

Educational Attainment Level: Examining Its Influences from an International Lens

Tulsi Shrivastava
Data Analytics Engineering
George Mason University
Fairfax, VA
tshrivas@gmu.edu

Abstract— *It is widely known that education holds the power to transform lives for a population. This research uses various data analytics methods and tools to examine the state and influence of the different educational attainment levels among different populations around the world. My primary aim was to identify and understand the relationship between educational attainment, government expenditure, and employment rates – as given in exports from the Organization for Economic Co-operation and Development (OECD). Additionally, I analyzed the data from a lens of gender parity. Results from the study both contradicted and fulfilled previously held hypotheses about the data. With more depth and time, furthering this research may yield key public policy insights.*

Keywords—education, employment, international, gender parity

I. INTRODUCTION

When an individual is educated – they carry that learning forward to better their families, neighborhoods, and communities – eventually transforming the entire well-being of a country. My primary research questions examine the role that level of education plays in determining positive outcomes. Conducting this analysis could potentially influence policy decision-makers who are on the fence about educational investments. If we can analyze the impact of higher education and identify roadblocks and gaps obstructing academic progress, that insight gained can be used to make key public policy decisions in governments. Lastly, comparing the data on an international level may be able to add context and highlight additional factors (culture, religion, systemic) that play a significant role in shaping education.

My interest in this field stems from my previous work experience as BI Analyst working with a variety of schools across the country. Given a front row seat to education in America at Learning Sciences International, I learned of the barriers that prevent children from accessing quality instruction and saw the transformative power of academics in framing positive outcomes into adulthood. In addition, as a woman of color, I’m particularly interested in studying the learning gaps that exist between males and females – which is an area of focus I hope to contribute to as a professional.

II. RESEARCH QUESTIONS

My first research question asks how does level of education (below upper secondary, upper secondary/ postsecondary, and tertiary) effect employment rates. Secondly, how does government expenditure in education affect the economic viability of a country? Lastly, which countries exhibit the largest gaps in gender equality, and what can we learn from the nations which have a more balanced playing field?

III. DATA SELECTION

A. Data Source

The datasets selected for this project were pulled from the Organization of Economic Co-operation and Development (OECD), which is an organization of several countries that helps to find solutions to a range of social, economic, and environmental challenges. OECD houses a large data warehouse with information on topics ranging from life satisfaction, economic status, health, and education.

B. Types of Data & Definitions

I utilized two primary exports from the OECD database, which I cleaned and renamed to “**edecon**” (relating to the employment rate and expenditure) and “**edgen**” (relating to gender parity values).

Table 1 below explains some commonly used words in the dataset.

Term	Definition
Below Upper Secondary	This indicates the range of school that is between less than primary school education and prior to high school.
Upper secondary	This indicates completion of high school or post-high-school program.
Tertiary	This type of education is considered a higher education degree: bachelors, masters, or doctorate.
Expenditure	Expenditure is the amount of money a government invests in public institutions, given in terms of USD Purchasing Power Parity (UPP) in thousands.

Table 1: Terms and Definitions

IV. LITERATURE REVIEW

To begin with, I explored the studies published by the OECD itself about the international educational arena. Their “Education at a Glance” provides summary tables and an analysis of education systems across the OECD countries and partner economies. In particular, they examined how educational attainment affects participation in the labor market and found that on average, “Employment rates for tertiary-educated young adults (25-34 year-olds) are 8 percentage points higher than those who have attained upper secondary or post-secondary non-tertiary education and 26 percentage points higher than those who have attained below upper secondary education across OECD countries” [2].

A 2017 article published in the International Journal of Public Health explored the effect of low education on

unemployment. This directly ties in with my first research question and the OECD study surrounding the labor market. Researchers used survey data from almost 70,000 respondents regarding education and if they were working less than 12 hours per week. They found a significant relationship between low education and poor health – both which tend to further aggravate the individual's unemployment status. This conclusion is one I also expect in my own analysis: lower educational attainment levels result in lower levels of employment across the board [3].

Lastly, an article published in the Feminist Economics peer-reviewed journal covers the topic of gender parity, which ties back to my second research question. Variables used to measure economic growth included the female-male education level ratios, GDP, population, and age. As expected, research concluded that gender gaps in education reduce economic growth. Additionally, even if gender gaps in education slowly decreased in some countries, the effect of that did not extend itself to show in employment data – indicating there is still gaps in research [1].

