***Preliminary Project Proposal***Ray Chandonnet and Barrett Viator  
July 22nd, 2025

**Basics**Team Lead: Barrett Viator (Week 3)  
Spokesperson: Ray Chandonnet (Week 3)  
Recorder: Ray Chandonnet (Week 3)  
Title: *Will Bridging the Digital Divide bring prosperity to the Second “Next Billion Users” and their home countries?*

**Background**

The Rural Electrification Act (“REA”), passed in the United States as part of President Roosevelt’s “New Deal” legislation in 1936, was designed to bring electricity (initially) and other infrastructure (later) to large portions of the country that were considered by industry to be unprofitable places for infrastructure investment. By all accounts, it was a rousing success because “the act allowed the U.S. to grow rapidly in the postwar period and to attain the position of economic dominance it now enjoys…by investing in the development of the country…(it) provided an economic springboard that propelled America ahead of other nations from the 1930s onward.” (Taylor, 2024)

Globally, over the last three decades, the expansion of internet connectivity and digital technologies has presented a parallel opportunity to address long-standing global disparities in livelihood, economic participation, and education. While a significant portion of the world's population has been online for several decades, efforts such as Google's Next Billion Users initiative have been focused on developing applications and ways of connecting with the internet for those who reside in under-developed countries. These users are in the developmental stages of global digital inclusion and often access the internet through devices with less computational power, reliability and connection speed, if they can access it at all (Ranjan, 2022).

“Advanced Meaningful Connectivity” is defined as having at least 4G access, unlimited data, access to a mobile device, and the ability to be online daily (A4AI, 2022). For many in the Global South, the cost of an entry-level mobile device can exceed 30% of monthly income (GDIP, 2022). Historically, a lack of access to information has perpetuated cycles of poverty and limited educational attainment. This phenomenon, termed “adverse digital incorporation,” highlights the need for humanitarian and policy-based interventions (Heeks, 2022).

The widespread availability of affordable internet may help dismantle these barriers, potentially improving both individual wellbeing and broader national economic outcomes—provided supportive infrastructure and policy frameworks are in place. The COVID-19 pandemic underscored this potential, as nations expanded connectivity and more individuals engaged in advanced digital access. Increased digital literacy and access to online learning allowed individuals to gain health knowledge, stay safer, and acquire skills for economic participation and civic engagement (A4AI, 2022).

The "digital divide" extends beyond mere connectivity and encompasses disparities felt by new internet users of all age groups. This gap stems from multifaceted issues including insufficient education, underdeveloped digital literacy, language bias, unreliable digital infrastructure and speed, as well as limited access to devices able to process modern mobile applications. Students, both children and adult learners, without technological access face considerable educational disadvantages which was exacerbated during the pivot to virtual learning amidst the COVID-19 pandemic (Kloza, 2023). Recognizing these challenges, initiatives like Google's "Next Billion Users,” IEEE, and A4AI emphasize the importance of user research in low-income and middle-income countries to develop inclusive tools and products that address specific needs, such as limited storage on older devices, high data costs, and a preference for mobile-first experiences. This demographic represents more than just a new market, and the nurturing of socio-economic development and global engagement for these members is now at the forefront of humanitarian efforts that can be enacted upon more readily than before.

This project intends to utilize a variety of data science methodologies to highlight patterns within technology adoption and to examine the relationship between digital inclusion and the wellbeing of emerging user populations. We will employ both supervised and unsupervised machine learning and predictive modeling methods to attempt to correlate growth in digital inclusion with improved population wellbeing in “clusters” of countries / populations. We further plan to explore whether, beyond pure correlation, we can also actually predict improvements in wellbeing based on growth in digital inclusion. In the absence of proof of that clear predictive relationship, we plan to explore how MUCH change in wellbeing can be predicted by change in digital inclusion.

