureau website; and
3. *2010 Poverty Rate by Race/Ethnicity,* extracted from the KFF website and based on the *Historical Poverty Tables: People and Families - 1959 to 2022* - United States Census Bureau Datasets.

This analysis will uncover trends and insights into educational inequalities across racial groups in various states, a problem persisting into the present day.

The chosen datasets are from the year 2000, as per NCES website more recent data is restricted and would need a Licence Application. However, the data analysis and pipeline of this project can be applied to more recent datasets. Moreover, the main topic of this project - the disparity of educational levels and poverty between races, is still a current issue that regardless of some progress over the years is far from being resolved, as already shown in the *2022 Poverty Rate by Race/Ethnicity* from KFF.

Another objective is, by transforming the datasets into linked data, to leverage the principles to enhance the richness and accessibility of the datasets, thereby facilitating more sophisticated analyses. As well as transforming the traditional data format into RDF, contributing to the semantic web and making the data more interpretable and machine-readable.

Scope

The template used for this project is "***CM3010*** *Databases and Advanced Data Techniques: Project Title 1 - Convert an existing Open Data resource into Linked Data and connect it to something else*", and its Scope will involve:

1. **Data Preparation:** The use of *Pandas* Library to manipulate, clean and align the three datasets. This will involve the use of a consistent format of standard labels and handling inconsistent values.
2. **Data Modelling and Transformation:** The use of *Pandas* Library to prepare the data for RDF conversion and *rdflib* to model the data.
3. **Linked Data Conversion:** The use of *rdflib* to produce the RDF triples, as well as define the URIs to create entities and establish relationships across the datasets.
4. **Data Analysis:** The use of libraries like *Pandas, seaborn and matplotlib* to perform the analyses of the dataset, plotting trends and calculations.
5. **Deliverables:** Jupyter Notebook with the source code and report and the RDF Serialisation.

Ethics

1. **Data Attribution:** The sources of the three datasets have been properly attributed, and the guidelines are being followed.
2. **Data Usage Agreement:** The three datasets are public data and no licences are required. However, an agreement to not make *"any effort to determine the identity of any reported case by public-use data*" is made before downloading the NCES data. Regarding the U.S. Census Bureau datasets, according to their website: "*All U.S. Census Bureau materials, regardless of the media, are entirely in the public domain. There are no user fees, site licenses, or any special agreements etc for the public or private use, and or reuse of any census title."*
3. **Data Privacy:** This analysis is based on non-identifiable data and no attempt to identify any individual or educational institution is made by combining the datasets.
4. **Responsible Analysis:** This analysis and interpretation do not attempt to result in misleading conclusions by creating any biases or misinterpretations of the data.

3. Literature Review

Linked Data Overview

Linked Data (LD) is a concept in the realm of web technologies that aims to connect and consolidate datasets that are originally isolated within various separate databases and often overseen by different organisations. (Bizer *et al*, 2009)

The aim is to transform traditionally isolated data (often in formats like CSV) into an integrated, accessible network. This interconnectedness improves data usability and fosters collaborative research across disciplines. (Bizer et al, 2009)

This practice not only improves data accessibility but also markedly enhances opportunities for collaborative research, interdisciplinary analysis, and thorough comprehension across a range of academic and professional disciplines. In other words, the data becomes more machine-readable. (Bizer *et al*, 2009)

Berners-Lee (2006) defined a set of four rules to be followed when publishing data so that it can be interconnected on the web:

- Use URI as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
- Include links to other URIs. so that they can discover more things.

Using Uniform Resource Identifiers (URIs) to name things, a unique identifier is provided for each item in the dataset. (Berners-Lee, 2006) This information should be provided in standard formats such as the Resource Description Framework (RDF), which encodes the data using triples that consist of *subject, predicate* and *object*. (Addlesee, 2018)

Let's take for example:

Table 01: Example of a table that provides information about Name and Age

Name	Age
John Doe	23

In this case, the triples would consist of *Subject: '*John Doe'; *Predicate: '*has age'; and *object: '*23'.

For instance, the URI could be something like: *http://example.org/names/01*, and the triple*: <http://example.org/names/01> <hasAge> <23>.*

Therefore, LD represents a significant advancement in web technology, aiming to bridge the gap between isolated datasets housed in disparate databases. The fundamental

objective of LD is to transform the way data is shared on the web, moving away from traditional formats to a more interconnected and semantically rich format. (Bizer *et al*, 2009)

Linked Data Modelling

The data model used in LD is RDF, which essentially captures a depiction of two separate entities and the nature of the connection between them. Collectively, such RDF statements can construct complex graph-based structures, transcending the limitations of simple hierarchical data configurations. (Wood *et al*, 2014)

As mentioned above, RDF encodes the data and defines the components of *triples* that consist of *subject, predicate* and *object* (Addlesee, 2018) and, as seen in Wood *et al* (2014), allows to name things using both URIs and literals.

The framework elaborates on key principles, encompassing strategies for restricting literals to designated data types and associating them with particular human languages. The data model's elements, denoted by Uniform Resource Identifiers (URIs), are capable of being organised into specific classes, which facilitates improved discoverability, enhanced search functionality, and more efficient querying. (Wood *et al*, 2014)

The RDF imposes no constraints on the size of literals. Nonetheless, it is observed that numerous database systems exhibit suboptimal performance when managing large objects. Consequently, it is recommended to employ a technique known as *page holding*, particularly when handling extensive datasets. (Wood *et al*, 2014) This approach is advocated to optimise database efficiency and ensure smoother data management in scenarios involving substantial volumes of information.

Another key concept introduced by Wood *et al* (2014) is the organisation of resources into groups, known as classes and each entity within a class is called an instance. These classes are not merely organisational tools; they are RDF resources in their own right and are categorised under the type ***rdfs:Class*** within the RDF Schema (RDFS).

The RDFS provides a structure to define these classes and their relationships. One significant aspect of this structure is the use of the ***rdfs:subClassOf*** property. This property is instrumental in establishing hierarchical relationships between classes, indicating, for instance, that a certain class is a subclass, or a more specific category, of another class. (Wood *et al*, 2014)

Such a classification system and the hierarchical arrangement of classes are pivotal in RDF, especially in the realm of semantic web technologies. They serve as the backbone for organising and delineating the relationships and attributes of data, thereby enabling a more nuanced and interconnected representation of information on the semantic web. (Wood *et al*, 2014)

Best Practices for Defining the URIs for Linked Data

Although in RDF both literals and URIs are utilised for naming entities, URIs offer a higher degree of flexibility compared to literals when it comes to expanding and

interconnecting data. (Wood *et al*, 2014) This flexibility stems from the capability of URIs to have additional information appended to them, a feature not applicable to literals. Essentially, URIs provide a more dynamic and extendable approach for data representation and linkage in RDF, facilitating the enrichment and interrelation of data elements in ways that literals cannot. (Wood *et al*, 2014)

A set of guidelines for designing URIs is provided by Wood *et al* (2014, p. 30):

- *"Name things with URIs most of the time."*
- *"Use a DNS domain that you control."*
- *"Use natural keys."*
- *"Make your URIs neutral to implementation details."*
- *"Be cautious with the use of fragment identifiers"*

Domain Control

In the sphere of RDF data, establishing trust is critically linked to the use of URIs derived from one's domain. This practice, whereby publishers generate and disseminate URIs on their domains, is in line with the implied social contract of Linked Data publication. (Wood *et al*, 2014) This method not only bolsters the trust in the data provided but also enhances its perceived authenticity and reliability, with the domain acting as a testament to the source's credibility. (Wood *et al*, 2014)

Natural Keys

The application of natural keys within URIs is strongly advised. These human-readable elements within a URI succinctly encapsulate the content or theme it represents. As intuitive components of a URI, natural keys significantly aid in deciphering RDF data, particularly for developers engaging with RDF in its raw format. (Wood *et al*, 2014) The incorporation of natural keys in the construction of URIs greatly improves the immediate comprehensibility of the data, thereby rendering RDF more accessible and decipherable for its diverse array of users. (Wood *et al*, 2014)

Neutral URIs

Utilising technology-neutral URIs is also understood as a strategy for ensuring the longevity and stability of web content links, irrespective of the underlying technological infrastructure. Neutral URIs are designed to be agnostic regarding the specific technology employed to serve their associated content. This characteristic ensures that these URIs remain constant and unaffected, even in the event of changes or upgrades to the web server infrastructure. (Wood *et al*, 2014)

This practice aligns with the principles of sustainable web design and digital preservation, which according to Tucker (2022) includes the permanence and consistent retrievability of web resources and long-term access and storage of data. As explained by Fielding (2000), a good web architecture smoother migration processes and system upgrades by allowing extensibility.

Fragment Identifiers

In the sphere of Linked Data vocabularies, fragment identifiers are frequently employed due to the nature in which vocabularies are served - often as documents where the fragment is utilised to reference a specific term within that document. (Wood *et al*, 2014) However, the utility of fragment identifiers is not universally applicable, particularly in the context of data identifiers. (Wood *et al*, 2014)

When a request is made to a server using a URI with a fragment identifier, the server does not process the fragment part. This lack of awareness about the fragment could result in the server providing a broad range of information, rather than the specific data intended by the request. (Wood *et al*, 2014)

Therefore, when it comes to naming resources, the use of fragment identifiers should be approached with caution. It's advisable to employ them only when there is a clear understanding of their implications and how they influence the retrieval of information. (Wood *et al*, 2014)

Project Background Research

Linked Data Conversion

Lebo and Willians (2010) present an innovative approach to transforming Comma Separated Values (CSV) datasets into Linked Data. The authors developed a methodology that emphasizes the conversion of widely used but limited CSV formats into more interconnected and semantically rich and stable RDF triples data that supports enhancement without affecting existing applications utilizing it, ensuring that previous data remains unaltered. However, the paper does not detail the specifics of the URI scheme.

