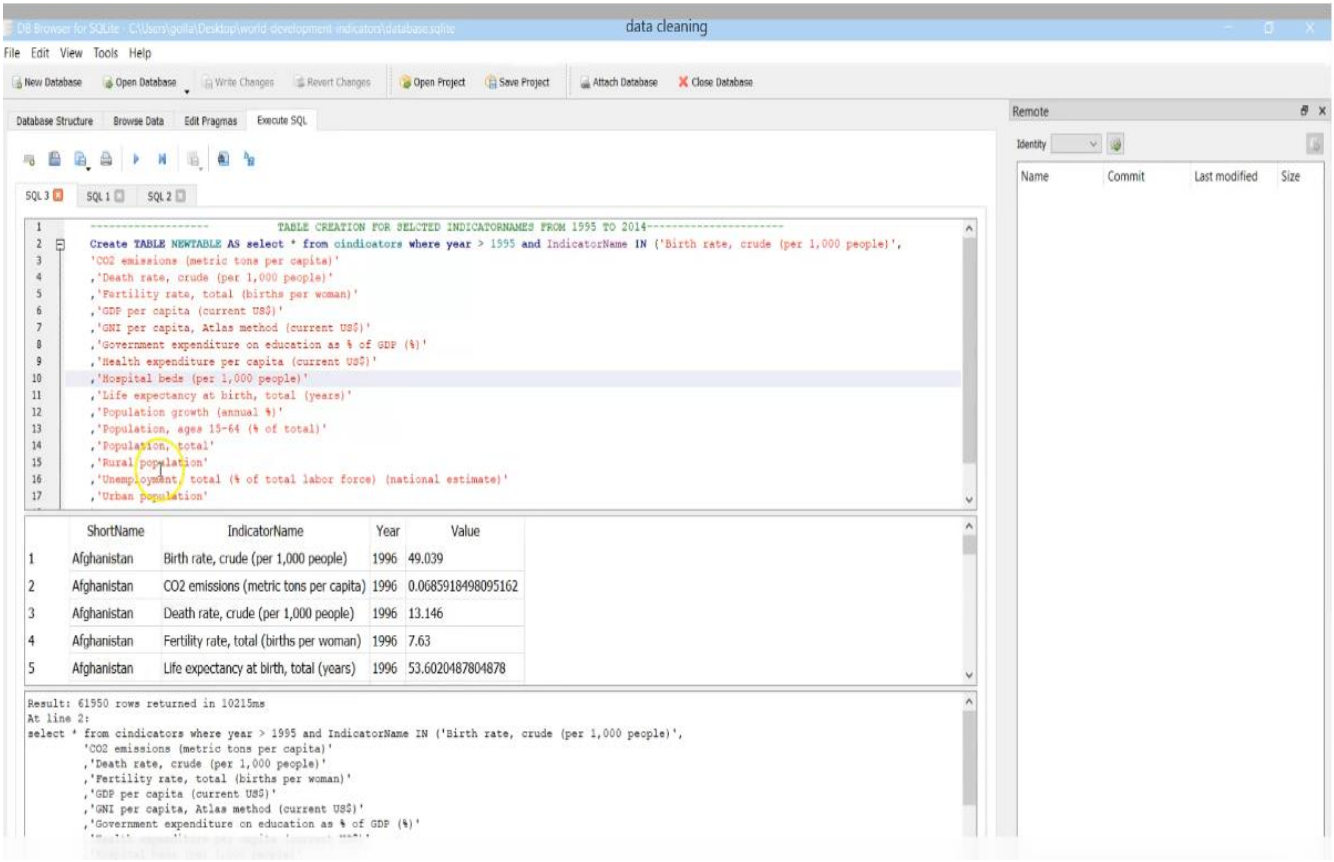
Data visualization final group project
World bank indicators
Group 3

