Impact of Internet Penetration on Economic Development: A Data Science Analysis

Cherelle Brigden, Jemma John, Katrina Rogala,
Amy-Louise Snelling, Louise Wright

December 17, 2023

1 Introduction

Aims and Objectives

The primary objective of this project is to investigate the hypothesis:

"Regions with higher internet penetration rates show more rapid economic development."

This study aims to analyse data to validate this hypothesis and understand the correlation between internet accessibility and economic growth. Additionally, we investigate the broader implications of internet accessibility, including its impact on education, political engagement, and environmental sustainability through secondary, complementary hypotheses.
Our primary dataset, sourced from the ITU [1], offers a comprehensive view of internet penetration rates from 2010 to 2022. We have paired this data with insights from reputable sources like the World Bank's World Development Indicators [6] and the World Happiness Report [5]. By merging these datasets, we seek to uncover genuine implications, opportunities, and challenges presented by internet penetration.

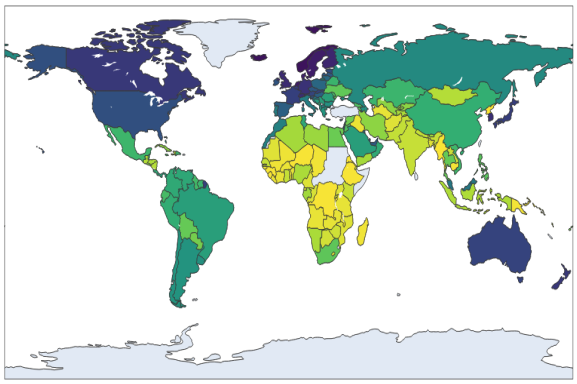

Figure 1: Internet Penetration Rate (%) mapped globally from the 2010 data

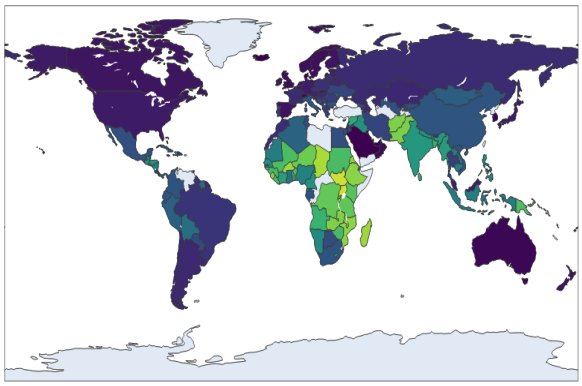

Figure 2: Internet Penetration Rate (%) mapped globally from the 2020 data

Chloropleths of global internet pentration rate between 2010 and 2020, with darker colours depicting a higher percentage.

Roadmap of the Report

Our report begins with a background section, setting the stage for our research and highlighting global changes in internet penetration over the last decade. The Specifications and Design section follows, outlining our methodology, data selection, and pre-processing strategies 3. In Implementation and Execution, we detail our project journey, including the challenges faced and the agile methodologies adopted 4. The Data Collection section describes the datasets used, their sources, and the collection methods, including API integrations 5. Each subsequent section delves into our findings for the various hypotheses, enhanced with relevant visualisations for clarity. Finally, the report concludes with a summary of our findings, discussing their implications and potential areas for future research 8.
Our report aims to provide a comprehensive, clear, and academically rigorous narrative, making our findings accessible and insightful, and underscoring the significant role of internet penetration in shaping contemporary societies.

2 Background

Overview of the Project Report

The study examines the evolution of internet connectivity over the past decade. We analyse trends in internet penetration rates globally, looking at how these rates have changed from 2010 to 2020 and their impact on different regions. This exploration is crucial in understanding the digital divide and its implications on global inequality. Our team draws parallels with historical economic transformations to provide context and depth to our analysis.
We seek to produce a comprehensive analysis of how internet penetration rates affect various aspects of economic and societal progress. We are particularly focused on investigating whether higher levels of internet access correlate with improved economic indicators, such as Gross National Income (GNI) per capita, Gross Domestic Product (GDP), and other measures of economic health. In addition to economic development, our research extends to understanding how internet penetration influences education, mental health, political engagement, and environmental sustainability. We examine if regions with better internet access show improvements in literacy rates or if there's a notable impact on political awareness and participation. We also explore the relationship between internet access and environmental factors, like CO2 emissions, to understand the broader implications of digital growth.

