

**TASK**

**Exploratory Data Analysis on the World Development Index( WDIData) & Human Development Index (HDI) Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

World Development Index (WDIData) data set from World Bank contains various development indictors for each country in the world and does this across a time period of 48 years. The data for literacy rates and average working hours of children were used in the analysis. Other indicators were not considered in the analysis to keep the task shorter. Many factors impact literacy rates and some of those compare gender and age on the rate. The Human Development Index was also combined to the World Development Index (WDIData) data set to provide data on countries rankings. The HDI can be used to classify whether a country is first or third world.

**DATA CLEANING**

# SUMMARY OF THE METHODS AND VISUALIZATIONS DONE DURING DATA CLEANING

To ensure the data was accurate, consistent and useable, the following steps were taken. The data that was selected with interest in the fields of children and education. The topics of interest were furthered drilled down to look at just the Literacy rates and the Average working hours of children in the age group of 70- 14 years. As the data set was big only the tail of worst countries was looked at with further analysis.

To ensure accuracy the data types were checked. No data types were needed to be changes. Duplicates were also checked for, but none were found. First the data was read into a dataframe and the datatypes were checked. By the nature of the data duplicates were expected but necessary so it was kept in. The arranged of the data was changed using melt to configure a new dataframe in the form desired. Pivot table were also created to view the data in a more desirable format to view. The data was then pivoted again with the indices in the columns of the pivot dataframe. Visualisations were generated from the melted and pivoted data set/

MISSING DATA

# ANY MISSING DATA? HOW DID YOU HANDLE IT

The isnull method was used to check how much data was missing. The original data set was large and as expected there were considerable fields left blank (NAN) or filled by “..”. Columns that had no data were determined and then dropped in the beginning. The next step was to decide which data was needed for the analysis and then filter based on that. Then the structure of the of how the data would look was changed. After that the data was further filtered to exclude ‘..’.

The missing data was dropped after making the pivot table

DATA STORIES AND VISUALIZATIONS

# THIS IS THE BULK OF THIS PROJECT. EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALIZATIONS OF THE DATA

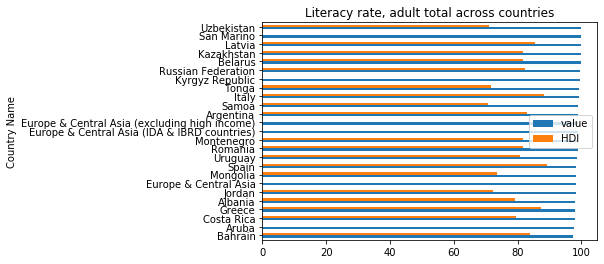


Figure 1

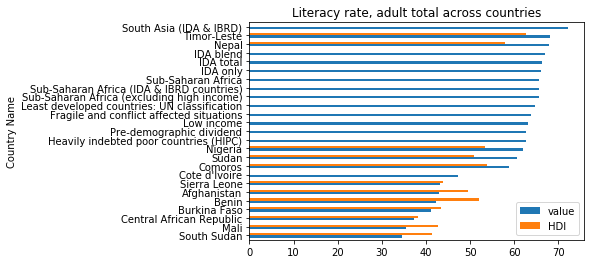
****

Figure 2

The graphs above indicate the Literacy rate of adults throughout the world. Data from the more recent year was plotted (2018).Only the head and tail of 25 countries was used. As can be seen in Figure 1 the country with best literacy rate in the world (based on the information provided) for adults is Ubekistan but is should also be noted that the top five countries are close in the literacy percentage. Also the Figure 1 indicates countries with a high HDI as expected. As can be seen in Figure2 the country with best literacy rate in the world for adults (based on the information provided) is South Sudan Also the Figure 2 indicates countries with a low HDI as expected (third world).

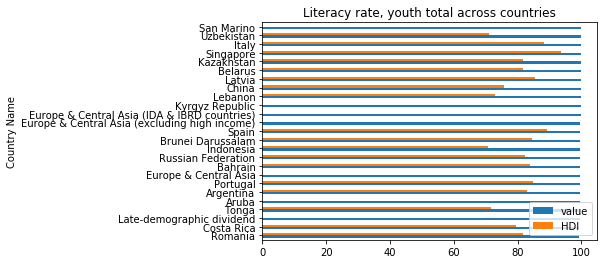


Figure 3

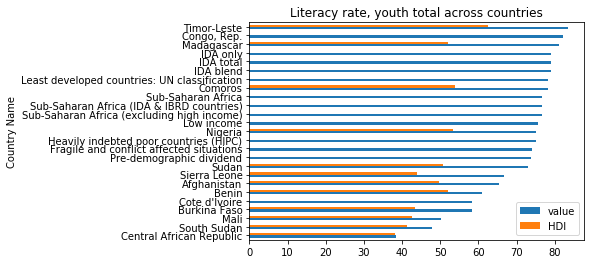
****

Figure 4

The graphs above indicate the Literacy rate of youth throughout the world. Data from the more recent year was plotted (2018) .Only the head and tail of 25 countries was used. As can be seen in Figure 3 the country with best literacy rate in the world (based on the information provided) for the youth is San Marino but is should also be noted that the top countries are close in the literacy value. Also the Figure 3 indicates countries with a high HDI as expected. As can be seen in Figure2 the country with best literacy rate in the world for the youth (based on the information provided) Central African Republic. Also the Figure 2 indicates countries with a low HDI as expected (third world).

