NOVA IMS Information Management School

PORTLINES

How the Country Conditions Influence their Population Progression

Spring Semester of 2022

WORK DONE BY GROUP O:

Luis Silvano r20201479
Miguel Tomé r20201644
Pedro Sousa r20201611
Vasco Reis r20201478

WORLD BANK GROUP

Index

1. Introduction

Nowadays, although exist a suppose increase of the lifestyle with the overall population, there is also an increasing of the inequalities around the world, the rich are getting richer, and the poor are getting poorer. Also, with the advance of technology the people have become more informed with trustworthy sources of information and became more aware of the influence of some factors regarding the development of their country. The data flow and the amount of data available also increased and this became a bit overwhelming for the average person to analyze and completely understand the major impact of the overall conditions of the different countries and realize the true inequalities.

So, the *World Bank Group* gave to *PORTLINES* the mission to recognize ***"How the Country Conditions Influence their Population Progression"*** focusing on countries all over the world. We analysed different areas and subjects throughout the time, showing which countries have greater values in each one of the indicators, by display their progression, and the differences between the territories using the program, Power BI. We selected indicators revolving education, economics, environment, population, among others, studying different variables in each one of them, to have more precise and detailed information and a better analysis of our project.

2. Data Source

PORTLINES collected datasets from the economy, demographic, education, employment, environment and pollution and the basic care. We decide to use these categories, to get bigger understanding of the concept, to achieve a deeper analysis and a consistent interpretation of their country's progression. With this, set the differences between the territories from the citizens of each country and how the territories are influenced by each variable from the different categories. Since it provides a vast list of datasets, we used The World Bank Group to take our data and we organized each dataset by a specific country with a time interval data measured in years. In addition, we also created new categories related to worldwide organizations like OCDE, NATO, CPLP, BRICS and other regions, giving an extra interpretation of the data. This allowed us to evaluate these organizations between each and compare the values of the countries' indicators inside each organization and region.

3. Data Pre-processing – Power Query M

After extracting the datasets from *The World Bank Group, we* organized, cleaned, and transform the data properly in Excel. After accomplishing these steps, we can conclude the datasets was ready to be prepared in Power BI. The main goal of this global analysis is to compare the countries, and which country needs more support according to the World Development Indicators.

Starting, we imported datasets to the program so we could use the business intelligence tools and make conclusions from the dataset. Secondly, we organize our table inside of the program Power BI, putting the first row to the head and we turned multiple columns into rows **(unpivot)** to easier identification the values according to their respective year (from 2001 until 2020).

Next up, we verified that we had more than 4 characters in the time interval (column of the years), and that being extra information, we decide to split it up to not have unnecessary information. Also, since there were some irrelevant columns that could contain wrong information, such as "Series Code" (which shows the code of each different variable, and for the purpose of the project that information does not matter), so we decided to remove them. Besides that, we also removed the extra information from the split year column.

Afterwards, we renamed the column year for a better analysis, and we treated all the missing values. For this process, we used the "Replaced values" option, for substitute all values **(".."** to missing values, eliminating any possible have errors in our dataset could had.

Later, we isolated each column of our dataset in different tables (by copying the respective table and merging the columns Country Name, Country Code, and Year), but since the dataset had only 1 variable it wasn't necessary to perform any merge.

It was time we put the different merged columns in a single table, change the types of each column and we converted the values into percentages, to facilitate the visual representation of our model.

We created two new dimensional tables, a "Calendar" with information referring to years and other periods of time, and a country table, with information, referring to each country and some extra information.

For the creation of the country table, we took the base of a fact table and deleted all the information besides the countries name and code, then we removed the duplicates to have one single row for each country. After that we added some extra information about the continent, region and organizations which each country

belongs to, we did this to apply a filter with the purpose of only selecting the countries that belongs to a specific continent, or region or organization. At some point we had to pay special attention in some cases where the name of some countries where different with small differences but that affected the connection between the initial table and the table with the information of the organizations (a good example was the Slovakia which sometimes was referred as Slovakia Republic). After verifying this we could finally join all tables together in only one, that we called country table. For the Organizations we put a column for each organization and specify as True if a country belongs to that, for the continents and regions we had each continent and region that each country it's in.

Lastly, for the creation of the calendar table we used the formula "**CALENDARAUTO()", from DAX**, to create a date column, to connect this column with other similar columns in each table that performs data intersection for every single table. Then, from the date column we added different columns for the year, quarter, month, weekday, and the number of the week, and from the year, a column for the decade and the lustrum (period of five years), this with the purpose of extra filtering the data and more specification.

