The Role of Infrastructure in Economic Growth: Analyzing Electricity Production and GDP Growth

Md Sohan Sorker
23074752

Economic growth has a significant impact on infrastructure, especially in developing nations like Latin America. This study looks at the correlation between GDP growth (yearly percentage) and power generation (kWh per capita) in a few countries in Latin America between 2015 and 2021. The study looks for trends, patterns, and connections between economic growth and energy infrastructure through analysis of two World Bank databases. Governments and investors will benefit from the insights produced, which will assist them maximize infrastructure development for the purpose to encourage sustainable growth.

I. QUESTION

How does electricity production (kWh per capita) correlate with economic growth (GDP growth) in Latin American countries from 2015 to 2021?

This inquiry aims to determine whether increased electricity production—a major indication of energy infrastructure—contributes to economic expansion. By examining the relationship between these two factors, the research aims to offer practical insights into how infrastructure supports economic stability and development.

II. DATA SOURCES

Two datasets that provide thorough information on economic growth and power production have been considered for this project: GDP Growth (annual percentage) and Energy Production (kWh per capita). The World Bank, widely recognized worldwide source of economic and development data, is the source of the datasets. The following are the primary reasons for choosing these datasets: -Their thorough coverage of important metrics, such as yearly GDP growth and power output per capita, which are crucial for examining the connection between infrastructure development and economic success.

A. *Data Structure*

The Electricity Production dataset is structured with temporal variables (year), categorical variables (country name and country code), and continuous variables (electricity production (kWh per capita)). Notably, the dataset is highly reliable, with minimal missing values, ensuring robust data quality for analysis.

The GDP Growth dataset is structured with temporal variables (year), categorical variables (country name and country code), and continuous variables (GDP growth (annual %)). The dataset is notable for being accurate and free of major missing values, providing high-quality data for precise analysis.

B. *Data Quality*

The **Electricity Production dataset** has several key dimensions of data quality. Since the data accurately represents actual electricity production levels across nations, including annual trends and electricity output per capita, accuracy is maintained. The dataset's timeliness covers many years (2015–2021), presenting a thorough perspective for identifying patterns and trends in the evolution of energy infrastructure. The data may not, however, accurately represent the effects of recent developments in energy production technology or regulations because it is based on previous years. By focusing on one important metric—electricity output (kWh per capita)—that is closely related to the development of infrastructure and the availability of energy supplies, two important determinants of economic growth, relevance is attained.

Histograms and Kernel Density Estimation (KDE) plots were utilized to visually analyze the distributions and identify anomalies within the Electricity Production dataset for its categorical variables. The data is evenly divided among the years (2015–2021), according to the histograms, offering balanced temporal coverage for trend analysis. Significant cross-country comparisons are made possible by the country-level distributions, which also display an even representation of records for Brazil, Mexico, Argentina, Colombia, and Chile. No significant anomalies were observed, ensuring that the categorical variables accurately reflect real-world distributions and are suitable for further analysis.

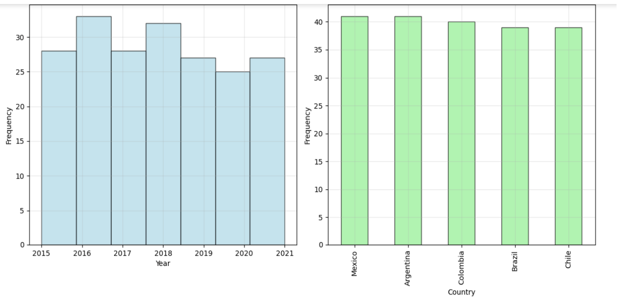

Fig. 1. Categorical features from the Electricity Production

Similarly, the KDE plots for continuous features in the Electricity Production dataset reveal distinct patterns for each country. The distribution in Brazil is somewhat bimodal, suggesting that there have been two main levels of electricity output over time, which could be related to changes in infrastructure or policy. Mexico's bell-shaped, balanced distribution shows stability in power generation, with most values centered around the mean. Argentina, on the other hand, exhibits a tighter curve, suggesting consistent infrastructure performance and less variation in production levels. A steep peak in Colombia's KDE plot indicates a high concentration of output levels close to the mean, most likely because of stable energy regulations. Finally, Chile displays a broader distribution, suggesting a wider range of production values that could be influenced by varying energy sources or changing dynamics. Overall, these KDE plots provide insights into the differences in electricity production across countries while confirming that the data is well-suited for analysis.