V. METHODOLOGY

A. Data Cleaning in Excel and R

The data exported from the OECD website required data cleanup. After converting the files to CSV and removing extra stylistic formatting using Excel, I pulled the data into RStudio. Using R, I first checked for null values and removed them from the dataset if I found that they would not be helpful. Also, with this particular type of data, it is dangerous to simply “guess” the values – because each country could vary greatly. For example, I did not want to simply put the mean value for public school expenditure by the government. However, in the case of some variables, the OECD provided the average value – so I used that value to fill a few nulls.

Additional work needed was renaming the column values, adding an index value to insert into a database, and preparing the datasets for modelling. A snapshot of the cleaning techniques used in R is shown in the picture below.

```
# Add index
edecon$index <- 1:nrow(edecon)

# Remove & Rename Columns
edecon <- edecon[, -7]
colnames(edecon)[2] = "lessthanprimary"
colnames(edecon)[3] = "primary"
colnames(edecon)[4] = "lowersecondary"
colnames(edecon)[5] = "uppersecondary"
colnames(edecon)[6] = "postsec-nontert"
colnames(edecon)[7] = "shorttert"
colnames(edecon)[8] = "bach"
colnames(edecon)[9] = "master"
colnames(edecon)[10] = "doctor"

# Average for 2022
edecon$avgemprate <- rowMeans(edecon[, c(12:15)], na.rm=TRUE)

##### CLEANING #####

# Create Subset for Modeling
edecon2 <- edecon[, -(12:15)]
edecon2 <- na.omit(edecon2)
summary(edecon2)
```

The output of this step was two clean datasets: “**edecon**” (relating to the employment rate and expenditure) and “**edgen**” (relating to gender parity values).

B. Exploratory Analysis and Regression in R

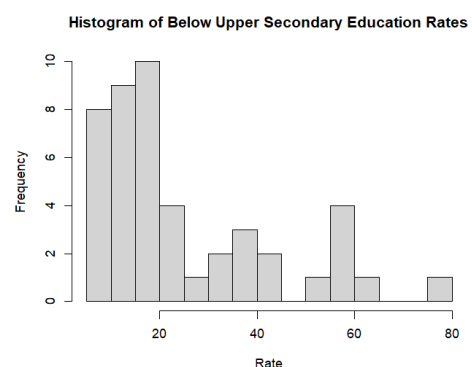
I began my study in R by analyzing the variables in the **edecon** dataset. Below is a sample look at the data. Country is

a nominal data value, representing 46 different countries. The independent variables used were the varying levels of education: variables 3, 4, 5, 7, 8, 9, 11, 12, and 13. Expenditure was also used as an independent variable. Average Employment Rate was calculated for 2022 and used as a dependent variable.

```
'data.frame': 46 obs. of 20 variables:
 $ Country      : chr  "Australia" "Austria" "Belgium"
 $ totalbelowupper : num  15.47 14.06 18.47 6.93 28.1
 $ lessthanprimary : num  0.287 1.947 2.538 2.538 1.
 $ primary       : num  3.61 1.13 3.86 1.74 3.79
 $ lowersecondary : num  11.57 12.93 12.07 5.19 11.
 $ totaluppersecond : num  34.8 51.3 36.7 31.1 40.5
 $ uppersecondary : num  29 48.5 34.9 21.3 40.5
 $ postsec-nontert : num  5.77 2.87 1.76 9.78 5.77
 $ totaltert     : num  49.8 34.6 44.9 62 31.4
 $ shorttert     : num  11.202 14.988 0.833 26.31
 $ bach         : num  28.12 4.91 24.16 24.17 11.
 $ master       : num  8.81 13.63 19.02 11.42 2.
 $ doctor       : num  1.631 1.071 0.853 1.251
 $ Expenditure   : num  10309 1454 1592 7765 196
 $ Q1.2022      : num  76.7 73.9 66.5 75.3 61.5
 $ Q2.2022      : num  77.2 74.4 66.1 75.8 62.2
 $ Q3.2022      : num  77.5 73.8 66.8 75.5 62.1
 $ Q4.2022      : num  77.5 73.9 66.7 75.8 61.9
 $ index        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ avgemprate   : num  77.2 74 66.5 75.6 61.9 ..
```

