Predykcja PKB per capita państw w zależności od czynników edukacyjnych, położenia geograficznego oraz populacji

Bartosz Rodowicz
MSiD, K01-21d
IST, W04N, semestr 4
Politechnika Wrocławska
266549@student.pwr.edu.pl

Czerwiec 2023

1 Wstęp

Tematem niniejszej pracy jest predykcja PKB per capita państw w zależności od czynników edukacyjnych, położenia geograficznego oraz populacji. Do badania wykorzystano następujące dane dla poszczególnych państw:

- PKB per capita i położenie geograficzne
- wskaźnik procentowy PKB wydawany na edukację
- populacja
- udział procentowy osób z wyższym wykształceniem w całej populacji
- ranking 1000 najlepszych uniwersytetów świata

Celem projektu jest zbadanie zależności między PKB per capita, a wyżej wymienionymi wskaźnikami. Analiza tych danych pozwoli na uzyskanie odpowiedzi na pytanie jak poziom wykształcenia populacji, jej liczba oraz inwestycje w edukację mogą wpłynąć na poziom życia i zamożność obywateli państwa.

2 Dobór danych i ich analiza

2.1 Gromadzenie danych

Dane zostały pozyskane łącznie z 6 stron internetowych. Dane o PKB per capita, podawane w dolarach amerykańskich (i nazywane w dalszych etapach dokumentu GDP_per_capita)[4], zostały pobrane z wikipedii za pomocą techniki scrapowania strony internetowej. Dane te zawierają także informacje, na

jakim kontynencie znajdują się poszczególne kraje (kontynenty określane są potem jako UN_Region). Techniki scrapowania stron użyto również do pobrania danych dotyczących wskaźnika procentowego wydawanego na edukację (określanych potem jako Expenditure_on_education_ (%_of_GDP)) [5], danych o populacji państw (inaczej Population) [3] oraz udziału procentowego osób z wyższym wykształceniem w całej populacji (Tertiary_edu_%) [6]. Wymienione zbiory danych także pozyskano z wikipedii. Ranking 1000 najlepszych uniwersytetów świata pobrano z kaggle w formie pliku csv [1].

2.2 Przetwarzanie wstępne (pre-processing)

Po weryfikacji pobranych danych okazało się, że dane o udziale procentowym osób z wyższym wykształceniem w całej populacji dotyczą jedynie 45 krajów, głównie z Europy oraz Ameryki Płn. i Płd., dlatego też pobrano dodatkowy uzupełniający zbiór danych ze strony World Banku, który zawiera potrzebne informacje o znacznie większej ilości państw [2]. Następnie przetwarzaniu poddano ranking 1000 najlepszych uniwersytetów świata. Należało utworzyć nowy zbiór danych zawierający spis krajów oraz policzyć dla każdego z nich liczbę uczelni wyższych znajdujących się w rankingu i będących z danego kraju. Brakujące dane (dla krajów, które nie posiadają żadnej uczelni znajdującej się w rankingu) uzupełniono wartościami 0. Dane o PKB per capita zawierają wartości PKB z trzech różnych źródeł (IMF, World Bank, United Nations) i z różnych lat (głównie 2020-2022). Niestety nie wszystkie kraje posiadają dane ze wszystkich trzech źródeł. Utworzono zatem nową kolumnę z wartościami PKB per capita obliczanymi jako średnia wartości ze wszystkich dostępnych dla danego państwa źródeł. Ameryka Północna i Południowa są przedstawiane razem jako "Americas"w kolumnie dotyczącej kontynentów, natomiast Australia jest przypisana do Oceanii. Dane o wskaźniku procentowym PKB wydawanym na edukację, populacji i udziale procentowym osób z wyższym wykształceniem w całej populacji (mniejszy zbiór [6]) nie wymagały większej ingerencji (oprócz zmiany nazw i usuwania niepotrzebnych kolumn). Brakujące dane o wskaźniku procentowym PKB wydawanym na edukację zostały uzupełnione poprzez wyliczenie średniej wartości wskaźnika na każdym kontynencie i następnie przemnożonej przez losowo wygenerowaną wartość z przedziału (0.8, 1.2) dla każdej uzupełnianej wartości. Dane o udziale procentowym osób z wyższym wykształceniem w całej populacji powstały poprzez połączenie dwóch pobranych zbiorów danych (brakujące dane w mniejszym zbiorze [6] uzupełniono danymi z większego zbioru [2]). Pozostałe brakujące wartości wyliczono analogicznie, jak brakujące dane o PKB per capita.

