Analysis of World Sustainability Dataset

[redacted]

Western Governors University

Table of Contents

A. Project Highlights

This project aimed to see if countries with higher GDPs are more sustainable than those with lower GDPs. The project involved cleaning the World Sustainability Dataset, analyzing GDP and sustainability trends, visualizing results with Tableau, and sharing my findings. It excluded new data collection, long-term analysis, and in-depth country studies. Using the CRISP-DM method, I systematically worked through each stage, from understanding the goal to sharing my findings.

I used various tools like Python, Jupyter Notebooks, and Tableau for detailed and clear analysis. Excel helped with quick data checks. This structured approach gave me reliable insights into the relationship between GDP and sustainability.

B. Project Execution

I stuck to the plan outlined in Task 2, all goals, objectives, and deliverables were met without variance. I planned to use the entire CRISP-DM methodology, but the process was faster than expected. Although the initial stages were as planned, later phases like modeling and evaluation overlapped, speeding up the work. While I initially projected a 14-day timeline, I completed the project in just 5 days. Here's the revised timeline:

Milestone	Duration (days)	Projected Start Date	Projected End Date	Actual End Date
Data Collection and Cleaning	2 days	9/25/2023	9/26/2023	9/25/2023
Initial Data Understanding	1 day	9/27/2023	9/27/2023	9/25/2023
Data Preparation	3 days	9/28/2023	9/30/2023	9/26/2023
Statistical Modeling	2 days	10/1/2023	10/2/2023	9/26/2023
Evaluation of Findings	1 day	10/3/2023	10/3/2023	9/27/2023
Visualization Creation	2 days	10/4/2023	10/5/2023	9/28/2023
Report Compilation	2 days	10/6/2023	10/7/2023	9/28/2023
Stakeholder Feedback	1 day	10/8/2023	10/8/2023	9/29/2023
Project Closure	1 day	10/9/2023	10/9/2023	9/29/2023

C. Data Collection Process

I followed my original data selection and collection plan, obtaining the World Sustainability dataset as a .csv from Kaggle. No significant issues were faced during data collection and no unexpected data governance problems occurred, either.

Data Set Pros and Cons

Pros:

Depth of Data: The dataset offered a wide view of sustainability over a 19-year period, including metrics like renewable electricity production, renewable energy consumption, income classification, and GDP. Complete data on 'Renewable energy consumption' enabled detailed analysis across 173 countries.

Cons:

Data Gaps: Some inconsistencies and missing data, like the 15.7% gap in 'Renewable electricity output', might influence analysis accuracy.

D. Data Extraction and Preparation

The data was downloaded from Kaggle in .csv format. Tools involved:

- **Web Browser (Chrome)**: Used to access Kaggle and download the dataset.
- **CSV Viewer (Excel)**: To initially view and inspect the dataset.

After extraction, data was prepared through:

- **Cleaning**: Filling in missing data points, such as in 'Income Classification'.
- **Formatting**: Making sure data is consistent.
- **Exploratory Analysis**: Briefly understanding the dataset's patterns.

Tools used during preparation:

- **Python with Pandas**: For cleaning and exploration.
- **Jupyter Notebook**: To run Python code and visualize data.

These tools were chosen because they provide a thorough method for preparing data, ensuring it's ready for analysis. The preparation steps were crucial for:

- **Accuracy**: Manual imputation for certain missing values, like 'Income Classification', ensured decisions were based on domain knowledge and data context.
- **Complete Analysis**: Techniques like mean filling and model predictions for missing 'Renewable electricity output' and ‘GDP’ values provided a more complete dataset for analysis.
- **Research Needs**: Clean and consistent data is a must, especially when focusing on GDP and sustainability.

E. Data Analysis Process

E.1 Data Analysis Methods

Independent Samples T-test:

- **Description**: Compares means of two groups to find statistically significant differences.
- **Application**: Used to see the difference in sustainability scores between high and low GDP countries.
- **Rationale**: It's best suited for comparing two separate groups on a single metric.

Logistic Regression:

- **Description**: A model for binary classification tasks, estimating probability of category membership.
- **Application**: Classified nations by their sustainability scores.
- **Rationale**: Ideal for predicting binary outcomes and understanding relationships between variables.