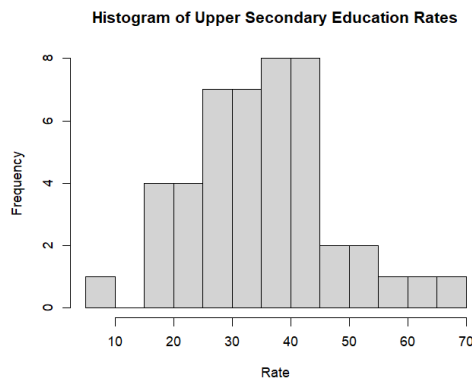
Next, we will further explore the interval variables for educational attainment level (totalbelowupper, uppersecondary, and tertiary). I plotted a histogram to examine the **totalbelowupper** variable (percent of the population that has completed Below Upper Secondary education) and found that it is slightly skewed to the left side. Once again, it is worth noting that the data is not highly-right skewed, as I initially expected. This indicates that there are still very low rates of basic education attainment among many countries. In fact, a minimum value of 5.58% was quite startling to me and shows that there is either not enough data collected or potentially a larger issue at hand for that country.

```
summary(edecon_orig$totalbelowupper)
Min. 1st Qu. Median Mean 3rd Qu. Max.
5.58 12.19 18.38 24.88 35.46 77.65
```



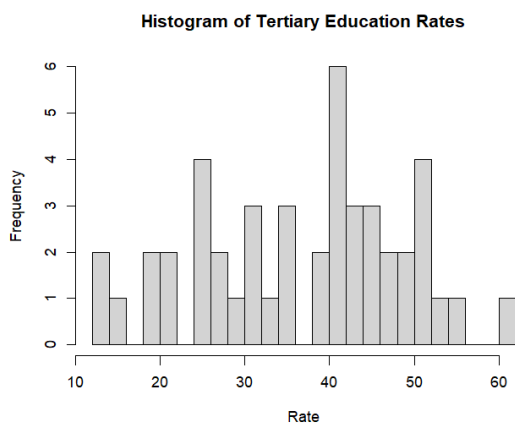
Next, I plotted a histogram to examine the **uppersecondary** variable (percent of the population that has completed Upper Secondary education). The chart below displays a normal distribution, which is not odd for this dataset.

```
summary(edecon_orig$uppersecondary)
Min. 1st Qu. Median Mean 3rd Qu. Max.
8.188 28.161 35.597 35.069 41.287 67.883
```



Below, I plotted a histogram to examine the **totaltert** variable (percent of the population that has completed Tertiary education). Like the previous graph, the chart below displays a fairly normal distribution as well.

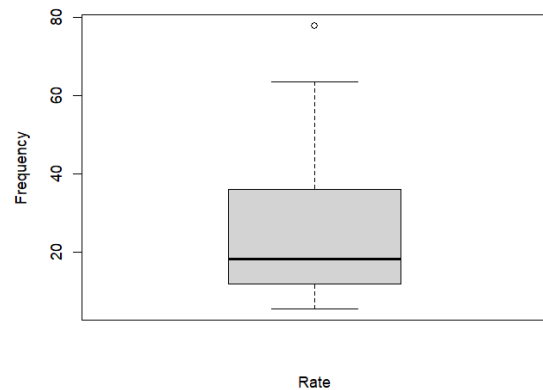
```
summary(edecon_orig$totaltert)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 12.95  26.88  40.01  36.73  45.26  61.99
```



By examining the boxplot and summary stats for average employment rate for 2022, we notice a highly left skewed distribution and a potential outlier close to 80% employment rate. To resolve this issue, we eliminated the value when running our models. In this case, it would be helpful to know how the OECD is calculating its labor force participation statistics and what constitutes the average value.

```
summary(edecon_orig$avgemprate)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
 38.56  67.34  73.13  70.68  76.60  83.26     7
```

