**Economic Factors vs. Population Distribution and Human Welfare**

Exploring the relationship between features of national economic factors and population, and the effect that has on human welfare.

Andrew Libby  
School of Computing and Data Science (not sure if proper department name)  
Wentworth Institute of Technology  
Boston, Massachusetts, USA  
libbya2@wit.edu

**ABSTRACT**

The purpose of this study is to explore the different relationships that exist between aspects of a country’s economy. Some aspects of a country’s economy examined in this research are number of exports, GDP per capita, types of exports, and the like. Gini coefficient, Youth Gender Parity Index (YGPI), and literacy rate are all examples of features used to outline and define human welfare in countries based on economic factors. The relationships between economic factors and human welfare are then explored using models like k-prototype clustering, linear regression, Gaussian models, and others. Based on the findings of this research, some relationships are easier to find between things like population growth and GDP per capita, and more difficult to find between military spending and YGPI.

**KEYWORDS**

GDP per capita, literacy, inequality, gender parity, exports

1**Introduction**

A country’s economy can provide useful insights into other phenomena that make up a country’s general development. This can include wealth inequality, GDP per capita, and gender disparity, for example. These kinds of insights can reflect on the distribution of wealth and trajectory of a country’s GDP per capita growth, for example, which can be an important metric to use when assessing standard of living**’’**, while a country’s Gini coefficient provides insight into the state of wealth inequality in said country. Research into this subject is key for other reasons, as well. For instance, not only can GDP per capita provide insight about a country’s ability to handle a crisis, or in other words, economic resilience**’’**, while also offering clues on potential future development**’’**. Moreover, this information can reveal the broader issues related to a country’s economy such as social mobility Understanding these economic behaviors is key for policymakers as well, particularly in regards to future development, since these clues can serve as tools for passing legislation in an aim to address wealth inequality, ensure sustainable growth, or easing demographic disparities between men and women, among other things**’’**.

2 **Data**

To satisfy the goals of this research, data was collected from various sources, and then preprocessed in the form of text or CSV files with data formatted in the same style for consistency in data collection and analysis. These datasets were sourced from reliable and reputable organizations that provided the information required for this study to gather insights on different economic and social indicators.

This data covers a range of important factors, such as GDP per capita, number of annual exports in a given year, top exports in a given year, population change over the course of a range of years, literacy rates, Youth Gender Parity Index, military spending, Gini coefficient, and average adult wealth.

2.1**Sources of datasets**

The datasets themselves were not directly downloaded, but rather a plaintext file was made for each dataset, and selected data from charts and datasets published elsewhere was then put into the text file and reformatted partially by hand and partially with Python. Most of the datasets were organized like charts or CSV files, with each data entry in its own cell. Some of the datasets had lots of extra information in other columns that were not needed for this research, so only certain columns were paid attention to and then selected to be put into a text file. Next, the sources of the datasets were vetted for credibility, then read and analyzed for their datasets, with relevant data being selected by hand and pasted into the corresponding text file for further preprocessing and reformatting.

2.2**Characteristics of the datasets**

Once pasted into the text file, the datasets still required more cleaning and formatting, for a few reasons. Some entries missed relevant data, and those entries had to be removed. Moreover, other formatting was applied to the datasets. A semicolon was used as a delimiter to separate each dataset's header into column names, and the subsequent rows below the header were then separated with a semicolon so that each entry in each row corresponded to the correct column header. Once these text files were properly generated, they were almost ready to be implemented for data manipulation and visualization. All numerical values were converted to float values, removing commas which would have caused the numerical values to be treated like strings, which was to be avoided for the most efficient data processing. Even after the datasets were changed to the ideal format, some of them were longer than others, as they had previously contained empty entries whose rows needed to be removed. This meant that not only were the datasets not the same size, but they also did not have all of the same countries. This would make proper data processing and visualization difficult without some additional changes, since to answer a lot of the questions posed in this research, many datasets needed to be merged before calculations could be made. This meant that datasets needed to be compared with one another to first check and even out the length of each dataset, but also to ensure that the names of the countries matched and were sorted properly, so that then they could be seamlessly merged according to the name of a country in the rows of each dataset. These merged datasets were the ones primarily used for the processing and visualization of the data.