Ze wszystkich zbiorów danych usunięto niepotrzebne w dalszej części kolumny oraz zmieniono nazwy pozostałych kolumn na bardziej informatywne. Przetworzone zbiory na końcu połączono w jeden zbiór (łączenie odbywało się po nazwach krajów).

2.3 Wstępna analiza danych (Exploratory data analysis)

Wstępna analiza danych została wykonana za pomocą narzędzia ydata_profiling, które jest zamiennikiem dla pandas_profiling dla Pythona 11. Poniżej przedstawione są ogólne informacje o przetworzonym już zbiorze danych, rozkłady poszczególnych zmiennych i zależności między nimi. W poniższych rozkładach danych zmiennych numerycznych dla lepszej wizualizacji rozkładu na wykresie usunięto skrajne 5% największych i najmniejszych wartości z danych.

Dataset statistics

Number of variables	7
Number of observations	193
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	10.7 KiB
Average record size in memory	56.7 B

Variable types

Text	1
Numeric	5
Categorical	1

Rysunek 1: Statystyki ogólne zbioru danych

	Country	Population	UN_Region	GDP_per_capita	Top_1000_Uni_Count	Tertiary_edu_%	Expenditure_on_education_(%_of_GDP)
0	Afghanistan	32890171	Asia	451.0	0	3.1	4.1
1	Albania	2793592	Europe	6649.0	0	12.9	4.0
2	Algeria	45400000	Africa	3957.0	0	8.0	4.3
3	Andorra	82623	Europe	42863.0	0	32.2	3.2
4	Angola	33086278	Africa	2472.0	0	2.6	3.5
5	Antigua and Barbuda	100772	Americas	16423.0	0	20.0	2.5
6	Argentina	46044703	Americas	11702.0	3	13.7	5.5
7	Armenia	2981200	Asia	5980.0	0	43.3	2.3
8	Australia	26508995	Oceania	64107.0	27	42.0	5.3
9	Austria	9120091	Europe	54760.0	12	30.0	5.5

Rysunek 2: Przykładowa próba zbioru danych po przetwarzaniu wstępnym

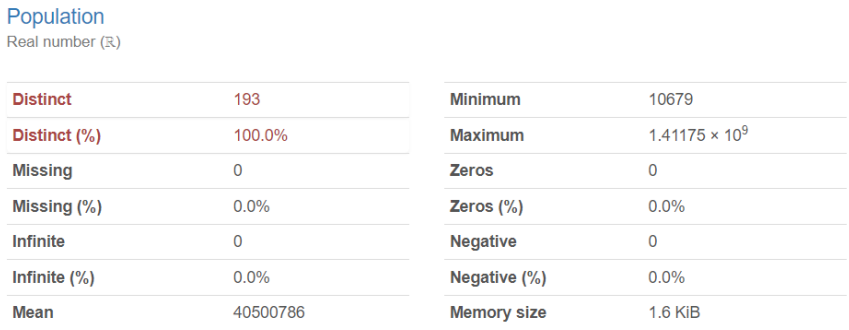
2.3.1 Analiza zmiennej Country

Statystyki zmiennej Country (zawierającej nazwy krajów po angielsku).

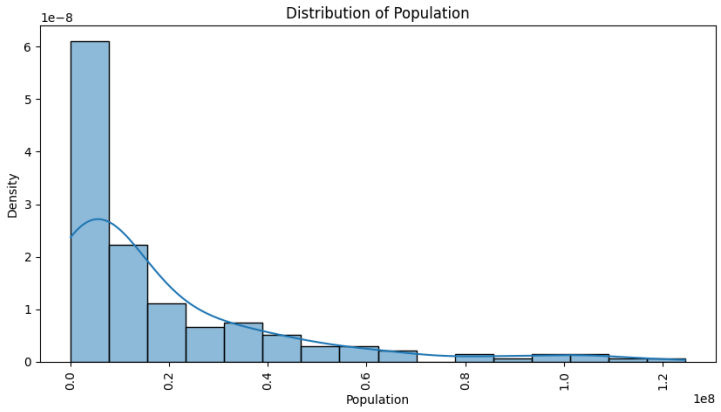
Rysunek 3: Statystyki zmiennej Country

2.3.2 Analiza zmiennej Population

Statystyki i rozkład danych zmiennej Population (zawierającej liczbę ludności państw).



