Justin Wang

Data 230

Term Project

6 December 2023

**Table of Contents**

Abstract………………………………………………………………………………………….2

Initial Data Exploration…………………………………………………………………………2-3

ETL……………………………………………………………………………………………..3

Setting up the Problem…………………………………………………………………………4-5

First Looks……………………………………………………………………………………..5-7

Spending……………………………………………………………………………………….7-8

Accessibility……………………………………………………………………………………8-10

Affordability………………………………………………………………………………..…10-12

Air Quality……………………………………………………………………………………12-15

Interpretation…………………………………………………………………………………15-19

Conclusion……………………………………………………………………………………19-20

Dashboard Interaction……………………………………………………………………….20-26

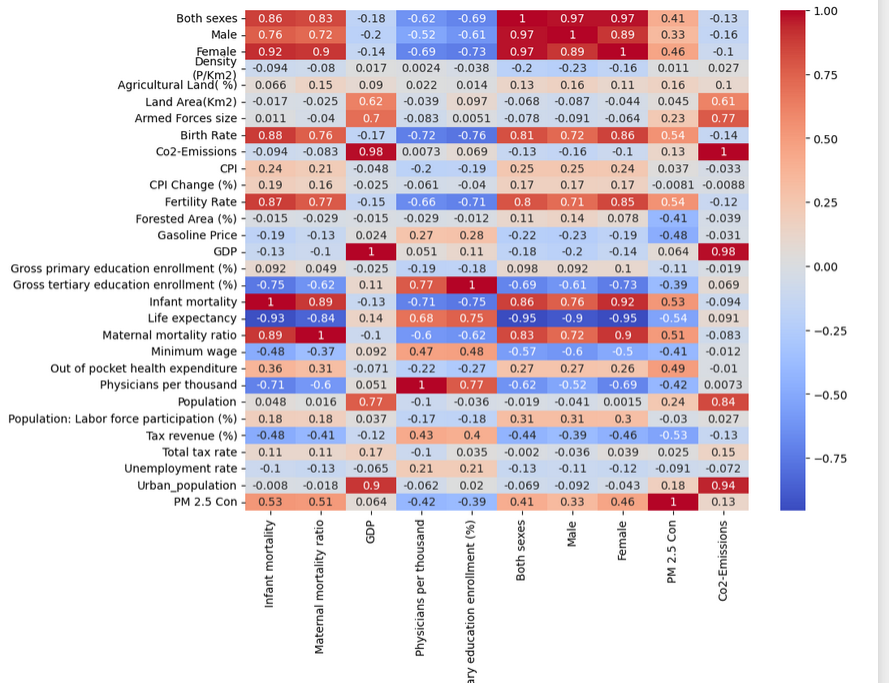
Sources………………………………………………………………………………………….26

Github…………………………………………………………………………………………..26

**Abstract**

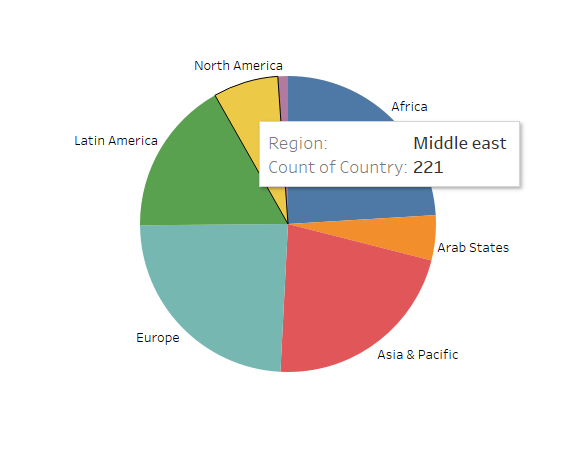
Regardless of what we do, mortality will always be hanging over our heads. While it cannot be avoided, it can be slowed down. This project explores some of the factors that may contribute to the mortality rate in an attempt to figure out ways to slow it down. I use data found from Kaggle as the basis for this research, as well as resources from the World Health Organization and the worldbank in order to find correlations between select factors and mortality rate. This report explores a few main factors: GDP/per Capita, physicians per thousand, tertiary enrollment ratio, minimum wage, out of pocket health expenditures, co2 emissions, PM 2.5 concentrations, and tax revenue. I found that there is correlation between mortality rates and each of these factors. Additionally, some of the factors are correlated with each other. As a result, it gives us an idea of some of the areas that can be focused on to decrease mortality rates.

**Initial Data Exploration**

For this project, a few different datasets were used. The main dataset was a dataset from Kaggle that looked at countries of the world. The second dataset explored air quality, namely pm 2.5 concentrations, for each country. The third dataset explored mortality rates by year for countries. The last dataset explored GDP per year by country. By combining some of these datasets, a correlation table was created to be used as the basis for the project.

This does not show every relationship, and not every relationship was used, but some of the relationships were used to try to find some of the reasons some countries and regions have higher mortality rates than others.

Additionally, there is a large discrepancy between the number of countries per region, so some of the data may be more skewed than others.

****

**ETL**

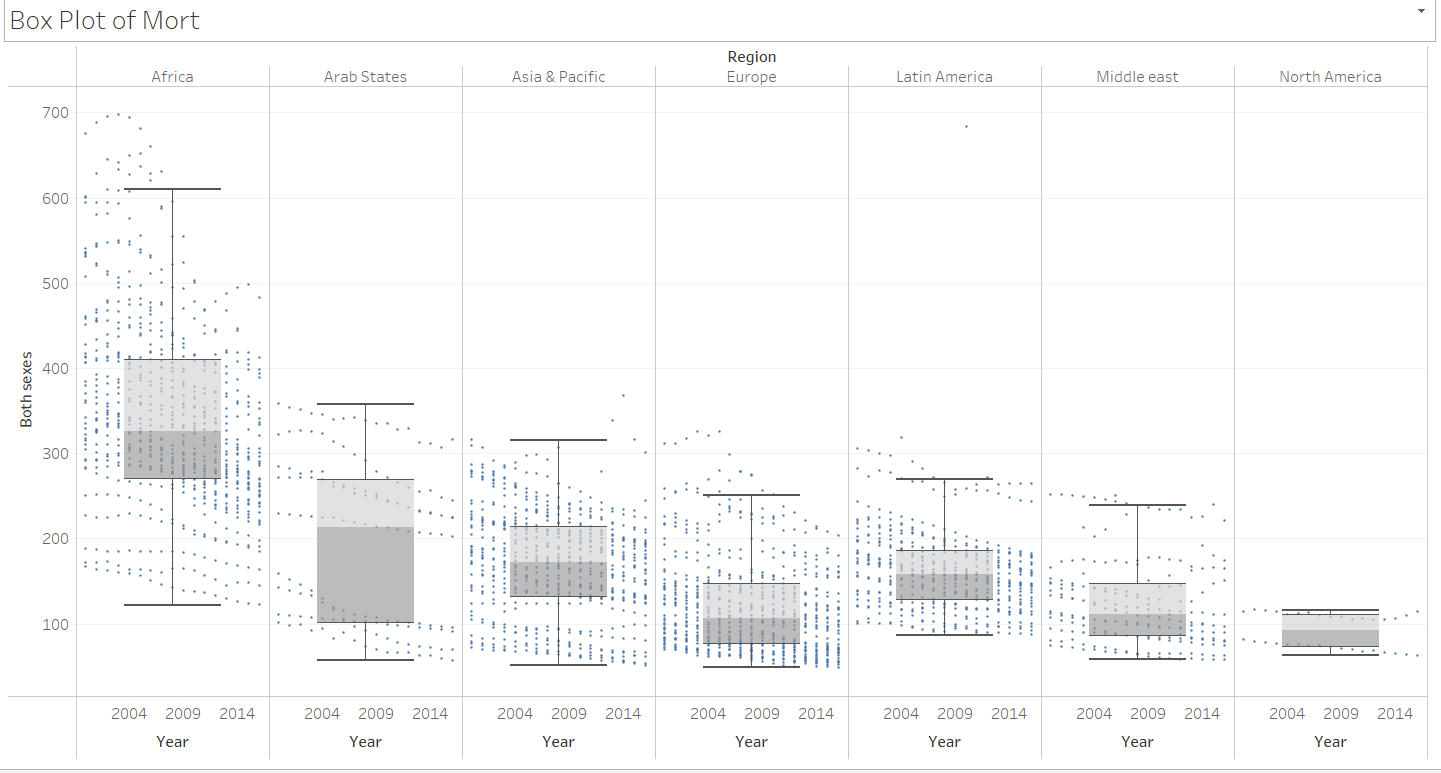
In terms of ETL tools, I primarily used Python and Excel. Some of the datasets listed above had missing/null values. To remedy this, I simply removed it, since I felt as if there was insufficient amounts of data to feasibly replace them. Additionally, many of the datasets did not exactly match each other, so when merging them together, some of the countries were skipped. This was mostly done with Python.

I had used Excel to add columns and edit formats. For example, I had added a “Region” column into a dataset for aggregation using the filter function in Excel and manually inserting the region names. Additionally, the date columns were not in the format that I had needed, so I had changed them as well.

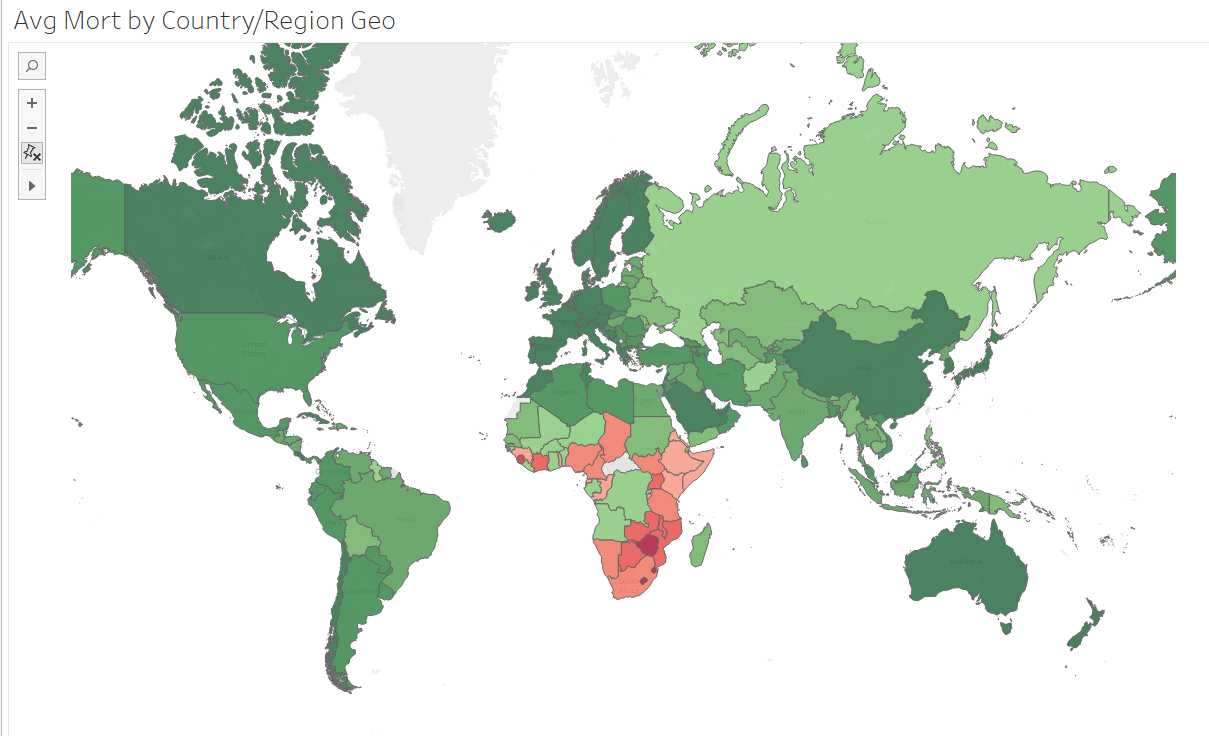
Also, some of the merging was done in Tableau itself.

**Setting up the Problem**

By visualizing the average mortality rate by region, it can be seen that most regions are within the 100 to 250 mortality rate range, with Africa towering above them at 350-400. One thing to note is that Africa’s 25% quartile is higher than every other region’s 75% quartile.



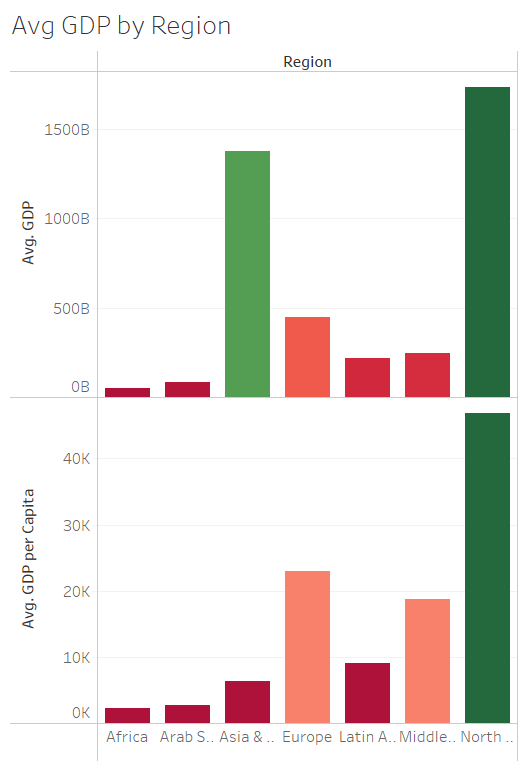
This can be further viewed in a symbol map with color based on the mortality rate, where it can be seen that the majority of African countries are in the red while almost every other country/region is a shade of green.



This puts into perspective how much higher the mortality rate in Africa is compared to other regions, which poses the question: Why is this so and what factors contribute to this difference? It also poses the question of how we can attempt to reduce mortality rates, not just in Africa, but overall. However, we will be using region statistics as well as their respective average mortality rates in order to find correlations.

