

Project Report - Group 2

What factors influence adult female literacy rates?

# Introduction

Female literacy rates remain a significant global issue, and understanding the factors that contribute to higher or lower rates can help identify areas for improvement. This project aims to study the key factors contributing to female literacy rates in the top 5 countries versus the bottom 5 countries between the years 2000 and 2021. The objective is to identify and compare the factors that have the most and least impact on female literacy rates, providing valuable insights for governments and charities working to address this issue.

The report will discuss the background of the project, where we will explain the type of analysis that took place in this project, how relevant this analysis was to our overall objective and our target audience. The step specifications will go through our approach to each of the steps in our data analysis process: question framing, data gathering and preprocessing, and data exploration and analysis. The implementation and execution section will go through our development approach, team member roles, implementation process and challenges. The result reporting highlights the key findings of our project and the conclusion section presents our final conclusions.

# Background

The project utilises data from various sources, including the UNESCO Institute for Statistics, World Bank, and national statistical agencies. These databases provide reliable and internationally accepted statistics. Our strategy was to investigate the variation in our different variables and measure their effect of female literacy rates for our chosen sample. Our conclusion details the most and least impactful factors that influence female literacy rates

We took the method of analysis for all of our notebooks. We started with basic data exploration to understand the data was presented and what it meant, we then preprocessed it in a way that suited how we wanted to present our final analysis, and this was in the form of a merged table with all of our different variables alongside the literacy rate with each of the ten countries that had been selected. After preprocessing the data, we used visualisations for descriptive analysis which allowed us to summarise the main findings of the data and present the data in a meaningful way. Finally when we had created our merged table, we performed regression analysis to see how significant our figures were. The approach we took aligned with our objectives because it allowed us to investigate each of our chosen factors that would contribute to female literacy, then merge them in our table so that we can measure how significant our model was as a whole.

Our target audience is varied as this is a topic that is useful for several groups to understand female literacy rate. They should be used to channel resources into bridging the gap in female literacy rate.Our target audience includes education providers/government authorities, charities focused on [female] education, edtech companies and anyone in academia focusing on female education.

For edtech companies, we would expect this project to provide a commercial use case for providing their services to countries they hadn’t considered before. These would be countries that are most affected by a low female literacy rate. They can work with local community partners in the most affected countries to help improve female literacy in these countries. Our report can also encourage other edtech companies to fund charities focused on improving female literacy. This report could also be used when private companies want to further their CSR initiatives, they can use this report to help them make decisions on areas to support.

For education providers or government authorities that manage education, this project would increase understanding of the different factors that may create a literacy rate. Government tax and spending budgets have economic, social and environmental implications on a country’s population. Reports like ours could help governments improve their gender budgets by supporting education providers and policy makers. An equally educated society helps shape people’s choices and impacts their contribution to the society at large.

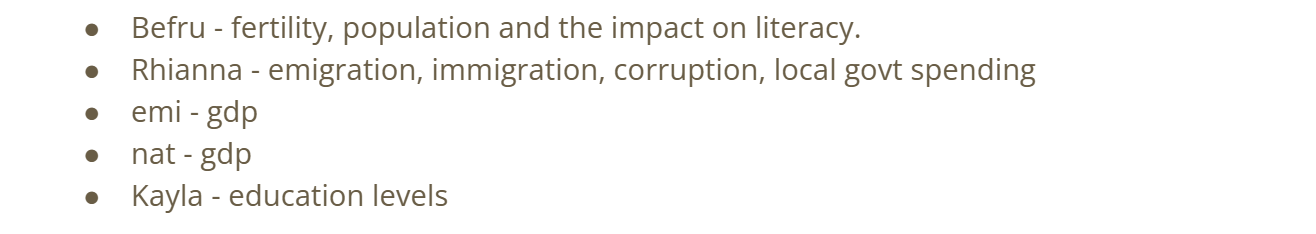
For charities, NGOs and women right activists that are focused on female education, the insights from this report can be used to design programs and initiatives that aim to improve female literacy rates in areas where it remains low.

