CST 2102
DATA VISUALIZATION

Final Project

Professor: Prableen Singh

Ifeowula Ibidun - 041049560
João Carolino de Oliveira - 041066120
Micky Mwiti - 041077182
Vinicius Gonzalez Caetano - 041009010

Executive Summary

Purpose of the Project

Analise and forecast world population, using different factors since it was known that other groups would analyse Fertility, Mortality and Migration, we wanted to use other factors that are GDP and Life expectancy while also forecasting than and visualise how they corelate while also analysing and taking conclusions from the expected changes.

Method

Process data available on multiple open data portals, using Python, Excel and google sheets to analyse using Python(Linear Regression model) and power BI(Exponential smoothing model), transforming data to visualise on Power point.

We are going to compare and select the best model to then process into vizualizations and conclusions for the rest of the class.

Target Audience

Data Visualization Class of the BISI course

Work Division

Vinicius: Data Processing, data organization and support on forecasting models.

Micky: Forecasting, support of data processing, Analysis of forecasting results.

Ife: Data research, support Vinicius providing data for formatting, help organize the presentation, Analise Life expectancy section of the work.

Carolino: Power BI importing and Data Visualization, Power point presentation organization, Final Report.

Contents

Executive Summary 2

Purpose of the Project 2

Method 2

Target Audience 2

Work Division 2

Background 4

Analysis 4

Data processing 4

Missing Values 5

Live Connection 5

Model Building 5

Results 8

.... 9

Life Expectancy 9

GDP 10

Conclusions and Recommendations 12

Bibliography 14

Data Sets 14

Articles 14

Background

The team selected the topic “Growth of world population and future prediction” with the purpose project being the application of multiple models in order to predict the future trend of the World Population, the aim is to achieve this by applying a total of two different models to the data, the first model being Linear Regression, which is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables, the second model being applied to the data is Exponential Smoothing Method (ETS) which are a family of time series forecasting models, this is a time series forecasting method for univariate data that can be extended to support data with a systematic trend or seasonal component. A reason for this topic selection was the variety of python modules that could be a good fit for this topic and aggregate in the team members’ knowledge.

Analysis

Data processing

Original CSV files with the data were processed with python, using Google Colab, which runs an instance of Jupyter Notebook in the cloud. Processing the original files is an important step, to combine them together and create just one output file with all the information needed, validated and modified as necessary. To create the output file, there is the need of searching for matching primary keys, which in this case are the country names (some country names can be written in more than one way, therefore it’s necessary to normalize those names, to be matched later, otherwise some rows will probably not match and will not be shown in the output file).

The files used in this processing contain information about GDP (Gross Domestic Product), total population and life expectancy for all the countries in the world and data starting from 1960 until 2019. Processing information with python makes the checks for empty and null values an easy task, and this indicated if there was any combination between country and year with that condition, so an additional step was taken to fill up those missing values.

Besides the country names and years, a column for country region was added as well during the processing, so information could be grouped later in the visualizations. Also, as there are three different series for the data (GDP, population and life expectancy), only one column for the series “label” was used, therefore we were able to have only one dataset containing all the information joined together accordingly.

Missing Values

The files used in this processing contained a few missing values that could be filled up with information fetched from other files, so they completed each other in order to create the final output to be used in the visualizations and forecasting.