3 Specifications and Design

Project Conception

At the start of our project, we convened on Slack in a Huddle meeting to brainstorm the topic we would explore in this analysis. We individually researched various topics of interest with the understanding that we needed a large, complete, and interesting dataset. This research and the subsequent meeting discussions led to the selection of a primary dataset on internet penetration. Additionally, we decided to compare the primary dataset to additional datasets to explore secondary hypotheses. Team members then gathered the data that would provide a multi-dimensional perspective on our evolving hypotheses.

Data Management Strategy

Central to our technical strategy was the use of a Jupyter notebook, which we stored on Google Colab, as GitHub was not considered ideal for the team collaboration on a Jupyter Notebook filetype. This notebook was optimised to handle null values and inconsistencies, ensuring a robust foundation for our subsequent analysis. In a concerted effort to streamline our data and eliminate redundancies, we merged all the collected data from multiple datasets (including data from APIs) into a singular standardised CSV file. This file became the primary dataset of our collective analysis, allowing each team member to draw from a unified data pool, thereby maintaining consistency across our analysis.

Design and Architecture

The architecture of our project was designed to be iterative and collaborative. Following the data preprocessing, we employed an agile approach to explore our datasets. We used Python scripts for data transformation, aligning different data sources into a coherent format suitable for analysis. This approach facilitated an efficient transition from raw data to actionable insights. Our design philosophy was to create a system that was both efficient in the short term and adaptable for future research. To this

end, we implemented a modular design for our data processing tasks, enabling us to refine our methods as the project evolved. This adaptability proved invaluable as new data became available and as our hypotheses underwent refinement.

Collaboration and Version Control

Colab emerged as a vital tool for collaboration and version control. It allowed us to manage our Jupyter Notebooks effectively, tracking changes and ensuring that all team members had access to the latest versions of our scripts and datasets. Additionally, we used GitHub to manage final files and our project Kanban, where we tracked all tasks and stored meeting minutes files. This setup was critical in creating a collaborative environment where team members could work asynchronously while staying aligned with the project's progression. In summary, our specifications and design phase laid a strong foundation for the project. Through a combination of strategic dataset selection, careful preprocessing, and a focus on collaborative tools and practices, we established a workflow that was robust, efficient, and capable of adapting to the evolving landscape of our research objectives.

4 Implementation and Execution

Development Approach and Team Roles

The execution of our project was methodically segmented into distinct phases, each with dedicated responsibilities, to ensure a coherent and efficient workflow.

Phase	**Description**
Phase 1 Data Sourcing	The entire team embarked on an expedition to identify and procure datasets that would form the basis of our study. This collaborative effort resulted in a comprehensive list of datasets, each chosen for its relevance to the multifaceted aspects of our hypotheses.
Phase 2 Pre-Processing	Katrina managed the initial stages of data transformation, ensuring the cleanliness and compatibility of the datasets. In parallel, all team members contributed to the integration of their respective datasets into the main data cleaning and merging notebook.
Phase 3 Evaluation	With the groundwork laid, the team collectively undertook the task of evaluating and analysing the data. This phase allowed each member to draw from the collective pool of metrics to inform their individual analysis.
Phase 4 visualisation	Amy and Louise provided the creative direction for the project's visual narrative, ensuring that each analysis was accompanied by clear and insightful visualisations.
Phase 5 Reporting	The synthesis of our efforts was orchestrated by Jemma and Amy, who led the report development.

Additional notable responsibilities:
Amy as Project Manager ensured smooth project progression.
Cherelle focused on consolidating our efforts for a compelling final presentation.

Tools and Libraries

- Data Manipulation: Pandas and Openpyxl
- Numerical Computation: NumPy
- visualisation: Matplotlib, Seaborn, and Plotly
- Advanced Analysis: Scikit-learn, SciPy, and Statsmodels
- Data Retrieval: Requests library for API interactions

Achievements and Challenges

One notable challenge was the variation in country name spellings across datasets. We overcame this by integrating datasets that matched country names to World Bank and ISO codes, ensuring accurate merging of data.
Agile development was a strength throughout our project, incorporating iterative development, regular refactoring, and thorough code reviews.
The implementation of our project, while grounded in meticulous planning, remained fluid to accommodate insights and adapt to challenges.