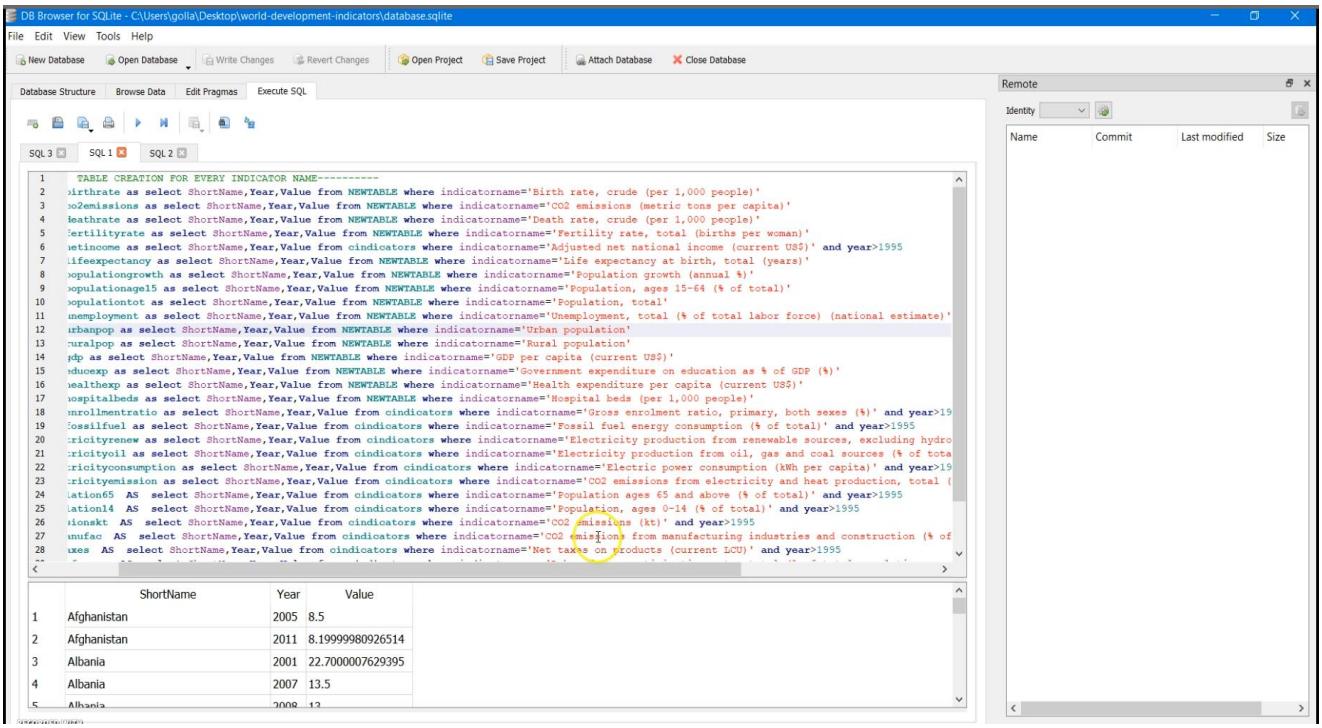
Phuong Bui
Sai Krishna Gollapally
Hsien-Lin Wu
Qihang Zhou