3 Methodology

Various different models were used to explore the different issues that this research attempted to address. Linear regression as well as other models including Gaussian process regression, data classification, k-means clustering, and k-prototype clustering were implemented. Linear regression models are useful tools for predictive modeling, as that type of analysis is performed by predicting the value of a variable based on another variable. Linear regression is useful because of its simplicity. It can provide insights into how predictor variables influence a target variable, which makes it an effective tool for forecasting and identifying relationships between data. However, challenges in its usage include the assumption of linearity, which is not always true for real-world data. It is also sensitive to outliers. Linear regression was used in this research primarily to make predictions on a country’s GDP per capita in 2028 by looking at their population change within the past 18 years. A particularly useful Python library for this part of the research was **from sklearn.linear\_model import LinearRegression**. Gaussian process regression is effective when there are several data points that must be grouped based on some non-linear relationship between variables. It assumes there is a correlations structure between observations that can be captured using a kernel function. This model is powerful due to its flexibility in capturing complex, non-linear relationships and its ability to provide uncertainty estimates. However, it can prove to be a liability due to its complexity, especially for large datasets. Gaussian process regression was primarily used to find a relationship between exports and literacy rate by country. Two Python libraries proved useful for implementing this model: **from sklearn.gaussian\_process import GaussianProcessRegressor,** and **from sklearn.gaussian\_process.kernels import RBF, ConstantKernel as C.** K-means clustering is an algorithm that divides data points into k clusters, where each data point belongs to the cluster with the nearest mean. The assumption behind k-means clustering is that the data points in each cluster are more similar to each other than to those in other clusters. K-means clustering is useful due to its simplicity and efficiency with large datasets. The disadvantages of k-means include the need to specify the number of clusters in advance, and it may struggle with the non-spherical clusters or clusters of varying sizes. This model was used to predict the adult literacy rate of a country in 2030 based on its most recently recorded literacy rates and YGPI values. Some Python libraries that proved useful for this investigation were **from sklearn.linear\_model import LinearRegression** and **from sklearn.metrics import mean\_squared\_error.** K-prototype clustering is similar to k-means clustering, only that it can also be used to handle categorical variables along with continuous ones, making it an effective approach to analyzing datasets of mixed data types. This gives it a small edge over k-means clustering in terms of flexibility, although it also experiences the same challenges as k-means clustering. This approach was mainly used to find commonalities between countries grouped according to their YGPI, or Youth Gender Parity Index. More specifically, the relationship between YGPI value, military spending, and top export per country was explored. The YGPI is a metric used to determine how much more disadvantaged girls are over boys, or vice versa. A YGPI of less than one suggests girls are more disadvantaged than boys, but a value greater than one suggests the opposite. For this portion of the research, the libraries **from kmodes.kprototypes import KPrototypes** and **from sklearn.metrics import silhouette\_score**. Data classification is a supervised learning technique where the goal is to assign data points to predefined categories based on certain characteristics. The assumption behind this approach is that there are patterns in the data that allow the assignment of new data points to the correct class based on the relationships between input features. Classification models are useful in that they are versatile in handling both binary and multi-class problems dealing with more than two variables, which makes it a good fit for analyzing relationships between countries based on different economic and social factors. Classification models can prove to be a disadvantage, however, in that they need labeled training data, and are liable for data overfitting. For this research, data classification was used to determine if we could separate countries of different Gini coefficient values into different groups based on wealth inequality in each country. In order to properly visualize this data in Python, the library **from sklearn.metrics import accuracy\_score, confusion\_matrix, classification\_report** was used.

4 Results

This research yielded some insightful information about the way certain economic factors affect human population and social welfare.

