STAT 250 PROJECT REPORT

ANALYZING THE EFFECT OF VARIOUS FACTORS ON INTERNATIONAL STUDENT MOBILITY

PRESENTED BY:

AIARU ABDIRAKHMAN 2603553
NURAY TAGHIYEVA 2603264
HADIYA KHAN 2602084

CONTENTS

ABSTRACT

This project aims to analyze international student mobility in tertiary education and the effect of factors like education expenditure and literacy rates on it. R-studio was employed to perform all the analysis and generate graphs and figures in this report. Firstly, in order to learn more about the overall structure of the data, exploratory data analysis was done. Then basic graphs and figures to visualize any possible relations between variables were generated. To further analyze and form a conclusion about these relationships, hypothesis testing was performed.

INTRODUCTION

Trends in International student mobility have been continuously shifting, especially in the past decade. Some countries like the US and Türkiye have come out to be growing hotspots for international students. We attempt to analyze whether the GDP expenditure on education and the literacy rates of a country affect this trend.

METHODOLOGY

1. Dateset Overview

The data used for this project was retrieved from multiple sources on the web. Data for international students based on country of origin was taken from the OECD database[1]. Data for education spending of each country as a percentage of the GDP was also taken from OECD[2]. Literacy rates for each country along with male and female literacy rates was obtained from UNICEF[3]. These datasets were merged with respect to the country column using R codes. Something to be noted is that since this is real-life data, there are null values in it. However, null values in the literacy rates column were later filled using data from The World Bank[4]. Another column containing the mean number of students from 2013-2021 was added. The final data obtained for this project has the following variables:

- **"Continent":** The continent to which the country belongs.
- **"Country":** The name of the country.
- **"2013-2021":** Yearly data of total number of students for each country.
- **"Mean_Students_Total"**: The mean number of students between 2013-2021.
- **"Exp.Year":** Expenditure year.
- **"Economy.code":** The economic code for the country.
- **"GDP":** Expenditure as a percentage of GDP.
- **"Total_literacy_rate":** The total literacy rate.
- **"Female_literacy":** The literacy rate among females.
- **"Male_literacy":** The literacy rate among males.

2. Descriptive Statistics

- Table 1 presents descriptive statistics for the number of international students from 2013 to 2021, including minimum, 1st quartile, median, mean, 3rd quartile, and maximum values.

Statistic	2013	2014	2015	2016	2017	2018	2019	2020	2021
Minimum	0	0	0	0	0	0	0	0	0
1st Quartile	6	0	0	0	0	0	0	0	0
Median	37	34	37	41	35	40	44	43	33
Mean	1182	1247	1378	1570	1764	2058	2345	2104	1720
3rd Quartile	368	356	414	439	511	469	478	453	340
Maximum	87980	90245	97387	112329	128498	143323	155594	128183	93437

Table 1 "Descriptive Statistics Summary for Student Data (2013-2021)"

According to Table 1, over the years, the minimum number of international students remained at zero, indicating some countries reported no international students. The 1st quartile also remained consistently at zero, showing that at least 25% of countries had no outgoing international students. The median values ranged from 33 to 44, while the mean number of students increased from 1182 in 2013 to a peak of 2345 in 2019, before dropping to 1720 in 2021. The 3rd quartile showed a similar trend, peaking at 511 in 2017 and decreasing to 340 by 2021. The maximum number of international students significantly varied, from 87980 in 2013 to a high of 155594 in 2019, then dropping to 93437 in 2021. This indicates a substantial growth in the number of international students until 2019, followed by a decline likely due to global events, one of which is COVID-19, impacting student mobility.

- Table 2 shows the descriptive statistics summary for recent literacy data across various countries, showing metrics for total literacy rate, female literacy, and male literacy.