There are four distinct sets of stakeholders who will benefit from this project:  
1) Disadvantaged populations, whose wellbeing would improve from targeted investments in digital inclusion motivated in part by our study  
2) Public sector leaders of underdeveloped nations, who can use our study to justify allocating resources toward digital inclusion as a drive of economic progress  
3) Private sector leaders, who may choose based on our study to invest more in projects that, and potentially support government leaders who wish to, increase digital inclusion in the developing world, in order to reap the commercial benefits of rising economic activity in these nations as a result.  
4) Global aid organizations, which can use the results of our study to target digital inclusion initiatives more confidently in order to improve population outcomes

**Proposed Research Question**

We are posing the following research question:

*To what extent has growth in internet access of populations influenced population wellbeing in more developed nations, and can that be used to predict improvements in wellbeing and economic progress among global minority populations (the "next billion users")?*

This topic is not novel in and of itself. A number of studies have attempted to quantify this relationship - for example, one study produced a correlation between a country’s % of population accessing the internet and that country’s wellbeing[[1]](#footnote-0) with a 0.77 R2 (Pratama and Al-Shaikh, 2012), and Jyad et al, while another study suggests that HDI can be predicted with a high accuracy (R2 of 0.94) using internet users, broadband access and electricity as predictors (Jyad et al, 2021). However, we seek to build on and improve these studies in three key ways:

1. By adjusting the HDI wellbeing measure for income inequality, we can better capture the impact of internet access on the broad population and the “drag” effect of income inequality to offset benefits that would otherwise accrue to the population from greater digital inclusion.
2. The previous studies do not take into account the time lag between investments in internet infrastructure and its economic benefits, which makes prescriptive policy recommendations difficult and calls into question whether the studies truly measure causality or mere correlation which could be caused by covariant predictors not considered in their studies. Our study introduces this time lag element in order to more justifiably causality if it exists and backstop policy prescriptions.
3. The previous studies are dated and reflect much smaller sample sizes (in time period), do not include the effect of the explosion in internet access tied to the Covid 19 pandemic, and have a far more limited list of countries in their data set than our study, which addresses all three of these shortcomings.

Our topic is worth the time and effort of exploring because if this correlation and predictability does in fact prove strong, it can significantly influence the efforts of governments and international aid organizations, and drive strategy and product design for many global businesses who will be engaging with the newly emerging user population. These policies and efforts span industries from traditional e-commerce efforts to education and to healthcare, to name only a few, and are important for improving global economic outcomes while driving industry profit growth.

**Proposed Hypothesis**

Our hypothesis is as follows:

* Increased access to the internet in the developing world is a measurably material contributing factor to improved population wellbeing - so much so that we may be able to predict how much wellbeing will improve based on increased digital inclusion.
* In addition, mobile broadband likely accelerates this impact due to its relative ease and inexpensiveness versus fixed broadband - portending well for newer technologies such as satellite-based broadband access.

**Prediction**

In summary, we expect performance of any model attempting to predict population wellbeing based solely on digital inclusion to be directionally useful but quite imperfect. Because there are many other factors that impact these absolute outcomes, it’s likely that our models will be best suited to infer a positive relationship, and perhaps to predict the INCREMENTAL impact of digital inclusion - which is still incredibly useful for guiding policy and investment in the developing world. Here’s why:

* We expect the inferential component of our study to show that growth in digital inclusion has strong positive predictive correlation with improvements in education, which is a significant component of overall population wellbeing. We also expect to show reasonably significant positive correlation between digital inclusion and OVERALL wellbeing.
* We expect that PREDICTING the rate of improvement in wellbeing based on internet inclusion alone will be ***extremely*** challenging and inaccurate due to the number of other factors, including things like public / private investment in infrastructure, political stability, conflict, food / water scarcity, prevalence of disease, cultural influences, the distribution of population across rural versus urban environments and other important factors, which are significant contributors to wellbeing.
* Despite these challenges, we expect to be able to isolate and estimate the predictive value of internet inclusion on wellbeing, either by controlling for other important factors as mentioned above, or by using predictive modeling techniques that can isolate and quantify the impact (“importance”) of internet inclusion within a broader predictive modeling exercise that seeks to predict changes in wellbeing based on a broader set of inputs.
* We expect that we may need to cluster countries using unsupervised learning to maximize predictive accuracy due to noisiness associated with different forms of government, social structure and norms, and geopolitical factors as well as other noise factors we cannot predict.