4.1 Predicting GDP per capita by examining a country’s population growth

When it comes to population growth and change as an indicator of GDP per-capita, the data visualizations, while showing a positive trend, were a bit under-fitted, as the data points did not fit consistently along the prediction line. This suggests that population growth is not the best predictor of a country’s future GDP per capita.

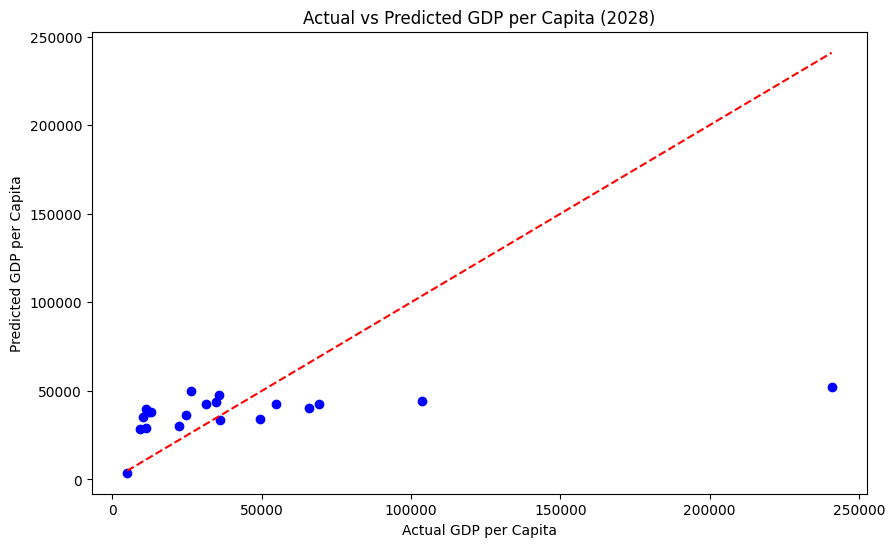


Fig 1 linear regression model: header label name is Actual vs Predicted GDP per Capita (2028); y axis is predicted GDP per capita with a range of 0-250000 and intervals of 50000; x axis is actual GDP per capita with a range of 0-250000 and intervals of 50000; the prediction line goes from the bottom left of the graph to the top right, and the data points are under fitted in the model, with a few outliers towards the far right, but most of the data point are congregated towards the bottom left.

4.2 Predicting Adult Literacy Rate Using YGPI

The visualization for the k-means clustering model was more promising. However, despite the data points being organized in clusters, they were not circular, and were rather clustered along straight lines, which suggests the model over-fitted the data. However, despite being over-fitted, the behavior of the clusters suggests that YGPI is a good indicator of future adult literacy rates in a country, but likely some extra training and fitting of the data needs to be done before that can be properly represented.

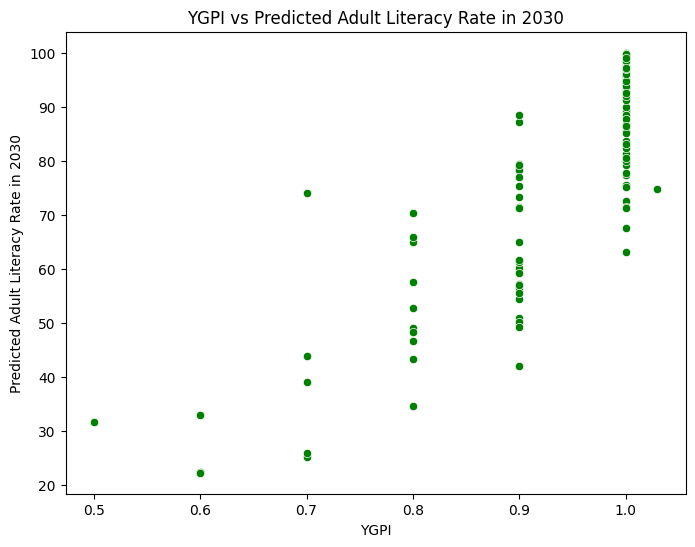


Fig 2 k-means clustering: header label name is YGPI vs Predicted Adult Literacy Rate in 2030; y-axis is predicted adult literacy rate in 2030 with a range of 20-100 and intervals of 10; x-axis is YGPI with a range of 0.5-1.0 and intervals of 0.1; the data is clustered in straight vertical lines, with most of them clustered in lines near 0.8, 0.9 and 1.0, with a few outliers out or below 0.7. In other words, the model overfit the data.