Lebo and Williams (2010) detail a dual-process approach for the transformation of CSV data into Linked Data. This approach comprises two distinct types of conversions: raw conversions and enhancement conversions. Raw conversions are responsible for the initial conversion of CSV data into a rudimentary RDF format. In contrast, enhancement conversions focus on the iterative refinement of this RDF data. This refinement process involves specifying data types, reorganising relational structures, and establishing connections with external ontologies and datasets.

Both conversion types operate under the guidance of declarative parameters articulated within RDF. These parameters serve a dual purpose: they steer the tools used for conversion and provide a historical record of the data transformation journey, from the original CSV format to the resulting Linked Data. For ease of reference and discussion in their work, the authors utilize a specific URI to detail the conversion ontology, a key component in these processes. (Lebo *et al*, 2010)

Their process involved surveying existing governmental datasets and resulted in the creation of over 5.3 billion RDF triples from 312 datasets. This project is significant in its

contribution to enhancing the accessibility and usability of government data, promoting better integration, and facilitating its use in various applications. (Lebo *et al*, 2010)

Two main challenges were identified in their approach. First, they highlight the necessity of creating a user-friendly interface to improve the accessibility of these conversion methods. Second, they advocate for the establishment of a conducive environment that actively encourages and incentivizes various decentralized communities to collaborate in enhancing and interconnecting CSV datasets. Such collaboration is deemed essential for ongoing and sustained contributions. (Lebo *et al*, 2010)

According to Paquet (2020), the project called The OCLC CONTENTdm Linked Data Pilot aims to enhance user accessibility to digital collections through an efficient search interface. This integration also capitalizes on CONTENTdm's digital preservation strengths, facilitating the conversion of collections into linked data using authoritative and local library vocabularies.

Key outcomes of this pilot include the demonstration of linked open data records by the Cleveland Public Library and the Minnesota Digital Library, exemplifying the user-end benefits of the CONTENTdm platform (Paquet, 2020). This project not only showcases the practical application of linked data in improving collection management and user experiences but also marks a significant step in the adoption of linked data technologies within the library sector. Despite its achievements, the pilot recognizes the need for further development in areas like technological infrastructure, user interface, and professional training.

Challenges such as data inconsistencies and the complexity of learning new systems have led to the initiation of OCLC’s Entity Management program (Paquet, 2020). This program aims to establish a uniform database for entity terms, addressing key issues in linked data mapping and user interface development, thereby facilitating smoother transitions of bibliographic data into the linked data framework (Paquet, 2020).

Importance of Linked Data in Social Science Analysis

In social science research, a fundamental aspect involves exploring social phenomena by examining quantitative data. Extensive and complex analyses of statistical information is an important work that the growing volume of digital data has reshaped. (Zapiko *et al*, 2011) In most cases, conducting detailed secondary analyses on diverse and widely dispersed datasets is not negotiable and a considerable amount of this data is accessible online, yet it is often presented in a wide array of formats. (Zapiko *et al*, 2011)

Therefore, there is a necessity to convert and amalgamate data from various sources into specific formats suitable for analysis. While the technical aspects of data conversion and integration are typically manageable, the time and effort invested in these processes can be burdensome. (Zapiko *et al*, 2011)

Zapilko (2014), after conducting extensive interviews with experts in social sciences, reveals the transformative impact that the integration of Linked Open Data (LOD) can have in the field of academic research. This approach is distinguished by its ability to intricate detail and nuance data, thereby meeting the rigorous demands and specific needs of academic

research. LOD is particularly adept at transforming complex datasets into formats that are not only accessible but also richly informative for researchers, providing them with the essential, detailed information necessary for conducting thorough scientific investigations.

One of the key benefits of LOD is its capability to enhance statistical data with contextually relevant information, adding depth and meaning, thus increasing the overall value of research endeavours. (Zapiko, 2014) However, this benefit is accompanied by challenges, particularly in selecting context information that aligns with the diverse range of research interests in the field, making the task of choosing universally applicable context information complex. (Zapiko, 2014)

Additionally, LOD significantly impacts the data matching process, especially in aligning code list entries for data enrichment and integration. (Zapiko, 2014) This process is made more efficient through LOD, thereby reducing the typically labour-intensive task of organising and matching structural data, such as geographical regions.

LOD also facilitates the discovery and integration of disparate datasets. Its comprehensive approach to data documentation allows for easier interconnection of different datasets and enables researchers to conduct effective comparative searches across datasets for similar or related variables. This leads to researchers having access to a wealth of information that is both thoroughly documented and supplemented with relevant literature and contextual insights. (Zapiko, 2014)

In conclusion, the field of social science research, which often deals with the challenge of managing diverse and complex datasets, is set to benefit substantially from the integration of Linked Data principles. This integration is expected to enhance data accessibility, improve documentation quality, and streamline the data integration process, thus significantly enhancing the effectiveness and efficiency of research methodologies in the social sciences. Researchers are thereby empowered to conduct more in-depth and precise investigations. (Zapiko, 2014)

In an educational data context, substantial prospects emerge for crafting novel solutions. According to Pereira (2017), this can be achieved through a detailed analysis of the prevailing educational landscape, thereby facilitating more informed and targeted decision-making processes addressing the identified challenges.

Alternatively, these opportunities may manifest in the creation of applications designed to harness the available data and resources, to enhance the overall quality of education. This approach emphasises the importance of data availability and underscores the potential for its utilisation in driving educational advancements and innovations. (Pereira, 2017)

4. Project Design

Project Overview

This exploratory and explanatory research is centred on a detailed analysis aimed at understanding racial disparities in educational access, particularly focusing on dropout rates and their association with poverty levels.

The primary goal of this research is to convert the following key datasets into Linked Data, thereby enabling the linkage and comprehensive analysis:

- **State-level public School Dropout Data (2002–2003 through 2004–05)** from the National Center for Education Statistics, which provides insights into dropout rates across different states.
- **ACS Demographic and Housing Estimates (2010)**, filtered by race, from the United States Census Bureau, provides valuable demographic information pertinent to the study.
- **2010 Poverty Rate by Race/Ethnicity** extracted from the Kaiser Family Foundation, based on the Historical Poverty Tables from the United States Census Bureau, provides a comprehensive view of poverty rates among different racial groups.

By integrating these datasets into a Linked Data structure, the research aims to conduct a thorough analysis that illuminates the complexities of the educational gap as influenced by race and socioeconomic factors.

Therefore, the project follows this scope:

1. **Data Preparation:** Involves cleaning and aligning the data for the analysis. This phase includes checking for any inconsistent values and formatting the labels to create a standard.
2. **Data Modelling and Transformation:** Prepare and model the data for RDF conversion.
3. **Linked Data Conversion:** Convert the data into RDF triples, defining the URIs and establishing the connections between the datasets.
4. **Data Analysis:** Perform the data analysis by combining the datasets to extract insights about the educational gap and define the correlation with the poverty levels.

Domain and Users

The **Domain** of the project is Educational Research, more specifically focused on educational outcomes and socioeconomic factors involved.

The findings from this research hold significant value for a diverse range of **users/stakeholders:**

- **Educational Data Scientists:** Experts in this field who are keen on applying advanced data analysis techniques to educational data can find the project's methodologies and insights particularly useful.
- **Organisations:** Both government and Non-governmental Organisations could use this data to work towards policies and programs aiming to address the educational gap and poverty.
- **Social and Educational Researchers:** Professionals who are exploring the intersections of education and poverty can find the research beneficial for deepening their understanding of these areas.

Considering these users, the selected features of this project are designed to ensure a precise and consistent data analysis, establishing it as a trustworthy resource in the domains of educational and social research. The **methodologies** include:

- **Using Pandas to prepare the Datasets:** Ensures accuracy and consistency in data, which is crucial for reliable analysis in educational research.
- **Data Modeling, RDF Conversion and Serialisation:** The use of RDF and linked data principles allows for the integration of disparate data sources, enabling a more holistic analysis which is essential in understanding complex social phenomena like education and poverty.
- **Analysis Tools:** The selection of statistical and data analysis libraries (*Pandas*, *seaborn* and *matplotlib*) is aligned with the need for analytical methods in educational research to derive meaningful insights from datasets.
- **Jupyter Notebook Deliverable:** Provides an interactive and comprehensive report that is essential for academic presentation and facilitates replication and further exploration by other researchers in the field.

In essence, the project leverages cutting-edge data science tools and methodologies to offer insightful contributions to educational and social research.

Project Structure

1. **Background and Introduction**
   This section articulates the project's objectives and the rationale behind the analyses, providing an in-depth background on the educational gap and its correlations with poverty.

2. **Project Overview with Datasets Description**
   A thorough description of each dataset, including their origins. This segment also elucidates how these datasets can be synergistically combined for comprehensive analysis.

3. **Methodology**
   As previously outlined, the methodology encompasses the data preparation, modelling, RDF conversion and serialisation, as well as the Linked Data creation and integration.