**Defining the “wellbeing” response variable**

Defining, investigating, and statistically measuring population "wellbeing" is a complex undertaking that requires a multidimensional approach. Focusing on a singular measure such as economic output (for example, GDP or even GDP per capita) due to regular access to the internet is not sufficient to make claims that “wellbeing” has improved. Historically, metrics like Gross National Income (GNI) per capita have served as primary indicators of a nation's ability to thrive. (GNI represents the total income earned by a country's residents and businesses, both domestically and from abroad, offering a larger picture of total economic activity. ) While it is useful for gauging the overall economic scale, GNI per capita does not reveal the true *distribution* of income, which can greatly and disparately impact certain groups within these countries (Auerbach, 2021a). It also fails to encompass non-monetary aspects that profoundly impact human lives, such as access to education as well as health and longevity.

In order to alleviate some of the insufficiencies of focusing on GDP and GNI, the United Nations Development Programme (UNDP) introduced the Human Development Index (HDI) in 1990. The HDI was implemented to create a more holistic understanding of development by recognizing that human wellbeing is not solely defined by traditional measures of income. HDI describes “wellbeing” via a composite score that accounts for life expectancy, years of education obtained, and standard of living based on GNI (Auerbach, 2021b). However, even the HDI, which is merely an aggregate average, can be a faulted measure as it does not take into consideration wealth inequality and access to services.

Due to the imperfections in HDI, and our study’s primary goal of benefiting as many people as possible (the “next billion users”) in the under-developed world, we will create our own customized measure of wellbeing which will use the HDI index, but apply a penalization factor for income inequality using the the “Gini index”, which is also tracked by UNDP. This Index ranges from 0 to 1 (with perfect equality being 0) and allows for comparison of wealth inequality across nations without being skewed by overall economic production as seen with GNI (Hayes, 2024). “Handicapping” HDI using the Gini index will give us the ability to indicate to what degree the benefits of development (including internet access) are not impacting all segments of the population equally when it comes to income distribution. This then allows for a more robust and nuanced understanding of progress and challenges in the evaluation of improvement of wellbeing.

**Data & Analysis**

* ***Data***
  + A full list of public data sources we are considering can be found in Appendix A
  + This includes data files containing time series and other data for most countries as well as groupings of countries:
    - “Development” indicators across the globe - things such as economic conditions, wellbeing, access to and use of resources, crime rates, pollution etc, going back 60+ years. (We will focus on data starting in 1995 where internet access became more generally available to the public)
    - Categorization of nations and nation groups as being Low Income, Lower Middle Income, Upper Middle Income, and High Income
    - Internet, fixed broadband and mobile broadband usage by country / year
    - UNDP data tracking the HDI index and its components, as well as the Gini index.
* ***Data Wrangling***
  + Cleanup / Deduplication / Dealing With Nulls
  + Data “Tidying”
    - Some of the data is in “wide” format - it needs to be made tidy to be usable for machine learning, predictive modeling, and visualization for EDA and other purposes
    - API retrieval of the data versus static file use will be implemented to enhance replicability of our analysis and to also allow for the data to be kept current as the original data sources are updated
  + Unification
    - Data will need to be joined into one contiguous data set by linking data files using common “Country codes” and year timestamps.
  + Feature Reduction, Transformation And Creation
    - As described above, a response variable will be created using a custom formula that adjusts the HDI index for income inequality using the Gini index for the same time period
    - Other variables:
      * The development data that serves as the core of the data set contains over 400,000 “wide” rows, which would translate into millions of “tidy” rows.
      * In addition, it contains over 1,500 discrete “development” data points, which fall into 88 “Topics” (categories).
      * This is an extremely large and complex dataset - as such we will reduce the size and dimensionality of the feature space using the following methods:
        + *Feature Reduction:* Features we deem duplicative, covariant, or simply not germane to our thesis will be eliminated. We will also eliminate all data prior to 1995.
        + *Dimension Reduction:*

We have distilled the list of potential variables to approximately 40 which we will include in our analysis:

The full list of variables we will initially consider (which is subject to change) can be found in Appendix B

Some of are granular variables related to wellbeing that we may use to construct our own wellbeing index, if using a reconstituted HDI index proves difficult or ineffective

The rest are variables we expect to be significantly correlated if not predictive versus wellbeing.