Rysunek 4: Statystyki zmiennej Population



Rysunek 5: Rozkład danych zmiennej Population

2.3.3 Analiza zmiennej UN_Region

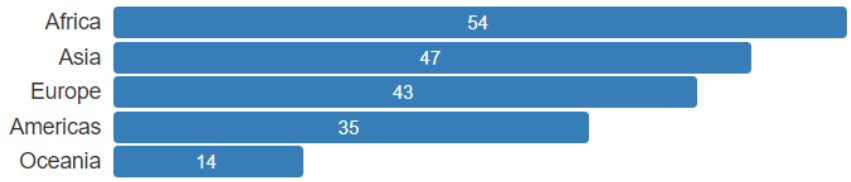
Statystyki zmiennej UN_Region (zawierającej nazwy regionów (kontynentów), na których znajdują się państwa).

UN_Region

Categorical

Distinct	5
Distinct (%)	2.6%
Missing	0
Missing (%)	0.0%
Memory size	1.6 KiB

Rysunek 6: Statystyki zmiennej UN_Region



Rysunek 7: Rozkład danych zmiennej UN_Region

2.3.4 Analiza zmiennej GDP_per_capita

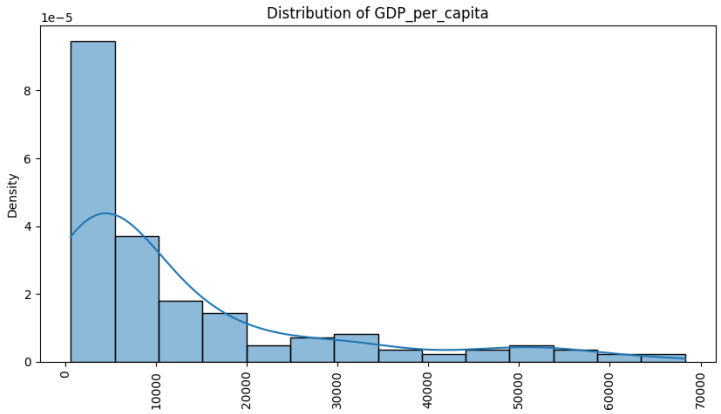
Statystyki i rozkład danych zmiennej GDP_per_capita (zawierającej wartości PKB per capita państw).

GDP_per_capita
Real number (ℝ)

HIGH CORRELATION UNIQUE

Distinct	193	**Minimum**	260
Distinct (%)	100.0%	**Maximum**	234316
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	17445.953	**Memory size**	1.6 KiB

Rysunek 8: Statystyki zmiennej GDP_per_capita



Rysunek 9: Rozkład danych zmiennej GDP_per_capita

2.3.5 Analiza zmiennej Top_1000_Uni_Count

Statystyki i rozkład danych zmiennej Top_1000_Uni_Count (zawierającej liczbę uniwersytetów będących wśród 1000 najlepszych na świecie dla każdego państwa).

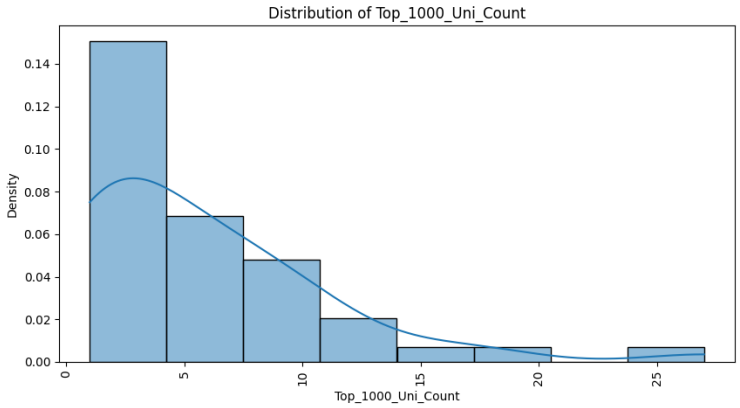
Top_1000_Uni_Count

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	27	**Minimum**	0
Distinct (%)	14.0%	**Maximum**	229
Missing	0	**Zeros**	138
Missing (%)	0.0%	**Zeros (%)**	71.5%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	5.0310881	**Memory size**	1.6 KiB

