Analyzing the GDP disparities between North and South American countries and the factors contributing to these differences

1 Introduction

Despite being part of America, North and South American countries have huge disparities in GDP. In this project, I took the United States as the clear leader and representative of North America, on the contrary, Brazil is often considered as the dominant country of South America. Then I aimed to investigate the GDP difference over the years 1989 to 2023 using metadata and data engineering approaches. This research and analysis includes significant changes in GDP over the decades and also feature**s** engineering for the factors finding which value most for both of the countries. By factor analysis, the data suggests that a wide range of selections can be found contributing to GDP. That includes sectoral contributions and investment patterns which gives a convex output for both cases. Two main questions will be answered through this data analysis process:

1. What are the primary economic, political, and social factors driving the GDP disparities between North and South American countries and is there any correlation that exists?

2. At which point significant GDP changes occurred for both countries over the years?

2 Used Data, Libraries and Methods

```
import pandas as pd
import numpy as np
import sqlite3
from sklearn.ensemble import
RandomForestRegressor
from sklearn.preprocessing
import LabelEncoder
import os
```

I. Four data sets are used to facilitate the analysis under the Open Database License (ODbL), Microdata Research License, and Creative Commons Attribution license (CC-BY 4.0). All the data are available in CSV format. Those data sets come with a wide range of attributes from the World Bank's official sites, which are mainly used to represent the global economy by yearly trends. Data

II. In the preprocessing phase of the project, I took the *'Country Name'*, and *'Country Code'* columns as relevant for GDP data, similarly for country Metadata *'Country Code'*, *'Region'*, and *'Income Group'* columns. For the indicator data processing used data. melt function to extract specific values for id_vars as "*Indicator Code"*, var_name as *"Year",* and value name as *"Value"*. Then merge the country meta data with GDP data where the "left" merging is used. To extract important features I used *RandomForestRegressor* to model the target variable (GDP) by assigning the scores to every independent variable. Therefore, *train_and_get_importance* defines the training of *RandomForestRegressor* while it learns the target columns and *label_encoder* used for the conversion of the country names and other text variables to the numeric value. Lastly, it produces a data frame where the weighted values are present to predict the factors associated most with the USA and Brazil's GDP. Furthermore, through filtering top indicator data was extracted and updated to the SQL lite database. Due to missing data on 3rd and 4th data set I had to eradicate and work with the data from 1989-2023 where in the meta data and the gdp data data can be found from earlier years.

III. Another question was to find the significant changes in gross domestic product value for both *countries find_top_gdp_change_year* defines this by taking the country metadata concerning the GDP data where it identifies the value year with the max values for absolute and percentage change, *country_data*

['GDP_Change'] = country_data['GDP'].diff() function calculate and compare the difference of both rows the current one and the previous one and therefore produce the changes in absolute. On the other hand, *country_data['GDP_Change_Percent'] = (country_data['GDP_Change'] / country_data['GDP'].shift(1)) * 100* to find the percentage values, it mainly computes the GDP continuously from one year to another and the shift function used to shift the data from current one to the previous year. Finally, *idxmax* finds the peak value for the GDP changes in the year. s*ave_to_sqlite()* function is used to save all the data in the different tables for later use and due to real-time accessibility. Additionally, I extracted all the data in CSV format for visible output. For dynamical allocation, I used Sq lite Database named *gdp_brazil_usa.db.*

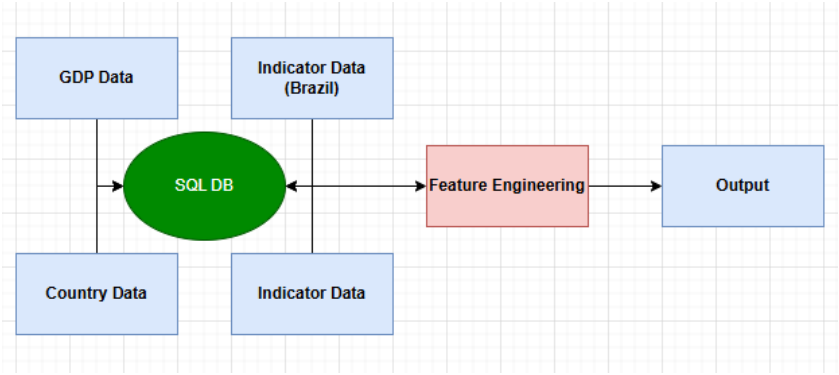

Fig 1 Pipeline Architecture

3 Analysis

What are the primary economic, political, and social factors driving the GDP disparities between the USA and Brazil, is there any correlation that exists?