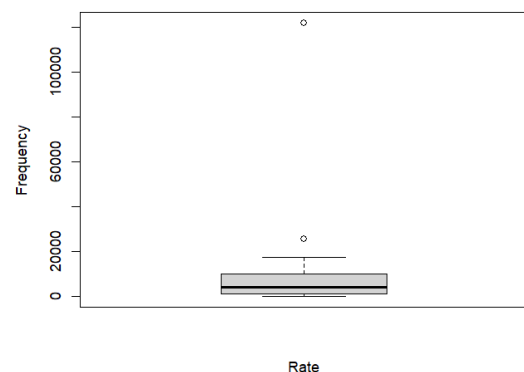
Boxplot of Avg. Employment Rates



Lastly, we look at the boxplot and summary stats for Expenditure (a ratio variable) and notice a very left skewed distribution and an outlier. After further exploration, we found that it is the United States with over \$100,000 (in thousands) in expenditure value and removed the value in order to ensure our dataset is fair. The summary stats also show that there are nine NA values, which we will delete. As previously mentioned, it would not be beneficial to the analysis to assume these values for each country.

```
summary(edecon_orig$Expenditure)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
 62.02 1418.58 4222.68 8836.06 10056.38 121910.19     9
```

Boxplot of Expenditure



C. Summary Statistics of Education-Gender Data in Python

In order to identify the distributions and check for outliers, we check the summary statistics and histograms for the **edgen** dataset using Python in Jupyter Notebook. We see that the dataset has 516 records. Years 2011 and 2021 are both represented, allowing us to make timely comparisons.

```
RangeIndex: 516 entries, 0 to 515
Data columns (total 6 columns):
#   Column    Non-Null Count  Dtype
---  -
0   id         516 non-null    int64
1   country   516 non-null    object
2   edlevel    516 non-null    object
3   gender     516 non-null    object
4   year       516 non-null    int64
5   percent    492 non-null    float64
dtypes: float64(1), int64(2), object(3)
memory usage: 24.3+ KB
```

However, after sampling the first five rows, we find that these are not 516 distinct rows. Rather, each country consists of 12

values – broken by education level, gender, and year. This will be important to remember when querying the data, as there is potential for a double counting error.

```
1 edgen.head(5)
2
```

	id	country	edlevel	gender	year	percent
0	1	Australia	Below Upper secondary	Men	2011	78.213898
1	2	Australia	Below Upper secondary	Men	2021	70.176048
2	3	Australia	Below Upper secondary	Women	2011	49.237667
3	4	Australia	Below Upper secondary	Women	2021	46.896439
4	5	Australia	Upper secondary	Men	2011	91.693428

Next, we take a look at the edlevel variable and find that it is an ordinal data type. That is, the values in the variable follow an order (lowest to highest) that is important to understanding the dataset. We find that each country has 172 values going from below upper-secondary education, to upper-secondary, and finally to tertiary education.

```
1 print(edgen['edlevel'].describe())
2 print('\n')
3 print(edgen['edlevel'].value_counts())
```

```
count          516
unique           3
top      Below Upper secondary
freq           172
Name: edlevel, dtype: object

Below Upper secondary    172
Upper secondary         172
Tertiary                 172
Name: edlevel, dtype: int64
```

Examining the country variable, we found 43 unique countries represented in the dataset – each with an equal number of 12 rows.

```
1 edgen['country'].describe()
```

```
count          516
unique          43
top      Australia
freq           12
Name: country, dtype: object
```

Examining our third categorical variable of gender, we found 2 unique values for Men and Women represented in the dataset, with 258 values for each.

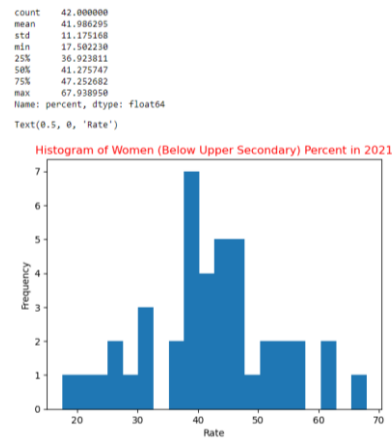
```
1 print(edgen['gender'].describe())
2 print('\n')
3 print(edgen['gender'].value_counts())
```

```
count          516
unique           2
top      Men
freq           258
Name: gender, dtype: object

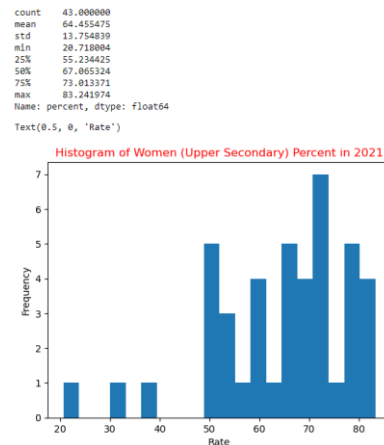
Men      258
Women    258
Name: gender, dtype: int64
```