4.3 Gini coefficient data classification

Classifying countries based on Gini coefficient data provided yet more useful insights into how economic factors affect societal wellbeing. The data had to be adjusted however to perform adequate analysis and classification of the dataset, as most of the countries in the dataset had Gini coefficients of 80 or higher, and outliers like Zimbabwe, which had a Gini coefficient of above 200, massively skewed the data. The chart below suggests that countries of higher inequality tend to also have higher average household income.

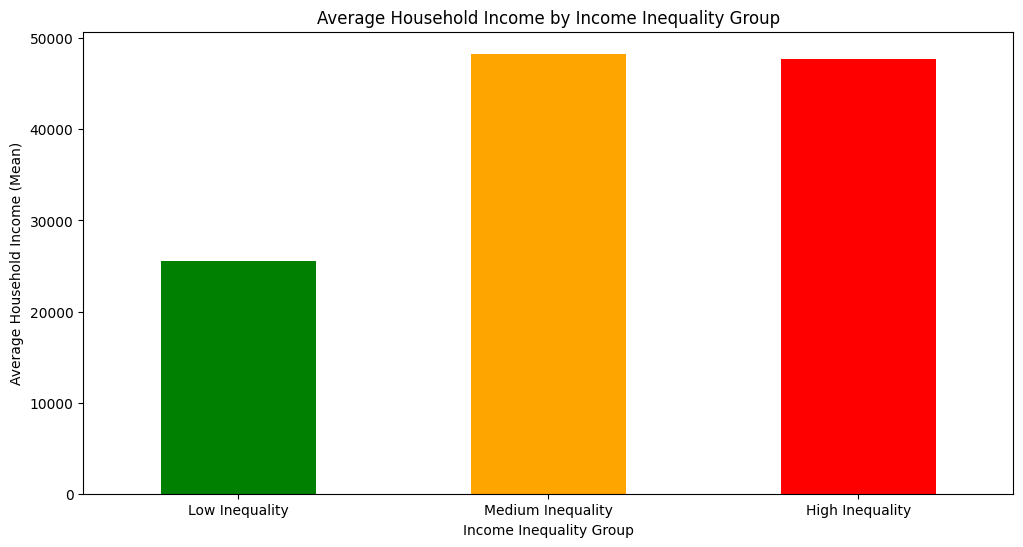


Fig 3 data classification: header label name is Average Household Income by Income Inequality Group; y-axis is average household income (mean) with a range from 0-50000 and intervals of 10000; x-axis is income inequality group and is split into three categories, those being low, medium, and high inequality; the medium and high inequality bars in this model are much higher than the low inequality bar, and the medium inequality group was the largest.

4.4 Exports vs. literacy rates

Gaussian process regression was used to compare the relationship between exports and literacy rates by country, and see if exports served as a good predictor of literacy rates. As it turns out, Gaussian process regression is not the most effective method for determining if exports by country are predictors of literacy rate, as the model markedly under-fitted the data, and the relationship between exports and literacy rate in the training data of the model was likely not adequately established for one to be identified.