4. **Data Analysis**

- **Analytical Techniques:** Use of libraries cited above in the Analysis tools section for data analysis.
- **Trend Identification:** This involves identifying trends in dropout rates and poverty levels and using statistical methods to quantify the relationship between these variables.
- **Comparative Analysis:** The project conducts comparative analyses across different states, racial demographics, and periods.
- **Data Visualisation:** Various graphical representations, including scatter plots, bar charts, and area diagrams, are employed for enhanced data interpretation and visualisation.

5. **Project Evaluation**
   The Project is evaluated based on the data quality and consistency, the success of the RDF conversion and the Linked Data Integration, the analytical accuracy and the relevance and impact of the insights.

   - **Data Quality Assessment:**
     The goal is to assess the accuracy, completeness, and consistency of the data after the cleaning and standardisation. To achieve that spot checks through data profiling and summary statistics analyses to ensure data integrity are performed.

   - **RDF Conversion Evaluation:**
     The objective is to make sure that all RDF triples and URI formations are correct. This is done by verifying the RDF triples for accuracy in representing the data relationships and validating URIs for correct referencing.

   - **Linked Data Integration Success:**
     The degree of effective linkage and integration between datasets is measured to ensure the success of the Data Integration. Test queries are plotted across the linked datasets to assess the integration and retrieve intended data successfully.

   - **Overall Project Success:**
     The project is reviewed against its initial goals to evaluate if they have been achieved. Any future potential for further research will also be evaluated.

6. **Conclusion**
   The conclusion recapitulates the project's goals and summarises its key findings. It also proposes areas for future research, considering the project's scope and limitations.

Project Key Technologies

- **Pandas Library (Python):** Pandas is used to prepare the data, check for any missing values and standardisation of any labels.

- **RDF (Resource Description Framework):** The data is converted into RDF, as it is the standard model for data interchange on the web and enhances linking and sharing across different systems.
- **rdflib (Python Library):** rdflib is used to convert datasets into RDF, define URIs, and create triples to establish relationships between data entities.
- **Matplotlib, and seaborn (Python Libraries):** These two Python libraries are used to conduct the data analysis by analysing trends, and correlations, and performing statistical tests on the datasets.
- **Jupyter Notebook:** Jupyter Notebook is used to document the entire data analysis process, including code, results, and visualisations.

5. Implementation

The source code, CSVs, RDFs files, and full Data Analysis of this project can be accessed on GitHub via this link.

This endeavour aligns with CM3010's Template 1, focusing on the transformative potential of Linked Data in comprehensively analyzing pivotal educational datasets. This project's core ambition is to deepen the understanding of educational disparities, particularly in school dropout rates, demographic variations, and poverty levels within the United States.

The implementation encompasses initial data cleaning and alignment, laying a solid groundwork for advanced analysis, followed by RDF conversations and extraction of the linked data for a Data Analysis.

Features

Preliminary Data Preparation

The initial phase, data preparation, involves cleaning and alignment of datasets. This stage is pivotal in converting raw data into a structured format suitable for analysis and RDF transformation. The process leverages Python's Pandas library, noted for its data manipulation prowess.

Data Cleaning and Transformation Techniques

1. **Loading Data:** The initial step involves loading data from different formats (Excel and CSV) into Pandas DataFrames. This is achieved using the pd.read_excel() and pd.read_csv() functions:

```
# Loading the National Center for Education Statistics (NCES) dropout data for the year 2002 into a Pandas DataFrame.
NCS2002_data = pd.read_excel('Original_Datasets/NCES/2002/2002.xls')
# Loading the NCES dropout data for the year 2003 into a Pandas DataFrame.
NCS2003_data = pd.read_excel('Original_Datasets/NCES/2003/2003.xls')
# Loading the NCES dropout data for the year 2004 into a Pandas DataFrame.
NCS2004_data = pd.read_excel('Original_Datasets/NCES/2004/2004.xls')
# Loading the 2010 Census data into a Pandas DataFrame from a CSV file.
census_data = pd.read_csv('Original_Datasets/Census/Census 2010.csv')
```

   For the poverty dataset, a try-except block is used to handle potential parsing errors due to different data formats or unexpected headers:

```
try:
                porvertyKFF_data    =    pd.read_csv('Original_Datasets/KFF Porverty/2010-porverty.csv', skiprows=2)
# If a parsing error occurs, reading the file again with a different delimiter and skipping the first two rows.
```

```
except pd.errors.ParserError:
                porvertyKFF_data    =    pd.read_csv('Original_Datasets/KFF
Porverty/2010-porverty.csv', delimiter=';', skiprows=2)
```

2. **Data Cleaning:** The NCES datasets contain placeholder values (-1) representing missing or undefined data. These are replaced with 0 to standardize and allow aggregation of the gender data without subtracting "-1", once the aggregation was done the values (0) were replaced with NaN using the same technique. Example:

```
# Replacing all occurrences of -1 with 0 in the NCS2002_data variable
NCS2002_data = NCS2002_data.replace(-1, 0)
# Replacing all occurrences of -1 with 0 in the NCS2003_data variable
NCS2003_data = NCS2003_data.replace(-1, 0)
# Replacing all occurrences of -1 with 0 in the NCS2004_data variable
NCS2004_data = NCS2004_data.replace(-1, 0)
```

3. **Data Transformation:**
   - Aggregation of Data: For the school dropout datasets (NCES datasets), the data is segregated by gender and unknown gender (M, F, U) for each grade. The data is aggregated to create a combined count for each grade irrespective of gender. Example:

     ```
     NCS2002_data[f'BL{i}']=NCS2002_data[f'BL0{i}M']+
     NCS2002_data[f'BL0{i}F'] + NCS2002_data[f'BL0{i}U']
     ```

     Creating a Structured DataFrame: Columns relevant to the analysis are selected and renamed for clarity and consistency. Example:

     ```
     NCS2002_table = NCS2002_data[['YEAR', 'STNAME']]
     NCS2002_table.columns = ['Year', 'State']
     ```

4. Filtering Data Based on Specific Criteria: For the census data, the dataset is filtered to include only rows of interest (specific races) starting from a particular index. Example:

```
race_section_index = census_data[census_data['Label (Grouping)'] == 'Race
alone or in combination with one or more other races'].index[0]

filtered_census_data = census_data.iloc[race_section_index:]
```

5. Data Formatting: Numeric data containing commas (as in the census dataset) is converted to float for accurate numerical analysis:

```
for col in census_table.columns[1:]:
    # Removing commas and converting to float
```

```
                census_table[col]    =    census_table[col].str.replace(',',
'').astype(float)
```

6. **Standardization of the Data:** A final standardising of the data is completed, by renaming the race/ethnicities to 'Black', 'Hispanic or Latino':'Hispanic' across all data sets, and transforming the various ethnic columns in the povertyKFF data and census data into only one column called “Ethnicity” and ‘Black’, ‘White’ and ‘Hispanic’ became values of that column:

	State	Ethnicity	Census
0	alabama	White	3382838
1	alaska	White	526642
2	arizona	White	5242273
3	arkansas	White	2336002
4	california	White	24611291

Figure 2: Result of Census table extracted from the rdf file

	State	Ethnicity	Poverty
0	alabama	White	414000
1	alaska	White	36200
2	arizona	White	407300
3	arkansas	White	312100
4	california	White	1420900

Figure 3: Result of Poverty table extracted from the rdf file

The process is repeated for the NCES data, transforming the various grades/Ethnicities columns into one grade and ethnicity column, and the rest in values:

	Year	State	Ethnicity	Grade	Enrollment	Dropout
0	2002	alabama	Black	7	22817	18
1	2002	alaska	Black	7	532	5
2	2002	arizona	Black	7	3612	187
3	2002	arkansas	Black	7	0	0
4	2002	california	Black	7	43647	63

Figure 4: Result of NCES table extracted from the rdf file

Data Modeling and Transformation and Linked Data Conversion

1. **Linked Data Conversion**

After cleaning the data, the census and poverty tables had the same structure, therefore I have done the conversion with the same codes, just saving them into different RDFs files.

First, namespace and graph are created for both census and poverty data:

```
ex = Namespace("http://educationaldata.org/")
census_graph = Graph()
poverty_graph = Graph()
```

The second step was to define the URIs and the triples, converting the table into triples and saving the triples into the RDFs files.

```
for idx, row in census_table.iterrows():
    state_uri = ex[row['State'].replace(' ', '_')]
    ethnicity_uri = ex[row['Ethnicity'].replace(' ', '_')]
    census_graph.add((state_uri, RDF.type, ex.State))
    census_graph.add((state_uri, ex.Census, Literal(row['Census'])))
    census_graph.add((state_uri, ex.hasEthnicityData, ethnicity_uri))
for idx, row in porvertyKFF_table.iterrows():
    state_uri = ex[row['State'].replace(' ', '_')]
    ethnicity_uri = ex[row['Ethnicity'].replace(' ', '_')]
    poverty_graph.add((state_uri, RDF.type, ex.State))
    poverty_graph.add((state_uri, ex.Poverty, Literal(row['Poverty'])))
    poverty_graph.add((state_uri, ex.hasEthnicityData, ethnicity_uri))
# saving the triples into rdf files
census_graph.serialize(destination='census.rdf', format='xml')
poverty_graph.serialize(destination='poverty.rdf', format='xml')
```

To create the URIs all spaces have been replaced with underscores. The triples followed the same logic for both files.

