Data Set Title
Exploratory Analysis

Name, student@bellarmine.edu
Name, student@bellarmine.edu

I. INTRODUCTION

Our data set covers a comprehensive list of countries worldwide. It covered a wide range of variables such as economic indications, healthcare measurements, education statistics, and demographics. We chose this data set as we are finance and economics majors and felt it tied in closely with our majors.
Dataset: https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023

II. DATA SET DESCRIPTION

The original data set contained 195 entries with 35 columns of various data types. Some examples of the type of data that was included in it was country name, land area, birth rate, GDP, currency code, national language, life expectancy and much more. **Table 1** displays this and shows how much of that data was missing.

Since our data is representative of the specific country based on numerous factors, we decided not to clean the data in pandas as it would mean using estimates and averages when that is highly unlikely to be representative of that country. Instead, we decided to research all the missing data using the World Bank Databank to fill in the missing values. As the dataset included all countries, certain countries were deleted based on the fact there was no/very little data ever reported. Columns that weren't going to be relevant to our analysis such as major city or capital city, forested area, calling code, and armed forces size were removed for the dataset as well. This left the cleaned data set with 184 entries with 20 columns. **Table 2** displays this and shows that none of the data is missing anymore. The rest of the work in this report will be represented by the cleaned data set.

Table 1 - Original Data (Uncleaned):

Index	Variable Name	Level of Measurement	Pandas Data Type	Missing Data (%)
0	Country	Nominal	object	0%
1	Density (P/Km2)	Interval	int64	100%
2	Abbreviation	Nominal	object	99.46%
3	Agricultural Land (%)	Ratio	float64	99.46%
4	Land Area(Km2)	Interval	int64	100%
5	Armed Forces size	Interval	int64	93.48%
6	Birth Rate	Ratio	float64	100%
7	Calling Code	Nominal	int64	100%
8	Capital/Major City	Nominal	object	98.91%
9	Co2-Emissions	Interval	int64	99.46%
10	CPI	Interval	int64	96.74%
11	CPI Change (%)	Ratio	float64	97.28%
12	Currency-Code	Nominal	object	97.82%
13	Fertility Rate	Ratio	float64	99.46%
14	Forested Area (%)	Ratio	float64	99.46%
15	Gasoline Price	Interval	float64	95.11%
16	GDP	Interval	int64	99.46%
17	Gross primary education enrollment (%)	Ratio	float64	99.46%

Index	Variable Name	Level of Measurement	Pandas Data Type	Missing Data (%)
18	Gross tertiary education enrollment (%)	Ratio	float64	98.91%
19	Infant mortality	Ratio	float64	100%
20	Largest city	Nominal	object	98.91%
21	Life expectancy	Interval	float64	98.37%
22	Maternal mortality ratio	Interval	float64	98.37%
23	Minimum wage	Interval	float64	81.52%
24	Official language	Nominal	object	100%
25	Out of pocket health expenditure	Ratio	float64	99.46%
26	Physicians per thousand	Ratio	float64	99.46%
27	Population	Interval	int64	99.46%
28	Population: Labor force participation (%)	Ratio	float64	95.65%
29	Tax revenue (%)	Interval	float64	91.85%
30	Total tax rate	Ratio	float64	96.74%
31	Unemployment rate	Ratio	float64	95.65%
32	Urban population	Interval	int64	97.28%
33	Latitude	Ratio	float64	99.46%
34	Longitude	Ratio	float64	99.46%

Table 2 - Cleaned Data:

Index	Variable Name	Level of Measurement	Pandas Data Type	Missing Data (%)
0	Country	Nominal	object	0%
1	Density (P/Km2)	Ratio	int64	0%
2	Agricultural_Land(%)	Ratio	float64	0%
3	Land_Area(Km2)	Interval	int64	0%
4	Birth_Rate	Ratio	float64	0%
5	Co2-Emissions	Interval	int64	0%
6	Currency-Code	Nominal	object	0%
7	Fertility_Rate	Ratio	float64	0%
8	Gasoline_Price	Ratio	float64	0%
9	GDP	Interval	int64	0%
10	Gross_primary_education_enrollment (%)	Ratio	float64	0%
11	Gross_tertiary_education_enrollment (%)	Ratio	float64	0%
12	Infant_mortality	Ratio	float64	0%
13	Life_expectancy	Interval	float64	0%
14	Minimum_wage	Ratio	float64	0%
15	Official_language	Nominal	object	0%
16	Physicians_per_thousand	Ratio	float64	0%
17	Population	Interval	int64	0%
18	Unemployment_rate	Ratio	float64	0%

Index	Variable Name	Level of Measurement	Pandas Data Type	Missing Data (%)
19	Urban_population	Interval	int64	0%

III. Data Set Summary Statistics

Moving in into the summary statistics from this dataset (cleaned data), we can see the count, mean, std, min, 25%, 50%,75%, max for all the continuous variables. Reference **Table 3** for the specific data.

Table 3: Summary Statistics for World Data 2023 (cleaned)

	Density\n(P/Km2)	Agricultural Land(_%)	Land Area (Km2)	Birth Rate	Co2-Emissions	Fertility Rate	Gasoline Price	GDP
count	184	184	1.84E+02	184	1.84E+02	184	184	1.84E+02
mean	207.1957	0.391196	7.26E+05	20.30277	1.82E+05	2.695326	1.02038	5.00E+11
std	662.8604	0.216786	1.97E+06	9.904643	8.48E+05	1.280172	0.388033	2.22E+12
min	2	0.01	1.81E+02	6.4	6.60E+01	0.98	0	1.95E+08
25%	32	0.2175	2.86E+04	11.375	2.92E+03	1.705	0.77	1.12E+10
50%	83	0.4	1.48E+05	18.01	1.28E+04	2.245	0.99	3.90E+10
75%	176	0.55	5.83E+05	28.6675	6.62E+04	3.575	1.2525	2.47E+11
max	8358	0.83	1.71E+07	46.08	9.89E+06	6.91	2.8	2.14E+13
	Gross tertiary education enrollment (%)	**Infant mortality**	**Life expectancy**	**Minimum wage**	**Physicians per thousand**	**Population**	**Unemployment rate**	**Urban population**
count	184	184	184	184	184	1.84E+02	184	1.84E+02
mean	0.386576	21.3788	72.25109	1.92875	1.801304	4.13E+07	0.071196	2.29E+07
std	0.295129	19.62584	7.511012	3.119375	1.626461	1.49E+08	0.053579	7.66E+07
min	0.01	1.4	52.8	0	0.01	1.82E+04	0	1.45E+04
25%	0.12	6.075	67.05	0.25	0.3325	2.46E+06	0.03	1.42E+06
50%	0.315	13.95	73.5	0.755	1.46	9.61E+06	0.055	4.72E+06
75%	0.6325	33.125	77.45	1.91	2.8975	3.05E+07	0.1	1.52E+07
max	1.37	84.5	84.2	17.45	8.42	1.40E+09	0.28	8.43E+08

Table 4: Proportions for Country

From table 4 we can see that each country is only represented once with proportion out of the whole dataset equal to 0.543478%.