For those in academia, this report can be used as a basis for further analysis, to test conclusions and increase literature on this topic.

As listed above, it can be seen that our target audience is varied therefore, we will make the report easy to understand for everyone with little to no technical knowledge by using graphs that can be interpreted easily and providing further analysis to aid further understanding in the subject matter.

# Steps Specifications

Our data analysis process consisted of research framing questions, data gathering, preprocessing, visualisation and then analysis. Once we had decided our topic of female literacy rates, we all researched factors that contribute to these rates. To make sure that we weren’t all researching the same thing, we broke down a non exhaustive list of possible factors that would contribute to female literacy rates. We assigned each person with a couple of topics to research as pictured below.



After this, we then had another meeting where we discussed what we found and how we could frame the question. From the literature, we found that government spending, global peace index were widely researched in relation to female literacy rate, whereas topics like fertility didn’t have an abundance of datasets on it.

Some of the sources of our data sets:

* Data per country from UNICEF (link [here](https://data.unicef.org/resources/data_explorer/unicef_f/?ag=UNICEF&df=GLOBAL_DATAFLOW&ver=1.0&dq=.PT_F_20-24_MRD_U15+PT_M_20-24_MRD_U18+PT_F_20-24_MRD_U18+PT_M_15-19_MRD+PT_F_15-19_MRD..&startPeriod=2016&endPeriod=2022)) <https://data.unicef.org/resources/data_explorer/unicef_f/?ag=UNICEF&df=GLOBAL_DATAFLOW&ver=1.0&dq=.PT_F_20-24_MRD_U15+PT_M_20-24_MRD_U18+PT_F_20-24_MRD_U18+PT_M_15-19_MRD+PT_F_15-19_MRD..&startPeriod=2016&endPeriod=2022> The dataset shows the percentage of women (aged 20-24 years) married or in union before age 15.
* Female literacy dataset used ([here](https://data.worldbank.org/indicator/SE.ADT.LITR.FE.ZS)) <https://data.worldbank.org/indicator/SE.ADT.LITR.FE.ZS>
* Literacy dataset that was cleaned: <https://ourworldindata.org/literacy>

Due to our extensive online research, we had a good grasp on key factors that were influential in affecting female literacy rates. In our third meeting, we discussed how we would frame the question. At first, we discussed basing the project on one factor that influenced literacy rates, but after looking at other examples of data analysis projects, we decided to change the objective to study various factors that would influence female literacy rates. This would allow for more data exploration. We decided to choose four factors: government expenditure, child marriage, global peace index and income levels.

The data for all of our variables for each country was collected and structured in a CSV file for analysis. Descriptive statistics were calculated for each variable. Visualisations were created to analyse the data distribution and identify potential relationships between the variables. Statistical models were employed to analyse the factors influencing female literacy rates.

The data sources used are from UNESCO Institute for Statistics, World Bank, UNICEF and national statistical agencies. Their databases are essential tools for providing key statistical information; reports are consistent and reliable as they follow internationally accepted standards of data statistics. When combining data from national statistics websites that have different columns and table structures, there were several strategies employed to overcome the challenges and effectively combine the data for the project. We started by identifying common attributes or variables across the different datasets. Look for columns or fields that contain similar information but might be named differently. Once we have identified common attributes, we transformed and standardised the data to ensure consistency. This could involve renaming columns, converting data types, or applying formatting rules. One of the most challenging parts is dealing with missing data appropriately. We first determined the reason for missing values and decide how to handle them based on the project's requirements. We may choose to impute missing values using statistical methods or remove incomplete records altogether, depending on the impact of missing data on the analysis. Lastly, we performed thorough data validation and quality checks to ensure the accuracy and consistency of the combined dataset. This includes checking for outliers, inconsistencies, and data integrity issues.