Rysunek 10: Statystyki zmiennej Top_1000_Uni_Count



Rysunek 11: Rozkład danych zmiennej Top_1000_Uni_Count

2.3.6 Analiza zmiennej Tertiary_edu_%

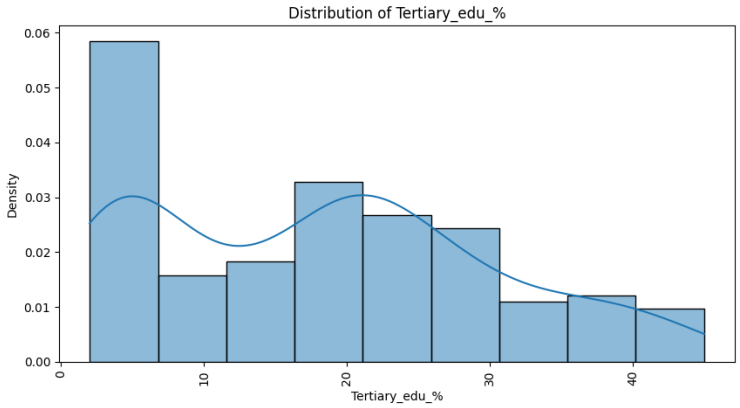
Statystyki i rozkład danych zmiennej Tertiary_edu_% (zawierającej procent populacji państw z wyższym wykształceniem).

Tertiary_edu_%

Real number (ℝ)

Distinct	109	**Minimum**	0.2
Distinct (%)	56.5%	**Maximum**	73.9
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	19.056995	**Memory size**	1.6 KiB

Rysunek 12: Statystyki zmiennej Tertiary_edu_%



Rysunek 13: Rozkład danych zmiennej Tertiary_edu_%

2.3.7 Analiza zmiennej Expenditure_on_education_(%_of_GDP)

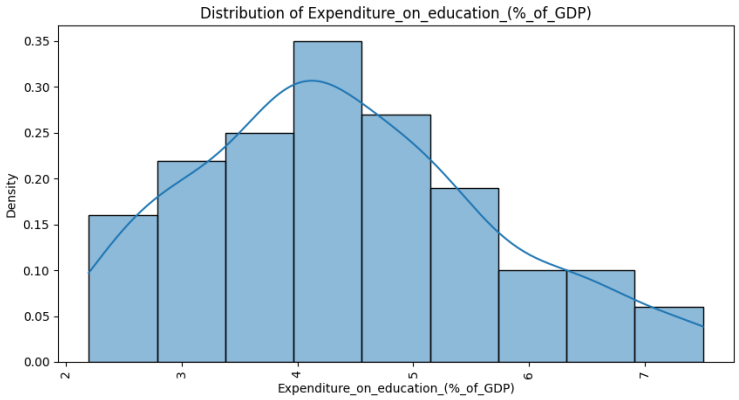
Statystyki i rozkład danych zmiennej Expenditure_on_education_(%_of_GDP) (zawierającej procent PKB państw wydawany na edukację).

Expenditure_on_education_(%_of_GDP)

Real number (ℝ)

Distinct	65	**Minimum**	0.6
Distinct (%)	33.7%	**Maximum**	15.8
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	4.5333679	**Memory size**	1.6 KiB

Rysunek 14: Statystyki zmiennej Expenditure_on_education_(%_of_GDP)