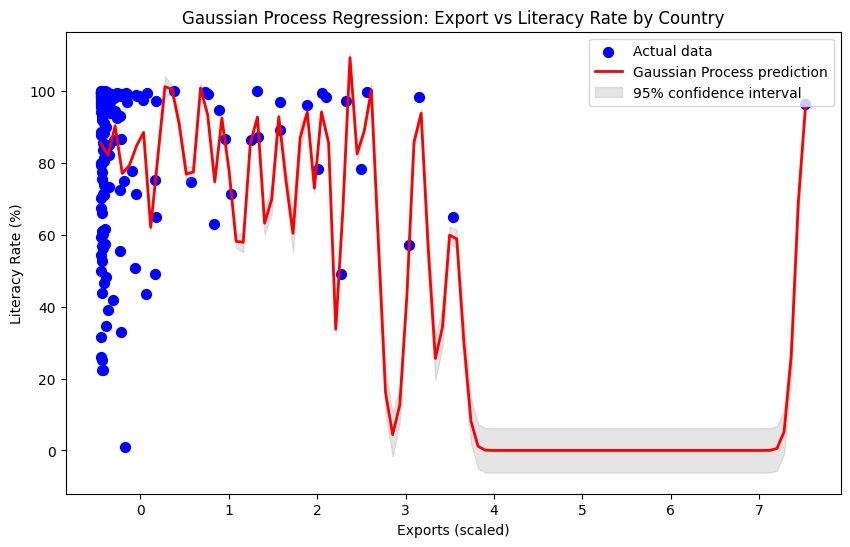


Fig 4 gaussian process regression: header label name is gaussian process regression: export literacy rate by country; y axis is literacy rate (%) with a range of 0-100 and intervals of 20; x axis is exports (scaled) with a range of 0-7 and intervals of 1; the data is mostly clustered around the top left of the graph, although there are some outliers going farther out towards the right and one data point out all the way to the top right; the gaussian process prediction line looks like a messy scribble as it tries to connect all the dots, suggesting GPR underfit the data

4.5 YGPI vs. military expenses

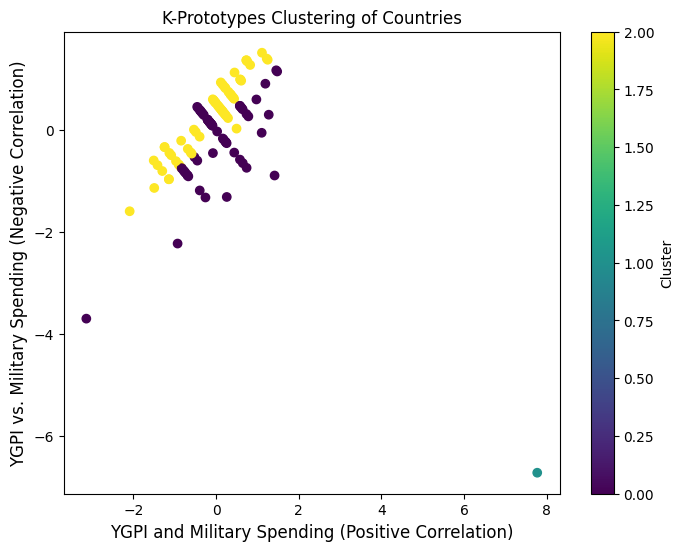
K-prototype clustering is generally used more for datasets of mixed data types, like numerical and categorical data. Due to this added flexibility in comparison to k-means clustering, the significance of a country’s YGPI value in determining what percent of its GDP is spent on its military and what it had as its top export in 2021 was explored. The resulting k-prototype cluster visualization and silhouette score, which is a metric used to determine how well the data points are clustered together, suggested that the data points were not very well or distinctly clustered. A weak silhouette score is below a 0.5, and this model yielded a figure of 0.31 for its silhouette score. This suggests that percent of GDP that goes into military expenses and the top export in a given year does not necessarily indicate the severity of gender disparities between youth in different countries.