The NCES data code is similar, however, the triples were a bit different as the table had Year, grade, enrollment and dropout data:

```
for idx, row in nces_table.iterrows():
    # creatin URIs based on: grades, states and ethnicities
    record_uri = ex[f"{row['State'].replace(' ',
'_')}-{row['Ethnicity'].replace(' ',
'_')}-{row['Year']}-{row['Grade']}"]
    state_uri = ex[row['State'].replace(' ', '_')]
    ethnicity_uri = ex[row['Ethnicity'].replace(' ', '_')]
    # Adding the triples
    nces_graph.add((record_uri, RDF.type, ex.Record))
    nces_graph.add((record_uri, ex.hasState, state_uri))
    nces_graph.add((record_uri, ex.hasEthnicity, ethnicity_uri))
    nces_graph.add((record_uri, ex.Year, Literal(row['Year'],
datatype=XSD.integer)))
    nces_graph.add((record_uri, ex.Grade, Literal(row['Grade'],
datatype=XSD.integer)))
    nces_graph.add((record_uri, ex.Enrollment,
Literal(row['Enrollment'], datatype=XSD.integer)))
    nces_graph.add((record_uri, ex.Dropout, Literal(row['Dropout'],
datatype=XSD.float)))
```

2. Linking the Data

To link the data a new graph was created for the linked_data and looped through all previous RDF files saving the data into the linked_data graph, for example:

```
# looping through the NCES rdf, adding the data into the new rdf linked
data file
for record_uri in nces_graph.subjects(RDF.type, ex.Record):
    state_uri = nces_graph.value(subject=record_uri,
predicate=ex.hasState)
    if not state_uri:
        continue
    enrollment = nces_graph.value(subject=record_uri,
predicate=ex.Enrollment)
    dropout = nces_graph.value(subject=record_uri,
predicate=ex.Dropout)
    year = nces_graph.value(subject=record_uri, predicate=ex.Year)
    grade = nces_graph.value(subject=record_uri, predicate=ex.Grade)

    if enrollment:
        linked_data_graph.add((state_uri, ex['Enrollment_' + str(year)
+ '_' + str(grade)], Literal(enrollment, datatype=XSD.integer)))
    if dropout:
        linked_data_graph.add((state_uri, ex['Dropout_' + str(year) +
'_' + str(grade)], Literal(dropout, datatype=XSD.float)))
```

The linkage across the datasets was facilitated through the commonality of State identifiers, serving as the pivotal point of connection among them. The following presents the outcomes derived from the linked data, as defined by its corresponding ontology file:

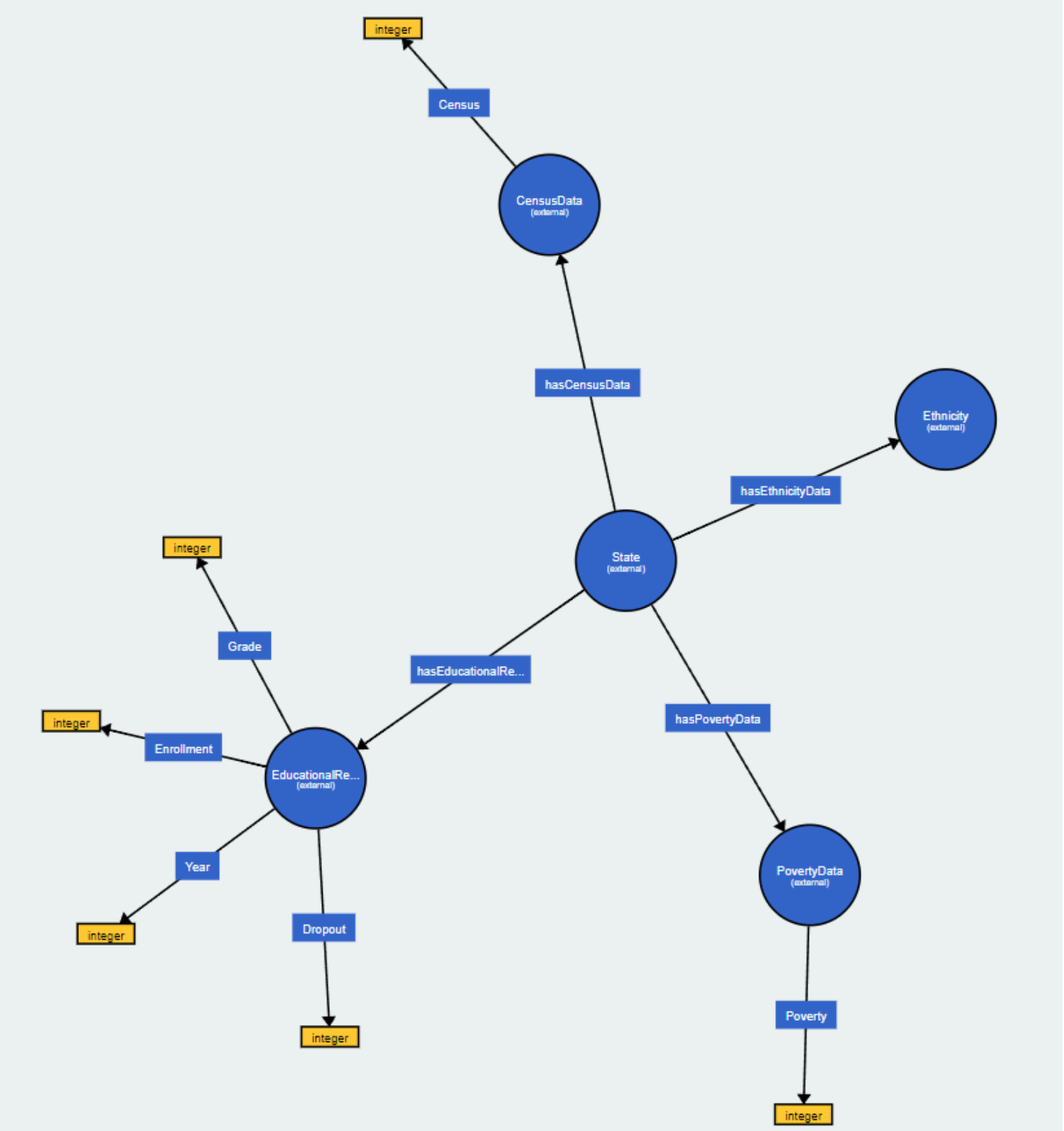

Figure 5: Ontology Graph of the linked_data_detailed.rdf file

Extracting the Data from the RDFs files for Further Analysis

To be able to perform the trend analysis, the data from the RDFs files was extracted and saved into a new table. The new tables have the same name as the original ones, but with a '2' on the end:

Table 2: Original Table vs Table Extracted from the RDF names

Original Table	Table Extracted from the RDF
census_table	census_table_2
porvertyKFF_table	povertyKFF_table2
nces_table	nces_table2

This approach also helped to evaluate the conversions by comparing both tables.

To accomplish this, a new graph was instantiated to load the RDF files, and the data was iterated through, capturing relevant information into variables before appending it to a list. The code snippet below demonstrates how data was extracted from the nces.rdf file:

```
# loopling through the graph
      for record in g.subjects(RDF.type, ex.Record):
          #  saving records of the state
          state = g.value(subject=record, predicate=ex.hasState)
          if state:  # spliting url to get only the name of the state
              state = state.split('/')[-1].replace('_', ' ')
          #  saving records of the ethnicity
          ethnicity = g.value(subject=record, predicate=ex.hasEthnicity)
          if ethnicity:  # spliting URI
              ethnicity = ethnicity.split('/')[-1].replace('_', ' ')
          #  saving records of the year
          year = g.value(subject=record, predicate=ex.Year)
          #  saving records of the grade
          grade = g.value(subject=record, predicate=ex.Grade)
          #  saving records of the enrollment
          enrollment = g.value(subject=record, predicate=ex.Enrollment)
          #  saving records of the dropout
          dropout = g.value(subject=record, predicate=ex.Dropout)
          # adding data to the list
          data.append({
              "Year": year.toPython() if year else None,
              "State": state,
              "Ethnicity": ethnicity,
              "Grade": grade.toPython() if grade else None,
              "Enrollment": enrollment.toPython() if enrollment else None,
```

```
            "Dropout": dropout.toPython() if dropout else None,})
```

Data Analysis

The data analysis was conducted using the tables extracted from the RDF files. To simplify the analysis as the poverty and census tables have the same structure, both tables were combined:

```
census_poverty_table2 = census_table_2.merge(povertyKFF_table2,
on=['State','Ethnicity'])
census_poverty_table2.head()
print(census_poverty_table2.to_string())
census_poverty_table2.to_csv('census_poverty2.csv', index=False)
census_poverty_table2['Census'] =
census_poverty_table2['Census'].astype(int)
census_poverty_table2['Poverty'] =
census_poverty_table2['Poverty'].astype(int)
```

Result:

	State	Ethnicity	Census	Poverty
0	vermont	White	607394	64200

Figure 6: Census and Poverty Table merged for Data Analysis

The proportions and percentages of the population and the population in the poverty line were also calculated to avoid any misleading conclusion. For example:

```
# Calculating the total proportion per ethnicity
population_ethnicity =
census_poverty_table2.groupby('Ethnicity')['Census'].sum()
# Calculating the total population in poverty per ethnicity
poverty_ethnicity =
census_poverty_table2.groupby('Ethnicity')['Poverty'].sum()
# Calculating the total population in poverty in relation to all
ethnicities
poverty_total = poverty_ethnicity.sum()
# Calcular a proporção de cada etnia em relação à população total
porportion_pop_ethn = (population_ethnicity /
population_ethnicity.sum()) * 100
# Calculating the proportion of each ethnicity in relation to the
total population
proportion_pov_ethn= (poverty_ethnicity / poverty_total) * 100
# Table to save the results
```

```
proportion_table = pd.DataFrame({'Population(%)':
porportion_pop_ethn, 'Population_In_Poverty(%)':
proportion_pov_ethn})

# Reseting indexes to make the 'Ethnicity' column regular
proportion_table.reset_index(inplace=True)
# Sorting the table by the 'Population(%)' column
proportion_table_sorted =
proportion_table.sort_values(by='Population(%)', ascending=False)
```

Also, *matplot* and *seaborn* libraries were used to plot all graphs used in the analysis. The results of the analysis are shared in the conclusion section.