Statistic	Total_literacy_rate	Female_literacy	Male_literacy
Minimum	22.30	14.00	31.30
1st Quartile	87.05	85.80	88.50
Median	98.00	98.00	98.00
Mean	89.39	87.77	91.03
3rd Quartile	99.00	99.00	99.00
Maximum	100	100	100

Table 2 "Descriptive Statistics Summary for Literacy Data"

Based on Table 2, the descriptive statistics summary for recent literacy data across various countries reveals that the minimum literacy rates are 22.30% for total literacy, 14% for female literacy, and 31.30% for male literacy. The 1st quartile values indicate that at least 25% of countries have literacy rates at or above 87.05% for total literacy, 85.80% for female literacy, and 88.50% for male literacy. The median values for total, female, and male literacy are all 98%, showing a high central tendency across the board. The mean literacy rates are 89.39% for total literacy, 87.77% for female literacy, and 91.03% for male literacy, reflecting slightly higher literacy rates for males. The 3rd quartile values are 99% for all categories, indicating that 75% of countries have literacy rates below this level. The maximum literacy rate is 100% for both genders and total literacy, suggesting some countries have achieved complete literacy. The data overall shows that global literacy rates are high, with minor gender differences that tend to favor males.

- Table 3 demonstrates the correlation of continuous variables, including Mean_Students_Total, GDP, Total_literacy_rate, Female_literacy, Male_literacy.

	Mean_Students_Total	**GDP**	**Total_literacy_rate**	**Female_literacy**	**Male_literacy**
Mean_Students_Total	1	-0.071401	0.098952	0.088584	0.106651
GDP	-0.071401	1	0.113423	0.110821	0.117673
Total_literacy_rate	0.098952	0.113423	1	0.995194	0.991865
Female_literacy	0.088584	0.110821	0.995194	1	0.975735
Male_literacy	0.106651	0.117673	0.991865	0.975735	1

Table 3 "Correlation matrix for continuous variables"

From Table 3, it can be concluded that Mean Students Total has a slight positive correlation with Total Literacy Rate (0.098952), Female Literacy (0.088584), and Male Literacy (0.106651), but a weak negative correlation with GDP (-0.071401). GDP shows a small positive correlation with Total Literacy Rate (0.113423), Female Literacy (0.110821), and Male Literacy (0.117673). The Total Literacy Rate is highly positively correlated with both Female Literacy (0.995194) and Male Literacy (0.991865), indicating strong relationships among these literacy variables. Similarly, Female Literacy and Male Literacy are also strongly positively correlated (0.975735), showcasing the close relationship between literacy rates among different genders. Overall, the table highlights the interconnectedness of these variables, particularly the strong associations between different literacy measures.

ANALYSIS AND RESULTS

Research questions:

1. ***Which continent is the most common place of origin for international students worldwide?***
2. ***What are the key differences in literacy rate distributions between females and males across different countries, and what factors contribute to the observed disparities?***
3. ***How do literacy rates vary across different continents?***
4. ***Is there a linear relationship between the number of internationally mobile students from African countries and the country's education expenditure?***
5. ***Are the mean number of internationally mobile students from Europe and Asia the same?***
6. ***What are the main factors that influence the number of international students that, between 2017 and 2021, choose to study across different nations? More specifically, what is the impact of the continent, overall literacy rate, and education spending as a % of GDP on the number of overseas students?***
7. ***How does the number of international students be distributed over different continents from 2017 to 2021?***
8. ***Is the proportion of international students from the UK and the US the same?***
9. ***Are more than 50% of the international students from Asia in 2021?***
10. ***What is the mean number of international students from Europe in 2021? Is it more than or less than 1000?***

1. Explarotary Data Analysis

Research question 1: Which continent is the most common place of origin for international students worldwide?