Natalie, Emi, Lilia, and Rhianna were responsible for specific notebooks. In these notebooks, they performed in-depth analysis which consisted of calculating the mean and/or median for our datasets. We were then able to explain the relationships in the data and summarise them through descriptive statistics, visualisations, and other exploratory techniques. This helped identify patterns, trends, relationships, and potential outliers in the data. Once we had collected all of our data, Lilia used regression analysis to understand how changes in the independent variables affect the dependent variable and to also quantify the strength and direction of these relationships. OLS regression is a form of regression analysis that assumes a linear relationship between the variables. In OLS regression, the goal is to fit a line to the data that minimises the sum of squared differences between the observed values of the dependent variable and the predicted values from the regression model. The line represents the best-fit linear approximation of the relationship between the variables.

# Implementation & Execution

Each team member was assigned some roles in order to help the project run smoothly (Table 1).

Table 1: Team member roles.

| **Names** | **Roles** |
| --- | --- |
| Befru | * Research & data gathering * Final report * Code review |
| Emi | * Adult female literacy notebook (cleaning, exploration, analysis & visualisations) * Income notebook using APIs (cleaning, exploration, analysis & visualisations) * Code review * Readme preparation |
| Natalie | * Adult female literacy notebook (cleaning, exploration, analysis & visualisations) * Child marriage notebook (cleaning, exploration, analysis & visualisations) * Code review * Presentation preparation |
| Kayla | * Research & data gathering |
| Rhianna | * Organised meeting times using Doodle and wrote up meeting notes * Government education expenditure notebook (cleaning, exploration, analysis and visualisations) * Final report |
| Lilia | * Regression notebook (exploration, visualisations & analysis) * Code review * GPI notebook (cleaning, exploration, analysis and visualisations) |

As we are considering several different factors and their influence on female literacy rates, we were able to combine resources of different datasets to expand their data series. The objective of this project is to study the key factors, by study we mean identify and compare the key factors, once we do this with the data, we would consider the core questions answered. We used Pandas to clean and deal with missing values in the data and Matplotlib and seaborn for creating visualisations and statistical analysis. Other tools and libraries that we used include: Slack, Doodle, Google notes and apps, Github, Jamboard, VScode, Jupyter Notebook, numpy, statsmodels, readme.so.

Our team's agile development approach allowed for flexibility, collaboration, and continuous improvement. Regular team meetings were scheduled to review progress, discuss findings, and address any challenges. The team members worked on their assigned tasks concurrently and shared their outputs for collaborative review and integration. This iterative approach enabled the team to adapt and make necessary adjustments throughout the project. By following this development approach and assigning specific roles and tasks to team members, the project progresses efficiently, allowing for thorough analysis and comprehensive reporting of the key factors influencing female literacy rates in different countries. GitHub was used to review code which allows for comments and feedback on necessary changes required.

The team encountered a few challenges during the implementation of the project. These challenges are listed below:

1. Missing countries. However, we could not do anything for the adult female literacy dataset as to the best of our knowledge, there were no other datasets available to combine with. For missing countries in other variables, we tried to fill in the gaps using either regional data (i.e., Sub-Saharan Africa for Niger) or by combining our DataFrame with another dataset which contained the target countries.
2. Nulls were present within the dataset. We had to decide how we would deal with these (i.e., deleting the nulls).
3. Outliers/skewness. We determined how to deal outliers/skewness with by visualising the data and comparing the means and medians to see if they were vastly different.
4. Small sample size limitation. We are limited to the 10 countries in our analysis and an ideal sample would be 50 and above to conduct a detailed statistical analysis and run a regression.
5. The small sample size limitation also applies to machine learning, meaning that we could not perform machine leaning within our study, as a slightly larger sample of 16 and above would be required to run successfully.
6. Availability and conflicting schedules, as everyone had differing availability so often we weren't able to meet altogether as a group during the project. However we used Doodle to determine availability to choose the best time slots for group meetings, as well as meeting after the daily sessions while we were already present.

# Result Reporting

Mean adult female literacy rates between 2000-2021 range from 15.38% (Chad) to 99.99% (Korea, Dem. People’s Rep.). The countries with the bottom five literacy rates are Chad, Afghanistan, Mali, Niger, and Guinea. The countries with the top five literacy rates are Cuba, Lithuania, Estonia, Latvia, and Korea, Dem. People’s Rep. Our dependent variable is adult female literacy rate, and our independent variables are: income level (split into low income and upper middle income), median female child marriage rate (%), mean government education expenditure rate (%), and mean GPI.