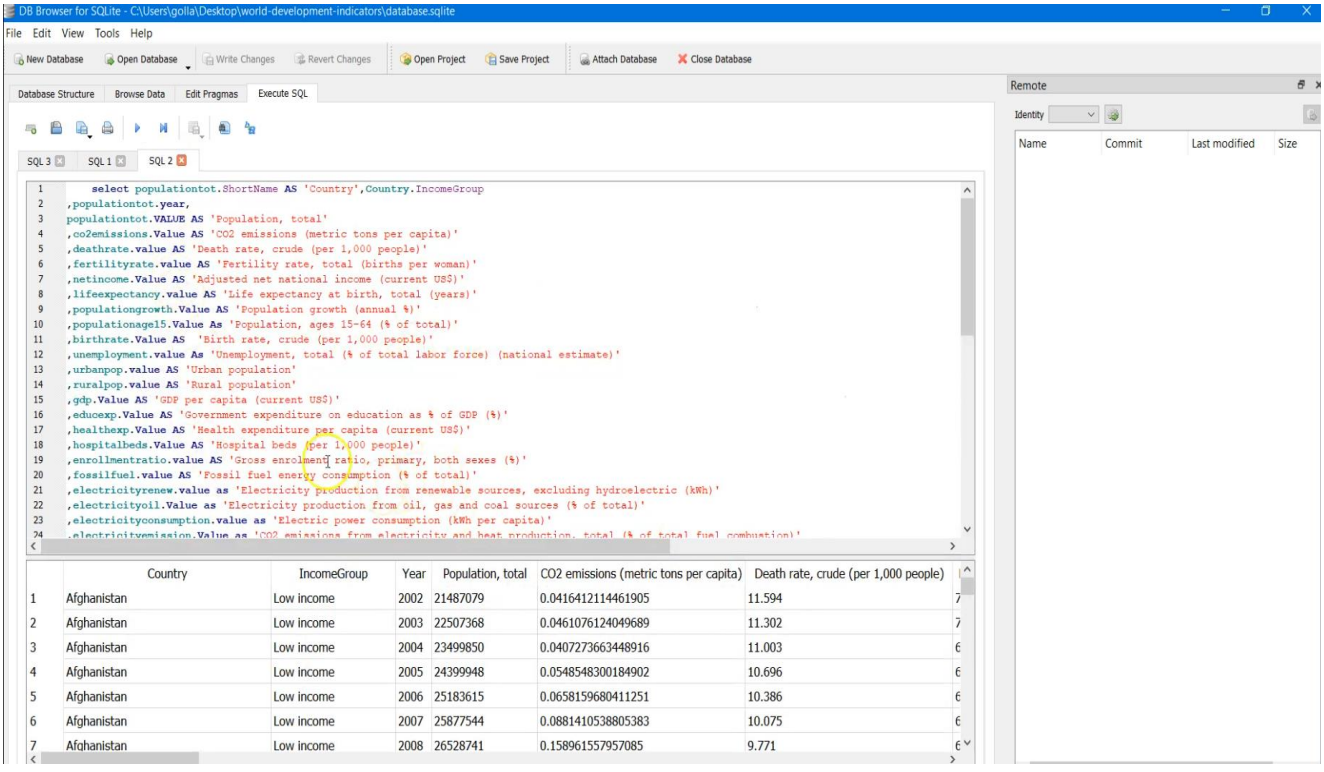
The dataset World Bank Development Indicator was obtained from Kaggle. The original dataset contains 6 tables with more than 2.5million records and uncountable null values.

The Data was in Entity Attribute value model, The tables structures and the data volume weren't ideal for us to use on tableau as they were. So we used SQL and excel to restructure and extract the original dataset into a smaller and variables oriented one. The final dataset contains 4 thousand rows with 22 variables.