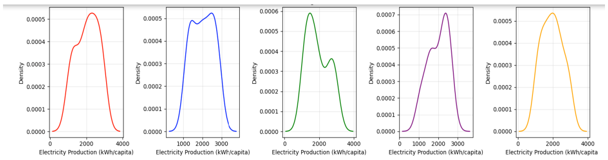

Fig. 2. *Continuous* features from the Electricity Production

The **GDP Growth dataset** demonstrates strong data quality, precisely representing the yearly GDP growth rates of Latin American nations including Argentina, Brazil, and Mexico. Its World Bank source provides reliability and accuracy with few missing data points. The dataset, which covers the years 2015–2021, offers a pertinent time for examining growth, stability, and volatility in the economy. The dataset remains beneficial for examining how policy and infrastructure changes affect economic development because it focusses on GDP growth, a crucial indicator of economic performance. Comprehensive cross-country comparisons and trend analysis are ensured by the temporal and country-specific structure. It becomes a beneficial instrument for producing actionable insights because of its organized structure, comprehensiveness, and compatibility to real-world circumstances.

I examined the distributions of the categorical features in the GDP Growth dataset using histograms. To ensure balanced temporal coverage for trend analysis, the histograms demonstrate that the data is evenly distributed across 'Year'. In the same way, 'Country' shows a balanced representation of the picked Latin American nations, such as Brazil, Mexico, Argentina, Colombia, and Chile. Thanks to this balanced distribution, the data is dependable for examining trends in economic growth and appropriate for cross-country comparisons.

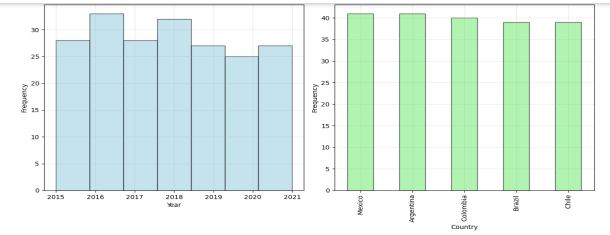

Fig. 3. Categorical features from the GDP Growth

Similarly, the KDE plots for continuous features in the GDP Growth dataset indicate distinct patterns for the data. The feature 'GDP Growth (annual %)' shows bell-shaped curves for most countries, which approximate a normal distribution and show stable economic performance with values centered around the mean. Nonetheless, there are differences between nations, and certain distributions exhibit a small degree of skewness. These trends demonstrate the dataset's wide range of economic circumstances and patterns**.**

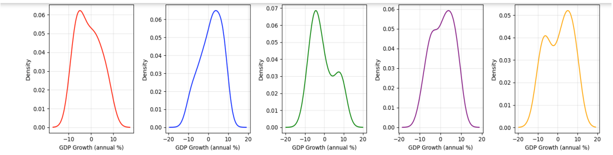

Fig. 4. Continuous features from the GDP Growth

C. License

Both datasets used in this project, Electricity Production (kWh per capita) and GDP Growth (annual %), are licensed under the World Bank Open Data License. This license allows the data to be used, shared, and modified for any purpose as long as the World Bank is properly credited as the source. I will make sure that the World Bank is properly credited in all uses, including reports, presentations, and publications that come from this analysis, in order to abide by the licensing requirements. Furthermore, in order to preserve transparency and allow for future usage by other scholars and practitioners, any derived or modified versions of the datasets will appropriately credit the World Bank as the original source.

III. DATA PIPELINE

For this project, I created an automated data pipeline using Python that retrieves the dataset from the internet, cleans and transforms it, and stores it in the /data directory. I used ETL (Extraction, Transformation, and Loading) as my data pipeline architecture. Extracting raw datasets from the internet is the first step in the procedure. After that, the data undergoes several transformation processes to guarantee usability, including cleaning, imputation missing values, and reformatting the data as needed (Transformation). In order to facilitate easy retrieval and additional analysis, the converted data is finally saved into a structured database (loading).