When each independent variable is considered individually in the Ordinary Least Squares (OLS) regression analysis, they do not tend to have a statistically significant impact on adult female literacy rates. Median female child marriage (%) is the anomaly, with a *p*-value of 0.07 which is below our chosen *p-*value of 0.1. However, when considering the combination of the non-significant variables, they collectively explain a large portion of the variation in the dependent variable, shown by the high adjusted R-squared value (0.75). This means that approximately 75% of the variability in the dependent variable is explained by the independent variables in the model. This indicates a strong relationship between the independent variables and the dependent variable, and suggests that the model is able to capture a significant portion of the variation in the data.

In simpler terms, although individually the factors didn’t show strong associations, when taken as a whole, they had a meaningful influence on adult female literacy. The only factor that stood out as important for adult female literacy was the rate of child marriages among girls. But when we looked at all the factors combined, they showed a significant overall effect on adult female literacy.

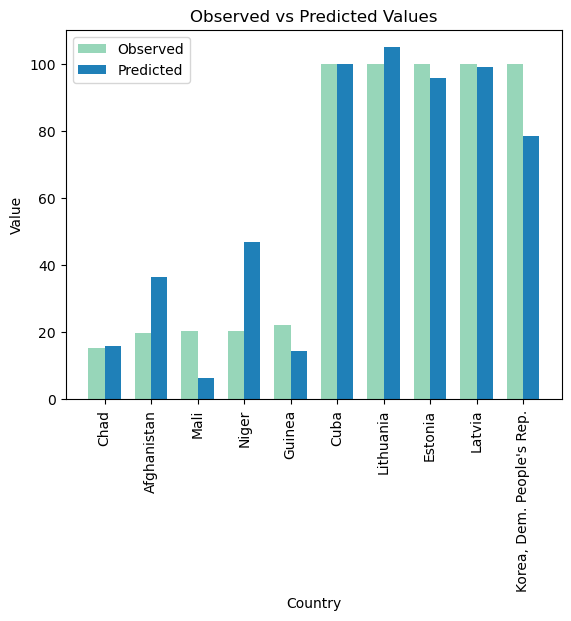
As shown in Figure 2 below, the pattern is captured relatively well, meaning that the model provides some useful insights. However, the increased volatility observed in the deviations between predicted and actual values further highlights the complexity of the relationship between the independent variables and the dependent variable. While the model provides valuable insights and predictions, there are certain cases where it may fall short in accurately estimating the adult female literacy rate. It is essential to interpret the model’s predictions with caution, considering the limitations and potential discrepancies it may have. Further research and analysis are needed to enhance the model’s performance and address the identified deviations. 

Figure 2 : The observed vs predicted values for each country extracted from the model.

Additionally, the correlation heatmap below (Figure 3) shows the relationship between adult female literacy rate and the other chosen variables, using the Pearson’s correlation coefficient (r) statistic which measures the linear correlation between two variables. The values within the heatmap represent this r-value; the closer this value is to zero, the less correlated the two variables are. Income level (upper middle income) has a moderate positive correlation with adult female literacy rate (r = 0.33). Income level (low income), mean GPI, and median female child marriage rate (%) have a strong negative correlation with adult female literacy rate, with an r-value of -0.82, -0.54, and -0.84 respectively ([source for correlation strength interpretation](https://www.scribbr.com/statistics/pearson-correlation-coefficient/)). This means that with increasing adult female literacy rate, mean GPI and median child marriage rates decrease, and less countries with low income have higher adult female literacy rates. As child marriage rate was also individually significant in our OLS regression analysis, the strong correlation between child marriage rate and adult female literacy rate reinforces the importance of addressing child marriage rates in order to promote higher literacy rates among adult women. Mean government education expenditure rate (%) is the least correlated with adult female literacy rate, with a r-value of 0.07 (indicating a very weak positive correlation). This is because the mean government education expenditure rate for all 10 countries go from 12% to 16.7% which means they're all spending similar amounts on education. Although this initially does not seem significant, the total amount spent will vary depending on their income classification (i.e., a low income country with GPI of less than $1045 will have less money to invest compared to high income countries), which is not captured in the correlation.

