

BUSAN 300 Project Report

Project Summary

This project explored the intersection of national economic performance and university-level metrics using two datasets: global GDP statistics and the World University Rankings. After cleaning and merging both datasets to resolve format inconsistencies and standardise country names, we developed a single analytical dataset. This enabled us to explore how economic indicators may relate to educational and research performance across countries.

Three core analytical questions guided this project:

1. Do countries with higher GDP per capita tend to have higher average university scores?
 - Yes. Countries with higher GDP per capita, such as Switzerland, Ireland, and the United States, consistently had universities with higher average scores. In contrast, countries with lower GDP per capita, like India and Uganda, had significantly lower university scores.

Insight:

- This trend suggests that a nation's economic prosperity likely enables stronger educational infrastructure, faculty, and research output. However, other contextual factors like national education policy and investment in higher education also play a role and merit further analysis.

2. Which countries have the highest university influence relative to their GDP?

- Smaller or less economically dominant countries like Cyprus, Uganda, and Lebanon demonstrated the highest research influence relative to the size of their economies. Conversely, countries like the United States and China, while dominant in raw output, scored lower on a per-dollar basis.

Insight:

- This finding reveals that smaller nations can achieve high research efficiency, suggesting that focused investment and academic policy can yield disproportionately strong results. It offers a new lens through which to evaluate and benchmark academic productivity beyond just total output.

3. Is there a relationship between population density and the number of patents produced by universities?

- No clear correlation was found. Although some densely populated countries like Taiwan had high patent activity, others like Singapore did not. Additionally, several less dense countries also produced high patent counts.

Insight:

- Population density alone is not a strong driver of innovation. Patent production appears more closely tied to factors like research funding, institutional priorities, and national intellectual property frameworks. Investment strategy rather than physical density is the likely determinant.

Overall Findings:

This analysis confirms that economic conditions, especially “GDP per capita” are broadly aligned with university performance. However, smaller nations can outperform expectations when measuring academic influence per GDP dollar. Meanwhile, innovation, as reflected by patent output, does not follow predictable geographic or demographic patterns. This highlights the importance of targeted education investment and smart research policy over simplistic economic or demographic assumptions.

Wrangling Details

World GDP Data Source Information:	World University Rankings Data Source Information:
<p><i>Dataset Origin:</i> https://www.kaggle.com/datasets/darknez/gdp-among-world</p> <ul style="list-style-type: none">• This data set was obtained from Kaggle, titled “GDP among world”	<p><i>Dataset Origin:</i> https://www.kaggle.com/datasets/darknez/gdp-among-world</p> <ul style="list-style-type: none">• This data set was obtained from Kaggle, titled “World University Rankings”
<p><i>General Characteristics</i></p> <ul style="list-style-type: none">• Format: JSON file• Data Year: 2020• structured as a spreadsheet-style table, containing various economic indicators by country. Is a flat structure, no nested dictionaries.• 184 rows<ul style="list-style-type: none">◦ 184 Countries• 16 columns<ul style="list-style-type: none">◦ world_rank◦ institution◦ country◦ national_rank◦ quality_of_education◦ alumni_employment◦ quality_of_faculty◦ publications◦ influence◦ citations◦ broad_impact◦ patents◦ score	<p><i>General Characteristics</i></p> <ul style="list-style-type: none">• Format: CSV file• Data Year: 2012• 2188 rows<ul style="list-style-type: none">◦ 2188 Institutions• 14 columns<ul style="list-style-type: none">◦ country◦ Population Rank◦ Density◦ 2020 Population Rank◦ Capital City◦ Region◦ Subregion◦ GDP (IMF) [\$Bn]◦ GDP (UN) [\$Bn]◦ GDP Per Capita [\\$]◦ Land Area [km²]◦ Total Area [km²]◦ Growth Rate [%]◦ World Population Share [%]◦ 2020 Growth Rate [%]◦ 2020 Population Share [%]

- | | |
|--------|--|
| ○ year | |
|--------|--|

Initial Data Audit:

GDP Dataset

15 Columns in the GDP dataset were classified as ‘objects’, including important numerical information which is key for data analytics. Such ‘objects’ include “World Percentage”, “Growth Rate”, and “GDP Per Capita”. Additionally, these values include commas, percentage and dollar symbols, or economic columns including symbols like “Tn” and “Bn” etc.

In order to perform analytics, these columns and rows need to be converted to floats, and cleaned.

In the GDP dataset, there are two columns which are majority filled with NaN, or are entirely useless in this analytical context, which are “Anthem” and “Government” which need to be removed.