When comparing the literacy of the adults to literacy of the youth the data suggests that the youth have a higher literacy rate than the adults.

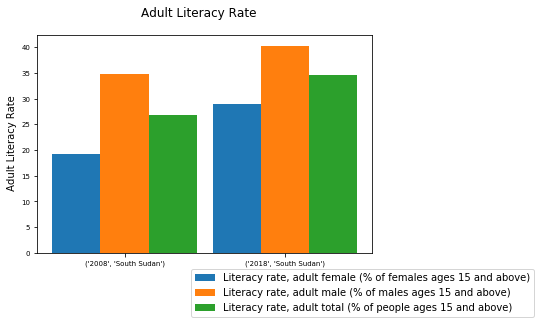
****

Figure 5

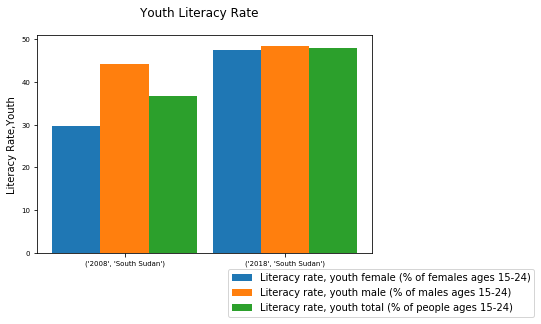
****

Figure 6

From Figures 2 and 4 it was shown that South Sudan faired very low in the literacy rate. Figures 5 and 6 are just further expanded across the timeline to look at South Sudan in detail. It can seen there has been improvement in the literacy rate from 2008 to 2018. The 2018 portion also shows an even split between males and females where in 2008 the male literacy rate was higher.

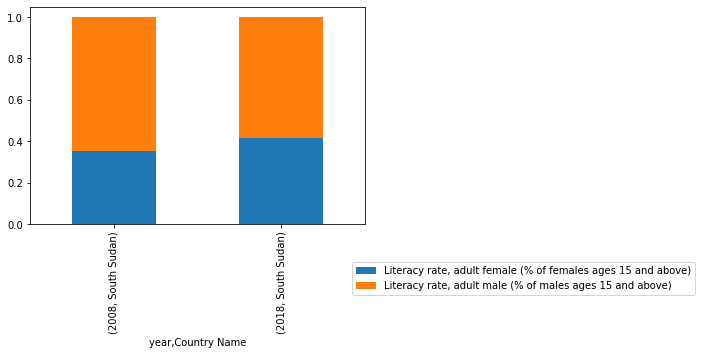
****

Figure 7

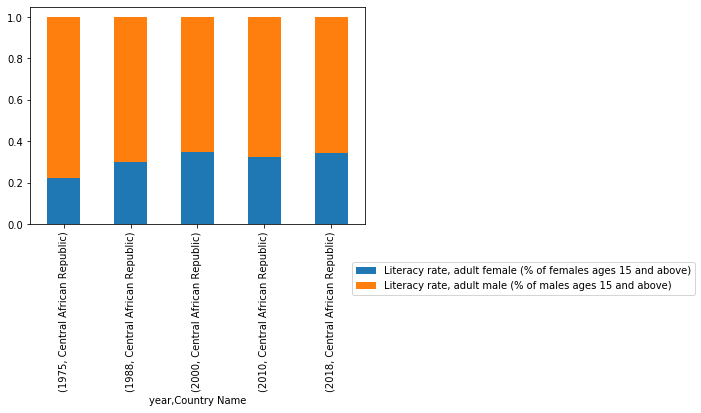
****

Figure 8

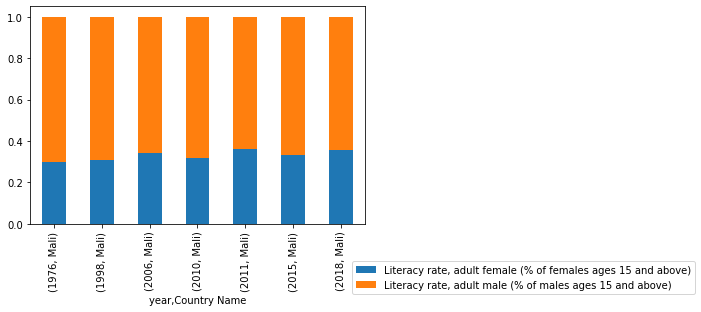
****

Figure 9

The bottom three countries were analysed under adult literacy. To understand the gender disparity that exists Figure 7 -9 were visualised. These graphs indicate that males are more literate than females even across the timelines. There has been a small change in the disparity over the years.