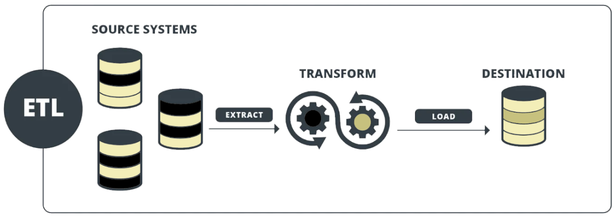

Fig. 5. The ETL data pipeline architecture

A. Data Extraction

The two datasets are downloaded as zip files using the requests library to start the extraction process. I used the Zip File module to access the zip archive after downloading, and then I extracted the pertinent CSV file. Next, use the pandas read csv method to read the file.

B. *Data Transformation and Cleaning*

After extraction, the Electricity Production and GDP Growth datasets go through several transformation and cleaning steps. Unnecessary columns, such as metadata and indicator codes, are removed to simplier To ensure reliability, rows with large missing values—such as data missing for several years in a row—are eliminated. In order to replace NaNs for other missing values, linear interpolation is used to estimate values from nearby data points. By maintaining the temporal linkages across successive years, this approach guarantees data continuity and consistency. To guarantee effectiveness and reproducibility, the pandas package in Python is used to implement all transformations.

C. *Loading Data into the Sink*

The cleaned and converted data must then be saved into an SQLite database as the last step. In order to connect to the database and store the datasets in two specific tables ("Electricity Production" and "GDP Growth"), the sqlite3 module is utilized. This method guarantees that the data is effectively saved and is always accessible for analysis or querying in the future.

During the development of the pipeline, one practical challenge was managing the temporary storage of downloaded files on the local machine. To address this, the pipeline was automated to delete temporary files after they were successfully processed and saved to the SQLite database, ensuring optimal use of local storage.

Now, though, the pipeline is not flexible enough to accommodate dynamic changes in the datasets. For example, a specific temporal range (2015–2021) is assumed by the static interpolation approach used to handle missing information. The pipeline will need to be manually updated to account for any new data for later years or additional nations, which could restrict its scalability. In order to increase the pipeline's versatility, this problem will be fixed in subsequent versions.

IV. Result and Limitations

A. *Output Data of the pipeline*

The output of the data pipeline is an SQLite database that contains two distinct tables ("Electricity Production" and "GDP Growth") with cleaned and transformed data from the respective datasets. Because there are no notable missing values, the output datasets are complete and reliable for analysis. With their temporal, categorical, and continuous variable structures, both tables offer thorough coverage of key elements including year, nation, and important metrics (GDP growth and power output). The datasets are guaranteed to be well-structured and prepared for additional analysis thanks to this arrangement.

B. *Why SQLite?*

I chose the SQLite format as the pipeline output due to its lightweight nature and ease of integration with various programming languages such as Python and R. Another reason is that SQLite databases are portable and easy to share with collaborators

C. *Critical Reflection on Data and Potential Issues*

There are very few missing values in the pipeline's energy production (kWh per capita) and GDP growth (annual percentage) databases, which are largely accurate and comprehensive and represent actual trends in electricity production and economic performance. To deal with missing data, the interpolation method's static nature may provide problems when adding new data, such future years or other nations. Future analyses could contain inconsistencies if the pipeline isn't updated dynamically.

The dataset's cross-country character enhances the research by providing a more comprehensive view of the relationship between infrastructure development and economic growth. However, variations in local energy systems, economic policies, and data gathering techniques among nations may create variability that could affect the consistency of the findings. Although the datasets offer insightful historical information, they may not adequately reflect more recent trends because they concentrate on the years 2015–2021, which could compromise their timeliness and usefulness.

Some features, such as minor variations in country-level GDP growth rates, may not initially appear significant but are retained in the dataset to allow for a more comprehensive analysis. By including these characteristics, it may be possible to find subtle patterns or correlations that are important for comprehending how the development of energy infrastructure affects long-term economic stability and growth in various geographical areas. For the final analysis, this method guarantees a comprehensive and nuanced viewpoint.

V. Conclusion

To sum up, the data pipeline effectively used Python and an ETL framework to handle the extraction, transformation, and loading of the GDP Growth and Electricity Production datasets. Effective and portable data integration was provided by using SQLite for data storage. The pipeline has drawbacks, including difficulties dynamically adjusting to new data inputs, despite its strength in processing structured datasets.

References

World Bank Open Data Licensing. (n.d.). Terms of Use for Datasets.
Retrieved from https://data.worldbank.org/about/terms-of-use

“Data Pipeline Architecture - A Deep Dive — StreamSets,” Software AG. (accessed Jun. 03, 2024).