DS-160 TEAM PROJECT 1:

2023 WORLD DATA ANALYSIS

Robert Pearson & Clay Jackson
3/21/2024

INTRODUCTION

- Our Dataset: *Global Country Information Dataset 2023*
- Project Objectives:
  - Clean the data and create visualizations using the tools we have learned in class (Pandas, Matplotlib, Seaborn, and Tableau).
  - Thoroughly explore, analyze, and summarize the data
  - Find correlation between variables and global trends
  - Clearly present our project's findings to the class

DATA SOURCE

- Data Source: Kaggle
- Our Dataset: Global Country Information Dataset 2023
- Dataset Structure:
  - 35 variables
  - Continuous and discrete variables
  - Qualitative and quantitative variables
  - All countries worldwide
    - Demographic stats, economic indicators, healthcare metrics, education statistics, environmental factors, etc.

DATA CLEANING AND PREPARATION

- Missing Values: Rather than using mean or median to fill empty values, we pulled data from the world bank's collection
- Unwanted Columns: After reviewing the data, we removed 10 columns that were unrelated toward our planned analysis
- Utilized .drop()

```
countries.drop(['Abbreviation', 'Armed_Forces_size', 'Calling_Code', 'Capital/Major_City', 'Largest_city', 'Out_of_pocke
        axis=1,
        inplace=True)
```

FAMILIARIZING OURSELVES WITH THE DATA

- .head()
- .tail()
- .info()
- .describe()

```
1 countries=pd.read_csv("world-data-2023.csv")
2 countries.head()
```

]:

	Country	Density\n(P/Km2)	Abbreviation	Agricultural Land(%)	Land Area(Km2)	Armed Forces size	Birth Rate	Calling Code	Capital/Major City	Co2-Emissions	...	Minimum wage	Official language	Ou...
0	Afghanistan	60	AD	0.58	652230	323000	32.49	93	Kabul	8672	...	0.43	Pashto	
1	Albania	105	AE	0.43	28748	9000	11.78	355	Tirana	4536	...	1.12	Albanian	
2	Algeria	18	AF	0.17	2381741	317000	24.28	213	Algiers	150006	...	0.95	Arabic	
3	Andorra	164	AG	0.40	468	0	7.20	376	Andorra la Vella	469	...	6.63	Catalan	
4	Angola	26	AL	0.48	1246700	117000	40.73	244	Luanda	34693	...	0.71	Portuguese	

5 rows × 30 columns

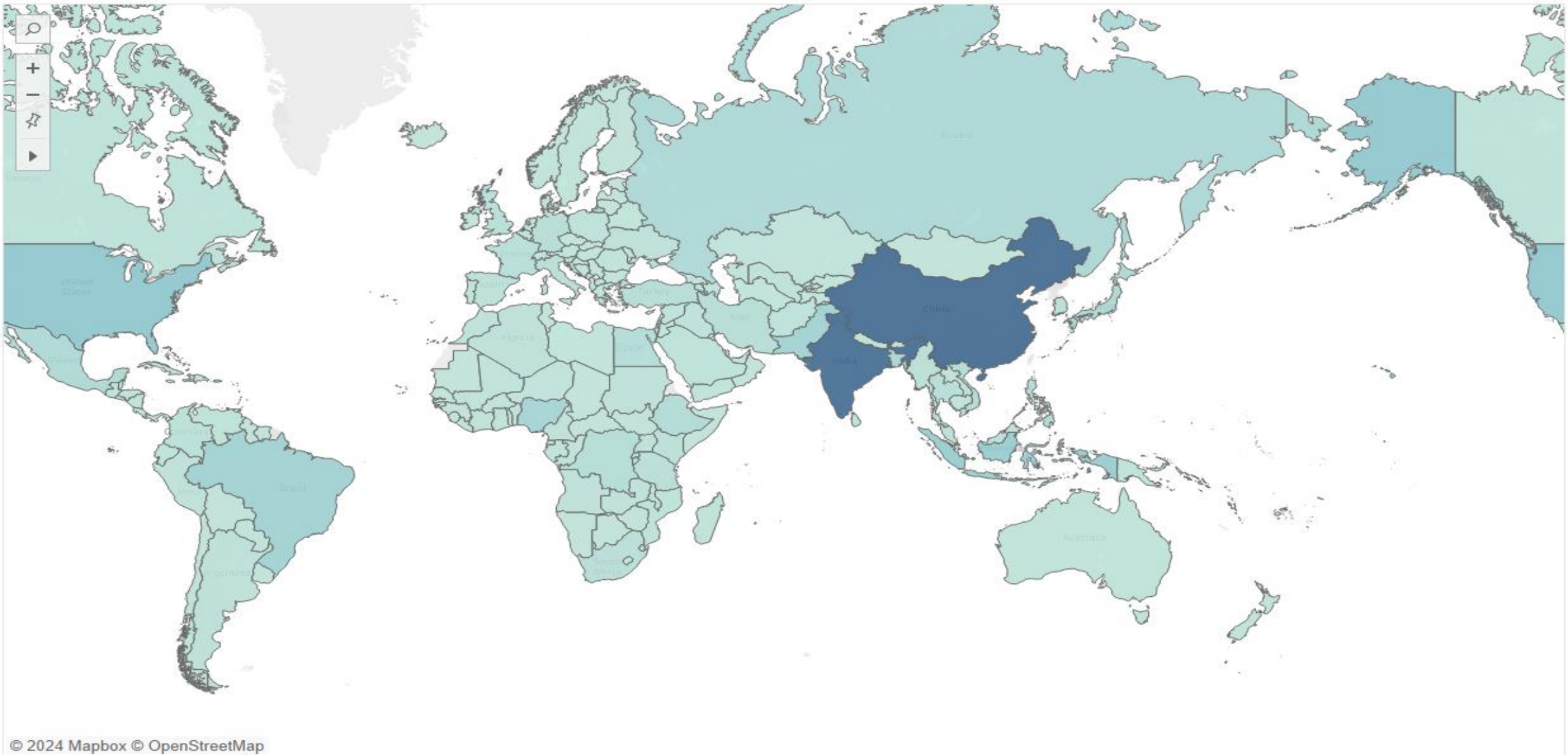
```
1 countries.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 184 entries, 0 to 183
Data columns (total 20 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   Country                                       184 non-null    object
 1   Density
(P/Km2)                                  184 non-null    int64
 2   Agricultural_Land(_%)                         184 non-null    float64
 3   Land_Area(Km2)                                184 non-null    int64
 4   Birth_Rate                                    184 non-null    float64
 5   Co2-Emissions                                 184 non-null    int64
 6   Currency-Code                                 184 non-null    object
 7   Fertility_Rate                                184 non-null    float64
 8   Gasoline_Price                                184 non-null    float64
 9   GDP                                           184 non-null    int64
 10  Gross_primary_education_enrollment_(%)        184 non-null    float64
```

```
1 countries.describe()
```

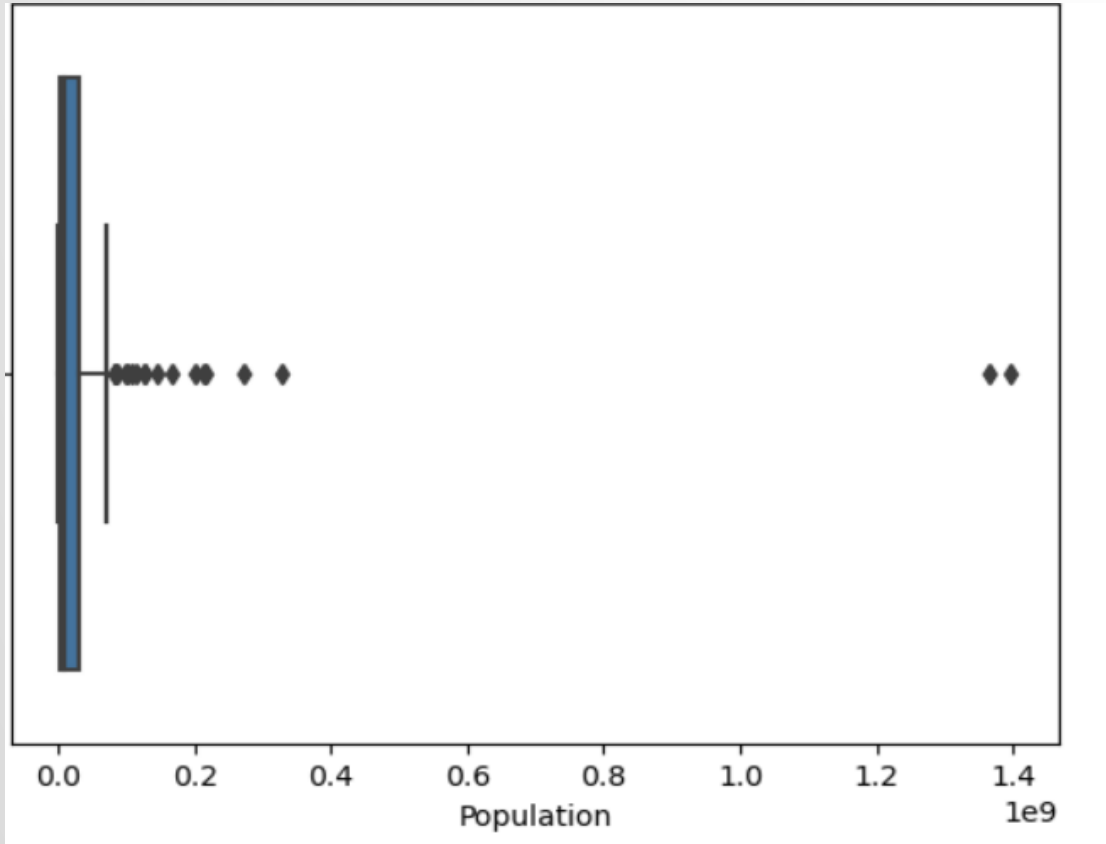
	Density\n(P/Km2)	Agricultural_Land(_%)	Land_Area(Km2)	Birth_Rate	Co2-Emissions	Fertility_Rate	Gasoline_Price	GDP	Gross_primary_edu
count	184.000000	184.000000	1.840000e+02	184.000000	1.840000e+02	184.000000	184.000000	1.840000e+02	
mean	207.195652	0.391196	7.256769e+05	20.302772	1.815457e+05	2.695326	1.020380	5.003513e+11	
std	662.860431	0.216786	1.966949e+06	9.904643	8.475146e+05	1.280172	0.388033	2.222368e+12	