**First Looks**

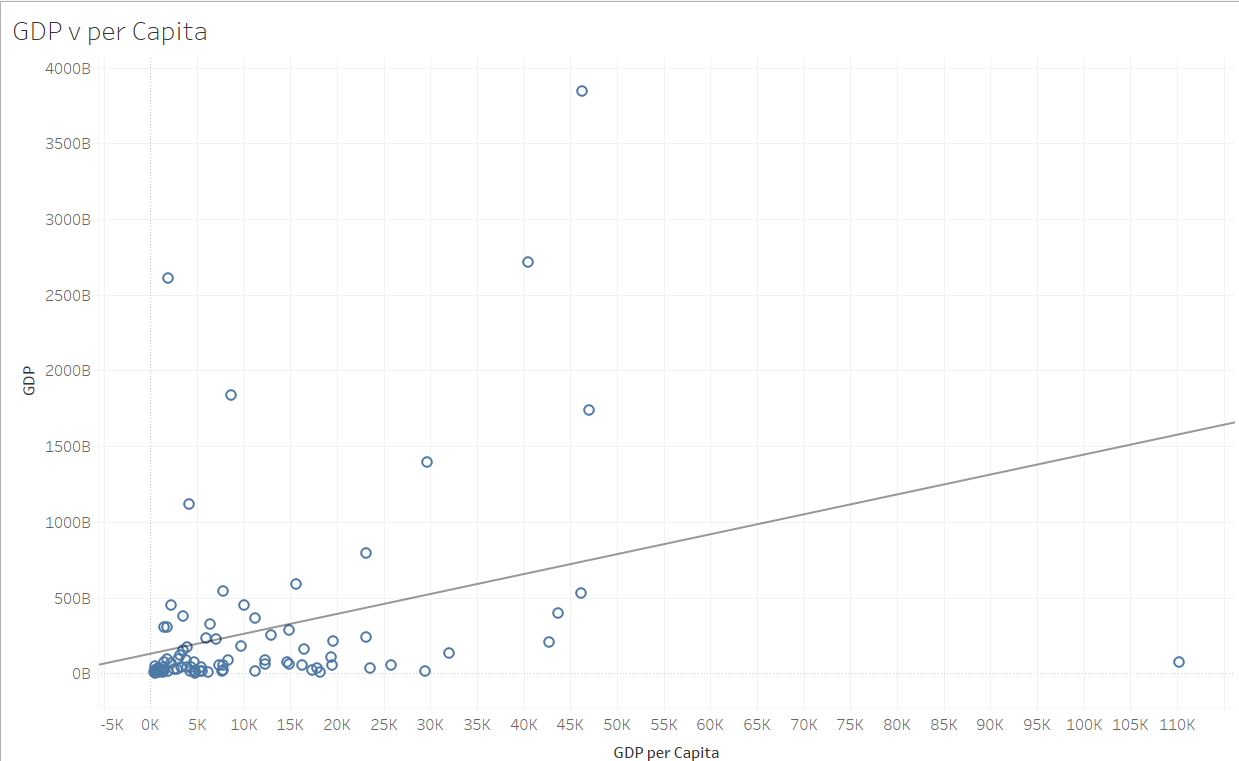
First and foremost, one of the most important features that could contribute to this is how rich a country is.



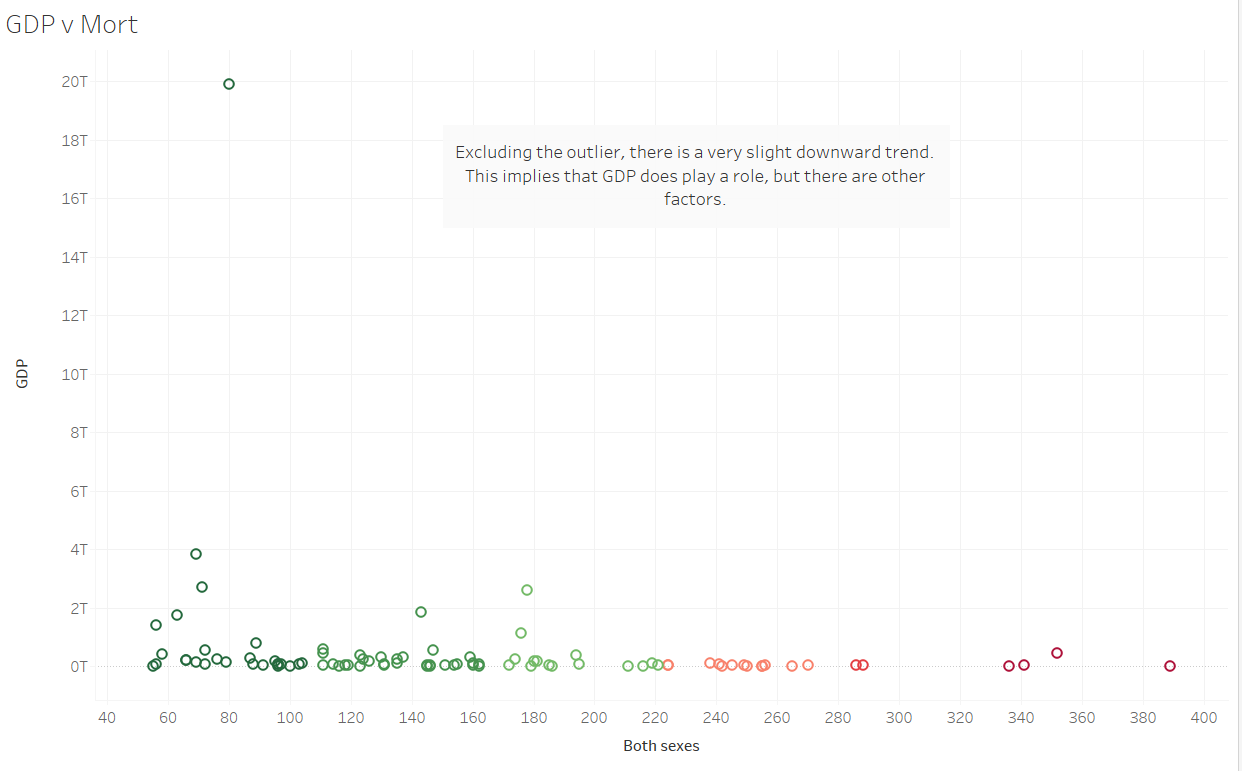
The visual on the top shows the average GDP per region, where we can see that Africa has significantly lower GDP than the rest of the regions. Based on the visual, it can also be seen that Asia & Pacific has a large GDP and Europe has a comparatively low GDP, but when it comes to the average mortality rate, Europe has one of the lowest.

On the other hand, the visual shows the average GDP per capita by region, which is a more clear indicator for country comparisons, where it can be seen that North America, Europe, and the Middle East have the highest average GDP per capita. Additionally, it would seem that these results reflect on the average mortality rate box plot shown above, where these three regions have the lowest overall mortality rates. The GDP per capita is based on the GDP and population, meaning that it is more or less a measure of the average citizen’s wellbeing. This implies that because the regions have lower GDP per Capita they have higher mortality rates.

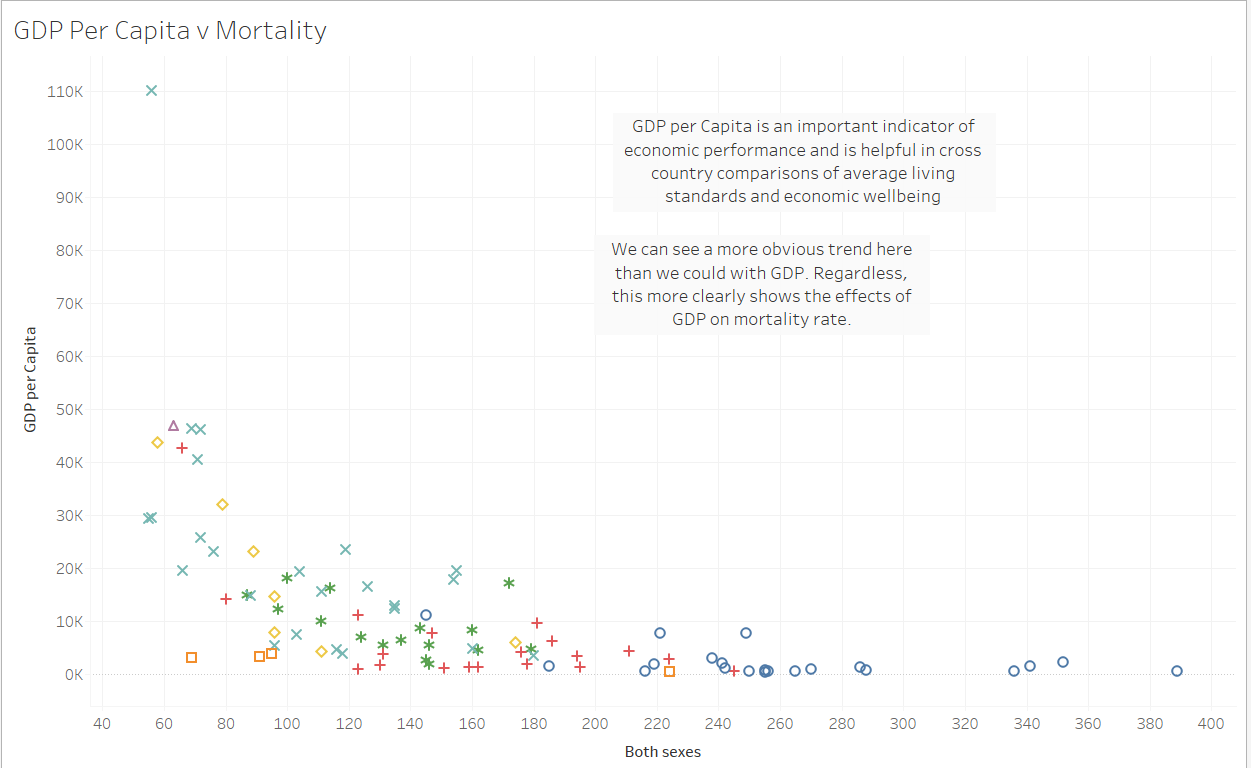
While GDP per capita shows more expected results, GDP is still an important factor to look at. Additionally, GDP and GDP per capita are slightly correlated, so they are not completely unrelated.



Below is the correlation between GDP and mortality rate, excluding a few outliers.



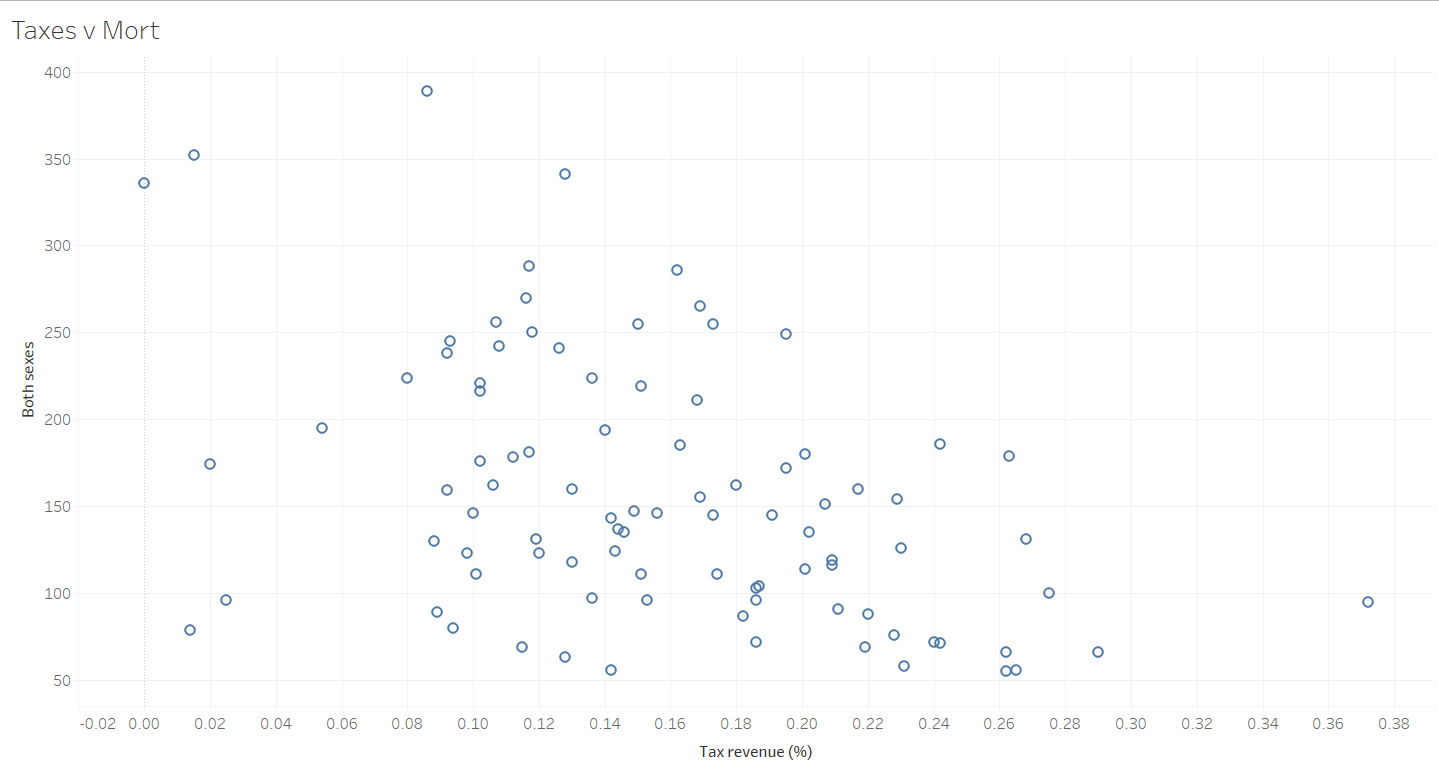
We can see that there is an extremely weak correlation here, implying that there may be other factors that may exist. On the other hand, the correlation between GDP per capita and mortality rate is much more apparent, featuring a much more noticeable trend.



