A Visual Exploration of Gross Domestic Product

Fasih Munir

Abstract— This paper focuses on understanding how Gross Domestic Product (GDP) has changed over time, how countries are grouped according to their GDP and if it is possible to reliably predict the GDP of a country. Our analysis focuses on data collected by the World Bank. We begin with a timeseries analysis of our data to see how GDP has changed between the years 2000 and 2022. Then, we build clusters through K Means to see how countries naturally end up being grouped. Finally, we will build regression models, specifically Linear Regression, to see if we can reliably predict GDP and what factors may have the greatest impact. Our analysis will continually be supported with visuals in the form of graphs and maps to guide our intuition and build our understanding.

1 Problem Statement

Gross Domestic Product (GDP) is a measure that was developed in the 1930s to understand the economic activity of a country[1]. The higher your GDP, the better your economic standing. However, over the years it has begun to incorrectly be used as an indicator of economic well being [2], and while that debate is yet to reach a conclusion, its role in understanding a country's output still holds significance.

Our research will aim to understand GDP and answer the following questions:

1. How did GDP evolve between the years 2000 and 2022 amongst different continents and countries?
2. What kinds of clusters were formed and how were countries grouped given the inputs of GDP? Which input of GDP determined the cluster formations?
3. Is it possible to reliably predict GDP given its inputs and other macro-economic variables? Is Linear Regression a viable method?

Working through the questions will help us compare what GDP was to what it is now. The answers can be informative for researchers looking to model predictions of GDP and useful for policy makers to understand what drives economic output and where to invest their resources.

The data has been sourced from Kaggle [3] and collected by the World Bank. It contains common GDP inputs such as expenses and gains from imports, exports and industry (amongst others). It also contains macro-economic variables such as inflation and unemployment rates. The data is shown country wise and collected yearly making it effective in understanding trends over time.

2 State of the Art

The papers that I have looked at focus on the GDP of America and China and take unique approaches to visualizing it. Estrada (2010) [4] proposed a multidimensional graphical model that aimed to visualize all the variables of GDP and run it in real time to visualize unexpected shocks that may occur. The variables that are used contain a breakdown of the GDP of the US economy between 1993 and 2008. Its main components include personal consumption, domestic investment, net trade and government expenses. These are further broken down into smaller categories. The majority of variables that our paper will be using can be found in Estrada (2010). A Mega Space Coordinate System is set up in which the z axis shows growth rates, the x axis shows the variables and the y axis shows the years. The plot extends positively and negatively showing good growth and poor growth. While this approach is not directly applicable to our paper as we are not dealing with growth rates and not just looking at GDP inputs, the idea of visual colour coding temporally will help our work as we will plot GDP temporally on a map for each country with a colour scale to investigate where and when high economic output was achieved.

Chong et al (2021) [5] takes a different approach to visualize the inputs of GDP. They focus on the economic system of China in 2018. Rather than taking a temporal view they decide to dive deep into the characteristics of an economy by creating a Sankey Diagram which is modelled on an input-output table. Sankey Diagrams [6] are usually used to depict quantity flows within a system and help visually identify the most significant contributions with wider paths from point A to point B. Chong et al (2021) uses such a diagram to visualize monetary policy expenditure which in turn visualizes the biggest drivers for GDP and which commodity brings this drive. While they have employed a unique approach, our paper will employ an alternative way to understand the drivers of GDP which involve a mixture of Principal Component Analysis (PCA) followed by clustering.

The final paper we look at is from Raghupathi and Raghupathi (2020) [7]. Their paper takes a visual approach to analysing how healthcare expenditure is related to economic performance. While their paper is specifically focused on one aspect of GDP, much inspiration can be drawn from their visuals. Their research involved looking at distributions of healthcare expenditure and how they rank by states in the US. We will employ a similar map technique to visualise GDP across countries. One area where their paper falls short is that their analysis is not causal which is to say that they do not observe whether healthcare expenditure plays a significant role in economic performance only that there is a positive association between them. Our paper will take this one step further by building regression models to understand causality.

