**Data Set Title**

**Exploratory Analysis**

Kerry Effinger, [keffinger2@bellarmine.edu](mailto:keffinger2@bellarmine.edu)

1. **INTRODUCTION**

[**https://databank.worldbank.org/reports.aspx?source=283&series=SE.TER.GRAD.FE.SC.ZS#**](https://databank.worldbank.org/reports.aspx?source=283&series=SE.TER.GRAD.FE.SC.ZS#)

**From the World Bank data, I found statistics on gender equity, specifically as it relates to education, for each country. I was able to create my own dataset by selecting variables from a long list, then organizing them by year, and saving them to a csv file. In the following exploratory analysis, I will present information about both the continuous and discrete variables in the Gender Statistics dataset and look for any correlations among the different variables.**

Short description of the data set including a reference to where it can be found and why you chose it.

1. **DATA SET DESCRIPTION**

**There were originally 31 variables of data in this set I compiled. After cleaning the data more, I decided to throw away a few of the variables (marked with \*\*) because they were either not useful due to too much missing data, or they were repetitive. After deleting variables, there are 26 total. Most of the missing data was noted with ‘..’ while some was noted with ‘na.’ I was able to count how many of each there were in each column and find the percent of missing data for each variable. Some of the variables have a high missing data value, but the dataset is large enough that I kept going.**

Narrative summary of the data set: e.g. this data set contains 398 samples with 7 columns with various data types. A complete listing is shown in **Table 1**. For data types you want to indicate two things (nominal, ordinal, interval, or ratio) and the Pandas data type. For example, age might be ratio/int32. For missing data, indicate what percentage of data from that column are missing. Ensure you check to for NaN, NA, or any other indicators that actually mean missing data.

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| Time | Datetime64, interval | .22% |
| \*\*Year | Datetime64, interval | .38% |
| Country | String, nominal | .38% |
| Country Code | String, nominal | .38% |
| Females in NatSci/Math/Stats (%) | Int32, ratio | 83.4% |
| GDP Per Capita (current $US) | Int32, ratio | 25.8% |
| Gov expenditure per student | Int32, ratio | 88.2% |
| Female Life Expectancy | Int32, ratio | 44.4% |
| Male Life Expectancy | Int32, ratio | 44.4% |
| Population | Int32, ratio | 20.7% |
| Female Population | Int32, ratio | 27.6% |
| Female Pop (% of total) | Int32, ratio | 27.6% |
| Females in STEM programs(%) | Int32, ratio | 83.2% |
| Female Primary Graduation(%) | Int32, ratio | 77.3% |
| Male Primary Graduation(%) | Int32, ratio | 77.3% |
| Female Tertiary Graduation(%) | Int32, ratio | 82.5% |
| Male Tertiary Graduation(%) | Int32, ratio | 82.6% |
| Youth Literacy | Int32, ratio | 78.1% |
| A woman can get a job in the same way as a man (1=yes; 0=no) | Int32, nominal | 28.5% |
| A woman can open a bank account in the same way as a man (1=yes; 0=no) | Int32, nominal | 28.5% |
| A woman can choose where to live in the same way as a man (1=yes; 0=no) | Int32, nominal | 28.5% |
| \*\*First Marriage female | Int32, ratio | 90.8% |
| \*\*First Marriage male | Int32, ratio | 91.3% |
| Sons and daughters have equal rights to inherit assets from their parents (1=yes; 0=no) | Int32, nominal | 28.1% |
| \*\*Employed children, Male | Int32, ratio | 100% |
| \*\*Employed Children, Female | Int32, ratio | 100% |
| Total Labor Force | Int32, ratio | 11.8% |
| Female Labor Force | Int32, ratio | 29.5% |
| Female Labor Force (% of total) | Int32, ratio | 29.5% |
| Female Retirement Age | Int32, ratio | 73.5% |
| Male Retirement Age | Int32, ratio | 73.5% |

1. **Data Set Summary Statistics**

The next table includes summary statistics for every continuous, or numerical, variable in the dataset. After changing all the missing data to zeroes, the summary statistics were thrown off and skewed to 0. Changing the 0s to NaN values made it so Pandas could ignore the missing data and only find summary statistics for the numerical data.

