Data Report

Correlation analysis between Female School Education and Female Political Representation in America (Canada, United States, Guatemala, Honduras)

Laura Dauti

Methods of Advanced Data Engineering

Introduction and Main Question

Gender Inequality in political representation is an important problem, because it hinders progress toward a more inclusive and representative governance.
Despite the progress made in various regions, women still remain underrepresented in political positions across the globe, which makes it crucial to address this imbalance.
This project analyzes if there is a correlation between female school education and female political representation across the Americas, focused on four countries, namely Canada, the United States, Guatemala and Honduras, using comprehensive datasets and statistical correlation. These countries were chosen to have a comparison between more developed countries in North America (Canada and the United States) and less developed countries in South America (Guatemala and Honduras).
The results of this project can give insights into the factors that promote or hinder women's political participation.
The main question for this project is: Is there a correlation between a woman's school education and female political representation across the Americas (Canada, United States, Guatemala, Honduras)?

Data Sources

For answering the main question, four different data sources were used. These data sources were chosen because they provide comprehensive and reliable information essential for analyzing the correlation between female education and political representation across the described selected countries in America. They were found by searching for suitable metadata to answer the main question.

The first data source contains data about the percentage of seats in parliament that are held by women, nearly for every country worldwide. This data source is from "Our World in Data", which focuses on large global problems. [1]
It was chosen because it offers valuable insights into women's representation in politics, particularly in parliament, which significantly helps to answer the main research question.
The data from this data source is structured, meaning that it is organized in a fixed schema and its data format is a CSV file.
Concerning the quality of the data, it is accurate since it reflects the real percentage of seats in parliament that are held by women. It is also complete, it contains all necessary information, such as the country name, the year and the percentage. Throughout the file the format is consistent and the age of data is appropriate, it contains data from 1997 to 2022. Last, the data is also relevant because it fits very well to answer the main question.
All data from "Our World in Data" are completely open access under the Creative Commons BY-4.0 license. They emphasize the permission to use, distribute and reproduce their data in any medium, provided the source and authors are credited and a link to the license is provided. The license for "Our World in Data" can be looked up here `https://ourworldindata.org/faqs#can-i-use-or-reproduce-your-data`.
But since the data was produced by a third-party provider and made available by "Our World in Data", it is subject to the license terms from the original providers, according to "Our World in Data". The original provider for this data source is the "World Bank Group". The license type of the "World Bank Group" is the same as for "Our World in Data", namely the Creative Commons BY-4.0, which has the same terms as described. To follow their obligations, it is necessary to fulfill these terms appropriately. The license for the "World Bank Group" can be looked up here `https://datacatalog.worldbank.org/public-licenses#cc-by`.

The second data source contains data about the learning years of school. It shows the expected years that a pupil will stay in school, by gender and countries.
This data source is from the "World Bank's Gender

Data Portal". [2]
It was chosen because it offers valuable insights into women's expected school education duration, which significantly helps to answer the main research question. The data from this data source is structured, meaning that it is organized in a fixed schema and its data format is a CSV file.
Concerning the quality of the data, it is accurate since it reflects the real expected years of schooling for women. It is also complete, it contains all necessary information, such as the country name, the year, and the expected years. Throughout the file the format is consistent and the age of data is appropriate, it contains data from 2010, 2017, 2018 and 2020. Last, the data is also relevant because it fits very well to answer the main question.
The data from the "World Bank's Gender Data Portal" are under the same license as the "World Bank", namely the Creative Commons BY-4.0 license, meaning that it allows copying, modifying and distributing data in any format for any purpose, provided that appropriate credit is given. Since the license is the same as for the "World Bank", it can be looked up at the same link `https://datacatalog.worldbank.org/public-licenses#cc-by`.

The third and fourth data sources contain data about the percentage of secondary and tertiary school enrollment, respectively. They show the percentage of how many pupils are enrolled in secondary or tertiary school, by gender and countries.
These data sources are also from the "World Bank's Gender Data Portal". [3, 4]
They were chosen because they offer valuable insights into women's degree of school education, which significantly helps to answer the main research question.
The data from these data sources are both structured, meaning that they are organized in a fixed schema and their data format are CSV files.
Concerning the quality of both the data, it is accurate since they reflect the real percentage of women's school enrollment in secondary and tertiary school. They are also complete, they contain all necessary information, such as the country name, the year and the value. Throughout the files the format is consistent and the age of data is appropriate. The data source for secondary school enrollment contains data from 2010 to 2018, which is not as current as the other data sources. And the data source for tertiary school enrollment contains data from 2009 to 2020. Last, the data are also relevant because they serve very well to answer the main question.
Since these two data sources are from the same source as the second data source, everything described about the license for "World Bank's Gender Data Portal" also applies here.

Data Pipeline

For implementing the data pipeline, Jayvee was used. Jayvee is a domain-specific language tailored for automated processing of data pipelines. The core concepts of Jayvee are pipelines, blocks and value types. A pipeline is a sequence of different computing steps, the blocks. The default output of a block becomes the default input of the next block, building a chain of computing steps. There are three types of blocks, namely Extractor blocks for modeling the data sources, Transformator blocks for modeling a transformation and Loader blocks for modeling a data sink. A value type is the definition of a data type of the processed data. [5]
For this project, four data pipelines were used, one for each data source. The structure of the pipeline for the first data source is shown in figure 1.