Given the GDP data set is a JSON file, immediate thoughts looking at the data show no need to normalise the data as it is already normalised, structured and clean. The structure of the data is flat and there are no nested dictionaries in the data.

WUR Dataset

The World University Rankings Dataset (WUR) is relatively clean, with numerical metrics being floats or integers.

However the key problem with this dataset is country name inconsistency. In the GDP dataset countries like the United States are represented as “United States” but in the WUR dataset, the United States is represented by “USA”. The most important aspect was that Hong Kong was not present in the GDP dataset. It was not present as “Hong Kong” or “Hong Kong SAR”, so decisions about whether to drop Hong Kong entirely are needed to be made.

Additionally with the ‘Country’ Column, in the WUR dataset country was labeled with a lowercase ‘c’ while GDP has a capital ‘C’ which need to be standardised if merging is to effectively work.

To merge these inconsistencies will need to be mended, as merging is intended to be done on the country column as this is common between both datasets.

Steps to Clean and Combine the datasets:

First loaded the datasets using Pandas. Pd.read_csv for the WUR dataset, and pd.read_json for the GDP dataset.

GDP dataset Cleaning:

1. First dropped the irrelevant columns (Anthem, Government)
2. Then I Cleaned the GDP (IMF) and GDP (UN) columns
 - a. This code defines a function convert_gdp to clean GDP values by removing symbols (like \$, ,) and converting strings with "Tn" (trillion) or "Bn" (billion) into numeric float values.
 - b. It then applies this function to the 'GDP (IMF)' and 'GDP (UN)' columns, creating two new cleaned columns: 'GDP_IMF_clean' and 'GDP_UN_clean'.

```
#Clean GDP (IMF) and GDP (UN) columns
def convert_gdp(value):
    if isinstance(value, str):
        value = value.replace('$', '').replace(',', '').strip()
        if 'Tn' in value:
            return float(value.replace('Tn', '')) * 1000
        elif 'Bn' in value:
            return float(value.replace('Bn', ''))

    return None # or np.nan

gdp_df['GDP_IMF_clean'] = gdp_df['GDP (IMF)'].apply(convert_gdp)
gdp_df['GDP_UN_clean'] = gdp_df['GDP (UN)'].apply(convert_gdp)
```

3. Then Cleaned GDP Per Capita by:
 - a. removing dollar signs (\$) and commas (,) using a regular expression, then converts the cleaned string values to floats, storing the result in a new column called 'GDP_Per_Capita_clean'.

```
#Clean GDP Per Capita
gdp_df['GDP_Per_Capita_clean'] = gdp_df['GDP Per Capita'].replace('[\$,]', '', regex=True).astype(float)
```

4. Then cleaned the 'Land Area' and 'Area' columns by removing commas from the string values and converting them to floats. The cleaned values are stored in new columns: 'Land Area_clean' and 'Area_clean'.

```
#Clean Land Area and Area
gdp_df['Land Area_clean'] = gdp_df['Land Area'].str.replace(',', '').astype(float)
gdp_df['Area_clean'] = gdp_df['Area'].str.replace(',', '').astype(float)
```

5. Then I cleaned the percentage columns by removing the % symbol from each value in the specified columns and converting the result to a float. The cleaned data is saved in new columns with _clean appended to their original names.

```
#Clean Percentage Columns
percentage_cols = ['Growth Rate', 'World Percentage', '2020 Growth Rate', '2020 World Percentage']
for col in percentage_cols:
    gdp_df[col + '_clean'] = gdp_df[col].str.replace('%', '').astype(float)
```

6. Then I renamed the cleaned columns to overwrite the original column names (e.g., replacing the old 'GDP (IMF) [\$Bn]' with the cleaned version). This ensures the DataFrame now uses the cleaned data under the same familiar column names. The inplace=True argument applies the changes directly to gdp_df.

```
#Rename Columns
gdp_df.rename(columns={
    'GDP_IMF_clean': 'GDP (IMF) [$Bn]',
    'GDP_UN_clean': 'GDP (UN) [$Bn]',
    'GDP_Per_Capita_clean': 'GDP Per Capita [$]',
    'Growth Rate_clean': 'Growth Rate [%]',
    'World Percentage_clean': 'World Population Share [%]',
    '2020 Growth Rate_clean': '2020 Growth Rate [%]',
    '2020 World Percentage_clean': '2020 Population Share [%]',
    'Land Area_clean': 'Land Area [km²]',
    'Area_clean': 'Total Area [km²]'
}, inplace=True)
```