In the analysis part mentioned before I used *Scikit-learn* for the feature engineering, to find the most important features contributing to the GDP of both countries. The function was limited to ten numbers (n) to facilitate the research, the *seaborn* library was used to visualize the top contributing features. The two plots (Fig 2 & Fig 3) represent the top ten indicators which contribute most to GDP growth for both countries and it is evident that the "Population" contribute most to the USA's GDP whereas "Export-import Business" is the primary driver for Brazil's GDP. Interestingly, the US GDP is influenced by demographic factors significantly, economic activities and technological aspects also make a major contribution. However, Brazil's economy mostly relies on trade, energy production and financial inclusion. A limited portion of the US GDP is also influenced by women's participation in parliament.

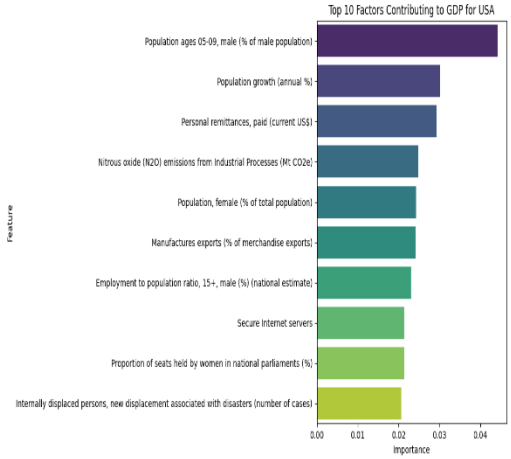

Fig. 2 Factors contributing To US GDP

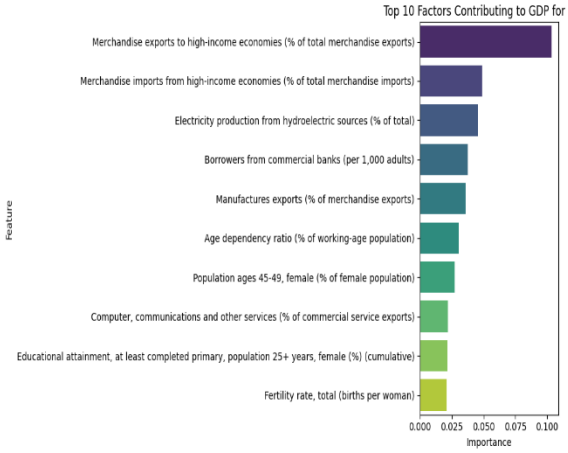

Fig .3 Factors contributing to Brazil's GDP

There is conclusive evidence that a positive correlation exists between population expansion and labour force participation rates in the USA. Additionally, a moderate correlation is visible between the Internet and financial flows. On the other hand, for Brazil Merchandise exports are explicitly related to each other, which reflects the economic structure of Brazil's GDP. However, there is limited correlation on the factors for both countries due to unique factors.

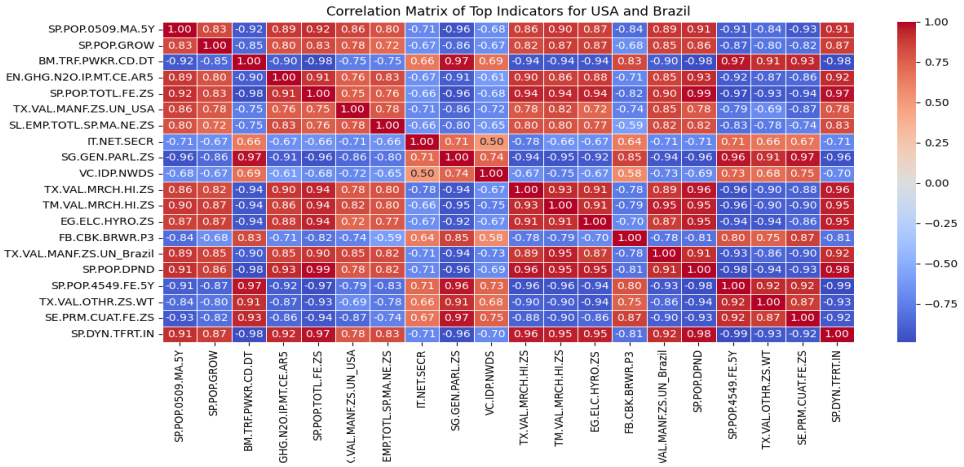

Fig. 4 Correlation Matrix of factors for USA and Brazil

Associated Codes	Factors Names	Value	Country
TX.VAL.MANF.ZS.UN TX.VAL.MRCH.HI.ZS	Manufactures exports (% of merchandise exports) Merchandise exports to high-income economies (% of total merchandise exports)	0.90	Cross Country
SE.PRM.CUAT.FE.ZS SP.POP.TOTL.FE.ZS	Educational attainment, at least completed primary, population 25+ years, female (%) Population, female (% of total population)	0.86	Cross Country
EG.ELC.HYRO.ZS TX.VAL.MRCH.HI.ZS	Electricity production from hydroelectric sources (% of total) Merchandise imports from high-income economies (% of total)	0.85	Brazil