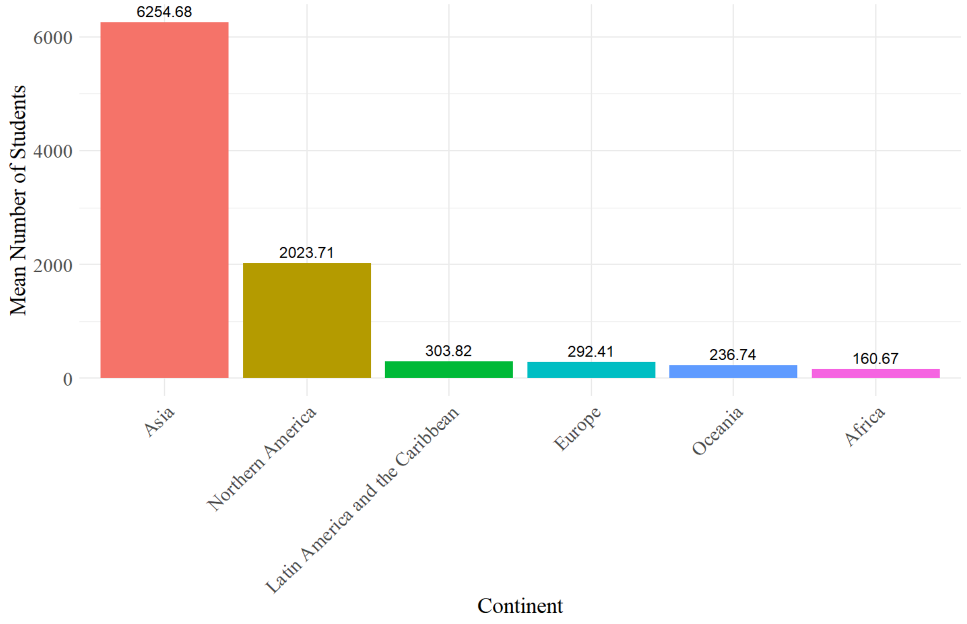

Figure 1 "Origin of international students by continent"

This bar chart(Figure 1) illustrates the mean number of international students originating from different continents. Asia stands out as the leading continent, with an average of 6254.68 students, significantly higher than any other region. Northern America follows with 2023.71 students, while other continents like Latin America and the Caribbean, Europe, Oceania, and Africa have considerably lower averages, all under 400 students. This indicates that a majority of international students worldwide come from Asia.

Research question 2: What are the key differences in literacy rate distributions between females and males across different countries, and what factors contribute to the observed disparities?

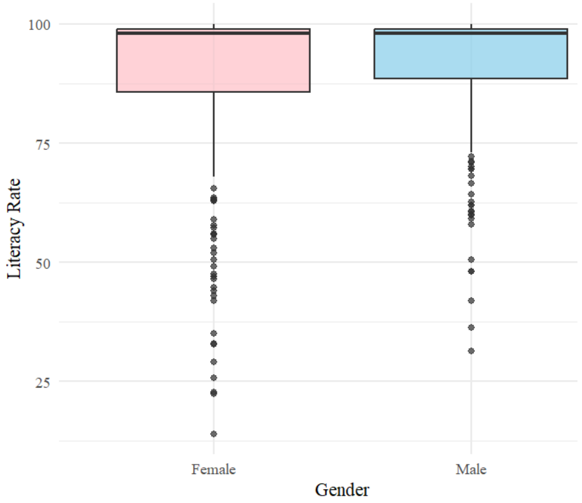

Figure 2 "Literacy rate by Gender"

The box plot (Figure 2) comparing literacy rates by gender shows that both males and females have high median literacy rates close to 100%, indicating that the majority of countries maintain high literacy rates for both genders. The interquartile range (IQR) is similar for both, suggesting comparable variability. However, there are more outliers on the lower end for females than for males, indicating that there are more countries with significantly lower literacy rates for females. This highlights the need to address gender-specific disparities in literacy rates in certain countries.

Research question 3: How do literacy rates vary across different continents?

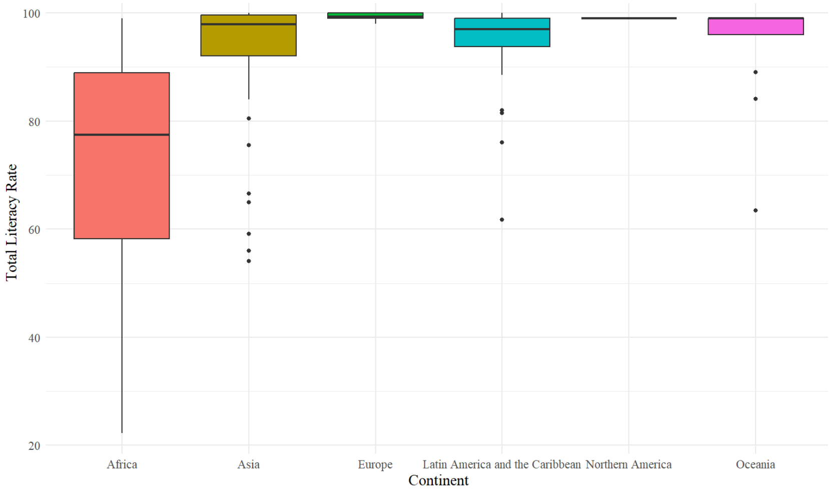

Figure 3 "Literacy rate by Continent"