6. Evaluation

Data Quality Assessment:

The data quality assessment was conducted to evaluate the accuracy, completeness, and consistency of the datasets. To streamline this process, a bespoke function, designated as ***data_evaluation***, was developed. This function accepts a dataset as input, executes the requisite assessments, and subsequently exhibits the outcomes.

```
def data_evaluation(table):

    # Displaying first rowns of the census table to unsdertand the new structure
    table.head()

    #Checking for any missing values
    missing_values = table.isnull().sum()

    # Data types
    data_types = table.dtypes

    # Check unique values for Unique Values
    unique_states = table['State'].unique()
    unique_ethnicities = table['Ethnicity'].unique()

     #Check for Duplicated Values
    table_duplicates = table.duplicated(subset=['State', 'Ethnicity'], keep=False).sum()

    #creating ti check the columns numbers of the table, if it is bigger then 3 it means that table is the
    #NCES one and chacks need to be performed in the Grade column

    if len(table.columns) > 3:
        nces_unique_grades = table['Grade'].unique()
        nces_duplicates = table.duplicated(subset=['State', 'Ethnicity', 'Year', 'Grade'], keep=False).sum()
```

Figure 7: Code snippet of the data_evaluation function

The ensuing findings per dataset are delineated as follows:

- **Census Table:**

```
Missing Values:
State        0
Ethnicity    0
Census       0
dtype: int64

data_types:
State        object
Ethnicity    object
Census        int32
dtype: object

Unique States Values:
['alabama' 'alaska' 'arizona' 'arkansas' 'california' 'colorado'
 'connecticut' 'delaware' 'districtofcolumbia' 'florida' 'georgia'
 'hawaii' 'idaho' 'illinois' 'indiana' 'iowa' 'kansas' 'kentucky'
 'louisiana' 'maine' 'maryland' 'massachusetts' 'michigan' 'minnesota'
 'mississippi' 'missouri' 'montana' 'nebraska' 'nevada' 'newhampshire'
 'newjersey' 'newmexico' 'newyork' 'northcarolina' 'northdakota' 'ohio'
 'oklahoma' 'oregon' 'pennsylvania' 'rhodeisland' 'southcarolina'
 'southdakota' 'tennessee' 'texas' 'utah' 'vermont' 'virginia'
 'washington' 'westvirginia' 'wisconsin' 'wyoming' 'puertorico']

Unique Ethnicities Values:
['White' 'Black' 'Hispanic']

Duplicated Values
0
```

Figure 8: Print Results of the Census Table Evaluation

- **PovertyKFF Table:**

```
Missing Values:
State        0
Ethnicity    0
Poverty      0
dtype: int64

data_types:
State        object
Ethnicity    object
Poverty       int32
dtype: object

Unique States Values:
['alabama' 'alaska' 'arizona' 'arkansas' 'california' 'colorado'
 'connecticut' 'delaware' 'districtofcolumbia' 'florida' 'georgia'
 'hawaii' 'idaho' 'illinois' 'indiana' 'iowa' 'kansas' 'kentucky'
 'louisiana' 'maine' 'maryland' 'massachusetts' 'michigan' 'minnesota'
 'mississippi' 'missouri' 'montana' 'nebraska' 'nevada' 'newhampshire'
 'newjersey' 'newmexico' 'newyork' 'northcarolina' 'northdakota' 'ohio'
 'oklahoma' 'oregon' 'pennsylvania' 'rhodeisland' 'southcarolina'
 'southdakota' 'tennessee' 'texas' 'utah' 'vermont' 'virginia'
 'washington' 'westvirginia' 'wisconsin' 'wyoming' 'puertorico']

Unique Ethnicities Values:
['White' 'Black' 'Hispanic']

Duplicated Values
0
```

Figure 9: Print Results of the Poverty KFF table Evaluation

- **NCES Table:**

```
Missing Values:
Year          0
State         0
Ethnicity     0
Grade         0
Enrollment    0
Dropout       0
dtype: int64

data_types:
Year           int64
State         object
Ethnicity     object
Grade         object
Enrollment     int32
Dropout        int32
dtype: object

Unique States Values:
['alabama' 'alaska' 'arizona' 'arkansas' 'california' 'colorado'
 'connecticut' 'delaware' 'district of columbia' 'florida' 'georgia'
 'hawaii' 'idaho' 'illinois' 'indiana' 'iowa' 'kansas' 'kentucky'
 'louisiana' 'maine' 'maryland' 'massachusetts' 'michigan' 'minnesota'
 'mississippi' 'missouri' 'montana' 'nebraska' 'nevada' 'new hampshire'
 'new jersey' 'new mexico' 'new york' 'north carolina' 'north dakota'
 'ohio' 'oklahoma' 'oregon' 'pennsylvania' 'rhode island' 'south carolina'
 'south dakota' 'tennessee' 'texas' 'utah' 'vermont' 'virginia'
 'washington' 'west virginia' 'wisconsin' 'wyoming' 'american samoa'
 'guam' 'northern marianas' 'puerto rico' 'virgin islands']


Unique Ethnicities Values:
['Black' 'White' 'Hispanic']

Unique Grades Values:
['7' '8' '9' '10' '11' '12']

Duplicated Values
0
```

Figure 10: Print Results of the NCES table Evaluation

Data Evaluation Interpretation:

The comprehensive data evaluation conducted across the three datasets—census, NCES, and povertyKFF— revealed insights into their quality and structure, ensuring their suitability for further analysis. The census dataset is characterized by its enumeration of populations across various states and ethnicities. It demonstrates a high level of completeness, as there are no missing or duplicate values.

Furthermore, the allocation of data types is found to be aligned with the nature of the dataset's content; with 'State' and 'Ethnicity' being categorized as strings, and 'Census' delineated as an int64 type. This classification is deemed apt for facilitating the requisite computations and analyses in subsequent stages. Moreover, the dataset demonstrates consistency by encompassing all U.S. states and focusing on 'White', 'Hispanic', and 'Black' ethnic groups, satisfying the specific requirements of the project.

Similarly, the NCES dataset offers insights into educational metrics such as enrollments and dropouts across different states and ethnic groups. It showcased no duplicated values, and the inclusion of the year and grade levels in the duplication check ensured an understanding of the educational landscape. Like the census data, it maintained consistency in the representation of states and ethnicities, extending the analysis with educational data without compromising integrity or coherence.

The povertyKFF dataset, focusing on poverty counts among different ethnic groups across states, mirrored the strengths of the previously mentioned datasets in terms of data completeness, accuracy, and consistency.

It is also possible to identify the *state* and *ethnicity* categories across all three datasets, due to the approach of data standardization during the preparation of the datasets, which facilitates the integration of the three datasets for comprehensive socio-economic and educational analyses.

RDF Conversion Evaluation

The initial phase involved a direct examination of RDF triples. While RDF triples are inherently straightforward, the Jupyter Notebook environment proved the task challenging due to display constraints.

To circumvent this limitation and enhance the clarity of our evaluation, a comparative analysis was conducted. This involved contrasting the original data tables, *census_table, porvertyKFF_table, and nces_table,* with their respective RDF-converted counterparts, *census_table_2, porvertyKFF_table2, and nces_table_2*:

census_table_2

	State	Ethnicity	Census
82	alabama	Black	1290667
83	alabama	Hispanic	182795
81	alabama	White	3382838
22	texas	Black	3159265
23	texas	Hispanic	9533880
21	texas	White	19191255

Figure 11: census_table_2 visualisation

census_table

	State	Ethnicity	Census
52	alabama	Black	1290667
104	alabama	Hispanic	182795
0	alabama	White	3382838
95	texas	Black	3159265
147	texas	Hispanic	9533880
43	texas	White	19191255

Figure 12: census_table visualisation

porvertyKFF_table2

	State	Ethnicity	Poverty
99	alabama	Black	395400
100	alabama	Hispanic	53400
98	alabama	White	414000
13	texas	Black	680700
14	texas	Hispanic	2475100
12	texas	White	1039400

Figure 13: porvertyKFF_table2 visualisation

porvertyKFF_table

	State	Ethnicity	Poverty
52	alabama	Black	395400
104	alabama	Hispanic	53400
0	alabama	White	414000
95	texas	Black	680700
147	texas	Hispanic	2475100
43	texas	White	1039400

Figure 14: porvertyKFF_table visualisation

nces_table2

	Year	State	Ethnicity	Grade	Enrollment	Dropout
1863	2002	alabama	Black	7	22817	18
482	2002	alabama	Black	8	20592	50
3007	2002	alabama	Black	9	22993	605
1011	2002	alabama	Black	10	18847	642
2102	2002	alabama	Black	11	16022	659

Figure 15: nces_table2 table visualisation

nces_table

	Year	State	Ethnicity	Grade	Enrollment	Dropout
0	2002	alabama	Black	7	22817	18
168	2002	alabama	Black	8	20592	50
336	2002	alabama	Black	9	22993	605
504	2002	alabama	Black	10	18847	642
672	2002	alabama	Black	11	16022	659

Figure 16: nces_table table visualisation

The comparative analysis yielded congruent datasets between the original tables and those derived from RDF conversions. This parity is evident across all tables, confirming the RDF conversion process's accuracy and reliability.