3 Properties of the Data

The data used in this paper has been collected by the World Bank and made available for the public. We observe historic trends for 215 countries between 2000 and 2022 for the following variables: Education Expenditure, Service Contribution, Import, Industry Contribution, R&D, Export, Health Expenditure, Agriculture Contribution, Population, Ease of Doing Business, Unemployment Rate, Inflation Rate and GDP. The contribution variables were originally shown as a percent of GDP and were transformed to reflect the actual value for that observation. Not every country had data for every variable. Countries which had any amount of data missing were dropped entirely leaving us with 131 countries. This was not the first choice to cleanse the data as we are losing valuable information. We would likely impute for an individual country either through the mean or an assumed growth rate, however a pattern was observed where when data was missing it was missing for all 23 years. So rather than dropping the variable, the observations were dropped.

GDP and its corresponding inputs are floats and are values rather than ratios. Ease of Doing Business is a number between 0 and 100 while the two rates are percentages shown as floats. GDP will be our dependent variable for our regression models. It should be noted that apart from the rates, population and business ease, the other variables are what make up GDP. Given this, we can expect multicollinearity when doing our analysis which we will address when we come to the relevant sections.

We used violon plots to investigate the distribution and variance of our data. Violon plots are useful in that they not only give us an idea about the underlying distribution of a variable but also show the quartile ranges and extent of outliers present.

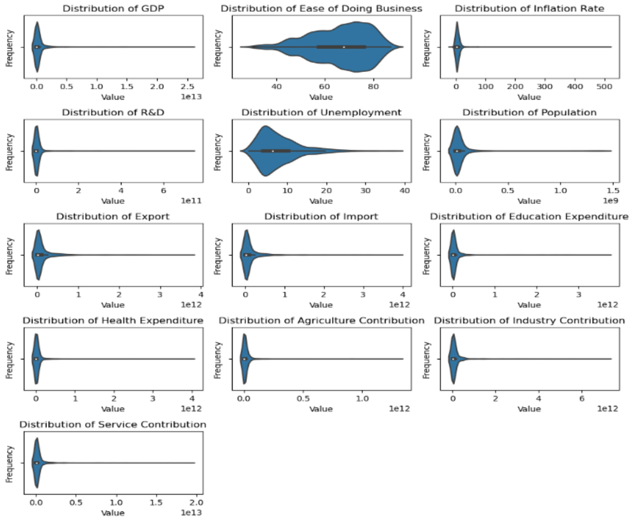

Figure 1

As seen in Figure 1 above, our data (all years and countries) is extremely skewed and no variables are following a normal distribution. There is a large presence of outliers. Given this, we will employ robust scaling to ensure that our clustering is not affected. We will not be dropping any values that deviate as this information is meaningful in the context of GDP given that some countries can have higher or lower expenditures on variables for example an agrarian economy will spend more on agriculture than a service based one.

Since we will be employing regression techniques, specifically linear regression, it is important to understand if the relationship between the dependent and independent variables is linear to satisfy the linearity assumption.

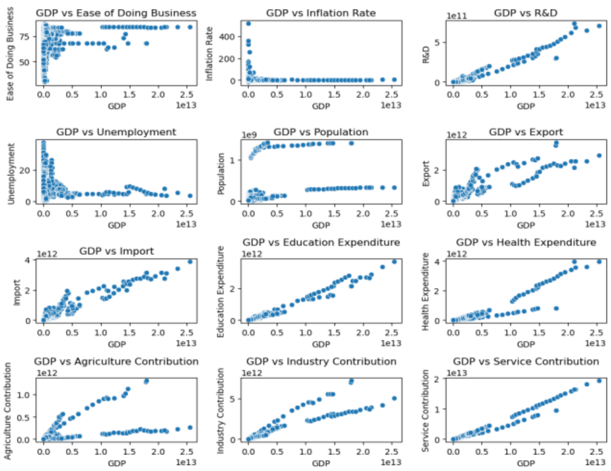

Figure 2

Figure 2 above shows scatterplots for GDP against the other variables. As expected, variables that make up GDP show strong positive linearity while the others do not. Since most variables are exhibiting a linear relationship we will continue with fitting a regression model and asses the outputs. It is interesting to note that while the GDP inputs are showing a positive relationship, they tend to branch out which likely means that there will be strong clustering or specific trends.