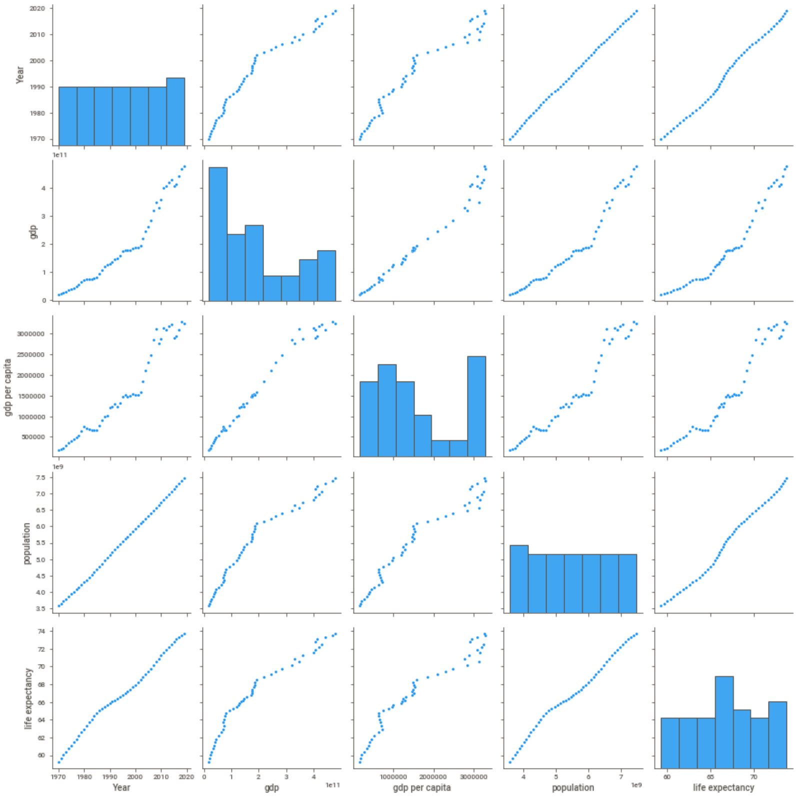
Live Connection

To display the output file in PowerBI, a live connection between the Microsoft tool and a shared Google Drive folder was created, so every time the output file got updated, there was no need to upload it again in PowerBI, only refreshing the data was sufficient to see all the new fetched data, working like an instance of a live database, but structured as a csv file.

Model Building

Before the creation and application of the models to the data, a quick correlation analysis was applied to the data in order to find which variables had the strongest correlation to the population. The results from the correlation are shown in the table below. As presented, the correlation between the variables is quite strong but doesn't show significant variance with the lowest correlated variable being gdp having a correlation level of 96%.

	Year	**GDP**	**GDP Per Capita**	**Life Expectancy**	**Population**
Year	1.000000	0.966448	0.969852	0.999840	0.996665
GDP	0.966448	1.000000	0.989883	0.968394	0.965568
GDP Per Capita	0.969852	0.989883	1.000000	0.970666	0.966619
Life Expectancy	0.999840	0.968394	0.970666	1.000000	0.995774
Population	0.996665	0.965568	0.966619	0.995774	1.000000

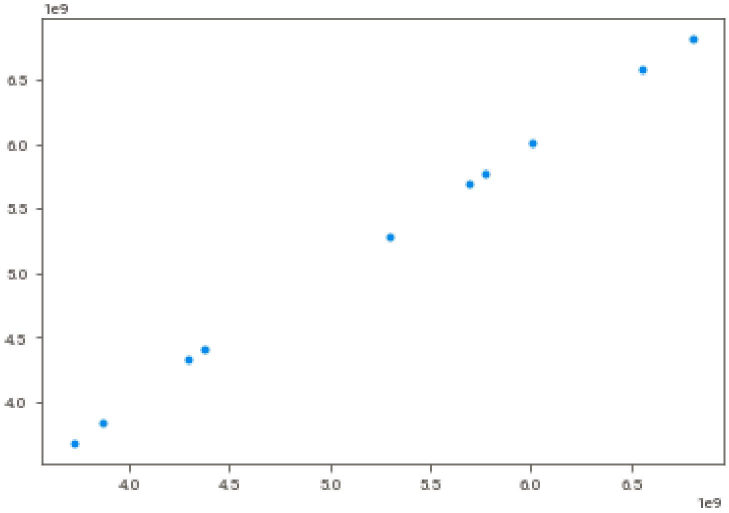


In order to further analyse the relationship between each variable, a visualisation was created as shown in the figure on the right to present what sort of relationship each variable has. According to the visualization, the majority of the variables present a linear relationship with some variables such as life expectancy and population showing a much stronger linear correlation in comparison to the other variables. The visualization also helps to understand each variables spread further such as, the largest life expectancy age presented in the data being 66 years.

Once the correlational analysis was complete, two models were created in order to achieve two outputs, one being the population prediction for the world and the second being the population prediction of each country, both predictions were made till the year 2050. The first model was created by initially allocating the dependent and independent variables, once this was done the data was then split into two, training and testing data set, the training set is then fed to the model and the test set was used to determine the accuracy and performance. The results are as shown in the figure below with the predicted values on the