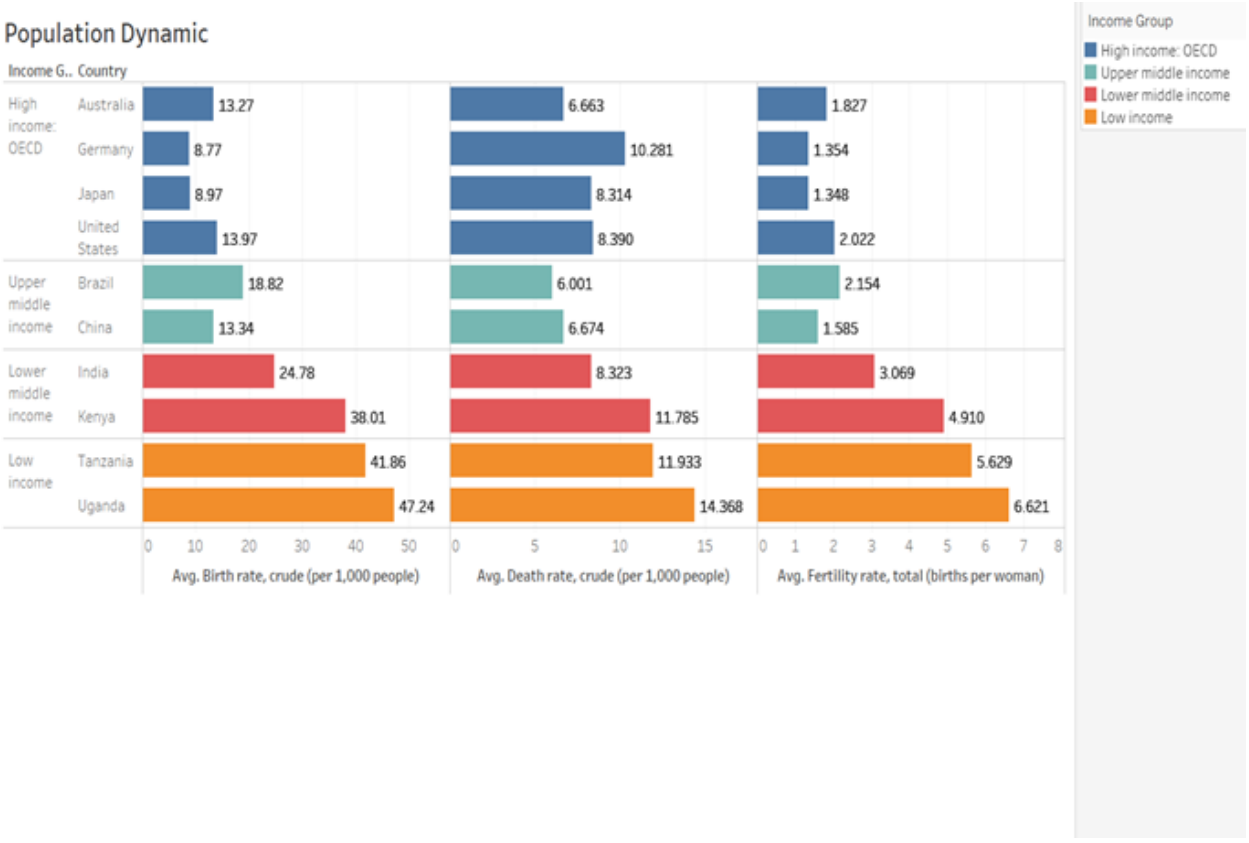
Below are the Steps involved in our Data Cleaning Process.