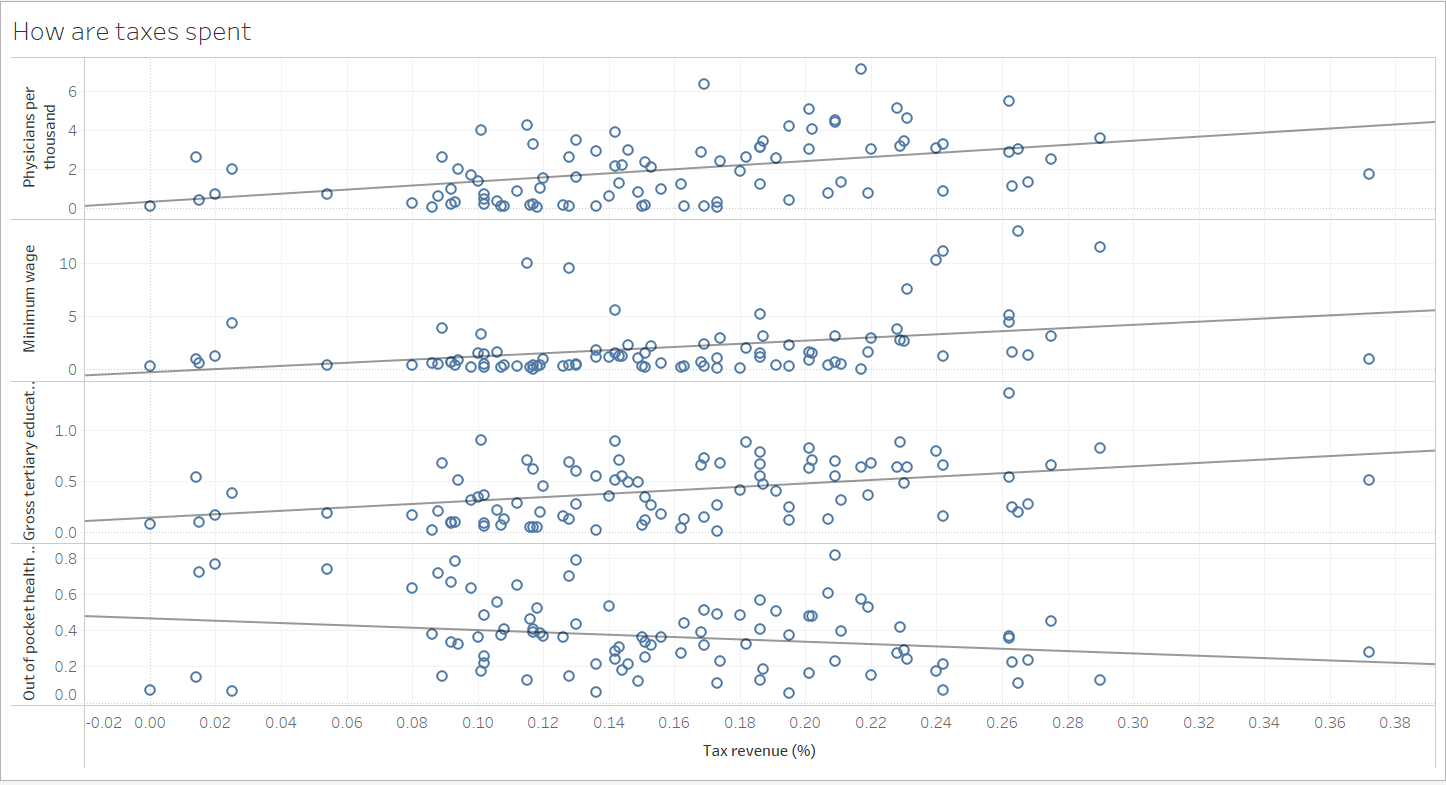
The richness of the region, however, does not necessarily mean much. Rather, I believe that how they spend their money is more important.

**Spending**

As stated before, how a country/region allocates their money to certain essential services likely has a big impact on mortality rates. Higher tax revenues mean that a country is able to spend more on essential services, and is a key factor for economic growth and poverty reduction.



Generally speaking, taxes are used to fund essential services. As seen above, a higher tax revenue has lower mortality rates. This is likely due to their ability to increase spending on essential needs.



As seen here, increased tax revenue has positive correlations with physicians per thousand, minimum wage, and gross tertiary education. In addition, it has a negative correlation with out of pocket health expenditures, or in other words, higher tax revenue leads to a decrease in out of pocket expenditures.

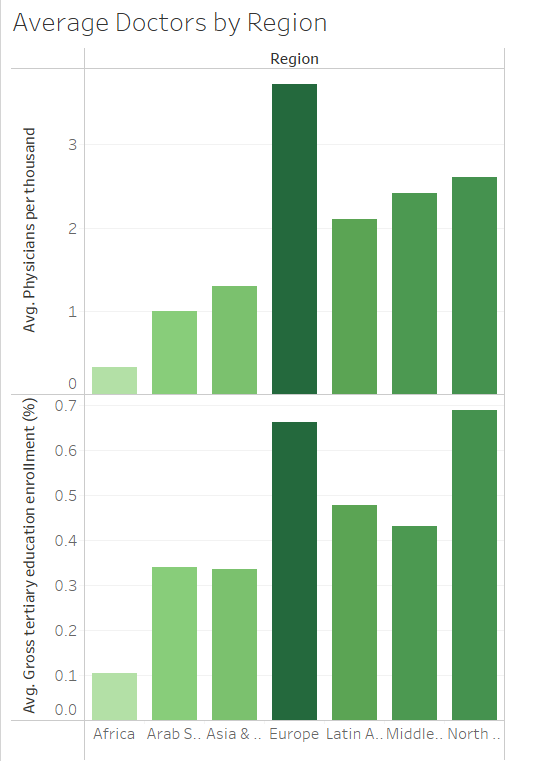
This is not to say that richer countries have the advantage in this regard, but rather it should be used as an indicator of which areas spending should be increased in.

**Accessibility**

After viewing the relationships that GDP and GDP per capita have with mortality rate, it is clear that while they do affect it, they are not the only factors that contribute. In addition, we took a look at how certain essential factors are related to tax revenue. In this section, we will observe those factors.

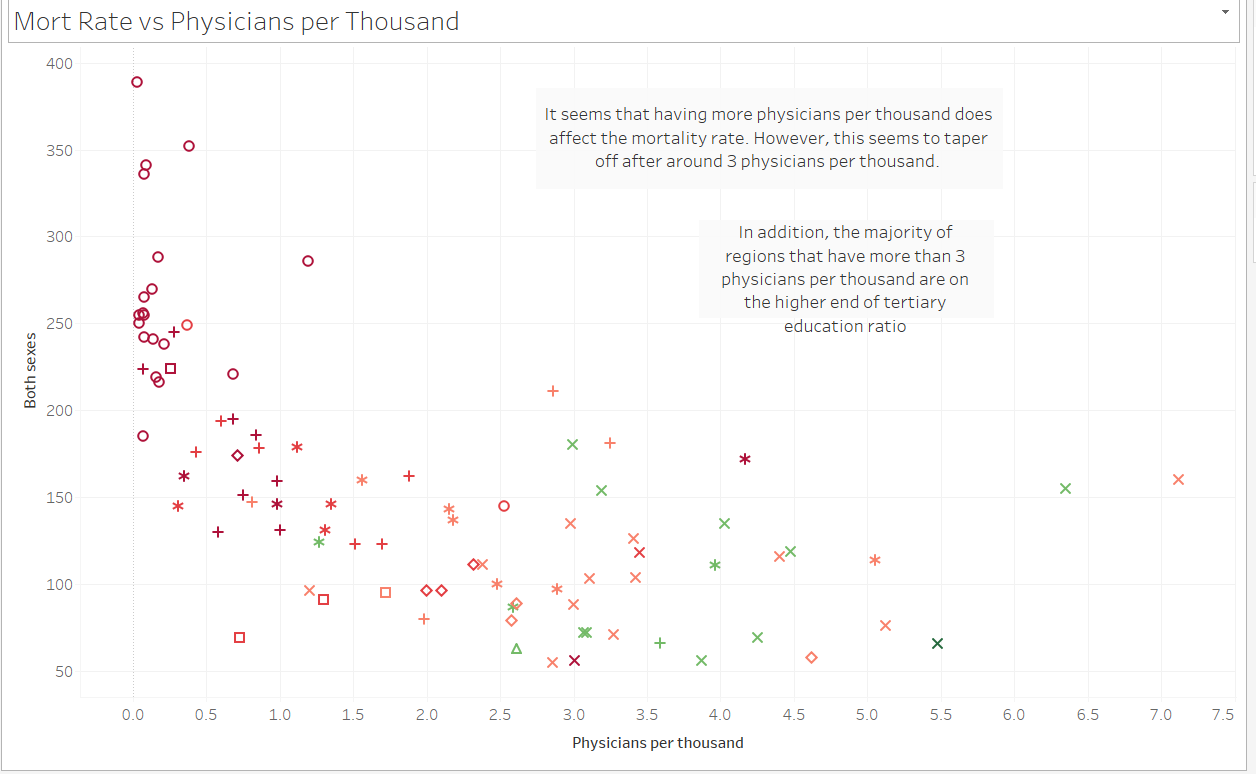
Outside of the overall richness of the regions and countries, a large factor would likely be related to health care. Of course, the quality of health care is essential, but accessibility for the average citizen is equally as important.

For our analysis, I will be taking a look at the physicians per thousand, the average minimum wage, and the average out of pocket health expenditures.



The visual above shows the average physicians per thousand for each region, as well as their gross tertiary education ratio. One thing to note is that there seems to be a correlation where higher tertiary education ratios lead to higher amounts of physicians per thousand.

Below shows the correlation between mortality rate and physicians per thousand, with the color being denoted by the tertiary education ratio.

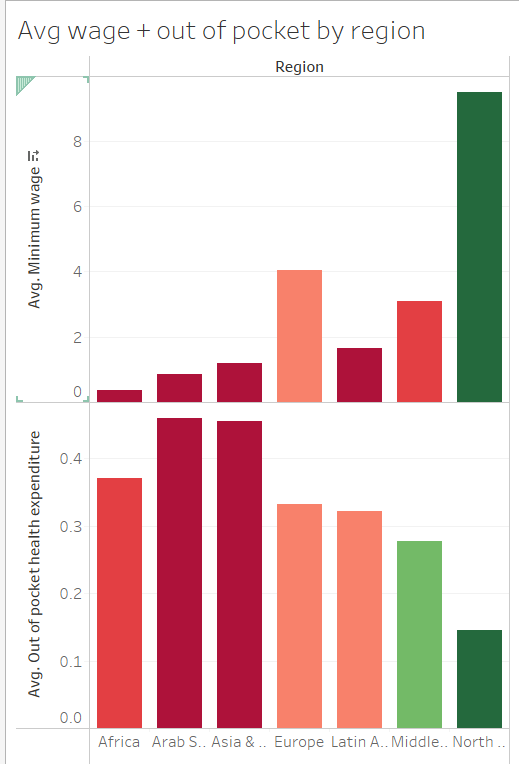


From this, we can conclude that the more physicians per thousand there are, the lower the overall mortality rate is. This is likely due to less tradeoffs in terms of travel time to reach a doctor and pricing due to the increased competition. Additionally, we can see that the tertiary education ratio increases as physicians per thousand increases (denoted by the red to green coloring). However, despite decreasing in mortality rate, it seems to taper off around the 3-4 physicians per thousand mark, implying that having too many physicians may be a waste.

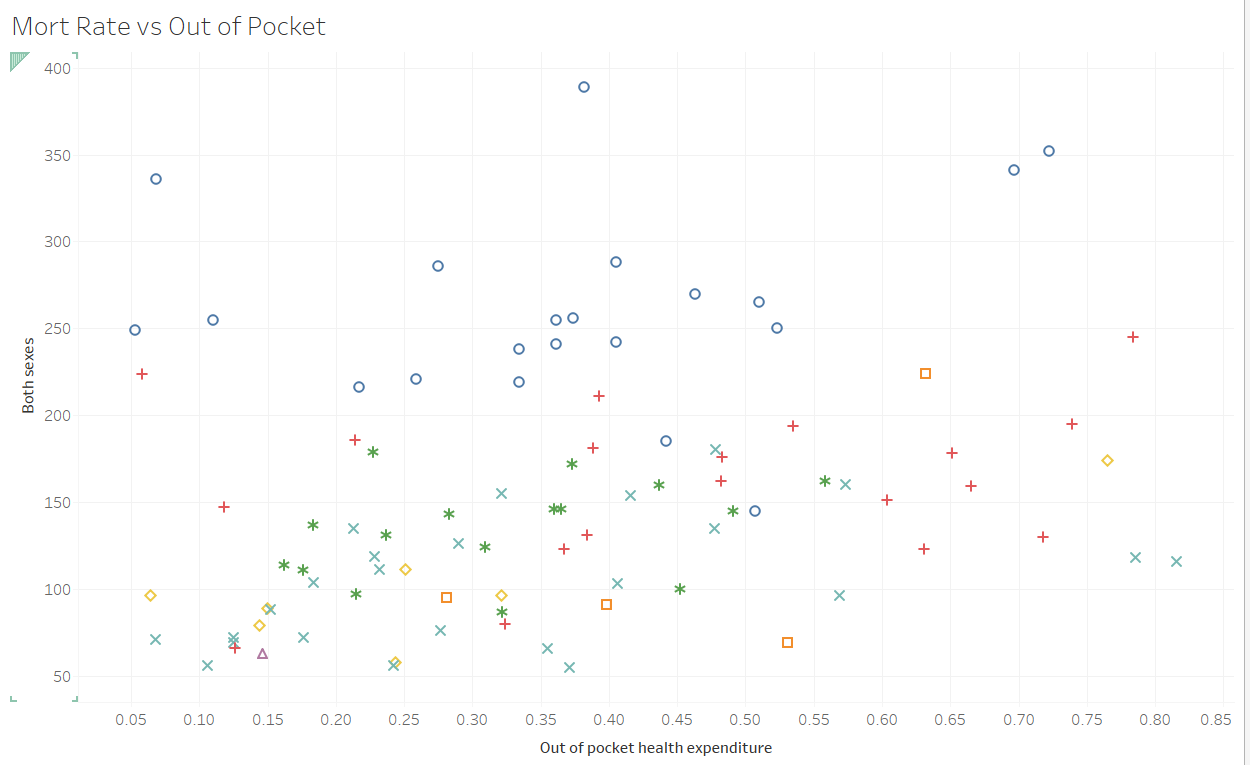
Another thing to note is that higher tertiary education ratios generally have more benefits for the individual as well as society as a whole. In this case, it could be said that the number of physicians increasing is a byproduct of increased higher education ratios.