4 Analysis

4.1 Approach

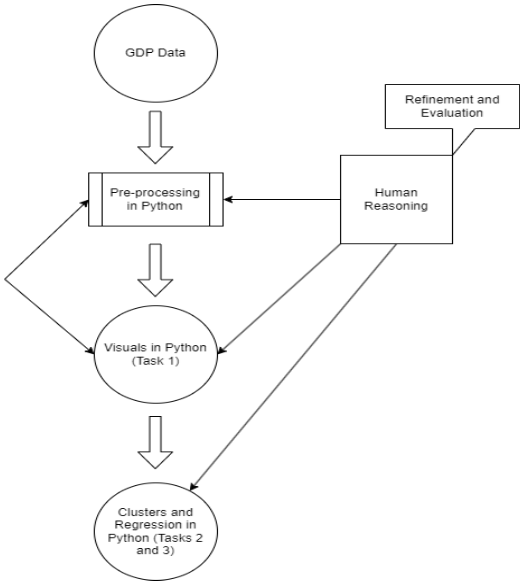
GDP Data
Refinement and Evaluation
Pre-processing in Python
Human Reasoning
Visuals in Python (Task 1)
Clusters and Regression in Python (Tasks 2 and 3)

Figure 3

Figure 3 above shows an overview of the analytical process.

The process starts with exploratory analysis and pre-processing to prepare the data. These steps build an understanding of the data that we are working with which will further guide human reasoning. This is an iterative stage, as exploration will define transformation which will define further exploration.

Our first task is to conduct a time series analysis to understand how GDP has changed. Our data has a column for Year which contains data from 2000 to 2022 allowing us to create visuals that can help the human eye discern temporal patterns. We will create 2 kinds of plots. The first will be a line graph across years with countries grouped into continents. This will give us a high level idea of how the world has evolved. We will follow this with an animated choropleth map to observe how GDP for countries has increased or decreased. We will also create a correlation heatmap to observe which variables might explain the movement of GDP. This heatmap while not built for every year will guide our thinking for the next 2 tasks.

Our second task involves constructing a K Means cluster. We will group countries based on the structures that the algorithm learns given the GDP input variables only (healthcare expenditure for example). However, clusters that involve 4 or more variables are not possible to visualize as we currently do not have a way to visualize 4 or more dimensions. This poses a challenge to involve human judgement. To solve for this we will use Principal Component Analysis (PCA) that will reduce the number of features. While we will lose some explainability as PCA reduces variables by finding linear combinations through eigenvectors, we will gain the power to visually analyze if our clusters are sensible. The optimal PCA components will be identified through an explained variance plot, the optimal number of clusters will be identified through an elbow plot and finally we will plot our clusters on a scatter plot. To observe how clusters have changed over time within countries, we will create another animated choropleth map.

Our third task will have us building linear regression models to understand if linear regression can be reliably used to predict GDP. Given the presence of multicollinearity (shown in the next section) and outliers, we will attempt to make 3 linear models. A model using all robust scaled variables (base), a model using robust scaled variables with PCA (attempt to reduce multicollinearity) and a lasso model with all robust scaled variables (feature selection). The robust scaling is employed given the outliers seen in the violon plots. While we will compute the R2 score to assess our models, we will additionally rely on residual and QQ plots to ensure that our models satisfy the assumptions of heteroskedasticity and residual normality. Specifically in the case of the lasso model, we will use the QQ plot to adjust our value of lambda (penalty term) through trial and error.

4.2 Process

We will now put our process into action, and look at answering our research questions mentioned before.
Firstly, let us take a look at the evolution of GDP over time. Our data contains yearly entries which makes it difficult to go down to further granular detail however given the scope of this research, we still have 23 years of data for each of the remaining 131 countries after we completed the data pre-processing. This is enough to understand the evolution up until recent times. We used Python along with the matplotlib and seaborn libraries to create the plots seen in this research.