DB Browser for SQLite - C:\Users\golla\Desktop\world-development-indicators\database.sqlite
File Edit View Tools Help
New Database
Open Database
Write Changes
Revert Changes
Open Project
Save Project
Attach Database
Close Database
Database Structure
Browse Data
Edit Pragmas
Execute SQL
SQL 3
SQL 1
SQL 2
-- TABLE CREATION FOR EVERY INDICATOR NAME----------
birthrate as select ShortName,Year,Value from NEWTABLE where indicatorname='Birth rate, crude (per 1,000 people)'
co2emissions as select ShortName,Year,Value from NEWTABLE where indicatorname='CO2 emissions (metric tons per capita)'
deathrate as select ShortName,Year,Value from NEWTABLE where indicatorname='Death rate, crude (per 1,000 people)'
fertilityrate as select ShortName,Year,Value from NEWTABLE where indicatorname='Fertility rate, total (births per woman)'
netincome as select ShortName,Year,Value from cindicators where indicatorname='Adjusted net national income (current US$)' and year>1995
lifeexpectancy as select ShortName,Year,Value from NEWTABLE where indicatorname='Life expectancy at birth, total (years)'
populationgrowth as select ShortName,Year,Value from NEWTABLE where indicatorname='Population growth (annual %)'
populationage15 as select ShortName,Year,Value from NEWTABLE where indicatorname='Population, ages 15-64 (% of total)'
populationtot as select ShortName,Year,Value from NEWTABLE where indicatorname='Population, total'
unemployment as select ShortName,Year,Value from NEWTABLE where indicatorname='Unemployment, total (% of total labor force) (national estimate)'
urbanpop as select ShortName,Year,Value from NEWTABLE where indicatorname='Urban population'
ruralpop as select ShortName,Year,Value from NEWTABLE where indicatorname='Rural population'
gdp as select ShortName,Year,Value from NEWTABLE where indicatorname='GDP per capita (current US$)'
educexp as select ShortName,Year,Value from NEWTABLE where indicatorname='Government expenditure on education as % of GDP (%)'
healthexp as select ShortName,Year,Value from NEWTABLE where indicatorname='Health expenditure per capita (current US$)'
hospitalbeds as select ShortName,Year,Value from NEWTABLE where indicatorname='Hospital beds (per 1,000 people)'
enrollmentratio as select ShortName,Year,Value from cindicators where indicatorname='Gross enrolment ratio, primary, both sexes (%)' and year>19
fossilfuel as select ShortName,Year,Value from cindicators where indicatorname='Fossil fuel energy consumption (% of total)' and year>1995
electricityrenew as select ShortName,Year,Value from cindicators where indicatorname='Electricity production from renewable sources, excluding hydro
electricityoil as select ShortName,Year,Value from cindicators where indicatorname='Electricity production from oil, gas and coal sources (% of tota
electricityconsumption as select ShortName,Year,Value from cindicators where indicatorname='Electric power consumption (kWh per capita)' and year>19
electricityemission as select ShortName,Year,Value from cindicators where indicatorname='CO2 emissions from electricity and heat production, total (
lation65 AS select ShortName,Year,Value from cindicators where indicatorname='Population ages 65 and above (% of total)' and year>1995
lation14 AS select ShortName,Year,Value from cindicators where indicatorname='Population, ages 0-14 (% of total)' and year>1995
sionskt AS select ShortName,Year,Value from cindicators where indicatorname='CO2 emissions (kt)' and year>1995
nufac AS select ShortName,Year,Value from cindicators where indicatorname='CO2 emission from manufacturing industries and construction (% of
xes AS select ShortName,Year,Value from cindicators where indicatorname='Net taxes on products (current LCU)' and year>1995
ShortName Year Value
1 Afghanistan 2005 8.5
2 Afghanistan 2011 8.19999980926514
3 Albania 2001 22.7000007629395
4 Albania 2007 13.5
5 Albania 2008 13
Remote
Identity
Name Commit Last modified Size


DB Browser for SQLite - C:\Users\golla\Desktop\world-development-indicators\database.sqlite
File Edit View Tools Help
New Database
Open Database
Write Changes
Revert Changes
Open Project
Save Project
Attach Database
Close Database
Database Structure
Browse Data
Edit Pragmas
Execute SQL
SQL 3
SQL 1
SQL 2
select populationtot.ShortName AS 'Country',Country.IncomeGroup
,populationtot.year,
populationtot.VALUE AS 'Population, total'
,co2emissions.Value AS 'CO2 emissions (metric tons per capita)'
,deathrate.value AS 'Death rate, crude (per 1,000 people)'
,fertilityrate.value AS 'Fertility rate, total (births per woman)'
,netincome.Value AS 'Adjusted net national income (current US$)'
,lifeexpectancy.value AS 'Life expectancy at birth, total (years)'
,populationgrowth.Value AS 'Population growth (annual %)'
,populationage15.Value As 'Population, ages 15-64 (% of total)'
,birthrate.Value AS 'Birth rate, crude (per 1,000 people)'
,unemployment.value As 'Unemployment, total (% of total labor force) (national estimate)'
,urbanpop.value AS 'Urban population'
,ruralpop.value AS 'Rural population'
,gdp.Value AS 'GDP per capita (current US$)'
,educexp.Value AS 'Government expenditure on education as % of GDP (%)'
,healthexp.Value AS 'Health expenditure per capita (current US$)'
,hospitalbeds.Value AS 'Hospital beds (per 1,000 people)'
,enrollmentratio.value AS 'Gross enrolment ratio, primary, both sexes (%)'
,fossilfuel.value AS 'Fossil fuel energy consumption (% of total)'
,electricityrenew.value as 'Electricity production from renewable sources, excluding hydroelectric (kWh)'
,electricityoil.Value as 'Electricity production from oil, gas and coal sources (% of total)'
,electricityconsumption.value as 'Electric power consumption (kWh per capita)'
,electricityemission.Value as 'CO2 emissions from electricity and heat production, total (% of total fuel combustion)'
Country IncomeGroup Year Population, total CO2 emissions (metric tons per capita) Death rate, crude (per 1,000 people)
1 Afghanistan Low income 2002 21487079 0.0416412114461905 11.594
2 Afghanistan Low income 2003 22507368 0.0461076124049689 11.302
3 Afghanistan Low income 2004 23499850 0.0407273663448916 11.003
4 Afghanistan Low income 2005 24399948 0.0548548300184902 10.696
5 Afghanistan Low income 2006 25183615 0.0658159680411251 10.386
6 Afghanistan Low income 2007 25877544 0.0881410538805383 10.075
7 Afghanistan Low income 2008 26528741 0.158961557957085 9.771
Remote
Identity
Name Commit Last modified Size