5 Data Collection

We identified the following primary sources for our data: Internet Penetration Rates from the International Telecommunication Union (ITU), World Development Indicators from the World Bank, Voter turnout records from the International Institute for Democracy and Electoral Assistance (International IDEA), Historical CO2 emissions data from the Global Carbon Project.
Additional sources included databases that provided internet pricing across different cities and countries, sourced from Kaggle, and mental health statistics, including the World Happiness Report, which, upon further inspection, was found to be more reflective of economic and social well-being than mental health.

Data Collection Methodology

Around 70% of our datasets were obtained through direct downloads of CSV files from the respective databases. These were then integrated with our main dataset through a process of data cleaning and merging, managed within our Jupyter notebooks.
For the remaining 30%, we utilised the World Bank's API to fetch the latest data. This API provides a wealth of global development data, accessible via HTTP requests. We specified indicator codes to retrieve data that corresponded to our variables of interest.

Approach to Analysis

Our analysis began with exploratory data analysis (EDA) to understand the distributions and patterns within the data. We then employed statistical methods to test the hypotheses, using regression models to assess the strength and nature of relationships between internet penetration and various socio-economic indicators.

6 Data Analysis

Economic Development Findings

A pivotal finding of our study confirmed the hypothesis that regions with higher internet penetration rates tend to have more rapid economic development. We found a strong exponential relationship between internet penetration from 2010 to 2016 and GNI per capita from 2017 to 2022. Interestingly, GNI per capita was more strongly correlated with internet penetration than GDP, suggesting that internet access might have a more direct impact on individual economic prosperity than on broader economic output. When we examined the data by region, Africa and the Asia Pacific showed the most robust relationship. We also observed that countries with the lowest internet penetration rates in 2010 experienced the largest increases in internet penetration, GNI per capita and GDP per capita by 2022. This pattern was consistent with the theory that initial investments in Internet infrastructure could lead to considerable economic gains. The detailed statistical approach and the visualisations supporting these findings can be explored in our notebook titled (GroupProject_Economics.ipynb).

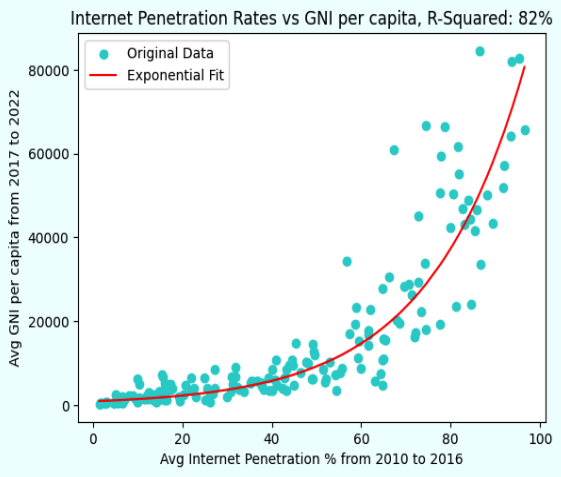

Figure 3: Exponential relationship between internet penetration rates before 2017 and GNI per capita after 2017

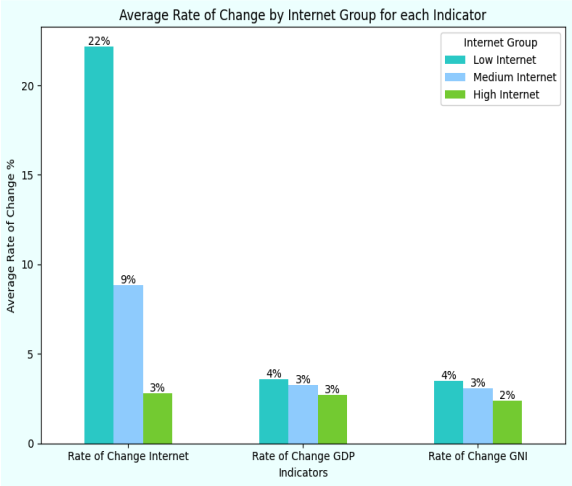

Figure 4: Countries with lowest internet penetration rates in 2010 experienced the highest internet, GDP and GNI rate of growth in the past 10 years

6.1 Political Engagement and Internet Penetration

"Areas with higher internet penetration rates show increased political engagement and political awareness."