Category: Country	Frequency	Proportion (%)
Afghanistan	1	0.543478
Panama	1	0.543478
New Zealand	1	0.543478
Nicaragua	1	0.543478
Niger	1	0.543478
Nigeria	1	0.543478
North Macedonia	1	0.543478
Norway	1	0.543478

Category: Country	Frequency	Proportion (%)
Oman	1	0.543478
Pakistan	1	0.543478
Palau	1	0.543478
Papua New Guinea	1	0.543478
Albania	1	0.543478
Paraguay	1	0.543478
Peru	1	0.543478
Philippines	1	0.543478

Poland	1	0.543478
Portugal	1	0.543478
Qatar	1	0.543478
Romania	1	0.543478
Russia	1	0.543478
Rwanda	1	0.543478
Netherlands	1	0.543478
Nepal	1	0.543478
Namibia	1	0.543478
Myanmar	1	0.543478
Libya	1	0.543478
Lithuania	1	0.543478
Luxembourg	1	0.543478
Madagascar	1	0.543478
Malawi	1	0.543478
Malaysia	1	0.543478
Maldives	1	0.543478
Mali	1	0.543478
Malta	1	0.543478
Marshall Islands	1	0.543478
Mauritania	1	0.543478
Mauritius	1	0.543478
Mexico	1	0.543478
Federated States of Micronesia	1	0.543478
Moldova	1	0.543478
Mongolia	1	0.543478
Montenegro	1	0.543478
Morocco	1	0.543478
Mozambique	1	0.543478
Saint Kitts and Nevis	1	0.543478
Saint Lucia	1	0.543478
Saint Vincent and the Grenadines	1	0.543478
Tanzania	1	0.543478
East Timor	1	0.543478
Togo	1	0.543478
Tonga	1	0.543478
Trinidad and Tobago	1	0.543478
Tunisia	1	0.543478
Turkey	1	0.543478
Turkmenistan	1	0.543478
Uganda	1	0.543478
Ukraine	1	0.543478
United Arab Emirates	1	0.543478
United Kingdom	1	0.543478
United States	1	0.543478
Uruguay	1	0.543478
Uzbekistan	1	0.543478
Vanuatu	1	0.543478
Venezuela	1	0.543478
Vietnam	1	0.543478
Yemen	1	0.543478
Zambia	1	0.543478
Thailand	1	0.543478
Tajikistan	1	0.543478
Samoa	1	0.543478
Syria	1	0.543478
Saudi Arabia	1	0.543478
Senegal	1	0.543478
Serbia	1	0.543478
Seychelles	1	0.543478
Sierra Leone	1	0.543478
Singapore	1	0.543478
Slovakia	1	0.543478
Slovenia	1	0.543478
Solomon Islands	1	0.543478
Somalia	1	0.543478
South Africa	1	0.543478
South Korea	1	0.543478
South Sudan	1	0.543478
Spain	1	0.543478
Sri Lanka	1	0.543478
Sudan	1	0.543478
Suriname	1	0.543478
Sweden	1	0.543478
Switzerland	1	0.543478
Liberia	1	0.543478
Lesotho	1	0.543478
Lebanon	1	0.543478
Brazil	1	0.543478
Bulgaria	1	0.543478
Burkina Faso	1	0.543478
Burundi	1	0.543478
Ivory Coast	1	0.543478
Cape Verde	1	0.543478
Cambodia	1	0.543478
Cameroon	1	0.543478
Canada	1	0.543478
Central African Republic	1	0.543478
Chad	1	0.543478
Chile	1	0.543478
China	1	0.543478
Colombia	1	0.543478
Comoros	1	0.543478
Republic of the Congo	1	0.543478
Costa Rica	1	0.543478
Croatia	1	0.543478
Cuba	1	0.543478
Cyprus	1	0.543478
Brunei	1	0.543478
Botswana	1	0.543478
Democratic Republic of the Congo	1	0.543478
Bosnia and Herzegovina	1	0.543478
Algeria	1	0.543478
Andorra	1	0.543478
Angola	1	0.543478
Antigua and Barbuda	1	0.543478
Argentina	1	0.543478
Armenia	1	0.543478
Australia	1	0.543478
Austria	1	0.543478
Azerbaijan	1	0.543478
The Bahamas	1	0.543478
Bahrain	1	0.543478
Bangladesh	1	0.543478

Barbados	1	0.543478
Belarus	1	0.543478
Belgium	1	0.543478
Belize	1	0.543478
Benin	1	0.543478
Bhutan	1	0.543478
Bolivia	1	0.543478
Czech Republic	1	0.543478
Denmark	1	0.543478
Latvia	1	0.543478
Guyana	1	0.543478
Honduras	1	0.543478
Hungary	1	0.543478
Iceland	1	0.543478
India	1	0.543478
Indonesia	1	0.543478
Iran	1	0.543478
Iraq	1	0.543478
Republic of Ireland	1	0.543478
Israel	1	0.543478
Italy	1	0.543478
Jamaica	1	0.543478
Japan	1	0.543478
Jordan	1	0.543478
Kazakhstan	1	0.543478
Kenya	1	0.543478
Kiribati	1	0.543478
Kuwait	1	0.543478
Kyrgyzstan	1	0.543478
Laos	1	0.543478
Haiti	1	0.543478
Guinea-Bissau	1	0.543478
Dominica	1	0.543478
Guinea	1	0.543478
Dominican Republic	1	0.543478
Ecuador	1	0.543478
Egypt	1	0.543478
El Salvador	1	0.543478
Equatorial Guinea	1	0.543478
Estonia	1	0.543478
Eswatini	1	0.543478
Ethiopia	1	0.543478
Fiji	1	0.543478
Finland	1	0.543478
France	1	0.543478
Gabon	1	0.543478
The Gambia	1	0.543478
Georgia	1	0.543478
Germany	1	0.543478
Ghana	1	0.543478
Greece	1	0.543478
Grenada	1	0.543478
Guatemala	1	0.543478
Zimbabwe	1	0.543478

Table 5: Proportions for Official Language

From table 5 we can see that there are 16 official languages that have a frequency greater than one and a proportion greater than 0.54% with English, French, Spanish, Arabic, and Portuguese being in the top 5. English has a frequency of 30 with a proportion of 16.3%, French has a frequency of 23 with a proportion of 12.5%, Spanish has a frequency of 19 with a proportion of 10.33%, Arabic has a frequency of 17 with a proportion of 9.24%, and Portuguese has a frequency of 7 with a proportion of 3.8%.