7. I then removed the original, uncleaned columns from the DataFrame (e.g., those with dollar signs, commas, or percentage symbols) since they've been replaced by cleaned versions. The inplace=True argument ensures the changes are applied directly to gdp_df.

```
#Drop Uncleaned Columns
cols_to_drop = [
    'GDP (IMF)', 'GDP (UN)', 'GDP Per Capita',
    'Growth Rate', 'World Percentage',
    '2020 Growth Rate', '2020 World Percentage',
    'Land Area', 'Area'
]

gdp_df.drop(columns=cols_to_drop, inplace=True)
```

8. I then extracted the unique country names from both the GDP dataset (gdp_df) and the World University Rankings dataset (WUR_df), then compared them to find which countries appear in the university rankings but are missing from the GDP data. It prints out this sorted list of mismatched countries for review.

```

# Unique countries in GDP data
gdp_countries = set(gdp_df['Country'].unique())

# Unique countries in University Rankings data
wur_countries = set(WUR_df['country'].unique())

# Check differences
country_diffs = wur_countries - gdp_countries
print("Mismatched Countries in WUR (not found in GDP):")
print(sorted(country_diffs))

Mismatched Countries in WUR (not found in GDP):
['Hong Kong', 'Slovak Republic', 'USA']

```

9. I then checked whether the exact strings "Hong Kong SAR" or "Hong Kong" appear in the 'Country' column of the GDP DataFrame, printing True if found and False if not. It helps confirm if Hong Kong is listed and under which exact name.

```

#Check if Hong Kong is in GDP
print("Hong Kong SAR" in gdp_df['Country'].values)
print("Hong Kong" in gdp_df['Country'].values)

False
False

```

10. I then dropped "Hong Kong" from the WUR dataset
- There are two choices, either drop Hong Kong or combine it with China.
 - Combining it with China is technically possible, however economically and educationally Hong Kong and China are treated separately.
 - Keeping it with China would inflate China's data making it less reliable
11. I then used a dictionary country_mapping to define how to rename or standardise country names mapping "USA" to "United States" and "Slovak Republic" to "Slovakia" to help align country names between datasets for easier merging or comparison.

```

#Country Mapping

country_mapping = {
    "USA": "United States",
    "Slovak Republic": "Slovakia"
}

#Change Country Names WUR_df

WUR_df['country'] = WUR_df['country'].replace(country_mapping)

```

12. I then checked to see if there were still any missing countries which there were not.

```
#Verify if all countries now match

mismatches_after_mapping = set(WUR_df['country'].unique()) - set(gdp_df['Country'].unique())
print("Remaining mismatches:", mismatches_after_mapping)

Remaining mismatches: set()
```

WUR Dataset Cleaning:

1. I first filled any missing (NaN) values in the 'broad_impact' column of WUR_df with the median value of that column, ensuring no gaps in the data for that feature.
2. I then reset the index to be safe
3. Finally I renamed the 'Country' column in gdp_df to lowercase 'country' so it matches the column name in WUR_df.

```
▶ #Handle Missing broad_impact
WUR_df['broad_impact'] = WUR_df['broad_impact'].fillna(WUR_df['broad_impact'].median())

[23] #Reset Index
WUR_df = WUR_df.reset_index(drop=True)

▶ #Rename country to match gdp_df 'Country'
gdp_df = gdp_df.rename(columns={'Country': 'country'})
```

Merge Datasets

1. I then performed a left join merge, keeping all rows from the university rankings DataFrame (WUR_df) and adding matching GDP data from gdp_df based on the 'country' column so that additional information will be added for the WUR dataset. If no GDP data exists for a country, those fields will have NaN.

```
#Merge Datasets
merged_df = pd.merge(WUR_df, gdp_df, on='country', how='left')
```

2. I then exported the dataset as a csv into my google drive without including the DataFrame's index in the saved file. I then load into excel using "get data".

```
merged_df.to_csv('/content/drive/My Drive/merged_dataset.csv', index=False)
```

FINAL PYTHON WORKING:

<https://colab.research.google.com/drive/1a-ecBCZEUmUUueT-BpyuigJoL0SS4pj?usp=sharing>