**Table 2: Summary Statistics for XXX (name of dataset)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* | *25th* | *50th* | *75th* | *Max* |
| Females in NatSci/Math/Stats (%) | *223* | *55.92%* | *15.71* |  |  |  |  |  |
| GDP Per Capita (current $US) | *981* | *16081.28* | *16081.28* | *0* |  |  |  |  |
| Gov expenditure per student | *159* | *16.24* | *6.52* | *0* |  |  |  |  |
| Female Life Expectancy | *736* | *74.75* | *7.75* | *0* |  |  |  |  |
| Male Life Expectancy | *736* | *69.94* | *7.20* | *0* |  |  |  |  |
| Population | *1048* | *306,183,000* | *963,940,000* | *0* |  |  |  |  |
| Female Population | *956* | *166,438,000* | *496,425,000* |  |  |  |  |  |
| Female Pop (% of total) | *956* | *49.90%* | *3.03%* |  |  |  |  |  |
| Females in STEM programs(%) | *224* | *34.82%* | *9.19%* |  |  |  |  |  |
| Female Primary Graduation(%) | *302* | *89.26%* | *21.91%* |  |  |  |  |  |
| Male Primary Graduation(%) | *302* | *88.48%* | *21.52%* |  |  |  |  |  |
| Female Tertiary Graduation(%) | *233* | *38.00%* | *21.05%* |  |  |  |  |  |
| Male Tertiary Graduation(%) | *232* | *25.67%* | *14.78%* |  |  |  |  |  |
| Youth Literacy | *293* | *97%* | *6.1%* |  |  |  |  |  |
| Total Labor Force | *1165* | *153,406,000* | *457,622,300* |  |  |  |  |  |
| Female Labor Force | *932* | *59,656,000* | *176,447,200* |  |  |  |  |  |
| Female Labor Force (% of total) | *932* | *40.93%* | *9.08%* |  |  |  |  |  |
| Female Retirement Age | *354* | *60.32* | *4.25* |  |  |  |  |  |
| Male Retirement Age | *354* | *61.64* | *3.74* |  |  |  |  |  |

*Table

Description automatically generatedTable

Description automatically generated*

*Graphical user interface, application

Description automatically generated*

There should be a table for **EACH** categorical variable.

Table 3: Proportions for ‘A woman can get a job in the same way as a man (1=yes; 0=no)’ (n=945)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *Yes* | *842* | *89.1%* |
| *No* | *103* | *10.9%* |

Table 4: Proportions for ‘A woman can open a bank account in the same way as a man (1=yes; 0=no)’ (n=945)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *Yes* | *909* | *96.2%* |
| *No* | *36* | *3.8%* |

Table 5: Proportions for ‘A woman can choose where to live in the same way as a man (1=yes; 0=no)’ (n=945)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *Yes* | *760* | *80.4%* |
| *No* | *185* | *19.6%* |

Table 6: Proportions for ‘Sons and daughters have equal rights to inherit assets from their parents (1=yes; 0=no)’ (n=945)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *Yes* | *738* | *78.1%* |
| *No* | *207* | *21.9%* |

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn’t make sense)

Table 7: Correlation Table/Tables

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

After the table with the raw data, include a heatmap of the correlation matrix as a figure.

1. **DATA SET GRAPHICAL EXPLORATION**

There were several relationships I was able to find in this dataset. The youth literacy rate was very much skewed to the right, meaning youth literacy worldwide is relatively high. As seen in scatterplot (a), males and females tend to have a similar graduation rate from primary school, since the plot is quite linear. While scatterplot(b) may have something going on as a confounding variable, it would seem to show that there is not a strong relationship between the percent of females in the population and the number of females in STEM programs. The bar charts for the categorical variables show that most countries feel that women and men have similar opportunities, but there are a select few that are still unequal in some areas.

* 1. *Distributions*

*Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated*

*Chart, histogram

Description automatically generated*

* 1. *ScatterPlots / Pairwise Plots (continuous variables)*

*Chart, scatter chart

Description automatically generated*

*Chart, scatter chart

Description automatically generated*

* 1. *Barcharts (categorical variables)*

*Chart

Description automatically generatedChart, bar chart

Description automatically generated*

*Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated*

* 1. *Other Plots - don’t skimp – there are likely other plots that would be useful that I haven’t already specified. Include those in this section.*

*A picture containing text, shoji, window

Description automatically generated*

1. **SUMMARY OF FINDINGS**

For the years from 2015 to 2020, men and women around the world have relatively equal opportunities in their lives, college educations and careers, according to this sample of a few variables. The data from the World Bank website had many more variables that could have been used for this analysis, but I wanted a wide variety of data types and also some data that could help give some context to the demographic of the countries in question (like the population stats).

There are a few things that could have been done to improve this analysis to make it more in depth and thorough. Cleaning the data a little more and formatting the numbered data, renaming and reordering some columns would help. Also, comparing the data from year to year would have made for more interesting analyses.