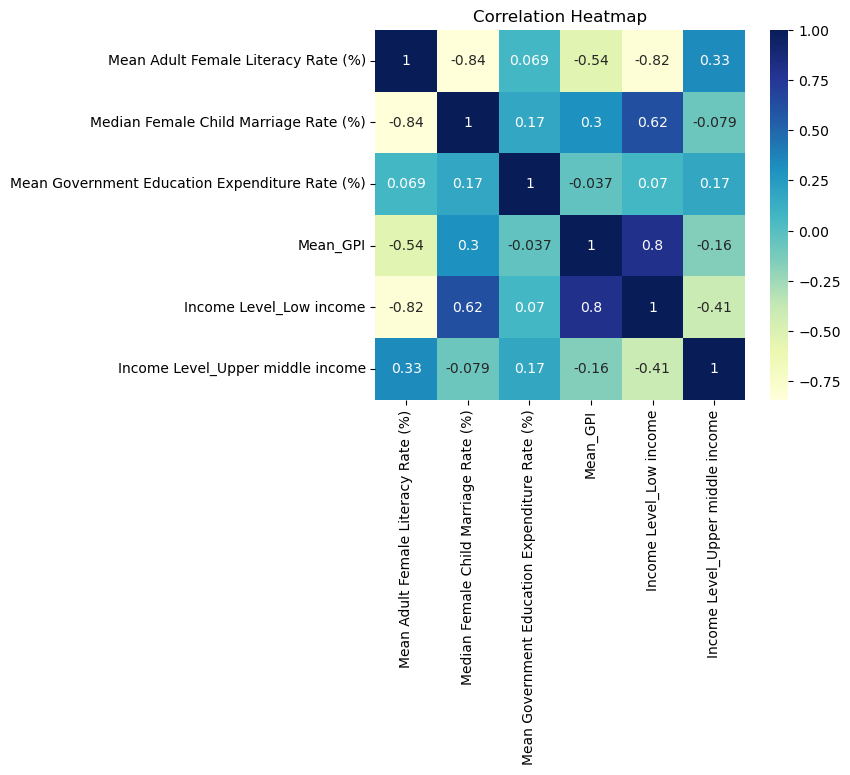


Figure 3: Correlation heatmap showing the correlation between each independent variable and the dependent variable (mean adult female literacy rate (%)).

# Conclusion

Based on the analysis, it is evident that median female child marriage rates play a significant role in determining female literacy rates. To address the issue of low female literacy rates, it is crucial for governments and charities to focus on reducing child marriage rates and promoting access to education for girls. Investments in educational infrastructure, awareness campaigns, and policies aimed at reducing gender inequality can help improve female literacy rates globally. Although it is important to note that the sample size in our project was too small for the regression, making this one limitation of our model. To conclude, although individually the factors didn’t show strong associations, when taken as a whole, they had influence on adult female literacy.

In terms of our experience on this project, our project gained us valuable insights into the topic and showed each of our strengths. We successfully explored various factors contributing to adult female literacy rates as well as refining the topic of our project to complete a successful and accurate analysis. One of our key strengths was our collaborative approach as we effectively used our diverse skills, knowledge and perspectives to conduct a comprehensive project and reach valuable conclusions. Our strength of being able to work cohesively as a team was key in enabling us to face the complex challenges mentioned earlier and generate solutions and a thorough understanding of the factors affecting female literacy. By determining a clear project timeline and structure we were able to complete every task and strengthened the validity of results as we made sure to check for any issues in data sets and conduct extensive research. Retrospectively, one area of potential improvement could have been adding additional variables such as demographic variables like population density or regional disparities to make this a broader perspective and analysis. This potentially could have strengthened our findings but ultimately a lack of updated information and data sets were what limited us from expanding this scope in the first place. We can confidently say we enjoyed this group project as we chose a meaningful topic and enjoyed gathering data and visualising it as a result.