Subsequent data quality assessments were conducted on the newly formed tables. These checks were pivotal in asserting the structural integrity and consistency of the data post-conversion. Although minor discrepancies in data types were observed, these did not detract from the overall structural alignment of the datasets.

RDF Evaluation Interpretation

The examination of RDF conversion efficacy, facilitated by a direct comparison between original and RDF-derived tables, underscores the robustness of RDF as a data interchange model. The congruence observed between the original and converted datasets not only attests to the successful execution of RDF conversions but also validates the integrity of the data post-conversion.

Despite minimal differences in data types, the structural consistency remained intact, affirming the reliability of RDF in preserving data fidelity across conversions. This study thereby concludes that RDF conversions, when executed within the specified parameters, maintain data integrity and structure, thereby serving as a reliable model for data interchange in diverse and schema-variant environments.

Linked Data Evaluation

The study embarked on an evaluation of RDF data to assess the efficacy of linked data conversions and the subsequent retrieval of integrated data using SPARQL queries.

The *rdflib* library was employed to execute SPARQL queries against the linked_data_detailed RDF data. The primary objective was to retrieve specific data subsets and evaluate the integration and accessibility of the intended data.

To facilitate this evaluation, specific SPARQL queries were constructed to extract subsets of the integrated data:

```
# Queries
query_poverty_data = """
PREFIX ex: <http://educationaldata.org/>
SELECT ?state ?poverty
WHERE {
  ?state a ex:State .
  ?state ex:Poverty ?poverty .
}
"""
query_census_data = """
PREFIX ex: <http://educationaldata.org/>
```

```
    SELECT ?state ?census
    WHERE {
      ?state a ex:State .
      ?state ex:Census ?census .
    }
    """
    query_nces_data = """
    PREFIX ex: <http://educationaldata.org/>
    SELECT ?enrollment
    WHERE {
      ?recordUri ex:Enrollment_2002_7 ?enrollment .
    }
    """
```

```
Poverty Data:
State: http://educationaldata.org/northcarolina, Poverty: 710900
State: http://educationaldata.org/northcarolina, Poverty: 533100
State: http://educationaldata.org/northcarolina, Poverty: 264000
State: http://educationaldata.org/mississippi, Poverty: 229100
State: http://educationaldata.org/mississippi, Poverty: 377000
State: http://educationaldata.org/mississippi, Poverty: 19500

Census Data:
State: http://educationaldata.org/northcarolina, Census: 6844118
State: http://educationaldata.org/northcarolina, Census: 2163190
State: http://educationaldata.org/northcarolina, Census: 804826
State: http://educationaldata.org/mississippi, Census: 1796181
State: http://educationaldata.org/mississippi, Census: 1124460
State: http://educationaldata.org/mississippi, Census: 73435

NCES Data:
Enrollment: 19911
Enrollment: 394
Enrollment: 21130
Enrollment: 36738
Enrollment: 3612
Enrollment: 26435
```

The execution of SPARQL queries demonstrated the capability to retrieve specific data segments from the linked_data_detailed RDF file effectively. This outcome underscores the successful linkage and establishment of relationships within the data, facilitating targeted data retrieval.

Project Overall Success

The comprehensive evaluation conducted across various facets of the project—spanning data quality assessment, RDF conversion integrity, and linked data querying efficacy—culminates in a resounding affirmation of the project's success. The rigorous data quality assessment ensured a solid foundation for analysis, while the fidelity of RDF conversions preserved data integrity, enhancing semantic interoperability. The effective retrieval of targeted data subsets via linked data queries further demonstrated the project's capability in data integration and accessibility..

Collectively, these evaluations not only validate the project's methodological approach but also highlight the potential of semantic web technologies in fostering data and enhancing analytical insights.

Furthermore, the successful completion of data analysis, the findings of which are elucidated in the conclusion section, underscores that all facets of the project were meticulously executed and comprehensively presented.

7. Conclusion

The primary objective of this research is to elucidate the educational disparities evident through dropout rates and their association with poverty levels across the states of America. This objective has been pursued by transforming the available datasets into linked data, facilitating their interconnection and subsequent data extraction for in-depth analysis.

Poverty Analysis

The initial phase of this study involves delineating the poverty distribution across the United States:

	Ethnicity	Population(%)	Population_In_Poverty(%)
2	White	71.125105	46.318253
1	Hispanic	16.162201	31.479479
0	Black	12.712694	22.202269

Figure 17: Results of Populations (%) vs Population in Poverty (%) per Ethnicity

The analysis delineates the demographic composition, revealing that the White population constitutes 71.12% of the total population, followed by the Hispanic population at 16.12%, and the Black population at 12.71%. Furthermore, a pronounced correlation between the total population and the poverty-stricken population is identified:

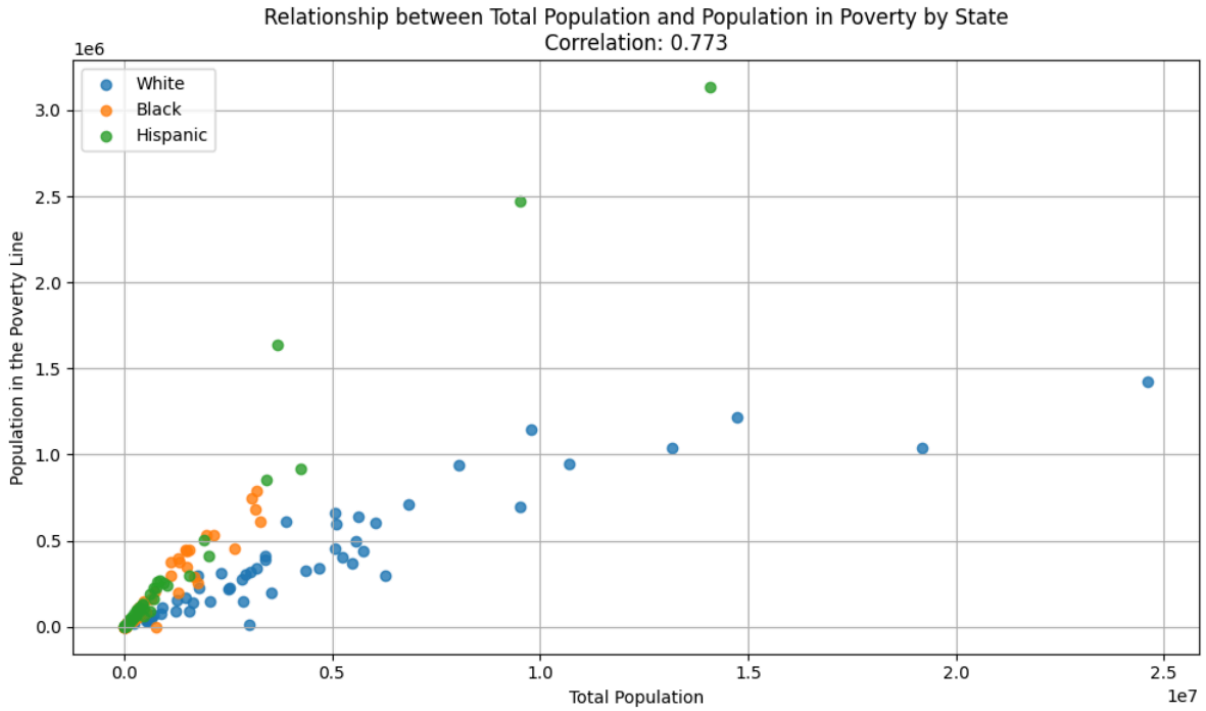

Figure 18: Scatter Graph Displaying the relationship between the total population and the Population in Poverty

The correlation coefficient of 0.773 between the total population and the prevalence of poverty indicates a significant positive relationship. This suggests that states with larger populations are more likely to have a greater number of individuals living in poverty.

This preliminary data allows for the anticipation of ethnic disparities in poverty rates. Despite the White population making up 71.12% of the total, their proportion of the poverty-stricken population does not surpass 50%.