The boxplot (Figure 3) displays the variation in literacy rates across different continents. Africa shows the most significant variability with a median literacy rate around 70% and a wide range from about 30% to nearly 100%, indicating diverse educational outcomes across the continent. Asia and Latin America and the Caribbean have high median literacy rates around 95%, but Asia has more outliers indicating some countries with lower literacy rates. Europe and Northern America show very high median literacy rates close to 100% with minimal variation, highlighting generally uniform high literacy across these regions. Oceania also has a high median literacy rate, though it shows slightly more variation than Europe and Northern America.

2. Hypothesis Testing

Research question 4: Is there a linear relationship between the number of internationally mobile students from African countries and the country's education expenditure?

H_0: Number of students leaving African countries for education is independent of the country's Literacy rates.
H_1: Number of students leaving African countries for education is dependent on the country's Literacy rates.

To visualize whether a possible relationship exists, we will conduct a simple linear regression. In order to do this, let us first state our assumptions:.

1. Linearity: The relation between the variables appears to be linear.
2. Residuals are assumed to be normally distributed.
3. Residuals have a constant variance (homoscedasticity). Since the points bounce randomly across the 0 line in Residuals vs Fitted plot.
4. In the Normal Q-Q plot, we observe that quantiles mostly lie along the line. Hence, we can assume that the residuals are independent.

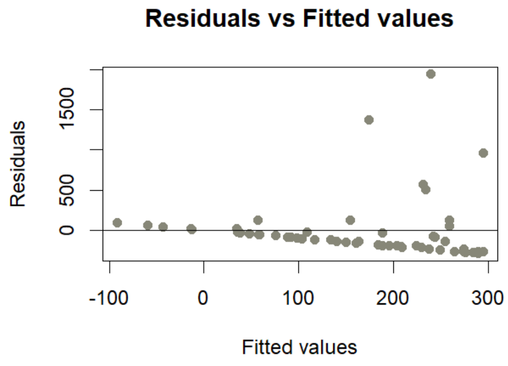

Figure 4

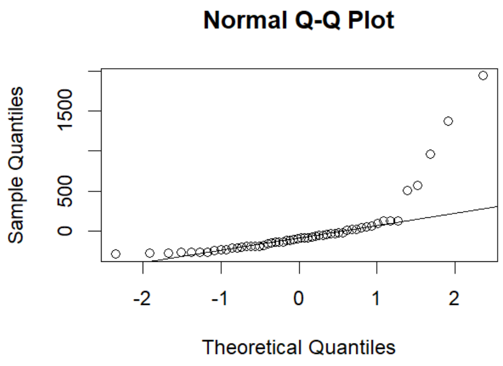

Figure 5

Proceeding with simple linear regression, we generate a linear model between the concerned variables. The equation for the fitted line is:

$$\hat{y} =- 203.854 + 5.031X$$

The p-value for this model is 0.06857, which is slightly more than 0.05. Hence, we fail to reject the null hypothesis.

This means that the number of students leaving African countries for tertiary education is not dependent on the country's literacy rates.

Research question 5: Are the mean number of internationally mobile students from Europe and Asia the same?

H_0: Mean Number of International students from Europe and Asia are the same.

H_1: Mean Number of International students from Europe and Asia are not the same.

We will conduct a test for two sample means to test our hypothesis. Firstly, we check for equality of variances. The p-value for F-test is $2.2 * 10^{-6}$ which is very less than 0.05. Hence, we reject the null hypothesis which means that the variances are unequal. We then perform a Welch two sample t-test to check equality of means. The p-value in this case comes out to be 0.02366. Since we are testing at the 95% confidence level, the alpha value is bigger than p-value in this case. So, we reject the null hypothesis. Our conclusion is that the mean number of Internationally mobile students leaving Europe are not the same as those leaving Asia. In addition, the confidence interval for the difference of means is (831.78, 11092.76). Again since the confidence interval does not include 0, we reject the null hypothesis.