We may create “indexes” of related data that reduce the dimensionality much further

We may create variables that transform our key variables (internet access, wellbeing) from static quantitative measures to “change from prior period” measures

We may consider transforming feature and/or response data from raw quantitative data into normalized data scores to create a classification problem

PCA and / or correlation matrices may be utilized to reduce dimensionality further as well as to reduce the amount of time and storage needed to process our models, as well as to reduce noise and to identify relationships between variables.

We will reflect a time lag in our analysis, which will likely be accomplished as part of the feature and/or response transformation process. This is crucial because changes to economic inputs (such as access to broadband, mobile, internet) take time to ripple through and be reflected in economic wellbeing data.

* ***Analytic Plan*** (Subject To Change)
  + *Phase 1:* We will develop both statistical and machine-learning (e.g. tree-based and neural network) models that look to infer and hopefully predict the relationship between internet access and population wellbeing.
    - *Predictors:* Subset of features representing digital inclusion
    - *Responses:* Wellbeing index
    - *Measuring Success:* We will measure success using overall model accuracy, an NxN confusion matrix, precision and recall. We will deem the first question (“can we predict change in population wellbeing based on digital inclusion”) answered if the model either is materially predictive (e.g. “YES”) or no better than at random (e.g. “NO”). The “YES” scenario would prove this most rigid component of our hypothesis
  + *Phase 2*: If the answer to question one is “NO,” we will proceed to modeling the data to answer the question differently, as follows:
    - We will build and run models seeking to predict wellbeing using the broader set of transformed features including digital inclusion PLUS additional potential inputs to the problem
    - We will determine the impact of digital inclusion using one (or both) of the following methods:
      * We will build and run a second model, using the first model but EXCLUDING the digital inclusion features, and comparing the results
      * Using various methods to directly isolate and quantify the importance of our digital inclusion features on the resulting model
    - Success or failure will be measured using the same approaches as Phase 1
  + *Phase 3*: We will build both a qualitative and a quantitative (algorithmic) clustering (unsupervised learning) model that seeks to cluster the countries based on common characteristics in order to control for certain important variables unrelated to internet inclusion and further reduce dimensionality.
  + *Phase 4*: We will look at building predictive models using only data from countries within certain segments as identified in Phase 3 to see if predictions improve, and then seek to apply learnings. For example:
    - We could create a cluster of countries that we know lifted themselves out of low income into higher income categories and assess the impact of digital inclusion, then extrapolate that to the remaining low income countries.
    - We could create clusters of countries with stable governments, lack of conflict, and/or reasonable access to food, water and healthcare to control for those factors
  + *Phase 5:*  We will use the top 2 preferred predictive models / approaches and reapply them, but for “digital inclusion” predictor variables, we will use the more granular “fixed” versus “mobile” broadband penetration data from ITU rather than a singular “% of households using the internet” variable, to see if we can isolate the accelerating effect of mobile broadband. We reserve this analysis for last because the ITU data will require a fair amount of data wrangling to make it usable, including dealing with the lack of country code, which is a key field we require to enable JOINS across data sets.
  + *Testing:* We will break our data into training and test sets in order to measure and limit overfitting of models, and measure model effectiveness on BOTH training and test data
  + *Types of Models:*
    - Statistical models:
      * We will run regression-type models first - using a full battery of model types, since these are the most interpretable, and see if we can prove the thesis this way
      * We anticipate poor results because of:
        + The breadth of the features in play
        + The relative challenges these models have in ignoring features that are not relevant
        + The likely material covariance we will have among features we have selected, which pose challenges for statistical models (we can try to minimize this using correlation matrices)
    - Tree-based models: These will likely be our preferred option for several reasons:
      * They do well at ignoring irrelevant features
      * We can measure the importance of our internet inclusion feature directly using a “feature importance matrix”
      * The results are still somewhat explainable, especially using feature importance matrices
    - Machine-learning models: These may perform best but be most challenging:
      * They do well at ignoring irrelevant features, but are difficult to explain
      * It will be difficult to isolate the impact of our internet inclusion feature
      * They may be most at risk of overfitting
  + *Technical Information*
    - Languages / platforms used: Python (primary), R (secondary / EDA), PostgreSQL (possibly for data prep), Tableau (presentation visuals)
    - *Github Repository:*  <https://github.com/Merrimack-Capstone/MSDS-capstone>
    - Other resources: GenAI as noted below