4. Modelling

Now it was the time of creating the model itself. We had the idea to follow a star schema but, we were limited with this option due to the quantity of dimensional table making our datasets unreadable. We decided then due to the complexity of our model, to put the fact tables on top and the dimensional table on the bottom to make a more comprehensive interpretation. To accelerate our program, we hid some variables for having the same meaning that were repeated in both tables, leaving in the fact table only the variables, and on the dimensional tables all the attributes, and connected each dataset to the country table by country code and to the calendar table by date with a many to one relationship.

After this step, we created two hierarchies Country and Year. Year Hierarchy containing Decade, Lustrum and Year and Country Hierarchy with Continent, Region and Country. Adding these hierarchies added more features to the model and facilitate the filtering and visualization, as well drill down and roll up of the dashboard.

5. Indicators

Education Indicator

Compulsory education, duration (years): *Duration of compulsory education is the number of years that children are legally obliged to attend school- by UNESCO Institute for Statistics.*

Pre-primary education, duration (years): *Pre-primary duration refers to the number of grades (years) in pre-primary school- by UNESCO Institute for Statistics.*

Primary education, duration (years): *Primary duration refers to the number of grades (years) in primary school- by UNESCO Institute for Statistics.*

Secondary education, duration (years): *Secondary duration refers to the number of grades (years) in secondary school- by UNESCO Institute for Statistics.*

Basic Needs Indicator

People using at least basic sanitation services (% of population): *The percentage of people using at least basic sanitation services, that is, improved sanitation facilities that are not shared with other households- by UNICEF Joint Monitoring Programme (JMP) for Water Supply, Sanitation and Hygiene.*

People using at least basic drinking water services (% of population): *The percentage of people using at least basic water services- - by UNICEF Joint Monitoring Programme (JMP) for Water Supply, Sanitation and Hygiene.*

People practicing open defecation (% of population): *People practicing open defecation refers to the percentage of the population defecating in the open, such as in fields, forest, bushes, open bodies of water, on beaches, in other open spaces or disposed of with solid waste- by UNICEF Joint Monitoring Programme (JMP) for Water Supply, Sanitation and Hygiene.*

Access to electricity, urban (% of urban population): *Access to electricity, urban is the percentage of urban population with access to electricity- World Bank Global Electrification Database.*

Access to electricity, rural (% of rural population): *Access to electricity, rural is the percentage of rural population with access to electricity- World Bank Global Electrification Database.*

Access to electricity (% of population): *Access to electricity is the percentage of population with access to electricity- World Bank Global Electrification Database.*

Access to clean fuels and technologies for cooking (% of population): *Access to clean fuels and technologies for cooking is the proportion of total population primarily using clean cooking fuels and technologies for cooking- by Global Health Observatory.*

Demography Indicator

Female Population: *Female population is based on the de factor definition of population, which counts all female residents regardless of legal status or citizenship- by United Nations Population Division's World Population Prospect.*

Male Population: *Male population is based on the de factor definition of population, which counts all male residents regardless of legal status or citizenship- by United Nations Population Division's World Population Prospect.*

Total Population: *Total population is based on the de factor definition of population, which counts all residents regardless of legal status or citizenship- by United Nations Population Division. World Population Prospects, also from national statistical offices, Eurostat, United Nations Statistical Division, U.S. Census Bureau and Secretariat of the Pacific Community.*

Suicide mortality rate, female (per 100,000 female population): *Suicide mortality rate is the number of suicide deaths in a year per 100,000 female population. Crude suicide rate (not age-adjusted)- by Global Health Observatory Data Repository.*

Suicide mortality rate, male (per 100,000 male population): *Suicide mortality rate is the number of suicide deaths in a year per 100,000 male population. Crude suicide rate (not age-adjusted)- by Global Health Observatory Data Repository.*

Age dependency ratio, old (% of working-age population): *Age dependency ratio, old, is the ratio of older dependents--people older than 64--to the working-age population -those ages 15-64 - by United Nations Population Division's World Population Prospects.*

Age dependency ratio, young (% of working-age population): *Age dependency ratio, young, is the ratio of younger dependents--people younger than 15—to the working-age population--those ages 15-64- 64 - by United Nations Population Division's World Population Prospects.*

Mortality rate, infant (per 1,000 live births): *Infant mortality rate is the number of infants dying before reaching one year of age, per 1,000 live births in a given year- by UNICEF, WHO, World Bank, UN DESA Population Division.*

Rural population: *Rural population refers to people living in rural areas as defined by national statistical offices- by United Nations Population Division's World Urbanization Prospects.*