Research question 6 : What are the main factors that influence the number of international students that, between 2017 and 2021, choose to study across different nations? More specifically, what is the impact of the continent, overall literacy rate, and education spending as a % of GDP on the number of overseas students?

Under certain assumptions, multiple linear regression is specialized to forecast the outcome of the response variable using a number of independent variables. Below is a description of the multiple linear regression model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_P + \varepsilon$$

For this analysis, the dependent variable is the number of international students. The independent variables in this analysis are the Total Literacy Rate, Education Expenditure as a Percentage of GDP, and Continent**.** We will use multiple linear regression to quantify the relationship between the number of international students and the selected independent variables. However, to use MLR for these questions, the required assumptions should be checked. We can state some issues by looking at a scatter plot of x and y values, and the linearity assumption is not satisfied. Shapiro-Wilk test is used to check normality, but it is not satisfied either (p-value < 2.2e-16). Homoscedasticity and independence are satisfied by the results of the Breusch-Pagan and Durbin-Watson tests (p-values are greater than 0.05). Considering the VIF values, there is no multicollinearity in the model (all of them are between 1 and 2). The transformation of variables method is used to deal with the unsatisfied assumptions. From the model results, since the p-value of the F statistic is 7.309e-08, we conclude that the overall model is significant. Using the same approach, the coefficients of the Total Literacy Rate in Asia and Europe from the continents are significant. Additionally, the model explains approximately 26.85% of the variability in the log-transformed number of international students (R-squared=0.2685). The finalized version of the model is:

$$y = -7.4773 + 2.4942x_1 + 0.1269x_2 + 3.6479x_3 + 1.7672\, x_4 + (-0.5021)x_5 + 0.5771\, x_6 + 1.6566\, x_7$$

Where:

y= Total Students, x_1= Total Literacy , x_2= Expenditure as % of GDP, x_4=Europe (dummy variable: 1 if Europe, 0 otherwise), x_5=Latin America and the Caribbean (dummy variable: 1 if Latin America and the Caribbean, 0 otherwise), x_6=Northern America (dummy variable: 1 if Northern America, 0 otherwise), x_7=Oceania (dummy variable: 1 if Oceania, 0 otherwise)

To sum up, the multiple linear regression model effectively predicts the number of international students using the processed data. Significant factors influencing the number of international students include literacy rates (marginally significant) and being located in Asia or Europe. Education expenditure as a percentage of GDP, Latin America and the Caribbean, Northern America, and Oceania are not significant predictors in this model.

Research question 7 : How does the number of international students be distributed over different continents from 2017 to 2021?

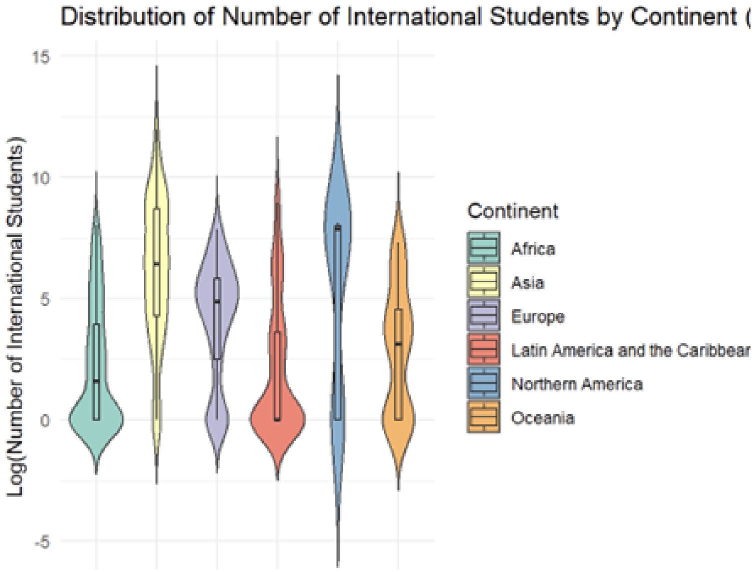

Figure 6 "Distribution of Number of International Students by Continent"

The data is transformed into a scale that reduces skewness and makes it easier to model and interpret by applying a log transformation. Each violin plot(Figure 6) represents the density of a continent's log-transformed number of international students, with the box plot inside. The violin plot shows that Asia has a wide distribution, which shows high variability in the number of international students among its countries, with a higher median value than other continents in the data. Europe and Northern America show wide variability with more concentrated distributions around their medians. In contrast, Africa, Latin America, and the Caribbean have narrower distributions.