For further analysis, I created subsets of the data in Python to focus on women in 2021. By examining the histogram for Percent of Women who completed Below Upper Secondary

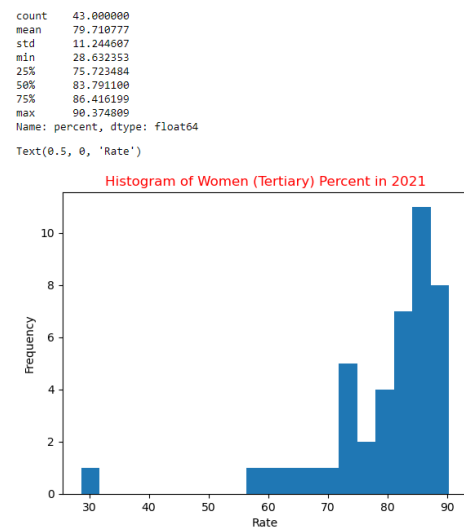
Education in 2021, we notice a normal distribution. Similar to the variable in the **edecon** dataset, it is worth noting that the data is not highly-right skewed, as I initially expected.



The histogram for Percent of Women who completed Upper Secondary Education in 2021 shows a slightly right-skewed distribution.



Finally, the histogram for Percent of Women who completed Tertiary Education in 2021 shows a very right-skewed distribution, with a potential outlier.



D. Analyzing Gender Data with SQL

The last method I used for analyzing was querying in MySQL. I created a table for **edgen** in my local database and performed queries on the data. Below is a snippet of the code used to define the dataset, followed by a sample of the first 9 rows. Further results are shown in the Results section of this paper.

```
1 CREATE TABLE `edgen` (
2   `id` int NOT NULL,
3   `country` varchar(45) CHARACTER SET utf8mb4 COLLATE utf8mb4_0900_ai_ci DEFAULT NULL,
4   `edlevel` varchar(45) CHARACTER SET utf8mb4 COLLATE utf8mb4_0900_ai_ci DEFAULT NULL,
5   `gender` varchar(45) CHARACTER SET utf8mb4 COLLATE utf8mb4_0900_ai_ci DEFAULT NULL,
6   `year` varchar(45) CHARACTER SET utf8mb4 COLLATE utf8mb4_0900_ai_ci DEFAULT NULL,
7   `percent` varchar(45) CHARACTER SET utf8mb4 COLLATE utf8mb4_0900_ai_ci DEFAULT NULL,
8   PRIMARY KEY (`id`)
9 ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb3
```

	id	country	edlevel	gender	year	percent
▶	457	Argentina	Below upper secondary	Men	2011	87.032547
	458	Argentina	Below upper secondary	Men	2021	82.685669
	459	Argentina	Below upper secondary	Women	2011	43.270638
	460	Argentina	Below upper secondary	Women	2021	40.976669
	461	Argentina	Upper secondary	Men	2011	88.044304
	462	Argentina	Upper secondary	Men	2021	80.598015
	463	Argentina	Upper secondary	Women	2011	59.071651
	464	Argentina	Upper secondary	Women	2021	58.717884
	465	Argentina	Tertiary	Men	2011	95.53331

VI. RESULTS

A. Research Question 1 & 2

To answer question 1, I created a linear model to examine which variables strongly contributed to the average employment rate. The results were not as I initially expected. My hypothesis, backed by the literature review, was that positive outcomes (increased employment rate) should increase as educational attainment went up. However, after fitting a linear model and reading the statistics, I found that only primary education completion was a significant factor in determining employment rate.

```
> fit1 <- lm(edecon3$avgemprate ~ ., edecon3)
> # Display the summary stats for the relationship
> summary(fit1)
```

```
Call:
lm(formula = edecon3$avgemprate ~ ., data = edecon3)
```