Advancing the analysis to the poverty line's demographic proportions validates these expectations:

	Ethnicity	Census	Poverty	percent_poverty
0	Black	42805394	9791800	22.875154
1	Hispanic	54420362	13883300	25.511223
2	White	239488042	20427600	8.529695

Figure 19: Proportion of the population in the poverty line per Ethnicity

The data reveal that 25.51% of the Hispanic population and 22.87% of the Black population live below the poverty line, compared to 8.52% of the White population. For a more tangible comparison, a segment from two graphical representations illustrates the distribution of the general and poverty-stricken populations by ethnicity across the states (comprehensive graphical analyses are included in the notebook):

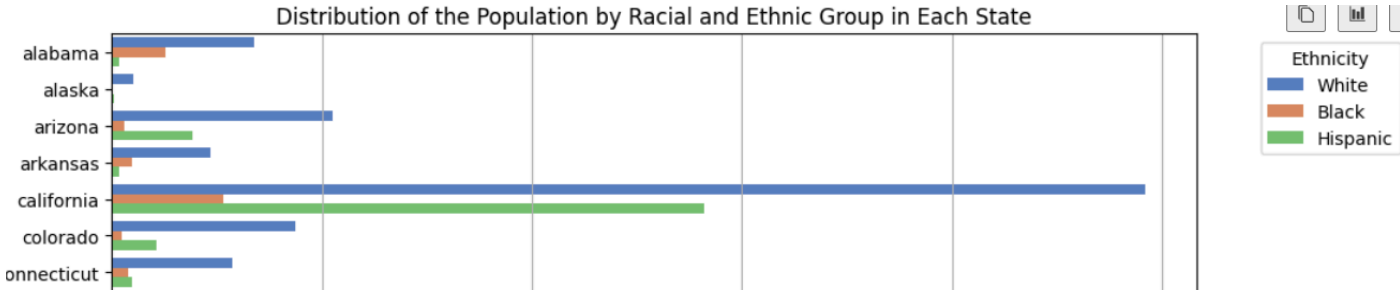

Figure 20: Horizontal Bar Graph Displaying Distribution of the Population by Racial and Ethnic Group in Each State of America

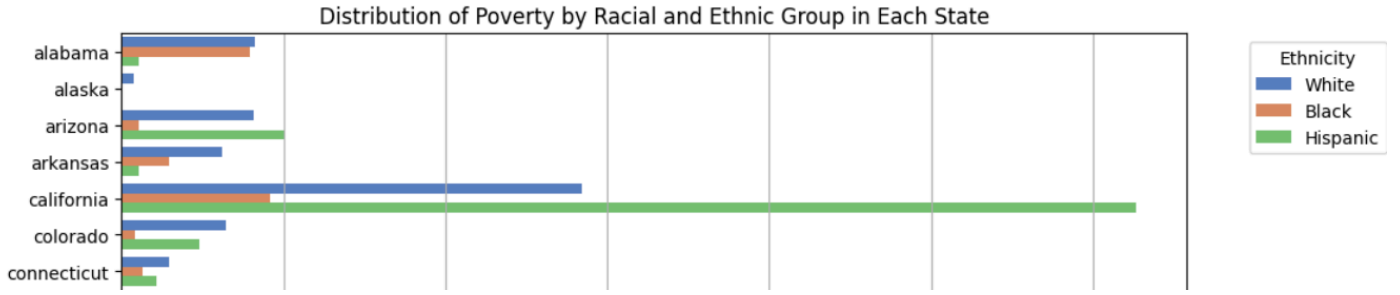

Figure 21: Horizontal Bar Graph Displaying the Distribution of the Poverty by Racial and Ethnic Group in Each State of America

This comparison underscores that, despite the larger size of the White population, Hispanic and Black communities endure a higher rate of poverty. This disparity accentuates the impact poverty can exert on educational outcomes, as children and adolescents from impoverished backgrounds are likely to encounter additional hurdles in accessing sufficient educational resources, academic support, and learning opportunities.

Dropout Analysis

The calculation of the dropout rate employed the subsequent formula, predicated upon the mean figures for "Enrollment" and "Dropout" across each "Ethnicity":

$$Dropout\ Rate\ =\ (Dropout/Enrollment)\ X\ 100$$

Between the years 2002 and 2004, observable consistency in the variation of dropout rates across different grade levels was discerned. This consistency suggests that specific educational stages regularly face hurdles in maintaining student engagement and transcending the confines of the academic year.

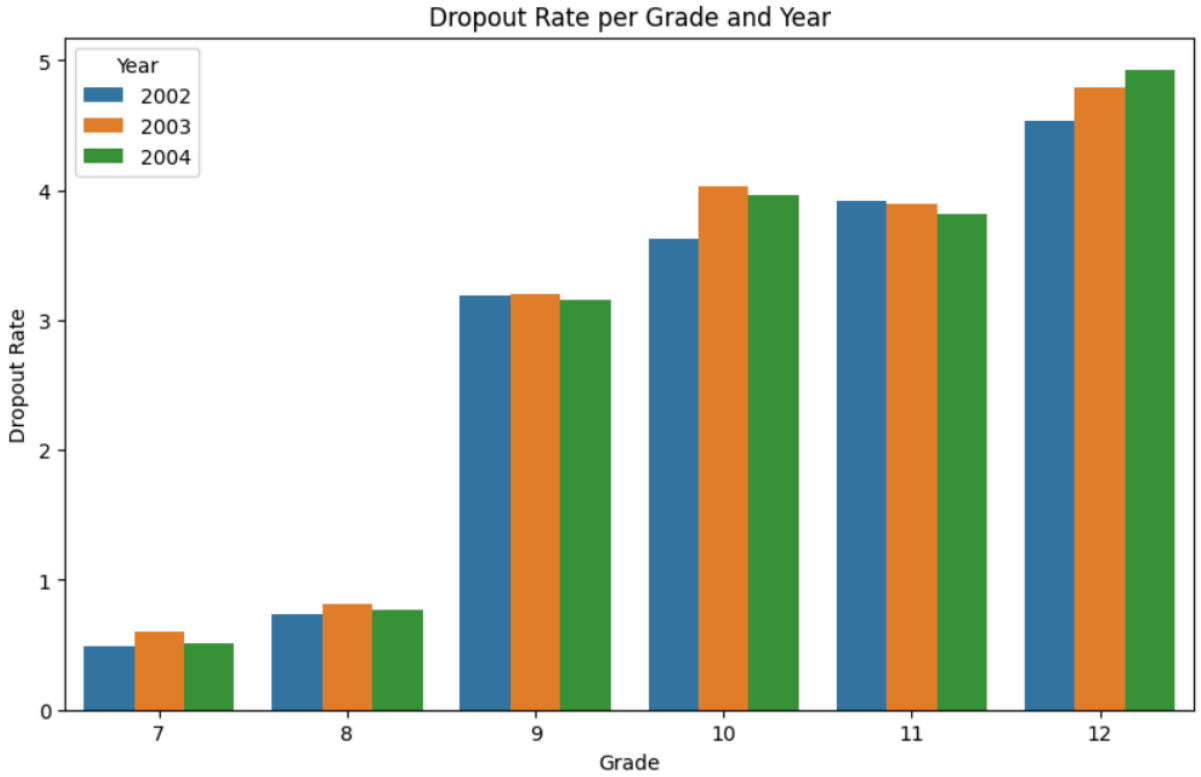

Figure 22: Vertical Bar Graph Displaying the Dropout Rate per Grade and Year

Upon examining the dropout rates across various grade levels over the years, distinct disparities become evident. Specifically, the dropout rates in upper grades—10th through 12th—are markedly higher than those in the 7th through 9th grades.

This phenomenon may correlate with legislative requirements in some States that mandate school attendance only up to the age of 16, typically the age of students in the 9th grade. Although there has been a policy shift towards extending compulsory education to the age of 18/19 since the year 2000, it is crucial to acknowledge that the dataset spans from 2002 to 2004.

When the analysis is pivoted to consider dropout rates by ethnic group, it becomes apparent that Black and Hispanic students exhibit elevated dropout rates, with the Black group recording the highest rates up to the 11th grade, at which point the Hispanic group

surpasses them. Concurrently, the White group consistently demonstrates lower dropout rates across all grades:

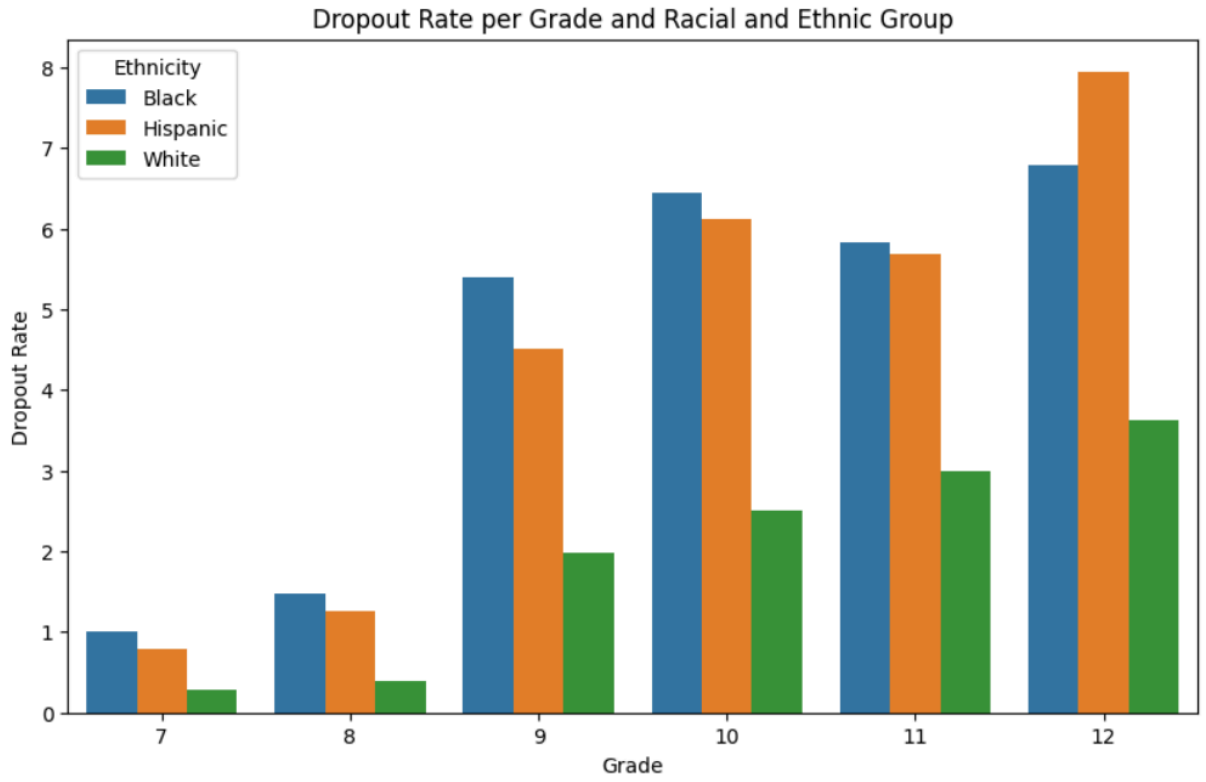

Figure 23: Vertical Bar Graph Displaying the Dropout Rate per Grade and Racial and Ethnic Group