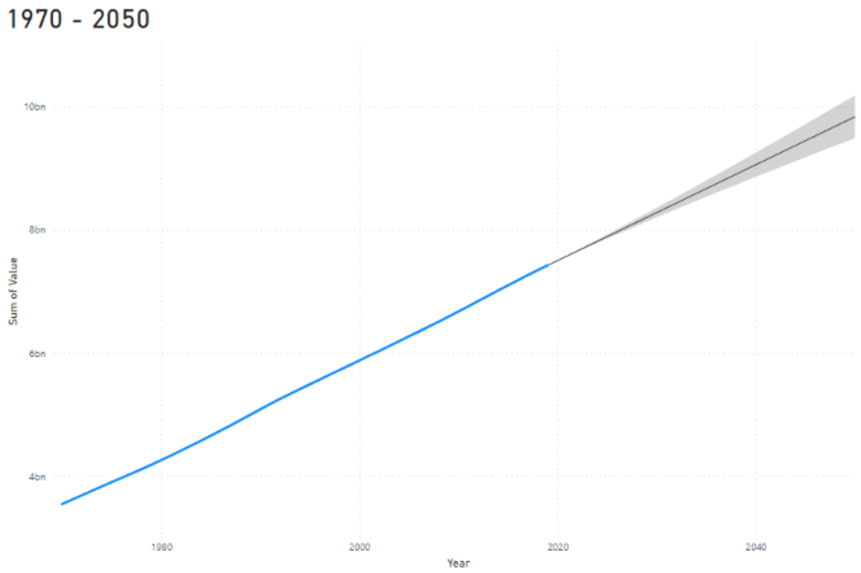
Y-Axis and the actual values on the X-Axis. Further performance evaluation was performed, and the model achieved the predictions with a coefficient of determination (R-Squared) Value of 99.96%.

The second model is capable of achieving two objectives, the first one being the ability to receive two inputs, a country name and a year the user would like the prediction for, the algorithm would then take the country name and confirm that the name is correct, then feed the name and the year into the model and output the population prediction for the country specified and the year. The second objective being creating a list of the countries that we would like the predictions for plus the years and feeding the list and years into the algorithm off a loop and saving the values in a dataframe before exporting into a csv file.

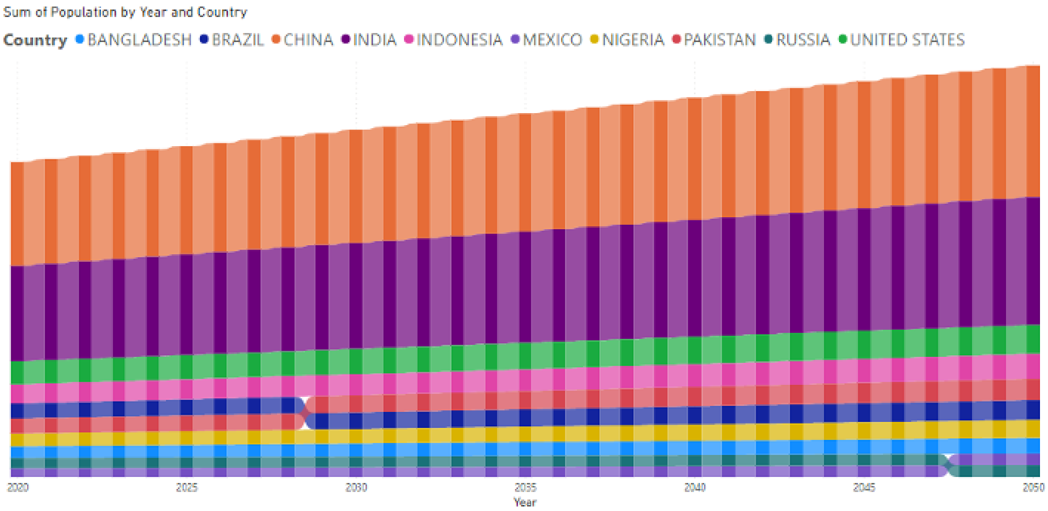


Results

With the two models producing the predictions for the world population until 2050, the results below show the visualisations to present the predictions. The first visualisation shows the general trend of the world population. As shown in the figure below, the prediction follows the already moving upwards linear trend with a confidence interval of 95%, estimated to reach a population value of approximately 9 billion by 2050.



Once we have a deeper look into the predictions and focus on individual countries, we can see a shift in population with some countries expected to pass others in terms of total population. As shown the figure below by a Ribbon Chart, the visualisation shows the top 10 countries and their growth in population between 2020 and 2050, as shown China and India are predicted to have the largest population but the chart also shows a shift in total population with the model predicting that approximately in 2028 Pakistan will pass Brazil in total population whilst also Mexico will pass Russia in total population by the year 2047.