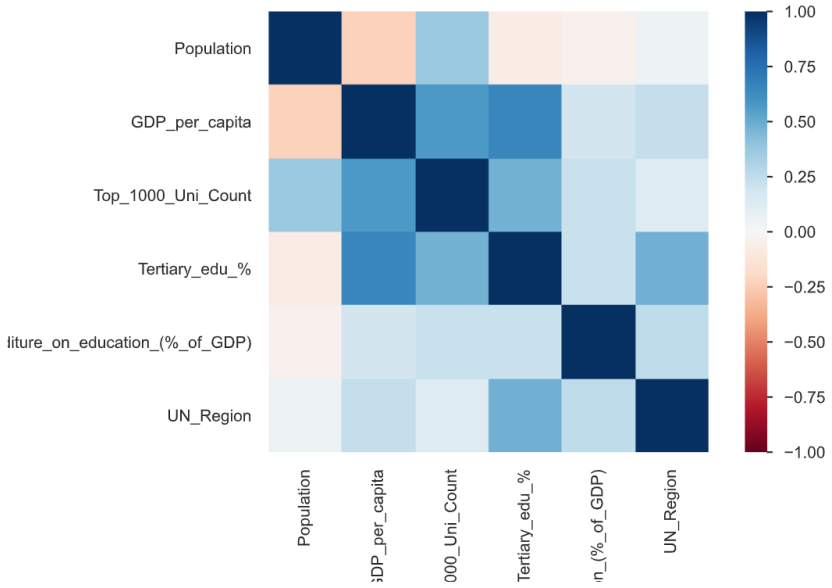


Rysunek 15: Rozkład danych zmiennej Expenditure_on_education_(%_of_GDP)

2.3.8 Zależności między danymi

Poniższa macierz korelacji (heatmap) pokazuje, jak silna jest korelacja między poszczególnymi zmiennymi. Po analizie macierzy możemy zauważyć, że zmienna GDP_per_capita jest w wysokiej korelacji ze zmiennymi Top_1000_Uni_Count i Tertiary_edu_%. Co ciekawe, jest w niewielkim stopniu powiązane z Expenditure_on_education_(%_of_GDP) oraz UN_Region i nie jest zupełnie powiązane z Population. Można więc wyciągnąć wstępne wnioski, iż największy wpływ na na PKB na osobę ma obecność wielu wysokiej jakości uczelni wyższych oraz aktualny poziom wykształcenia populacji. Na tym etapie jednak nie da się jednak jeszcze jednoznacznie określić tych zależności, a wnioski mogą być mylące ze względu na wielkość zbioru danych.

Można także przedstawić powyższą macierz w formie tabelarycznej.



Rysunek 16: Macierz korelacji między danymi (heatmapa)

	Population	GDP_per_capita	Top_1000_Uni_Count	Tertiary_edu_%	Expenditure_on_education_(%_of_GDP)
Population	1.000	-0.231	0.374	-0.075	-0.043
GDP_per_capita	-0.231	1.000	0.578	0.656	0.187
Top_1000_Uni_Count	0.374	0.578	1.000	0.476	0.223
Tertiary_edu_%	-0.075	0.656	0.476	1.000	0.226
Expenditure_on_education_(%_of_GDP)	-0.043	0.187	0.223	0.226	1.000
UN_Region	0.049	0.237	0.126	0.478	0.255

Rysunek 17: Macierz korelacji w formie tabelarycznej

3 Rozwiązanie problemu

Po wstępnej analizie danych wiadomo już, że GDP_per_capita jest w wysokiej korelacji ze zmiennymi Top_1000_Uni_Count i Tertiary_edu_%. Wykazuje natomiast niewielką lub żadną zależność od Expenditure_on_education_(%_of_GDP), UN_Region i Population. Przed wyborem i trenowaniem modeli należy podzielić dane na treningowe (train data) oraz testowe (test data). Podział został dokonany w proporcjach 3:1 (train:test), aby zminimalizować ryzyko overfittingu modelu do danych oraz zwiększyć generalizację modelu dla podobnych problemów. Wadą takiego podziału może być zwiększona niedokładność w estymacji wartości (czyli większy średni błąd). Do rozwiązania problemu wybrane zostały 3 modele, które służą do rozwiązywania problemów typu regresji liniowej:

- Model liniowy (Linear Regression)

- Maszyna wektorów losowych (Support Vector Machine, SVR)
- Uogólniony model liniowy (Generalized Linear Model, GLM)