**Affordability**

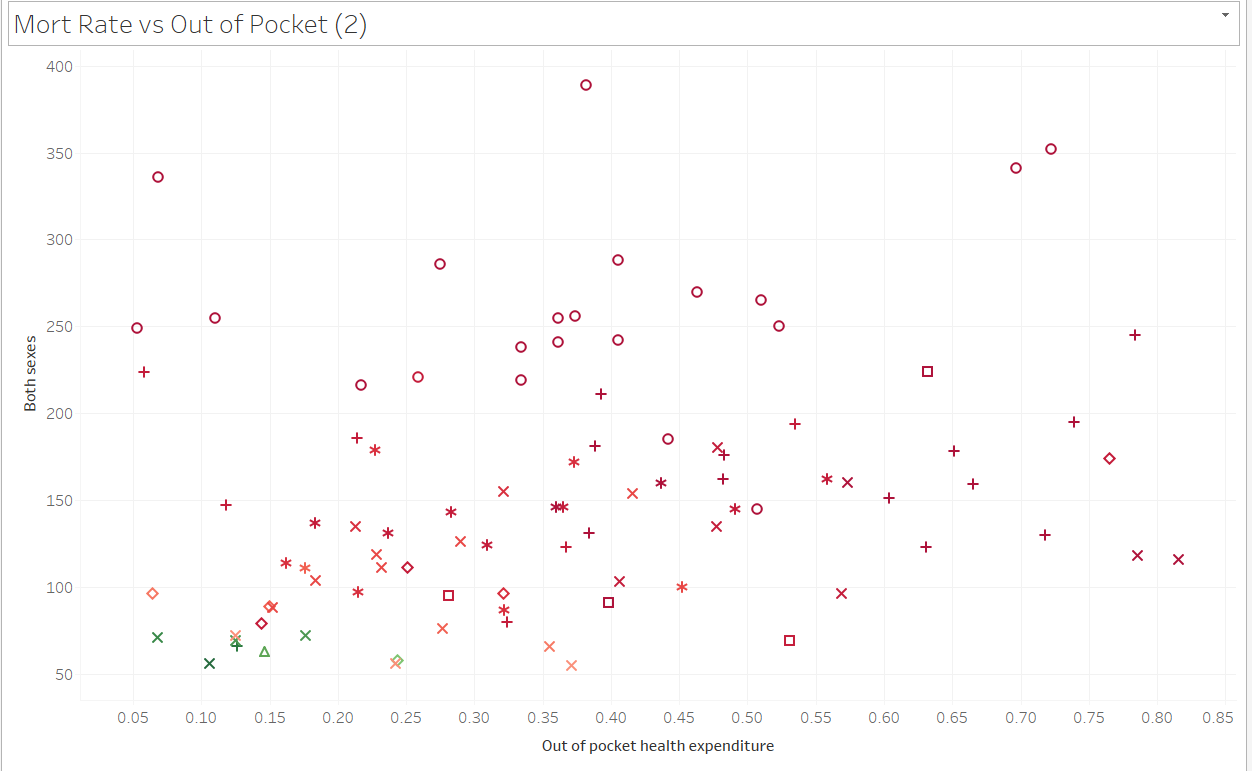
Next, we can take a look at the average minimum wage, as well as the out of pocket health expenditures per region.



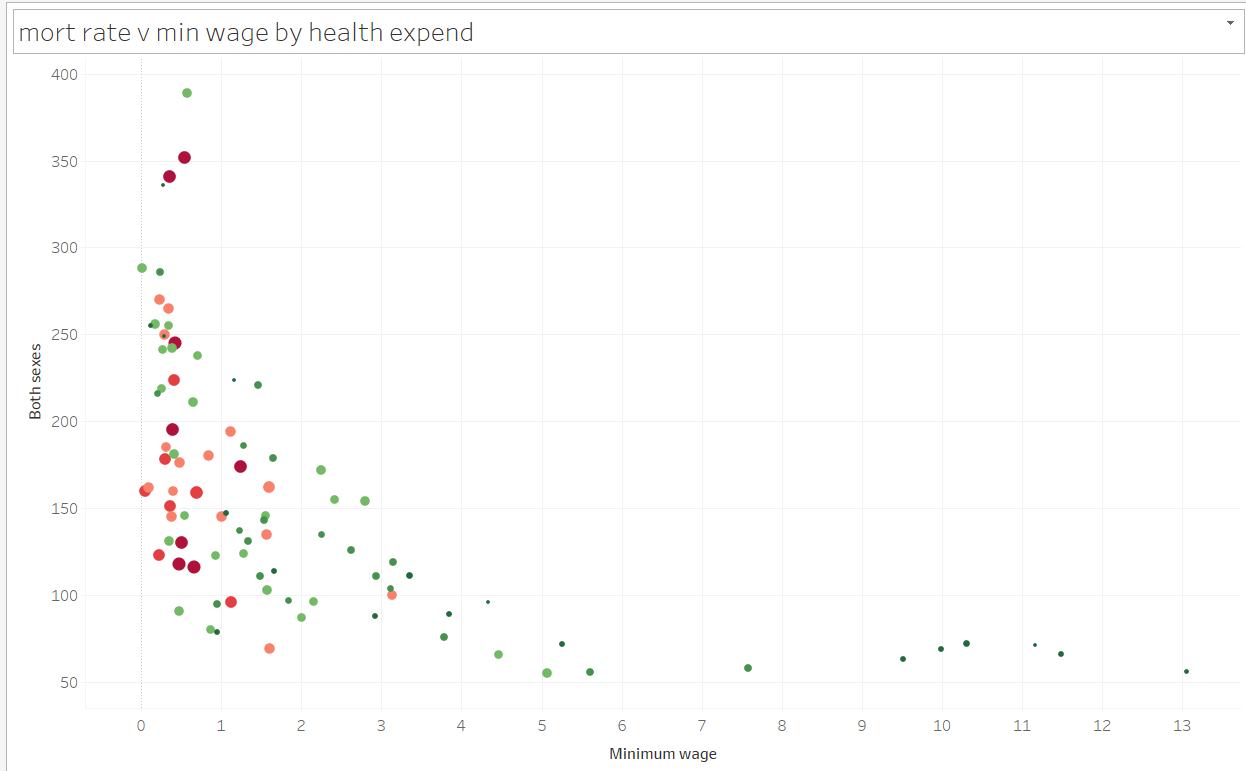
We can see that many of the regions that have high mortality rates also have high out of pocket health expenditures. For regions, such as Europe, this expenditure is offset by their higher than average minimum wage.



The scatter plot above shows the relationship between mortality rate and out of pocket health expenditures. There is a clear trend of higher expenditure meaning higher mortality rates, but the spread is big. This spread can be explained by the visual below, which shows the same scatter plot with the color equating to minimum wage.



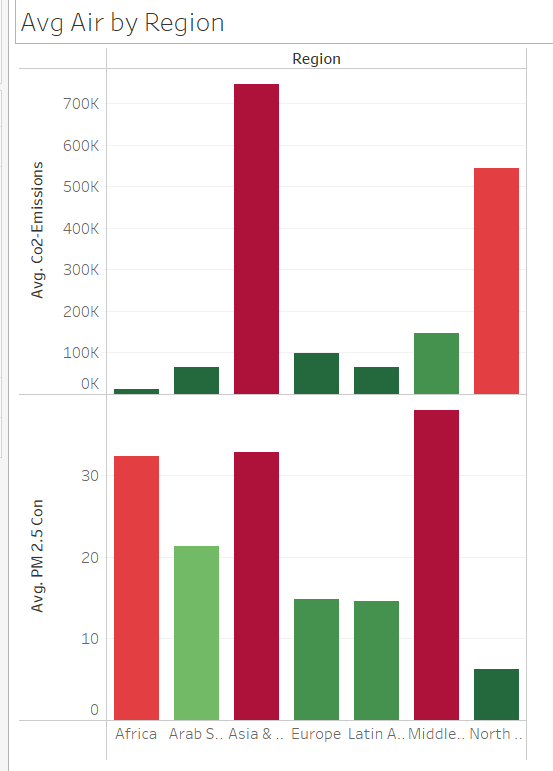
With the new colors, we can see that lower minimum wages (darker red) tend to have higher mortality rates than those with higher wages, even for those with similar health expenditures. We can also visualize this the other way around, with mortality rate vs minimum wage where the color is denoted by health expenditure.



This alternative scatter plot visualizes a similar thing, where low minimum wage combined with high expenditures have higher mortality rates.

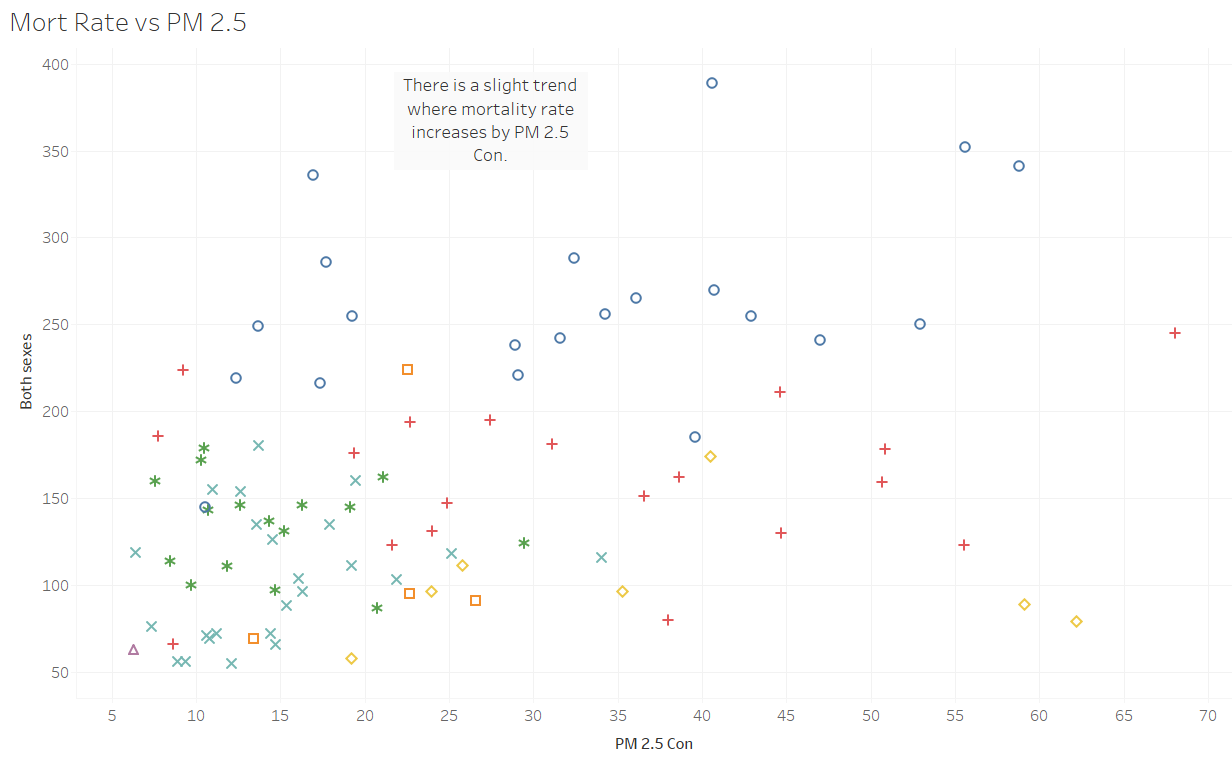
**Air Quality**

Finally we will look at the effects of co2 emissions and pm 2.5 concentrations on mortality rate.

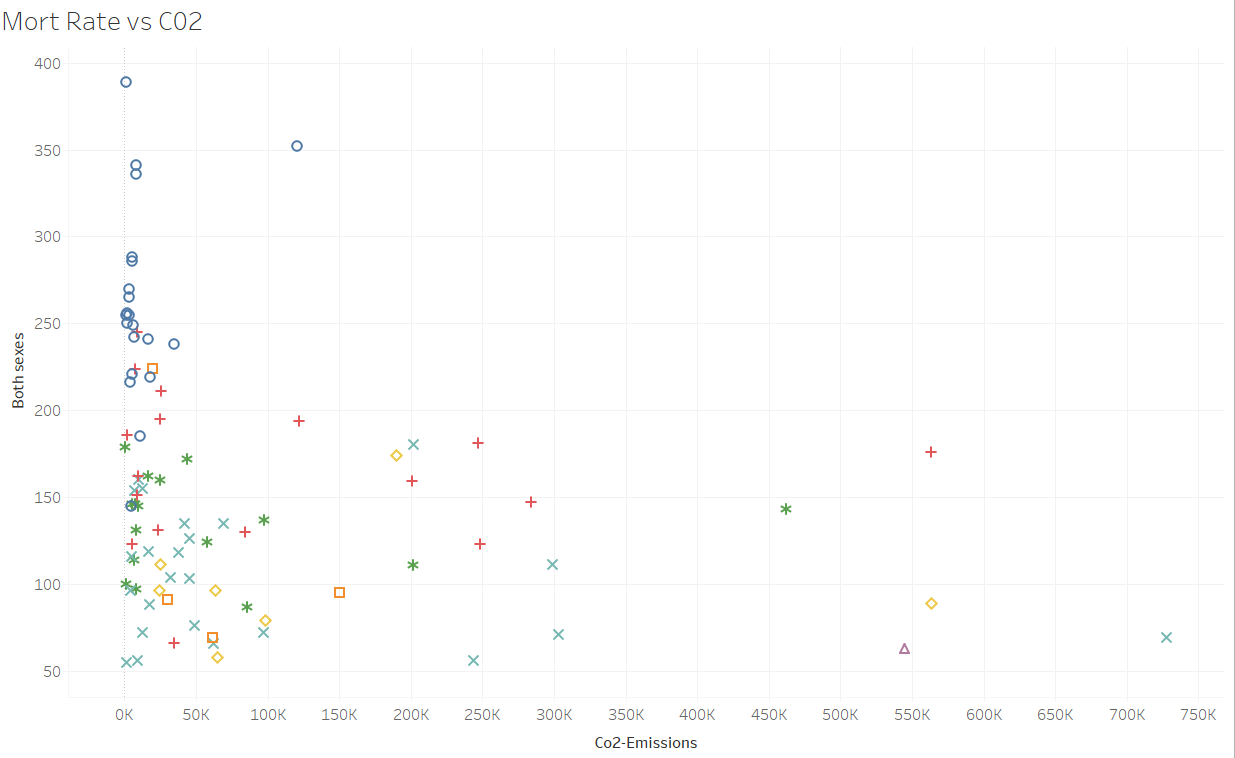


Unlike what I had expected, there does not seem to be much correlation between co2 emissions and PM 2.5 concentrations. Regions with high co2 emissions do not necessarily have high PM 2.5 pollution and vice versa. This implies that there are other factors that play into these two features.

When visualizing the correlation between these features and mortality rate, we get the following:



Mortality rates have an upward trend with PM 2.5. This is likely due to the adverse health effects that PM 2.5 pollution has.

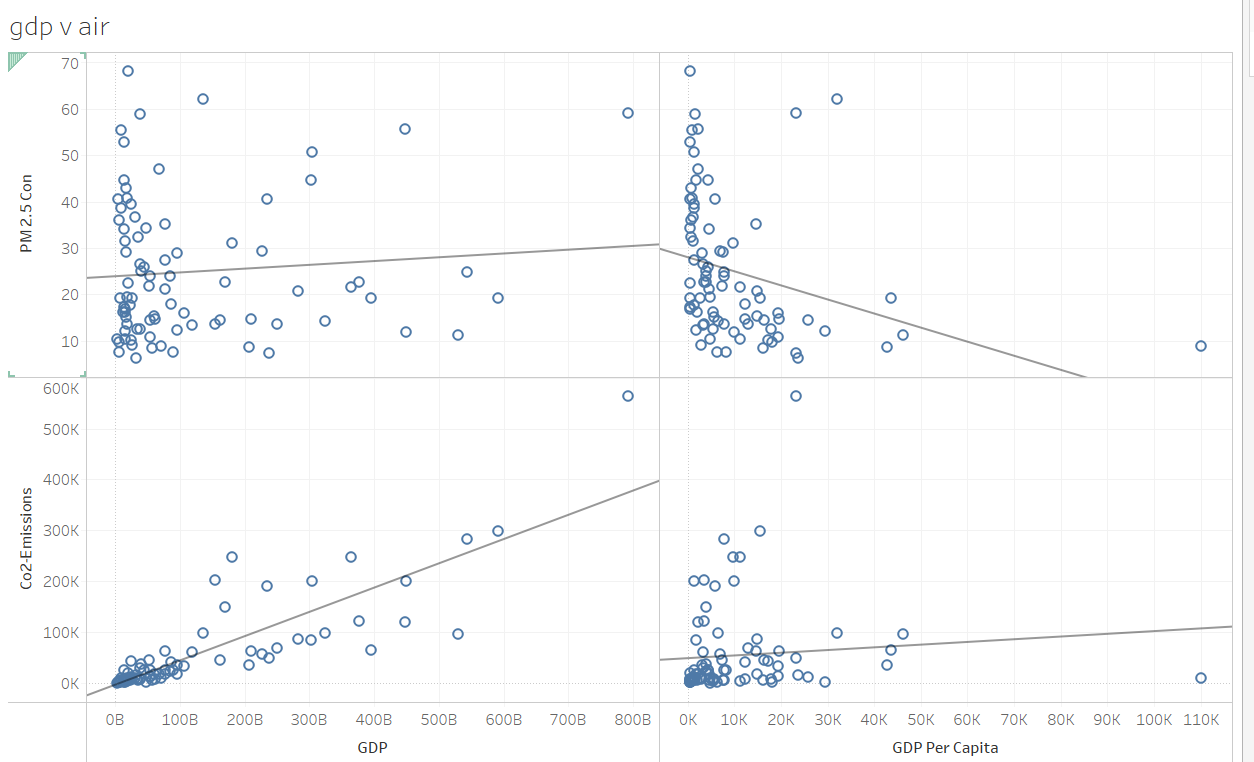


On the other hand, co2 has a less apparent trend that becomes more visible with the removal of outliers. There is likely a reason for this lesser trend, but we will discuss that in the next section.

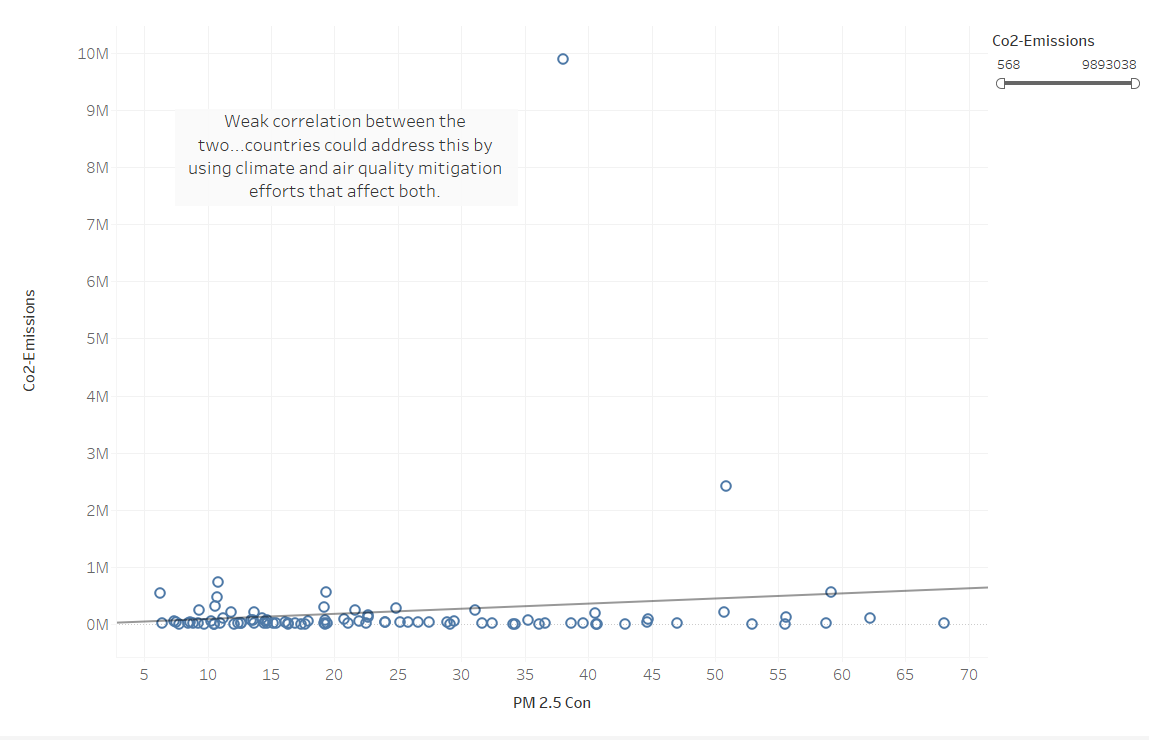
**Interpretation**

We have observed that a variety of factors play into mortality rates. To try to find ways to reduce mortality rates, we can attempt to find how correlated certain factors are in order to determine which methods are the most effective.

First, we will take a look at correlations for the air quality.

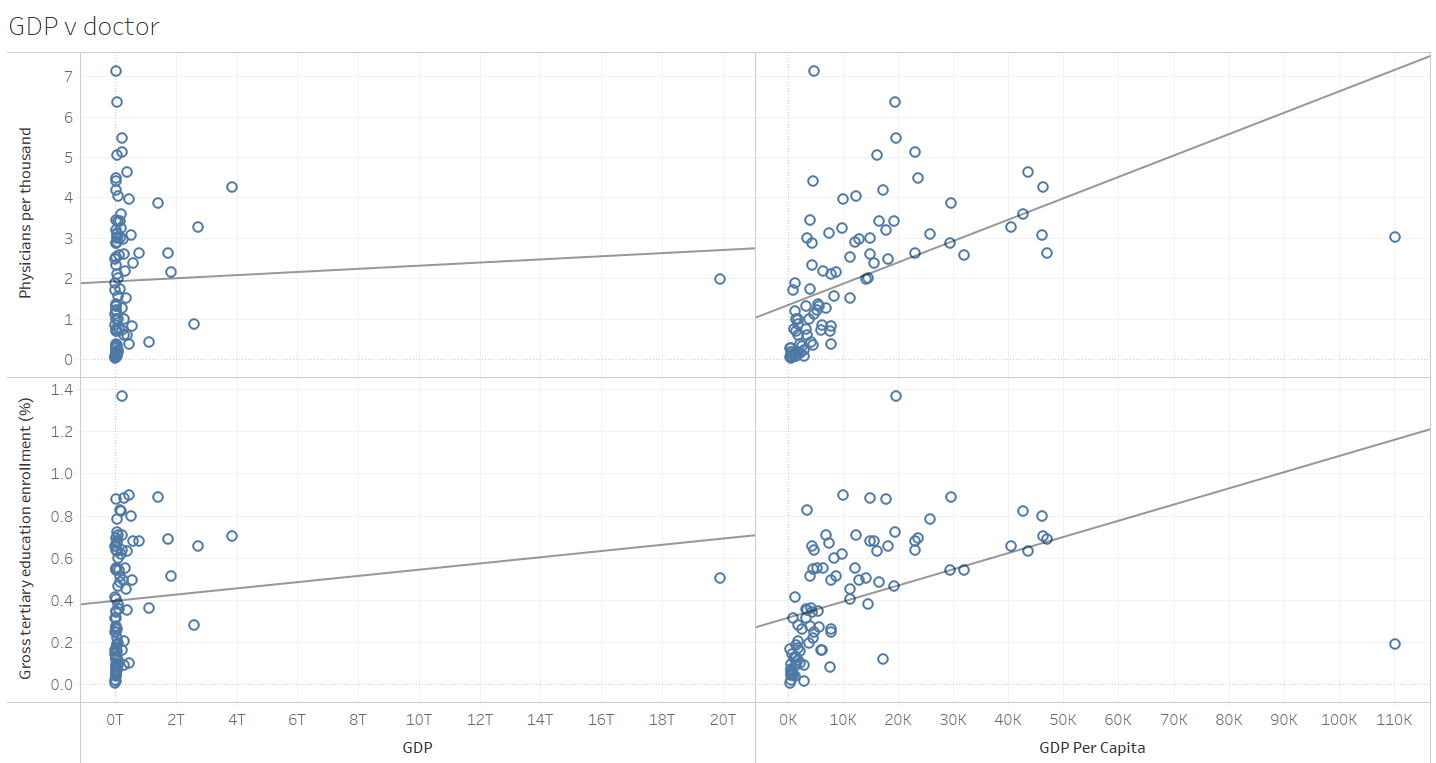


This visual shows the correlation between GDP/per Capita and PM 2.5 and co2 emissions. The PM 2.5 correlations do not seem to have any meaningful insights given how spread out it is. However, something to note is that CO2 emissions have a very positive correlation with GDP. Additionally, we noted earlier that co2 emissions did not have any correlation with mortality rates. This could be due to the fact that richer countries may have more methods to deal with these emission levels (like increased spending on essential services for example).

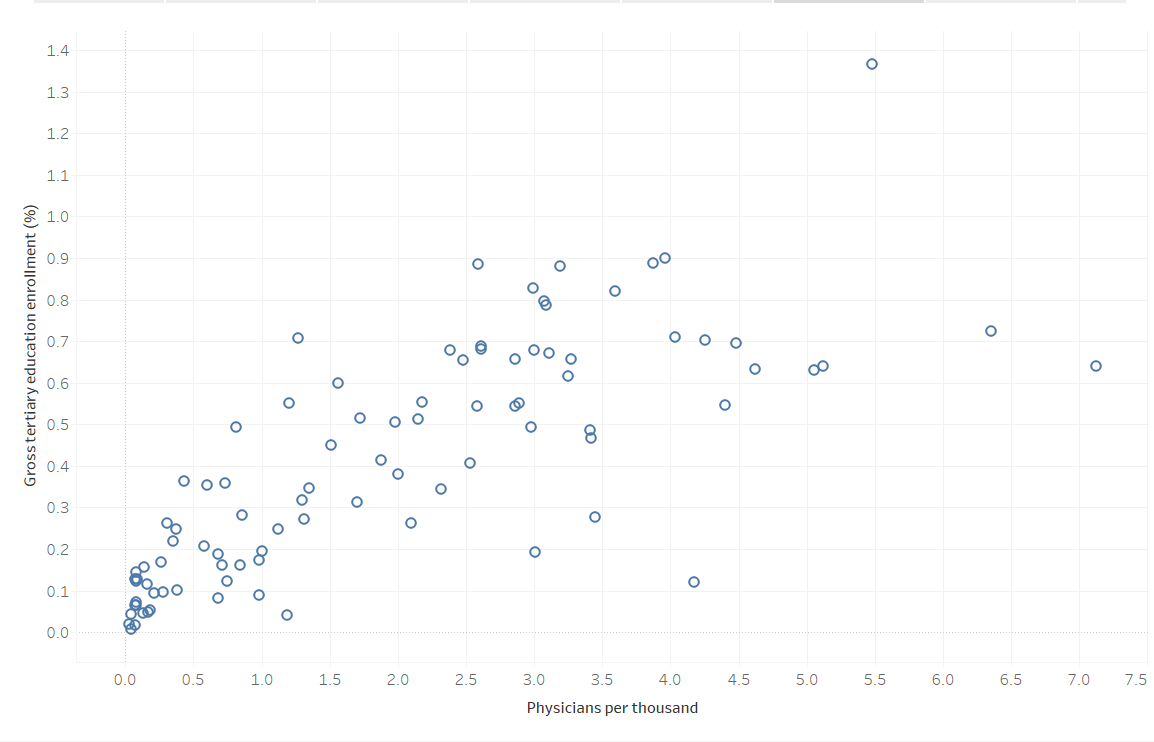


From this visual, we can see that there is a small correlation between co2 emissions and PM 2.5 concentrations. While co2 emissions by itself do not seem to have much correlation with mortality rates, we know that PM 2.5 does. As a result, countries should try to lower their PM 2.5 concentrations if possible. Since there is a slight correlation between these two factors, efforts to reduce emissions should also help the PM 2.5 situation.

Next, we will look at the physicians per thousand and tertiary education enrollment ratios.

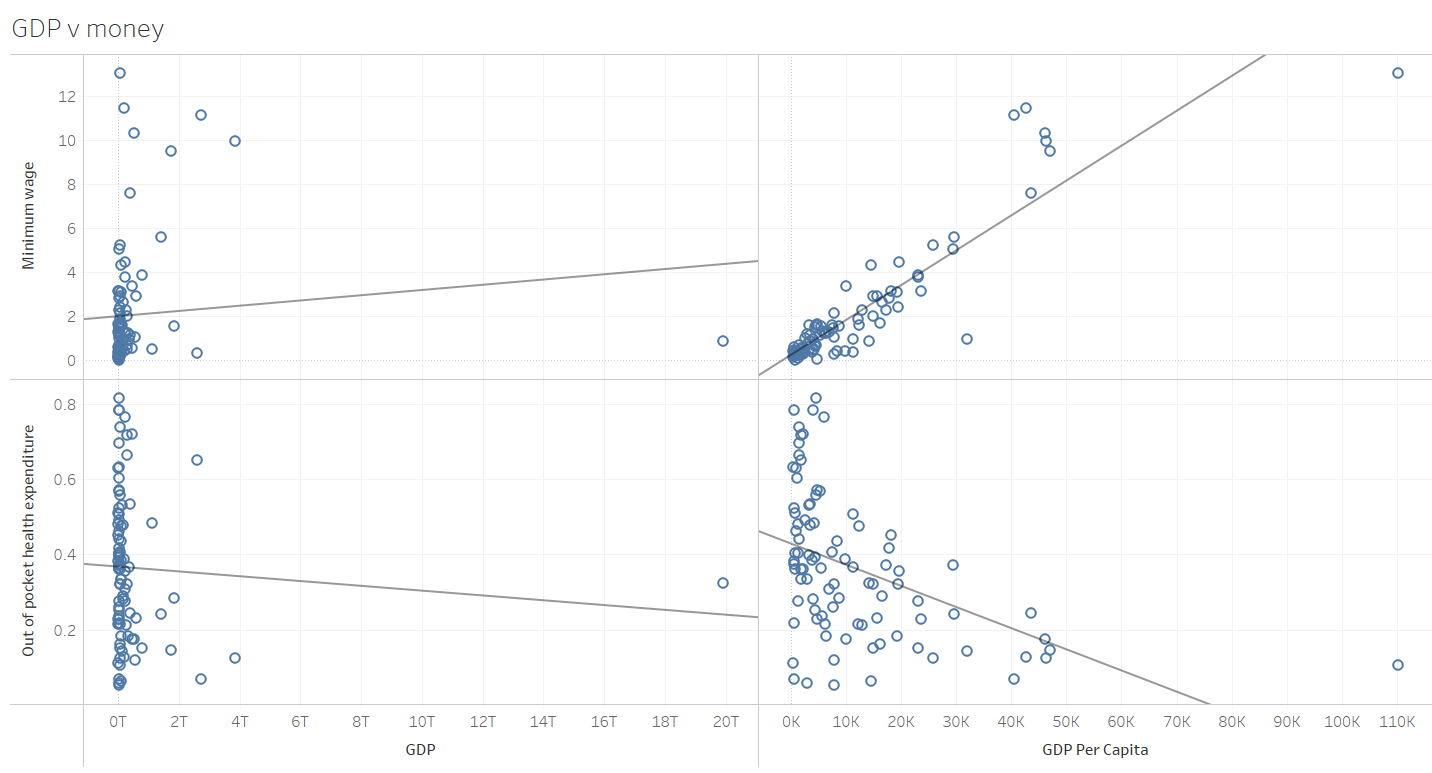


From what we can tell, GDP does not have any impact on these two factors. However, GDP per Capita has a more significant correlation with both of them. One thing to note is that the correlations look almost the same shape wise, something we will explore in the next visual.

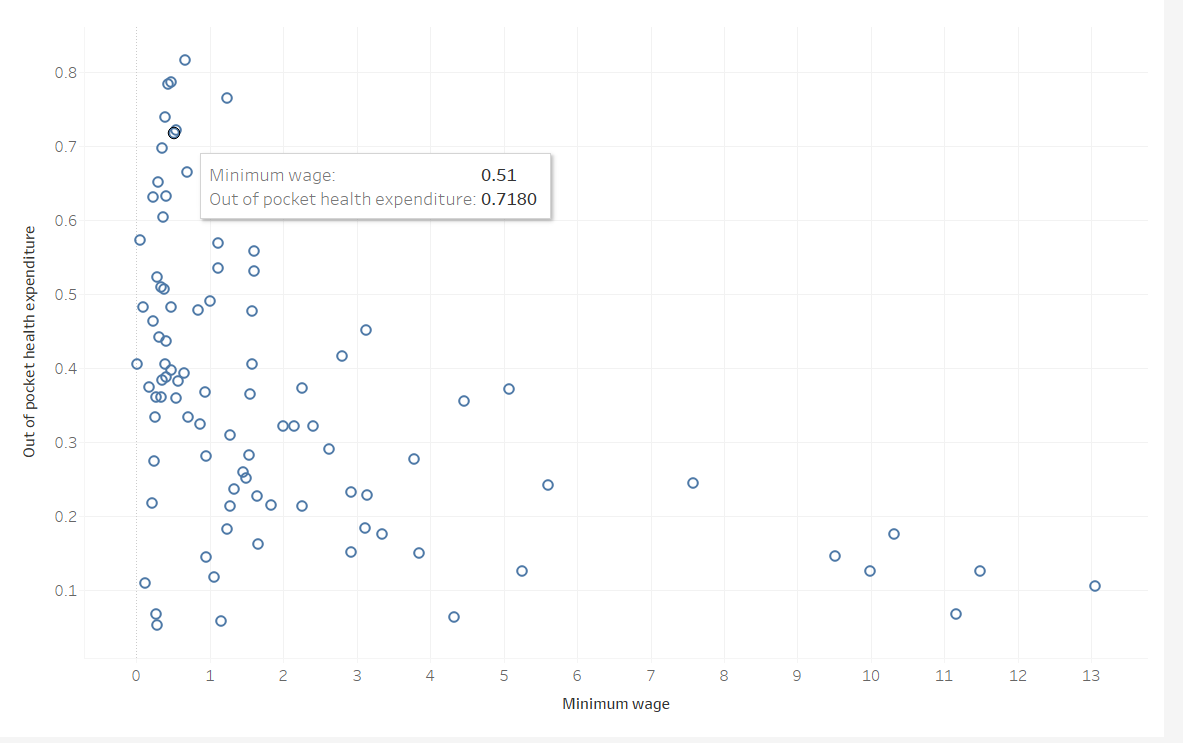


Here, we can see that physicians per thousand and tertiary education enrollment ratio is highly correlated. This, in tandem with the previous visual, implies that higher GDP per Capita allows for easier access to tertiary education, which in turn increases the amount of physicians per thousand. This makes sense since GDP per Capita is essentially the “GDP” per person.

Finally, let us take a look at the affordability trends.



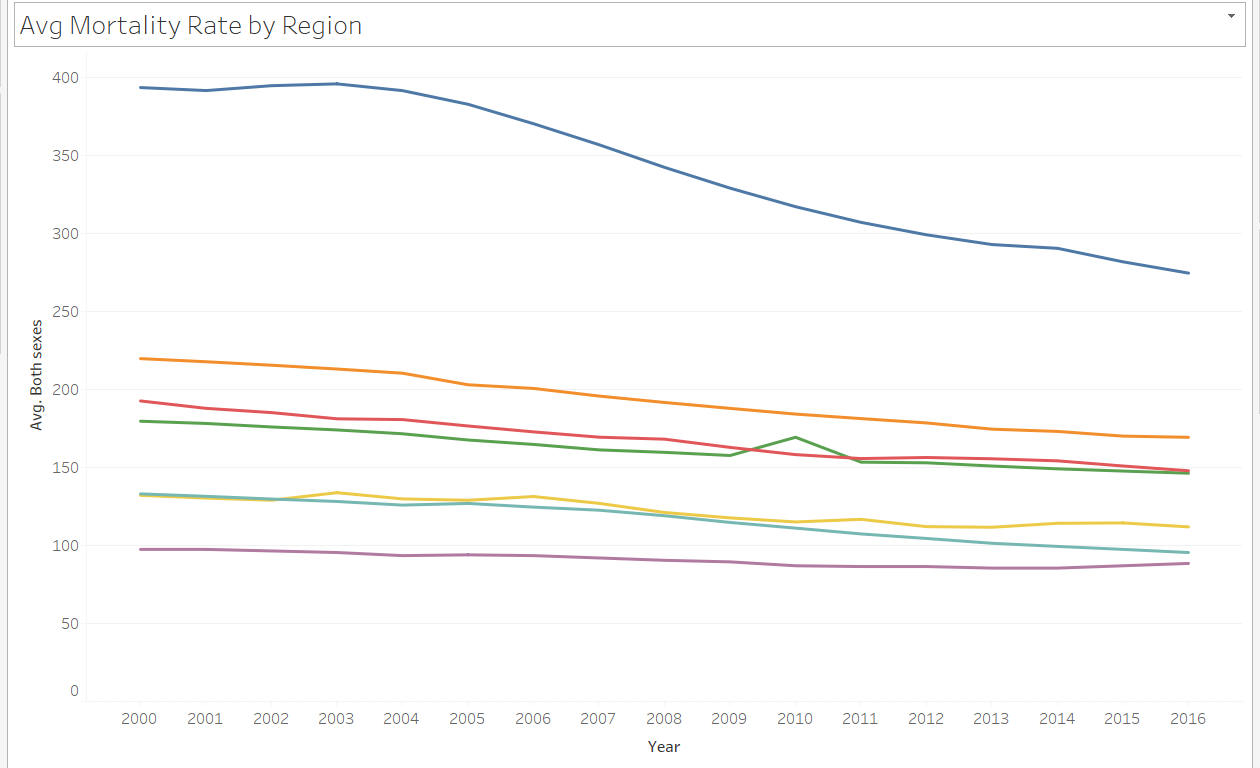
Again, GDP has no impact on minimum wage and out of pocket health expenditures. On the other hand, GDP per capita has a big correlation with minimum wage and a lesser one with out of pocket expenditures. It is doubtful that minimum wage and out of pocket health expenditures are very correlated, but we will take a look.



Here we can see the correlation between out of pocket expenditures and minimum wage. There is a slight downward slope, but the spread is very large. This implies that minimum wage does not have a significant impact on health expenditures. However, as we know from earlier, when both are taken into consideration together, there are large effects on mortality rates.

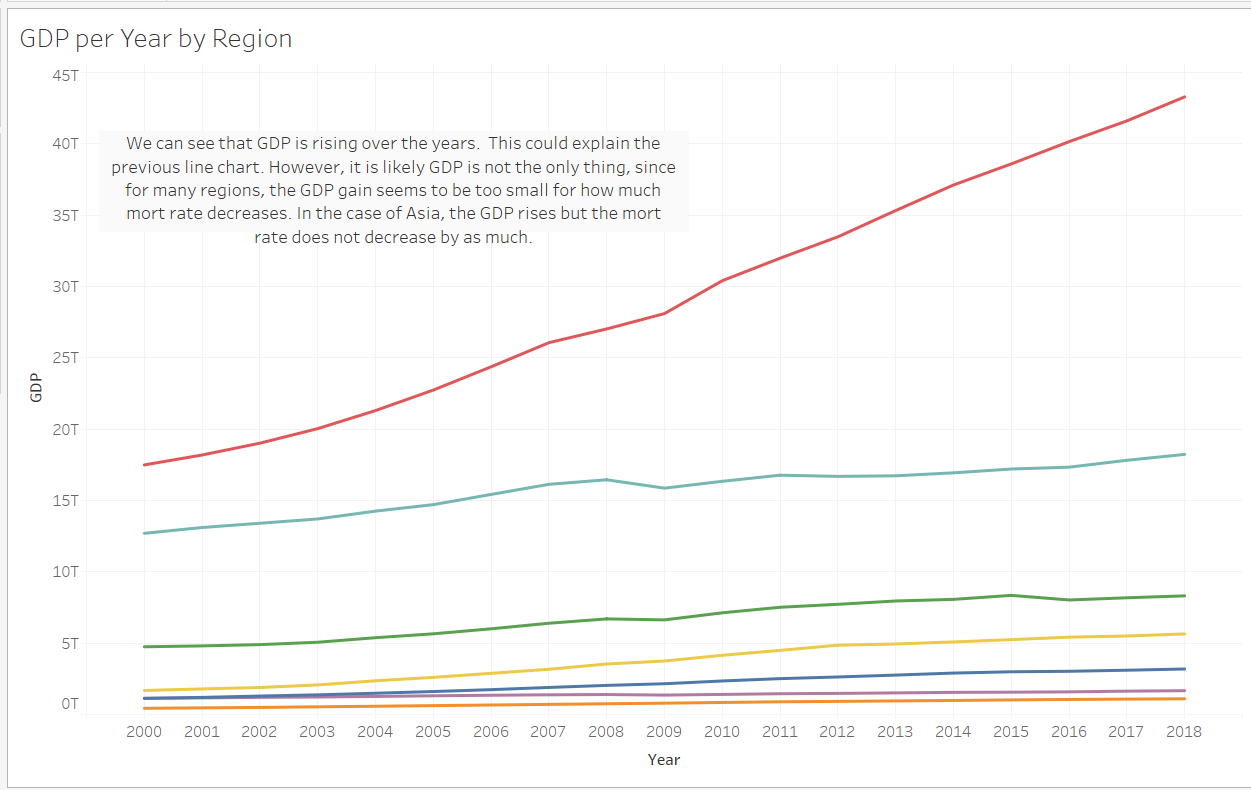
**Conclusion**

Over the years, outside of outliers like the COVID pandemic, mortality rates have decreased.



*Note that this visual does not include the pandemic*

We can see that mortality rates have constantly decreased. Conversely, GDP has increased over the years, as pictures below.

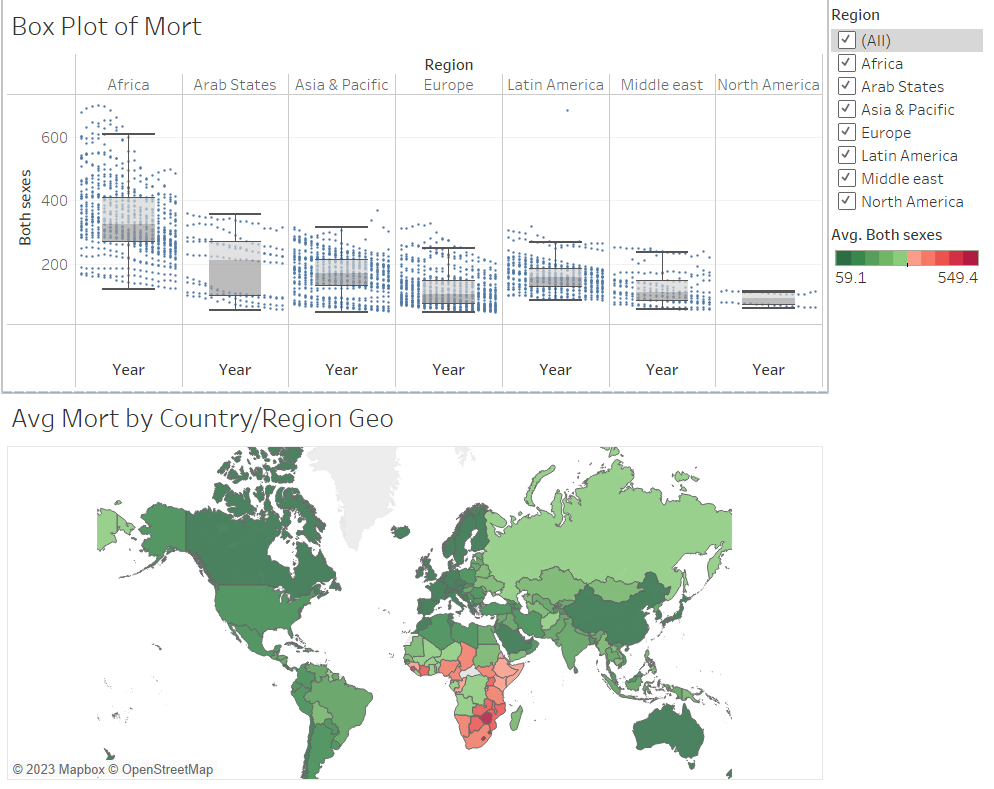


Note that the increase in GDP is not exactly one to one with the decrease in mortality rates, but likely still plays a role. Nonetheless, this shows that mortality rates can be decreased, and to do so, countries should choose wisely what they are spending on.

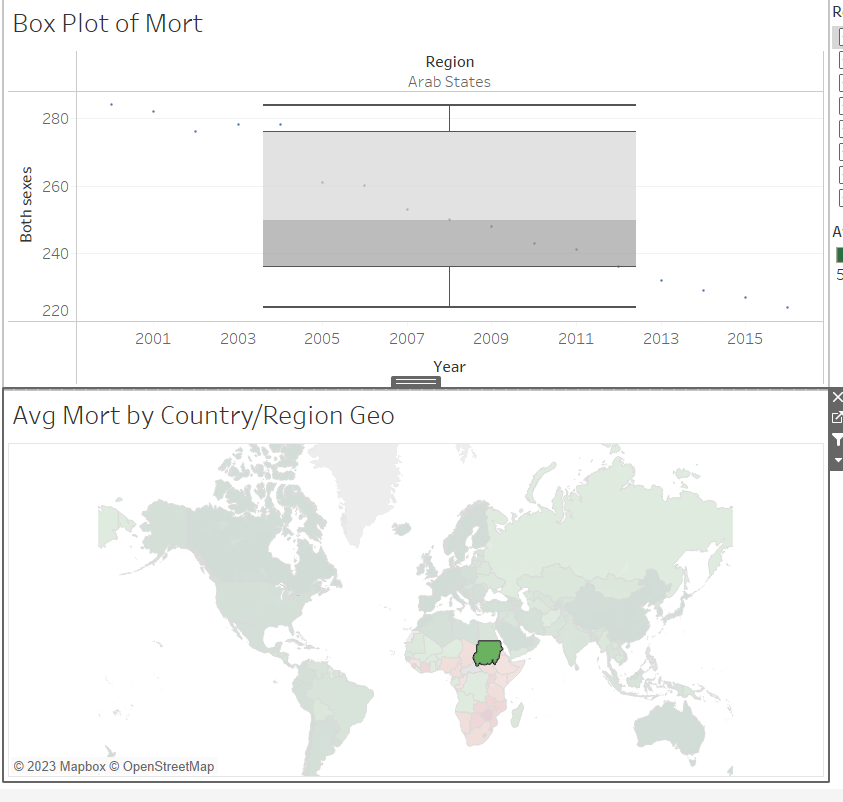
**Dashboard Interaction**

For more visual clarity, I split up my dashboard into multiple, smaller dashboards that focus on a specific topic: Mortality Rates, GDP, Taxes, Physicians + college, Min Wage v Out of Pocket health expenditures, Air Quality, and GDP + mortality trends.

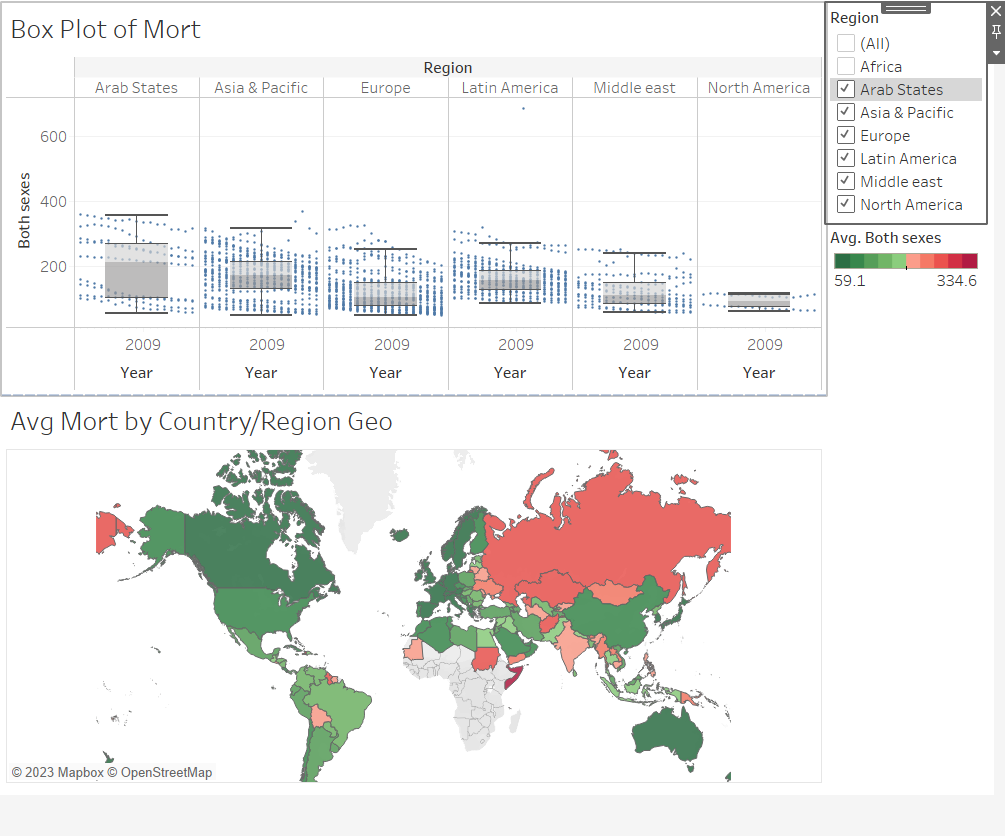
For the Mortality Rates dashboard:



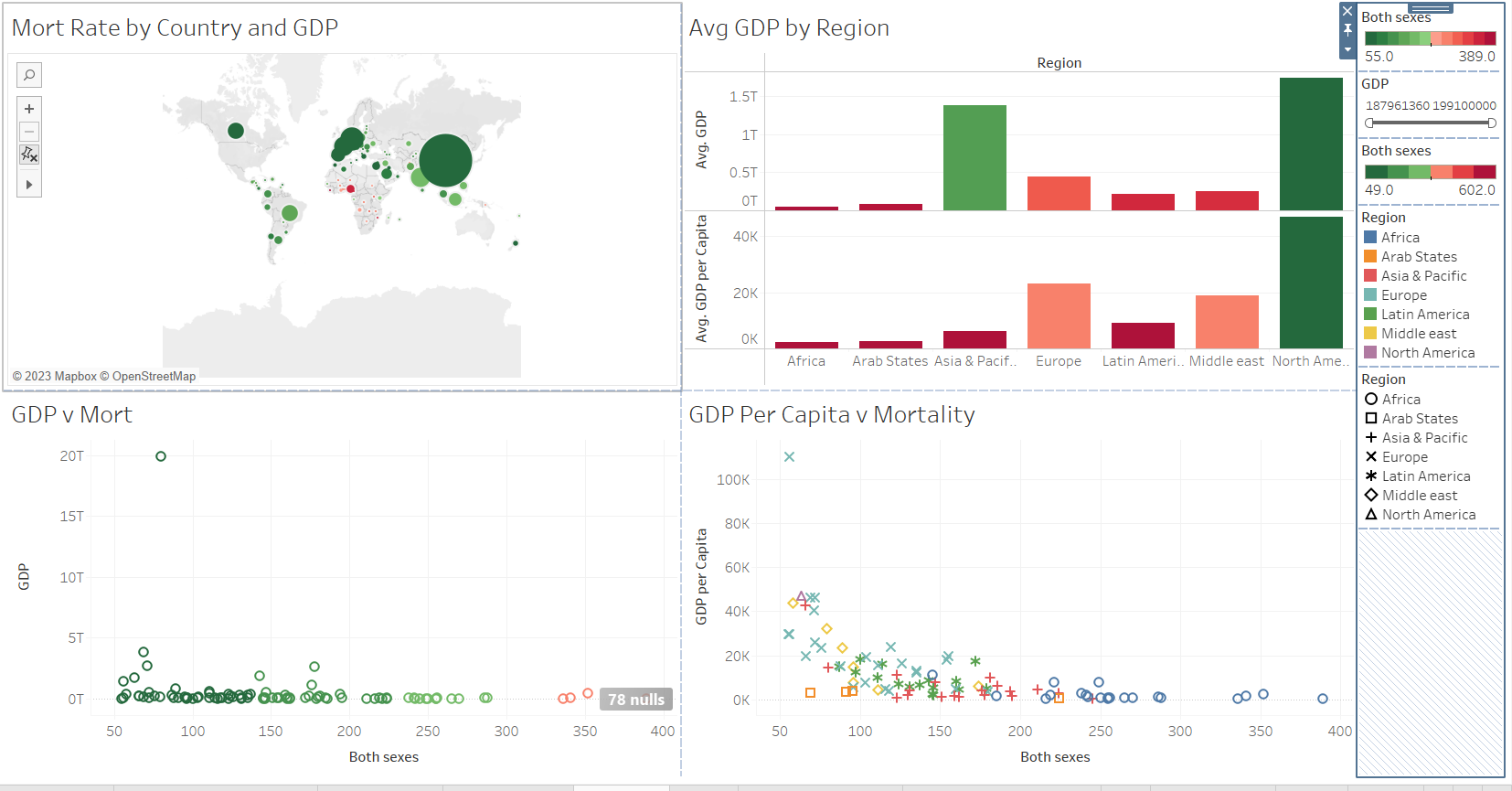
This is what the dashboard looks like normally. If you click on a country in the geographical map below, then you will be able to visualize the box plot using data only from that country, shown below.



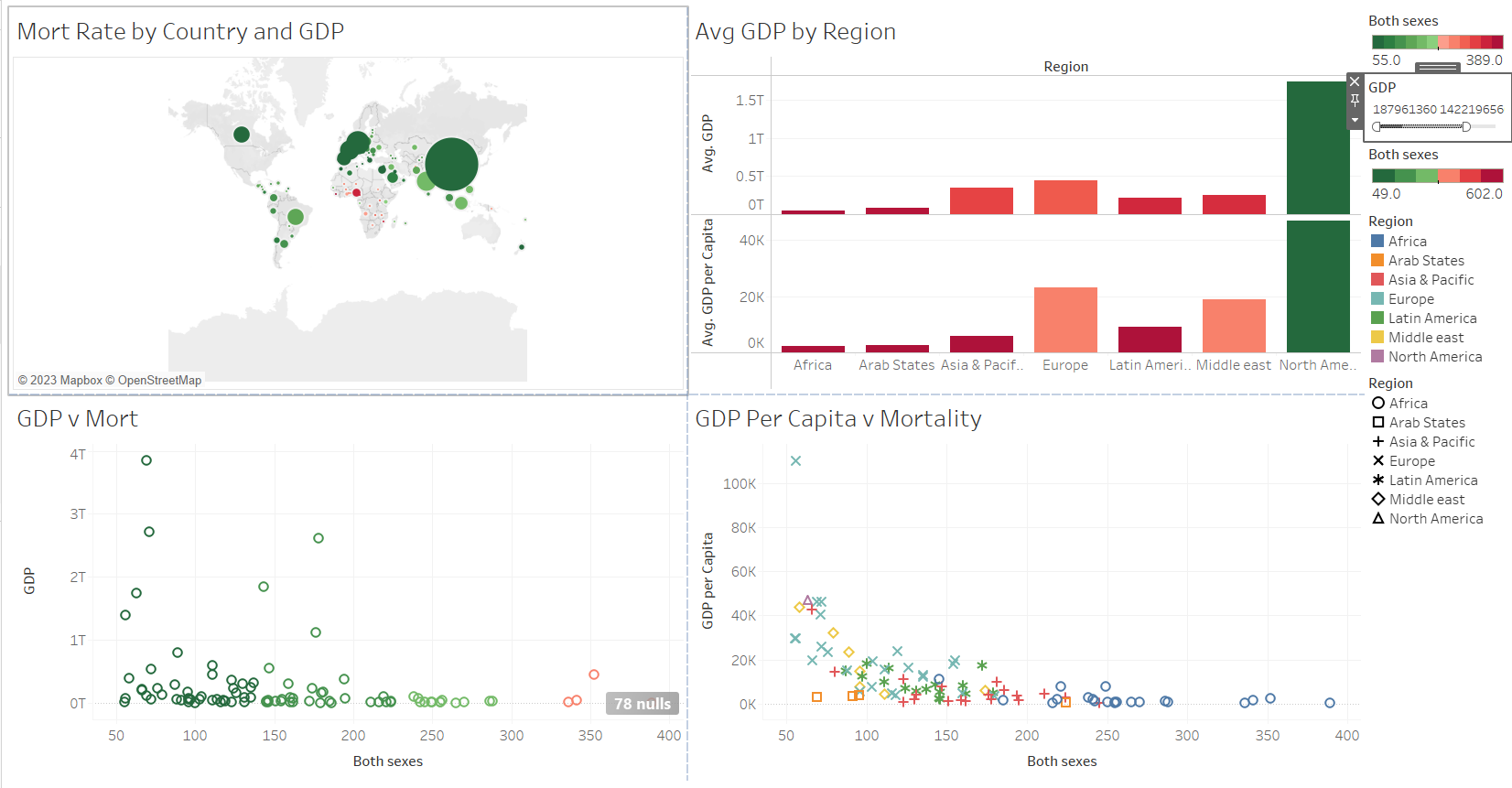
Additionally, you are able to use the region filter on the right to edit both the map and box plot.



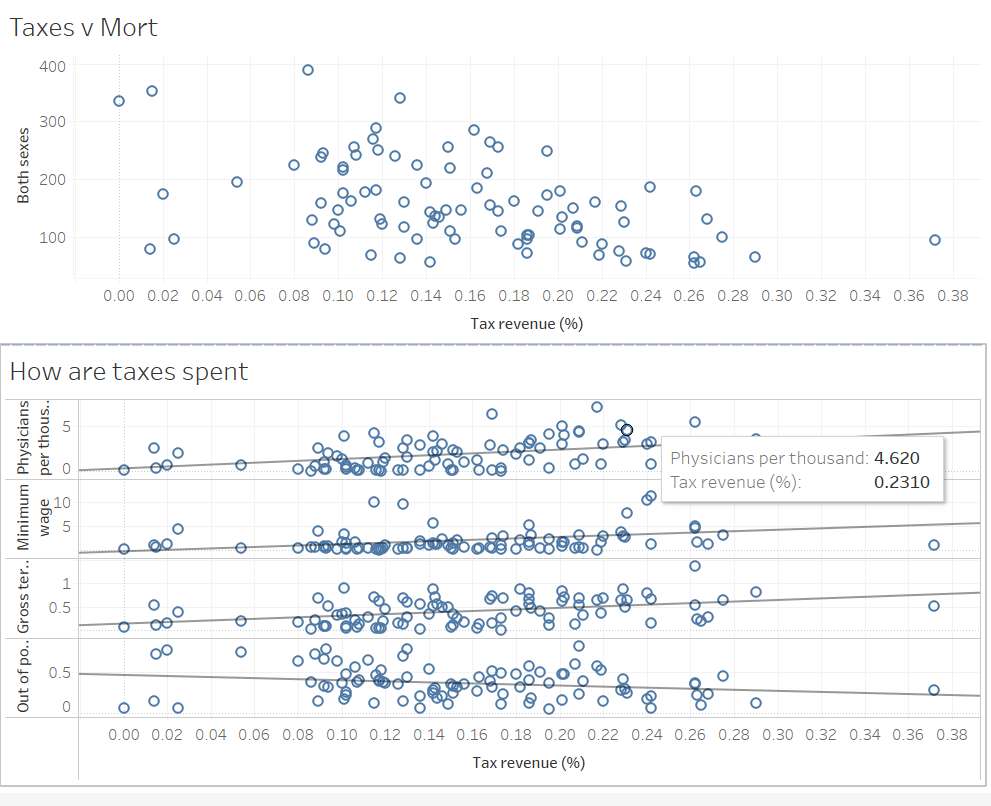
For GDP:



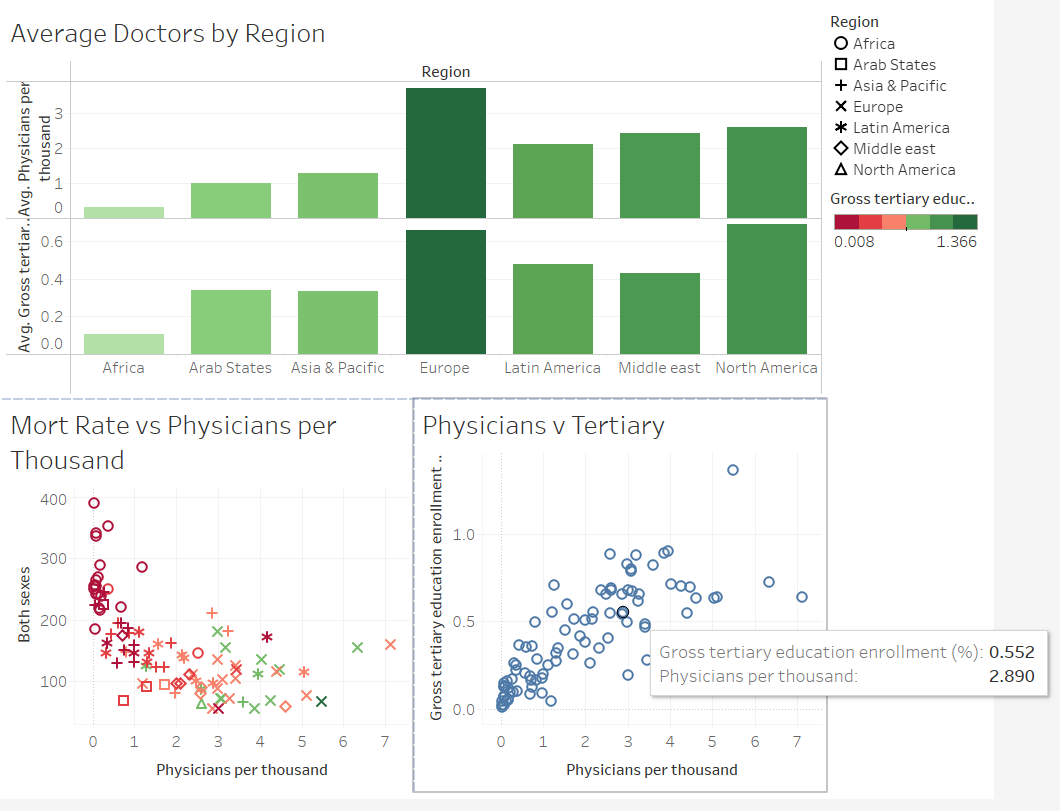
This is what the dashboard looks like normally. You are able to change the GDP range to change the visuals on the bottom left and the top right. This is done to visualize GDPs without specific outliers, such as China and the US, as seen below.



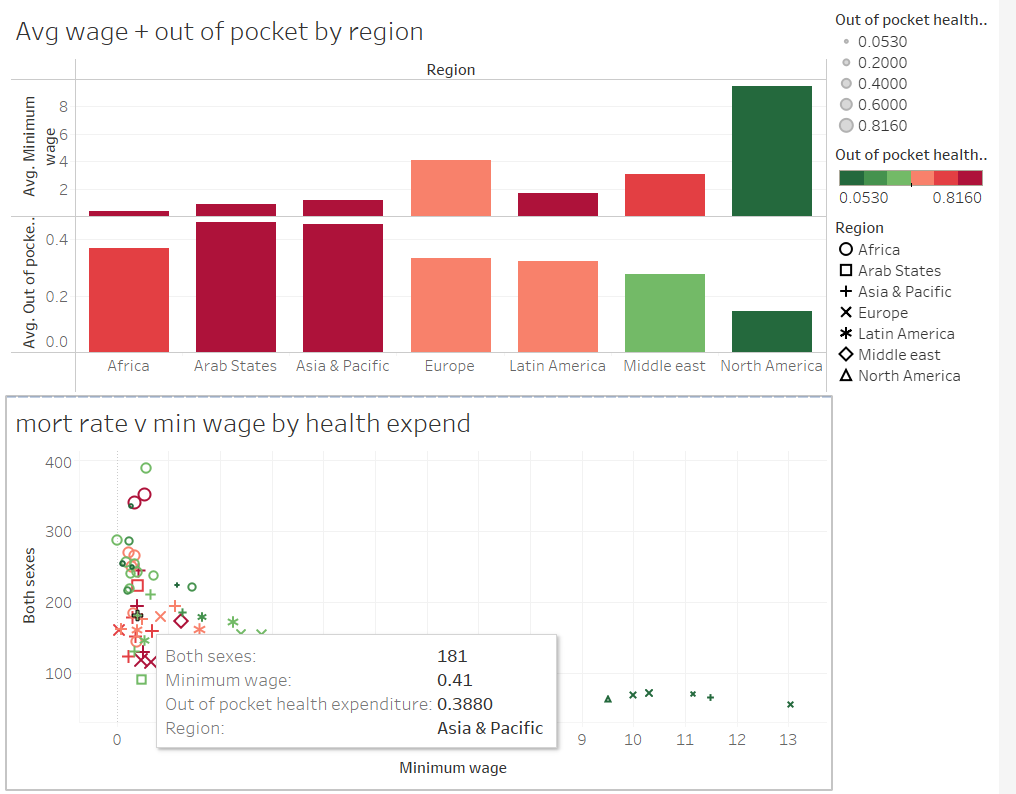
For taxes: There is no interaction outside of being able to hover over the trend line and points



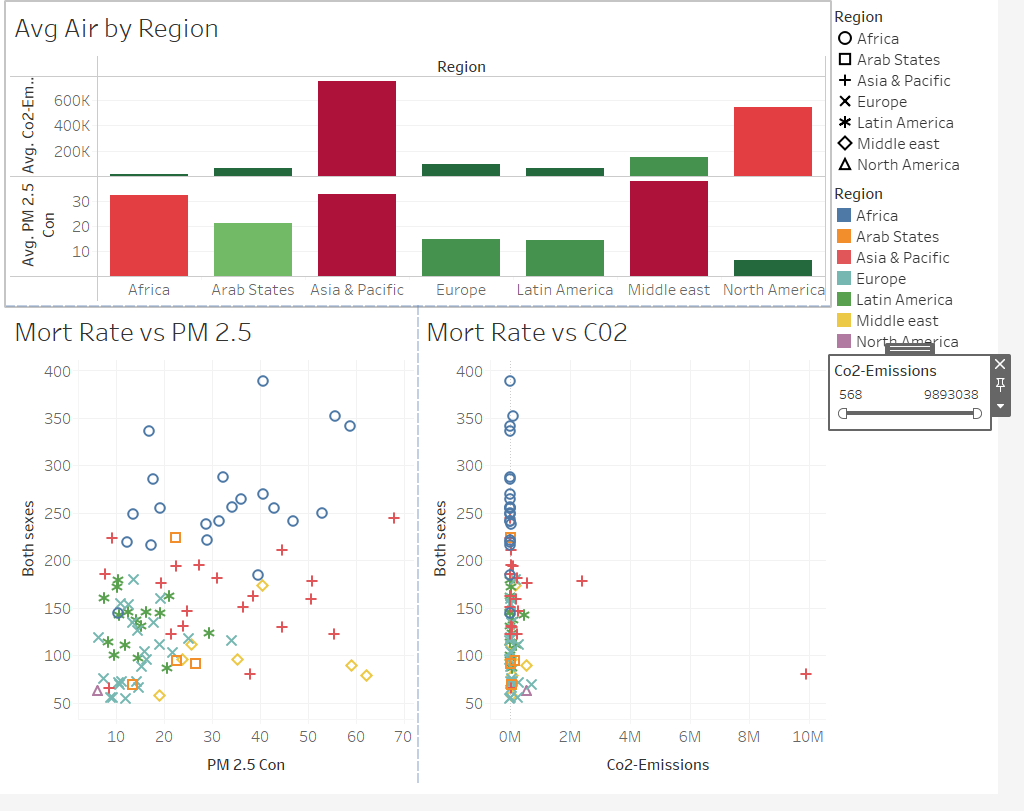
For Physicians + College: There is no interaction outside of hover data

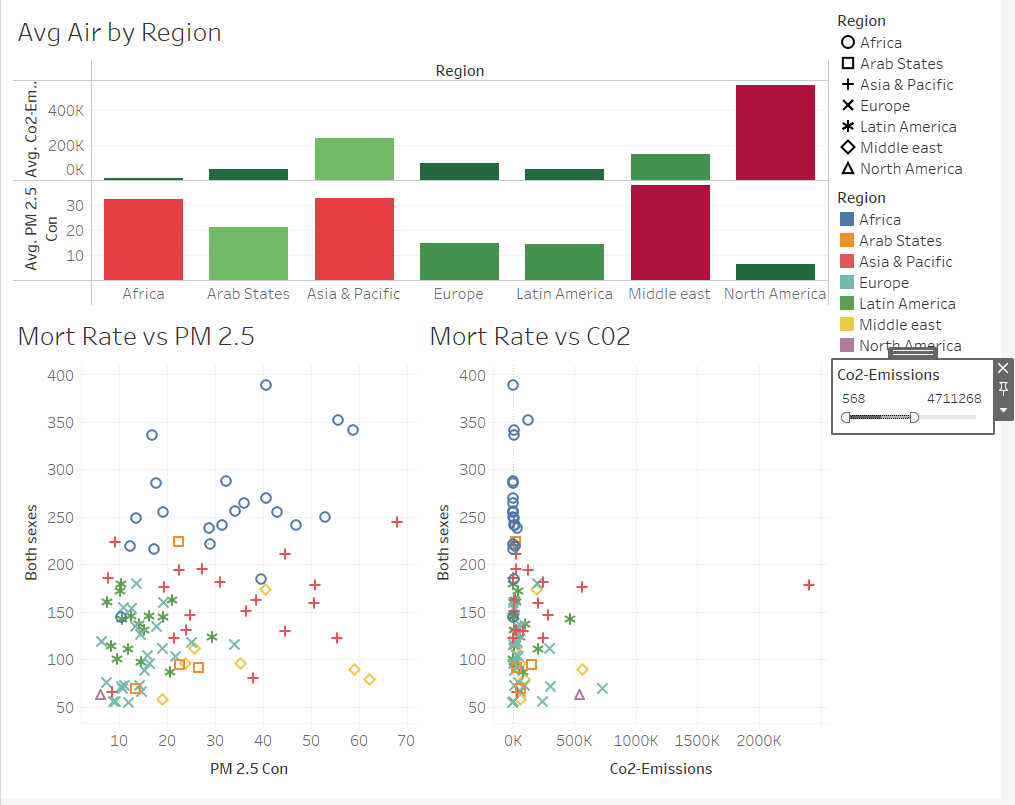


For minimum wage + out of pocket: There is no interaction outside of hover data

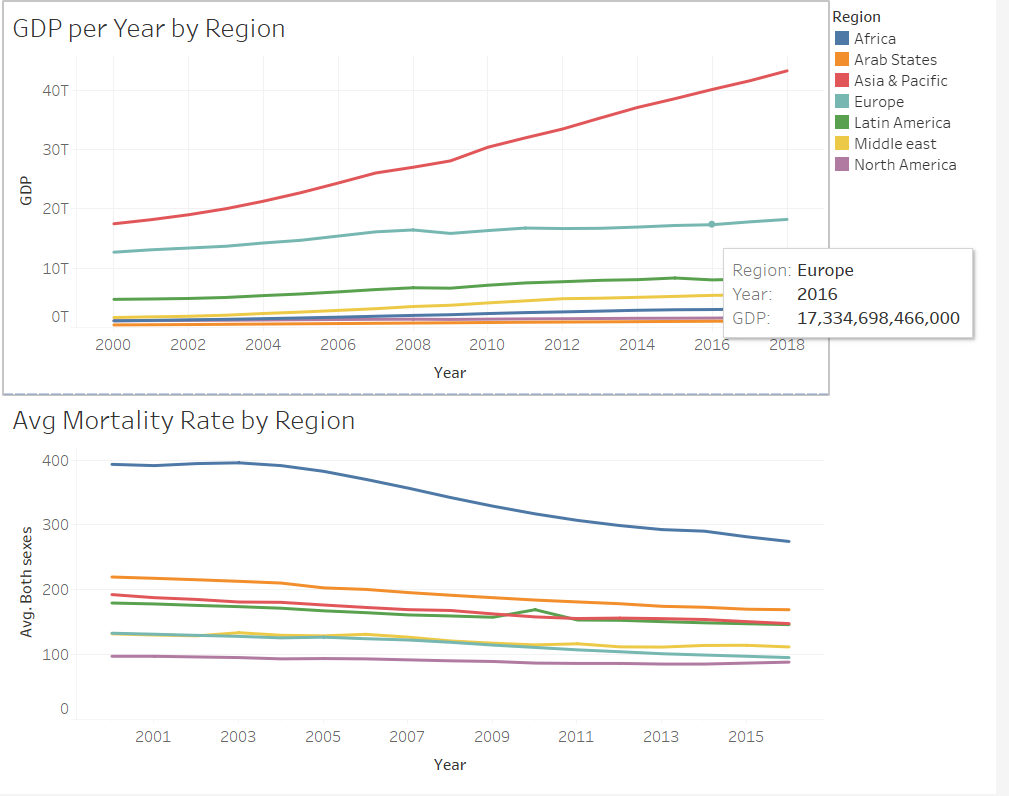


For air quality: You are able to use a slider to filter co2 emissions for outliers. The two images below are used to show the differences when applying the filter.





For GDP v Mortality: No interaction outside of hover data



**Sources (Data)**

<https://wits.worldbank.org/CountryProfile/en/country/by-country/startyear/ltst/endyear/ltst/indicator/NY-GDP-MKTP-CD>

<https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023/data>

<https://www.who.int/data/gho/data/themes/air-pollution/who-air-quality-database>

<https://www.who.int/data/data-collection-tools/who-mortality-database>

**Github**

<https://github.com/judowang/Data-230-Term-Project>

The github includes: tableau workbook, pptx presentation, docx report, the ETL + early analysis ipynb, CSVs (some may not be used, or were combined in ETL).

Note that not all ETL processes were done in the notebook. Some, such as adding columns like regions, was done in Excel. Additionally, some of the data was merged in Tableau.