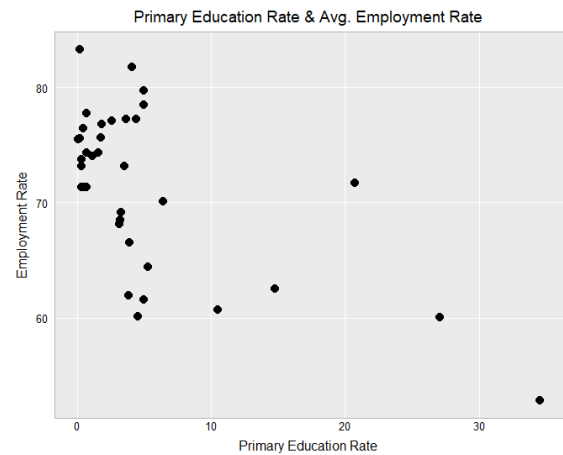
```
Residuals:
    Min       1Q   Median       3Q      Max
-7.5055 -3.6980 -0.5522  3.7769  9.8188
```

```
Coefficients:
(Intercept)      72.4277129  15.1675331  4.775 7.36e-05 ***
lessthanprimary    0.0884814   0.6586168  0.134  0.8943
primary          -0.5384946   0.2408806 -2.236  0.0349 *
lowersecondary   -0.1898084   0.1794863 -1.058  0.3008
uppersecondary   -0.0885587   0.1630009 -0.543  0.5919
'postsec-nontert' -0.0528959   0.2983700 -0.177  0.8608
shorttert        -0.0948780   0.2294338 -0.414  0.6829
bach              0.1434856   0.2035666  0.705  0.4877
master           0.2403603   0.2739117  0.878  0.3889
doctor            0.6707803   1.3095466  0.512  0.6132
Expenditure       0.0001634   0.0002167  0.754  0.4583
index             0.0254802   0.1095681  0.233  0.8181
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.889 on 24 degrees of freedom
Multiple R-squared:  0.5175,    Adjusted R-squared:  0.2964
F-statistic: 2.34 on 11 and 24 DF,  p-value: 0.03957
```

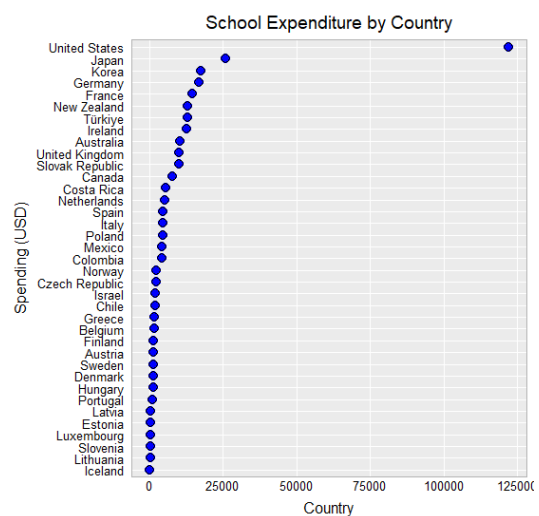
After plotting the values in the chart below, I noticed that there is a inverse relationship between the two variables. We examine the Multiple R-Squared value to understand the model's accuracy. This statistic shows that only 51.75% of the variation can be explained by the variables, which is not strong enough to assert that there is a strong correlation. There may be additional factors contributing to the employment rate, apart from simply education.



Scatterplot 1

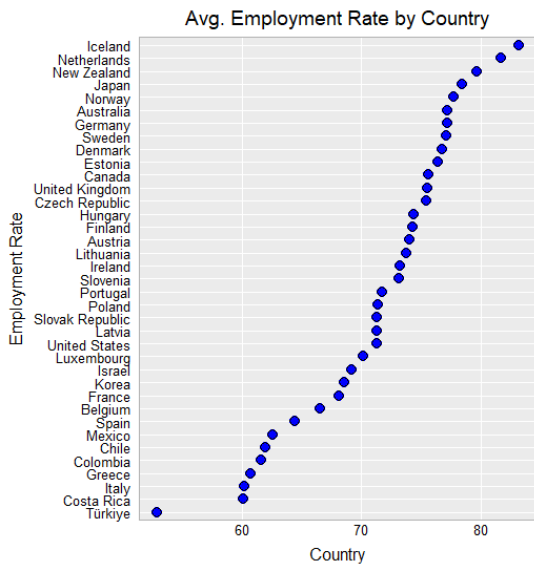
In the chart above, you can see a slight linear relationship. However, I would have expected that as the Primary Education Rate increased, the employment rate increased as well. Perhaps there is an error that I overlooked in the dataset. Thus, I will refine my model in a future study.