In the realm of political engagement, our findings were less definitive. The data did not reveal a significant global correlation between internet penetration and voter turnout. However, when analysed regionally, densely populated regions showed a stronger correlation, suggesting that internet penetration may influence political engagement more in these areas. it is worth noting that in regions with sparser data, such as the CIS countries, no reliable conclusions could be drawn due to the complexity of political systems and external factors such as conflict. For an in-depth exploration of the methodologies and regional analyses, refer to GroupProject_PoliticalEngagement.ipynb.

6.2 Education and Internet Access

"Internet penetration rates are positively correlated with access to quality education and educational resources."

Our analysis indicated a strong positive correlation between internet penetration rates and literacy rates, with a significant p-value and an R-squared value of 0.74. While the statistical association is clear, it is important to note that this analysis does not establish causation. Further exploration and research would be needed to understand the causal factors and implications behind this relationship. The full analysis, including regional breakdowns, is documented in GroupProject_Literacy.ipynb.

6.3 Environmental Impact

"Areas with increased internet penetration are positively correlated with higher CO2 emissions."

The relationship between internet penetration and environmental sustainability, specifically CO2 emissions, was complex. While a general positive correlation existed globally, indicating that higher internet penetration might be associated with higher CO2 emissions, regional variations were significant.
Regional subplots showed that the strength of this correlation varies, with the Asia & Pacific (0.74 correlation) and Arab States (0.71 correlation) regions displaying a stronger correlation, while Europe (0.37 correlation) and CIS (0.21 correlation) regions exhibited a weaker one. This suggests that the environmental impact of internet penetration is mediated by a multitude of factors, such as energy sources and industrialisation levels. The comprehensive analysis, which unpacks the correlation and the contributing factors, is available in GroupProject_CO2Emissions.ipynb.

6.4 Internet Price Dynamics

> *"In countries where internet penetration is low, the internet price is considered more of a luxury"*

Our investigation into internet pricing revealed that lower internet prices are not consistently associated with higher penetration rates. While a negative correlation was initially observed, the relationship did not fit a linear model. This led us to categorise internet affordability and analyse its impact on penetration rates over time. We found that since 2010, price has become a less significant barrier to internet access, suggesting other factors now play a more substantial role in influencing internet penetration. This part of our analysis, including the implications of internet affordability on global connectivity, can be found in GroupProject_InternetPrice.ipynb.

6.5 Mental Health Considerations

> *"Areas with increased internet penetration exhibit a change in the prevalence of certain mental health conditions, such as anxiety or depression."*

Our initial hypothesis that internet penetration would correlate with mental health metrics such as the suicide rate did not hold up to scrutiny. The data did not demonstrate a significant relationship, leading us to conclude that the complexity of mental health outcomes cannot be captured by internet penetration rates alone.
For additional visualisations and in-depth analysis pertaining to these hypotheses, please refer to the visualisations and detailed explanations in the corresponding notebooks.

7 Insights from the Machine Learning Model

In our advanced analysis, we developed a machine learning model, as documented in GroupProject_MLModel.ipynb, to predict GDP per capita using the metrics collected for this project. Noting that these metrics were not collected for the purposes of a predictive model and hence not comprehensive enough to achieve a very high level of prediction accuracy of GDP.

Model Development and Performance

We utilised a Linear Regression model with a Log Transformation on GDP per capita, addressing disparities in GDP across regions. This approach effectively normalised the data, resulting in an R-Squared of 81% for the training set and 79% for the testing set, indicating a strong model fit.

Data Handling and Feature Selection

The model's integrity relied on precise data handling, particularly the imputation of missing values and the systematic elimination of features with non-significant p-values. This process refined our model to include only the most impactful variables.