Official Language	Frequency	Proportion (%)
English	30	16.3
French	23	12.5
Spanish	19	10.33
Arabic	17	9.24
Portuguese	7	3.8
None	4	2.17
Swahili	4	2.17
Russian	4	2.17
German	3	1.63
Malay	2	1.09
Modern Standard Arabic	2	1.09
Swedish	2	1.09
Dutch	2	1.09
Persian	2	1.09
Greek	2	1.09
Romanian	2	1.09
Montenegrin language	1	0.54
Mongolian	1	0.54
Burmese	1	0.54
Polish	1	0.54
Nepali	1	0.54
Macedonian	1	0.54
Norwegian	1	0.54
Urdu	1	0.54
Marshallese	1	0.54
Maltese	1	0.54
Tok Pisin	1	0.54
Pashto	1	0.54
Samoan	1	0.54
Serbian	1	0.54
Slovak	1	0.54
Malaysian language	1	0.54
Slovene language	1	0.54
Afrikaans	1	0.54
Korean	1	0.54
Tamil	1	0.54
Thai	1	0.54
Tongan Language	1	0.54

Turkish	1	0.54
Turkmen	1	0.54
Ukrainian	1	0.54
Uzbek	1	0.54
Vietnamese	1	0.54
Divehi	1	0.54
Hebrew	1	0.54
Luxembourgish	1	0.54
Lithuanian	1	0.54
Catalan	1	0.54
Armenian	1	0.54
Azerbaijani language	1	0.54
Bengali	1	0.54
Dzongkha	1	0.54
Bosnian	1	0.54
Bulgarian	1	0.54
Kirundi	1	0.54
Khmer language	1	0.54
Standard Chinese	1	0.54
Croatian	1	0.54
Czech	1	0.54
Danish	1	0.54
Estonian	1	0.54
Amharic	1	0.54
Fiji Hindi	1	0.54
Georgian	1	0.54
Hungarian	1	0.54
Icelandic	1	0.54
Hindi	1	0.54
Indonesian	1	0.54
Irish	1	0.54
Albanian	1	0.54
Italian	1	0.54
Jamaican English	1	0.54
Lao	1	0.54
Latvian	1	0.54
Shona	1	0.54

Table 6: Proportions for Currency Code

From Table 6, we can see that there are 6 currencies with a frequency greater than one and proportion greater than 0.543478%. The top 5 are the Euro (EUR), West African CFA franc (XOF), United States Dollar (USD), East Caribbean Dollar (XCD), and Central African CFA franc (XAF). EUR has a frequency of 21 with a proportion of 11.413%, XOF has a frequency of 8 with a proportion of 4.34783%, USD has a frequency of 6 with a proportion of 3.26087%, XCD has a frequency of 6 with a proportion of 3.26087%, and XAF has a frequency of 6 with a proportion of 3.26087%.

Category: Currency-Code	Frequency	Proportion (%)
EUR	21	11.413
XOF	8	4.34783
USD	6	3.26087
XCD	6	3.26087
XAF	6	3.26087
AUD	2	1.08696
PLN	1	0.543478
PAB	1	0.543478
PGK	1	0.543478
PYG	1	0.543478
PEN	1	0.543478
PHP	1	0.543478
RON	1	0.543478
QAR	1	0.543478
OMR	1	0.543478
RUB	1	0.543478
RWF	1	0.543478
WST	1	0.543478
SAR	1	0.543478
RSD	1	0.543478
PKR	1	0.543478
AFN	1	0.543478
NOK	1	0.543478
MAD	1	0.543478
MVR	1	0.543478
MRU	1	0.543478
MUR	1	0.543478
MXN	1	0.543478
MDL	1	0.543478
MNT	1	0.543478
MZN	1	0.543478
MKD	1	0.543478
MMK	1	0.543478
NAD	1	0.543478
NPR	1	0.543478
NZD	1	0.543478
NIO	1	0.543478
SCR	1	0.543478
NGN	1	0.543478
SBD	1	0.543478
SLL	1	0.543478
TTD	1	0.543478
ZMW	1	0.543478
YER	1	0.543478
VND	1	0.543478
VED	1	0.543478
VUV	1	0.543478
UZS	1	0.543478
UYU	1	0.543478
GBP	1	0.543478
AED	1	0.543478
UAH	1	0.543478
UGX	1	0.543478
TMT	1	0.543478
TRY	1	0.543478

TND	1	0.543478
TOP	1	0.543478
SGD	1	0.543478
THB	1	0.543478
TZS	1	0.543478
TJS	1	0.543478
SYP	1	0.543478
CHF	1	0.543478
SEK	1	0.543478
SRD	1	0.543478
SDG	1	0.543478
LKR	1	0.543478
SSP	1	0.543478
KRW	1	0.543478
ZAR	1	0.543478
SOS	1	0.543478
MWK	1	0.543478
MYR	1	0.543478
LSL	1	0.543478
MGA	1	0.543478
BRL	1	0.543478
CZK	1	0.543478
CUP	1	0.543478
HRK	1	0.543478
CRC	1	0.543478
KMF	1	0.543478
COP	1	0.543478
CNY	1	0.543478
CLP	1	0.543478
CAD	1	0.543478
KHR	1	0.543478
CVE	1	0.543478
BIF	1	0.543478
BGN	1	0.543478
BND	1	0.543478
BWP	1	0.543478
DKK	1	0.543478
BAM	1	0.543478
BOB	1	0.543478
BTN	1	0.543478
BZD	1	0.543478
BYN	1	0.543478
BBD	1	0.543478