Urban population: *Urban population refers to people living in urban areas as defined by national statistical offices- by United Nations Population Division's World Urbanization Prospects.*

Economic Indicator

Current health expenditure per capita (current US$): *Estimates of current health expenditures include healthcare goods and services consumed during each year- by Health Organization Global Health Expenditure database.*

Adjusted net national income (current US$): *Adjusted net national income is GNI minus consumption of fixed capital and natural resources depletion- by World Bank's "The Changing Wealth of Nations: Measuring Sustainable Development in the New Millennium.*

Unemployment, female (% of female labor force) (national estimate): Female *unemployment refers to the share of the labor force that is without work but available for and seeking employment- by International Labour Organization, ILOSTAT database.*

Unemployment, male (% of male labor force) (national estimate): Male *unemployment refers to the share of the labor force that is without work but available for and seeking employment- by International Labour Organization, ILOSTAT database.*

Unemployment, total (% of total labor force) (national estimate): Total *Unemployment refers to the share of the labor force that is without work but available for and seeking employment- by International Labour Organization, ILOSTAT database.*

Inflation, GDP deflator (annual %): *Inflation as measured by the annual growth rate of the GDP implicit deflator shows the rate of price change in the economy as a whole- by World Bank national accounts data, and OECD National Accounts data.*

Income share held by highest 20%: *The percentage share of income or consumption that accrues to the fifth (wealthiest) quintile- by primary household survey data obtained from government statistical agencies and World Bank country departments.*

Income share held by the lowest 20%: *The percentage share of income or consumption that accrues to the first (poorest) quintile- by primary household survey data obtained from government statistical agencies and World Bank country departments.*

GDP (current US$): *GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus, any product taxes and minus any subsidies not included in the value of the products- to World Bank national accounts data and OECD National Accounts data.*

Environment and Pollution Indicator

CO2 emissions (metric tons per capita): *Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement- by the World Resources Institute.*

6. Dax Measures

Current Gross Domestic Product (GDP) = SUM (GDP[GDP (current US$)]).

- For obtain the Gross Domestic Product Total, and when using a year slicer, getting the correct Gross Domestic Product for each year, the same can be said when using a country slicer.

PriorYearGDP = CALCULATE (SUM(GDP[GDP (current US$)]), PREVIOUSYEAR('Calendar'[Date])).

- To obtain the total Gross Domestic Product for the previous year that we are analyzing, so for the year of 2019 we will have the Gross Domestic Product of 2018, because of this, the year of 2001 will return an empty value since it's the first year available.

Gross Domestic Product Difference = GDP Difference = VAR _CurrentYearGDP = [Current GDP]

VAR _PreviousYearGPD = [PriorYearGDP]

VAR _Result = CurrentYearGDP - PreviousYearGPD

RETURN _Result:

- For obtain the difference between two years we associated both measures created before to two temporary measures, and then we created another one which we subtracted to them, and finally we returned that variable ('Result').

Gross Domestic Product Growth in Percentage = DIVIDE ([GDP Difference], [PriorYearGDP],0).

- For obtain the percentual growth of the Gross Domestic Product by year, we divided the measure created before by the value of the previous year.

Average Difference in Age Dependency = DIVIDE (SUM('Age Dependency Ratio'[Age dependency ratio, young (% of working-age population)]) - SUM('Age Dependency Ratio'[Age dependency ratio, old (% of working-age population)]), COUNTA('Age Dependency Ratio'[Age dependency ratio, young (% of working-age population)])).

- For obtain the average difference between the age dependency ratio of young people and the age dependency ratio of old people, we divide the difference between both by the number of values present in the dataset (in this case we used the young columns, but we would achieve the same result if using the old one). It's to say that this difference it's a difference in percentual points.

Cases of Deflation = CALCULATE (COUNTA('Inflation, GDP'[Country Name]),'Inflation, GDP deflator'[Inflation, GDP deflator (annual %)]<0).

- For obtain the number of times that occur deflation over the years, we counted the number of values presents in the table of Inflation where the annual inflation was negative.

Count of countries without Pre-primary = DISTINCTCOUNTNOBLANK ('Education Duration'[Preprimary Education Duration]).

- For obtain the total number of countries that don't have a Pre-primary Education, or at least it's not said in the dataset, we used a distinct count for blank values, so even if the same country is listed multiple times, it only counts one time.

Count of high infant mortality = CALCULATE (COUNTX ('Infant Mortality','Infant Mortality'[Mortality rate, infant (per 1,000 live births)]), 'Infant Mortality'[Mortality rate, infant (per 1,000 live births)] > 50).