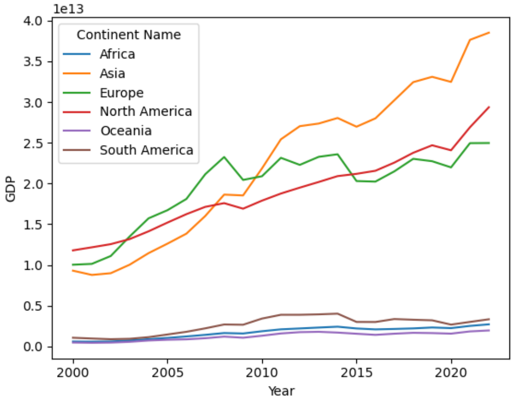

Figure 4

Figure 4 above, shows us at a continent level, how the GDP of the world has moved. With 6 continents in total, there seems to be a clear separation between the group of Asia, Europe and North America, and the group of Oceania, South America and Africa. The first group has steadily increased their GDP over time which is reflective of the current status of the world. This group leads in terms of innovation and political power while the latter group has remained stagnant. After 2010, Asian countries seem to dominate in terms of growth while North America seems to have a more steady and stable increase. The first group has always maintained a higher economic output that highlights the disparity amongst nations.

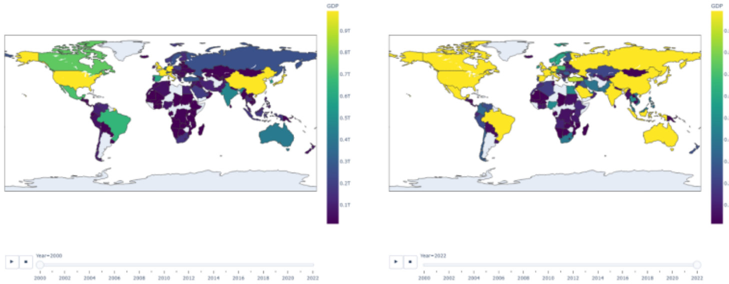

Figure 5

Figure 5 above, breaks the continent level analysis down to the country level analysis. GDP values are shown in USD trillions. While the animation will not work here (we will not be able to see how the clusters change over time) we have shown the start and end of the animation (the years 2000 and 2022). It is clear from the plot that many of the countries have been able to move towards a trillion USD or more in GDP (shown in yellow). What was masked in figure 4 before was that Russia is a country that was much closer to the likes of the African continent, but over the past 2 decades has been able to turn itself around. Additionally, while previously the South American continent is shown to have one of the lowest GDP, Brazil is a member country that is an outlier case comparable to its North American counterparts (although it still has a lower GDP in 2022 compared to the North American countries).

These visual aids have given us our stopping criteria as our time series analysis is reflected in the map, where in fact the map has given us further insight into how the world has evolved.

Before we move to answering our next research question, we will take a quick look at a correlation heatmap to understand what may be driving some of this growth.

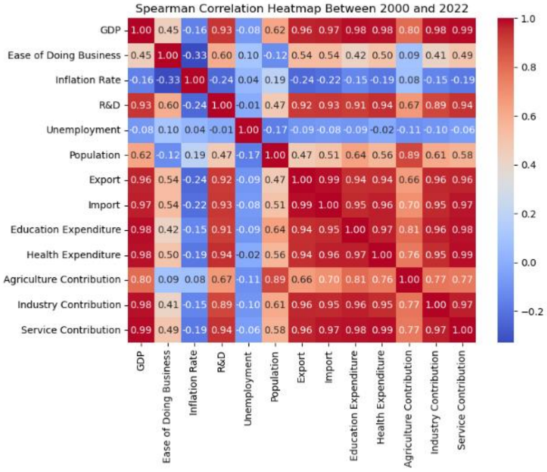

Figure 6

Figure 6 above, shows us a Spearman correlation plot of the numerical variables that are found in the data. From the plot, high multicollinearity is evident for the inputs of GDP. The macroeconomic variables are either showing a medium positive correlation or a weak negative correlation. It is likely that our following analysis will be affected mostly by the GDP inputs. While this will not affect our predictions [8], it will affect how we interpret the effect of each individual variable. For these reasons we will construct more than one linear regression model. The heatmap itself is not indicative of any one variable that may be the most important in driving GDP growth. We will aim to understand further as we answer the next research questions.