Questions and Answers

1. Do countries with higher GDP per capita tend to have higher average university scores?

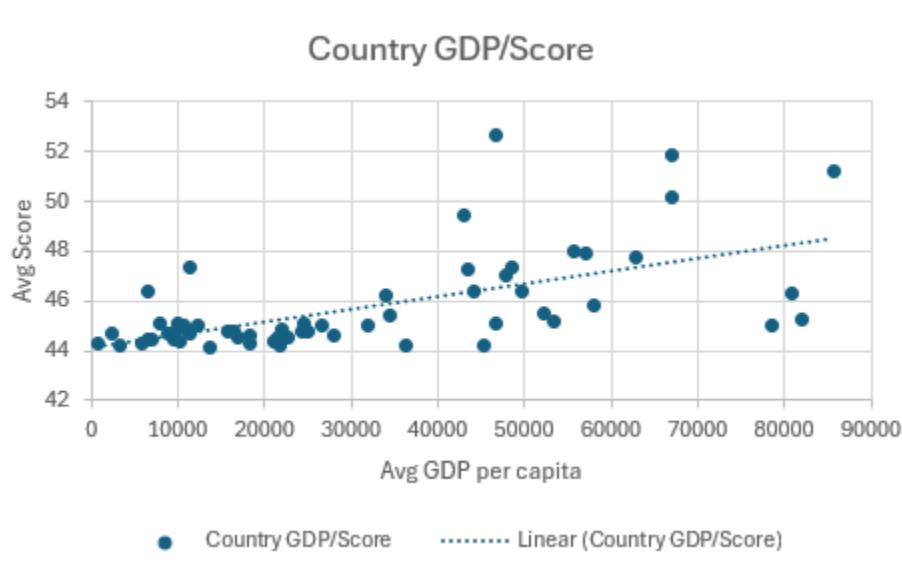
(This analyses the potential relationship between a country's wealth per person and the performance of its universities)

Answer:

There is a positive correlation between a country's GDP per capita and the average university ranking score. Countries with higher GDP per capita tend to host universities with higher average scores to determine world rank in the World University Rankings dataset.

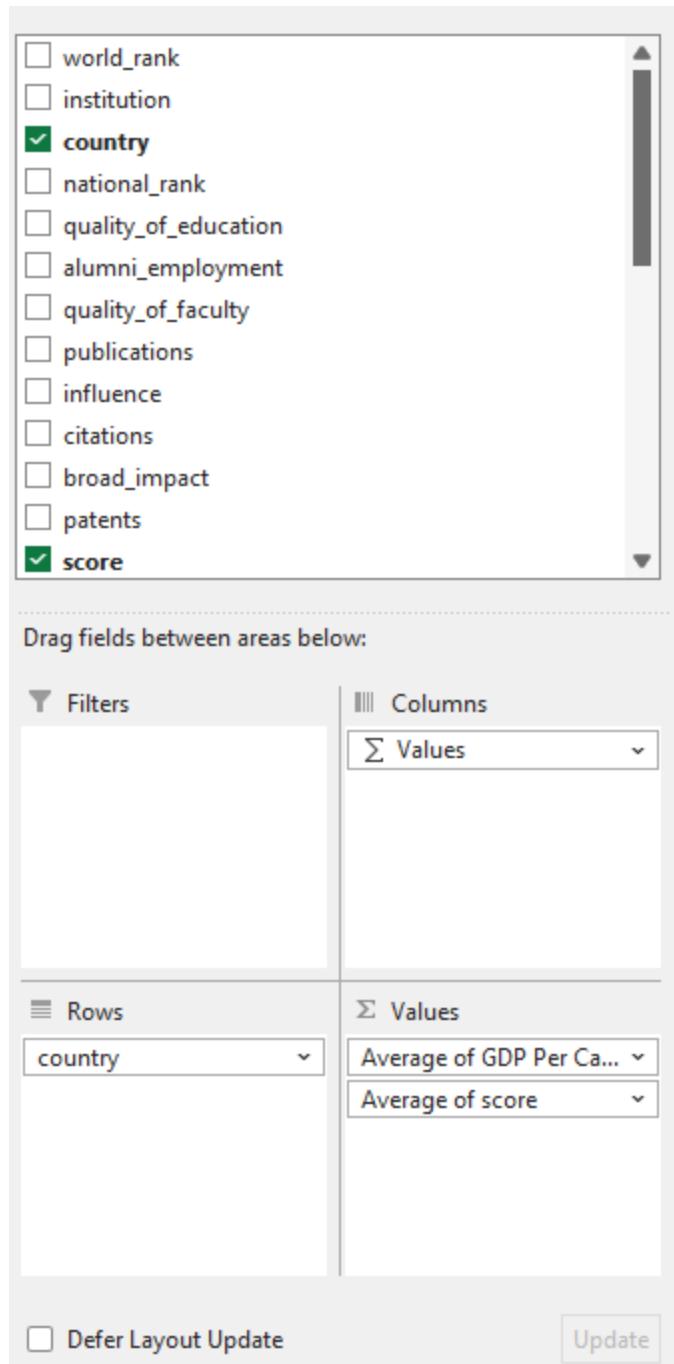
The analysis shows a clear trend: countries with high GDP per capita such as Switzerland (\$85,584), Ireland (\$82,058), Norway (\$80,908), and the United States (\$67,063) also have relatively high average university scores, many above 50.