Jako miarę jakości modeli przyjęto pierwiastek z błędu średniokwadratowego (w skrócie MSE). Wyniki przedstawione są w poniższej tabeli (wygenerowana za pomocą `https://www.tablesgenerator.com/`): Niestety wyniki nie są satys-

Model	**Pierwiastek błędu średniokwadratowego**
Linear Regression	1.82e+04
SVR	2.01e+04
GLM	1.69e+04

Tabela 1: Tabela jakości modeli, mierzonej za pomocą błędu średniokwadratowego

fakcjonujące, MSE wynosi od 17 do 20 ty, stąd decyzja o wytrenowaniu dwóch kolejnych modeli i porównaniu wyników z trzema modelami powyżej. Do rozwiązania dobrano zatem:

- Model Drzewa Decyzyjnego (Decision Tree)
- Model Lasów Losowych (Random Forest)

Wyniki modeli Decision Tree i Random Forest zostały przedstawione w poniższej tabeli:

Model	**Pierwiastek błędu średniokwadratowego**
Random Forest	1.13e+04
Decision Tree	1.48e+04

Tabela 2: Tabela jakości modeli, mierzonej za pomocą błędu średniokwadratowego

4 Wnioski

Miara jakości modeli waha się w przedziale od 11 do 20 tysięcy dolarów. Najlepszym modelem okazał się Random Forest, najgorszym SVR. Niestety nie są to wyniki akceptowalne i zadowalające, zważywszy na fakt, że średnia światowa PKB per capita wyliczona z analizowanego tutaj zbioru danych wynosi 17,44 tysiące dolarów. Dla wielu krajów, zwłaszcza z niskim wskaźnikiem PKB per capita, taki margines błędu oznacza brak możliwości wyznaczenia względnie dokładnej wartości. Przyczyn takich wyników można upatrywać w kilku źródłach. Jednym z głównych problemów jest wielkość zbioru danych. Zebrane zostały dane dla większości krajów świata, co po pre-processingu dało liczbę 193 wierszy z danymi. Jest to zdecydowanie za mało, aby móc skutecznie wytrenować model i predykować wartości PKB per capita. Rozwiązaniem mogłoby być

zebranie danych historycznych dla tych samych krajów, np. na przestrzeni 20 lat.

Jednakże napotykamy również tutaj na drugi problem. PKB per capita jest wartością, która zależy od bardzo wielu czynników, związanych z sytuacją ekonomiczną, geopolityczną czy kulturową państwa. Predykowanie tego wskaźnika jedynie na podstawie danych o wykształceniu obywateli, czy liczby uczelni wyższych o wysokim poziomie nauczania nie może dać oczekiwanego wyniku.

Dodatkowo, problemem podczas pozyskiwania i pracy na danych dotyczących państw na świecie jest często ich brak, niedokładność lub wątpliwe metody ich pozyskiwania.

Jak się okazało w trakcie badania, PKB per capita zależy między innymi od liczby światowej klasy uczelni wyższych w kraju, a także od procentu populacji z wyższym wykształceniem. Wskaźnik ten z kolei ma niewielki lub zerowy związek z parametrami jak populacja czy procent PKB wydawany na edukację. Ponownie, związki te mogą być niedokładne ze względu na niewielki zbiór danych.

Literatura

[1] Myles O'Neill. World University Rankings. https://www.kaggle.com/datasets/mylesoneill/world-university-rankings. Accessed on 30 May, 2023.

[2] The World Bank Group. Educational attainment, at least completed short-cycle tertiary, population 25+, total (%) (cumulative). https://data.worldbank.org/indicator/SE.TER.CUAT.ST.ZS?name_desc=true. Accessed on 30 May, 2023.

[3] Wikipedia contributors. List of countries and dependencies by population. https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population. Accessed on 30 May, 2023.

[4] Wikipedia contributors. List of countries by GDP (nominal) per capita. https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita. Accessed on 30 May, 2023.

[5] Wikipedia contributors. List of countries by spending on education (% of GDP). https://en.wikipedia.org/wiki/List_of_countries_by_spending_on_education_(%25_of_GDP). Accessed on 30 May, 2023.

[6] Wikipedia contributors. List of countries by tertiary education attainment. https://en.wikipedia.org/wiki/List_of_countries_by_tertiary_education_attainment. Accessed on 30 May, 2023.