Life Expectancy

According to current projections, the global population will reach eight billion by 2026, and will likely reach around 9 billion by 2042. Alternative scenarios for 2050 range from a low of 6.4 billion to a high of more than 10.6 billion. This increase can be affected by a lot of factors like life expectancy, GDP, Religion, migration. The birth and mortality rates around the globe were roughly equal up until 1998, which helped to maintain a constant population size. However, dramatic gains in human lifespan and the number of people living (and at greater ages) than in the past have been brought about by advancements in health and safety in many fields, as well as other generational growth factors. The following developments in science and technology during the Industrial Revolution significantly decreased the number of fatalities: An increase in food distribution and production, Betterment of the public's health (water and sanitation), Technology in medicine (vaccines and antibiotics).

"The wail of the newborn was heard across the land" occurred after World War II. Since there were more babies born between 1946 and 1964 than at any other time in history, those who were born during that time are referred to as "baby boomers." The youngest baby boomers will turn 69 this year, and the proportion of Americans ages 69 and older in the overall population will increase quickly in the United States. Life expectancy was 68.2 years old in 1950. That age is currently 79.12 years old, and by the year 2050, it is predicted to reach 83.9 years old. Ages 85 and older make up 8% of the world's 65 and older population, with 12% living in more developed nations and 6% in less developed nations. These people represent the fastest-growing segment of the population in many nations. The ageing population will experience various changes and difficulties during the coming decades. Most elderly Americans continue to live at home, frequently with the help of family, friends, and/or professional home-care services as their functional abilities deteriorate. Assisted living facilities have grown in popularity over the past ten years and have shown to offer people who need some help with daily tasks a pleasant living environment. Nursing homes are still an option for people who need more intense help, especially with daily tasks that are fundamental. Elderly people may decide to move or be forced to as they age and experience functional deficits. As a result, seniors can choose living arrangements that maximise their health, security, and functionality.

GDP

As we had gathered the data on GDP to correlate and compare with our population forecast and make a GDP growth forecast, on our own forecasting model it's shown that using either GDP to predict population growth or the inverse.

As we moved away from using GDP as a tool for forecasting GDP growth, by using the power BI model, we lean into analysing and comparing the growth of Population and GDP and what is the most probable impact it will have. Continents were the most appealing part of the data, specially because we could note some trends and combine countries of similar backgrounds that would lead to more informing conclusions and learning.

Europe Is where the power BI model were able to forecast with more nuance the evolution of the GDP, we can assume that this comes as most countries in Europe follows the same trend both during prosperous times and crisis as a unit, the GDP is expected to grow 70%, while the population will grow only a minuscule 3.7%, this will probably lead to a even more prosperous half century for this continent as GDP per capita will increase on a much faster pace than any other continent.

South America comes as an impressive surprise, after Europe it is the place where its predicted to improve the most where population will grow 34%, and GDP will grow an impressive 113%, a much expected relief for the quality of life of populations in there where if this happens as foreseen we can infer to see much less people in poverty or lack of infrastructure from governments.

North America will have a very moderate development in both areas remaining between 25% and 30%, the status quo on this area should not suffer any major changes.

Asia will see a very prospering half century as their economy is set to grow an astounding 86% percent while population will slow down and grow only 24.5% this can maybe have then getting close to Europe, NA and Oceania in quality of life in some countries but that depends on how this growth will be spread out in such a large and populous continent with quite large differences in moments for its biggest countries, we can see China for example getting much richer than India.

Africa will see its population growing almost 75% and GDP 66%, also its one that has the biggest range of on the forecast, even with a 75% confidence interval, their GDP could either grow almost 200% or decline 90%, that would be probably a symptom of many unstable countries both politically and economically unstable in the dataset that showed huge variance in the past. Most outside population and GDP growth forecasts like the UN Population Report[1], use only the most optimistic forecast for the economies and Africa, ignoring the downsides of these populations growths and expecting most countries to use it to its advantage, letting their hopes for a better Africa impact their analysis.