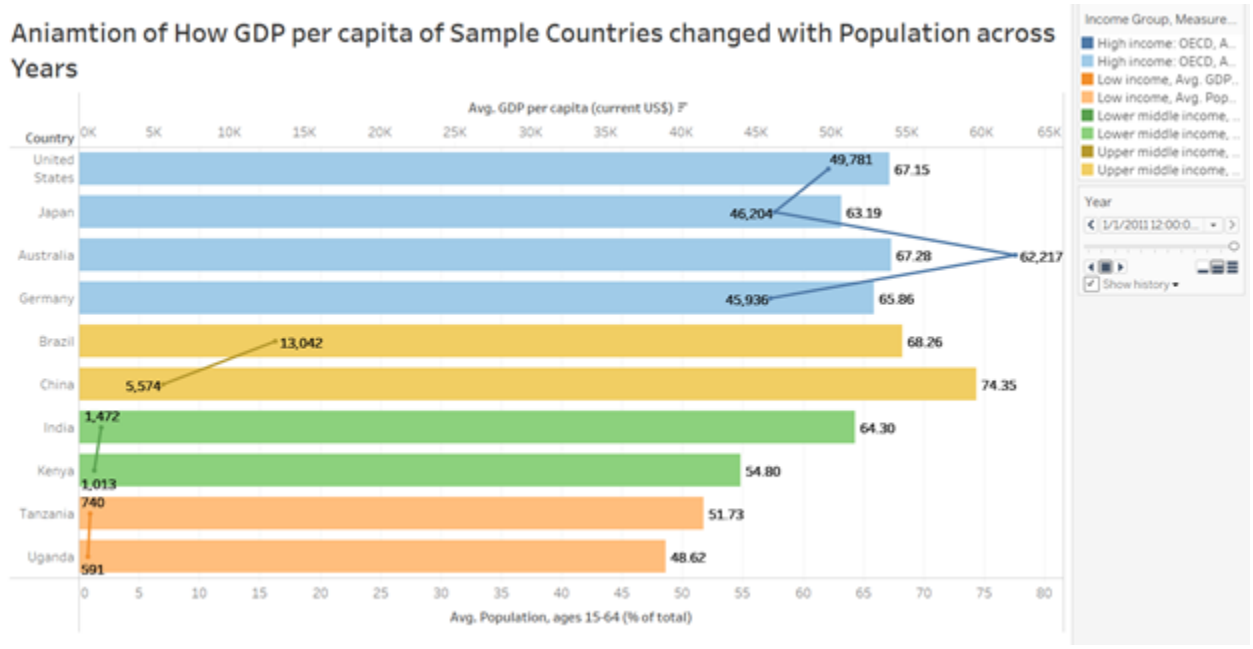
We looked into the population. Ten countries were chosen based on the income level over more than 200 countries to represent. Four countries from high-income level, which are Australia, Germany, Japan and United States. Two countries from upper middle income level are Brazil and China. Two from lower-middle-income level are India and Kenya and two from low-income level are Tanzania and Uganda.