To study research question 2, I created two charts: one with all the countries plotted by expenditure amount and two, with all the countries plotted by average employment rate. This chart includes the outlier in Expenditure, which was identified to be the United States (as found when reviewing the summary statistics). Japan and Korea trail behind as the next highest investors in education.

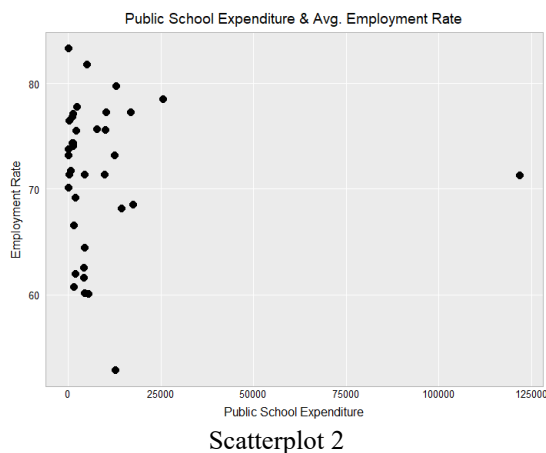


Given this view, one would expect that the U.S., Japan, and Korea also appear at the top for employment rates. However, looking at chart 2 below, this does not seem to translate into a direct correlation with the employment data

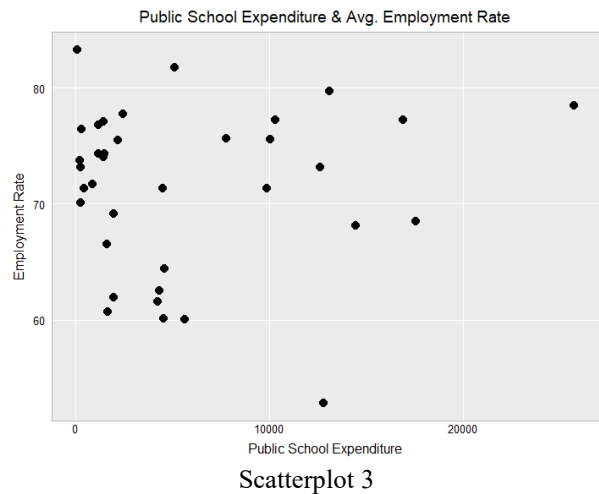
as only Japan is ahead in the Employment Rates. Surprisingly, Iceland leads the way in Average Employment Rate with Turkey trailing in last. With Iceland being near zero in school monetary investment yet at the top of employment rate, this may indicate a flaw in the dataset.



Prior to removing United States from the dataset, I plotted the relationship between School Expenditure and Employment Rate (shown below). No strong conclusions could be made with this data, so I tried again after removing the outlier.



After removing the U.S. and creating Scatterplot 3, I found no strong linear relationship between school expenditure and average employment rate. Thus, it would not be advisable to try and fit a linear model to this data. again, this deterred from my hypothesis that increased government investment in education heavily influenced employment rate.



B. Research Question 3

The last research question requires the need to break the education data into statistics for men and women. To conduct this analysis, I primarily used SQL queries and looked for any highlights in the results.

To begin with, I found the Top 10 countries with the highest rate of women being educated. The edlevel value I chose for this analysis was tertiary because it indicates the highest form of educational attainment.

	country	gender	percent
▶	Lithuania	Women	90.374809
	Netherlands	Women	89.405571
	Hungary	Women	89.003967
	Belgium	Women	88.898888
	Switzerland	Women	88.795265
	Norway	Women	88.415237
	Poland	Women	88.24192
	United Kingdom	Women	88.125328
	Slovenia	Women	87.02
	Ireland	Women	86.882576

Top 10 Women Education Countries in 2021

Since we have the data for year, I also plotted the same values for 2011, to provide us a perspective of any ten-year changes.

	country	gender	percent
▶	Netherlands	Women	91.627907
	Belgium	Women	87.877731
	Norway	Women	87.801933
	Lithuania	Women	87.116852
	Slovenia	Women	87.05
	Germany	Women	86.232231
	Brazil	Women	85.086113
	Austria	Women	84.798309
	Sweden	Women	84.720367
	Argentina	Women	84.674995

Top 10 Women Education Countries in 2011

We can see that Lithuania comes out on top in 2011 and the Netherlands is first for 2021, with both countries appearing in the top five for both years. Belgium also situates itself in the top five standings over the ten-year gap. It would be beneficial to study what additional factors Lithuania, the Netherlands, and Belgium have in place to achieve such a high percentage of the population achieving tertiary education. It is reasonable to conclude that other countries may want to follow their lead.