Secondly, let us now try and understand what kind of clusters form for countries given only the GDP input variables and which input may drive this formation. We prepare a K Means clustering algorithm using the variables Education Expenditure, Service Contribution, Import, Industry Contribution, R&D, Export, Health Expenditure and Agriculture Contribution. Given that such algorithms rely on distances between points to group observations, our data is scaled using the robust scale method to reduce the impact of outliers and create consistency between the data points. As mentioned before, there are more than 3 variables that we are using to cluster on. This will make visual analysis difficult. To counteract this, we employ a PCA which will enable us to visually analyse the data. While we lose some interpretability when creating the principal components, we are still able to access which variable in the first component explained most of the variation which can help us answer our question of which variable is most effective in pushing GDP. We will calculate the explained variance and use a variance plot and use human judgement to identify the best number of components given our general understanding of how many to use [9].

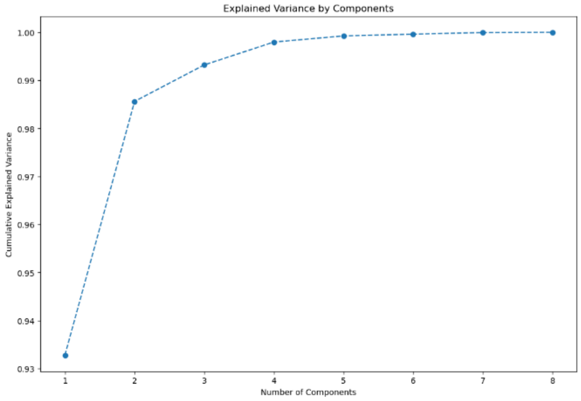

Figure 7

Figure 7 above shows our variance plot. Nearly 99% of the information in the data is captured by the first 2 components. This will be our number of components going forward.

We will then use an elbow plot to identify a mathematically optimal number of clusters.

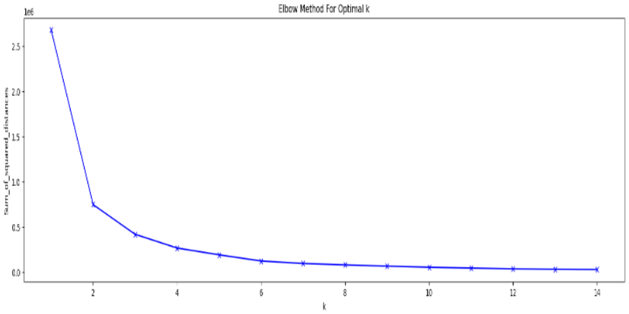

Figure 8

Figure 8 above is the elbow plot, typically used to identify the value of K in K Means algorithms. Using human judgement we can see that the curve kinks at the mark of 3, after which the curve begins to smooth out. This indicates that our data will likely cluster into 3 groups.

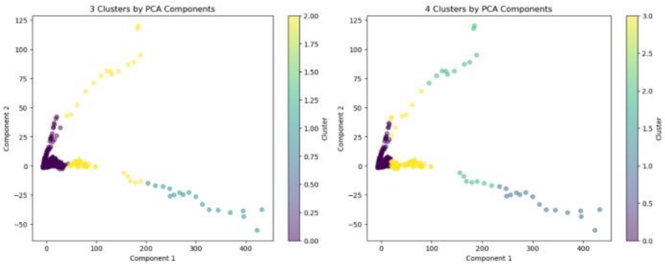

Figure 9

Figure 9 above shows the results of our clustering. Here human intuition was greatly involved. Using the elbow plot, we identified 3 to be the number of clusters we will likely have. However, looking at the outcome of our plot there looks to be a branching out of values which signifies that there are likely more than 3 clusters present (see the yellow clusters to the left of the figure). There is still some bleed between the deep purple and yellow clusters, and it is likely that the presence of outliers is causing some of the larger clusters to stretch (they may not be clusters but outliers that are distorting the center). However, the way the points are branching seem to suggest that there is at least one more cluster that may form. When we increase the clusters to 4, the yellow cluster seems to gain strength and when looking at the raw counts in each cluster, we seem to have better distribution (although the deep purple cluster still dominates the count). We will conclude that 4 clusters shows a better separation of the data although other techniques may need to be employed to reduce the impact of the outliers.

In our results section, we will discuss how the 4 clusters are also useful when visualized on the country level.

Thirdly, and finally, we will build our regression models to asses if we can create reliable predictions and if linear regression is a viable method. Given that we have outliers present and there is multicollinearity, we will robust scale all the data and create 3 models. We will create a base line linear regression using all the variables, then we will use our PCA components to create a second regression (to mitigate multicollinearity indirectly) and lastly create a Lasso Regression model to remove features that prove redundant (directly mitigate multicollinearity). For all these models GDP will be our dependent variable.

To assess our models we used a plot for predicted versus actual points and the R2 score. We further used residual plots and QQ plots to understand heteroscedasticity and normality as without validating the other assumptions, we can not confidently recommend a linear regression [10].

When calculating the R2 score, all 3 models gave a result of 0.99. On the surface this may sound like the models have captured 99% of the variability in the dependent variable, however, knowing that multicollinearity and outliers exist, it is likely that there is over fitting or some linearity assumptions not being met which can invalidate the result. Observing the residual plots show that there is little to no heteroskedasticity however, the QQ plots show violations of the normality assumption. While some violations are tolerable, it will not be good practice to recommend a model where knowingly the assumptions have not been validated. For this reason we use the QQ plots for our Lasso Regression to adjust our value of Lambda (through trial and error, increasing by a factor of 10) to bring the residuals close to normal giving us more trust in our result.

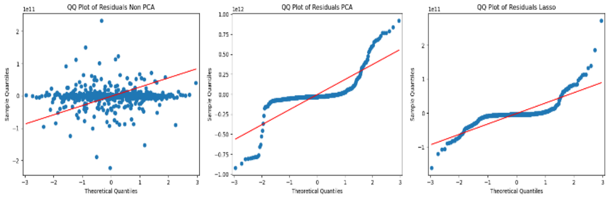

Figure 10

Figure 10 above shows the final QQ plots produced. From the plots, the Non PCA and PCA regressions can be disregarded given their violations of the normality assumption. Ideally we would want the residuals to be similar to the imposed regression line. The Lasso QQ plot shows the most normality which was achieved by adjusting the value of lambda through trial and error and observing the QQ plot becoming more normal. This removed some of the variables and reduced the multicollinearity. The Lasso model seems to be the most successful although given the data there is still likely overfitting that would need to be investigated.

4.3 Results

We observed significant changes over time when doing our times series analysis. Asia, North America and Europe showed the most increase in GDP overtime and were consistently higher in value compared to the other continents which remained stable. We did however see certain countries like Russia and Brazil prove to be outliers and grow.

Accurate clustering was seen using 4 clusters. Looking at the loading values of component 1, RnD (Research and Development) was the most important variable. This intuitively makes sense given that countries have grown the most from breakthroughs such as the industrial revolution. Looking at Figure 11 below provides further evidence for this. In 2022, the yellow and purple clusters represent locations that have spent the most on RnD on average (India, North America and Europe). Compared to the year 2000, they have also grown the most.

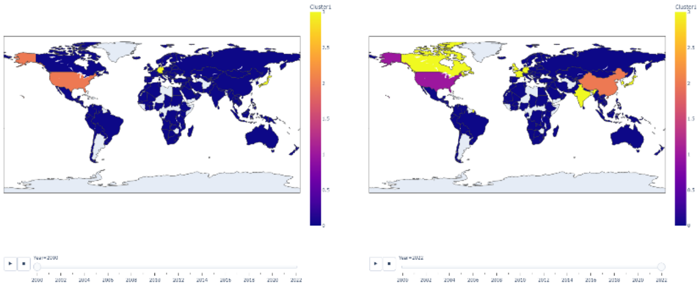

Figure 11

Our linear regression provided impressive results (R2 = 0.99) although, given some violated assumptions and potential overfitting it is difficult to recommend. Lasso Regression eliminated the variables Import, Health Expenditure, Ease of Doing Business, Unemployment and Inflation Rate. It kept

RnD providing additional evidence of its importance. With extra effort in understanding potential overfitting, Lasso Regression can be recommended.

