Analysis of World Sustainability Dataset

[redacted]

Western Governors University

Table of Contents

A. Proposal Overview

A.1 Research Question or Organizational Need

Do nations with higher GDP's operate in a more sustainable way compared to countries with much lower GDP's? This will inform organizations involved in sustainable development in decision-making and future planning.

A.2 Context and Background

The world's climate is changing quickly and, as a result, a global emphasis on sustainability has increased in popularity. It would be very useful to see which nations' practices result in effective sustainability and how that relates to their GDP. If it's true that wealthier nations are more sustainable, it would challenge the idea that more wealth equals more environmental damage. Furthermore, wealthier nations could potentially support poorer countries in best practices to improve global sustainability.

The connection between wealth and sustainability could also provide a road map for developing countries to become more economically robust and eventually adopt sustainable practices without harming the well-being of their populations. Increasing public awareness about this connection will also likely lead to an increase in political will for effective policy change.

A.3 and A3A Summary of Published Works and Their Relation to the Project

Review of Work 1

"Taking a Comprehensive View of Wealth to Meet Today's Development Challenges."

Summary: This article emphasizes that only measuring GDP doesn't fully capture a nation's economic health. GDP can encourage short-term gains at the cost of long-term well-being. The "Changing Wealth of Nations 2021" report referenced in the article suggests considering other factors, like a country's natural and human resources, alongside GDP. (DataBank | the World Bank, n.d.)

Relation to research question: This article introduces the idea of national wealth, encompassing resources and human capital, and implies that a focus on GDP alone might sacrifice sustainability for immediate growth. This highlights that a country with a high GDP might not always operate sustainably. To fully understand sustainability, factors beyond GDP, such as national wealth, should be evaluated.

Review of Work 2

"A radical shift to working with nature."

Summary: This article stresses the need to work with nature instead of exploiting it for economic and social progress. It highlights the rapid decline of natural resources and warns of economic risks if this continues. Brazil and Indonesia are commended for their sustainable approaches to nature, and other countries are encouraged to do the same. (United Nations Environment Programme, n.d.)

Relation to research question: The article argues that GDP growth doesn't necessarily indicate sustainable or healthy economies, especially when nature is exploited. It points out that even with rising GDP, natural resources are declining, highlighting a gap between economic metrics and environmental well-being. This suggests that high-GDP nations might not be more sustainable than low-GDP ones.

Review of Work 3

"Countries can tap tax potential to finance development goals."

Summary: This article highlights how emerging markets and developing economies can use untapped tax potential to fund their development and climate goals. By improving tax designs and bolstering public institutions, countries can raise their tax-to-GDP ratios by 9%. This increase can boost financial growth and private business activity, leading to sustainable development. (Countries Can Tap Tax Potential to Finance Development Goals, 2023)

Relation to research question: This article suggests that developing economies can achieve sustainability by improving their taxation methods, irrespective of their current GDP. Beyond GDP, the effectiveness of a nation's tax design and the strength of its institutions are key indicators of sustainable operations. Therefore, a high GDP doesn't directly imply sustainable practices or the capability to fund development sustainably.

A.4 Summary of Data Analytics Solution

Using the CRISP-DM approach, I explored whether countries with higher GDPs use more sustainable practices than those with lower GDPs, a useful insight for sustainable development. I examined the World Sustainability Dataset, then cleaned and prepared the dataset for statistical analysis.

My main tool was the Independent Samples T-test. With an alpha (α) of 0.05, I assessed the sustainability differences between the two GDP groups: low-income countries and high-income countries. Additionally, I used Logistic Regression to classify nations by sustainability, focusing on their GDP. I targeted an accuracy of 80% and an F-score of 0.75 or above.

A.5 Benefits and Support of Decision-Making Process

My data analytics solution delivers the following key insights for sustainable development organizations:

1. **Better Decision-making**: By understanding the tie between GDP and sustainability, leaders in developing countries may prioritize economic growth as a route to sustainability.
2. **Resource Efficiency**: Identifying high-GDP nations with low sustainability ensures that international aid is used where it's most needed.
3. **Targeted Efforts**: Insights highlight where sustainability efforts are needed most.
4. **Guided Policy Making**: The insights inform policymakers to frame rules that bolster both economic and sustainable growth.

B. Data Analytics Project Plan

B.1 Goals, Objectives, and Deliverables

Goal: Determine the relationship between a nation's GDP and its sustainability practices.

- **Objective 1:** Collect and clean the World Sustainability Dataset to ensure accurate and consistent data is used for analysis.
    - **Deliverable 1.1:** A cleaned and processed version of the World Sustainability Dataset in Excel format.
- **Objective 2:** Analyze the dataset to identify trends, correlations, or patterns between GDP and sustainability metrics.
    - **Deliverable 2.1:** A Jupyter Notebook detailing the statistical analysis, including correlation coefficients, p-values, and any regression models if applicable.
- **Objective 3:** Visualize the findings in a comprehensible manner to communicate results to stakeholders.
    - **Deliverable 3.1:** A Tableau dashboard showcasing the relationship between GDP and sustainability, broken down by regions, GDP brackets, and other relevant dimensions.

- **Deliverable 3.2:** A presentation summarizing key findings, implications, and recommendations for stakeholders.

B.2 Scope of Project

B.2.A Included in Project Scope

1. **Data Collection and Cleaning:** Obtain and clean the World Sustainability Dataset.
2. **Statistical Analysis:** Use regression models to find trends between GDP and sustainability.
3. **Visualization:** Make clear visuals with Tableau to show GDP and sustainability connections.
4. **Documentation:** Write a detailed report about the entire project, from start to finish, including all tools and methods used.
5. **Stakeholder Communication:** Hold a meeting to share findings with stakeholders.

B.2.B Not included in Project Scope

1. **Primary Data Collection:** I won't collect new data myself; I'll use the World Sustainability Dataset.
2. **Longitudinal Analysis:** I'm only looking at the current dataset, not tracking changes over time or predicting future trends.
3. **In-depth Case Studies:** I won't be doing detailed studies or visits to specific countries.

B.3 Standard Methodology

CRISP-DM (Cross Industry Standard Process for Data Mining)

1. **Business Understanding (Analysis Phase)**

- **Project Activity:** Defining the objective of our analysis: understanding the relationship between GDP and sustainability.
- **Output:** Clear project goals, objectives, and deliverables.

2. **Data Understanding (Analysis Phase)**
    - **Project Activity:** Reviewing the World Sustainability Dataset, identifying its structure, quality, and potential gaps or anomalies.
    - **Output:** Initial report on data quality and characteristics.
3. **Data Preparation (Design Phase)**
    - **Project Activity:** Cleaning and preprocessing the dataset. This includes handling missing values, outliers, and potential discrepancies.
    - **Output:** A clean, processed dataset ready for modeling and visualization.
4. **Modeling (Design & Development Phase)**
    - **Project Activity:** Applying statistical methods to identify trends, patterns, and correlations between GDP and sustainability metrics.
    - **Output:** Statistical models, correlation matrices, and initial insights.
5. **Evaluation (Testing Phase)**
    - **Project Activity:** Assessing the validity and reliability of the statistical findings. Confirming if the models fit the data well and if the insights align with the project goals.
    - **Output:** A report on model performance, insights validation, and any required refinements.
6. **Deployment (Implementation Phase)**
    - **Project Activity:** Translating insights into visualizations using tools like Tableau, creating reports, and presenting findings to stakeholders.
    - **Output:** Finalized Tableau dashboards, comprehensive reports, and presentations.

B.4 Timeline and Milestones

Milestone	Duration (days)	Projected Start Date	Anticipated End Date
Data Collection and Cleaning	2 days	9/25/2023	9/26/2023
Initial Data Understanding Report	1 day	9/27/2023	9/27/2023
Data Preparation	3 days	9/28/2023	9/30/2023
Statistical Modeling	2 days	10/1/2023	10/2/2023
Evaluation of Findings	1 day	10/3/2023	10/3/2023
Visualization Creation	2 days	10/4/2023	10/5/2023
Report Compilation and Documentation	2 days	10/6/2023	10/7/2023
Stakeholder Communication & Feedback	1 day	10/8/2023	10/8/2023
Final Adjustments & Project Closure	1 day	10/9/2023	10/9/2023

B.5 Resources and Costs

Personnel:

- **Me:** $0/hour (estimated 70-80 hours for project completion)

Technology:

- **World Sustainability Dataset:** Free (CSV File format)
- **Python:** Free (Open Source)
- **Tableau:** Free (Public Edition)
- **Excel:** Included with Microsoft Office Suite (Already owned)
- **Other Libraries (pandas, numpy, seaborn, etc.):** Free

Infrastructure:

- **Personal Computer:** N/A (Already owned)
- **Internet Connection:** N/A (Fixed monthly rate, unrelated to this project's duration)

Additional Resources (if required):

- **Backup Storage (e.g., Cloud storage):** $0 (using existing storage solutions)
- **Online Forums & Community Support (Stack Overflow, Tableau Community):** Free

The main investment for this project is my time, roughly 70 to 80 hours across two weeks. All the tech tools used are free. I'm using my own computer and internet, which weren't purchased for this project, so there's no cost tied to them here.