Since the continent variable has more than two levels, ANOVA can be conducted to determine whether the levels of the continent variable have at least one different effect on the number of international students.

Hypothesis:

Null Hypothesis (H_0:): There is no significant difference in the number of international students across different continents.

Alternative Hypothesis (H_1: **)**: There is a significant difference in the number of international students across at least one pair of continents.

Required assumptions to conduct ANOVA are checked. Levene's test satisfies the homogeneity of variance assumption (p-value=0.07886). However, the result of the Shapiro-Wilk test shows that the residuals are not normally distributed(p-value=4143e-13). By using Box-Cox Transformation, the normality assumption is satisfied.

	Df	Sum Sq	Mean Sq	F value	P value
Continents	5	2751	550.2	75.79	<2e-16
Res.	1039	7542	7.3		
Total	1044				

Table 4 "ANOVA"

The ANOVA results(Table 4) show that the continent statistically affects the number of international students (p-value < 2e-16). This means that at least one of the continents has a different average log-transformed number of international students compared to the others. Given the significant F-value and very small p-value, we reject the null hypothesis that the means of the log-transformed number of international students are the same across all continents. The next step is using Tukey's Honest Significant Difference (HSD) test to make pairwise comparisons between the continents. According to this test's results, Asia, Europe, and Northern America have significantly higher numbers of international students than Africa. Latin America and the Caribbean have significantly fewer international students compared to Asia. Oceania differs significantly from Europe and Northern America. These findings indicate that the continent is a significant factor in the distribution of international students from 2017 to 2021.

<u>*Research question 8: Is the proportion of international students from the UK and the US the same?*</u>

H_0: Proportional of International students from the US and the UK is the same in 2021

H_1: Proportional of International students from the US and the UK is not the same in 2021

To conduct this hypothesis testing, we do a two sample t-test for proportions with continuity corrections. The p-value for this comes out to be $2.2 * 10^{-6}$ which is way less than 0.05. Hence, we reject the null hypothesis and conclude that the proportion of International students from the US and the UK was not the same in 2021. The confidence interval calculated for the proportions is (0.048, 0.064). Since this confidence interval does not include 0, our conclusion is correct.

Research question 9 : Are more than 50% of the international students from Asia in 2021?

H_0: Proportional of International students from Asia in 2021 is less than or equal to 0.5

H_1: Proportional of International students from Asia 2021 is greater than 0.5

To test our hypothesis, we employ a one sample proportion test to check whether the proportion of students from Asia is more or less than 50%. The results of the test give a p-value of less than $2.2 * 10^{-6}$. Hence, we reject the null hypothesis, meaning, the proportion of students from Asia in 2021 is more than 0.5. The test coupled with the sample proportion of 0.866 verifies our claim. This tells us that more than half of the international students were from Asia in the year 2021.

Research question 10 : What is the mean number of international students from Europe in 2021?   Is it more than or less than 1000?

H_0: Mean number of international students from Europe in 2021 is equal to 1000

H_1: Mean number of international students from Europe in 2021 is less than 1000

To conduct a one sample mean test, we first check our assumptions of normality.