My curiosity led me to analyze the top two countries from an added perspective: gender. Thus, I queried the percent of both male and female populations attaining each of the three levels of education.

	country	edlevel	gender	percent
▶	Lithuania	Below upper secondary	Men	62.164875
	Lithuania	Below upper secondary	Women	39.943581
	Lithuania	Tertiary	Men	92.658409
	Lithuania	Tertiary	Women	90.374809
	Lithuania	Upper secondary	Men	85.917534
	Lithuania	Upper secondary	Women	66.42701
	Netherlands	Below upper secondary	Men	77.352715
	Netherlands	Below upper secondary	Women	61.724499
	Netherlands	Tertiary	Men	93.090141
	Netherlands	Tertiary	Women	89.405571
	Netherlands	Upper secondary	Men	89.667458
	Netherlands	Upper secondary	Women	80.501472

Lithuania and Netherlands in 2021

The results were an interesting but not unexpected find. Even though Lithuania and the Netherlands are at the forefront of countries with high percentages of educated women, the numbers *still* trail behind the percentages for men. For example, looking at Lithuanian tertiary educational attainment, women fall behind a mere 2% behind the men in that field. These results point to confirming my initial hypothesis. Even though we are improving the levels of educated women, there remains a gender gap even in the best of scenarios (Lithuania and the Netherlands).

Taking another perspective, I calculated how many countries have more than 70% of women at each level of education (shown in the outputs below). I was extremely surprised to find that only Upper Secondary and Tertiary were included in the query output. This is something to examine in a future study. However, it was affirming to see that the number of countries achieving more women education increased from 2011 to 2021. These results indicate we are moving towards a fairer educational playing field, yet at a slow rate.

	gender	edlevel	numcountries
▶	Women	Tertiary	37
	Women	Upper secondary	19

Countries with > 70% of Women Pop. Education: 2021

	gender	edlevel	numcountries
▶	Women	Tertiary	33
	Women	Upper secondary	11

Countries with > 70% of Women Pop. Education: 2011

VII. CONCLUSION

A. Conclusions

1) There was a slightly inverse relationship between primary school educational attainment level and average employment rate.

2) There was not a strong linear relationship between the average employment rate and government investment in education.

3) Gender parity in education has not been full achieved, but it has improved over the last ten years, which points to slow growth.

B. Limitations & Future Study

1) The model created to answer Research Question 1 can be improved. Perhaps additional variables (indicators of positive outcome) could be added to the dataset from the OECD database. Because of the null values, we had to remove many values from the dataset, which was not ideal for fitting a linear regression model. Future studies should involve several variables to create a more robust model.

2) Some of the output was surprising and did not match well-established conclusions about the education industry. For example, it is widely known that more education leads to better economic output (employment rate). However, results were not strong enough to support these assertions. So I would conduct this study again with perhaps a different dependent variable.

REFERENCES

- [1] F. Lamanna and S. Klasen, "The Impact of Gender Inequality in Education and Employment on Economic Growth: New Evidence for a Panel of Countries," *Feminist Economics*, vol. 15, pp. 91–132, Jul. 2009. Accessed: Apr. 2, 2023. [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/13545700902893106>
- [2] OECD, *Education at a Glance 2022: OECD Indicators*, Paris: OECD Publishing, 2022. [Dataset]. Available: <https://stats.oecd.org/#>. [Accessed: May 8, 2023].
- [3] S. K. R. van Zon, S. A. Reijneveld, C.F. Mendes de Leon, and U. Bültmann. The impact of low education and poor health on unemployment varies by work life stage. *International Journal of Public Health*, vol 62, pp. 997-1006, Dec. 2017. Accessed: Apr. 2, 2023. [Online]. Available:

