WHAT AFFECTS THE ECONOMY OF A COUNTRY?

Data Science Project Report

Submitted by:
Zhen Wong - 4576198
Vikram Niranjan - 45484450
Kirti Khade - 45733130
Tejeshvini Ashre - 45603402

TABLE OF CONTENTS

The Story

The main aim of our project is to understand what affects the economy of a nation and what could be the factors that cause the economic gap between countries. We wish to understand what factors influence the economies and how they are different.

We also wish to recommend what economies should do differently so as to increase there economic status and economic growth

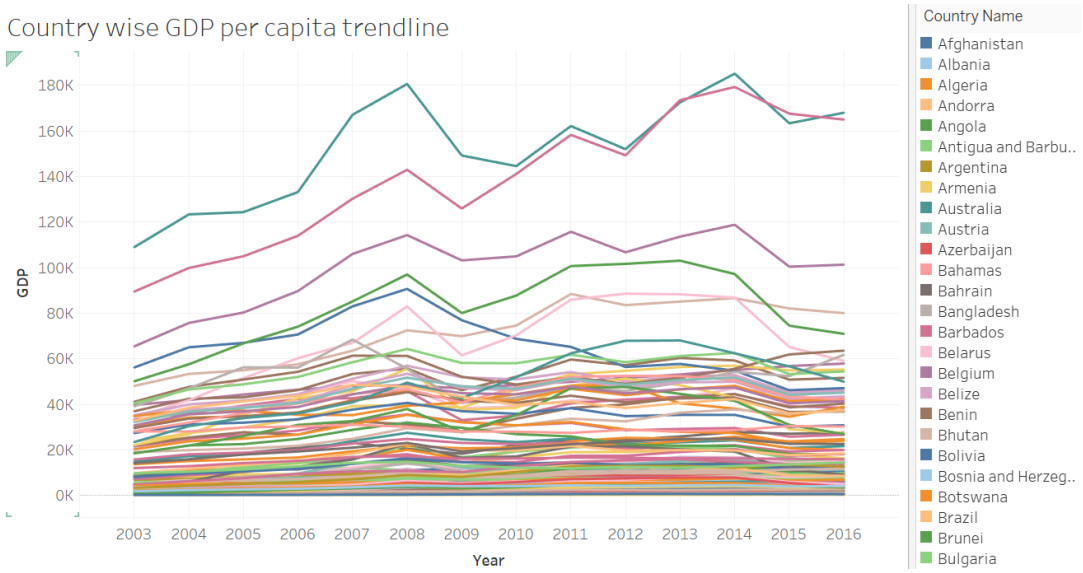
GDP as a Measurement

According to Investopedia

> "Gross Domestic Product (GDP) is a broad measurement of a nation's overall economic activity. GDP is the monetary value of all the finished goods and services produced within a country's borders in a specific time period. "

We have taken GDP per capital as a measure to understand the economic status of a country, as GDP per capita is not only measure of a country's economic output, but it also accounts for its number of people. It divides the country's GDP by its total population. That makes it the best measurement of a country's economic condition. It tells you how prosperous a country feels to each of its citizens.

The graph below shows how GDP per capita had been changing for different countries from 2003 to 2016.



MAJOR INSIGHTS

The distribution of economic sector in a country seemed to have a significant effect on the GDP of a nation. We observed that the countries with higher percentage of income from the industrial sector seemed to have a positive impact on GDP when compared to countries with lower percentage of income from the industrial sector.

While Industrial sector seemed to have a positive impact when compared to agricultural sector, the income from services sector seemed a prominent contributing factor in all the countries.

When we segregated the countries based on their GDP (top 10 countries with high GDP vs bottom 10 countries with low GDP), we observed that there seemed to be a higher percentage of trade, import and export activities in countries with high GDP. Henceforth we can say that apart from regular income from services, trade, import and export activities play a vital role in determining the GDP of a country.

On analysing logistics, we observed good relationships with GDP. Our suggestion would be to concentrate on Quality of trade and transport related infrastructure, Ease of arranging competitively priced shipments and Efficiency of customs clearance which play a major in determining a country's GDP.

In regards to technology, we found a relationship between cellular connections per user and GDP. We drilled down to check if fixed telephone lines had a similar impact but the indicator did not have a good relationship with GDP. Then we narrowed it down to access to internet having a big impact on GDP. Almost all records with GDP greater than 20k had more 50% access to the internet.

DATA SCIENCE PROCESS

Problem Solving with data

Design thinking:

What is the data:

The Data used for this project was from the World Bank which had data for the GDP per capita and many other such factors bucketed in 16 major categories (Like Agriculture and Rural Development, Economy, Education, Trade etc). This data was available from 1960-2018 for all the countries of the world. This data also had several countries grouped together based on different criteria like high income, low income etc. Each of these indicators were defined in other files (meta data). The total number of indicators in all the categories were around 2000.

Stakeholders:

The major stakeholder of this data is the world bank (owns the data). However, the other stakeholders who could benefit from this data would be the decision makers of economies of nations, financial sectors of nation and the government of countries.

Scenarios of use:

The prime scenario of use of this data in our case would be understanding factors affecting GDP and how much impact they have on the GDP per capita of a nation. This data set can also be used to individually view various factors of a nation like population growth, development in education, expansion of trade and industries etc. This can be utilized be specific departments of countries to concentrate on improving their performance and therefore benefiting the entire nation.

Limitations:

The limitation from this data set was that the data was missing for some countries like Kiribati, while for other countries data for some years was missing. The second challenge we faced in using this data was that it was difficult to find large number of factors significantly affecting GDP per capita of a nation.

GETTING THE DATA I NEED

The dataset used for this project was the world bank data. This data collected and stored with an intention to provide all users with the access to the world data. This data had around 2000 odd indicators belonging to different sectors of the economy – Agriculture, education, finance etc. We tried to correlate each of these factors to GDP and find the factors that had a strong relationship with GDP.

After choosing GDP as a criterion to measure a country's financial stability, we went in search of information regarding the GDP of each country split by specific time frames. This would allow us to analyse and observe trends in the target variable which then can be linked to independent variables. We started searching kaggle for reliable data sources and found just one reliable data source which was maintained by Kaggle. The dataset was updated regularly by Kaggle(GDP dataset- Kaggle). But the data was not available in the format we expected. The data source ranked countries based on the GDP but lacked the GDP value of the countries. The source also was live making it difficult to go through the older versions using kaggle APIS to build a dataset split by time frames.

We did come across a couple of datasets in Kaggle which not just gave information on the GDP but other factors which we could use as independent variables. But most datasets were limited to specific year. The reliability of the datasets was an issue too since most of them were hosted by individuals. And most datasets had several other factors along with GDP, but there was no proper information on how these factors were chosen. We wanted to make the data we used was reliable. So we started researching more on the dataset maintained by Kaggle.

We understood that kaggle pulled data hosted from World Bank using APIS. The daily update from kaggle was done through APIS and World Bank was the primary owner of the data. The World Bank is a vital source of financial and technical assistance to developing countries around the world and their Open Data website offers free access to comprehensive, downloadable indicators about development in countries around the globe. The website hosted information about over 2000 indicators for more than 200 countries. The data was also split by years with information from 1960 to 2018 being listed.

Data was available in various categories – 16 in total, namely: Agriculture & Rural Development, Aid effectiveness, Climate change, Economy & Growth, Education, Energy & Mining, Environment, External Debt, Finance sector, Gender, Health, Infrastructure, Poverty,

Private Sector, Science & Technology, Social Development, Social Protection & Labour, Urban development. The website did not have one file with all data but individual files for each indicator which was downloadable from individual webpages. Since there 2000 such files, we planned to automate the process.

Initial plan to pull data from the source:

1. Scrap the website for the links to other web pages hosting the downloadable files (Home Page).

2. Scrap the individual web pages for the links to the hosted files. (Sample Web Page)

3. Download csv files and convert them into data frames. Merge the data frames together.

But the downloaded files were compressed folders with metadata file and the dataset. We had issues pulling out just the dataset by uncompressing the folder. When searching for alternatives in the World Bank website, we came across the same data hosted in a different format. There were datasets with indicators grouped into 20 different categories. So each file had data for around 60 to 120 indicators. We manually downloaded these files and planned on integrating them later during the project.

IS MY DATA FIT FOR USE?

IS THE DATA RELIABLE?

1) International System: World bank works closely with international communities, including United Nations, International Monitory Fund to get the right data to the users
2) Reliable & Relevant: For every country (especially developed and underdevelop) there is a lot of awareness that needs to be raised about statistics, therefore World bank works in close partnership with these countries to get the data right.
3) Training & Client Services: World bank also engages in Training and client services to get better and better data for use.

DATA QUALITY

The data taken for this project was in its published format hence it was already suitable for use without much cleaning or changes. This means that the majority of the work such as data

transformation isn't necessary. The data (such as GDP, GNP, etc.) were compared to other sources and found to be the same, which means there shouldn't be any problem with our dataset accuracy. While our data covers from 1960-2017 for most indicators, for our project we picked a range of recent years to ensure that the data is still relevant.

The only issue was the missing values for certain years of some indicators (features). This problem was resolved through the use of multiple imputation by chained equation (more details in the Data Enrichment section).

Data Exploration

Exploratory data analysis was performed on multiple datasets to determine the potential correlation between each indicator to a nation's GDP per capita. These included box-plots and histograms, which were used to understand the distributions of each indicators as well as any potential outliers. Scatterplots were also used to explore any potential correlation an indicator may have to GDP per capita.

One of the challenges of data exploration was how time consuming it would be to do exploratory data analysis on the hundreds of indicators. This was not practical which meant that a dimensionality reduction technique had to be adopted to reduce the number of indicators. Techniques such as Principal Component Analysis (PCA) were considered at the start, but the issue with PCA was that it reduced the indicators down to principal components which contained parts of different indicators, which made it hard to interpret. Sparse Principal Component Analysis was also suggested but the same issue with interpretability was still present. Sparse PCA produces a linear combination that contains just a few of the input variables, and while this makes it easier compared to PCA, it would still be difficult to interpret parts of an indicator, (e.g. the indicator for the number of people with access to internet per 1000 people, would be difficult to interpret if one of the principal components only contain a certain fraction of this indicator).

This project required indicators to be in their original form to draw insights from them, hence feature selection was chosen instead of dimension reduction techniques. The simple feature selection method chosen was to choose a set number of indicators that had the highest correlation to GDP per capita. Each indicator was plotted on a scatter plot to investigate for trends. Histograms were also used to determine what the normality of each indicator is.

DATA ENRICHMENT

For the project's purpose, multiple datasets were required to cover as many areas as possible correlating to GDP of a nation. Areas such as agriculture, economy, education, energy, trades, finance were all included. The integration of multiple datasets into a single main data warehouse was not as complex as it could have been as the structure of each database is mainly the same. The integration part was done with Python and R using Jupyter Notebook/RStudio.

The missing values of our datasets were imputed using the multiple imputation by chained equation (MICE) package available on R. Through our exploratory data analysis, we found that most of our data were not normally distributed. This made the default method for R's MICE package, predictive mean matching (PMM), an attractive option since it could be used for variables that are not normally distributed.

Predictive mean matching works by choosing two variables, where one has missing values while the other doesn't, and then estimating a linear regression of the variables with missing values on the variables that does not have missing values, to generate a set of coefficients. A random draw is made from the posterior predictive distribution of these coefficients to produce a new set of coefficients which are used to predict all values for the missing variable (missing and not missing). For each missing values of the variable, its predicted values are compared to other predicted values of non-missing values in the same variable. A set with k number of observed non-missing values are chosen based on how close its predicted values are to the missing predicted values. One value from this set of observed values is chosen and imputed for the missing value. These steps are then repeated for all missing values.

This method of imputing values does not generate any values, and follows the range of previously observed values, e.g. if all observed values are either 0 or 1, the new imputed values would also be either 0 or 1. Similarly if the observed values all sit within the range of 0-10, the new imputed values would follow suit. The linear regression done in this method is not used to generate new values, rather as a metric to determine which previously observed values could be used. The randomness of the method also introduces variability in the imputed values which is desirable to stimulate real values.

MAKING THE DATA CONFESS

After missing values were imputed, we still had a major problem – The number of features or indicators we had were ~2000. The quantity of data that we had was an hindrance between data to storytelling.

Our main aim is to reduce it, for this we tried two methods:

1) (Failed approach) Principle Component analysis: This method uses orthogonal transformation to convert values that are correlated into set of values that are linearly related. We later figured that we were using it wrongly and if we do get the right implementation, we won't be able to interpret the variables for analysis.
2) (Failed approach) Spatial Principle Component analysis: Thomas suggested us to use spatial PCA, instead of PCA. This method is useful as it helps to easily interpret the output from special-PCA. When we applied this into dataset, the reduced feature we got were not very correlated with GDP (0.05 was. the max correlation). Due to little understanding on this topic, we decided to try a basic correlation method.
3) (Final) Correlation: When PCA and spatial PCA failed, we went to the most basic algorithm we knew - Correlation. For each of the 16 dataset we had, we saw the correlation of the indicator with GDP per capital, indicators more than 0.5 correlation were selected. This gave a total of 60 indicators.

The indicators could be seen as a division of 4 broad categories from here, which we have detailed in story telling.

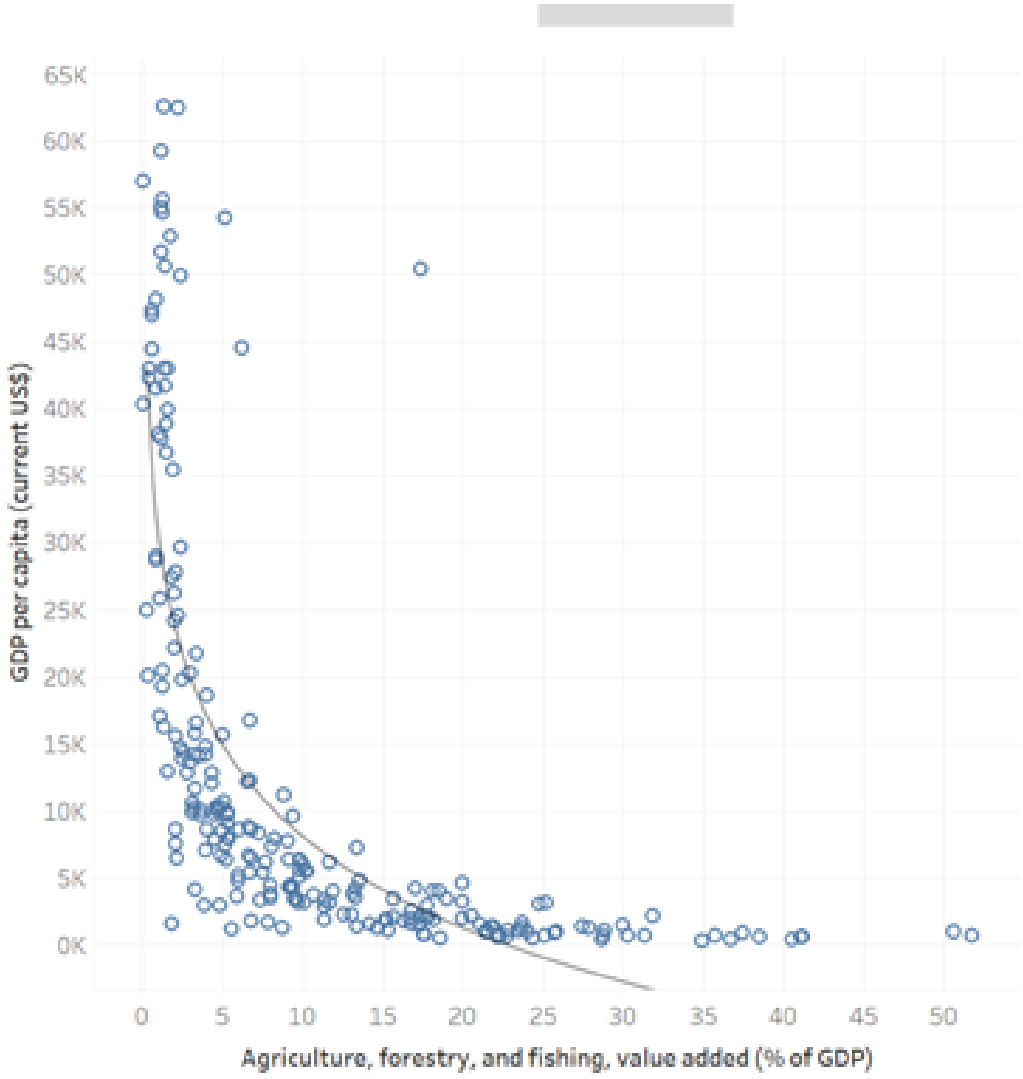
STORYTELLING WITH DATA

After reducing the number of indicators from 2000 to about a 100. We performed EDA on these factors to understand their relationship with GDP

We formed four major buckets where we saw a significant relationship between the factors and GDP – How the economies were distributed, technological advancement, impact of logistics (and it's distribution) and other human resources factors like employment and education

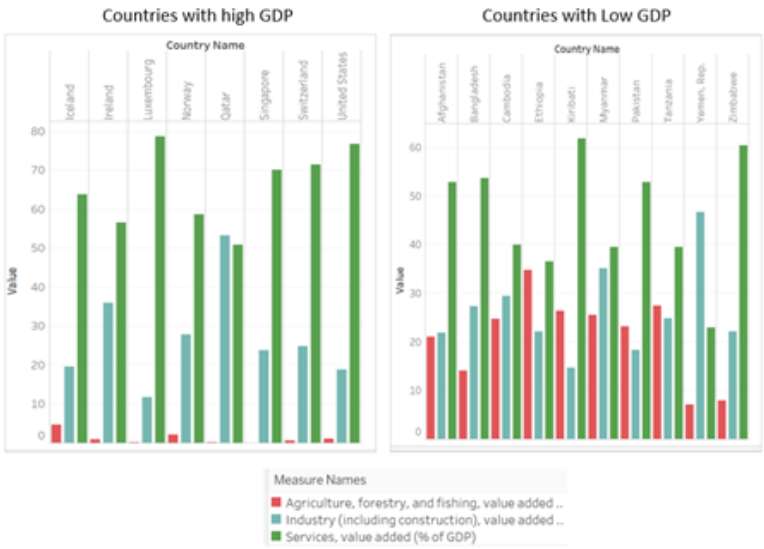
DISTRIBUTION OF ECONOMIES

When agriculture, fishing, forestry and value added (as a % of GDP) was plotted against GDP, there was a negative logarithmic relationship between the two factors.

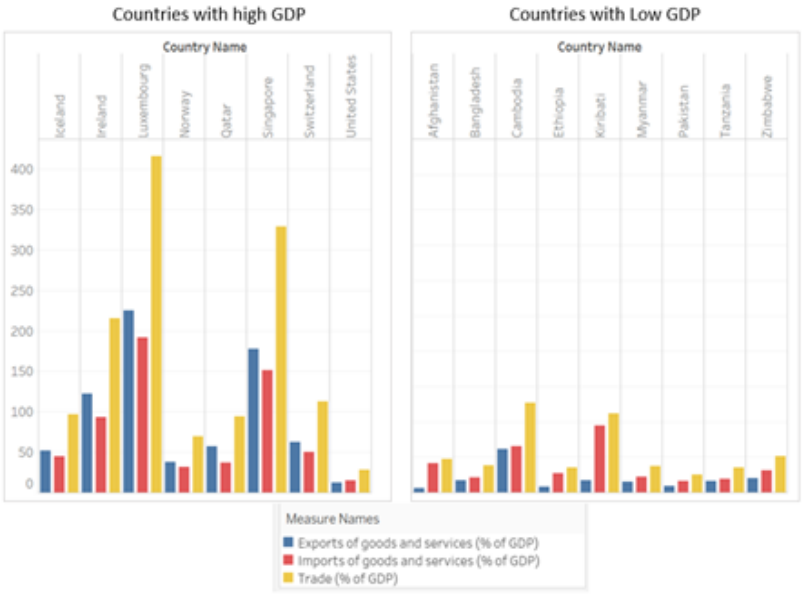


We then plotted each of agriculture, trade, imports and exports against GDP. We made 2 buckets. Top 10 countries with high GDP and bottom 10 countries with low GDP. We observed that when agriculture, industrial and services sector we grouped together for these countries, services sector seemed to be most significantly impacting GDP amongst the three.

When industrial and agricultural sectors were compared, the countries with high GDP seemed to have a significantly higher percentage of income from industrial sector when compared to the agricultural sector. Whereas, countries with low GDP have agricultural sector almost equal or very slightly lesser than industrial sector (Ex - Afghanistan)



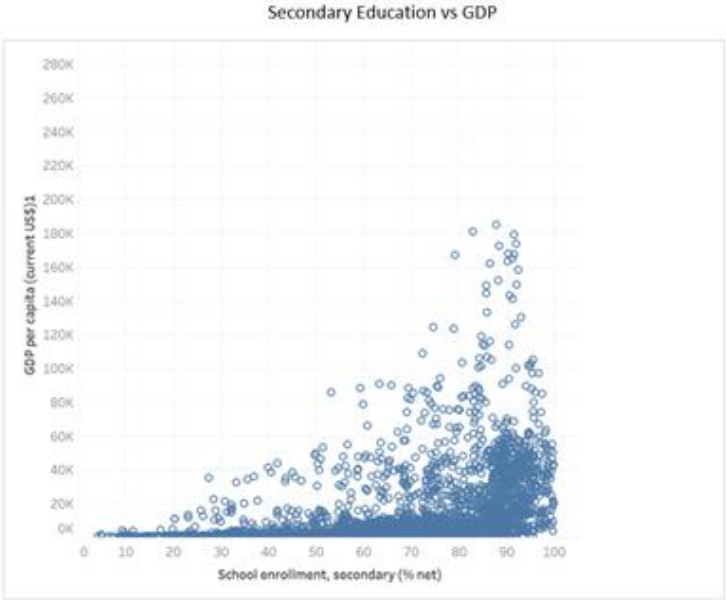
We then plotted trade, imports and exports for these two buckets. We observed that for countries with high GDP, the percentage of trade import and export activities were significantly high (400, highest trade % for Luxembourg) when compared to the countries with low GDP. This means that countries need to increase their trade, import and export activities to improve their GDP. This also helps in cross country interaction and strong international trade relationships.



HUMAN RESOURCES

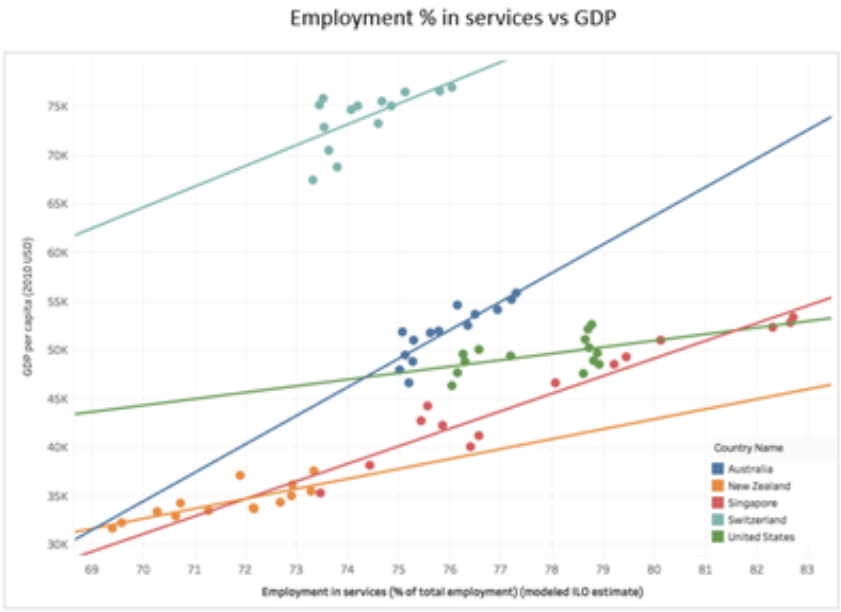
We tried to relate specific factors under the education sector which might have an impact on GDP.

We observed that enrollment in secondary education seemed to have a partial linear relationship with GDP whereas primary and tertiary education had no significant relationship with GDP. We then looked for trends of investment in education (overall and for secondary education alone) by the government, but there wasn't any significant relationship.



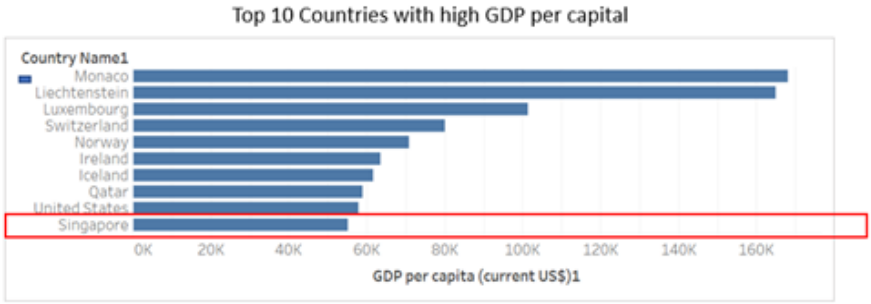
We then tried to relate how Employment in the services sector impacted the GDP of a nation.

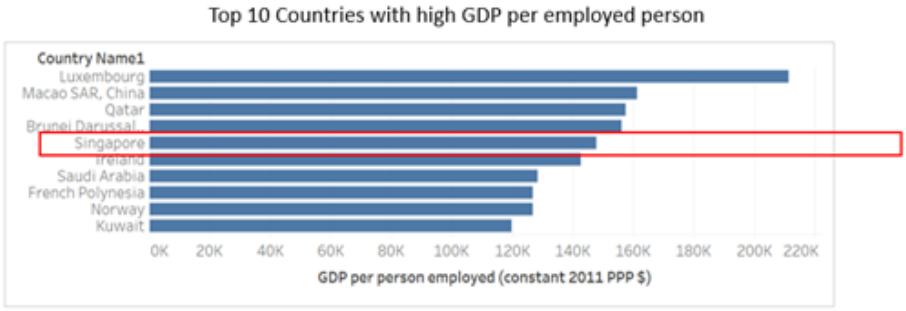
Initially it was difficult to determine a relationship as there were too many outliers. We removed outliers and isolated data for 5 countries with high rating in service sector. We then observed a positive correlation with GDP per capita and Employment in the services sector.



We then tried to relate GDP per capita of a country with GDP per employed person. We observed that not all the countries with high GDP per capita had high GDP per employed person, which meant that there were other factors contributing to the GDP of a nation.

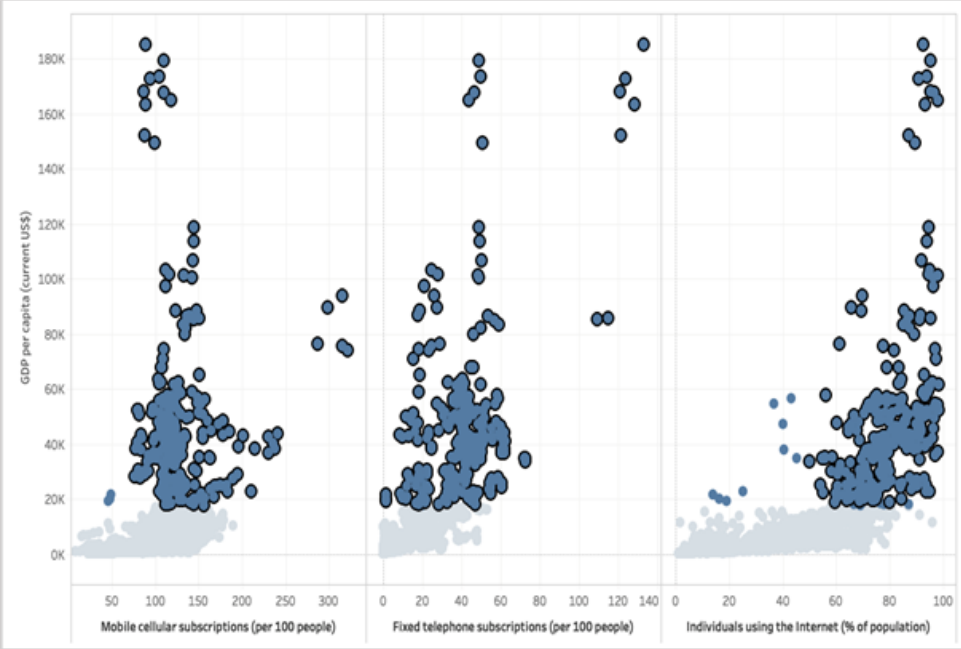
For example, Singapore which was low in countries with high GDP per capita, had a comparatively higher GDP per employed person. Countries like Brunei, Macao, Kuwait come in top ten which GDP per employed person is taken into account, but they are not present in top ten countries when GDP per capita is considered. On the other hand, countries like Liechtenstein, Monaco, Switzerland and Iceland which come in the top 10 countries with high GDP per capita, do not appear in the top 10 countries with high GDP per employed person.





TELECOMMUNICATIONS

An indicator that also stood out was mobile cellular subscriptions per 100. It can be seen in the scatter plot below that for countries with GDP per capita greater than 20k, the mobile cellular subscription rates were greater than 76%. (i.e. more than 76 out of 100 people have subscriptions).



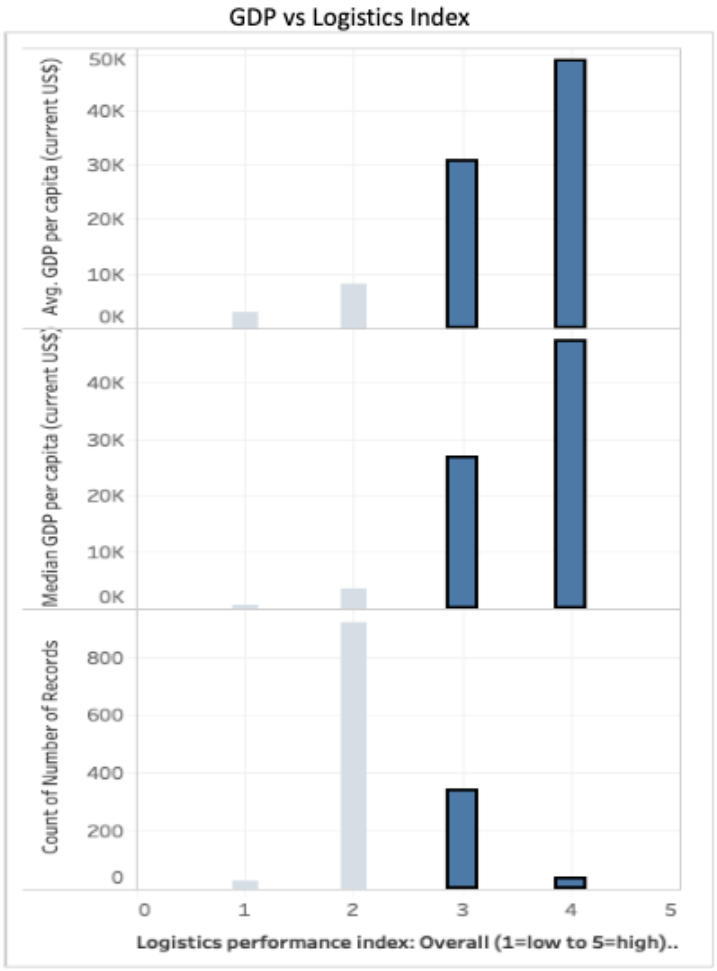
A similar indicator was also explored to determine whether communication services in general would follow the same trend. However, the indicator for fixed telephone lines per 100 people did not follow this trend. There were countries above 20k GDP per capita that had no fixed telephone subscriptions as seen in the graph above. A few possible explanations included: potential outlier countries masking the trend, landlines not as common with newer technologies (higher accessibility to cheap phones and mobile subscriptions).

An indicator that did follow similar trends to the mobile subscriptions was the percentage of individuals in the population using the internet. The graph shows that for countries above 20k GDP per capita, the majority of their population (>50%) had access to and uses the internet.

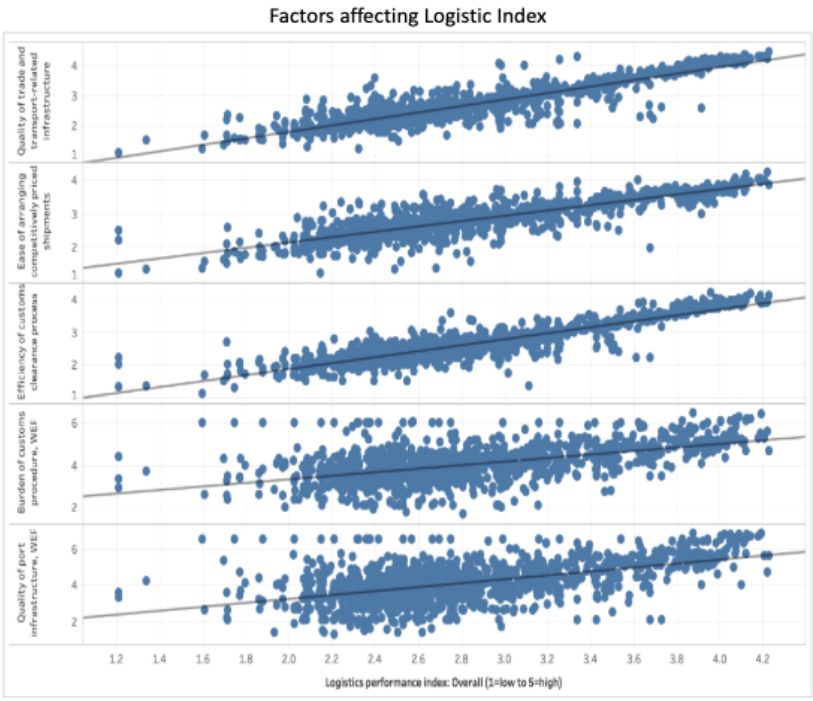
These insights are significant and can potentially be used as an indicator for countries that wishes to improve their GDP per capita.

Effect of Logistics index on GDP:
We found a good correlation value between the GDP and the logistic performance index which we wanted to drill down on. Logistics performance index is a measure of how well a country is performing in Logistics. It is a value between 1 and 5 with 1 being low and 5 being high. The initial analysis showed that mean GDP value was high when the performance index of the record was high. The mean GDP for countries with performance index less than 3 was less than 8k while the mean gdp for countries with performance index greater than 3 was

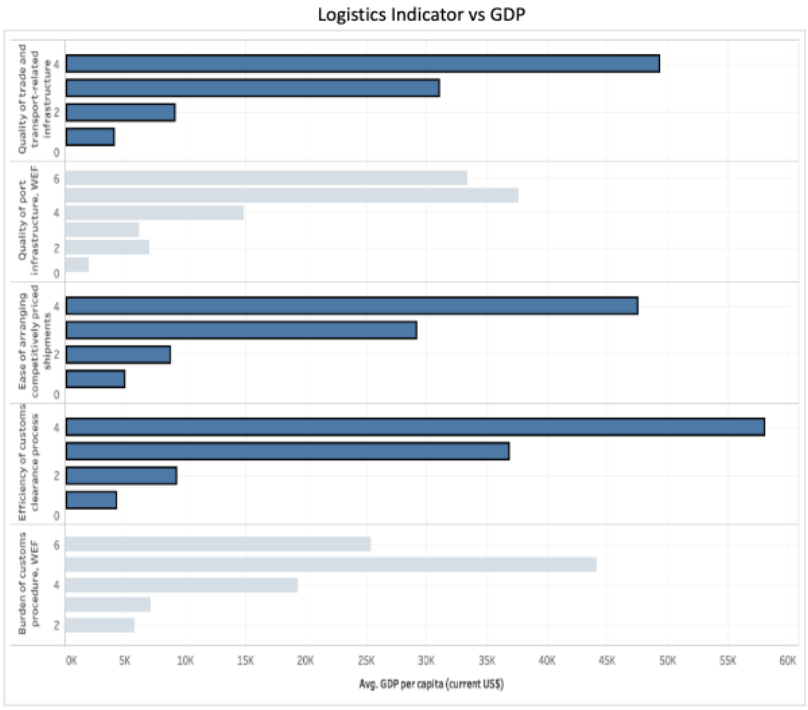
more than 30k. We also saw a trend in the decrease in the number of records with performance index greater than 3.



We then drilled down on various factors affecting Logistic performance index and we found relationships with Quality of trade and transport related infrastructure, Ease of arranging competitively priced shipments, Efficiency of customs clearance, Burdon of customs procedure
and Quality of port infrastructure. Some factors had linear relationship with the Logistics performance index while the plots of other factors were distorted.



On further analysis between factors affecting logistic index and GDP, we narrowed down the factors to three. These factors had a direct impact on the GDP. Our suggestion would be to concentrate more on improving these factors which in turn should result not just in a better logistic performance but a growth in gdp too. Quality of trade and transport related infrastructure, Ease of arranging competitively priced shipments, Efficiency of customs clearance playing a key role in logistics.



ADDRESSING FEEDBACKS

SUGGESTIONS ADDED TO THE REPORT

We compared the Export/ Import ratio in our report. We could see, higher ratio in countries with higher GDP .

GDP overtime graph was added, we could see few countries with increased GDP when compared to their GDP in 2000. However, greater difference could not be found.

Yes. Generally, it has been seen that neighbouring countries have high GDP. As shown in the below image.

GDP per capita of countries in 2003:

GDP per capita of countries in 2016:

SUGGESTIONS NOT ADDED TO THE REPORT

For the presentation, not following the order of data science process was done on purpose to first attract attention by providing the insights found before presenting the methods and technicalities. This was done on the assumption that this presentation would be for stakeholders instead of the general public. This also addresses another common feedback that the topic was too hard to follow and needed more explanation. The report does contain more details than the presentation as to hopefully provide a clearer explanation on some confusing topics.

A linear regression model was not built due to the assumptions needed to build a linear model not being met. For example, the distribution of most indicators, were not normal, hence some form of transformation was required. The problem then would be an added layer of difficulty in interpreting the newly transformed indicators, therefore it was decided best to not build a linear model.

One particular feedback was for our imputation method. It mentioned that mean imputation was used, but recommended other methods such as stochastic regression imputation. The method we implemented for imputation was not mean imputation, it was predictive mean matching with multiple imputation by chained equations. The variance generated are all within the bounds of the original data and are random, hence there is no need to run multiple different imputation methods to determine which one is best.

APPENDIX

The codes and scripts used are saved in their respective files, (e.g. r scripts, jupyter notebooks, etc.) and included with the submission

The dataset used for this project are also included with the submission

List of libraries and packages used:

Python:

- Numpy
- Pandas

R:

- MICE