Fig 5 k-prototype clustering: header label name is k-prototypes clustering of countries; y axis is YGPI vs. military spending (negative correlation) with a range of -7-1 and intervals of 1; x axis is YGPI and military spending (positive correlation) with a range of -2-8 and intervals of 2; there are three clusters, two of them are touching each other and the third is a single data point outlier, suggesting this data is underfit

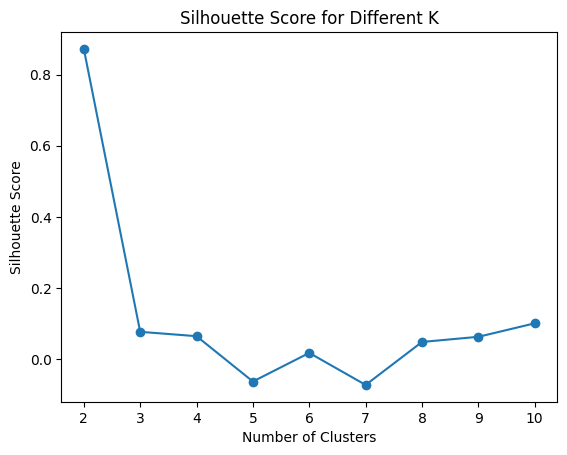


Fig 6 silhouette score: a visualization of a silhouette score of 0.31, indicating poor clustering

5 Discussion

While this research did provide some insights into how economic factors can play a role in the population growth and social welfare of a country, a lot of models were indicating that relationships between some phenomena were relatively weak. This may mostly be due to the fact, however, that the data used to train these models was not robust or comprehensive enough. Models like gaussian process regression and linear regression do a lot better with continuous data, and a lot of the datasets used in this study either likely did not use large enough value ranges for certain entries, like dates for example, which is a likely cause for the regression models in particular to under fit the data. Both clustering models used in this research were not without their problems either. There was likely not enough noise introduced in the data these models were tasked with testing, which resulted in data over fitting. Had there been larger variation in the YGPI and Gini coefficient data, for example, the clusters in both models would have been more distinct, with more pronounced shapes as there’s slightly higher variation among data points in one cluster, but not enough similarity among other data points outside their cluster for those clusters to overlap. For future work, the models would benefit from testing on data that included more continuous numerical values, like greater ranges between dates, or to test a field of data that has more variation to avoid overfitting, or introduce data to the models that has a more pronounced correlation with another set of data, so that underfitting is not as big of an issue as it was in this study.

6 Conclusion

PROMPT: In this part, you should summarize your project. What important results did you find for your topic and what’s the effect of this result on the real-world?

ANSWER:

This study aimed to explore the relationships between various economic indicators and human welfare in countries,using different models such as linear regression, gaussian process regression, and clustering methods. The key findings of this research revealed both useful insights as well as limitations in understanding these correlations.

One of the more notable results was the weak predictive power of population growth as an indicator of GDP per capita. The linear regression model did show a positive trend, although the underfitting of the data suggests that population change alone is not enough as a predictor of GDP per capita growth. It could also be the case that the population growth data was likely not robust enough and did not include long enough time ranges of population growth to properly assess how much of an effect it has on GDP per capita. If population change over time is indeed a weak predictor of economic growth as the model suggested, policy makers should likely be wary of relying on population growth as sufficient metrics for economic resilience.

On the other hand, however, the k-means clustering model provided more promising results for predicting adult literacy rates based on YGPI. The overfitting of the clusters, however, suggests that YGPI could still be a valuable factor, although more refinement in the model would be necessary to produce more accurate forecasts. Addressing gender disparities may in this case based on these findings have a non-trivial impact on future literacy rates, which may highlight the importance of gender equality in education, among other things.

**ACKNOWLEDGMENTS**

Insert paragraph text here. Insert paragraph text here.

**REFERENCES**

Use the following ACM Reference format for your citation

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK’18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

[1] Patricia S. Abril and Robert Plant, 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan, 2007), 36-44. DOI: <https://doi.org/>10.1145/1188913.1188915.

[2] Sten Andler. 1979. Predicate path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL '79)*. ACM Press, New York, NY, 226-236. DOI:https://doi.org/10.1145/567752.567774

[3] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.

[4] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY..