CIA World Factbook Analysis

Luisa Kalkert, Luke Richard, Moïse Placier, Noé Coursimaux, & Nicholas Chandler

2025-07-24

Table of Contents

Introduction

In this report, we analyze a dataset created from the CIA World Factbook[1] with the aim of visualizing the data therein. Specifically, the dataset we use was scraped from the CIA website in 2023 and is available on Kaggle[2] or on our GitHub Repository[3]. This dataset contains information about many countries and territories around the world and was compiled by the CIA from the United States. In its unprocessed form, it has 258 rows and 980 columns where each row corresponds to a country/territory and each column corresponds to an attribute associated with the given country/territory. In the following analyses, however, we primarily focus on different subsets of this dataset to target the analysis on specific aspects of the data. In each of the sections of this report, there was much effort devoted to dataset cleaning and string processing to remedy missing values and the default string data type from our dataset.

Technical details aside, the content of our report concerns itself most broadly with the state of the world in numbers (2023). We begin by examining connections various countries in the world have with one another through trade and the CIA description of each. Then we examine energy production, consumption, and trade. Next we analyze the state of agriculture and how it relates to water and energy use. We then examine relationships between various human geographic variables. Finally, we look at the difference between men and women in several demographic variables. All analyses examine the various topics at the country and continent levels to compare regions of the world.

Throughout these analyses we utilize a multitude of visualization techniques including several classical statistical graphics such as boxplots and bar charts, as well as some modern techniques such as chord plots and choropleths. We collaborated to ensure that we used a variety of techniques while maintaining clear analysis of each of the key questions.

[1] https://www.cia.gov/the-world-factbook/

[2] https://www.kaggle.com/datasets/lucafrance/the-world-factbook-by-cia

[3] https://github.com/chandlerNick/BHTDataVisualizationCourse

Question 1

Question 1: How is the World Connected?

This question analyzes the relationships between the different countries and world regions in our dataset. We use two main indicators to look at these relationships: 1. Mentions of other countries in each country's background information. 2. Imports and exports of each country.

1.1 Relations from the Countries' Introduction Background

The first part of the analysis is based on mentions in each country's background information. Our dataset has a short description in natural language for each country, containing information about the country's history, politics, and other relevant information. This gives us a bit of context to the country, but is not in tabular format.

To go from the description to a more structured format, we find mentions of other countries in the description of each country. For example, Argentina's background description text starts like this:

> "In 1816, the United Provinces of the Río de la Plata declared their independence from Spain. After Bolivia, Paraguay, and Uruguay went their separate ways, the area that remained became Argentina. ... "

Here, we find mentions of "Spain", "Bolivia", "Paraguay", and "Uruguay".

For each country, we also define a list of alternative names, e.g., "United Kingdom" is also known as "Britain", "UK", "England", "British". In this way, we can search for the mentions of other countries in the description of each country, and store the relationships between countries in a matrix $A = (a_{i,j})$, where:

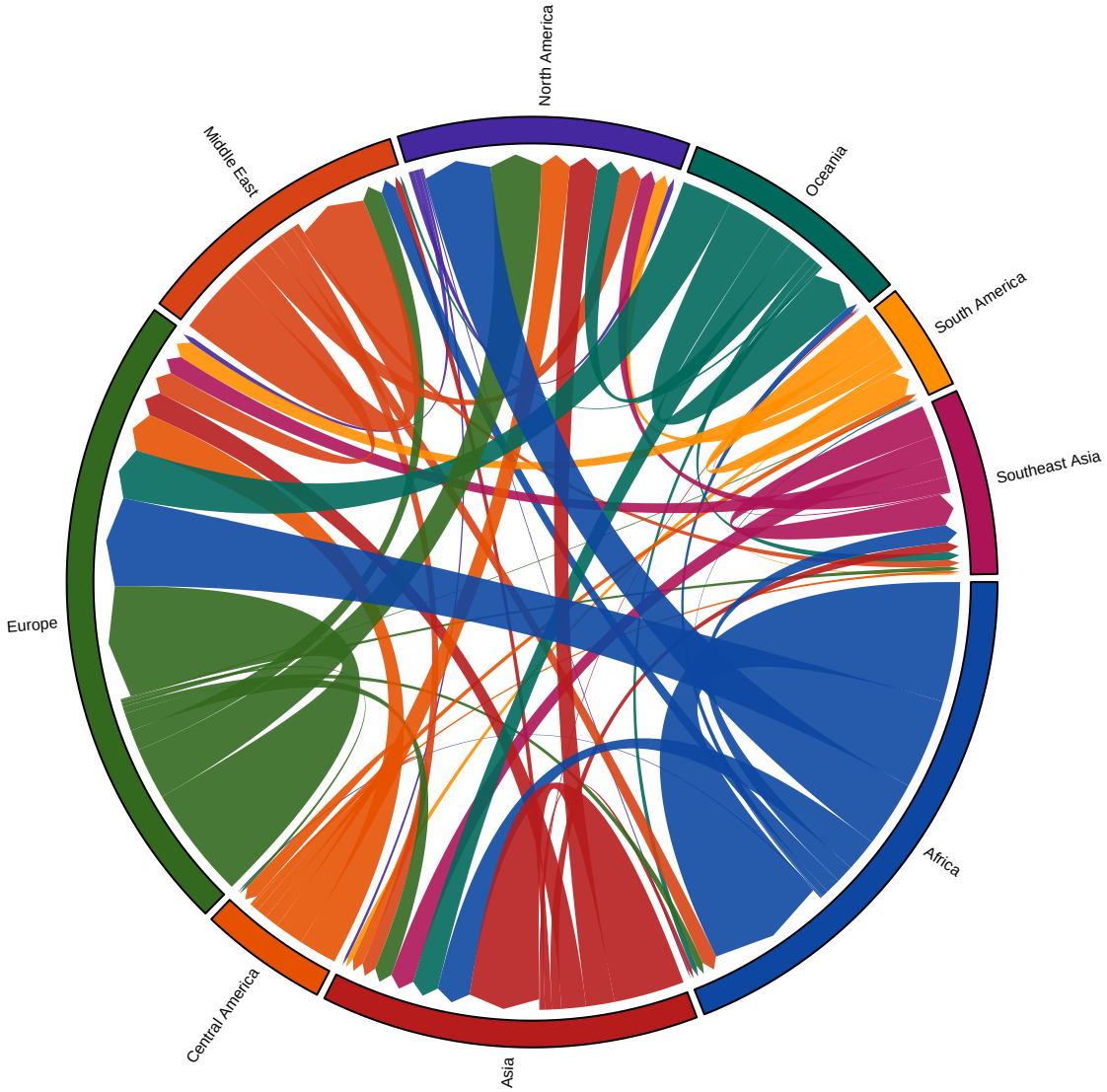
$$a_{i,j} = \begin{cases} 1 & \text{if country } j \text{ mentions country } i \\ 0 & \text{otherwise} \end{cases}$$

where n is the number of countries. This matrix is used for our subsequent analysis.

1.1.1 Chord Diagram of World Region Relationships

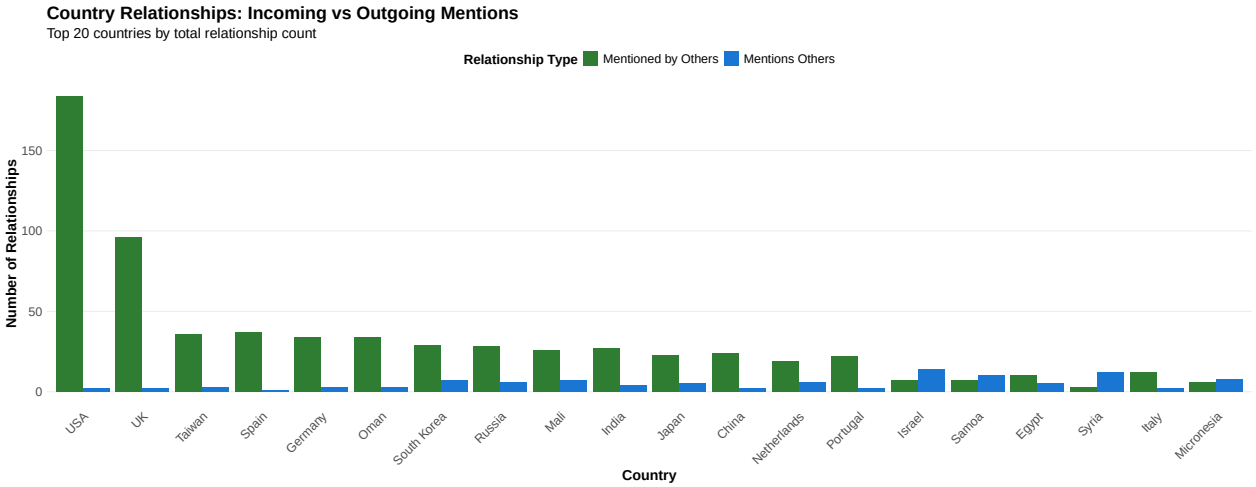
First of, we start with a high-level view of the relationships between different world regions. To do this, we create a chord diagram to visualize the relationships between the different regions using the circlize package. When a country is mentioned in the background information of another country we'll add (or increase the thickness of) an arrow from the region of the first country to the region of the second country. E.g. if the background information of Argentina mentions Spain, we'll add an arrow from South America to Europe. If Argentina mentions Paraguay, we'll add an arrow from South America back to South America.

World Region Relationships Chord Diagram

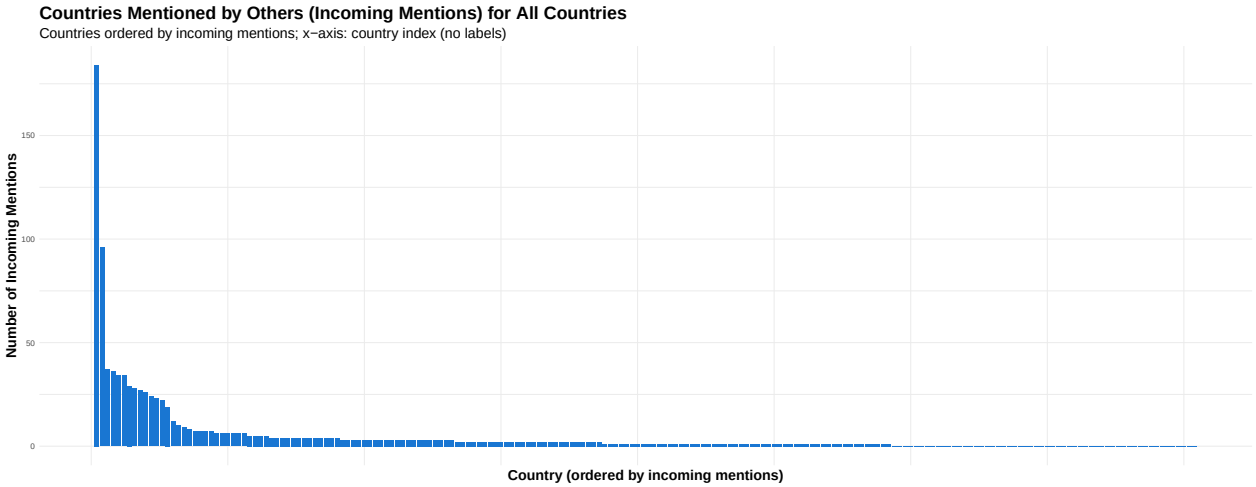


As we can see, not all regions are connected to each other in the same way. For example, North America has many incoming connections, but only a few outgoing connections. This means many countries mention Nort American Countries in their background information, but not the other way around. Also we can see that there are some regions wich are relevant to many other regions, e.g. Europe, Asia, North America and to a lesser extent Southeast Asia. This might be due to the colonial past of many european countries, as well as current trade relations. At the same time, there are other regions that are only relevant to a few other regions, e.g. Oceania, Africa, South and Central America as well as the Middle East. Especially for Africa you can clearly see that african countries mention countries from many other regions but the is not mentioned the other way around. Africa is a continent that has been influenced or colonized by many other regions, but is yet to play a significant role internationally.

1.1.2 Country Relationships Barplot



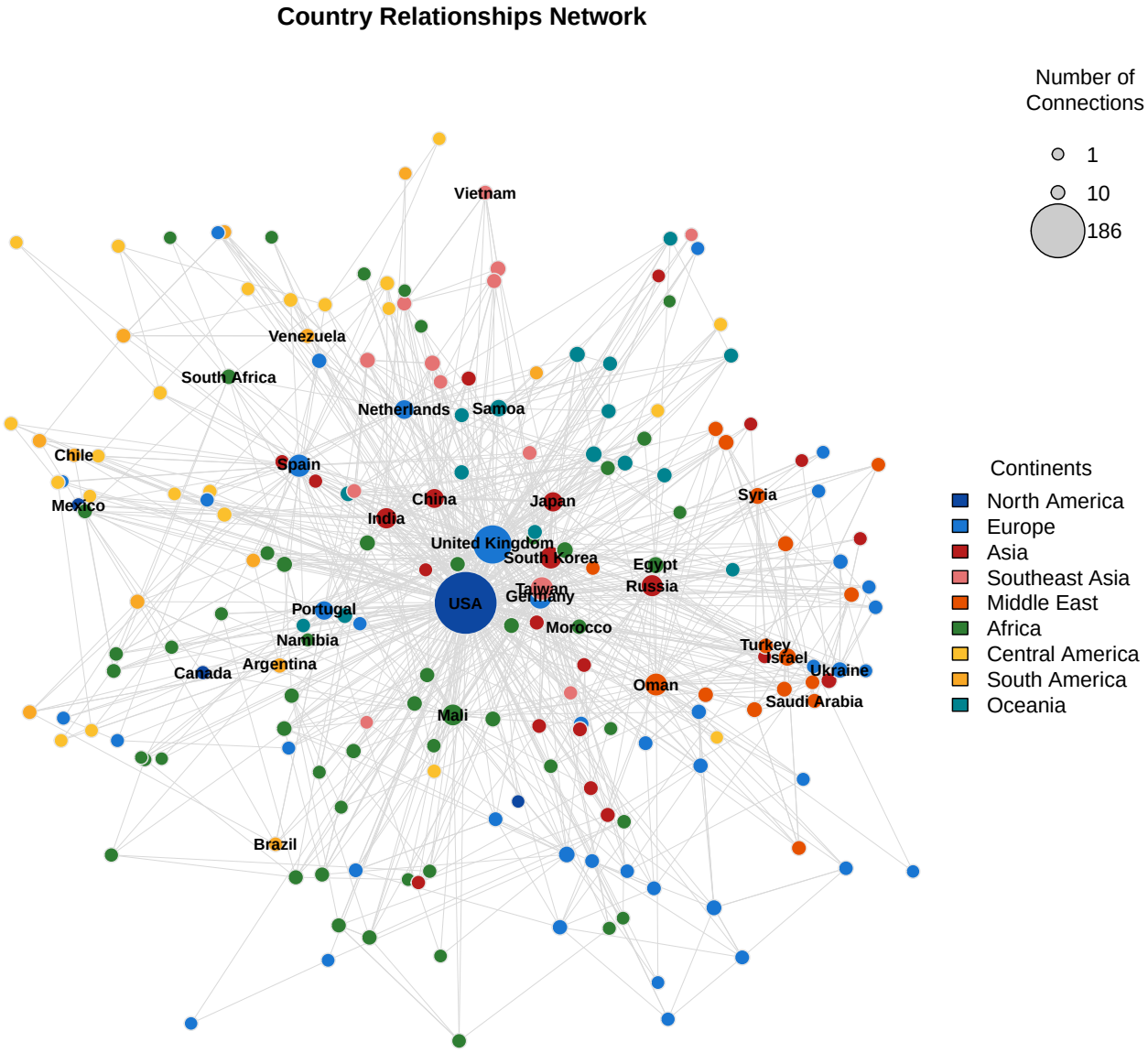
Here, we can see that the notion from the chord diagram is confirmed. The US is mentioned in the background information of 184 other countries, so almost every country in the world. It is still open, whether this is due to the US's influence on the world or due to the fact that the data set is created by an US agency. With the past of the british empire spanning over the globe, it is not surprising to see the UK in second place. Taiwan as a relatively small country ranking #3 was initially surprising, but might be due to its central role in regional border disputes, global semiconductor trade, and geopolitical tensions with China.



Now we are looking at the incoming mentions (how many other countries mention a country in their introduction text) for all the countries. We can see that there are two countries that are mentioned by many other countries (The US and the UK again). We have about 15 countries that have a still high number of mentions, which suggests these countries are highly important on the international stage. We then see a long tail for the rest of the countries, roughly folowing a zipfian distribution, wich suggest these countries have only a minor interational relevance. More than half of the countries world wide are mentioned once or even not at all in the introduction text of another country.

1.1.3 Network Visualization

Here, we vizualise the mentions in a network graph, based on the mention-matrix. Each node represents a country and the edges represent the mentions between countries. The Fruchterman-Reingold algorithm arranges network nodes by simulating physical forces. From an initial random configuration, nodes repel each other and edges act as springs pulling the nodes together. Forces are applied iteratively until the nodes are in a stable configuration (or the maximum number of iterations is reached).

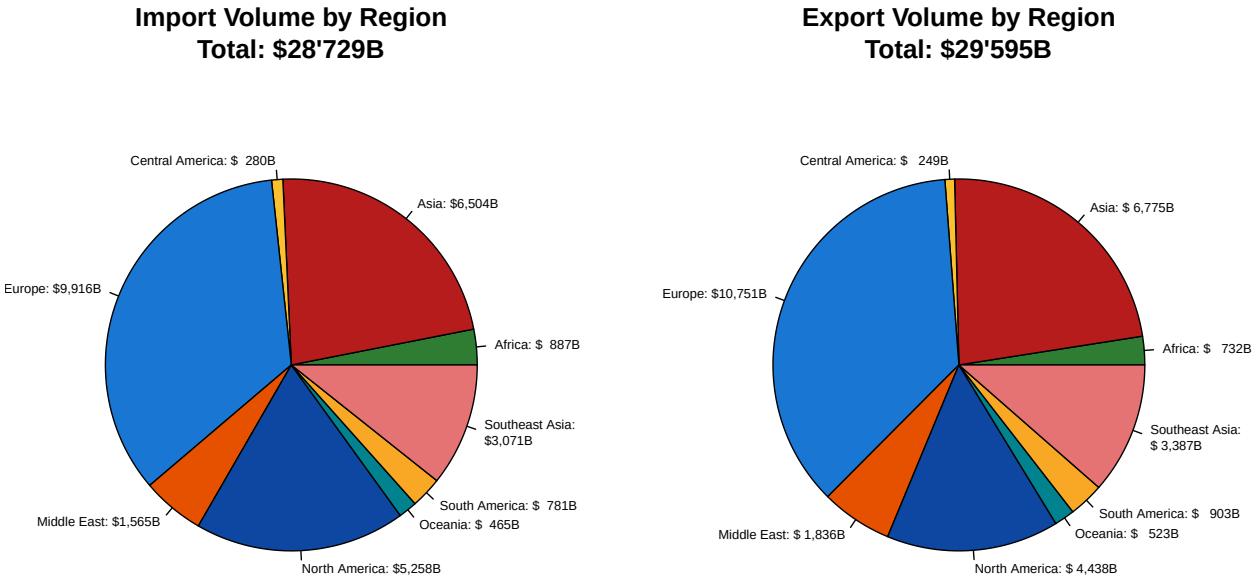


We can see that the network is quite dense, with many countries having multiple connections. There are no clear sub clusters forming, which showcases the globalized and interconnected nature of todays world. We can also see that the network is quite imbalanced with again the US and the UK dominating in central position. Continents are loosely grouped together. Some notable exceptions are Spain and the Netherlands, which are both close to many South and Central American countries, hinting at their past as colonial powers.

1.2 Trade Relations

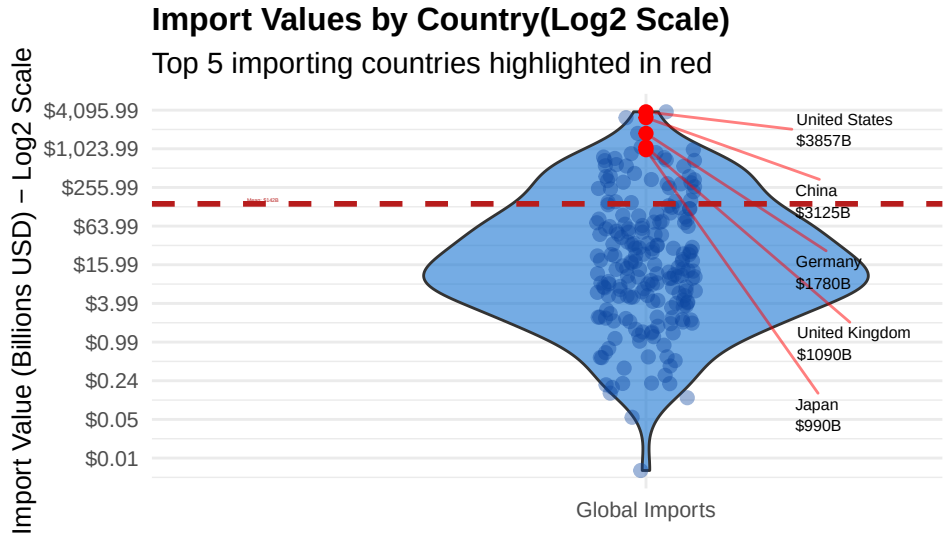
After looking at the qualitative analysis of the mentions, we can now look at the quantitative analysis of the trade relations. Our data set contains information about the monetary value of imports and exports of each country. We will start by looking at the trade distribution by region.

1.2.1 Trade by Region Piecharts



Looking at the pie charts, we can see that there are 4 regions dominating the trade: Europe, Asia, South-East Asia and North America. These are coincidentally also the regions that have the most incoming relations in the Chord diagram in 3.1.1. We can see that these regions play a mayor role both politically as well as in global trade. Interestingly the totals for imports and exports are a bit different, this might for example be due to inaccuracies in reporting. Also, we used the most recent data available, which is not always from the same year. We can see that the imports and exports are very similar per region, with some minor differences. For example we can see that North America has a higher import volume than export volume.

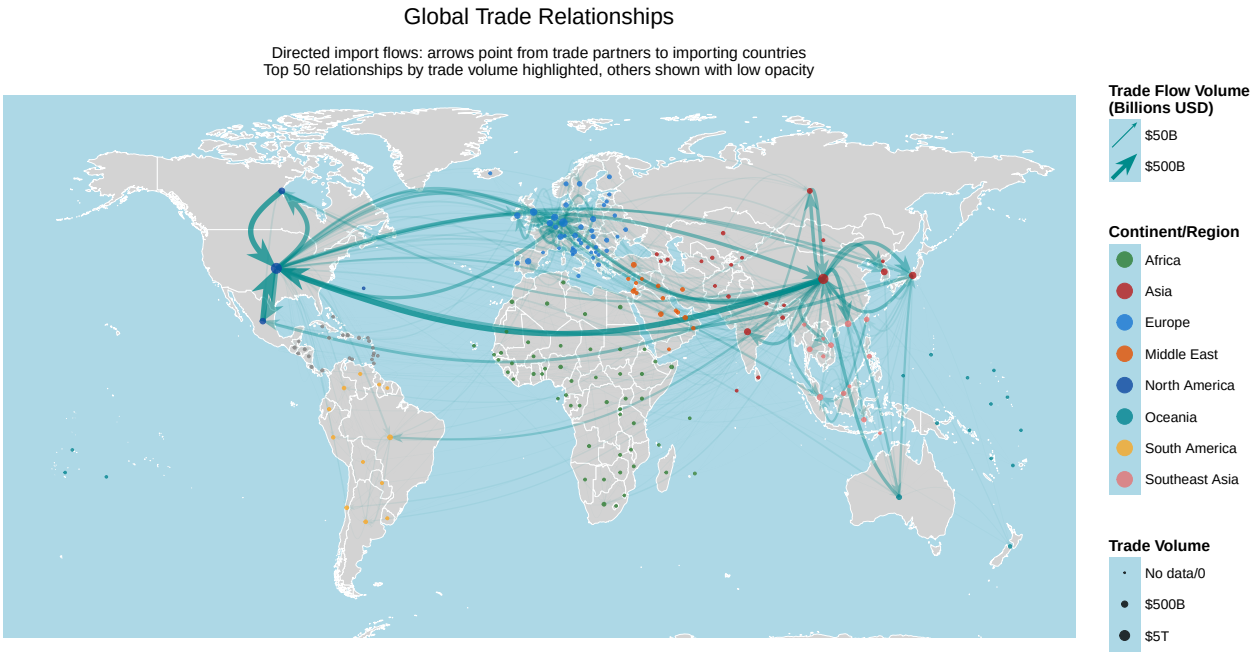
1.2.2 Top Importers Violin Plot



Next we look at the top importers. Note, that the y axis is log scaled. We can see that there is a few countries highlighted at the very top. Coincidentally these are also countries we have seen in the center of the network graph. The top importers are also those that are most relevant to understand the background of other countries. Also analougous to our analysis of country mentions, we can see that there are a few countries importing a lot and then a long tail of countries importing just a little. Chinas and Japans positions at the top of the distribution of importers are not reflected in their numbers of mentions, but are reflected in their position in the network graph. This could be another hint of the western centric nature of the dataset. It

is interesting to see that even though Japan and China have not been mentioned as much in the natural language descriptions of countries, there position in the network graph seems to have been self correcting due to their global trade relations.

1.2.3 Trade Flow Map



Finally we look at a flow map of the trade relationships. We can clearly see the trade hubs of the world: Europe, Asia and South-East Asia as well as North America. We can also see that the trade flows are not evenly distributed, but rather concentrated in a few countries. China and the US are the two countries with the largest trade flows. We can also clearly see the infamous trade deficit between the US and China. Even though trade is globalized and trade relations can be strong across continents, the majority of large trade flows are still locally between neighbouring countries. Long distance trades are primarily dominated by Europe, the US and China.

1.3 Conclusion

Todays world is highly interconnected and globalized. However, world regions still remain loosely grouped together by their importance to each other and the trade between neighbouring countries still plays an important role. Not all of the worlds regions are equally connected to each other. North America and Europe as well as Asia and South-East Asia stick out here as most globallized. Those regions are strongly connected to each other trough trade but also to the rest of the world. The Middle East, Africa, Oceania as well as South and Central America are the regions that are least connected to the rest of the world. The countries that are most relevant to the background of other countries tend to also be the ones most connected trough trade. These can be seen as the most important global players both historically as well as in todays world. As an additional note, we can see that either the influence from non-Western countries is not reflected in their historical and political impact, or the data is biased towards Western countries. We lean towards the latter conclusion, because Japan and especially China did historically have a strong influence on the world, and are also pulled toward the center of the network graph.

Question 2

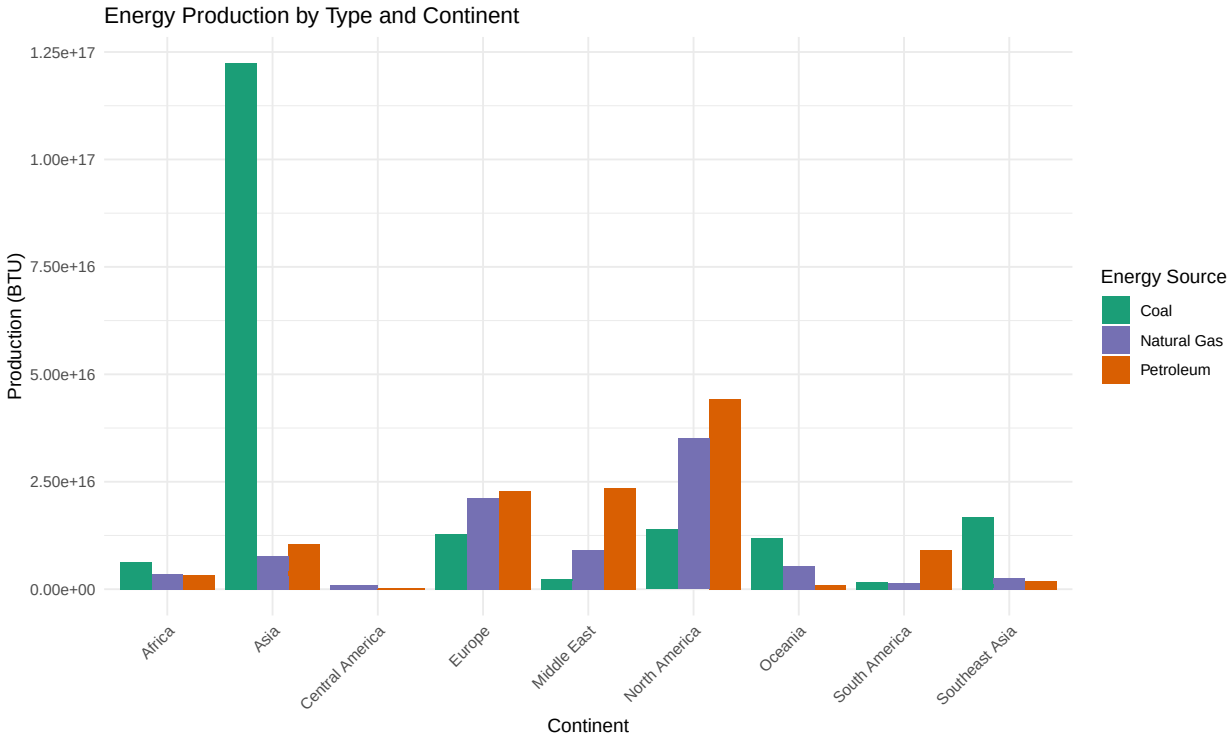
Introduction

In this section we explore the question, "How do countries' energy profiles impact their economy?"

Dirty Energy

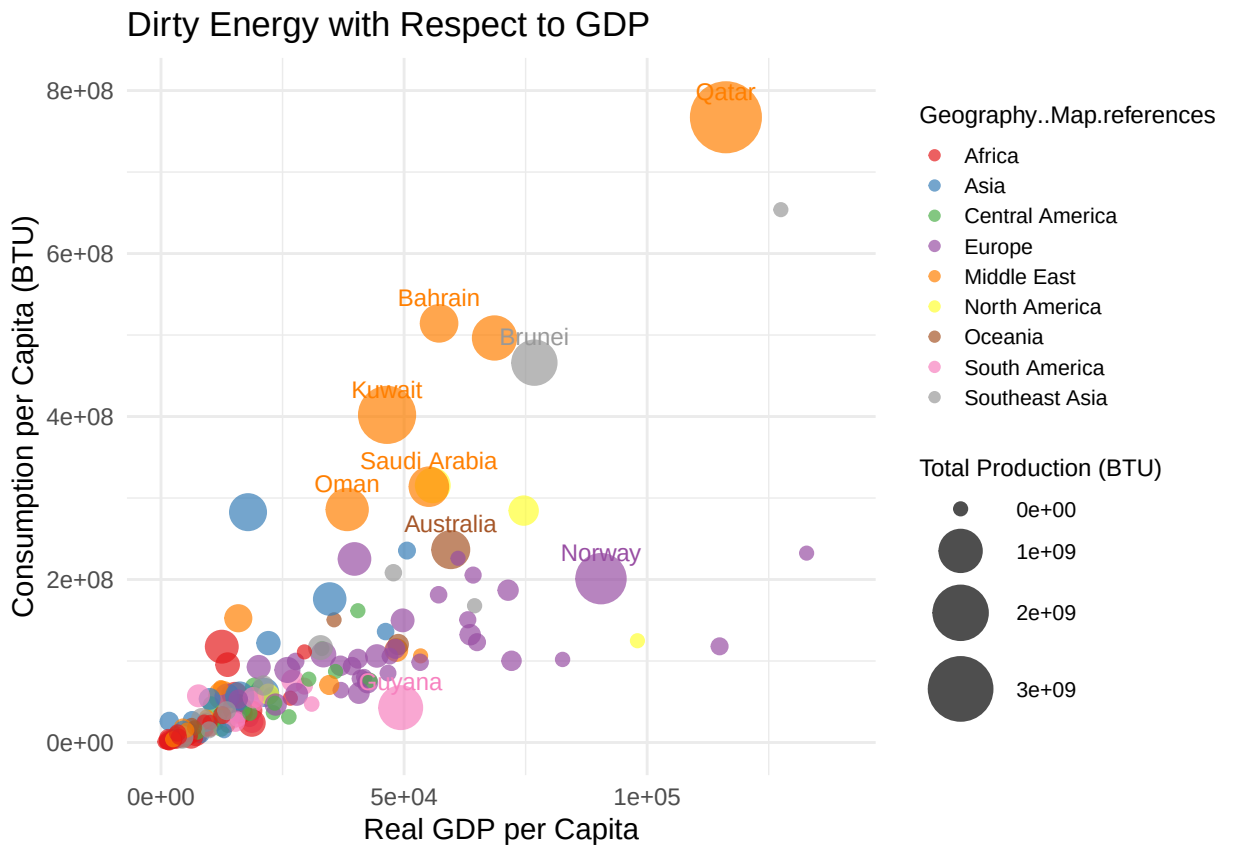
We first investigate countries energy profiles in relation to "dirty" energy. By "dirty" energy, we mean energy derived from petroleum, natural gas, and coal. We investigate both the amount of dirty energy that is produced and consumed. We then relate these characteristics to GDP per capita.

For data preparation, each source of dirty energy was measured in its own units. These were converted to BTU(British Thermal Unit). 1 BTU is equivalent to the energy in a single match. This allowed us to directly compare and aggregate the different types of energy. Note that gdp per capita in our data represents the real gdp calculated in US Dollars in 2023. By real gdp, we mean the purchase power parity.



First, we wanted to get a general overview of the amount of dirty energy produced. Here, we can see that Asia produces the majority of dirty energy by means of coal production. China, India are the top two producers of coal. North America is the largest producer of petroleum and natural gas. This is because The United States is the world's largest producer of both petroleum and natural gas. Meanwhile, Russia is the second largest producer of natural gas and petroleum. They are also the fifth largest producer of coal, and Russia accounts for the majority of dirty energy production in Europe. Also of note is the Middle East

which is the second largest producer of petroleum by continent. Lastly, we see that Africa, South America, and Southeast Asia produce little energy considering their size.

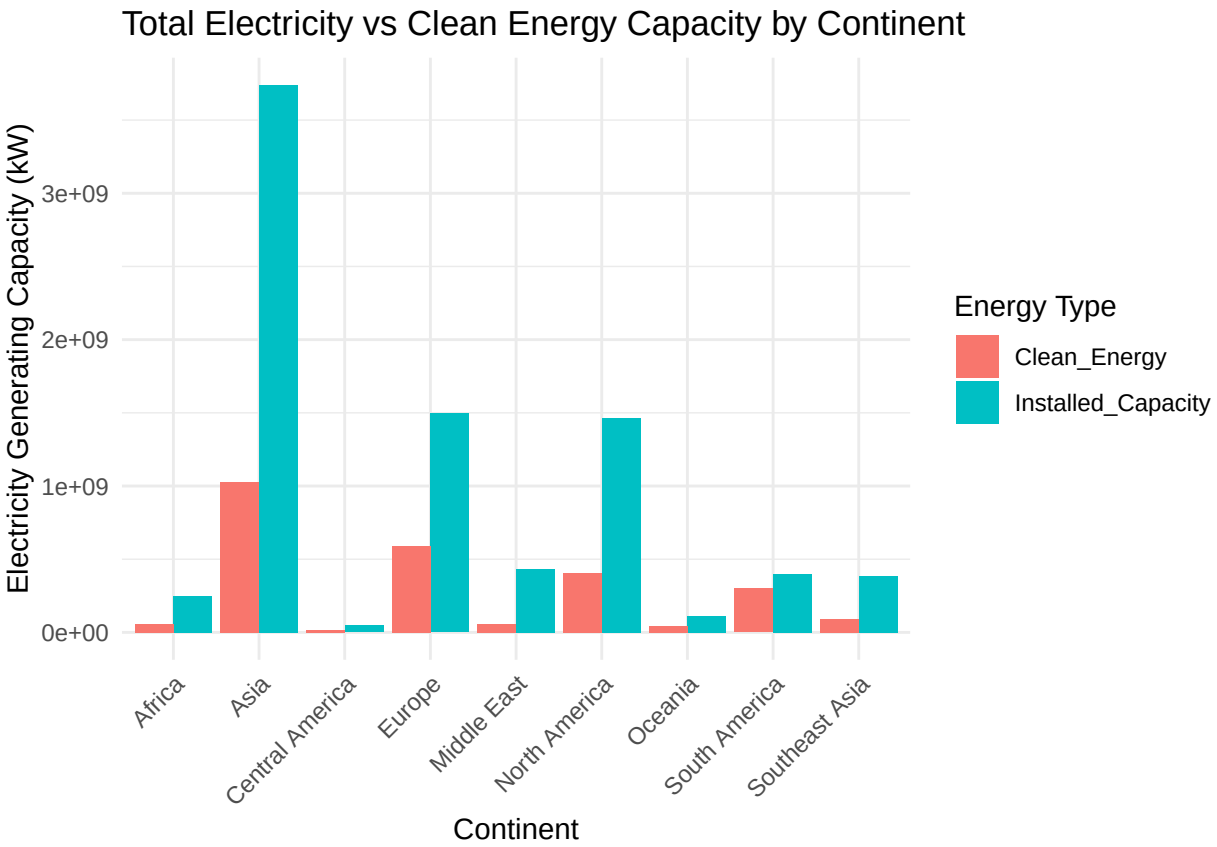


Next, we examine how these energy profiles relate to gdp per capita. Specifically, we look at how consumption changes in relation to gdp per capita, and we additionally observe the total production of each country based on the size of the bubbles. The graph has some linearity especially among countries who are not large producers. For large producers, we observe an additional trend. It seems that countries who produce alot of dirty energy also tend to consume alot. For example, we see the Turkmenistan has a relatively low gdp per capita, yet they are large consumers of dirty energy likely dut to their large production. We see two countries going against the trend. Norway is a relatively large producer of petroleum, yet they consume less than Turkmenistan. Finally, we note that Guyana consumes less than some other South American countries despite being the largest producer in the continent, prducing 100 times as much per capita compared to Columbia, the second largest producer.

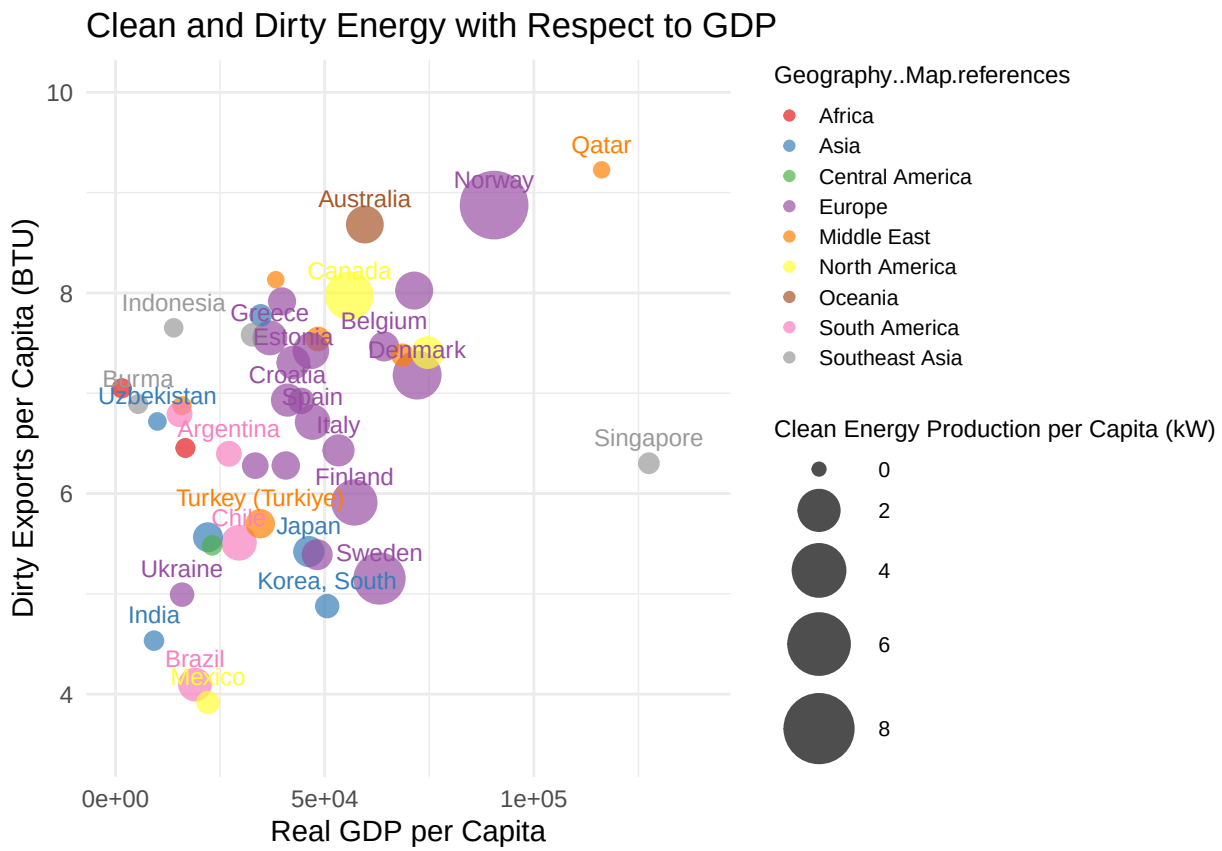
"Clean" Energy

We now consider the "clean" energy profiles of different countries. By clean, we mean energy produced by solar, hydroelectricity, nuclear, wind, biomass, geothermal, or tides and waves. Again, we also consider how this relates to gdp per capita. Additionally, we consider the amount of dirty energy imports and exports related to clean energy production.

Sources of clean energy were only given as percentages of a countries total electricity generating capacity in kW. This represents the amount of energy that can be instantaneously produced at any given time if all electricity generating sources in the country are operating at full capacity. The productive capacity from clean energy source was also converted to kW.



We first examined the total electricity generating capacity of continents compared to the total sources of clean energy production. We see that although Asia has the highest overall capacity, they are also the largest producer of clean energy. Europe and North America share strikingly similar profiles with Europe producing the second highest amount of clean energy in the world. Also of note is South America who produces the majority of their electricity through clean sources.



We now turn our attention to the connection to gdp percapita while also considering countries imports and exports. Note that the y-axis, Dirty Exports per Capita(BTU), is log scaled. Here we observe some correlation between gdp per capita and exports. This is likely because exporting dirty energy is a lucrative business. We also see that wealthier countries produce more clean energy in general. Maybe because the have more wealth to invest into new energy infrastructure. We also note that South American countries produce as much clean energy as countries with larger gdp. Norway is the largest producer of clean energy per capita, yet they are also one of the highest exporters of dirty energy. In general, we see that countries who are producing large amounts of clean energy may still be producing and exporting large amounts of dirty energy. Therefore, they are still contributing to the overall use of dirty energy worldwide.

Conclusion

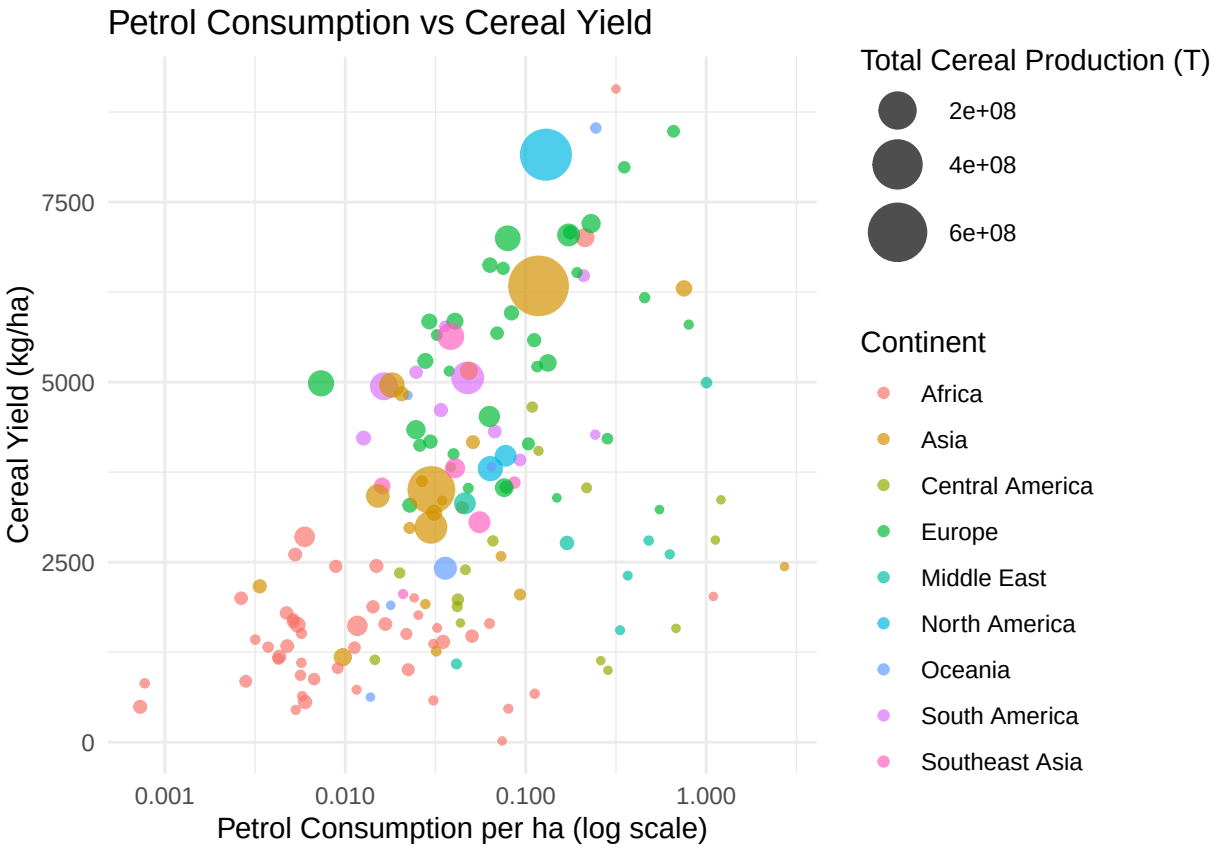
We conclude that a countries energy profile has a significant impact on their economy. Wealthier countries typically enjoy larger consumption rate. They are also better enabled to produce clean energy reducing dependence on fossil fuels. Countries who produce large amounts of dirty energy also see a positive benefit to their gdp, in part due to their ability to export their resources.

Question 3

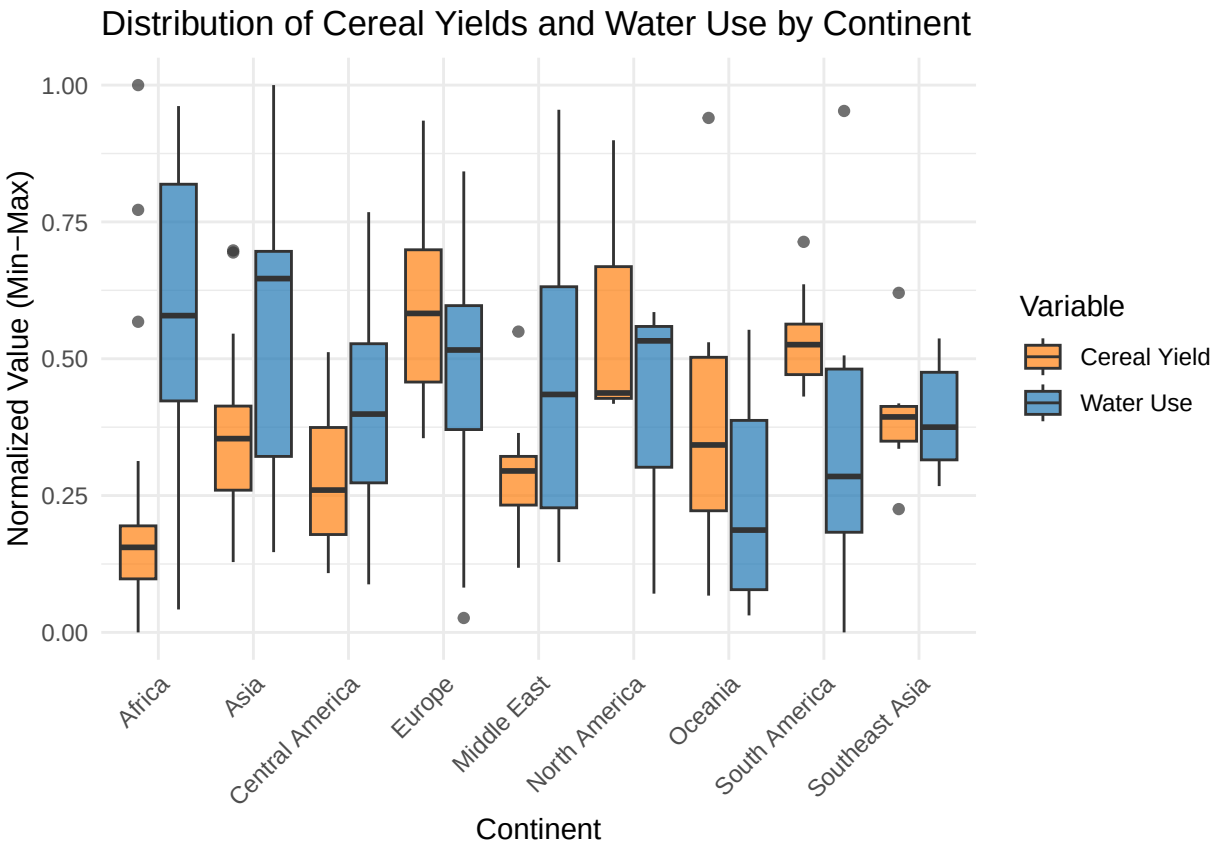
Introduction

We focus our attention on the question, "Can we identify global patterns in agricultural performance and energy-water usage among countries ?"

Data Exploration



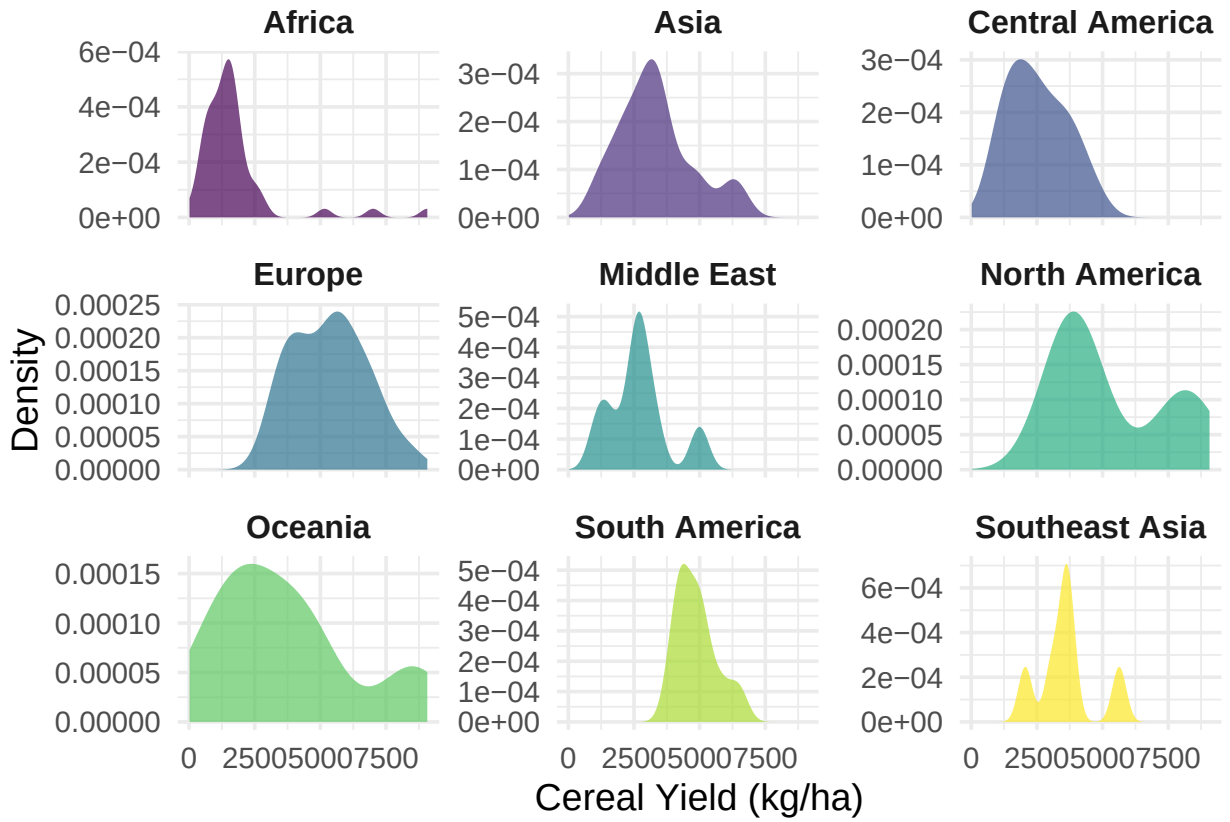
This first plot suggests a potential positive correlation between petrol consumption per hectare and cereal yield. Additionally, countries with higher total cereal production (indicated by the size of the points) also tend to exhibit higher petrol usage and yield levels. We can also observe clusters by continent, suggesting similarities within the same continent. A plausible hypothesis emerging from this visualization is that countries with higher energy inputs may achieve more productive agricultural systems.



These box plots represent the distribution of cereal yields (in orange) and the distribution of water consumption (in blue) for all continents. These values have been normalized between 0 and 1 for comparison purposes.

From an agronomic point of view, cereals are very water-intensive crops, and agriculture is one of the activities that exploits this resource the most. This graph seems to show a slight correlation between these two parameters. However, in Africa, the opposite situation can be observed. This graph seems to confirm agronomic knowledge by suggesting that a country's water consumption enables it to increase its cereal yields.

However, we cannot rule out a confounding variable. For example, we can imagine that developed countries have a rich economy, partly produced by strong agriculture. These same countries have developed water supply systems for industry and residents.



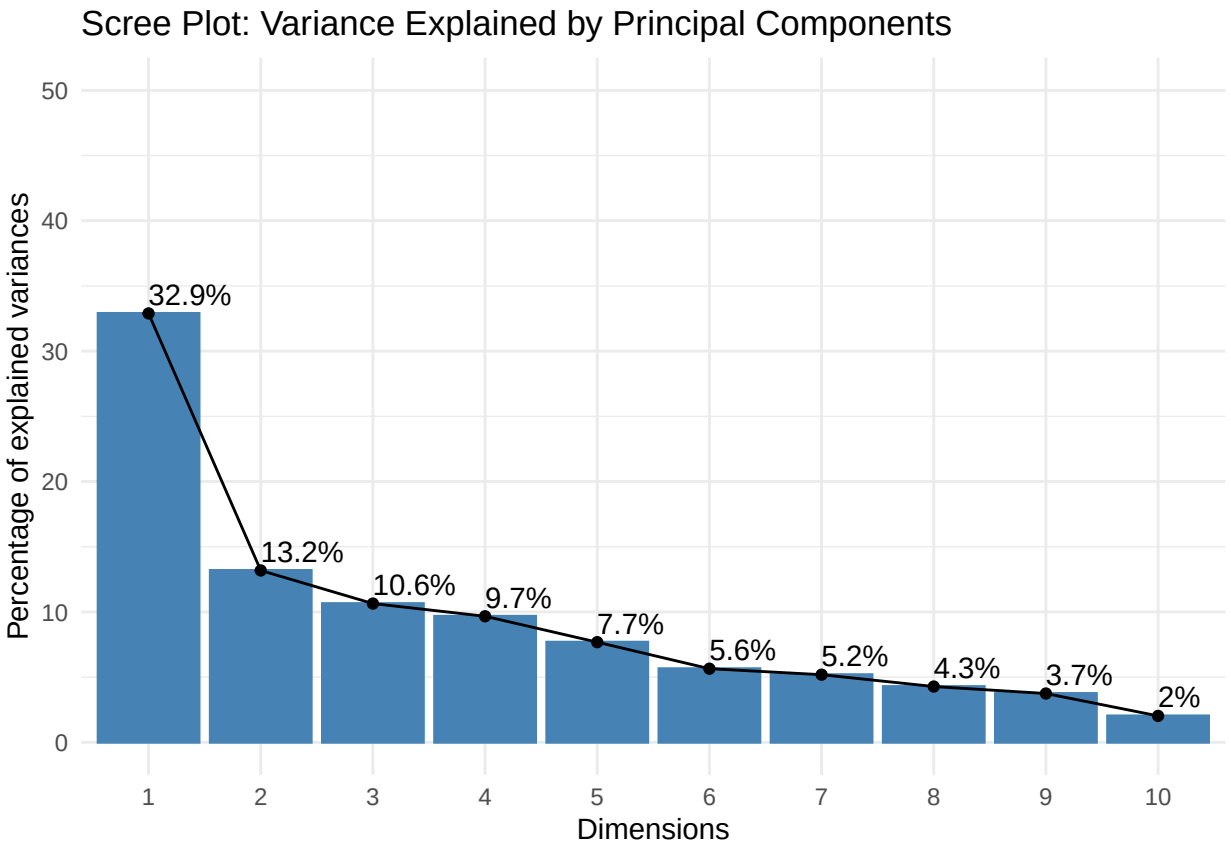
This graph also shows the distribution of cereal yields for each continent, but in a more detailed way by showing its distribution.

It allows us to see significant differences in productivity between continents. Africa appears to be the least productive, while Europe and North America are the most productive. The study of modes allows us to observe continental within differences. For example, European yields do not show a clear mode, suggesting uniformity of production between European countries. In Southeast Asia, there appear to be three yield groups. For North America, consisting solely of Canada and the United States, these two modes are clearly observable.

Principal Component Analysis (PCA)

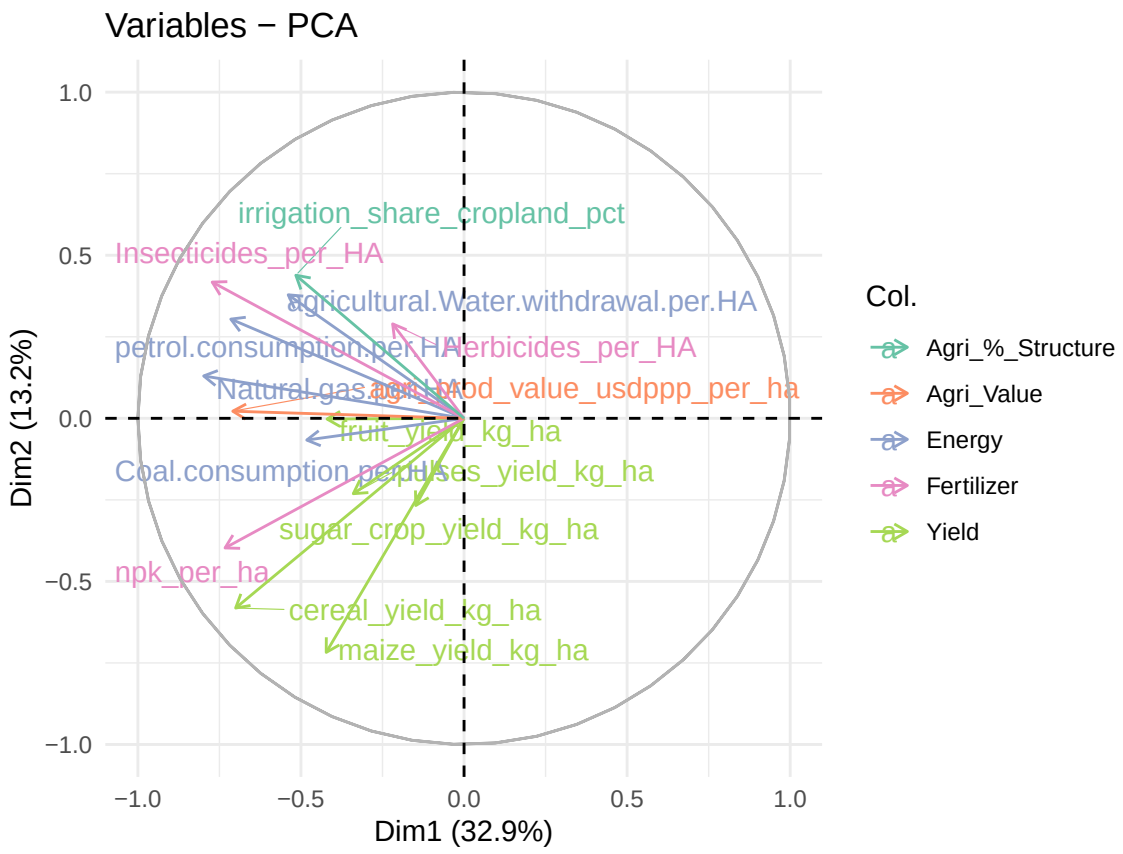
In this analysis, we performed a Principal Component Analysis (PCA) on a set of standardized (per hectare of arable land) agricultural, environmental, and input-related variables across countries. The variables include several yield measures (e.g., cereal, maize, fruit, pulses, sugar crops), energy consumption per hectare (coal, petrol, natural gas), agricultural water withdrawal per hectare, fertilizer input (NPK, herbicides, insecticides), agricultural production value per hectare, and the share of irrigated cropland.

These variables were grouped into thematic categories—Yield, Energy, Fertilizer, Water, Agricultural Structure, Value, and Production—to better understand the multidimensional structure of agricultural efficiency and resource use across countries.

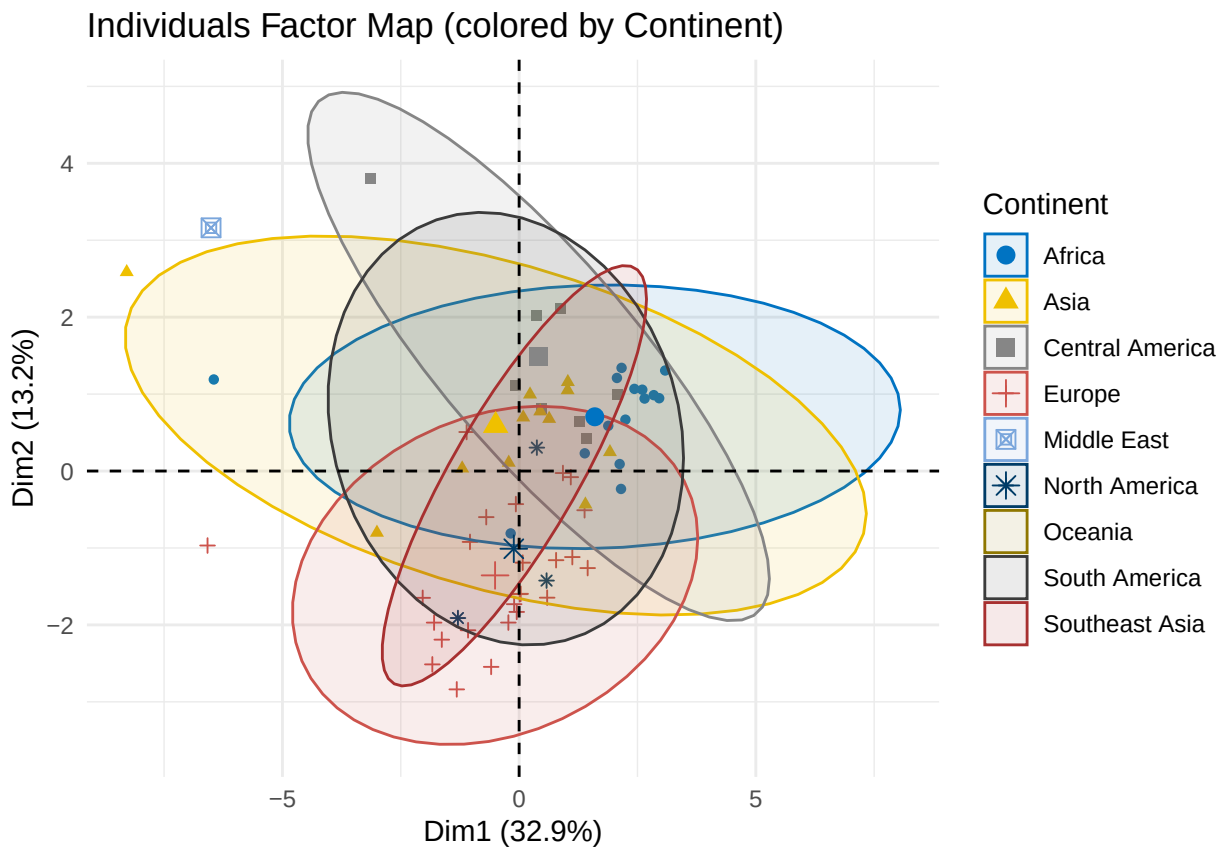


The scree plot below shows the percentage of total variance explained by each principal component. The first two principal components explain approximately 33% and 13% of the variance, respectively, accounting for nearly 46% of the total variability.

While the elbow in the scree plot is not very pronounced (with a slight inflection point after the first component), we proceed by focusing on the first two dimensions to explore the main patterns in the data.



This plot shows how each variable contributes to the two main principal components. The direction and length of each arrow represent the contribution and correlation of the variable with the components. Variables related to agricultural yields cluster in the bottom-left quadrant, suggesting they contribute strongly to Dimension 2. In contrast, variables associated with energy consumption (coal, petrol, gas) are mostly orthogonal to the yield variables and point toward the upper-left quadrant, indicating a strong contribution to Dimension 1. This suggests that yield and energy consumption are not strongly correlated in this dataset. Similarly, most input-related variables (insecticides, herbicides) align with energy variables, with the exception of NPK fertilizers, which point in the same direction as the yield variables. This may indicate that NPK use is more closely associated with crop productivity than other inputs. Interestingly, water consumption variables align with energy usage rather than with yields, suggesting that higher water withdrawal does not necessarily translate into higher agricultural performance at a global scale. The variable representing the average agricultural production value per hectare (USD PPP) points in an intermediate direction. A reasonable interpretation is that energy consumption is more reflective of general development level (e.g., high energy use in oil-exporting countries with low crop yields), rather than of agricultural performance. In contrast, countries with developed agricultural systems and access to irrigation might achieve both higher yield and higher value-added production per hectare. Hence, the average production value captures both productivity and the diversity of economically valuable crops.



Bottom-left quadrant (−Dim1, −Dim2):

Characterized by high agricultural yields, intensive NPK fertilizer use, and moderate to lower energy/water consumption.

Dominated by European and Southeast Asian countries.

This quadrant may represent efficient and productive agricultural systems, where outputs are achieved with relatively optimized input use.

Upper-left quadrant (−Dim1, +Dim2):

Associated with high energy and water consumption, but lower agricultural yields.

Includes some countries from North America and Oceania.

This could reflect input-heavy systems with lower overall efficiency or different agricultural priorities (e.g., livestock over crops).

Right side of the map (+Dim1):

Countries in this region exhibit lower energy and water use, though their yield or input profiles are more mixed. Several African and Asian countries cluster here. This may reflect less resource-intensive agriculture, either due to structural constraints or different development levels.

Center of the map:

Countries from Central America, the Middle East, and others are dispersed around the origin, indicating no strong association with either axis. These may represent diverse or transitional profiles, balancing input use and productivity.

Agricultural Value per Hectare: A Crosscutting Variable

The variable representing the average agricultural production value per hectare (USD PPP) appears to point in an intermediate direction—neither fully aligned with input intensity nor with yield exclusively. This suggests that:

Value creation may result from a combination of yield and input optimization, rather than from raw input intensity alone.Countries with both moderate yields and access to water/energy infrastructure (e.g., some European nations) may achieve better economic valorization of their land.

Caution and Final Remarks

While the PCA helps differentiate agricultural input-output profiles, several caveats apply:

Energy and water consumption do not directly correlate with higher performance globally; contextual factors (e.g., climate, technology, export orientation) likely moderate these effects.The overlap of countries from different continents indicates high internal heterogeneity, suggesting that development models and agricultural strategies vary widely within regions. Lastly, the PCA simplifies multidimensional relationships into two axes—so non-linearities or interaction effects may be obscured.

Question 4

Introduction

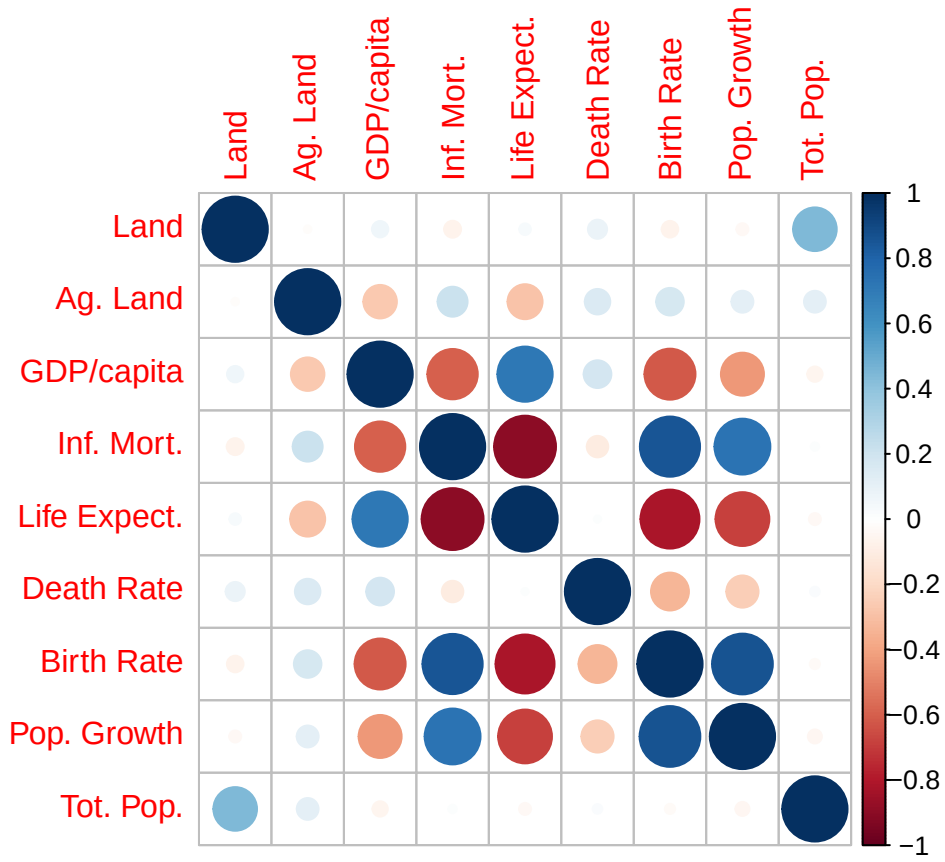
In this section, we focus on the question: "What do patterns in land use, demographics, and economics reveal about global inequality and regional development?"

We examine the variables: Total Land, Arable Land, Real GDP per capita USD, Infant Mortality Rate, Life Expectancy at Birth, Death Rate, Birth Rate, Population Growth Rate, and Total Population.

The techniques we use are: Corrplots, Biplots, Boxplots, and Scree Plots (Constituent of PCA).

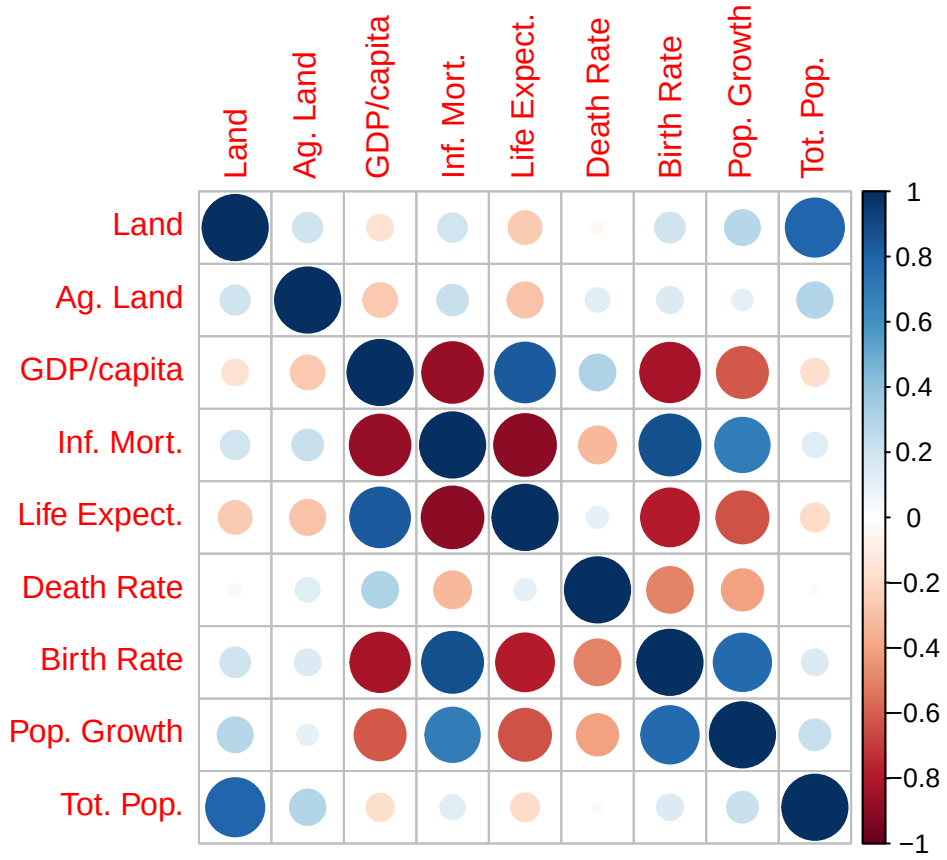
Correlation

In this section, we examine the correlation between numeric variables by means of a corrplot.



In this plot we see the Pearson correlation between the aforementioned numeric variables. Specifically, one can see that land, agricultural land, death rate, and total population have little correlation with other variables whereas the remaining variables seem to have complex correlation patterns among themselves. In this plot, one sees a strong negative correlation between life expectancy and infant mortality, a strong

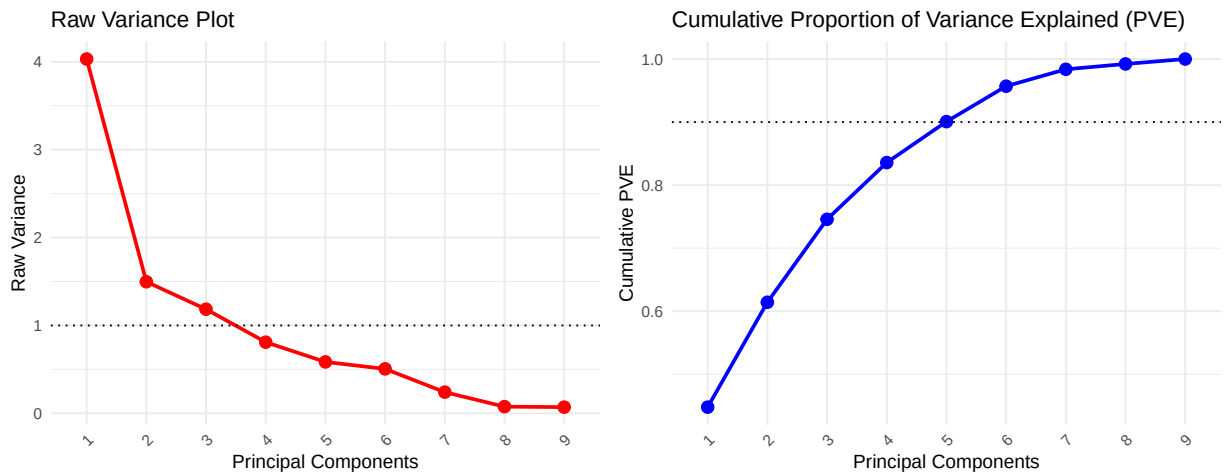
positive correlation between birth rate and population growth, a strong positive correlation between birth rate and infant mortality rate, and a strong negative correlation between birth rate and life expectancy. These correlations are linear and could perhaps be dependent on an underlying phenomenon such as access to medical care.



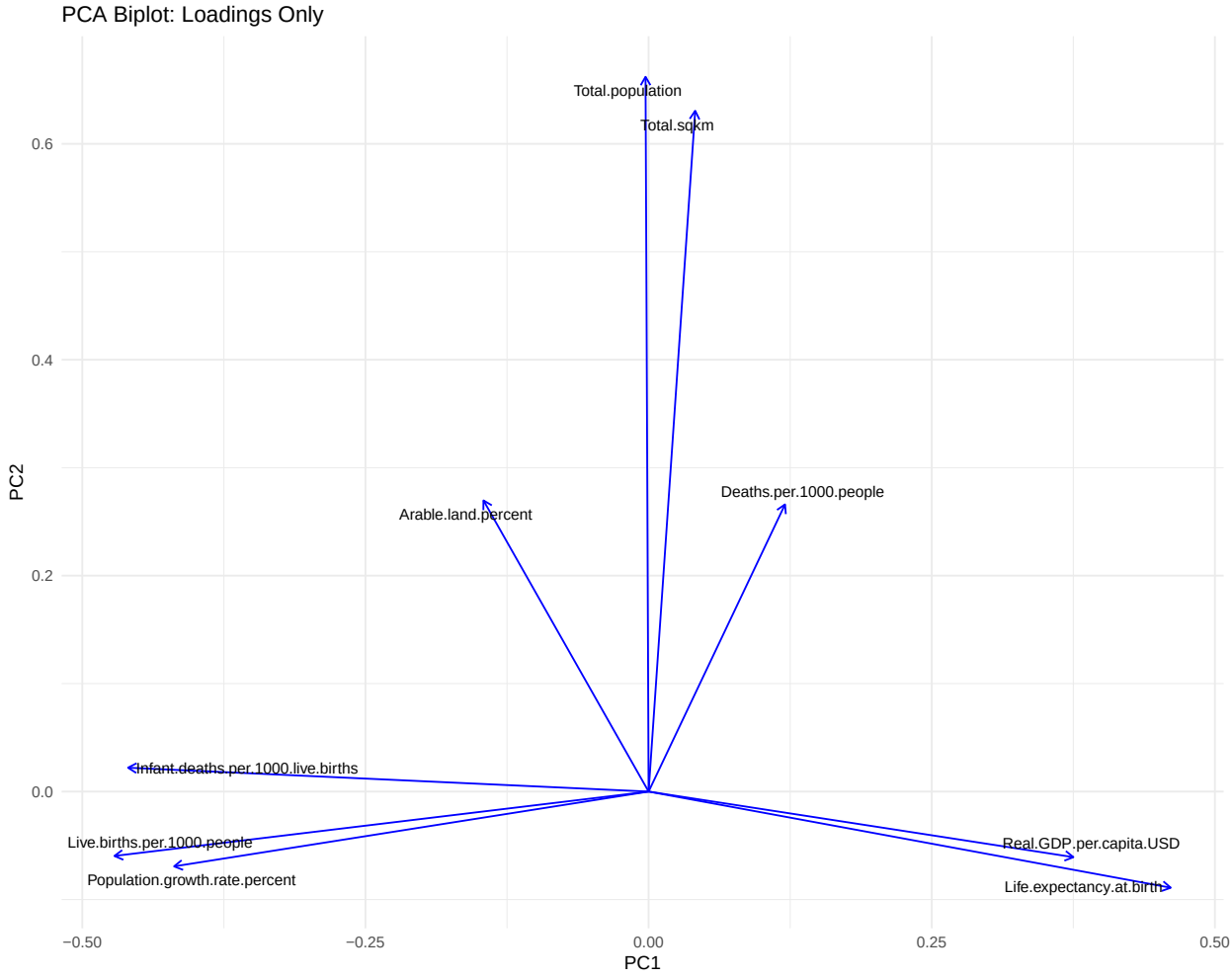
In this plot we see the Spearman correlation between the numeric variables. Although similar to the above plot, one can see higher positive correlation between land area and total population, suggesting that there is a monotonic relationship between the two variables though perhaps nonlinear. One can also see a stronger negative correlation between infant mortality and real GDP per captia, again suggesting that there is a strong, but non-linear relationship. There appears to be a weaker positive correlation between birth rate and population growth, perhaps suggesting that this is a linear relationship.

Principal Components Analysis

In this section we do principal components analysis on the same variables as in `Correlation`. We begin the analysis by looking at the variance contained in each of the principal components, then we examine the principal component loading vectors and formulate interpretations of each of the first two principal components based on the loadings of each original variable. Finally, we look at the scores of the observations and examine per-continent trends based on our interpretation of the first two principal components.



There are 3 principal components that "aggregate information" due to the variance being larger than 1 while the PCA was run on scaled data. Additionally, to capture ~90% of the variance, we need 5 PCs, and to capture ~75% of the variance, we need only 3 PCs.

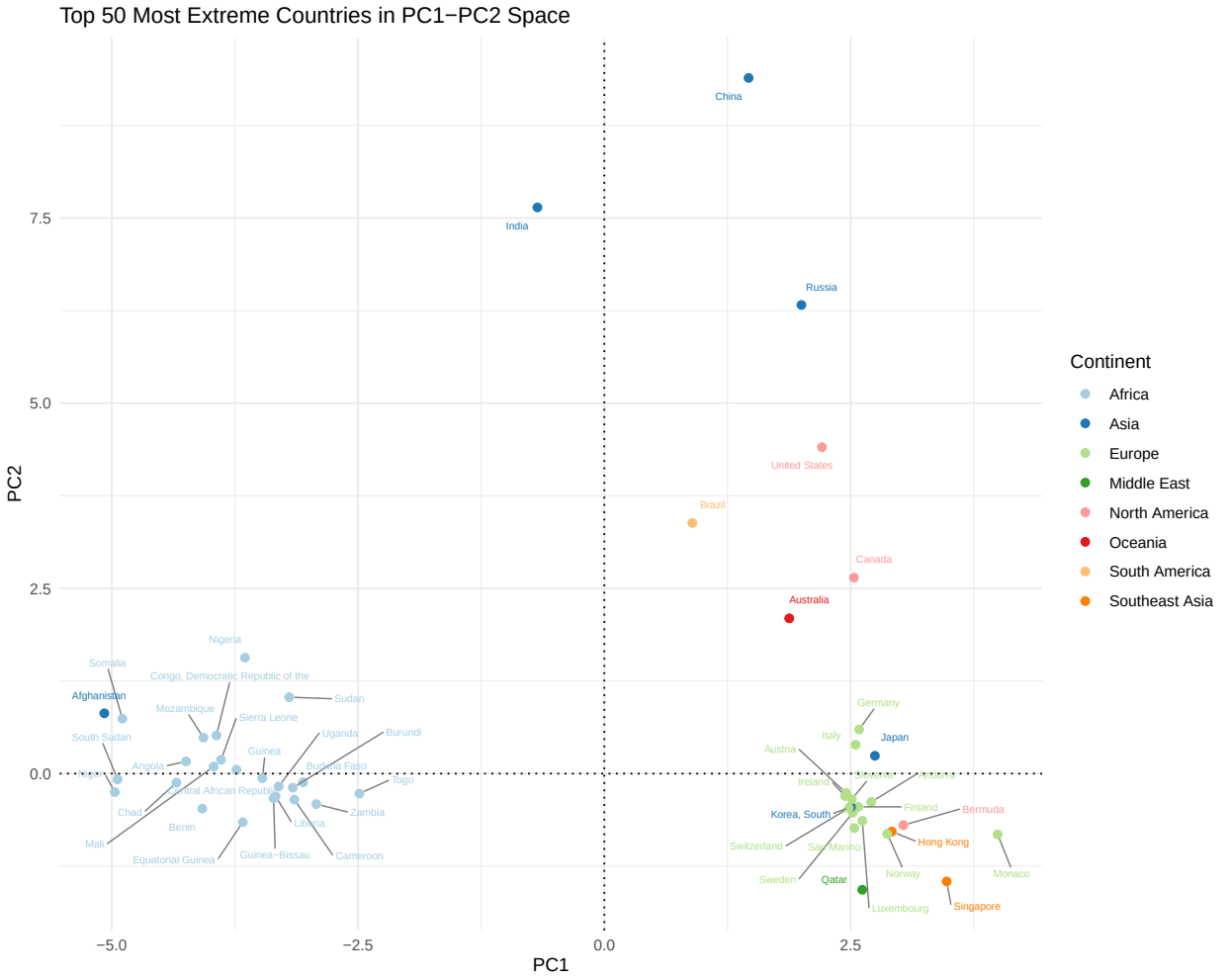


In the above biplot of the principal component loading vectors we make the following observations:

One can see that for small values of PC1 we see large values of Infant Mortality Rate, Birth Rate, and

Population Growth Rate. For large values of PC1 we see large values of Real GDP per capita USD and Life Expectancy at Birth. One could interpret PC1 as the developedness of countries with larger values corresponding to more developed countries.

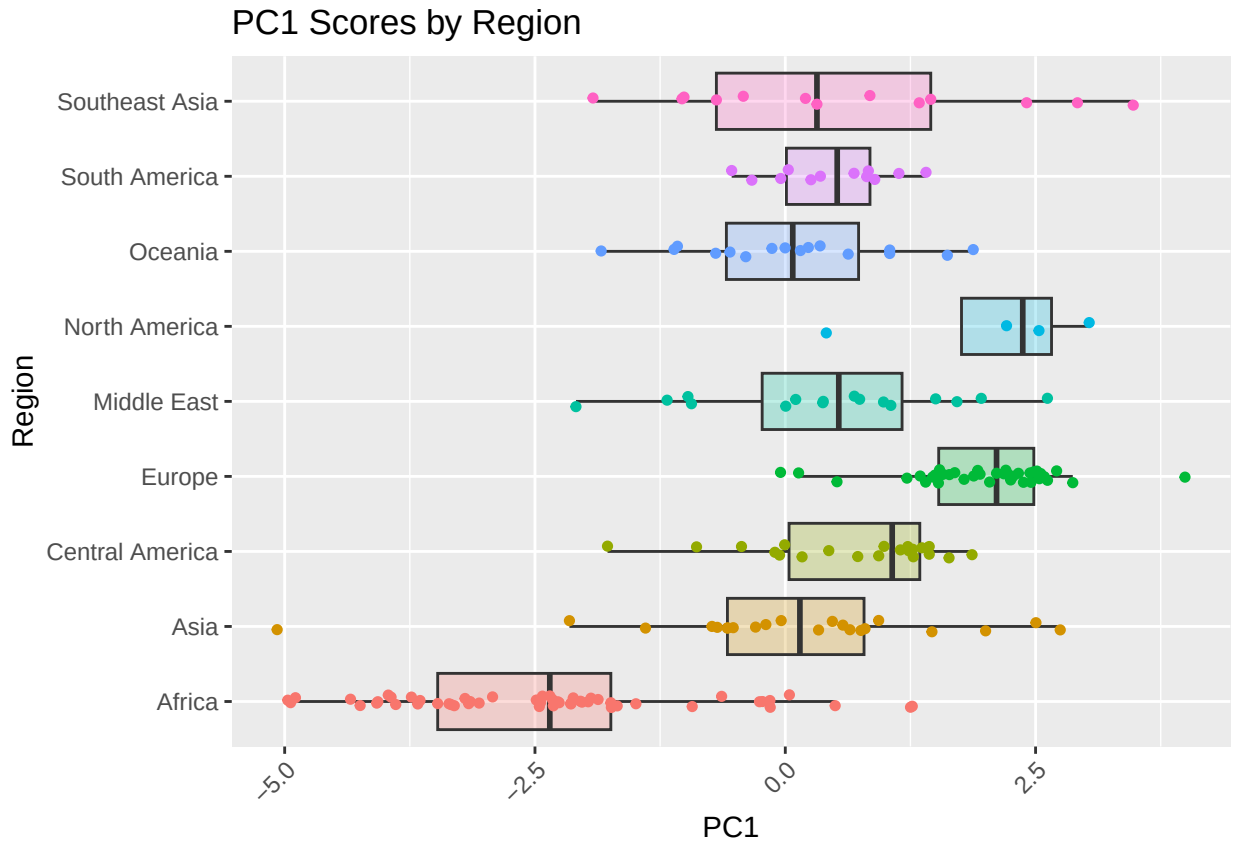
For large values of PC2 we see that the countries typically have large values of Total Poplation and Total SqKm. To a lesser degree Arable Land and Death Rate also contribute to the PC2 direction. An interpretation of PC2 could be the size of a country in terms of population and area.



In the scores plot we plot the scores of the 50 countries farthest from the origin. We also make the following observations and conclusions:

We see groupings along the PC1 axis of more developed countries on the side with larger PC1 values and less developed countries with smaller PC1 values. There are several outliers with large values for PC2 that correspond with larger countries in terms of both population and land mass. Specifically, countries from Africa tend to have lower values of PC1 whereas countries from North America, Europe, and Southeast Asia tend to have larger values for PC1. Additionally, it appears that strong outliers account for the majority of the variance in PC2 as we see India, China, Russia, and the United States with the largest values whereas most other countries group near 0 with less variation.

Examination of PC's by Continent



In these boxplots, one can see that Africa tends to have smaller values for the PC1 with the lowest median. This indicates less developedness by our earlier interpretation. North America has the highest median. The country with the highest value for PC1 is from Europe (Monaco, from above) and the country with the lowest value is from Asia (Afghanistan, from above).

Conclusion

Through this analysis, we sought to answer the question: "What do patterns in land use, demographics, and economics reveal about global inequality and regional development?" To do so, we examined linear and monotonic correlation between the variables themselves with Corrplots and continent/country-specific information, aggregated with PCA.

The main takeaways are as follows: The variables we examined tend to cluster into a few groups such as "size" of countries, "wealth" of countries, and "poorness" of countries. There are only a few "large" countries in terms of land mass and population, namely India, China, Russia, and the United States. Based on the variables examined and our interpretation of the Principal Components, Africa tends to be less developed than the others. From this analysis and the geographic grouping seen in the anaylsis, one could speculate that the geographic location (among other factors such as colonialization) plays a large role in the developedness of nations.

Question 5

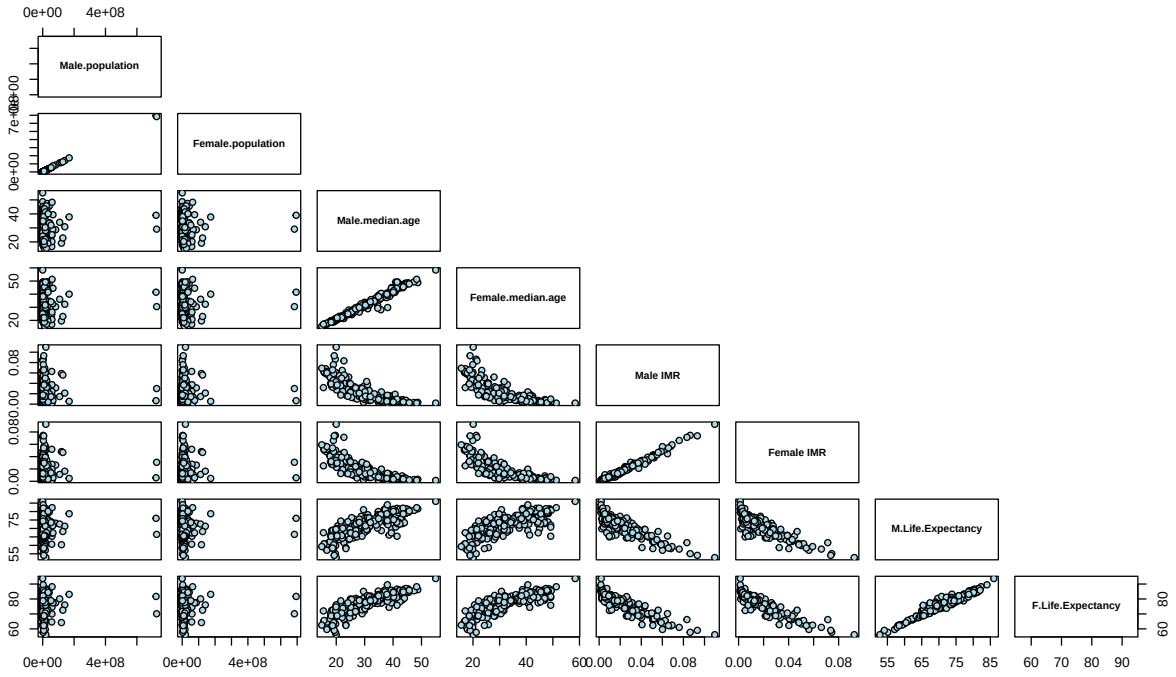
Introduction

In this section we seek to answer the question: "How do selected demographic variables differ between males and females?"

The variables we examine are: Population, Life Expectancy at Birth, Infant Mortality Rate, and Median Age. These are all measured for both males and females.

The techniques we use are Pairs Plots, Kernel Density Estimation + Rug Plots, and Choropleths.

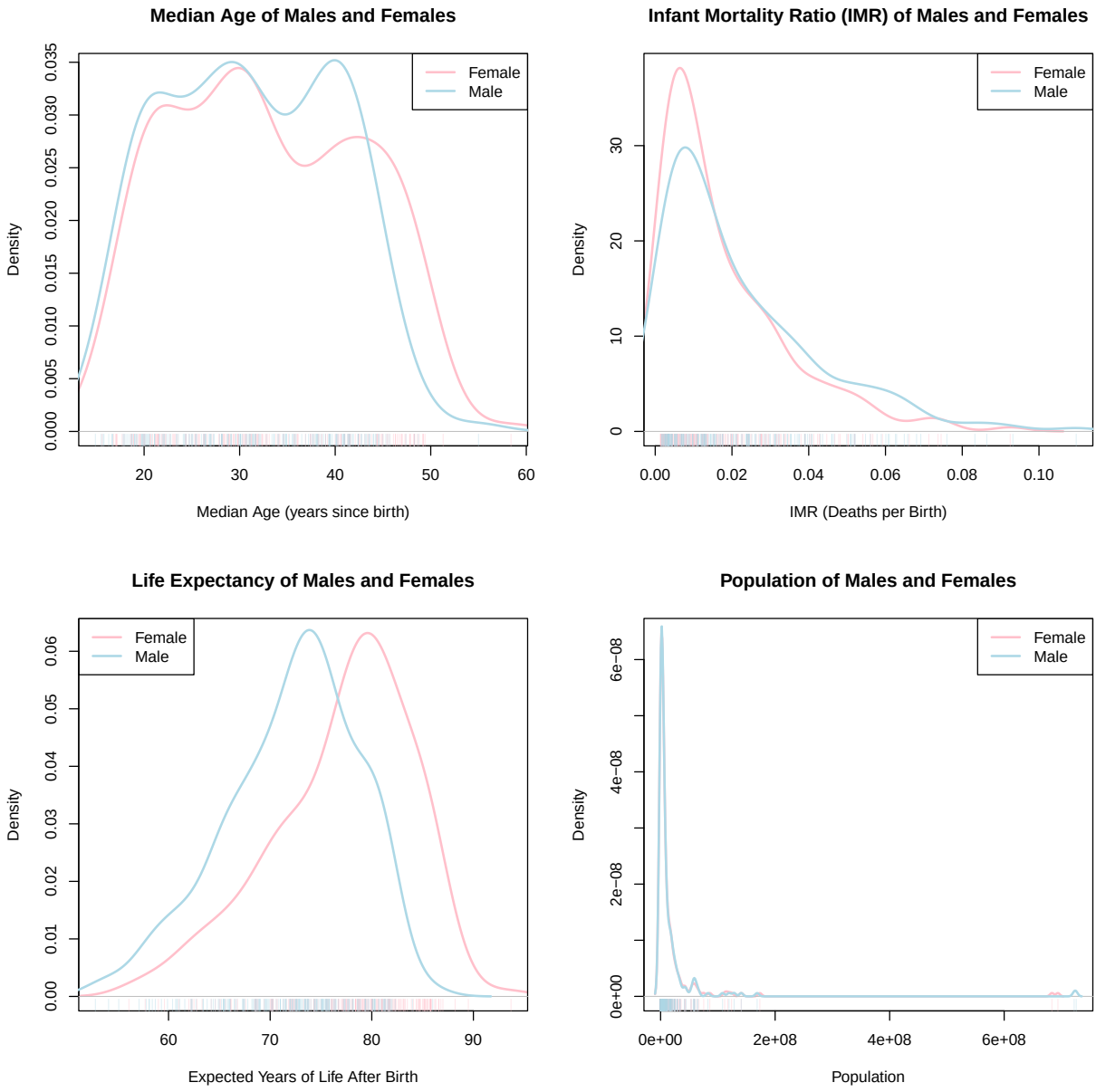
Correlation



In these plots, one can see a strong, positive linear relationship between the male and female values for each of the variables: population, median age, infant mortality rate, and life expectancy per country. Because of this, we focus on these variables for the remainder of the analysis.

Univariate Statistical Graphics

Kernel Density Estimates (KDE)



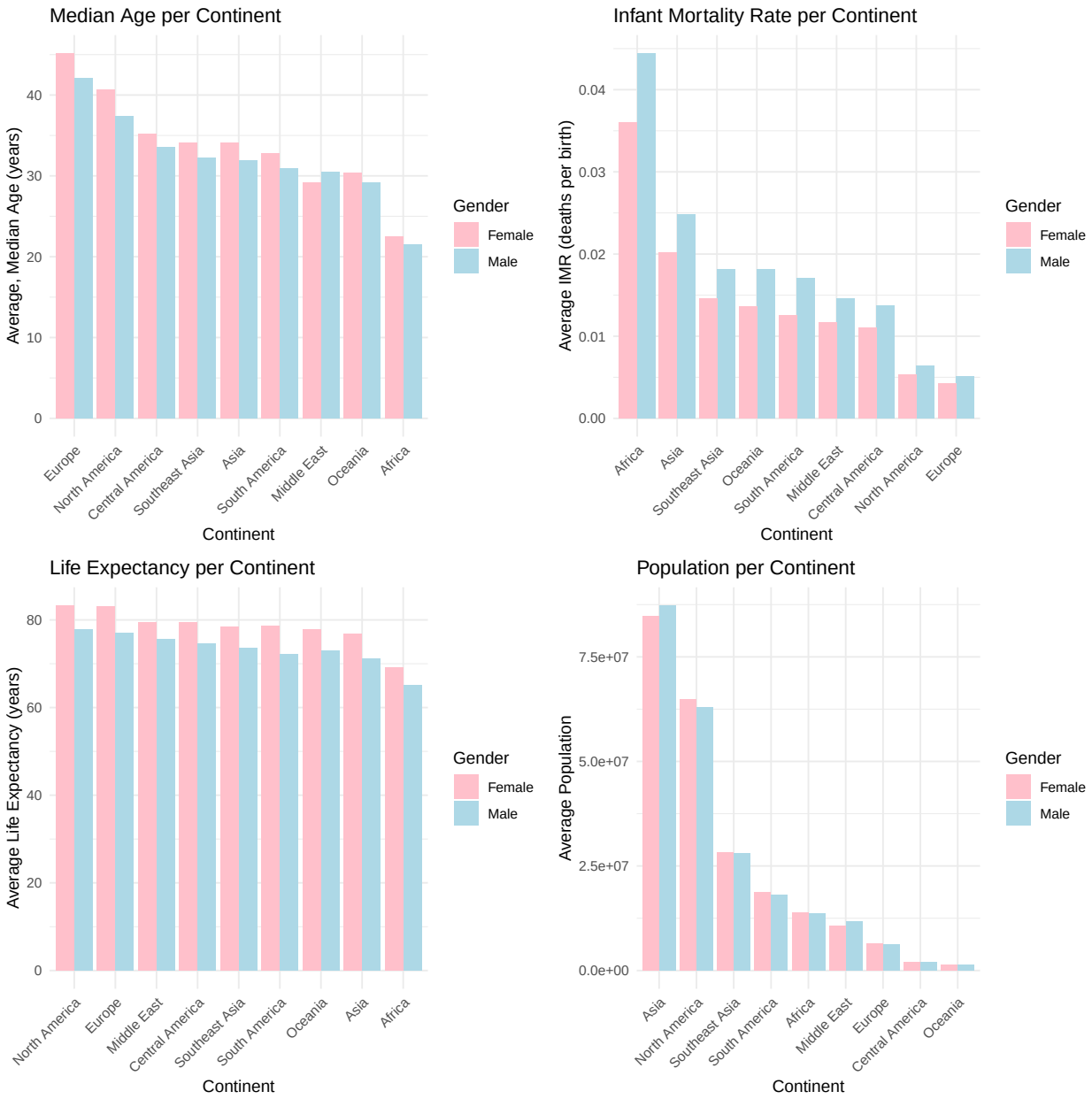
These graphics allow us to understand the distribution of statistics and variables from various countries. Specifically, we can see the relative shape, scale, and location of the distributions corresponding to males and females.

- **Top-Left:** It appears that the male density is slightly higher until approximately age 32 where the densities of males and females are approximately equal. From approximately age 33 to 45 the male density is larger. From approximately 45 and onwards, the female density is larger.
- **Top-Right:** The female density is larger at lower values of IMR, until about 0.02 where the male density is larger. This means that there is tendency for the female IMR in the examined countries to

be smaller.

- **Bottom-Left:** The shape of the male and female densities appear to be approximately equal though the male density is shifted slightly lower indicating a lower male life expectancy in many countries examined.
- **Bottom-Right:** The densities for male and female are approximately equal. Still, using the rug plot, one can see a shift in the right tail of the distribution with a smaller population of women present in the outlying countries, India and China. This could be due to preference for male children in both of these societies.

Grouped Bar Charts

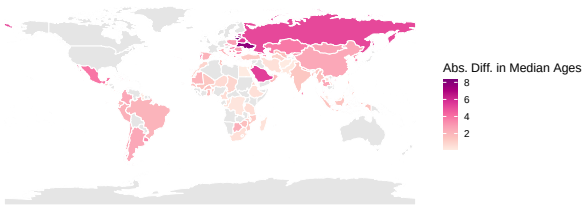


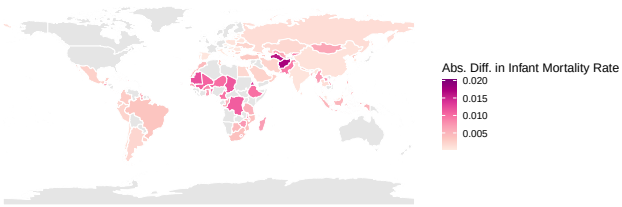
These graphics allow comparison of per-continent aggregates so one may glean various insights such as the ordering which different continents follow.

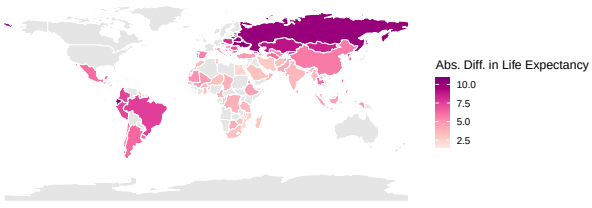
- **Top-Left:** One sees an ordering of average median ages by continent/region. Europe has the population with the highest median age while Africa has the population with the lowest median age. In each region except the Middle East, the median age of females is higher than that of males. Note that the collected statistic was the median age and we took the average of these.
- **Top-Right:** One sees an ordering of average infant mortality rate by continent/region. Africa has the highest infant mortality rate while Europe has the smallest. The ordering is the same for both males and females though males tend to have a higher infant mortality rate than females.
- **Bottom-Left:** One sees an ordering of average life expectancy by continent/region. North America has the highest average life expectancy and Africa has the lowest. Female life expectancy, as seen in the KDE plots tends to be higher than male life expectancy.
- **Bottom-Right:** One sees an ordering of average population per continent/region. Asia has the highest average population and Oceania has the lowest. North America, South America, and Europe have more women than men whereas the other regions tend to have more men than women.

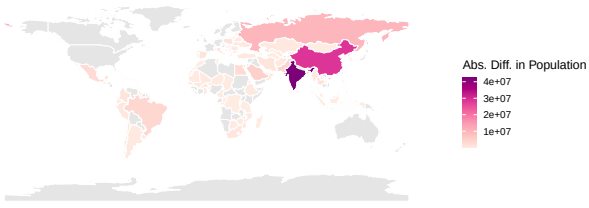
Choropleths

Here we present choropleths that display the absolute difference between the variables for each country. Note that many countries are gray since our dataset did not provide gender-wise information on these areas.









These graphics allow us to visualize geospatial trends in the data at a finer granularity than classical statistical graphics such as bar charts. Additionally, by using the absolute difference, we are more able to visualize inequality across the world. Finally, since we take the absolute value, we remove information concerning who the inequality effects as the absolute value function is not surjective.

- **Top-Left:** One sees that there is a large difference in median age in Eastern Europe, Russia, Saudi Arabia, and Mexico. There is a moderate difference in East Asia, South America, and parts of Africa. There is a smaller difference in the Middle East and other parts of Africa.
- **Top-Right:** One sees that there is a large difference in the infant mortality rate in the Middle East and parts of Africa. There is a smaller difference elsewhere.
- **Bottom-Left:** One sees that there is a large difference in life expectancy in Eastern Europe, Russia, Kazakhstan, Mongolia, and South America. In Africa and the Middle East, there tends to be a smaller difference.
- **Bottom-Right:** One sees that there is a large difference in population in India and China. There is a moderate difference in population in Russia, and smaller differences elsewhere. Note that we consider raw counts here, not proportions.

Conclusion

Through this analysis, we draw the following conclusions about the selected statistics and variables: Women tend to have longer life expectancy than men in general. On average countries tend to have older women than men. There are stark differences in the populations of men and women in China and India. Finally, men typically have a higher infant mortality rate, regardless of geographic region.

Conclusion

Throughout this report, we explored different aspects of global development and interaction using a wide set of visualization techniques. While the dataset has a noticeable US bias, it still provided valuable insight into global connections and disparities. We began by looking at international relationships, and observed that a small number of countries, most notably the US and UK, play a (disproportionately) large role in both. In our analysis of energy profiles, we found that wealthier nations tend to consume more and are better equipped to invest in clean energy, while energy-exporting countries benefit economically even with lower GDP per capita. In agriculture, we observed that higher resource input does not always correlate with higher yield, and regional differences suggest varying degrees of efficiency and development. Our analyses of demographics and land-use revealed clusters of countries along dimensions such as population size, economic output, and health indicators, with Africa generally displaying lower development scores. Finally, we investigated sex-based demographic differences and confirmed consistent global patterns, such as higher life expectancy and lower infant mortality rates for women. These analyses show how visualizations can reveal underlying structures, trends, and inequalities in the modern world.