Poverty and Dropout Rates Analysis

The analysis reveals a correlation between poverty levels and school dropout rates by ethnicity. Ethnic groups with higher poverty rates are observed to have correspondingly higher school dropout rates. For instance, the "Black" and "Hispanic" ethnicities, which have a higher percentage of their populations in poverty, also experience higher school dropout rates compared to the "White" ethnicity, which has a lower percentage of its population in poverty:

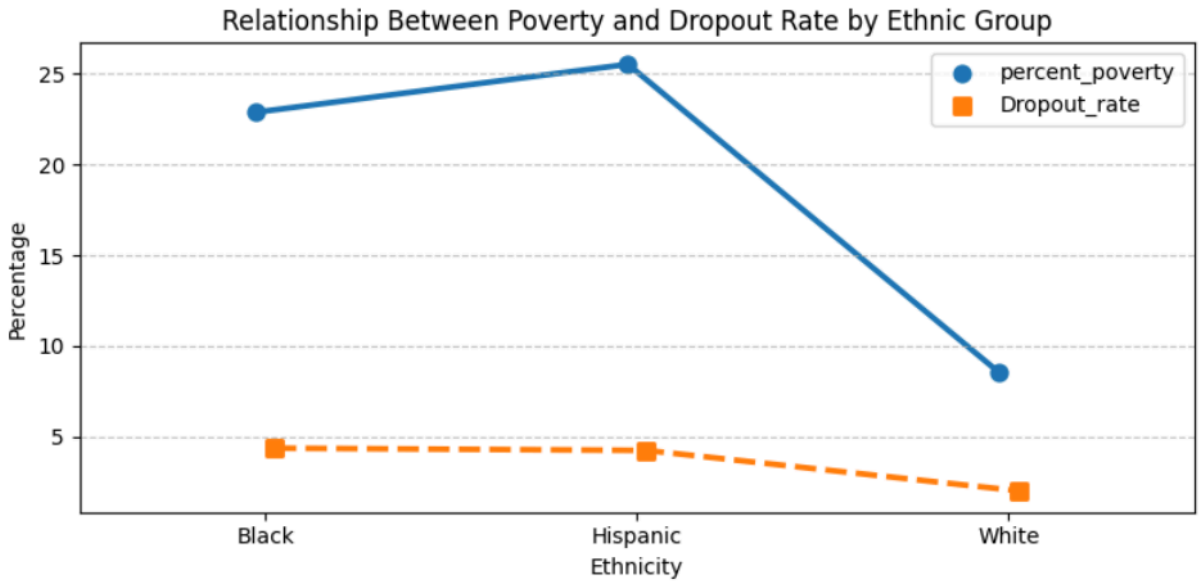

Figure 24: Line Graph Displaying the relationship between Poverty and Dropout Rates by Ethnic Group

Final Considerations

In conclusion, this research has embarked on an analytical journey to uncover the multifaceted nature of educational disparities. Through the transformation of datasets into linked data, a comprehensive analysis was conducted, shedding light on the disparities that exist between different ethnic groups in terms of both poverty rates and educational outcomes.

The demographic analysis indicates a disproportionate representation of poverty among ethnic groups, with the White population having a lower percentage of individuals living in poverty compared to the Black and Hispanic populations. This disparity is further magnified when examining dropout rates, where Black and Hispanic students are found to have higher rates than their White counterparts.

Moreover, the study highlights the strong correlation between poverty levels and dropout rates across ethnicities, suggesting that poverty serves as a substantial barrier to educational attainment.

The findings from this research underscore the need for targeted interventions and policies aimed at addressing the causes of poverty and its impact on educational equity. By understanding the interplay between poverty, ethnicity, and educational outcomes, stakeholders can better devise strategies to promote inclusivity and support for underrepresented and disadvantaged communities.

The project would benefit from further development in data analysis techniques to include more advanced statistical models and machine learning algorithms for predictive analysis.

Including such techniques could offer predictive insights into future trends in educational disparities, providing a more forward-looking analysis. To incorporate this, future iterations of the project could leverage additional Python libraries like scikit-learn for machine

learning and deeper statistical analysis to predict potential outcomes based on current data trends.

References:

- ACS DEMOGRAPHIC AND HOUSING ESTIMATES. (2010). Census Bureau. Available from: https://data.census.gov/table?q=DP05&g=010XX00US$0400000&y=2010 [17 December 2023].
- Addlesee, A. (2018) *Understanding Linked Data Formats: RDF, XML vs Turtle vs N-Triple*s. [Online] Medium. Available from: https://medium.com/wallscope/understanding-linked-data-formats-rdf-xml-vs-turtle-vs-n-triples-eb931dbe9827 [Accessed 17 December 2023].
- Berners-Lee, T. (2006). *Linked Data*. Available from: http://www.w3.org/DesignIssues/LinkedData.html [Accessed 17 December 2023].
- Bizer, C., Heath, T., and Berners-Lee, T. (2009). *Linked Data - The Story So Far*. International Journal of Semantic Web and Information Systems, 5(3), pp.1–22.
- Fielding, R., n.d. *Architectural Styles and the Design of Network-based Software Architectures.* (PhD), University of California, Irvine.
- Fram, M., Miller-Cribbs, J., Van Horn, M. (2007). P*overty, Race, and the Contexts of Achievement: Examining Educational Experiences of Children in the U.S. South*. Social work. 52. 309-19. 10.1093/sw/52.4.309.
- *Frequently Asked Questions*. n.d. U.S. Census Bureau. Available from: https://ask.census.gov/prweb/PRServletCustom/app/ECORRAsk2_/YACFBFye-rFIz_FoGtyvDRUGg1Uzu5Mn*/!STANDARD [Accessed 17 December 2023].
- Historical Poverty Tables: People and Families - 1959 to 2022. n.d. U.S. Census Bureau. Available from: https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-people.html#par_list [Accessed 17 December 2023].
- Ho Lee, S., (2013). *John Dewey's View on School and Social Reform*, pp. 125. Chung-Ang University.
- Lebo, T., Williams, G. T. (2010). *Converting governmental datasets into linked data. Proceedings of the 6th International Conference on Semantic Systems* (I-SEMANTICS '10). Association for Computing Machinery, New York, NY, USA, Article 38, 1–3. https://doi.org/10.1145/1839707.1839755
- Paquet, A. (2020) *Linked Data and Linked Open Data Projects for Libraries, Archives and Museums: Constructing Pathways to Information Discovery and Cultural Heritage Sector Collaboration.* John Hopkins University, Baltimore. AS.460.674.81- Digital Curation Research Paper.
- Pereira, C. *et al* (2017). *Linked Data in Education: A Survey and a Synthesis of Actual Research and Future Challenges*. IEEE Transactions on Learning Technologies, vol. 11, no. 3, pp. 400-412, 1 July-Sept. 2018, doi: 10.1109/TLT.2017.2787659.
- Poverty Rate by Race/Ethnicity. (2002). Kaiser Family Foundation. Available from: https://www.kff.org/other/state-indicator/poverty-rate-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D [Accessed 17 December 2023].
- Poverty Rate by Race/Ethnicity. (2010). Kaiser Family Foundation. Available from: https://www.kff.org/other/state-indicator/poverty-rate-by-raceethnicity/?dataView=1&currentTimeframe=11&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D [Accessed 17 December 2023].
- Ruth, L., Wood, D., Zaidman, M. (2014). *Linked Data - Structure data on the web*. New York: Manning Publications Co.
- State-Level Public School Dropout Data: 2002–2003 through 2004–05. n.d. National Center for Education Statistics. Available from: https://nces.ed.gov/ccd/ccddata.asp [Accessed 17 December 2023].
- Tucker, A. (2022). *Facing the challenge of digital sustainability as humanities researchers*. The British Academy. Journal of the British Academy, 10: 93–120.
- Weir, K. (2016). *Inequality at school: What's behind the racial disparity in our education system?* [Online] American Psychological Association. Available from: https://www.apa.org/monitor/2016/11/cover-inequality-school [Accessed 17 December 2023].
- Zapilko, B., Harth, A., Mathiak. (2011) *Enriching and Analysing Statistics with Linked Open Data*. NTTS - Conference on New Techniques and Technologies for Statistics.
- Zapilko, B. (2014). *Methods for Matching of Linked Open Social Science Data*. (PhD), University of Mannheim, Mannheim.