Conversely, countries like India (\$2,361) and Uganda (\$735) have much lower average scores, clustered just above 44. This suggests a general correlation between a country's economic prosperity and the quality of its universities. However, the variation within some middle-income countries indicates that other factors like education policy or research funding may also play significant roles.

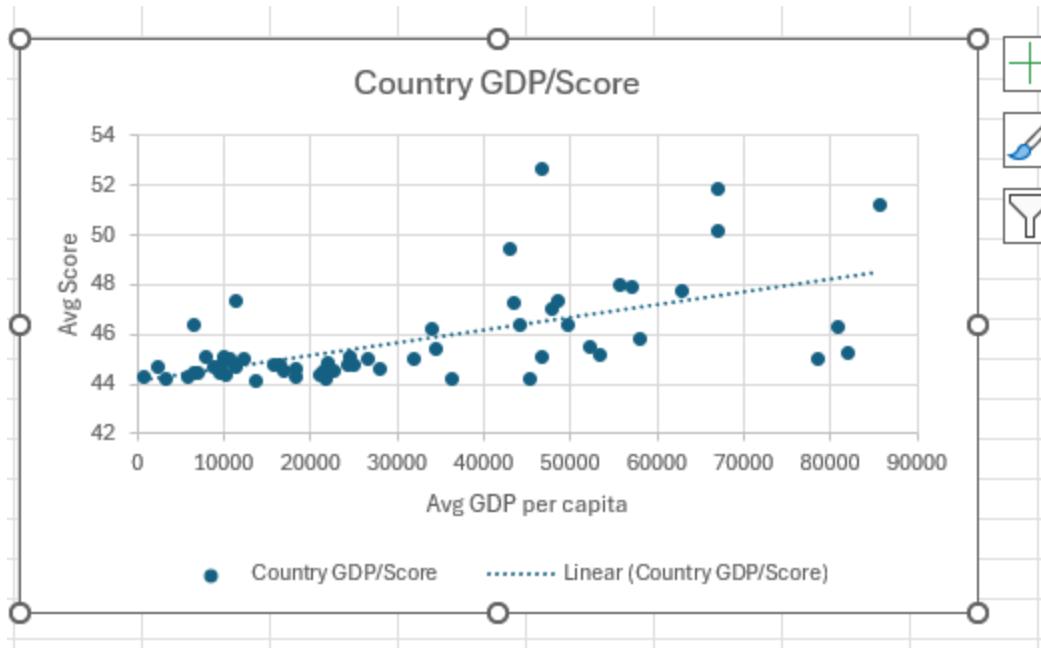


Steps to achieve result:

- Using the merged dataset in its final CSV format, I loaded the data into Microsoft Excel.
- I created a pivot table with countries as rows, and calculated the average GDP Per Capita [+] and average university score as values.



- To visualise the relationship, I created a scatter plot with average GDP per capita on the X-axis and average university score on the Y-axis.
- I then added a trend line by clicking on the + icon on the right, and adding trend line.



- The chart showed a clear upward trend, supporting the conclusion that wealthier countries (per capita GDP) generally have higher scoring universities.
- Excel's filtering options were used to identify outliers and ensure the data was not skewed by any single country.

Potential Setbacks:

- The analysis is correlational; it does not prove causation. Other factors (such as education policy, research funding) may mediate this relationship.
- GDP per capita is an average measure and may mask income inequalities within countries.
- The university ranking score is an aggregate metric that combines several criteria; isolating which factors drive the correlation would require further analysis.

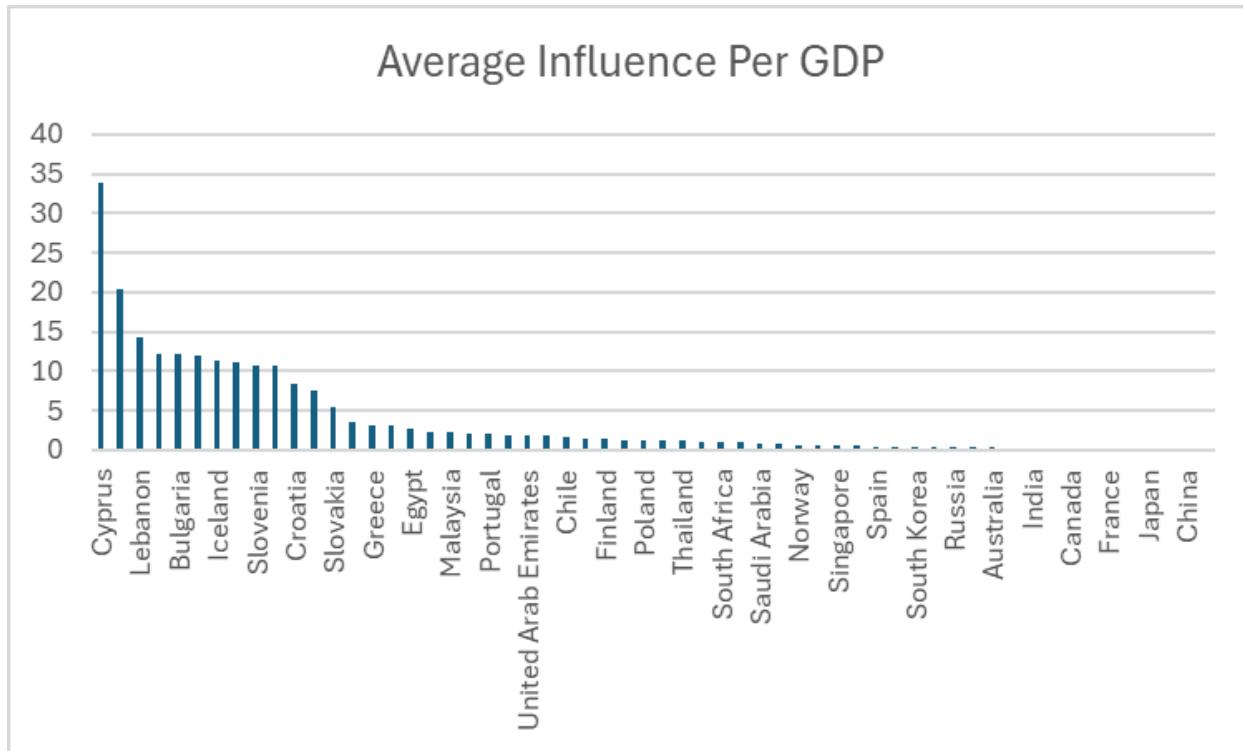
2. Which countries have the highest university influence relative to their GDP (IMF)?

(This shows which countries "punch above their weight" having universities with high influence scores even with smaller economies)

Answer:

The result of this was surprising with Cyprus, Uganda, and Lebanon who led the rankings, indicating that institutions in these countries exert high research influence relative to their economic size. Compare this to the lower down countries who have large, research-heavy economies like the USA, China and Germany ranked lower.

While these major economies produce significant research, their vast GDPs dilute the influence-per-GDP measure. This reveals that smaller nations can achieve high research efficiency, offering a valuable lens for evaluating academic productivity beyond raw output.



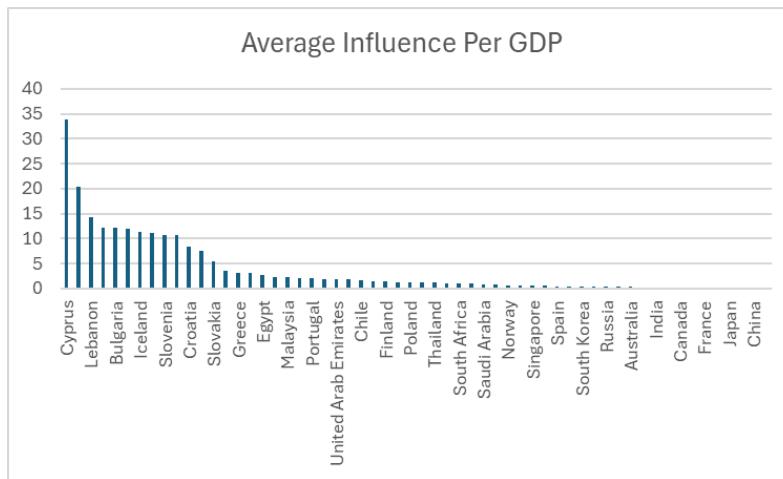
Steps to achieve result:

- Added a derived column in the merged dataset, Influence per GDP, calculated as influence / GDP (IMF) [\$Bn].
 - Column I was the “Influence” metric and Column U was the GDP (IMF) [Bn]

AD
Influence per GDP
=I2/U2
0.00018018
9.00901E-05
0.005460751
0.000990991
0.001486486
0.00443686
0.00027027
0.000540541
0.000225225
0.000900901

- Refreshed the final data set and created a pivot table with Region as rows.

- Summarised by calculating the average Influence per GDP for each region.
- Sorted the pivot table descending by this value to identify the regions with the highest relative influence.
- To confirm, generate a bar chart from the pivot table for visual clarity.



Setbacks:

- GDP values are reported at the country level, but influence scores are at the institution level, which introduces some aggregation challenges.

- Differences in reporting years and data source discrepancies between GDP and university data might affect precision.
- This metric does not account for the absolute number of institutions in each region, only averages.

3. Is there a relationship between population density and the average number of patents produced by universities in each country?

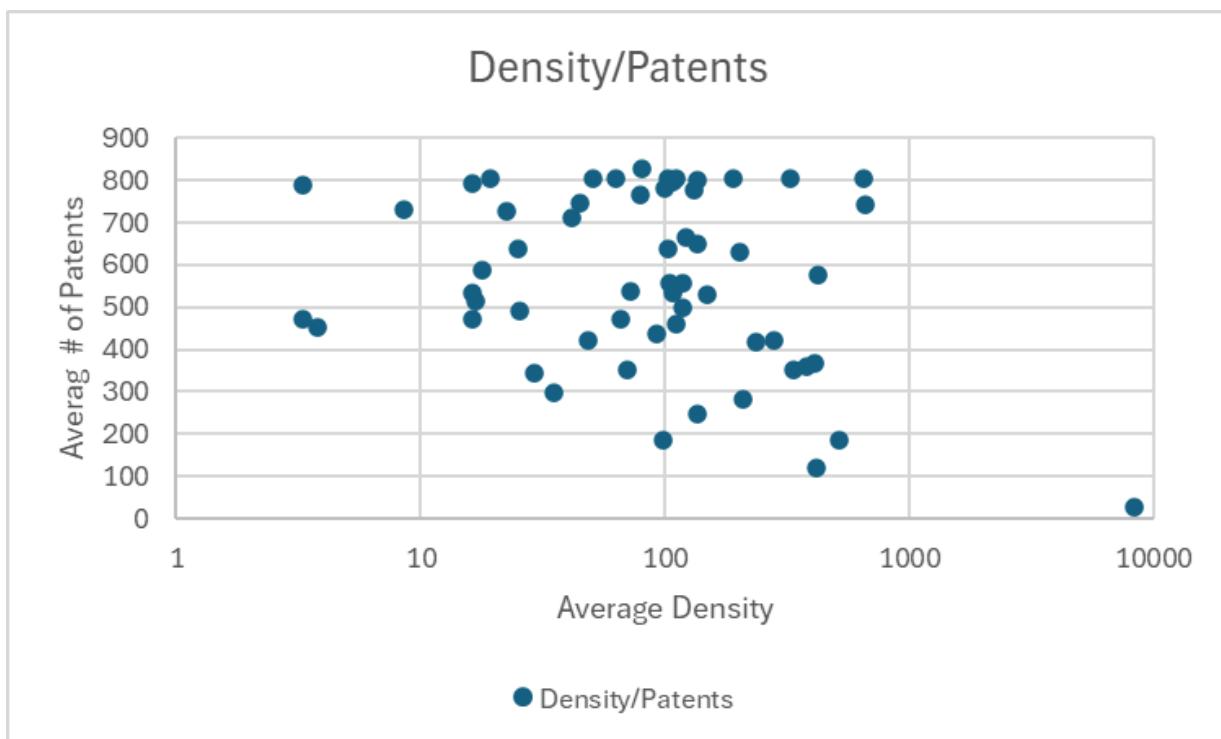
(This explores if more densely populated countries have more innovation (as measured by university patent production).

Answer:

The results showed the countries like Taiwan (658/km²) not only have high densities but also have high average patent counts (eg: 742 for Taiwan). However, there are outliers like Lebanon and Puerto Rico who also ranked high in patents despite smaller sizes. Looking at the scatter plot, there appears to be no correlation between average density and average number of patents as they are all clustered in the middle. There are countries who have low Average Density and a high average number of patents, while countries with high density and low avg patents.

With the data there is an outlier on the far right, of Singapore who had the highest Average density at (8,240/km²) while only having an average of 26 patents.

Overall there appears to be no relationship between population density and the average number of patents produced by universities in each country.



Steps to achieve result:

- Using the merged dataset, I created a pivot table grouping data by country.

PivotTable Fields

Choose fields to add to report:

|

world_rank
 institution
 country
 national_rank
 quality_of_education
 alumni_employment
 quality_of_faculty
 publications
 influence
 citations
 broad_impact
 patents
 score

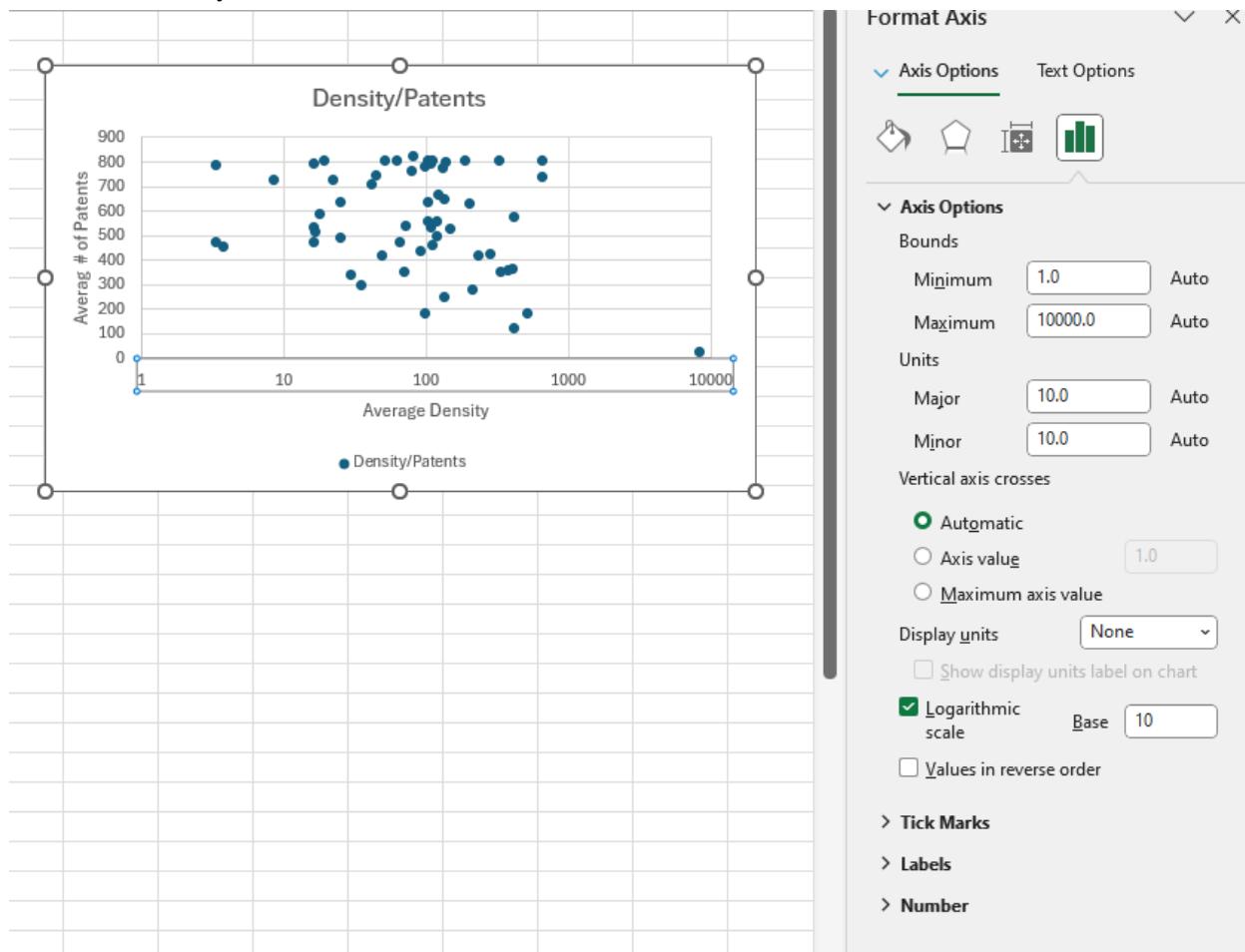
Drag fields between areas below:

Filters	Columns
	<input type="button" value="Σ Values"/>

Rows	Values
country	Average of Density Average of patents

Defer Layout Update

- Calculated the average Density (population density) and average patents per country's universities.
- Generated a scatter plot charting average density (X-axis) versus average number of patents (Y-axis).
 - Found that the result was troubled with many points clustered on the left side of the graph unable to see a relationship.
- I then changed the x axis scale to a logarithmic scale with a base of 10 in order to effectively see the trend better.



- Observed clustering of most countries at low to moderate density with varied patent counts, but noted outliers such as Singapore with extremely high density but relatively low patent counts.
- This suggests population density alone is not a strong predictor of patent output in universities.

Setbacks:

- Patents data might be influenced by university size, focus on research, and country-level IP laws, none of which are captured here.
- Population density is a general country-level metric and may not reflect the urban concentration around universities.

- Further data on research funding and university size would improve analysis.

References:

- Kakkar, R. (2020). GDP among world [Data set]. Kaggle.
<https://www.kaggle.com/datasets/rishikakkar/gdp-dataset>
- O'Neill, M. (n.d.). World University Rankings [Data set]. Kaggle.
<https://www.kaggle.com/datasets/mylesoneill/world-university-rankings>