5 Critical reflection

A time series analysis is useful to identify changes over time, especially in our case of GDP. We had countries grouped as continents initially to get an overall view and then proceeded to observe changes at a country level. The country level breakdown was important as if we had just stuck to a continent level (or if we engineered specific regions), we would have disregarded certain trends in countries such as Brazil and Russia. Reflecting on the animated choropleth map, it could have been wiser to breakdown the values of GDP further to further subdivide countries that achieved a GDP of greater than 1 trillion USD.

The combination of PCA and K Means clustering proved useful in bringing back the human element into analysis. It may have reduced some interpretability, however given that important variables from principal components can be still deduced mathematically, this was of little concern. We needed to critically examine the formation of the clusters with the human eye to see that in fact there are likely 4 clusters forming given how the data is branching despite the presence of outliers. We did observe some bleed and clusters being stretched from their center which shows the limitations of the basic K Means model. Alternate clustering algorithms that are more robust to outliers could be employed for better results such as DBSCAN.

Our human intuition was greatly required when building the linear regressions models. Our dependent and independent variables were linearly related although there was multicollinearity and outliers present. A case could also be made that many of the variables that are GDP inputs such as health expenditure are not truly independent however, variation in these will always exist across countries for example an agrarian economy spending more on agriculture versus a service based economy. Although, given the results of our Lasso regression, it is likely that all inputs are not required. If certain macro-economic variables such as population, and some GDP input variables such as RnD are used, these would eliminate our hesitancy around independence and still provide accurate results. The visual trial and error with the QQ plots to adjust our value of Lambda for the Lasso regression was important as it allowed us to improve the reliability of the model despite potential overfitting. It should also be noted that we used a simple regression that was not cross validated. Employing cross validation along with the lasso regression could further combat overfitting and make us confident in recommending linear regression as a viable method to predict GDP.

Table of word counts

Problem statement	248
State of the art	490
Properties of the data	494
Analysis: Approach	498
Analysis: Process	1495
Analysis: Results	199
Critical reflection	421

References

[1] R. Costanza, M. Hart, I. Kubiszewski, and J. Talberth, “A Short History of GDP: Moving Towards Better Measures of Human Well-being,” Solutions-for a sustainable and desirable future, vol. 5, no. 1, pp. 91–97, 2014, Available: https://researchprofiles.anu.edu.au/en/publications/a-short-history-of-gdp-moving-towards-better-measures-of-human-we

[2] S. Cohen Kaminitz, “The significance of GDP: a new take on a century-old question,” Journal of Economic Methodology, pp. 1–14, Jan. 2023, doi: https://doi.org/10.1080/1350178x.2023.2167228.

[3] “World Bank Data on Countries,” www.kaggle.com. https://www.kaggle.com/datasets/yusufglcan/country-data

[4] M. A. Ruiz Estrada, “The Visualization of the GDP from a Multi-Dimensional View,” SSRN Electronic Journal, 2010, doi: https://doi.org/10.2139/ssrn.1611622

[5] C. Chong et al., “A Visualization Method of the Economic Input–Output Table: Mapping Monetary Flows in the Form of Sankey Diagrams,” Sustainability, vol. 13, no. 21, pp. 12239–12239, Nov. 2021, doi: https://doi.org/10.3390/su132112239

[6] R. C. Lupton and J. M. Allwood, “Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use,” Resources, Conservation and Recycling, vol. 124, pp. 141–151, Sep. 2017, doi: https://doi.org/10.1016/j.resconrec.2017.05.002.

[7] V. Raghupathi and W. Raghupathi, “Healthcare Expenditure and Economic Performance: Insights from the United States Data,” Frontiers in Public Health, vol. 8, no. 156, May 2020, doi: https://doi.org/10.3389/fpubh.2020.00156.

[8] K. P. Vatcheva and M. Lee, “Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies,” Epidemiology: Open Access, vol. 06, no. 02, 2016, doi: https://doi.org/10.4172/2161-1165.1000227

[9] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: https://doi.org/10.1098/rsta.2015.0202

[10] M. Williams, C. Alberto, G. Grajales, and D. Kurkiewicz, “Assumptions of Multiple Regression: Correcting Two Misconceptions,” Practical Assessment, Research & Evaluation, vol. 18, no. 11, 2013, Available: https://files.eric.ed.gov/fulltext/EJ1015680.pdf

[11] All plots and code was written in jupyter notebook