E.2 Advantages and Limitations of Tools and Techniques

Independent Samples T-test:

- **Advantage**: Directly compares means of two groups, offering clear statistical insights.
- **Limitation**: Requires normally distributed data with equal variances in both groups.

Logistic Regression:

- **Advantage**: Gives both classification and a confidence score, capturing linear relationships well.
- **Limitation**: Assumes a linear boundary; not great for nonlinear relationships and needs a good sample size.

E.3 Application of Analytical Methods

Independent Samples T-test:

- **Steps**:
    1. Grouped data by GDP.
    2. Checked for normal distribution and equal variances.
    3. Ran the T-test.
    4. Interpreted the p-value.
- **Requirements**:
    1. Normal data distribution.
    2. Equal variances.
    3. Separate samples.
- **Verification**: Used histograms for normality.

Logistic Regression:

- **Steps**:
    1. Split data into train and test sets.

2. Trained the model.
3. Tested the model.
4. Evaluated using accuracy and F-score.

- **Requirements**:
  1. Binary outcome.
  2. No strong correlations between predictors.
  3. Adequate sample size.
- **Verification**: Checked predictor correlations; ensured sufficient sample size.

F Data Analysis Results

F.1 Statistical Significance

Statistical Test Overview:

- **Null Hypothesis (H0)**: No difference in sustainability scores between high and low GDP countries.
- **Test**: Independent Samples T-test.
- **Metrics**:
  - T-statistic: -8.13
  - P-value: 0.00
  - Alpha (α): 0.05
- **Conclusion**: Since the p-value is less than the alpha (α) value of 0.05, there is sufficient evidence to reject the null hypothesis. There is a statistically significant difference in sustainability scores between high and low GDP nations.

Model Analysis:

- **Model**: Supervised Classification on sustainability scores.
- **Algorithm**: Logistic Regression.
- **Performance**:

    - Accuracy: 29.48%
    - F-score: 0.46
- **Benchmark**:
    - Accuracy: ≥ 80%
    - F-score: ≥ 0.75
- **Conclusion**: The model's accuracy and F-score did not meet the benchmarks of success. The model's accuracy of 29.48% is significantly below the benchmark of 80%, and the F-score of 0.46 is also below the benchmark of 0.75. Based on these results, the model does not support the hypothesis that there is a relationship between GDP and sustainability scores.

Reasoning Behind Choices:

- **T-test**: Used to see if there's a notable difference between two group means.
- **Logistic Regression**: Chosen for binary tasks, like classifying countries by scores.
- **Metrics**:
    - **Accuracy**: Measures the model's success in making correct predictions.
    - **F-score**: Balances precision and recall, especially useful when class classes are uneven.

F.2 Practical Significance

This study identified a significant difference in sustainability scores between high and low GDP countries. Here's what this means practically:

Findings and Their Implications:

1. **Policy Changes**: The data suggests that high GDP countries might prioritize rapid growth over sustainability. Governments and global entities can use this insight to create policies that balance both economic growth and sustainable practices.

2. **Resource Distribution**: Aid organizations can allocate resources more effectively by focusing on the GDP group (either high or low) that requires more sustainable support.
3. **Global Agendas**: Recognizing the GDP-sustainability relationship can encourage inclusive strategies, ensuring all nations, irrespective of their GDP, participate actively.

Application for the Client:

Imagine our client is a global environmental NGO. Using our findings:

1. **Campaigns**: They could initiate awareness campaigns in affluent countries, emphasizing the significance of sustainability alongside economic growth.
2. **Partnerships**: Collaborate with companies in prosperous nations to endorse sustainable practices, showing that growth and sustainability aren't mutually exclusive.
3. **Training**: For lower GDP countries, offer training on sustainable growth methods, ensuring they progress economically without sidelining sustainable practices.

F.3 Overall Success

The goal of this project was to understand the relationship between GDP and sustainability scores, and despite some unexpected results, the project yielded plenty of useful insights.