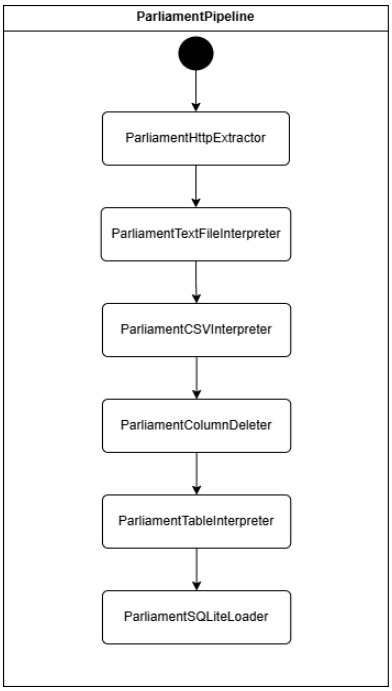

Figure 1: *Pipeline of the first data source for women in parliament*

Here, the first block "ParliamentHttpExtractor" is the Extractor block for downloading a file from the internet. The following four blocks are for modeling transformation and finally the last block "ParliamentSQLiteLoader" is a Loader block for creating a SQLite database with the processed data. The structure of the other three pipelines are similar to the pipeline seen in figure 1 with a few other Transformator blocks.

To improve the data quality and reliability, some transformation and cleaning steps were done.
By manually looking at the structure of the CSV files, it was apparent that all tables in the CSV start in the first row. Therefore, no rows had to be deleted.
Some columns were deleted that are not relevant for the output data like the country code. With this cleaning step, the output data is more adjusted to find out the really relevant columns for answering the main question.
It is ensured that there are no important values missing, invalid or duplicate in a column by doing validity checks, namely by defining constraints on numerical values, such as on the years and on the values.
The last cleaning step of the data sources that was done is to clean out non-relevant countries, since the data sources contain information for almost all countries worldwide. To have more adjusted output data, only the countries Canada, the United States, Guatemala and Honduras are kept from the data source by defining a regex constraint.
The only problem that appeared while processing the data is that there were some difficulties directly extracting the data sources from the internet from "World Bank's Gender Data Portal". However, for the data source "Our World in Data", it worked. By writing a Bash script that downloads the three data sources from "World Bank's Gender Data Portal" and stores them locally and then extracting them in the data pipeline with the local file extractor; the problem was solved.
Apart from the described data integrity checks, retries were implemented in the "Parliament Pipeline" since it is the only one extracted from the web directly. Retries ensure robustness and fault tolerance and are used to resolve some errors, like network timeouts, by retrying the operation after a brief delay. The maximum number of retry attempts was set to three. The wait time before executing a retry is defaultly set to 1000 milliseconds.

Result and Limitations

The output data of each data source contains the country name (Canada, Guatemala, Honduras or the United States), the most recent years available and the values for these years. The data structure of the output data is structured, meaning that it is organized in a fixed schema, since a relational database is used. The output data is stored in a SQLite database as SQLite files for efficient querying and storage and for structured data. Each pipeline is stored in one table in the same database. The fields are defined with appropriate data types to ensure consistency.
Concerning the quality of the output data, it is accurate and correct, consistent in its format and relevant for answering the main question. It provides an insight into the development of women's school education and the number of women's seats in parliament, since it contains data not only for one year, but partially starting from 2007 to 2022. This enables a better analysis and a better answer to the question if there is a correlation between women's school education and political representation, since a better comparison can be made.
Regarding the completeness of the data, there are some irregularities in the number of data collected for some countries that can be a potential issue for the analysis. The output data of the expected years of school education has only data for four years available, namely for 2010, 2017, 2018 and 2020. With this data an analysis can be made more difficult to achieve right conclusions. The output data of the secondary school enrollment contains more data for more years but the latest year data is available for is 2017, which represents slightly older data. Hopefully, this won't be a big issue for the final report since the other data are more current. On the other hand the other two output data are available for a minimum of ten years and the latest data available is from 2022. There is also no difference between the amount of data available for each country, meaning the same amout of data is available for the smaller countries like Guatemala and Honduras and for the bigger countries Canada and the United States. With that a good analysis and right conclusions can be achieved. The only limitation introduced by data processing is that every data available before 2007 was thrown away to achieve a more current result. No limitations in the original data are known. For the final analysis and report it is crucial to implement more methods into the pipeline for dealing with errors or changing input data, since until now only the retry, described in the chapter before, was implemented.

References

[1] Multiple sources compiled by World Bank – processed by Our World in Data. *"Share of women in parliament" [dataset]. Inter-Parliamentary Union (via World Bank), "World Development Indicators" [original data]*. 2022. URL: https://ourworldindata.org/grapher/share-of-women-in-parliament-ipu (visited on 11/27/2024).

[2] World Bank Group - Gender Data Portal. *Learning-Adjusted Years of School*. 2020. URL: https://genderdata.worldbank.org/en/indicator/hd-hci-lays (visited on 11/27/2024).

[3] World Bank Group - Gender Data Portal. *School enrollment, secondary (%)*. 2020. URL: https://genderdata.worldbank.org/en/indicator/se-sec-enrr (visited on 11/27/2024).

[4] World Bank Group - Gender Data Portal. *School enrollment, tertiary (% gross)*. 2024. URL: https://genderdata.worldbank.org/en/indicator/se-ter-enrr (visited on 11/27/2024).

[5] Jayvee. *Core Concepts*. 2024. URL: https://jvalue.github.io/jayvee/docs/user/core-concepts (visited on 11/27/2024).