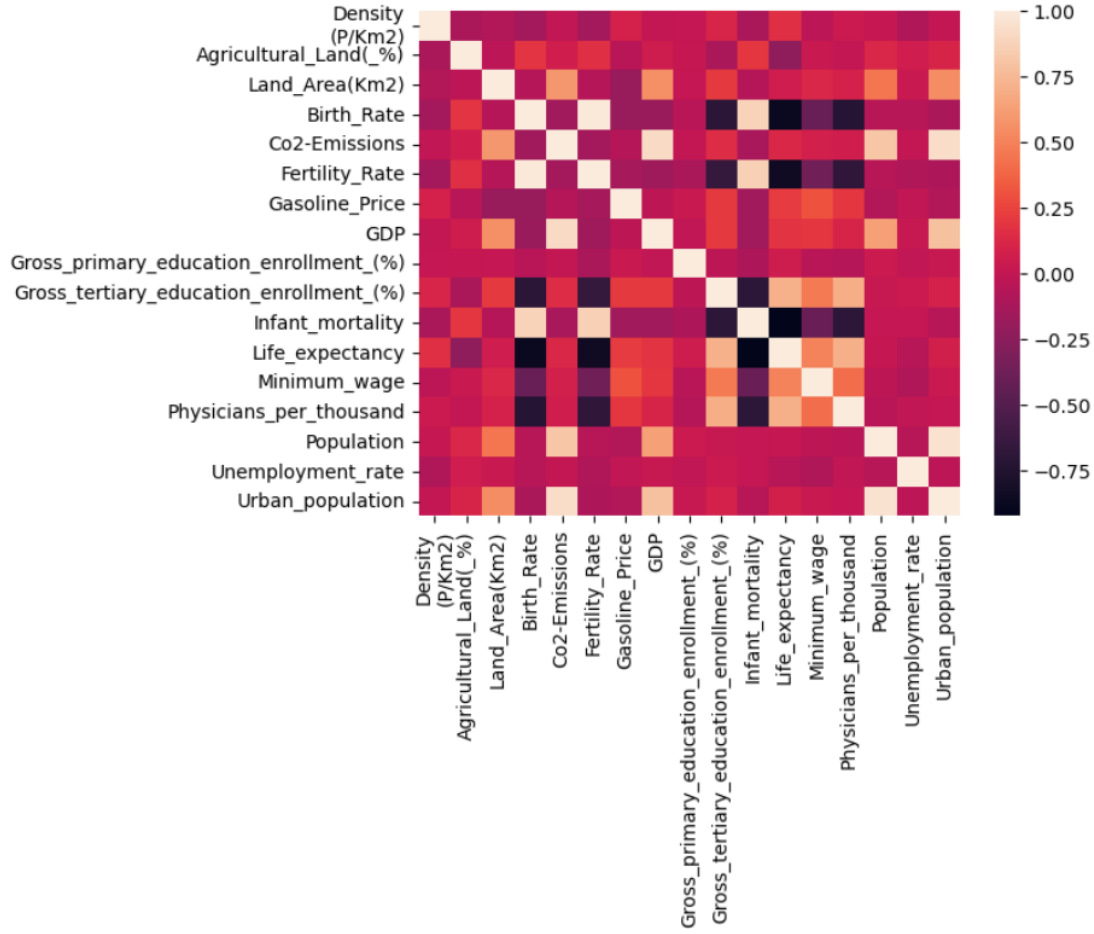
BDT	1	0.543478
BHD	1	0.543478
BSD	1	0.543478
AZN	1	0.543478
AMD	1	0.543478
ARS	1	0.543478
AOA	1	0.543478
DZD	1	0.543478
CDF	1	0.543478
DOP	1	0.543478
LYD	1	0.543478
IDR	1	0.543478
LRD	1	0.543478
ALL	1	0.543478
LBP	1	0.543478
LAK	1	0.543478
KGS	1	0.543478
KWD	1	0.543478
KES	1	0.543478
KZT	1	0.543478
JOD	1	0.543478
JPY	1	0.543478
JMD	1	0.543478
ILS	1	0.543478
IQD	1	0.543478
IRR	1	0.543478
INR	1	0.543478
EGP	1	0.543478
ISK	1	0.543478
HUF	1	0.543478
HNL	1	0.543478
HTG	1	0.543478
GYD	1	0.543478
GNF	1	0.543478
GTQ	1	0.543478
GHS	1	0.543478
GEL	1	0.543478
GMD	1	0.543478
FJD	1	0.543478
ETB	1	0.543478
SZL	1	0.543478
SVC	1	0.543478
ZWL	1	0.543478

Table 7: Correlation Table/Heatmap

Table 7 displays the correlation table for all the continuous variables within the cleaned dataset. With some basic observations we can see that there is a pretty strong negative correlation between life expectancy and birth rate/infant mortality rate/fertility rate. There is also pretty strong positive correlation urban population and population, life expectancy and physicians per thousand, Co2 emissions and GDP, and Infant mortality rate and fertility rate

The heatmap doesn't include any numbers as due to the number of variables, it makes it impossible to read.

	Density (P/Km2)	*ultural Land*	*Area(Km2*	*med Forces*	*Birth Rate*	*2-Emissior*	*ertility Rate*	*ested Area*	*asoline Pric*	*GDP*	*education*	*education*	*ant mortalit*	*e expectan*	*nimum wage*	*t health e*	*ians per tho*	*Population*	*x revenue (*	*mployment*	*an_popula*
Density (P/Km2)	1																				
Agricultura	-0.10753	1																			
Land Area(	-0.08207	-0.0322	1																		
Armed For	-0.00018	0.058264	0.585012	1																	
Birth Rate	-0.14236	0.181182	-0.07102	-0.13055	1																
Co2-Emiss	-0.01399	0.061849	0.590239	0.773325	-0.15557	1															
Fertility Ra	-0.14559	0.161018	-0.06504	-0.1363	0.980422	-0.14103	1														
Forested A	-0.10061	-0.43774	-0.01814	-0.0558	-0.09306	-0.02935	-0.08172	1													
Gasoline P	0.088394	-0.04841	-0.19012	-0.16303	-0.19094	-0.07464	-0.12901	0.244445	1												
GDP	-0.00781	0.051257	0.54916	0.635032	-0.1865	0.916778	-0.16179	-0.00226	-0.0288	1											
Gross prim	0.005736	0.014557	-0.00019	0.033316	-0.04576	-0.00317	-0.1129	0.127051	0.027891	-0.00887	1										
Gross terti	0.105659	-0.10974	0.207925	0.114879	-0.70176	0.148617	-0.65703	-0.00068	0.203527	0.204344	-0.02813	1									
Infant mort	-0.11812	0.193293	-0.06806	-0.06758	0.872796	-0.12265	0.858649	-0.06325	-0.15531	-0.155	-0.10031	-0.69724	1								
Life expect	0.163694	-0.23229	0.056932	0.080864	-0.87472	0.120268	-0.84794	0.020092	0.214437	0.17747	0.052135	0.71382	-0.92572	1							
Minimum v	-0.0379	0.023745	0.119513	-0.01357	-0.41252	0.078043	-0.36234	-0.06687	0.300535	0.190874	-0.05055	0.459275	-0.40572	0.503468	1						
Out of pocl	0.018253	0.133694	-0.02325	0.151721	0.250968	-0.03608	0.210956	-0.25397	-0.2889	-0.10767	-0.19091	-0.20374	0.351283	-0.31384	-0.31769	1					
Physicians	0.030065	-0.00184	0.082351	0.008076	-0.73662	0.061388	-0.68082	-0.06633	0.190039	0.10142	-0.06686	0.69891	-0.69756	0.704734	0.412821	-0.19496	1				
Population	0.007952	0.119089	0.444215	0.911698	-0.05642	0.80978	-0.05533	-0.05865	-0.07863	0.631404	0.034169	0.018837	0.00426	0.010149	-0.02796	0.126026	-0.04624	1			
Tax revenu	-0.03417	-0.01537	-0.16123	-0.19523	-0.35997	-0.13626	-0.35619	0.115605	0.434742	-0.10808	0.18333	0.270012	-0.3661	0.329821	0.263297	-0.35519	0.326033	-0.17994	1		
Unemployr	-0.09261	0.059248	0.028315	-0.02981	-0.05443	-0.00632	-0.09011	-0.05871	-0.01011	0.017709	-0.01075	0.032399	0.001713	-0.05581	-0.0954	0.01841	-0.01308	-0.05734	0.213228	1	
Urban_pop	-0.00507	0.103329	0.545584	0.886352	-0.11174	0.926249	-0.10423	-0.03773	-0.08192	0.78396	0.021968	0.089607	-0.05839	0.072117	0.024159	0.058269	0.006344	0.954114	-0.17484	-0.03182	1



IV. DATA SET GRAPHICAL EXPLORATION

When beginning to explore this data set, we started off by looking at more economical factors such as GDP, gasoline price per liter, and minimum wage. We checked if there was any correlation between these variables and the distributions of them. After looking at some of the economical factors, we noticed that several variables are tied to general health or well-being of the population, so we transitioned into that. We looked through a variety of variables such as life expectancy, gross tertiary education enrollment (%), birth rate, and physicians per thousand. Finally, we wanted to do some kind of analysis with countries that were on the EUR as they made up 21 countries out of the whole entire dataset. This was focused on the distribution of the data within in euro and comparing them alongside histograms of the entire world. (Details and analysis of graphs is included in the summary)