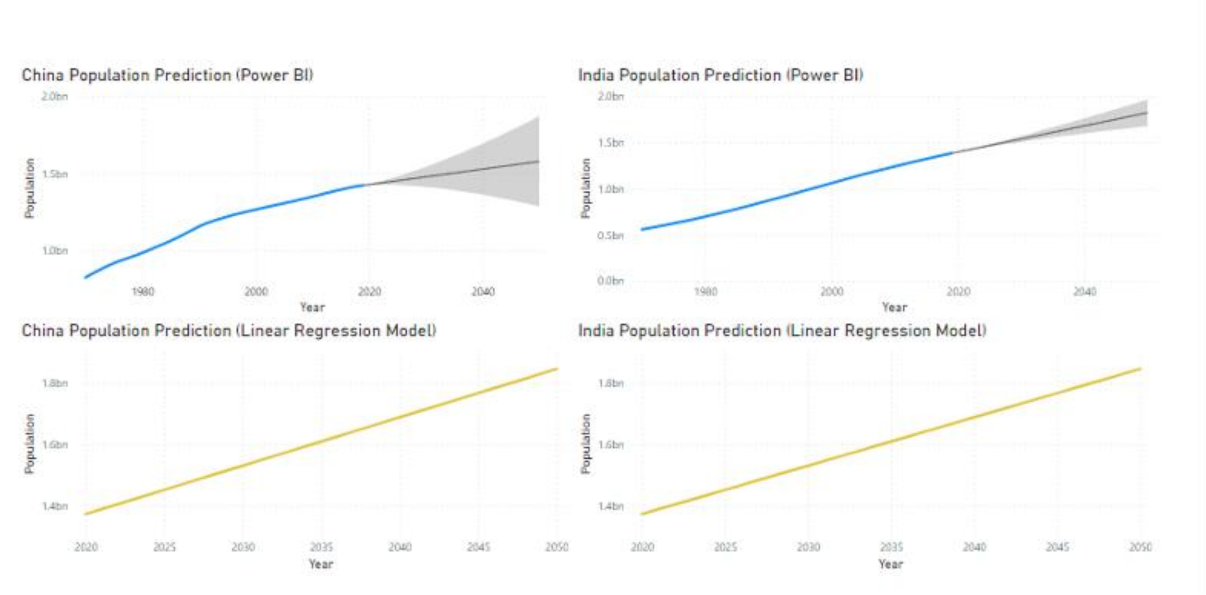
Oceania will have an very small Population growth of only 9.1%, but the GDP will raise as far as 52.8%, as a region with already quite a high standard of GDP per capita and quality of live we can only foresee it improving it even more and maybe even surpassing NA and EU as the most developed region.

Conclusions and Recommendations

Understanding the data is one of the most important parts of the analytical process, as we can check and validate missing values, so further analysis becomes more accurate, otherwise the missing data would probably have some influence in the results. Using a programming language combined with a virtual online environment (Google Colab) makes it easier to exchange the code between all the team members, where everyone can edit and run code and generate again the output file to be used in the analysis.

The data processing step is highly recommended to be the first step taken just after understanding the data. Spending some time checking, validating and manipulating data can save a lot of time ahead

When it comes to the performance of the two models, they both perform the task in completely different ways. The linear regression model takes the data from an overall perspective and predicts the future trend based on that, whilst the ETS model predicts the future points by considering the last two data points in order to predict the next. This gave a slight variance in the results as the regression model showed the general trend more as the ETS model had more accuracy on the smaller details. This can be proven by the two predictions of the total population for China and India below.



Between 2000 and 2019, the average life expectancy grew by more than 6 years, from 66.8 years in 2000 to 73.4 years in 2019. While the number of years lived with a disability has decreased, the number of years lived in good health has increased by 8%, from 58.3 in 2000 to 63.7 in 2019. This increase in healthy life expectancy is due to declining mortality. In other words, the 5.4-year rise in the length of a healthy life has not kept up with the overall increase in life expectancy (6.6 years).

The continents will see their population and GDP grow on different ratios, Africa despite most forecasts from other sources maybe will have the population growth more as a curse then a blessing, we can see that Asia, South America and Oceania will challenge even further Europe and North America on the Economic front, and that from what is seeing on these 3 continents having a slower population growth will have then focus n developing their economies on a higher pace.

Bibliography

Data Sets

UN department of Economic and Social Affairs Data Portal

https://population.un.org/wpp/

UN Data

https://data.un.org/

World Bank Open Data

https://data.worldbank.org/

International Monetary Fund – Open Database

https://www.imf.org/en/Data

Articles

[1] UN – 2019 World Population prospects

https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf

World Heath Organization mortality rate and life expectancy estimates

https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-life-expectancy-and-healthy-life-expectancy