**Time constraints (budget)**

This is an extremely ambitious project, rendered even more so by the limited time available to complete it. We have gotten a significant headstart on data wrangling which is helpful. However, here is our best estimate of the completion probability of each proposed modeling phase from above:

| **Modeling Phase** | **Probability Of Completing** | **Comments** |
| --- | --- | --- |
| Phase 1 | High | Directly predicting wellbeing based on internet usage penetration is central to our thesis so building these models is essentially “table stakes” to at least test our hypothesis that these are correlated. |
| Phase 2 | High | Assuming that the predictive model in Phase 1 fails (i.e. we cannot directly predict changes in Wellbeing based solely on changes in internet access), then attempting to identify how important internet access is to a broader predictive model is also central to our thesis and must be done |
| Phase 3 | Medium-High | Clustering is a fairly important priority for our thesis - assuming that we struggle to find predictive models that work, it’s possible that clustering to reduce dimensionality and control for other variables will lead to predictive models that DO prove our thesis - so it’s our intention to complete this but modeling complications in Phases 2 and 3 may render this impractical |
| Phase 4 | Low | We think it unlikely that we will get to this - but if we complete Phase 3 then the model frameworks created in Phases 1 and 2 may make it less time-consuming to build new models on clusters of data identified in Phase 3. This work would be necessary to DEFINITIVELY prove or disprove our hypothesis if the results of Phase 1 and 2 are inconclusive. |
| Phase 5 | Very Low | We think it highly unlikely that we will get to this. Fortunately, this phase is not required to prove or disprove our hypothesis - this is really an extension of the hypothesis that focuses on whether investing in MOBILE internet access accelerates improvements in wellbeing. The work done in Phases 1-4 set us up well to continue this, and this enhancement might render the results of this study publishable should we wish to continue the project after the class is over. |

**Proposed Use of LLM / Generative AI to assist with this project**

LLM can be extremely effective at assessing the existing “topics” (categories) and collapsing them into a much smaller set of categories for dimension reduction. Given the complexity of the data set, we may use a LLM for suggestions (NOT CODE) on how to visualize the EDA to see if our hypotheses have a chance of holding water. Away from this, use of LLM will be limited to suggestions on debugging code we wrote, and grammar / spelling / flow of our presentation materials - **no code or presentation / report text or other content included in the project will be LLM-generated.** See Appendix C for specific chat prompt and results that was used to get commentary on our draft proposal, which was then incorporated into the final preliminary proposal

**References**  
A4AI (2022). Advancing Meaningful Connectivity: Towards Active & Participatory Digital Societies.

Alliance for Affordable Internet.  
<https://a4ai.org/research/report/advancing-meaningful-connectivity-towards-active-and-participatory-digital-societies/>

Auerbach, M. (2021a). Gross national product and gross national income: EBSCO. EBSCO Information Services, Inc. | www.ebsco.com. <https://www.ebsco.com/research-starters/economics/gross-national-product-and-gross-national-income>