A. Distributions

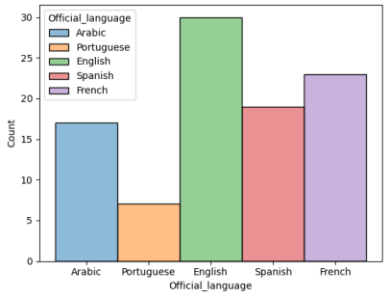


Figure 1: Count of top 5 languages (histogram) top 5 official languages

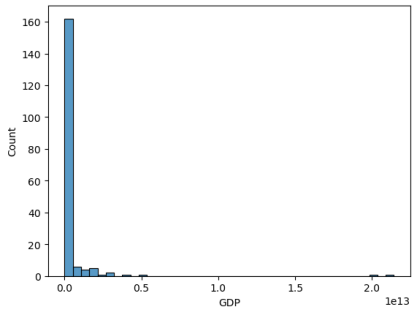


Figure 2a: Distribution of every countries GDP in USD($) (histogram)

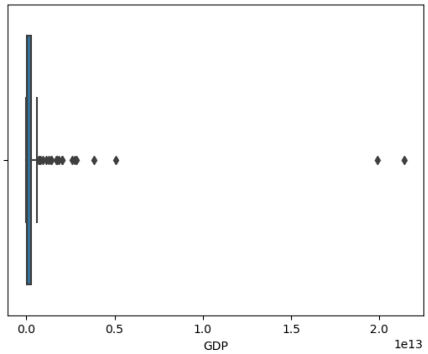


Figure 2b: Distribution of every countries GDP in USD($) (box plot) can see the outliers within the GDP range

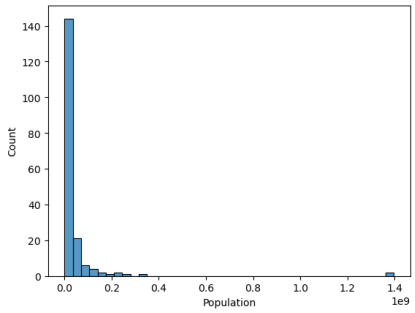


Figure 3a: Distribution of every countries Population (histogram)

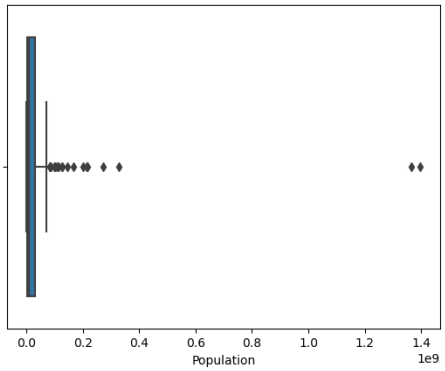


Figure 3b: Distribution of every countries Population (boxplot) can see the outliers within the population of dataset

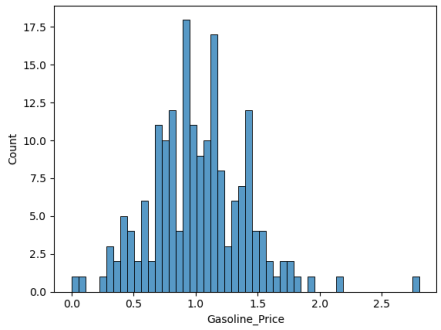


Figure 4: Distribution of every countries Gasoline price per liter in USD (histogram) pretty normally distributed

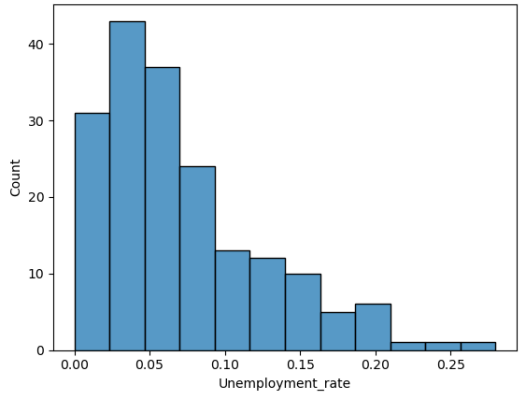
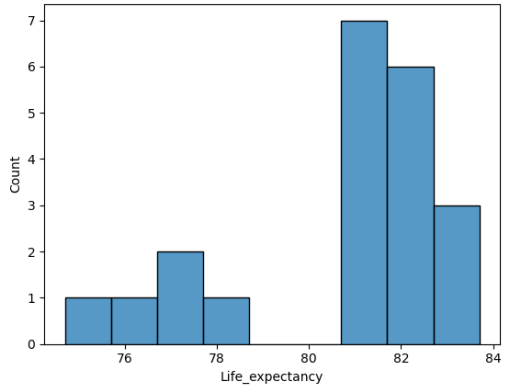
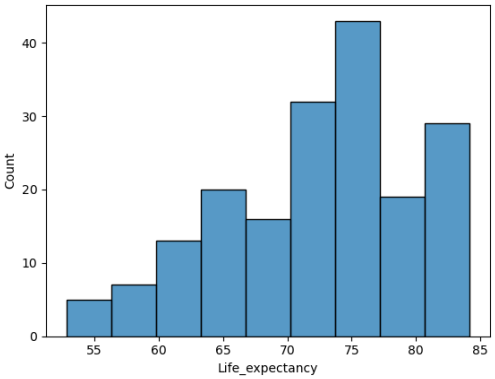


Figure 5a: Distribution of every countries unemployment rate (histogram)

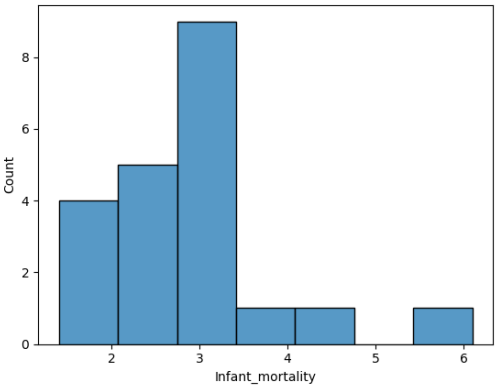


a)

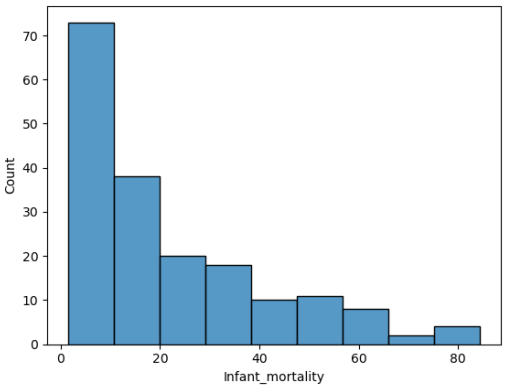


b)

Figure 6: a) Distribution of countries on the euro life expectancy alongside b) Distribution of every countries life expectancy (histogram) life expectancy also significantly higher

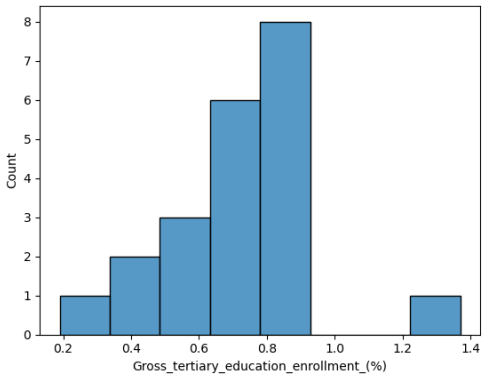


a)

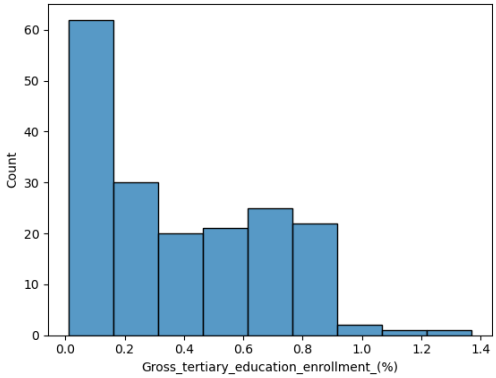


b)

Figure 7: a) Distribution of countries on the euro infant mortality rate alongside b) Distribution of every countries infant mortality rate (histogram) infant mortality rate significantly lower with countries on euro.

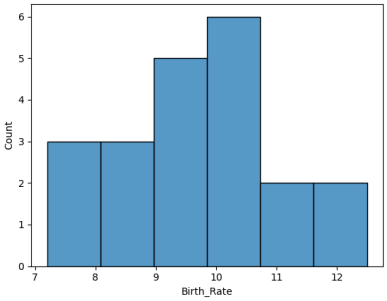


a)

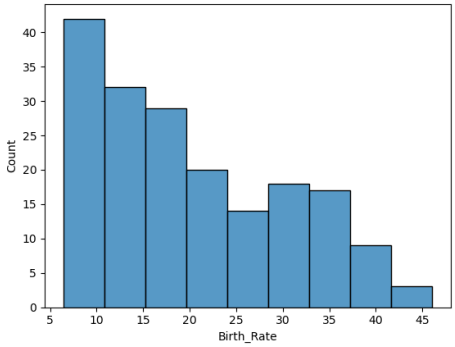


b)

Figure 8: a) Distribution of countries on the euro gross tertiary education enrollment(%) alongside b) Distribution of every countries gross tertiary education enrollment(%) (histogram)



a)



b)

Figure 9: a) Distribution of countries on the euro birth rate alongside b) Distribution of every countries birth rate (histogram) birth rate lower with countries on euro than globally.

B. ScatterPlots / (continuous variables)

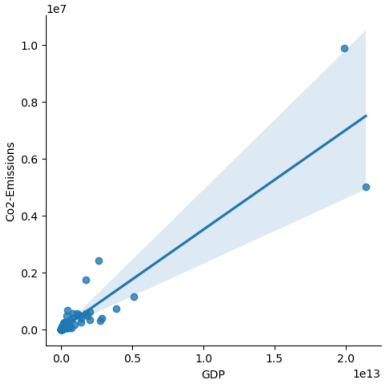


Figure 10: Comparison of GDP/C02-Emissions from the dataset (scatter plot with trend line) strong linear correlation

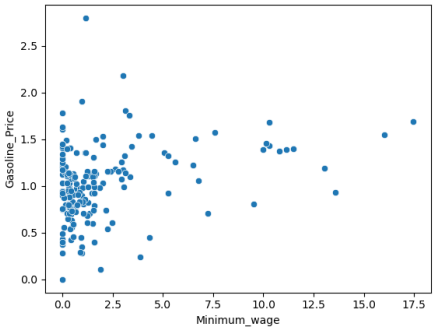


Figure 11: Comparison of Minimum wage/Gasoline Price from the dataset (scatter plot)

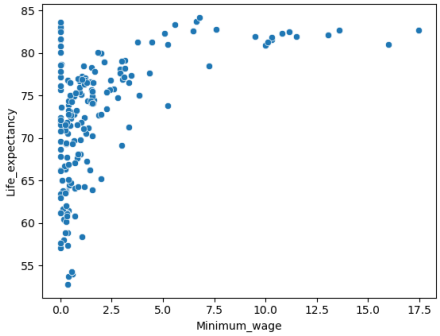


Figure 12: Comparison of Minimum wage/Life expectancy from the dataset (scatter plot) Can see the outliers within Europe as many countries don't have a mandated hourly minimum wage but high life expectancy

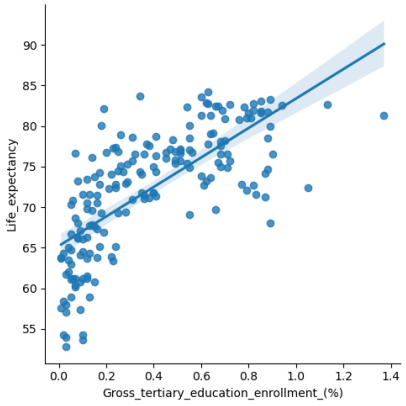


Figure 13: Comparison of gross tertiary education enrollment(%) and life expectancy from the dataset (scatter plot with trend line)

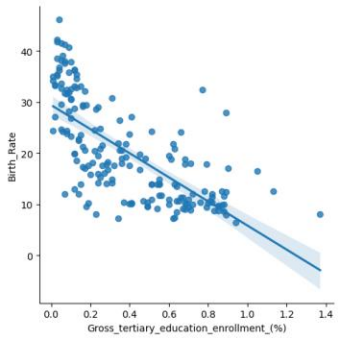


Figure 14: Comparison of gross tertiary education enrollment(%) and birth rate from the dataset (scatter plot with trend line)

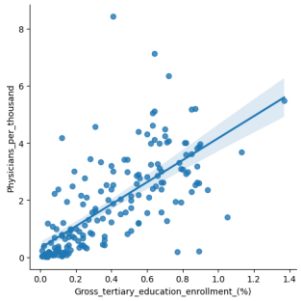


Figure 15: Comparison of gross tertiary education enrollment(%) and physicians per thousand from the dataset (scatter plot with trend line) Strong linear correlation

C. Barcharts (categorical variables)

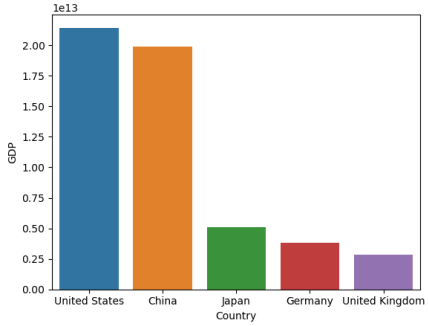


Figure 16: Top 5 countries GDP in USD($) (bar plot) USA and China's GDP significantly larger

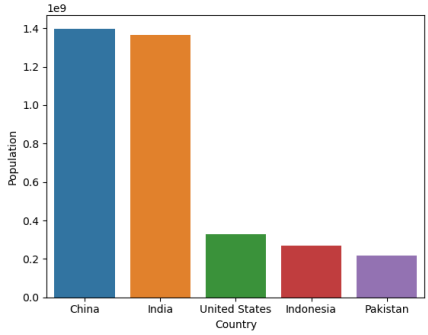


Figure 17: Top 5 countries in population (bar plot) China's and India's population significantly larger than rest of world

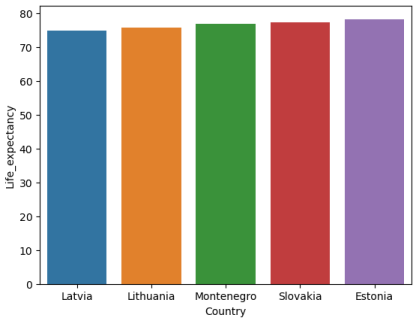


Figure 18: Bottom 5 countries on euro by life expectancy to show that where the distribution changes on the histogram(bar plot)

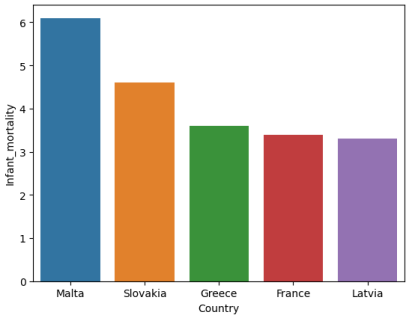


Figure 19: Top 5 countries on euro by infant mortality to show that Malta is the outlier on the histogram(bar plot)

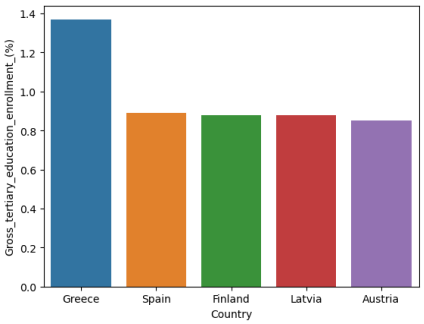


Figure 20: Top 5 countries on euro by gross tertiary education enrollment(%) to show that Greece is the outlier on the histogram(bar plot)

D. Other Plots - don't skimp – there are likely other plots that would be useful that I haven't already specified. Include those in this section.

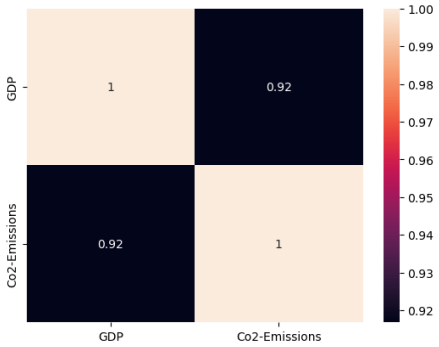


Figure 21: Correlation between GDP and Co2-Emissions (Heat map) really strong correlation between GDP and Co2-Emissions so as GDP goes up so does Co2-Emissions.

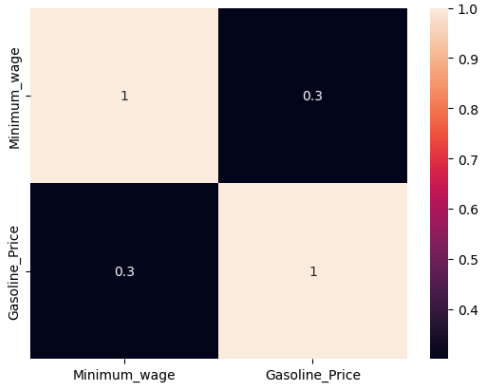


Figure 22: Correlation between GDP and Co2-Emissions (Heat map)

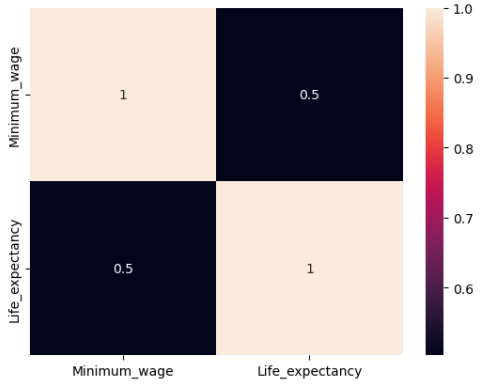


Figure 23: Correlation between Minimum Wage and Life expectancy (Heat map)

Figure 24: Correlation between GDP and Unemployment rate(Heat map) No correlation between GDP and Unemployment rate.

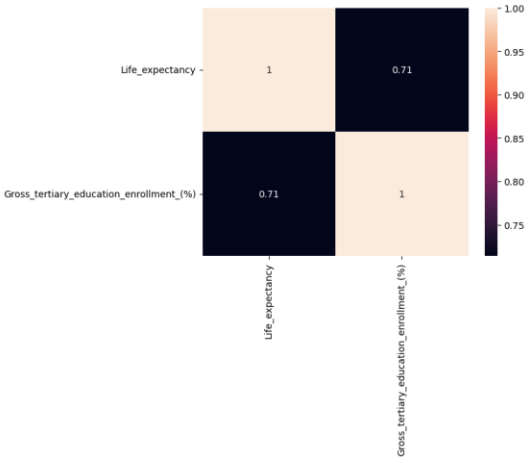


Figure 25: Correlation of gross tertiary education enrollment(%) and life expectancy from the dataset (heatmap) Strong positive correlation between the two variables

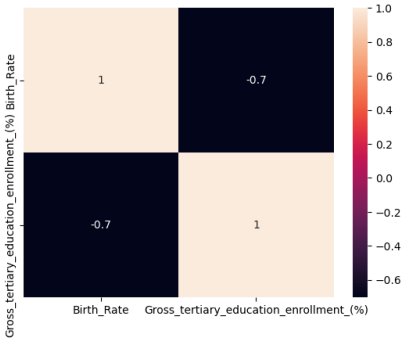


Figure 26: Correlation of gross tertiary education enrollment(%) and birth rate from the dataset (heatmap) Strong negative correlation between the two variables

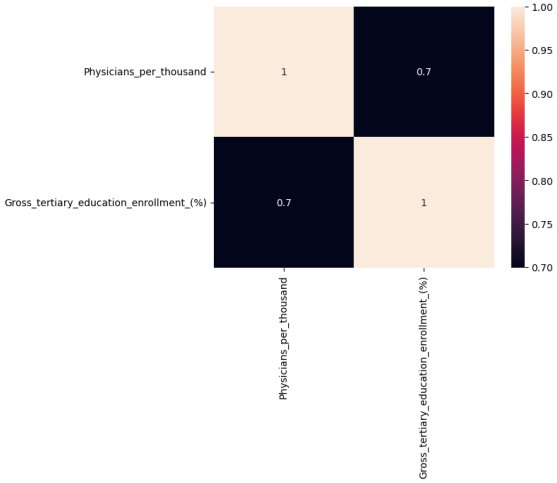


Figure 27: Correlation of physicians per thousand and gross tertiary education enrollment(%) from the dataset (heatmap) Strong positive correlation between the two variables

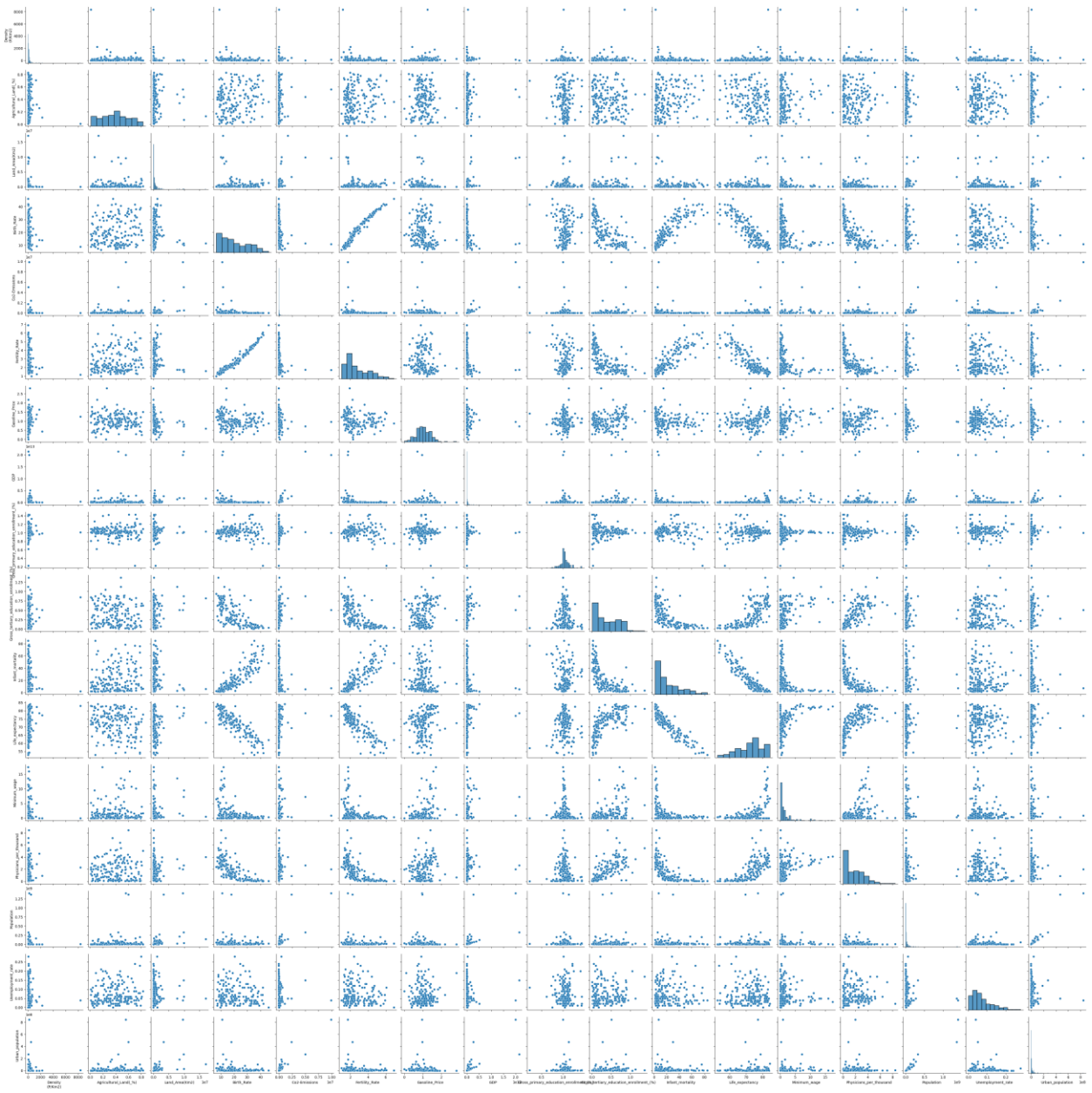


Figure 28: All continuous variables plotted against each other (Pair plot)

V. SUMMARY OF FINDINGS

Within our distribution graphs, we started by looking at the distribution of every countries GDP in USD($) as seen in **Figure 2a** and **2b**, with this analysis we noticed that there was a right skew in the data and several outliers. These outliers subsequently affect the mean of the dataset for GDP pulling it towards a larger number. We wanted to see what the top 5 outliers were so we plotted them on a bar plot as seen in **Figure 17**, from this graph we can see the top 5 which are USA, China, Japan, Germany, and United Kingdom. The US's and China's GDP were significantly greater than Japan or Germany's, so they would have the most effect on the mean.

We followed the same pattern with population as seen in **Figure 3a** and **3b** and found that it was similar to that of the histograms and boxplots of GDP. The mean would also be largely affected by the outliers present. With a box plot, as seen in **Figure 18**, we found that the top 5 outliers for population were China, India, USA, Indonesia, and Pakistan with China and India being significantly greater.

After the first few graphs, we transitioned into looking at the economic data within our dataset. We started with looking at the distribution of gasoline prices and saw it looked somewhat normally distributed as seen in **Figure 4**. With this graph we wanted to know if minimum wage affected the gasoline price per liter within the country. We started with a scatter plot (**Figure 11)** and noticed there might be some weak correlation between the two variables. This was confirmed with our heat map of the two variables (**Figure 22)**. Another question that we had was would GDP influence the amount of Co2- Emissions and we found that there was a strong positive correlation of 0.92 between the two variables (**Figure 10** and **21).** Another big economic question we had is whether there is any correlation between the unemployment rate and a country's GDP. We expected that there would be, however we found that there is no correlation between the two variables with a number of 0.018 (**Figure 24)**.

From here, we transitioned into looking at variables surrounding health and general wellbeing. When looking at this data one thing we noticed was the big impact of gross tertiary education enrollment (%) on life expectancy, birth rate, and physicians per thousand. The correlation between gross tertiary education enrollment (%) and life expectancy and physicians per thousand was a moderately strong positive correlation of .7 for both. While gross tertiary education enrollment (%) and birth rate had a moderately strong negative correlation of -.7. This data can be seen in **Figure 13, 14, 15, 25, 26, 27** displayed as heat maps, scatter plots, and scatter plots with trendlines.

The last set of data we looked at were countries that were on the euro and comparing a number of variables through histograms with the global dataset. The first variable we looked at was life expectancy with our assumption that life expectancy will be with countries on the euro than globally. We found this to be true as seen in **Figure 6a** and **6b**. We then went to compare infant mortality rate, with the assumption that it will be lower with countries on the euro than globally. This was also true as seen in **Figure 7a** and **7b.** The data had a range of 1-6 while globally it was between 0-85. We also found that tertiary education enrollment was higher in Europe than globally, which makes sense as many of Europe's countries offer cheap higher education as seen in **Figure 8a** and **8b**. Another one that was significant was the birth rate, we assumed that it would be significantly lower in Europe as their populations are getting older. We found that this was also true with mean of Europe's being 9.6 while globally it was 20.3. This is represented in **Figure 9a** and **9b**.