B.6 Criteria for Success

1. **Data Integrity and Cleaning:**
    - **Criterion/Metric:** Percentage of data inconsistencies and missing values addressed.
    - **Required Data:** Data discrepancy and null value reports.
    - **Cut Score for Success:** 100% of data discrepancies and missing values resolved.
2. **Statistical Significance:**
    - **Criterion/Metric:** P-value of the correlation between GDP and sustainability metrics.
    - **Required Data:** Results of correlation testing.
    - **Cut Score for Success:** p-value < 0.05, indicating the relationship is statistically significant.
3. **Documentation Comprehensiveness:**
    - **Criterion/Metric:** Coverage of every step in the project process in the documentation.
    - **Required Data:** Documentation review.
    - **Cut Score for Success:** Complete documentation of all steps, from data collection to recommendations.

C. Design of Data Analytics Solution

C.1 Hypothesis

Null Hypothesis (H_0): The average sustainability score is the same for nations with high GDPs as it is for nations with low GDPs.

Alternative Hypothesis (H_1): The average sustainability score is higher for nations with high GDPs compared to nations with low GDPs.

C.2 and C.2.A Analytical Method

The statistical method I will employ for this project is the **Independent Samples T-test**. This method checks for significant differences in averages between two separate groups. For this project, I'm comparing sustainability scores of high GDP nations to those of low GDP nations.

My goal is to see if the sustainability scores differ notably between these two groups. The T-test is a good fit for this, as it will tell me if any observed differences are likely real or just due to chance. Using this test, I'll determine if high GDP nations actually have better sustainability scores than low GDP ones.

C.3 Tools and Environments

I'll be using the following tools and environments for this project's data analytics:

1. **Python:** The main language for my analysis due to its strong data and statistical libraries.
2. **Jupyter Notebooks:** A simple-to-use platform for coding and visualizing the Python script because of its interactive feedback.
3. **Pandas:** Used for efficiently managing our CSV dataset, from loading to cleaning and analysis.

4. **Tableau:** For advanced visualizations after our initial Python analysis.
5. **Excel:** For quick data checks and minor adjustments.
6. **SciPy and StatsModels:** These Python libraries will support statistical tests.

C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance

Statistical Test Details:

- **Null Hypothesis (H_0):** No difference in sustainability scores between high and low GDP nations.
- **Test:** Independent Samples T-test.
- **Metric:** T-statistic leading to p-value.
- **Alpha (α):** 0.05. If p-value < 0.05, the null hypothesis is rejected, indicating a significant difference in scores.

Model Details:

- **Type:** Supervised Classification (based on sustainability scores).
- **Algorithm:** Logistic Regression.
- **Performance Metrics:** Accuracy, F-score.
- **Success Criteria:** Accuracy ≥ 80%, F-score ≥ 0.75.

Method and Metric Rationale:

- **T-test:** To compare sustainability scores of two GDP groups, the T-test is ideal. It determines if mean differences are significant or random.
- **Logistic Regression:** Perfect for binary classification, like classifying nations by sustainability scores. It also highlights influential features like GDP.
- **Metrics:**
    - **Accuracy:** Indicates correct predictions.
    - **F-score:** Balances false positives and negatives, useful for imbalanced classes.

C.5 Practical Significance

I'll be weighing practical significance using the following:

1. **Difference Size**: Does the gap in sustainability scores between high and low GDP nations matter in practical terms? A small statistical difference may not call for policy shifts, but a big one might.
2. **Insight Actionability**: Can we act on the results? If high GDP countries score significantly better in sustainability, it may show that a strong economy fosters sustainable actions. Although instead, I may find that the inverse is true. Either result can guide countries aiming to enhance both areas.
3. **Decision-making Integration**: Can findings smoothly fit into decision-making? For example, if a nation sees a tie between GDP and sustainability, they might lean towards strategies benefiting both.

If there is a strong correlation with a high GDP and sustainability scores, a nation wishing to uplift its sustainability might focus on both economic and sustainable growth. This could involve backing green tech, supporting eco-friendly enterprises, or incentivizing sustainable actions.