Auerbach, M. (2021b). United Nations Human Development Index: EBSCO. EBSCO Information Services, Inc. | www.ebsco.com. <https://www.ebsco.com/research-starters/social-sciences-and-humanities/united-nations-human-development-index>

GDIP. (2022, December 22). The cost of smartphones falls, but they remain unaffordable for billions around the world - global digital inclusion partnership. <https://globaldigitalinclusion.org/2022/12/22/the-cost-of-smartphones-falls-but-they-remain-unaffordable-for-billions-around-the-world/>

Hayes, A. (2024, April 14). Gini index explained and Gini coefficients around the world. Investopedia. <https://www.investopedia.com/terms/g/gini-index.asp>

Heeks, R. (2022). Digital inequality beyond the digital divide: conceptualizing adverse digital incorporation in the global South, Information Technology for Development, 28:4, 688-704, DOI: 10.1080/02681102.2022.2068492

Jyad, A et al (2021, February 20) Connectivity for All! Assessment of the Digital Divide and Its Impact on Human Development (Applying machine learning to better understand the impact of the digital divide.)

<https://medium.com/omdena/connectivity-for-all-assessment-of-the-digital-divide-and-its-impact-on-human-development-168bc93ce009>

Kloza, B. (2023, January 23). Digital Divide in developing countries: Why we need to close the gap. Connecting the Unconnected. <https://ctu.ieee.org/blog/2023/01/23/digital-divide-in-developing-countries-why-we-need-to-close-the-gap/>

Pratama, A. and Al-Shaikh, (2012, April). Relation and Growth of Internet

Penetration Rate with Human Development Level from 2000 to 2010; <https://www.researchgate.net/publication/264893520_Relation_and_Growth_of_Internet_Penetration_Rate_with_Human_Development_Level_from_2000_to_2010/download?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoiX2RpcmVjdCJ9fQ>

Ranjan, P. (2022, March 15). An anthology of insights, for a more inclusive internet. Google. <https://blog.google/technology/next-billion-users/anthology-insights-more-inclusive-internet/>

Taylor, K. (2024, September 11) Rural Electrification Act: What It Is and How It Works <https://www.investopedia.com/rural-electrification-act-5119177>

| **File Name** | **Underlying Data Source** | **Contains** |
| --- | --- | --- |
| Example:[[2]](#footnote-1)  HDR25\_Statistical\_Annex\_Tables\_1-7.xlsx | United Nations Development Programme  Data Access Page: <https://hdr.undp.org/data-center/documentation-and-downloads>  Example: 1 <https://hdr.undp.org/sites/default/files/2025_HDR/HDR25_Statistical_Annex_Tables_1-7.xlsx> | Human Development Index (HDI), Life expectancy at birth, Expected years of schooling, Mean years of schooling,Gross national income (GNI) per capita,GNI per capita rank minus HDI rank, HDI rank |
| active-mobile-broadband-subscriptions\_1753224220545.csv | ITU Data Hub  <https://api.datahub.itu.int/v2/data/download/byid/11632/iscollection/false> | Total and % of mobile broadband users by country / year |
| Fixed Broadband Subscription.xlsx | ITU Data Hub  <https://api.datahub.itu.int/v2/data/download/byid/19303/iscollection/false> | Total and % of fixed broadband users by country / year |
| World Development Indicators.xlsx | WORLD BANK GROUP  <https://datacatalogfiles.worldbank.org/ddh-published/0037712/DR0095336/WDI_EXCEL_2025_07_02.zip> | Over 1,500 “development” indicators within 85 “topics” for over 200 countries going back 60 years[[3]](#footnote-2) |