Key Predictive Insights

The final model highlighted Internet Penetration as having the highest coefficient, underscoring its significant relationship with GDP per capita. The Happiness Index, CO2 Emissions, and Literacy Rates each represent other important economic drivers:

- **Internet Penetration:** Suggests a link to entrepreneurial opportunities and global economic integration.
- **Happiness Index:** Could reflect societal well-being and its impact on productivity and economic output.
- **CO2 Emissions:** Associated with industrial and business activities, often correlating with economic production.
- **Literacy Rates:** Indicative of human capital development, crucial for enhancing workforce skills and productivity.

These insights from our ML model reinforce our hypothesis about the pivotal role of internet connectivity in economic development, offering further insight into its interplay with various socio-economic factors.

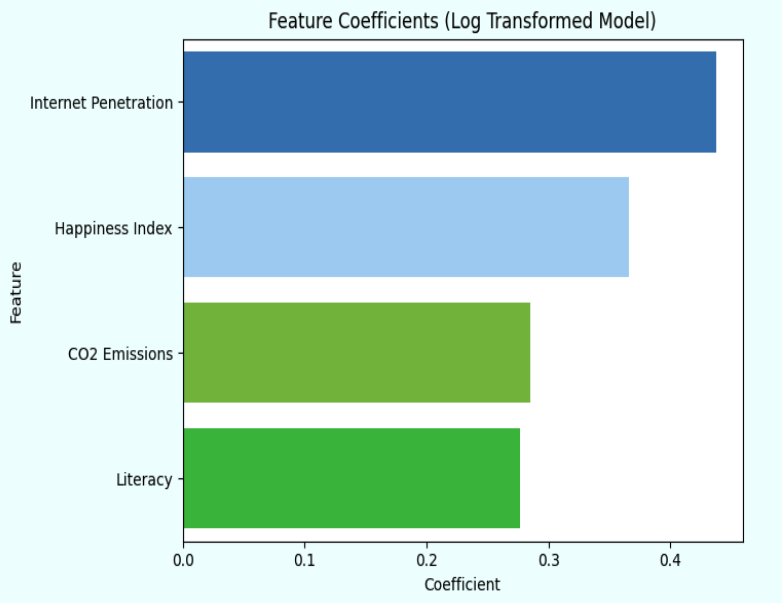

Figure 5: Regression Model Coefficients for Predicting GDP per Capita: Internet Penetration with the highest coefficient, followed by Happiness Index

8 Conclusion

In concluding our study on the impact of internet penetration on economic and social development, we have uncovered several key insights. Our investigation, grounded in data science methodologies, has led us to a deeper understanding of the complexities surrounding digital connectivity and its implications.
The core finding of our research supports the hypothesis that higher internet penetration rates are closely associated with improved economic development, particularly in terms of GNI per capita. This relationship, most pronounced in regions like Africa and Asia Pacific, is clearly detailed in our GroupProject_Economics.ipynb notebook. It underscores the potential of internet access as a significant factor in economic growth.
Our exploration of secondary hypotheses revealed diverse outcomes. The studies on political engagement, education, environmental impact, and internet pricing (each documented in respective notebooks specified in the analysis above), provide a multifaceted view of how internet penetration intersects with various aspects of society.
The addition of an ML model to predict GDP per capita, detailed in GroupProject_MLModel.ipynb, further enriched our analysis. The model's results, especially the notable influence of internet penetration on economic outcomes, align with our primary hypothesis and offer a glimpse into potential future trends.
This project represents a thorough effort by a dedicated team to analyse a complex and timely topic. We believe that the insights gained will be valuable for further exploration and research. As we present this report, we hope it serves as a useful resource for those looking to understand the role of Internet connectivity in shaping economies and societies in the digital age.

References

[1] https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators

[2] https://www.itu.int/en/Pages/default.aspx

[3] https://www.kaggle.com/datasets/cityapiio/world-cities-average-internet-prices-2010-2020

[4] https://www.kaggle.com/datasets/mvieira101/global-cost-of-living?select=cost-of-living.csv

[5] https://www.kaggle.com/datasets/simonaasm/world-happiness-index-by-reports-2013-2023/

[6] https://www.investopedia.com/terms/g/gini-index.asp

[7] https://www.idea.int/data-tools/data/voter-turnout-database

[8] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HTTWYL