C.6 Visual Communication

I'll be using the following visual tools to communicate my findings:

1. Global Map with Sustainability Scores:

- **Description:** A map where countries are colored based on sustainability scores. Dark green signifies high scores, while light green indicates low scores, providing a quick global view of sustainability.
- **Purpose:** Allows stakeholders to quickly grasp global sustainability patterns.
- **Tool:** Tableau, for its strong mapping features.

2. Bar Chart: High-Income Countries' Sustainability Scores:

- **Description:** Vertical bars for each high-income country, with height and color depth representing sustainability score.
- **Purpose:** Lets stakeholders compare sustainability in high-income countries.
- **Tool:** Tableau, for its detailed visualization capabilities.

3. Bar Chart: Low-Income Countries' Sustainability Scores:

- **Description:** Vertical bars for each low-income country, with height and color depth representing sustainability score.
- **Purpose:** Offers a view of sustainability in low-income countries, highlighting areas for improvement.
- **Tool:** Tableau, for uniformity in visual presentations.

D. Description of Dataset

D.1 Source of Data

I will be using the following dataset from Kaggle:

- World Sustainability Dataset (2021) – World Bank Data Bank; United Nations Statistics; Our World in Data (https://www.kaggle.com/datasets/truecue/worldsustainabilitydataset?select=WorldSustainabilityDataset.csv)

D.2 Appropriateness of Dataset

The World Sustainability Dataset (2021), derived from trusted organizations like the World Bank and the United Nations, is ideal for this study:

1. **Credibility**: Backed by reputable sources, the dataset offers accurate and comprehensive data, reducing biases.

2. **Data Scope**: It compiles data from multiple institutions, providing a holistic view on global sustainability.
3. **Current**: From 2021, this dataset reflects recent sustainability patterns, ensuring timely insights.
4. **Research Alignment**: Focusing on the GDP-sustainability relation, the dataset's economic and sustainability details match our needs.

D.3 Data Collection Methods

The .csv file was downloaded directly from the Kaggle website (https://www.kaggle.com/datasets/truecue/worldsustainabilitydataset?select=WorldSustainabilityDataset.csv).

D.4 Observations on Quality and Completeness of Data

The World Sustainability Dataset is, for the most part, accurate and complete. Here's an overview:

Quality:

- **Country Name and Year**: Entries are consistently categorized with no missing values.
- **Income Classification**: Only 0.06% of the data is missing, highlighting a reliable categorization of countries by income.

Completeness:

- **Renewable electricity output**: About 15.7% of data is missing, suggesting potential reporting gaps or unavailable data for certain years or countries.
- **GDP**: Only 1.25% data is missing, marking it as a reliable metric.
- **Renewable energy consumption**: Fully populated, reflecting comprehensive reporting.

Accommodations:

- For the Income Classification, I'll manually impute the null values based on my own judgement.
- For the 'Renewable electricity output' column, imputation methods like mean filling or model predictions could be used.
- For the minor gap in the GDP data, using mean or median values can fill in the gaps.

D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

Data Governance:

- **Relation**: Data governance ensures our dataset is managed and traceable.
- **Precaution**: Use standardized protocols for data handling.

Privacy:

- **Relation**: The dataset focuses on country-level data, not personal information.
- **Precaution**: Don't mix with other datasets that risk identifying individuals.

Security:

- **Relation**: Safeguarding against data breaches protects reputation and legal standing.
- **Precaution**: Use password-protection for PC access.

Ethical, Legal, and Regulatory Compliance:

- **Relation**: Ethical use of data avoids misrepresentation, and public data can have usage terms.
- **Precaution**: Review dataset terms regularly. Present findings neutrally, and consult legal advice as needed.

D.5.A Precautions:

- **Data Governance**: Maintain clear records of data changes.
- **Privacy**: Use anonymization if combining datasets. Log data access.

- **Security**: Store data securely with authentication measures.
- **Ethical & Legal Considerations**: Ensure findings are reviewed legally before public release.

References

World Bank Group. (2021). Taking a Comprehensive View of Wealth to Meet Today's Development Challenges. World Bank. https://www.worldbank.org/en/news/feature/2021/10/27/taking-a-comprehensive-view-of-wealth-to-meet-today-s-development-challenges

United Nations Environment Programme. (n.d.). A radical shift to working with nature. UNEP. https://www.unep.org/news-and-stories/speech/radical-shift-working-nature

Countries can tap tax potential to finance development goals. (2023, September 19). IMF. https://www.imf.org/en/Blogs/Articles/2023/09/19/countries-can-tap-tax-potential-to-finance-development-goals