**Appendix A  
*Potential Data Sources***

**Appendix B  
*Variables***

| **Short Name** | **Category / “Topic”** | **Long Name** | **Type** | **Source** |
| --- | --- | --- | --- | --- |
| seriesCode[[4]](#footnote-3) | N/A | N/A | Predictor | ITU - Mobile |
| entityIso | N/A | N/A | Predictor | ITU - Mobile |
| dataValue | N/A | N/A | Predictor | ITU - Mobile |
| dataYear | N/A | N/A | Predictor | ITU - Mobile |
| seriesCode[[5]](#footnote-4) | N/A | N/A | Predictor | ITU - Fixed |
| entityIso | N/A | N/A | Predictor | ITU - Fixed |
| dataValue | N/A | N/A | Predictor | ITU - Fixed |
| dataYear | N/A | N/A | Predictor | ITU - Fixed |
| Country | N/A | N/A | Response | UNDP |
| Year[[6]](#footnote-5) | N/A | N/A | Response | UNDP |
| HDI Index | N/A | N/A | Response | UNDP |
| Gini Coefficient | N/A | N/A | Response | UNDP |
| Country Name | N/A | N/A | Identifier | World Bank (Data) |
| Country Code | N/A | N/A | Identifier | World Bank (Data) |
| Income Level | N/A | N/A | Predictor / Control | World Bank (Country) |
| Year5 | N/A | N/A | Identifier | World Bank (Data) |
| SE.SEC.CUAT.PO.ZS | Education: Outcomes | Educational attainment, at least completed post-secondary, population 25+, total (%) (cumulative) | Response | World Bank (Data) |
| SE.SEC.CUAT.UP.ZS | Education: Outcomes | Educational attainment, at least completed upper secondary, population 25+, total (%) (cumulative) | Response | World Bank (Data) |
| SE.TER.CUAT.BA.ZS | Education: Outcomes | Educational attainment, at least Bachelor's or equivalent, population 25+, total (%) (cumulative) | Response | World Bank (Data) |
| SE.TER.CUAT.DO.ZS | Education: Outcomes | Educational attainment, Doctoral or equivalent, population 25+, total (%) (cumulative) | Response | World Bank (Data) |
| SE.TER.CUAT.MS.ZS | Education: Outcomes | Educational attainment, at least Master's or equivalent, population 25+, total (%) (cumulative) | Response | World Bank (Data) |
| SE.TER.CUAT.ST.ZS | Education: Outcomes | Educational attainment, at least completed short-cycle tertiary, population 25+, total (%) (cumulative) | Response | World Bank (Data) |
| AG.PRD.FOOD.XD | Environment: Agricultural production | Food production index (2014-2016 = 100) | Predictor | World Bank (Data) |
| EN.POP.DNST | Environment: Density & urbanization | Population density (people per sq. km of land area) | Predictor | World Bank (Data) |
| EN.POP.SLUM.UR.ZS | Environment: Density & urbanization | Population living in slums (% of urban population) | Predictor | World Bank (Data) |
| SP.RUR.TOTL.ZG | Environment: Density & urbanization | Rural population growth (annual %) | Predictor | World Bank (Data) |
| EG.ELC.ACCS.ZS | Environment: Energy production & use | Access to electricity (% of population) | Predictor | World Bank (Data) |
| ER.H2O.FWST.ZS | Environment: Freshwater | Level of water stress: freshwater withdrawal as a proportion of available freshwater resources | Predictor | World Bank (Data) |
| FX.OWN.TOTL.ZS | Financial Sector: Access | Account ownership at a financial institution or with a mobile-money-service provider (% of population ages 15+) | Predictor | World Bank (Data) |
| SN.ITK.MSFI.ZS | Health | Prevalence of moderate or severe food insecurity in the population (%) | Predictor | World Bank (Data) |
| SH.XPD.CHEX.PC.CD | Health: Health systems | Current health expenditure per capita (current US$) | Predictor | World Bank (Data) |
| SH.XPD.GHED.PC.CD | Health: Health systems | Domestic general government health expenditure per capita (current US$) | Predictor | World Bank (Data) |
| SH.STA.WASH.P5 | Health: Mortality | Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population) | Predictor | World Bank (Data) |
| SP.DYN.LE00.IN | Health: Mortality | Life expectancy at birth, total (years) | Response | World Bank (Data) |
| SH.UHC.SRVS.CV.XD | Health: Universal Health Coverage | UHC service coverage index | Predictor | World Bank (Data) |
| IT.NET.BBND.P2 | Infrastructure: Communications | Fixed broadband subscriptions (per 100 people) | Predictor | World Bank (Data) |
| IT.NET.USER.ZS | Infrastructure: Communications | Individuals using the Internet (% of population) | Predictor | World Bank (Data) |
| GB.XPD.RSDV.GD.ZS | Infrastructure: Technology | Research and development expenditure (% of GDP) | Predictor | World Bank (Data) |
| IS.SHP.GCNW.XQ | Infrastructure: Transportation | Liner shipping connectivity index (maximum value in 2004 = 100) | Predictor | World Bank (Data) |
| SI.POV.GINI | Poverty: Income distribution | Gini index | Response | World Bank (Data) |
| VC.IHR.PSRC.P5 | Public Sector: Conflict & fragility | Intentional homicides (per 100,000 people) | Predictor | World Bank (Data) |
| CC.EST | Public Sector: Policy & institutions | Control of Corruption: Estimate | Predictor | World Bank (Data) |
| GE.EST | Public Sector: Policy & institutions | Government Effectiveness: Estimate | Predictor | World Bank (Data) |
| PV.EST | Public Sector: Policy & institutions | Political Stability and Absence of Violence/Terrorism: Estimate | Predictor | World Bank (Data) |
| RL.EST | Public Sector: Policy & institutions | Rule of Law: Estimate | Predictor | World Bank (Data) |
| VA.EST | Public Sector: Policy & institutions | Voice and Accountability: Estimate | Predictor | World Bank (Data) |
| SM.POP.TOTL.ZS | Social Protection & Labor: Migration | International migrant stock (% of population) | Predictor | World Bank (Data) |