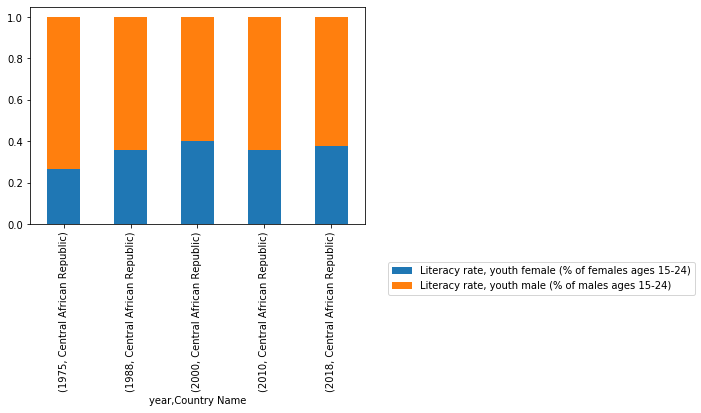
****

Figure 10

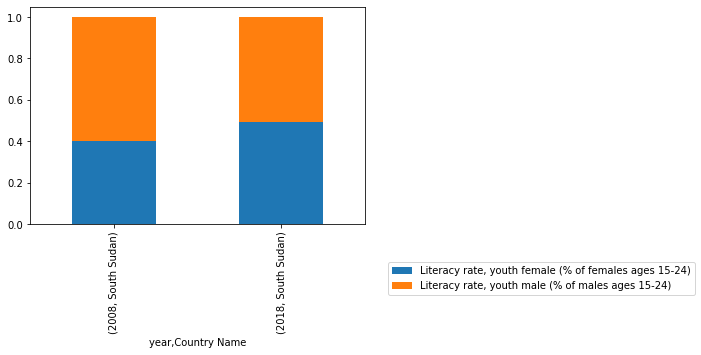
****

Figure 11

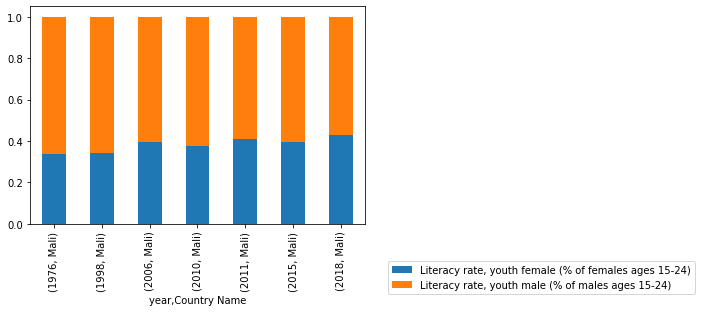
****

Figure 12

The bottom three countries were analysed under youth literacy. To understand the gender disparity that exists Figure 10 -12 were visualised. These graphs indicate that males are more literate than females even across the timelines. There has been a small change in the disparity over the years. The youth literacy rate for females is higher in the youth than adults.

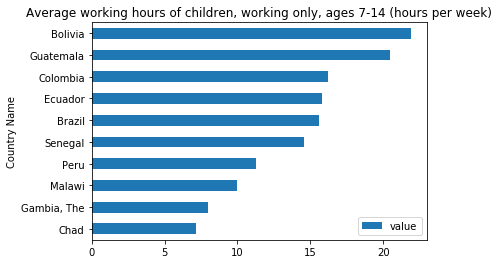
****

Figure 13

Figure 13 indicates the average working hours for children. Data was extracted from the most recent year. The country with the highest rate is Bolivia. The graph indicates that South American countries have the higher rates.

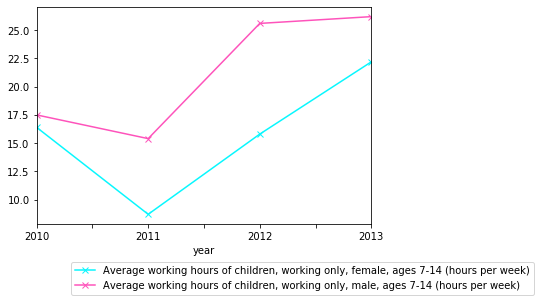
****

Figure 14

Figure 14 indicates that males work more hours per week than females.

**THIS REPORT WAS WRITTEN BY : Krishnie Thaver**