GENERAL ANALYSIS: POPULATION

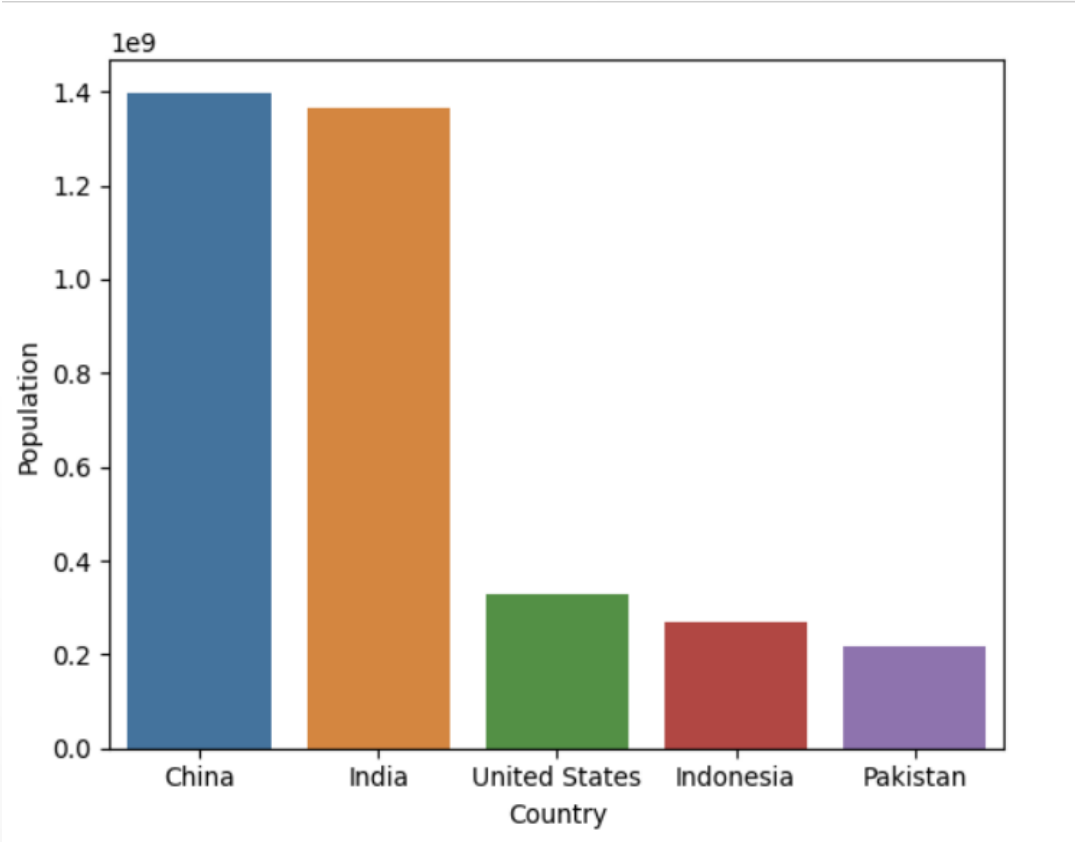


GENERAL ANALYSIS: POPULATION

WORLD POPULATION BY COUNTRY

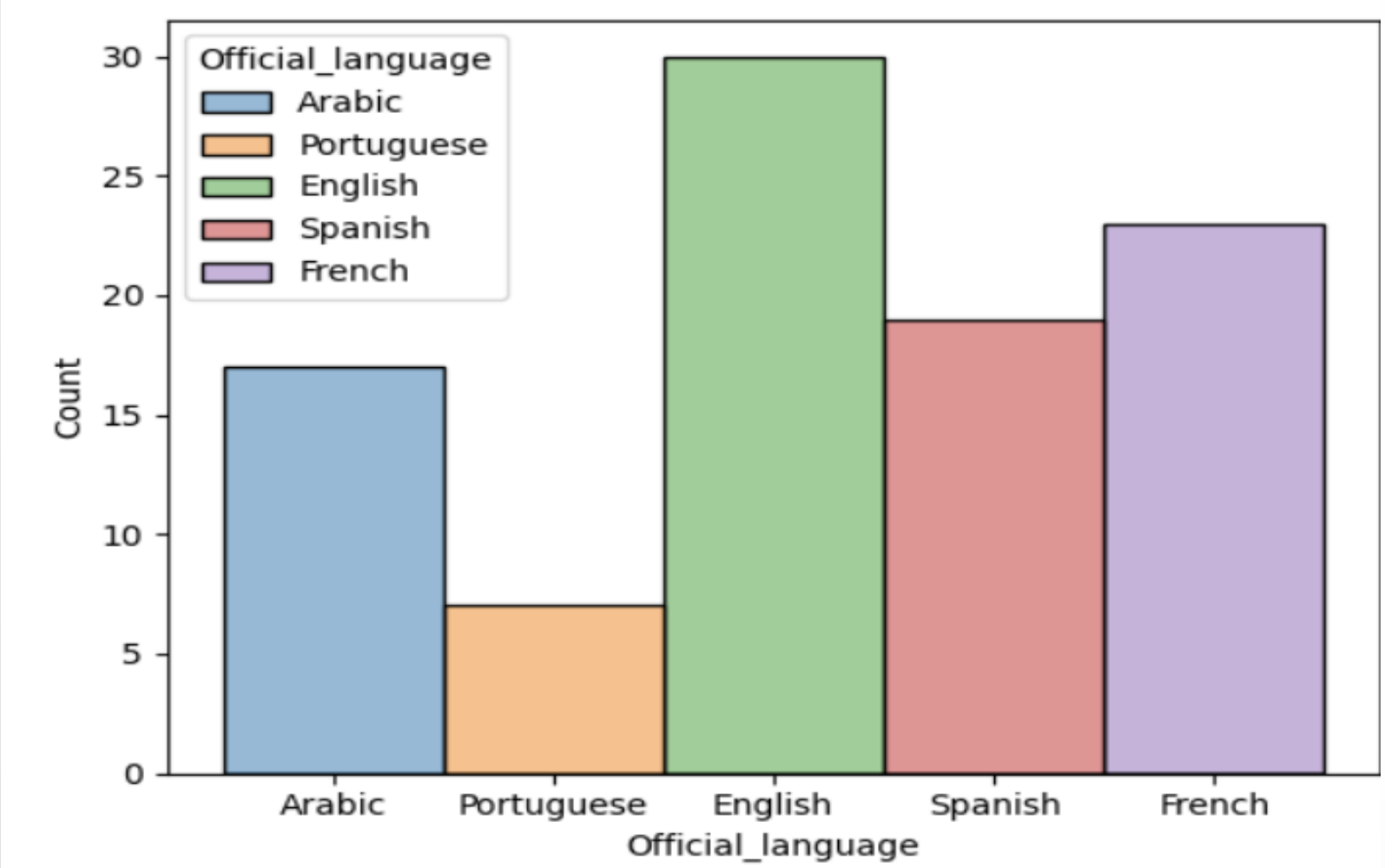


TOP 5 COUNTRIES BY POPULATION



GENERAL ANALYSIS: LANGUAGES

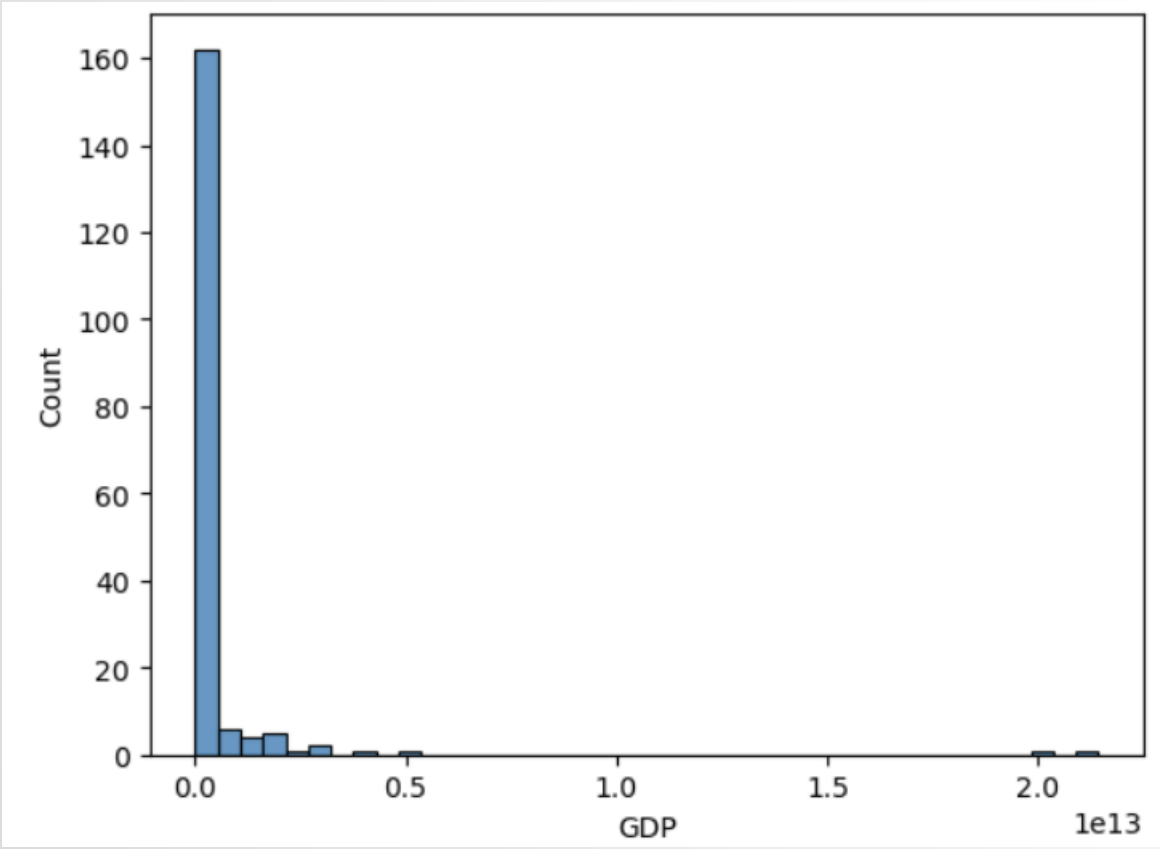
TOP 5 OFFICIAL LANGUAGES



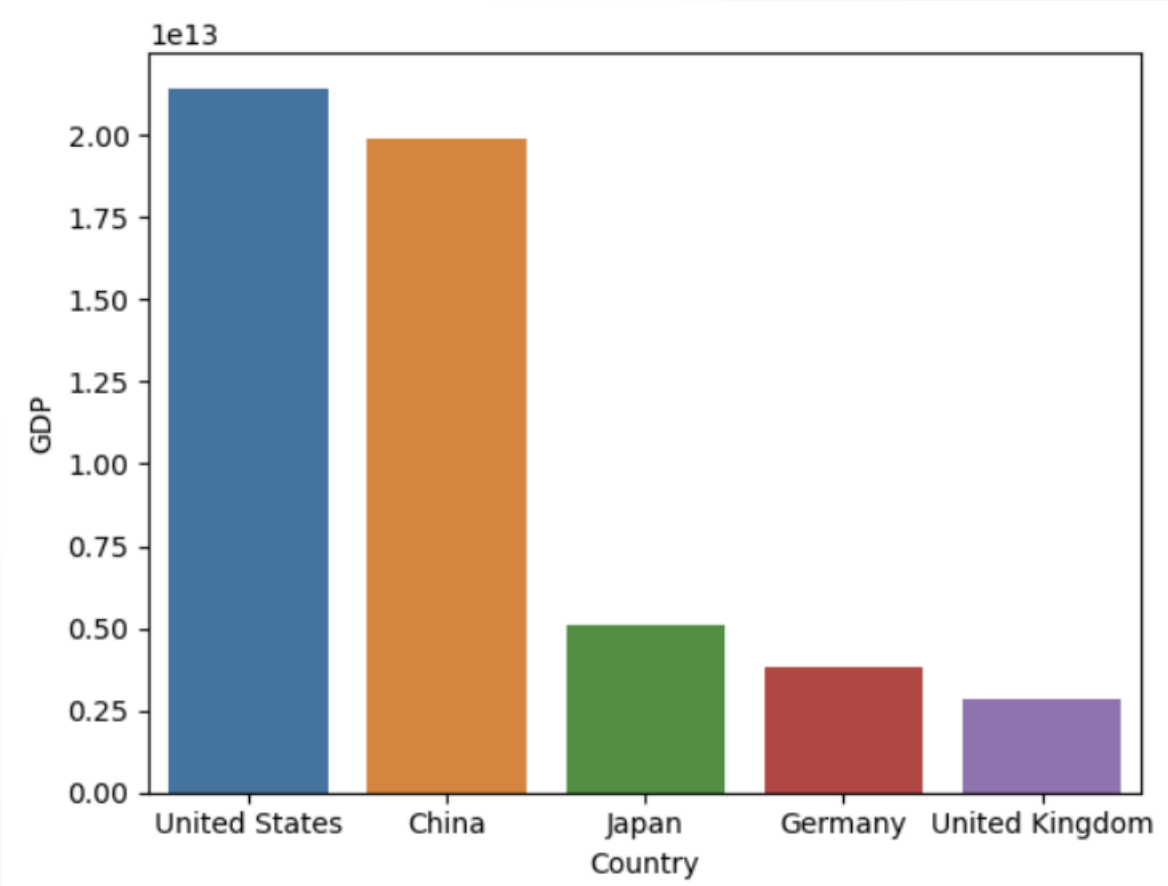
GENERAL ANALYSIS: GROSS DOMESTIC PRODUCT

GENERAL ANALYSIS: GDP

HISTOGRAM OF COUNTRY'S GDP

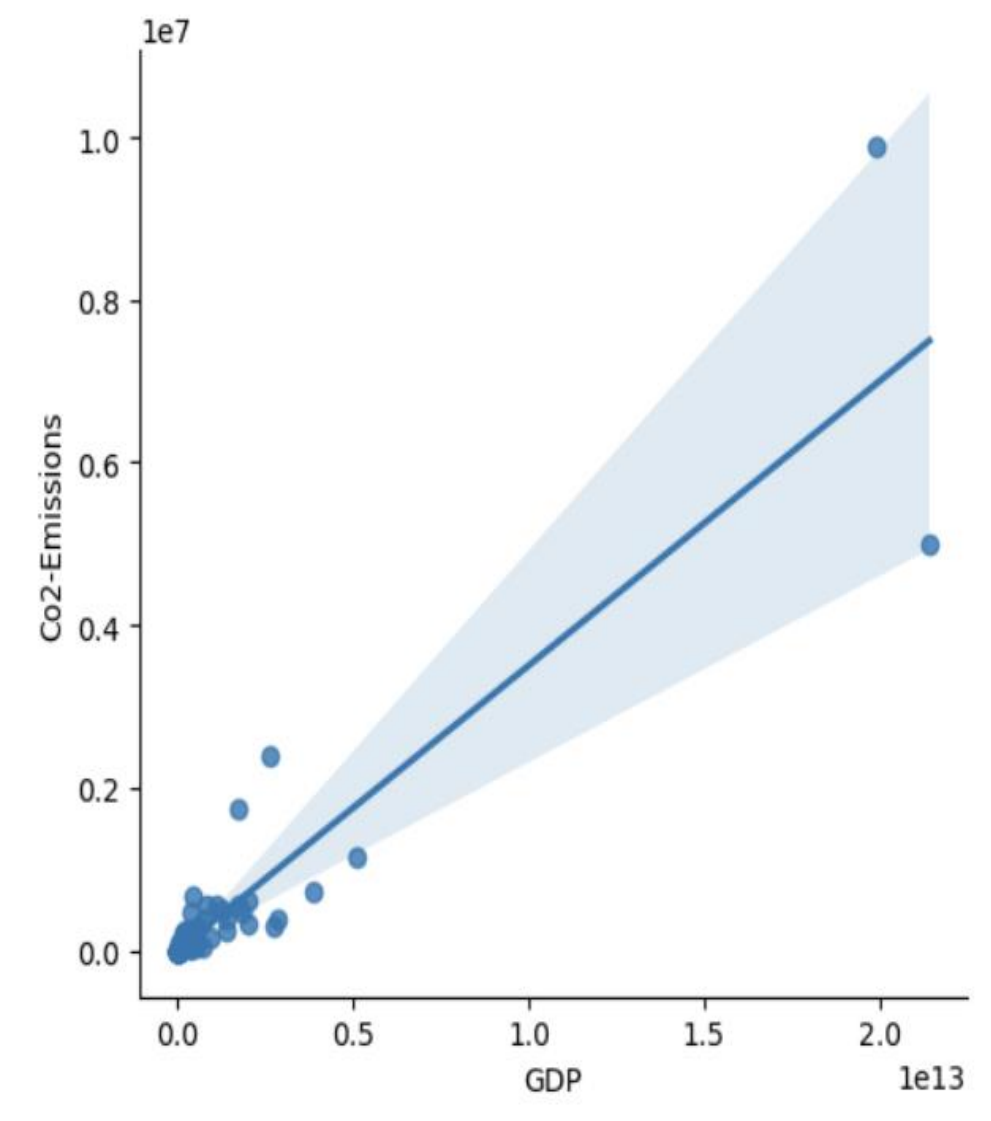


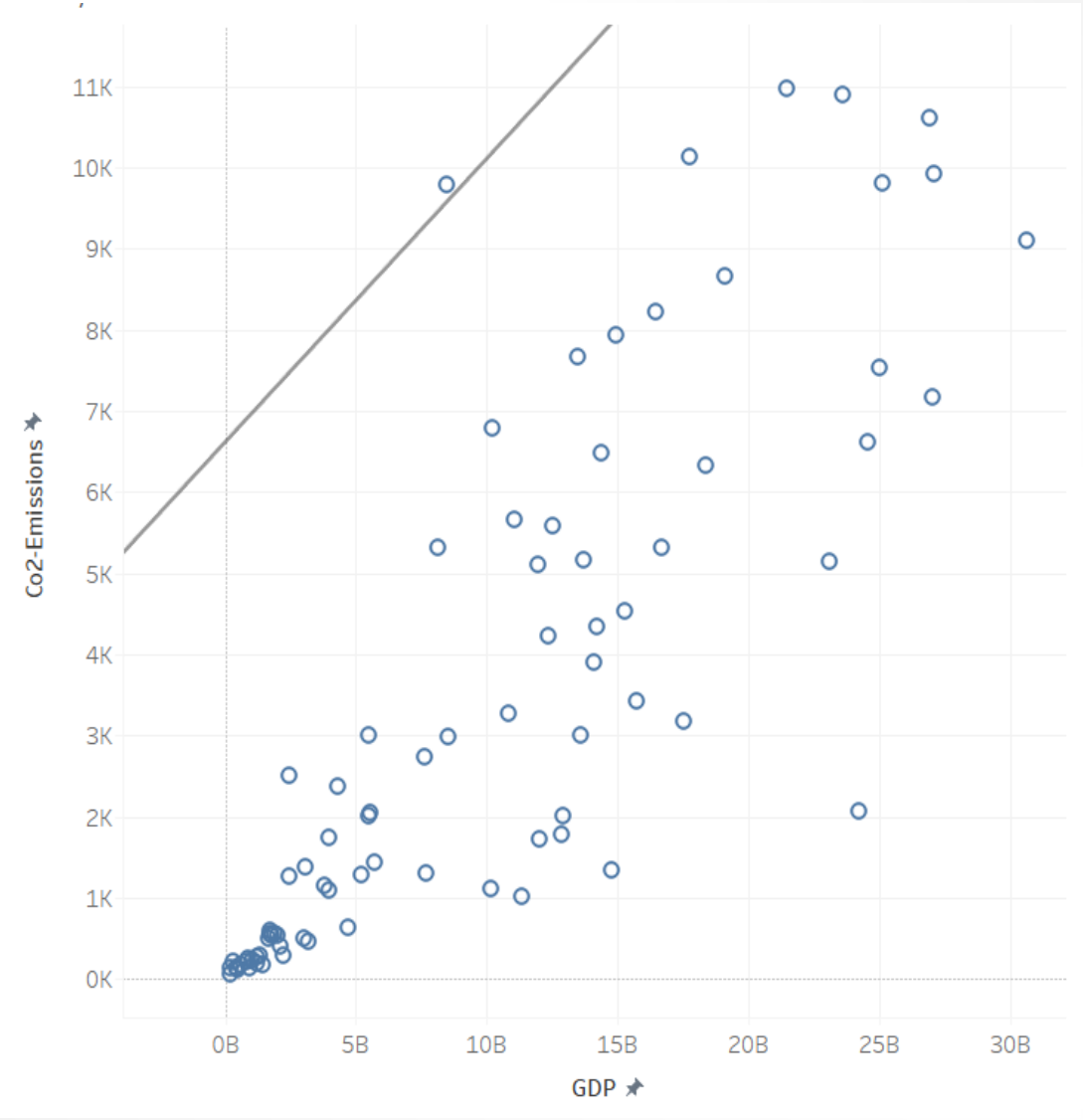
TOP 5 COUNTRIES BY GDP



ECONOMIC ANALYSIS

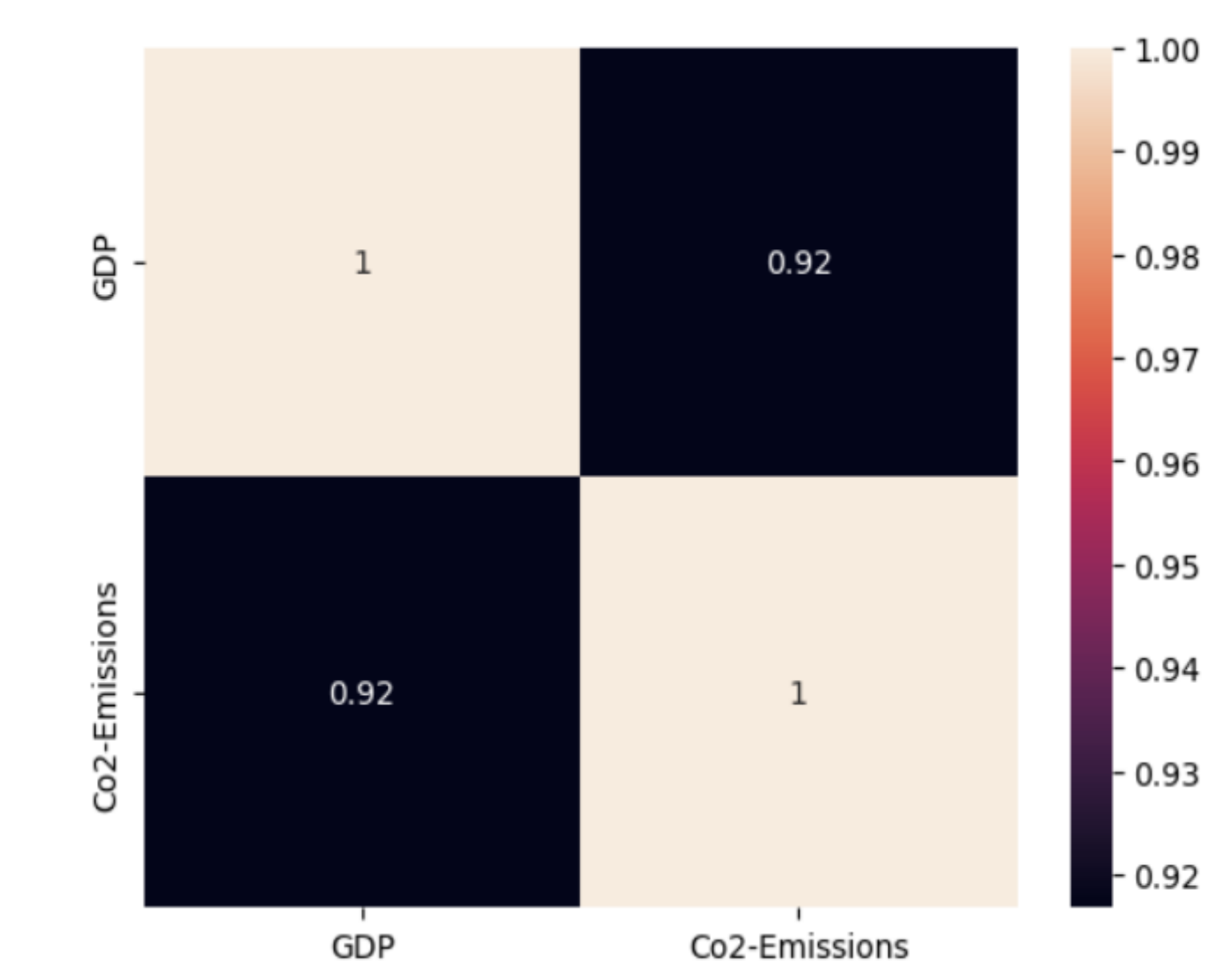
IS THERE CORRELATION BETWEEN A COUNTRY'S GDP AND CO2 EMISSIONS?





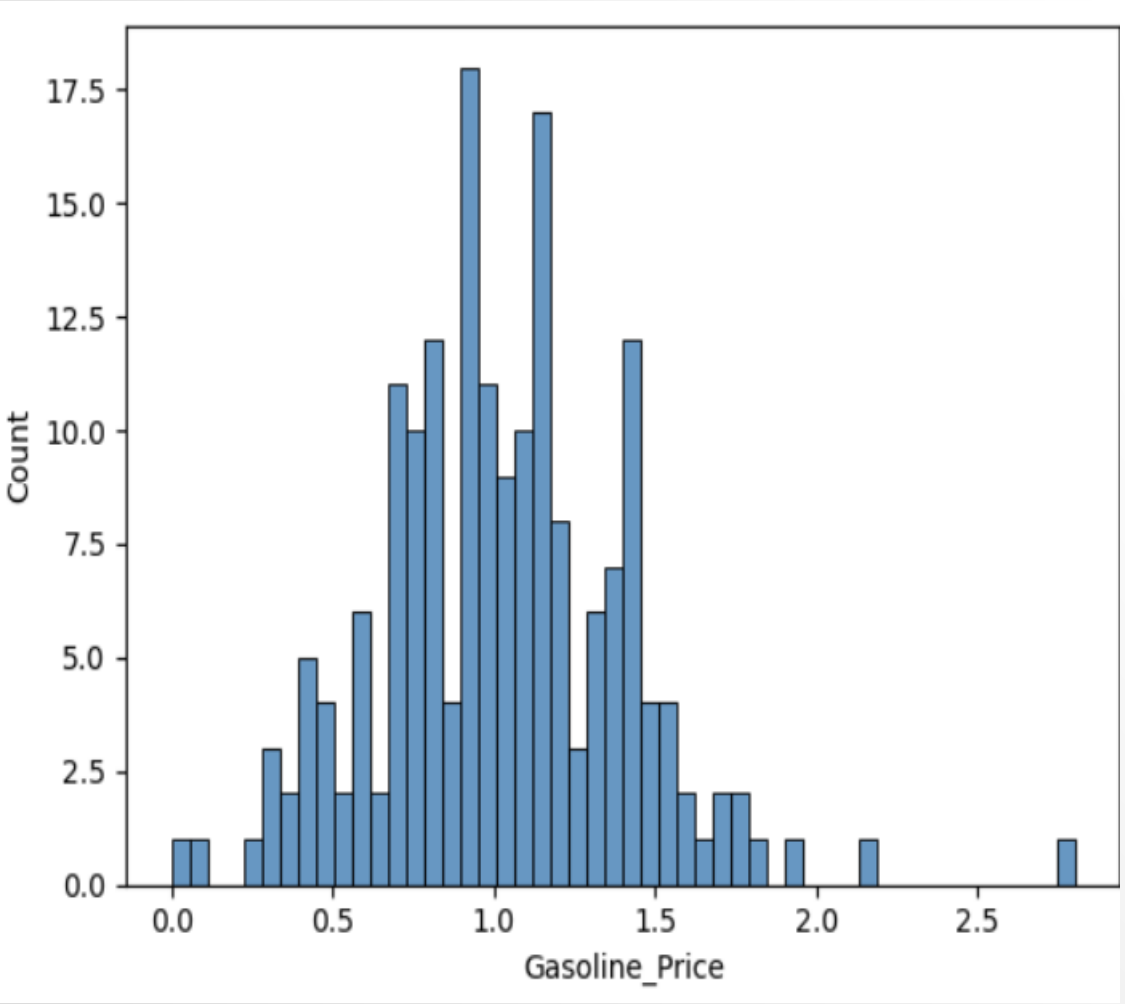
ECONOMIC ANALYSIS

IS THERE CORRELATION BETWEEN A COUNTRY'S GDP AND CO2 EMISSIONS?

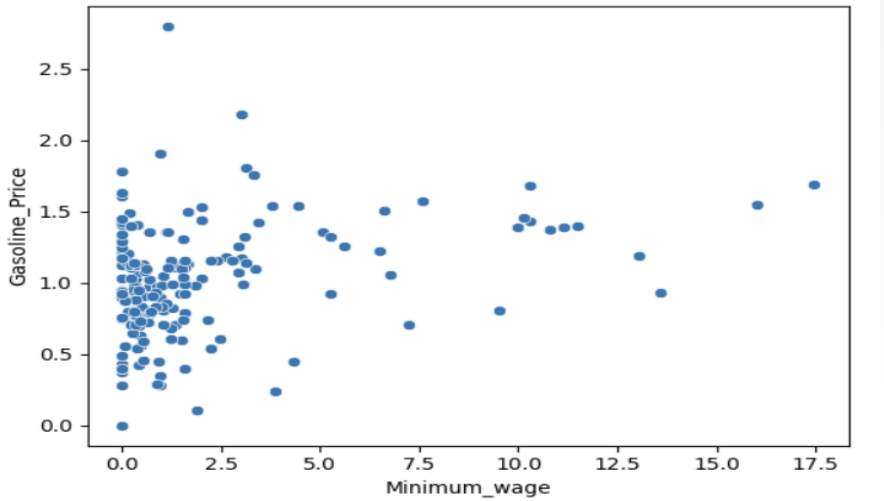


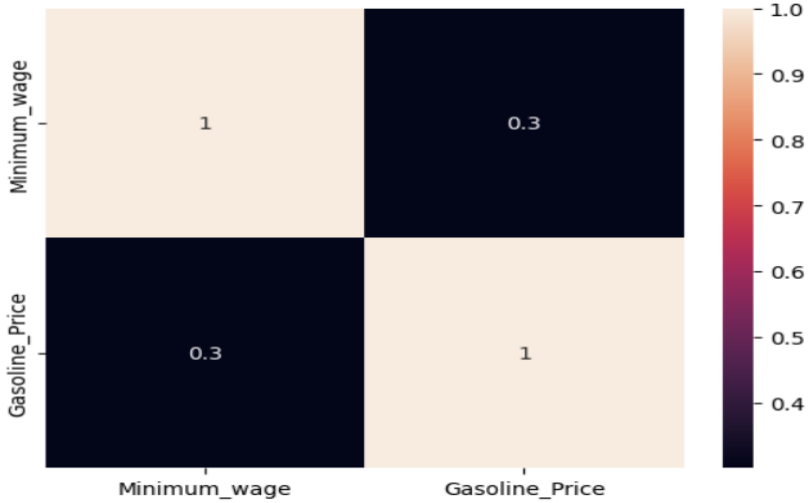
ECONOMIC ANALYSIS

GAS PRICES PER LITER (USD)



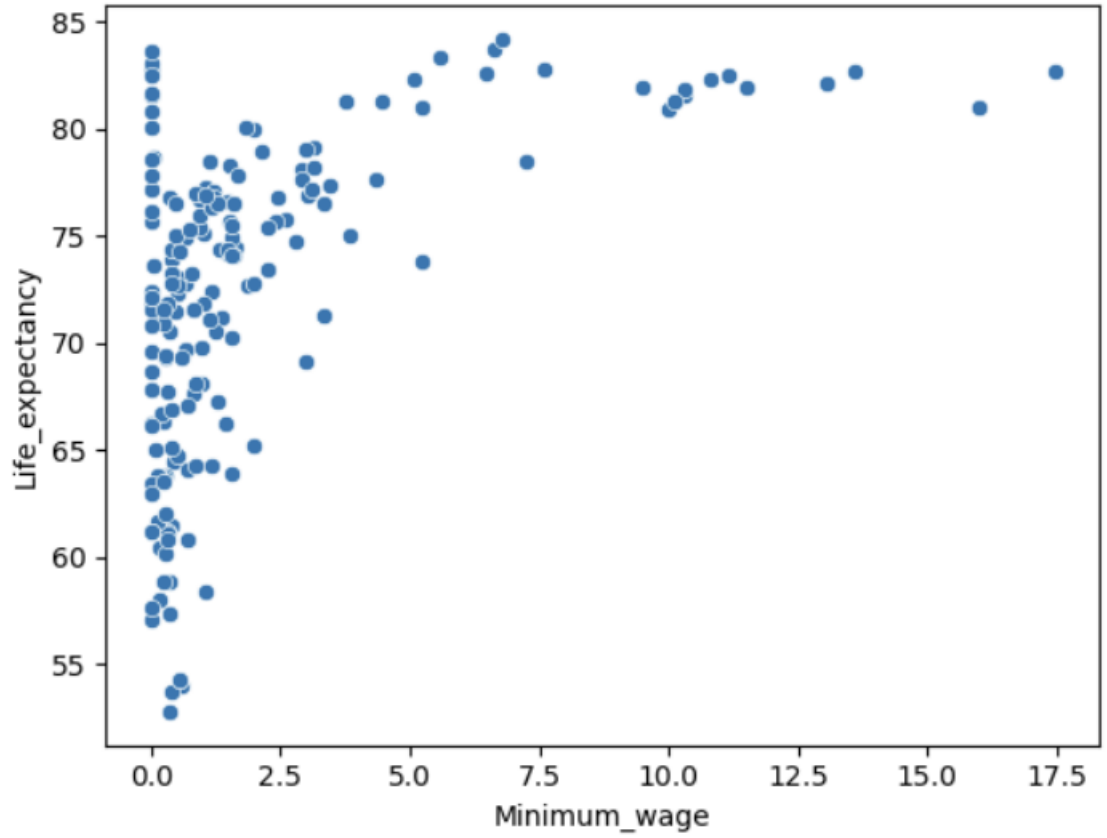
MINIMUM WAGE/GAS PRICES

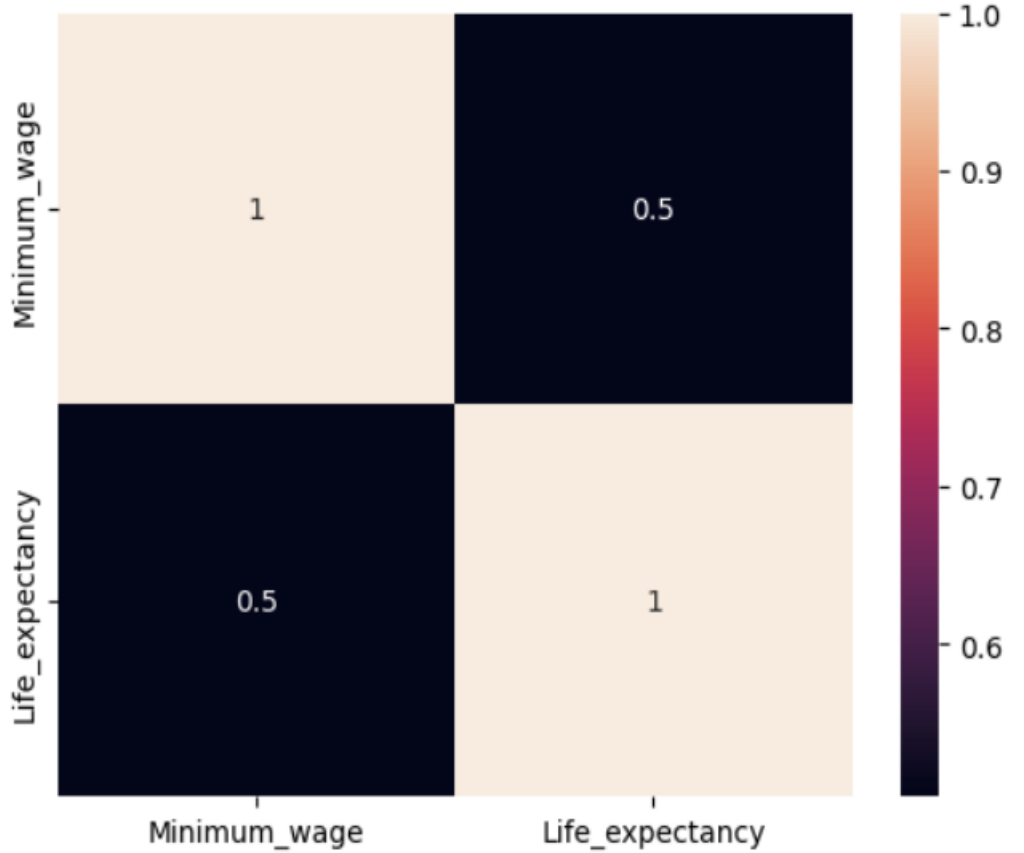




ECONOMIC ANALYSIS

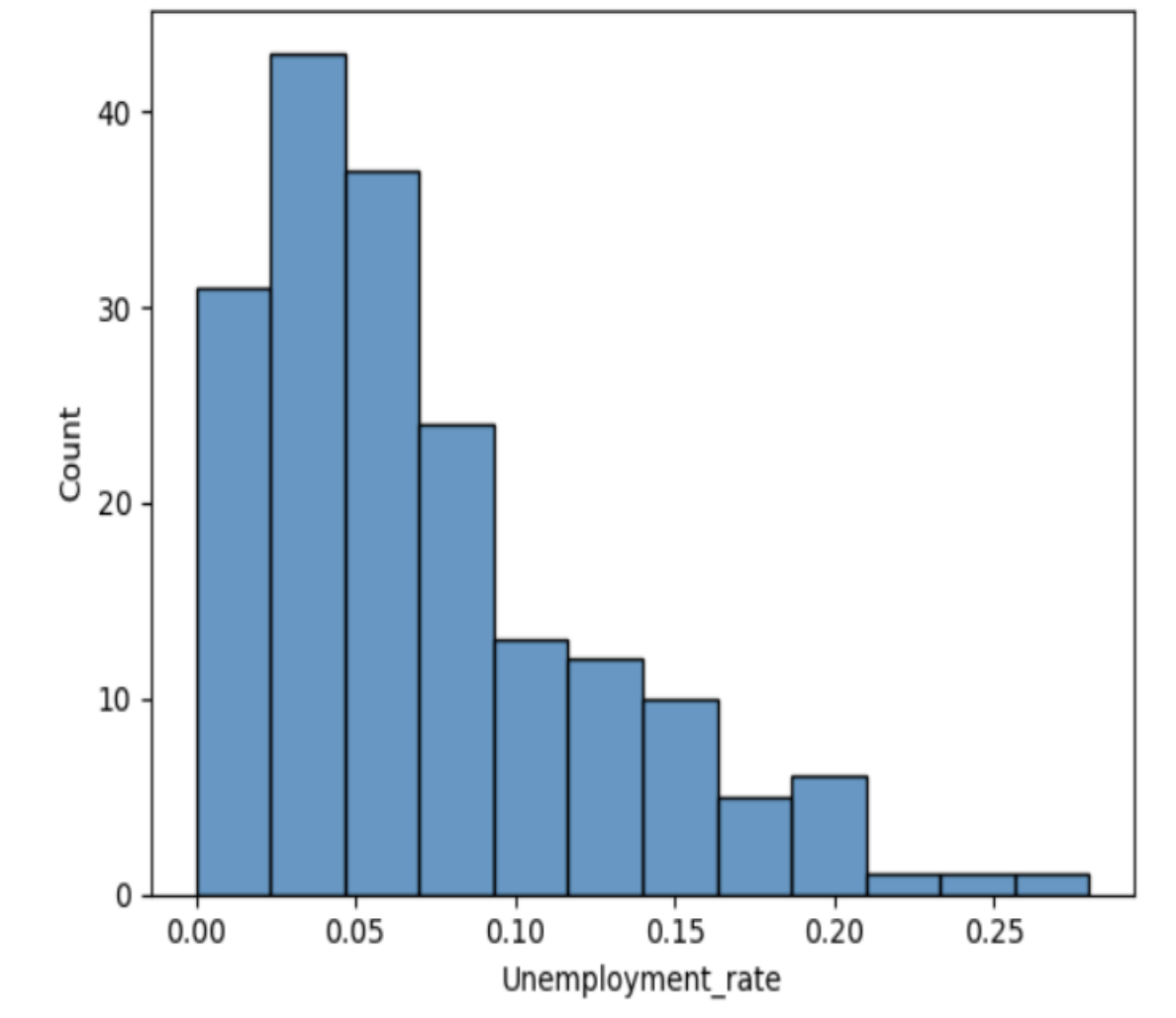
IS THERE CORRELATION BETWEEN MINIMUM WAGE AND LIFE EXPECTANCY?





ECONOMIC ANALYSIS

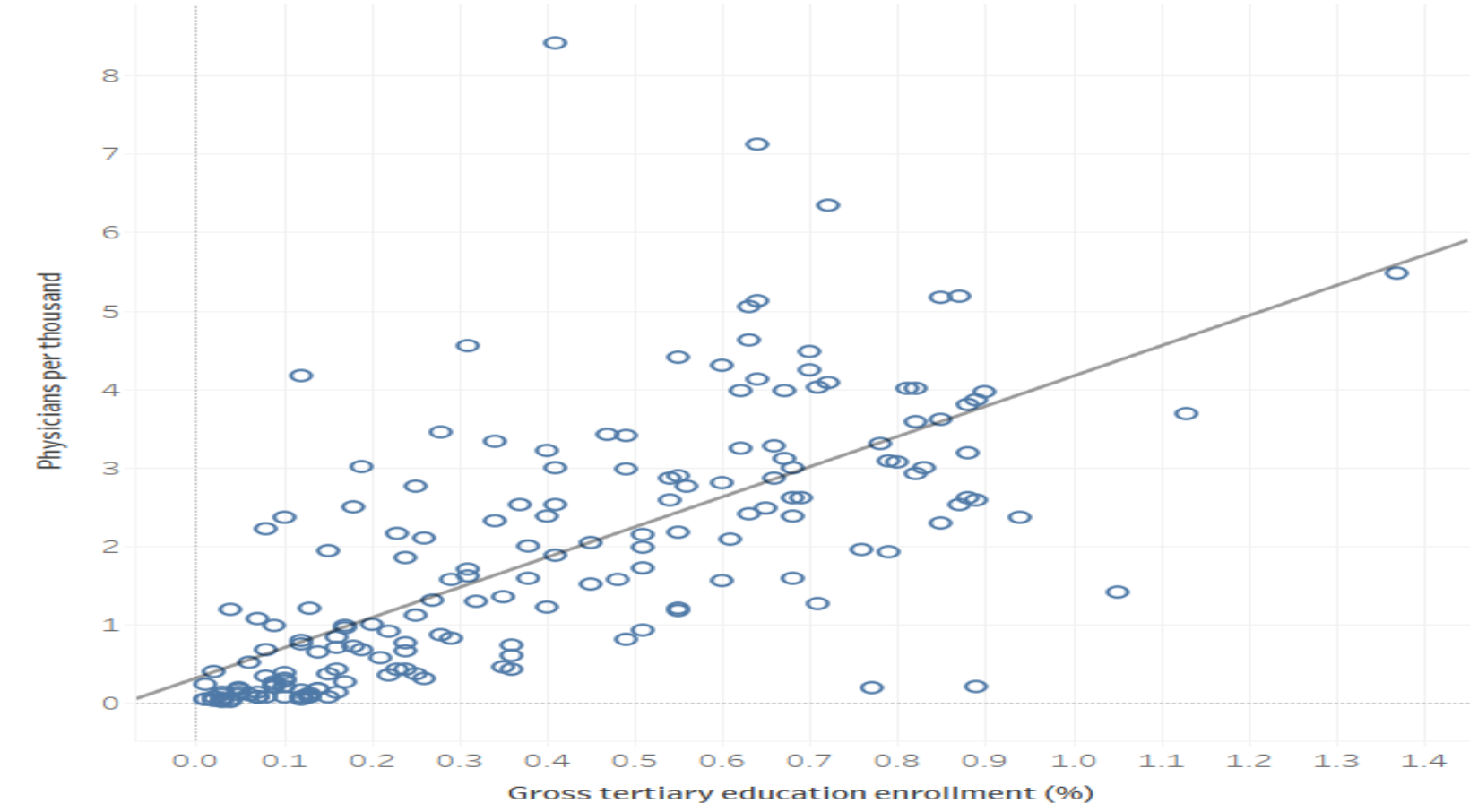
IS THERE CORRELATION BETWEEN GDP AND UNEMPLOYMENT RATE?



EDUCATION ANALYSIS

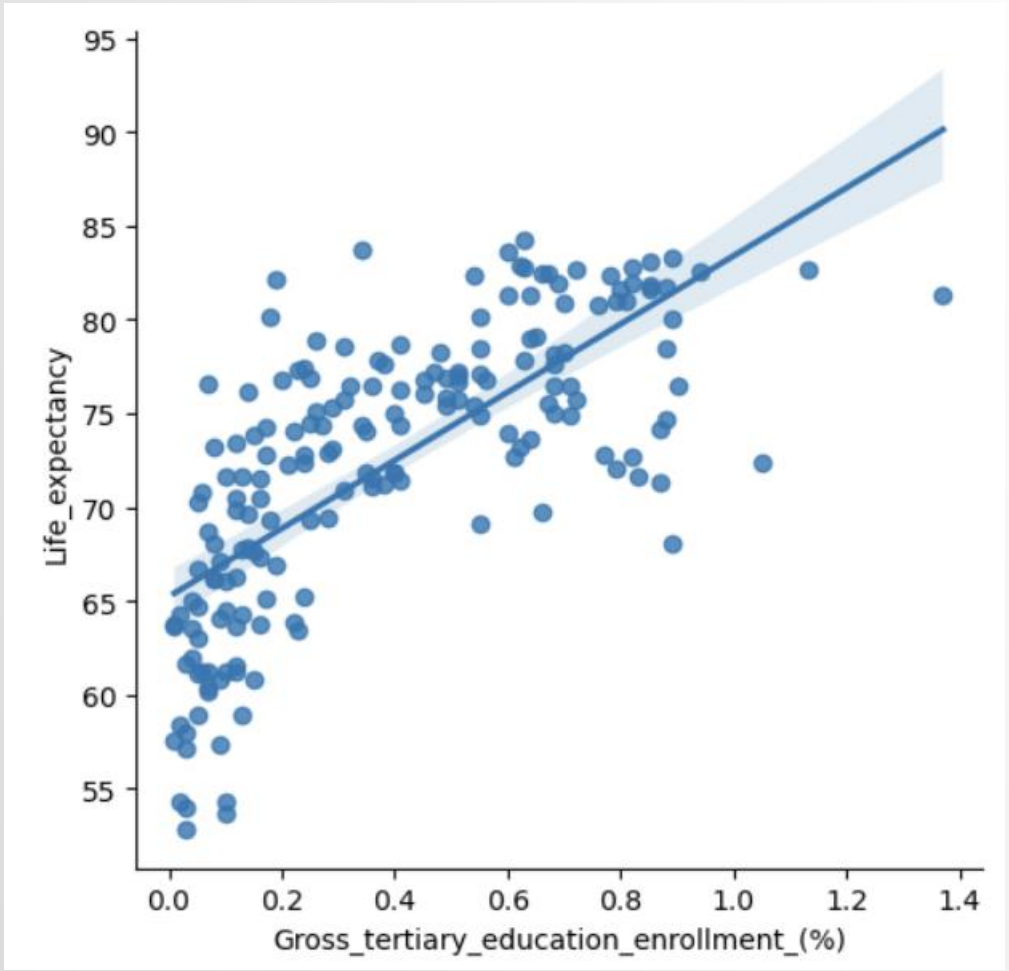
IS THERE CORRELATION BETWEEN A COUNTRY'S HIGHER LEVEL EDUCATION ENROLLMENT AND THE NUMBER OF PHYSICIANS?

- Correlation: 0.7

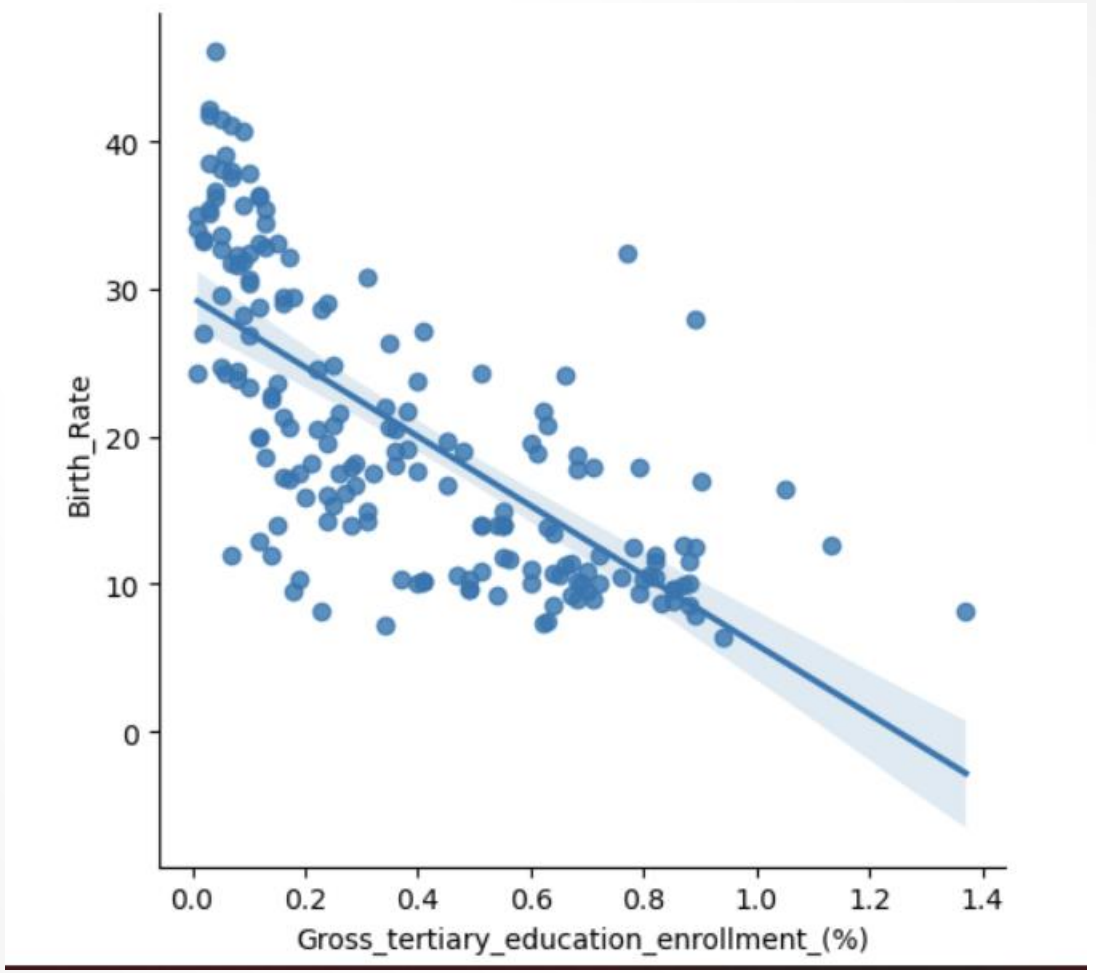


EDUCATION ANALYSIS

CORRELATION BETWEEN TERTIARY EDU ENROLLMENT & HEALTHCARE METRICS



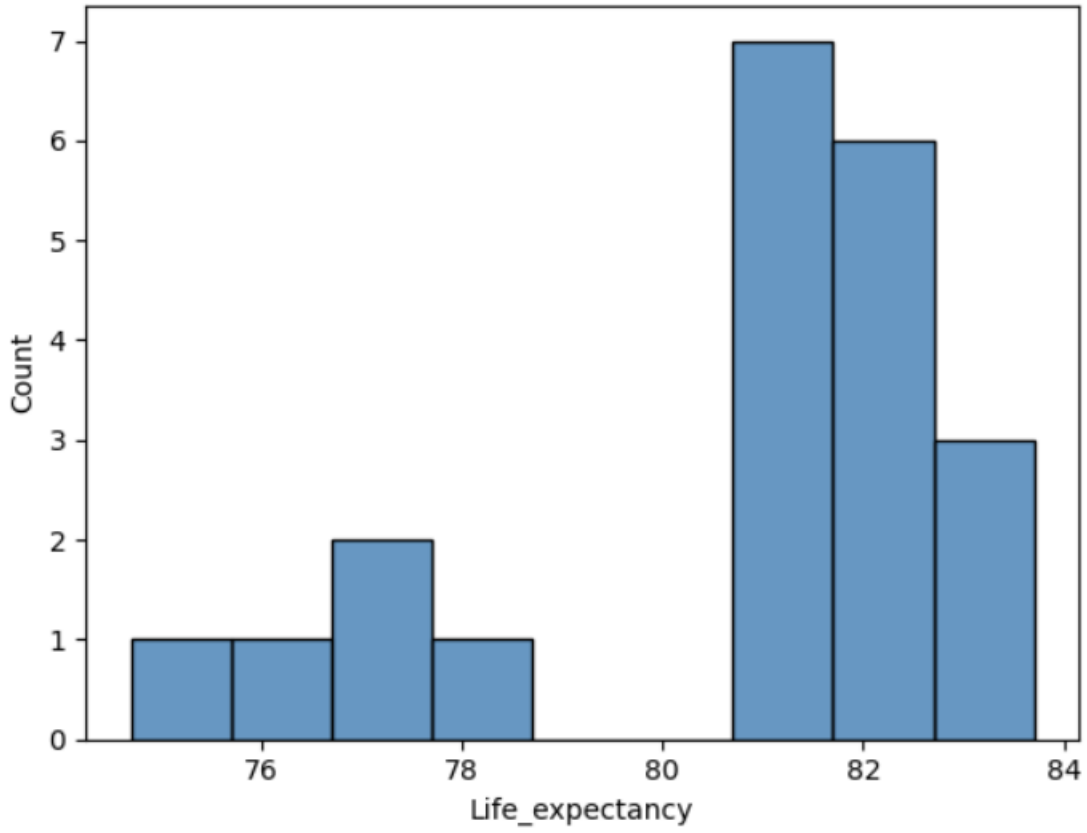
Correlation: 0.71



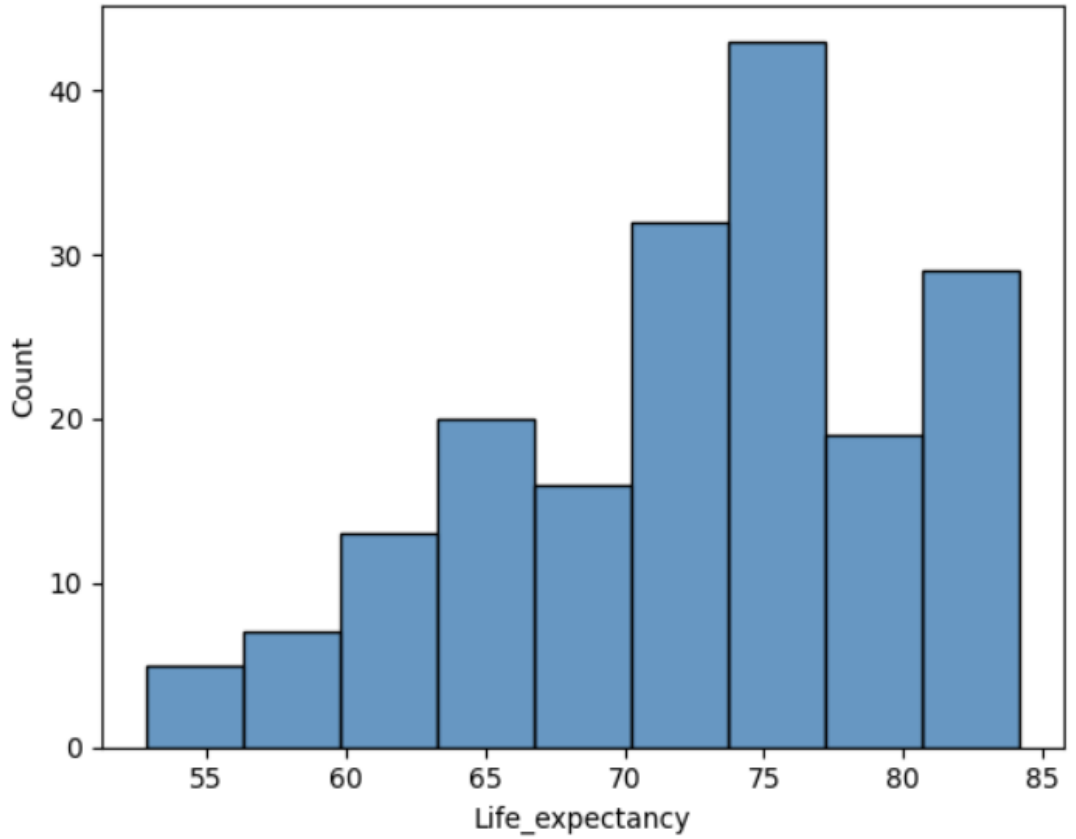
Correlation: -0.7

COMPARING A SAMPLE TO THE POPULATION: LIFE EXPECTANCY

COUNTRIES USING THE EURO

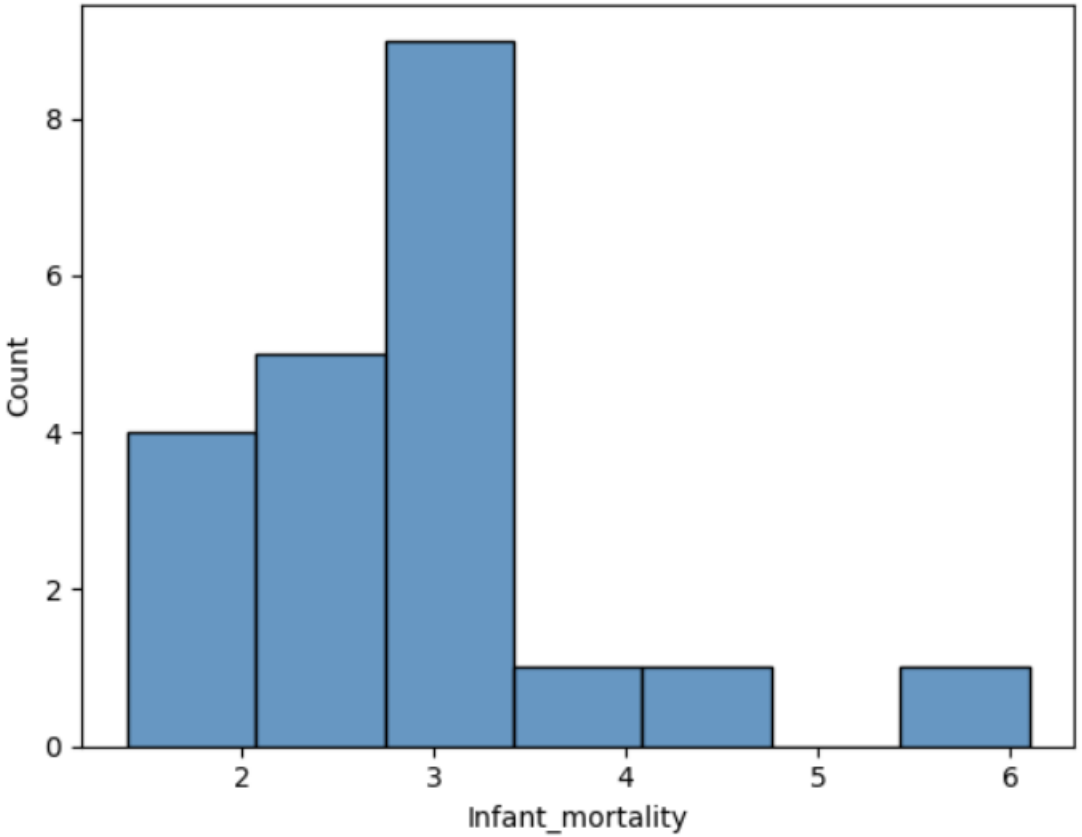


REST OF THE WORLD

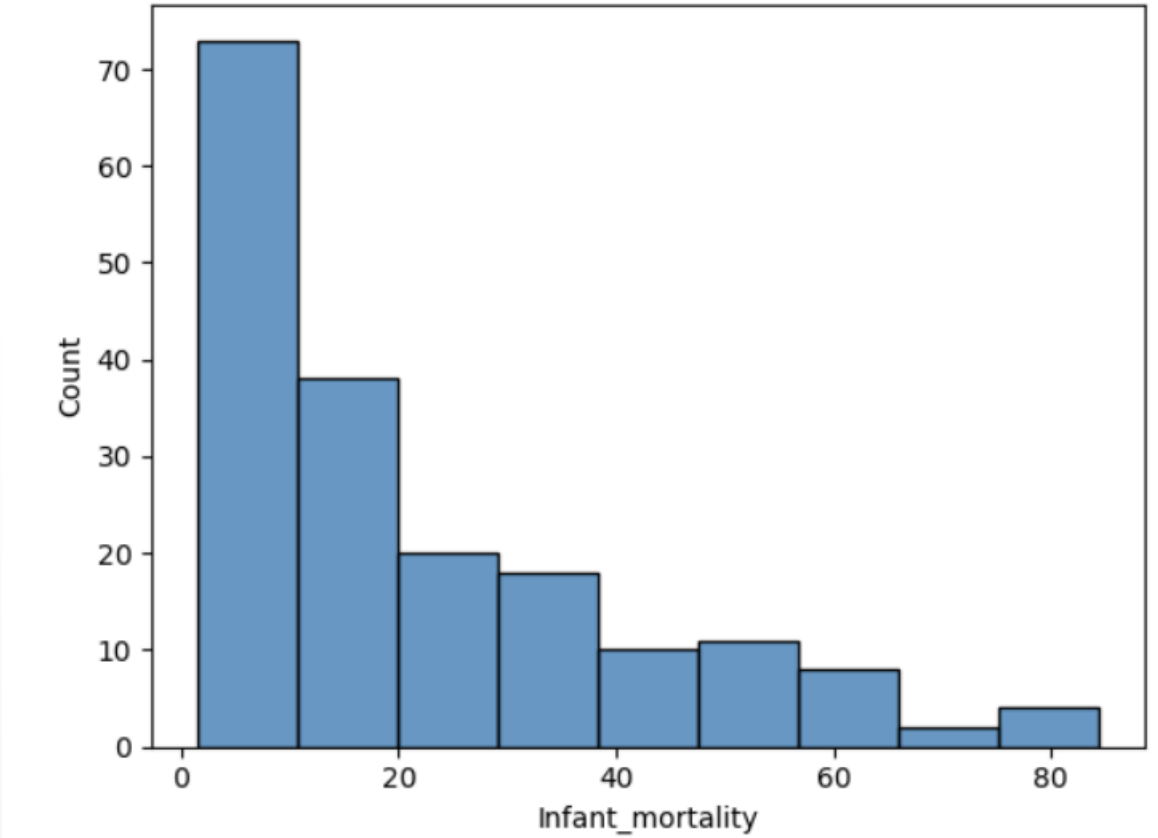


COMPARING A SAMPLE TO THE POPULATION: INFANT MORTALITY

COUNTRIES USING THE EURO

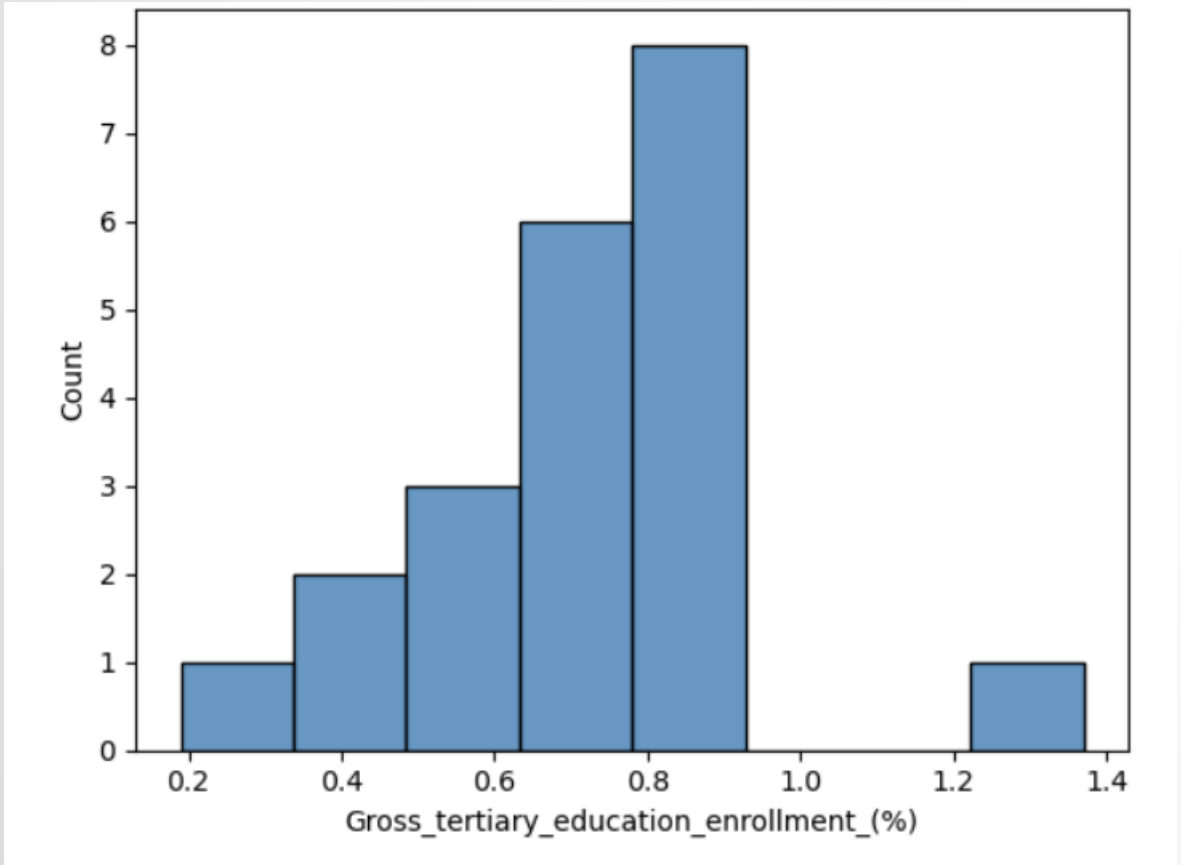


REST OF THE WORLD

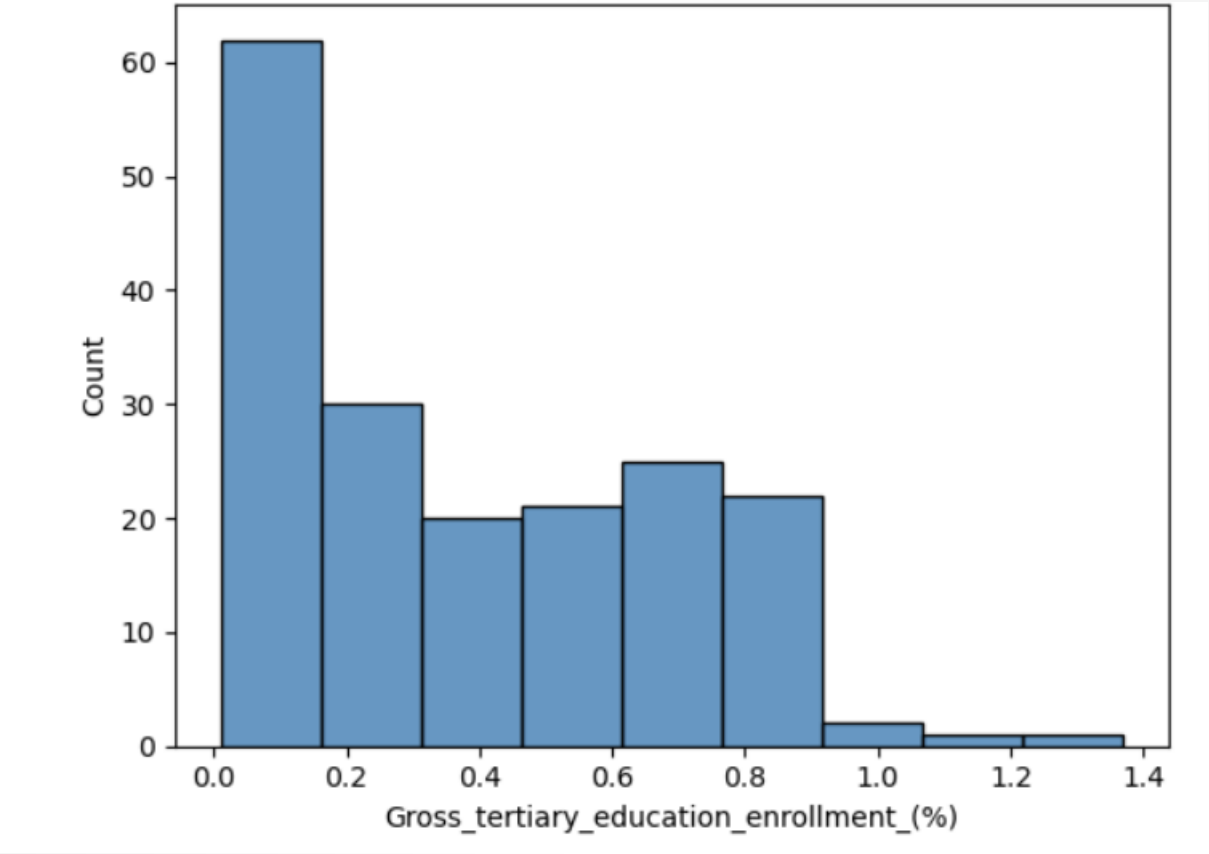


COMPARING A SAMPLE TO THE POPULATION: TERTIARY EDU ENROLLMENT

COUNTRIES USING THE EURO



REST OF THE WORLD



DISCUSSION

- Findings:
  - Found strong correlation between several variables in the dataset (Demographics, economic indicators, healthcare metrics, education, and environmental factors)
  - Compared to the rest of the world, countries on the Euro were found to have higher education enrollment and life expectancy

- Challenges
  - Limited variables
  - Data only for 2023
  - Correlation is not causation

- Application
  - Economists/world leaders can use data to analyze current trends and set goals

QUESTIONS?