**Appendix C  
*Use Of Generative AI For This Proposal***

* As described above, use of Generative AI was limited to asking for a review of our completed draft Preliminary Proposal with comments on:
  + Spelling
  + Grammar
  + Narrative flow
  + Problem/Hypothesis/Approach Quality
* ChatGPT was selected for this task since that particular LLM is well-suited for evaluating language in this way, and its weaknesses with regard to subject matter expertise are not an issue here since we are not asking for technical help or for evaluation of plausibility. For those types of questions, we would use CoPilot or Claude
* You will note that in my prompt, I specifically requested that ChatGPT not be excessively “flattering, since a known weakness of ChatGPT is its bias towards empathy and “agreeability’ with the user
* Some of the feedback we received especially regarding spelling, grammar and “flow” was incorporated into this proposal. Some we chose to ignore.
* The prompt we used follows:   
  *My partner and I have written our proposal for our data science capstone project. This is to be Masters Degree level work, and the basic requirement is to find extensive data, come up with a hypothesis and predicted result, and build predictive models to either prove or disprove your hypothesis, with all the appropriate data wrangling, feature and dimensionality reduction etc, EDA including through visualizations, and a final report with visualizations from Tableau. I would like you to review what we wrote and provide me with specific feedback on: - Spelling - Grammar - "Flow" - The overall quality and logic of the problem set, hypothesis and proposed approach. Since I know your settings are tuned to default to being agreeable, I am asking you to not just rubber stamp this but provide real feedback along those four axes.*

1. Wellbeing in these studies is defined as that country’s HDI index - see discussion in the ‘Defining the “wellbeing” response variable section below [↑](#footnote-ref-0)
2. We were unable to create a singular link to all the year-by-year data - we will either create individual CSV file downloads by year or (preferably) import the granular data directly into a Python dataframe within our code [↑](#footnote-ref-1)
3. All World Bank data can be found on the “Data” tab with the exception of Income Group which is found on the Country tab. [↑](#footnote-ref-2)
4. Only keep series code i911mw which is users per 100 people, and rename to “Pop%” [↑](#footnote-ref-3)
5. Only keep series code i992b which is users per 100 people, and rename to “Pop%” [↑](#footnote-ref-4)
6. Year data may need to be created via a “pivot narrow” of the underlying data [↑](#footnote-ref-5)