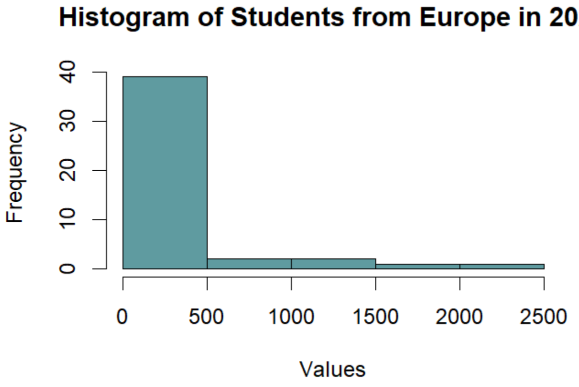

Figure 7

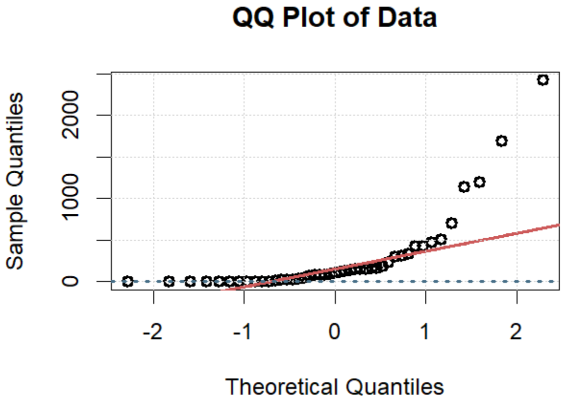

Figure 8

As we can see from the histogram above(Figure 7), the data is highly skewed to the right. The QQ plot(Figure 8) also shows the same thing, that the data is not normally distributed. Shapiro-Wilk normality test was also applied, and the p-value comes to be less than $5.4 * 10^{-10}$ which is way less than 0.05. Hence, we conclude that the data is not normally distributed. However, the data has 45 observations, so we can assume normality by Central Limit Theorem (CLT).

Since we do not know the population standard deviation, we use t-test. The p-value for which is $1.85 * 10^{-13}$ which is very small. We reject the null hypothesis and conclude that the mean number of students leaving Europe for tertiary education is less than 1000. This makes sense in the fact that most of the highly recognized educational institutions are situated in Europe and not a lot of students from Europe move outside of the continent for tertiary education.

CONCLUSION

Our project of analysis on international student mobility reveals some entrancing insights into what drives students to study abroad and the factors that influence these decisions. By examining data from the OECD, UNICEF, and The World Bank, we focused on education expenditure, literacy rates, and the students' continents of origin. The following essentinal points are highlighted:

1. Asia's Dominance: Asia stands out as the primary source of international students, far exceeding other regions. This highlights Asia's significant role in global education and student exchange programs.

2. Gender Literacy Disparities: Although literacy rates worldwide are high, there are noticeable differences between genders in some countries. Female literacy rates show more variability and lower outliers compared to males, pointing to areas where educational policies could be improved to support gender equality.

3. Continental Literacy Variations: Literacy rates differ widely across continents. Africa shows the most considerable variation, with some countries having very low rates and others quite high. In contrast, Europe and Northern America have uniformly high literacy rates, indicating more consistent educational outcomes.

4. Key Influencing Factors: Our regression analysis found that a country's total literacy rate and whether it is located in Asia or Europe significantly impact the number of its students studying abroad. Surprisingly, the percentage of GDP spent on education did not have a significant direct effect.

5. Asia's Majority: More than half of the international students in 2021 came from Asia, reaffirming the region's prominence in global student mobility. Additionally, there is a clear difference in the proportion of international students coming from the US and the UK, reflecting varying trends in student mobility from these countries.

6. European Student Mobility: We found that the average number of international students from Europe in 2021 is less than 1000. This makes sense considering many top educational institutions are located in Europe, reducing the need for students to seek education elsewhere.

In summary, our project highlights the critical role of literacy rates and geographic origin in shaping international student mobility. While education expenditure is essential, its impact appears less direct compared to other factors. These insights can help policymakers and educational institutions better understand and support international student trends. Future research could explore additional factors and long-term trends to provide a deeper understanding of global student mobility.

REFERENCES

1. OECD. (2023). Enrolment of international students by country of origin. OECD. Retrieved on April 14, 2024, from https://stats.oecd.org/Index.aspx?DataSetCode=EDU_ENRL_MOBILE
2. OECD. (2024). Public spending on education (indicator). doi: 10.1787/f99b45d0-en. Retrieved on May 5, 2024 https://doi.org/10.1787/f99b45d0-en
3. UNICEF. (2019). Education and literacy data. UNICEF. Retrieved on May 21, 2024 from https://data.unicef.org/resources/dataset/education-literacy-data/
4. World Bank. (2024). Literacy rate, adult total (% of people ages 15 and above). World Bank. Retrieved on May 21, 2024 from https://data.worldbank.org/indicator/SE.ADT.LITR.ZS