- For obtain an idea of which countries or cases of high infant mortality we count every row where the mortality rate was higher than 50 per 1000 live births.

Population Density = DIVIDE (SUM ('Total Population'[Pop Total]), SUM('Land Area (sq km)'[Area sqr/km])).

- For obtain the population density, we divided the total population of each country by the area of each country. To do this we added a small dataset with the land territory of each country in square kilometres.

Total C02 Emissions in metric tons = SUMX (NATURALINNERJOIN ('CO2 emissions (metric tons per capita)','Total Population'), ('Total Population'[Pop Total] * 'CO2 emissions (metric tons per capita)'[C02 emissions per capita])).

- Since we wanted to evaluate the Total CO2 Emissions over the years, we create this measure to iterate by each row and don't return wrong values. For this measure we did a natural join, between CO2 emissions and the table of the population, to multiply each CO2 emissions per country of each country by the correspondent population.

Number of countries with inequalities of income share = CALCULATE (DISTINCTCOUNT('Income Share'[Country Name]) , 'Income Share'[Income share held by highest 20%] > 0.5, 'Income Share'[Income share held by lowest 20%] < 0.05).

- For obtain this measure we believed that we were witnessing cases of inequalities when a country had more than 50% of the income share held by the 20% people richer and at the same time the poorest only held less than 5%.

Missing Values in Income Share = COUNTBLANK ('Income Share'[Income share held by highest 20%]).

- For obtain these values we counted every row that had missing values in the column of income share held by the highest 20%, for understand how much information we were working with in the previous measure.

Adjusted Net Income per capita = DIVIDE (SUM ('Adjusted Net Income - $'[Adjusted Net Income]), SUM ('Total Population'[Pop Total])).

- For obtain these values we adjusted net income per capita, and we divide our total adjusted net income by the total population.

Unemployment Evolution from the first year to the last in Percentage =

CALCULATE (SUM (Unemployment [Total Unemployment]),

LASTDATE (CALCULATETABLE(VALUES('Calendar'[Date]),'Calendar'[RowsNotEmplty- unemployment] =TRUE)))-

CALCULATE (SUM (Unemployment [Total Unemployment]),

FIRSTDATE(CALCULATETABLE(VALUES('Calendar'[Date]), 'Calendar'[RowsNotEmplty - unemployment] =TRUE))).

- For obtain this evolution we though in a measure that could show the growth of the unemployment from the first year available on the dataset, or on the filter if it is the case, to the last year. This measure shows the change between the two dates in percentage, and it can help to show the change in unemployment by continent for example.

Female Unemployment Evolution from the first year to the last Percentage =
CALCULATE(SUM(Unemployment[Female Unemployment]),
LASTDATE(CALCULATETABLE(VALUES('Calendar'[Date]),'Calendar'[RowsNotEmplty - unemployment]=TRUE))) -
CALCULATE(SUM(Unemployment[Female Unemployment]),
FIRSTDATE(CALCULATETABLE(VALUES('Calendar'[Date]),'Calendar'[RowsNotEmplty - unemployment]=TRUE))).

- This measure was created with same intuit as the total unemployment and has a similar code but for the female population, so this measure shows the change between the two dates in percentage, and it can help to show the change in female unemployment by continent for example.

Male Unemployment Evolution from the first year to the last in Percentage =
CALCULATE(SUM(Unemployment[Male Unemployment]),
LASTDATE(CALCULATETABLE(VALUES('Calendar'[Date]),'Calendar'[RowsNotEmplty - unemployment]=TRUE))) -
CALCULATE(SUM(Unemployment[Male Unemployment]),
FIRSTDATE(CALCULATETABLE(VALUES('Calendar'[Date]),'Calendar'[RowsNotEmplty - unemployment]=TRUE))).

- This measure shows the change between the two dates in percentage, and it can help to show the change in male unemployment by continent for example. And the code created is similar but directed to the male unemployment.

Average Access to Clean Fuels and Technology for Cooking in Percentage = AVERAGE ('Access to clean fuels and technologies for cooking (% of population)'[Access to clean fuels and technologies for cooking]).

- For obtain this measure we used the average function of the access to clean fuels and technologies for cooking, to be able to evaluate in world scale or over the years.

Difference of Access to Electricity between Urban population and Rural Population in Percentage = DIVIDE (CALCULATE(SUM('Acess to Electricity'[Urban Pop with Acess to Electricity]) - SUM('Acess to Electricity'[Rural Pop with Acess to Electricity])), SUM('Acess to Electricity'[Rural Pop with Acess to Electricity])).

- For obtain this measure we wanted to see the difference between the access to electricity by urban population and rural population, this difference it's in percentage, so if the values are for example 100% it means that urban population as 100% more of the rural ones (twice in this case).

Education Expenditure = SUMX(NATURALINNERJOIN(Expenditures,'GDP'),(Expenditures[Government Education Exp - % of GDP]*[GDP (current US$)])).

- Our dataset only had the education expenditure in percentage of the value of the GDP, to get the total expenditure, we use the natural inner join function, to join these two tables and multiply the value of the education by the total GDP.

Health Expenditure = SUMX (NATURALINNERJOIN(Expenditures,'Total Population'), ([Health Expenditure] * [Pop Total])).

- In this case our dataset had the health expenditure per capita only, and since we wanted to have the total expenditure, so we use the inner join to join the table of expenditures and total population, to after multiplying the value per capita by the total population and get the total health expenditure.

Total Expenditure = [Health Expenditure] + [Education Expenditure] + SUMX (Expenditures, (Expenditures [Military Expenditure] + Expenditures[Final Consumption Expenditure])).

- For obtain this measure we add all the total of sub-categories of expenditures.

Average Suicide Mortality Rate = DIVIDE ((SUM ('Suicide Mortality Rate'[Suicide mortality rate, female (per 100,000 female population).1]) + SUM('Suicide Mortality Rate'[Suicide mortality rate, male (per 100,000 male population).1])), COUNTA('Suicide Mortality Rate'[Suicide mortality rate, female (per 100,000 female population).1]) * 2).

- For obtain the average suicide mortality rate between females and males we did a summatory of both columns and then divided this total by twice the number of values in the rows of the table.

7. Conclusion

Summing up, to analyze ***"How the Country Conditions Influence their Population Progression"***, we resorted to World Bank, and we treated the data, changing the values of miss values, organized our tables, rearrange, and even make new ones for a better in this investigation of our data. From modeling our project to have it really defined indicators, with respective variables, we were able to use DAX and make a further and more detailed study of our indicators in each country, continent, and organization.

We can conclude that even though the overall state of the world and lifestyle is increasing and improving, not all the countries are keeping up and are getting left behind, with too financial discrepancies and other reasons such as management of resources, countries like United States of America China in Russia perform higher in most of the categories like GDP, inflation, but also CO2 emission, comparing to most African countries and some Asian countries whom cannot follow this super powers and have low access to clean fuels, electricity and basic hygiene needs. This along with other factors, show that today still exist first world countries and third world countries and that is not much that we can do about it, if we do not help each other out.

8. References

Eurostat. (n.d.). Retrieved from Demographic Statistics.

International Labour Organization. (2022, February 8).

International Monetary Fund. (n.d.). Retrieved from International Financial Statistics and data files.

Secretariat of the Pacific Community. (n.d.). Retrieved from Statistics and Demography Programme.

U.S. Census Bureau. (n.d.). Retrieved from International Database.

UNESCO Institute for Statistics. (2021, September). Retrieved from http://uis.unesco.org/

UNICEF, WHO, World Bank, UN DESA Population Division. (n.d.). Retrieved from Estimates developed by the UN Inter-agency Group for Child Mortality Estimation: www.childmortality.org

United Nations Population Division. World Population Prospects: 2019 Revision. (n.d.).

United Nations Population Division's World Population Prospects. (n.d.). Retrieved from Total population and age/sex distributions.

United Nations Statistical Division. . (n.d.). Retrieved from Population and Vital Statistics Reprot (various years).

WHO Global Health Observatory. (n.d.). Retrieved from https://www.who.int/data/gho/data/themes/air-pollution/household-air-pollution

WHO/UNICEF Joint Monitoring Programme (JMP) for Water Supply, Sanitation and Hygiene. (n.d.). Retrieved from washdata.org

World Bank. (n.d.). Retrieved from Poverty and Inequality Platform: http://pip.worldbank.org

World Bank. (2011). Retrieved from The Changing Wealth of Nations: Measuring Sustainable Development in the New Millennium.

World Bank. (2018). Retrieved from Estimates based on the United Nations Population Division's World Urbanization Prospects.

World Bank. (2019). Retrieved from Estimates based on age distributions of United Nations Population Division's World Population Prospects.

World Bank Global Electrification Database. (n.d.). Retrieved from Tracking SDG 7: The Energy Progress Report.

World Health Organization. (n.d.). Retrieved from Global Health Observatory Data Repository.

World Health Organization Global Health Expenditure database. (2022, January 30). Retrieved from http://apps.who.int/nha/database