At which point significant GDP changes occurred for both countries over the years?

The GDP metadata for the US is available from 1960, yet Brazil's data is available from 1989. Therefore, we ploted the graph from 1989 onward. Clearly, The United States of America's GDP consistently demonstrates an upward trend. However, here we found two major breakdowns of US GDP in 1998 and 2019. The USA had the highest changes in 2021**,** which increased by 12.96% of gross GDP. On the other hand, Brazil's GDP has always been on fluctuations. It acquired a significant upraise in 2010 when it changed 46.44% of total GDP compared to the previous year, driven by infrastructure investment. There were visible declines for both countries during the global economic crisis in 2009 and also during the COVID-19 pandemic in 2020.

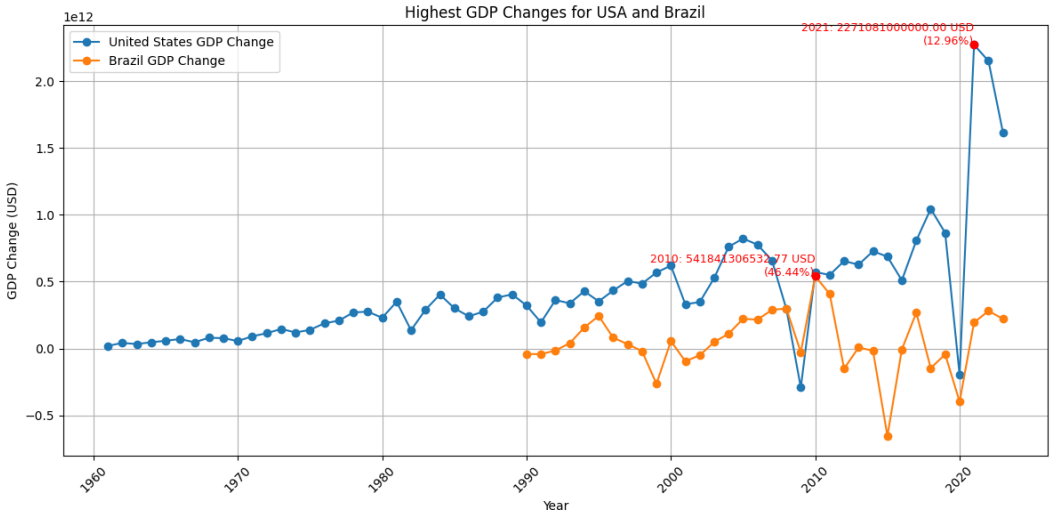

Fig. 5 GDP changes for Brazil and USA

4 Prospective Developments and Limitations

This project has reasonable limitations by reason of the smaller area selection and to avoid computational complexity. You may find overfitting problems due to missing data, randomising data and scalability. More generalized output can be acquired through sophisticated learning. Though this analysis is in the optimal position due to limited attribute selections, further factor analysis is possible if we take into account more variables. For example, statistical analysis for every factor, predictive forecasting for decision-making[1], sustainability analysis[2] and many more. Here I worked with only Indicator Data, GDP data, and Country Meta data. In addition, Income group, Region for every state, Age grouping and other attributes are possible selections and I believe dynamic insights can be extractable. To draw an accurate conclusion at every step data-driven approaches and training processes are crucial.

5 Conclusion

This study reveals that in the USA's diversified economy where human capital plays a vital role, to bridge the gap, Brazil should prioritize enhancing the educational sector and adopting innovation-driven approaches rather than depending on external factors. For instance, ensuring stable internet service nationwide. Aditionally, Brazil should focus on hydroelectric energy production relates to economic imports.

[1] https://www.researchgate.net/publication/326108053_DID_Analysis_on_the_Impact_of_Policies_on_the_Rural-Urban_Income_Disparity_in_Resource-Dependent_Regions_A_Case_Study_of_Ordos

[2] https://www.sciencedirect.com/science/article/pii/S0921800924002052