Statistical Outcomes:

Statistically, the T-test confirmed a significant difference in sustainability scores between high and low GDP countries with a p-value below 0.05. While the logistic regression model didn't meet the set benchmarks, the T-test results remain a foundational success.

Real-world Relevance:

The practical significance emphasizes the project's real-world value. Even if the model wasn't perfect, the insights are critical. These findings can guide governments, NGOs, and international organizations in policy-making, resource allocation, and global cooperation.

Meeting Success Criteria:

- **Data Integrity**: Data quality was maintained throughout the study. All inconsistencies were resolved.
- **Statistical Significance**: The T-test results aligned with our success criteria, indicating a valid relationship between GDP and sustainability scores.
- **Documentation**: Each step of the project, from initial data collection to recommendations, was thoroughly documented.

Conclusion:

Despite not meeting all benchmarks, the project successfully revealed a link between GDP and sustainability. Through solid data management, T-test results, and thorough documentation, the project's value is apparent. These findings can guide future research or improve existing models.

G. Conclusion

G.1 Summary of Conclusions

This study investigated the link between GDP and sustainability scores.

Statistical Insights:

- **T-test**: Found a clear difference in sustainability scores between countries based on GDP, with a significant p-value.
- **Logistic Regression**: Despite the T-test results, the regression model wasn't fully aligned with our benchmarks, suggesting other influencing factors on sustainability scores.

Real-world Impact:

- High GDP countries might value growth over sustainability, highlighting a need for balanced policies.
- Environmental NGOs can reshape strategies using these findings, focusing more on high GDP countries and aiding lower GDP ones.
- Recognizing this GDP-sustainability relationship can inform global strategies to ensure growth and sustainability coexist.

Data Approach:

Strict data standards and transparency were maintained, facilitating future research. Even if some elements fell short, the project gave key insights into GDP's role in sustainability. It emphasized the need to align economic goals with sustainable actions.

G.2 Effective Storytelling

The visualizations used in this project play an important role in effective storytelling by presenting complex data in a straightforward, easily digestible manner. Here's a closer look into the chosen representations:

1. **Global Map with Sustainability Scores**:
    - **Description**: An interactive map shows countries color-coded by sustainability scores.
    - **Purpose**: Gives a quick global perspective on sustainability, prompting questions about regional differences.
    - **Tool**: Tableau, chosen for its mapping capability and interactivity.
2. **Bar Chart: High-Income Countries' Sustainability Scores**:
    - **Description**: Displays sustainability scores of high-income countries using bars.
    - **Purpose**: Focuses on high-income countries, encouraging inquiries about their sustainability efforts.

- **Tool**: Tableau, for its versatility in crafting detailed visuals.

3. **Bar Chart: Low-Income Countries' Sustainability Scores**:
    - **Description**: Highlights sustainability scores of low-income countries.
    - **Purpose**: Spotlights challenges and achievements in sustainability within low-income nations.
    - **Tool**: Tableau, maintaining consistency in visual representation.

G.3 Recommended Courses of Action

1. **Advocate for Balanced Policies in High GDP Countries**:
    - **Rationale**: High GDP countries often emphasize quick growth, sometimes sidelining sustainability. This choice has potential long-term global environmental repercussions.
    - **How it Relates**: The objective of this project was to observe the relationship between GDP and sustainability. With the discovery that economic growth in high GDP countries might overshadow sustainability, it's vital for governments and global bodies to champion policies that harmonize growth with sustainable practices.
2. **Tailored NGO Interventions Based on GDP**:
    - **Rationale**: The findings of this project have invaluable implications for global environmental NGOs. There's a clear opportunity to make a difference by focusing interventions based on a country's GDP status.
    - **How it Relates**: Given the organizational need to guide NGOs, the findings of this study can be instrumental. By understanding the sustainability landscape in relation to GDP, NGOs can craft strategies that have the most impact. For high GDP countries, this could mean increased awareness campaigns, partnerships with businesses promoting sustainable practices, and lobbying

for eco-friendly policies. In contrast, for low GDP countries, this might translate to training, resource allocation, and capacity building for sustainable economic growth.

H Panopto Presentation

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id [redacted]

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