The population dynamic graph is visualizing how countries in different income groups differ in their population Dynamics (Average Birth Rate, Average Death Rate, Average Fertility rate).

Uganda and Tanzania have the highest Birth rate, death rate, and fertility rate among all chosen countries. However, high-income countries have the lowest rate. This could be driven from the fact that women in well-developed countries have higher education level. These influence the birth rate and fertility rate. Besides, those countries are equipped with a better medical system and sanitary condition, result in the low death rate.

Based on the visualization, it is evident that the United States has a higher birth rate and fertility rate than China. This is related to immigrants overseas, especially young immigrants. On the other side, China has implemented the “One-child policy” since 1980.

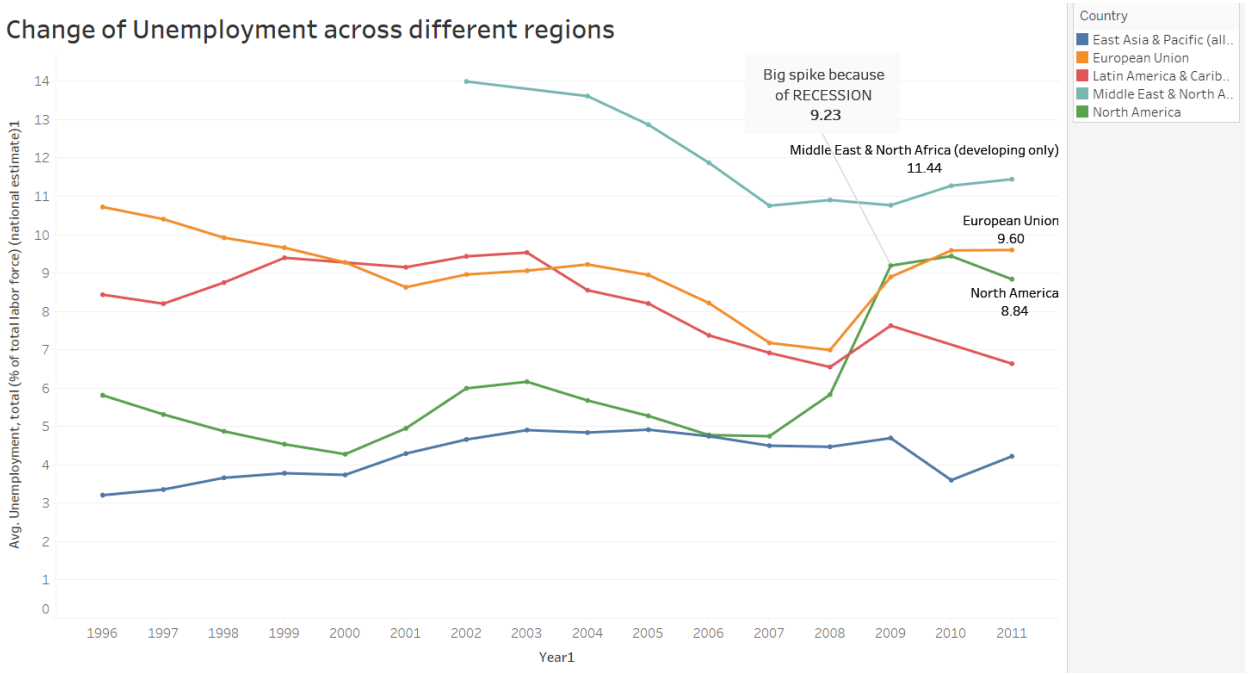


The population ages between 15-64 are related to the economy of the country. They are the labor force in the whole population who drive the economy.
For example, China has the largest labor population with a five times growth in GDP per capita during the past decade from 2001 to 2011. The GDP per capita grow from 1047 to 5574.

In the previous visualization, we saw how the GDP of countries varied with a change in their labor force. Let’s look at what factors might have influenced the growth of GDP of our sample countries. Unemployment Rate is one factor which significantly affects the growth of the economy. The unemployment rate could be defined as follows:

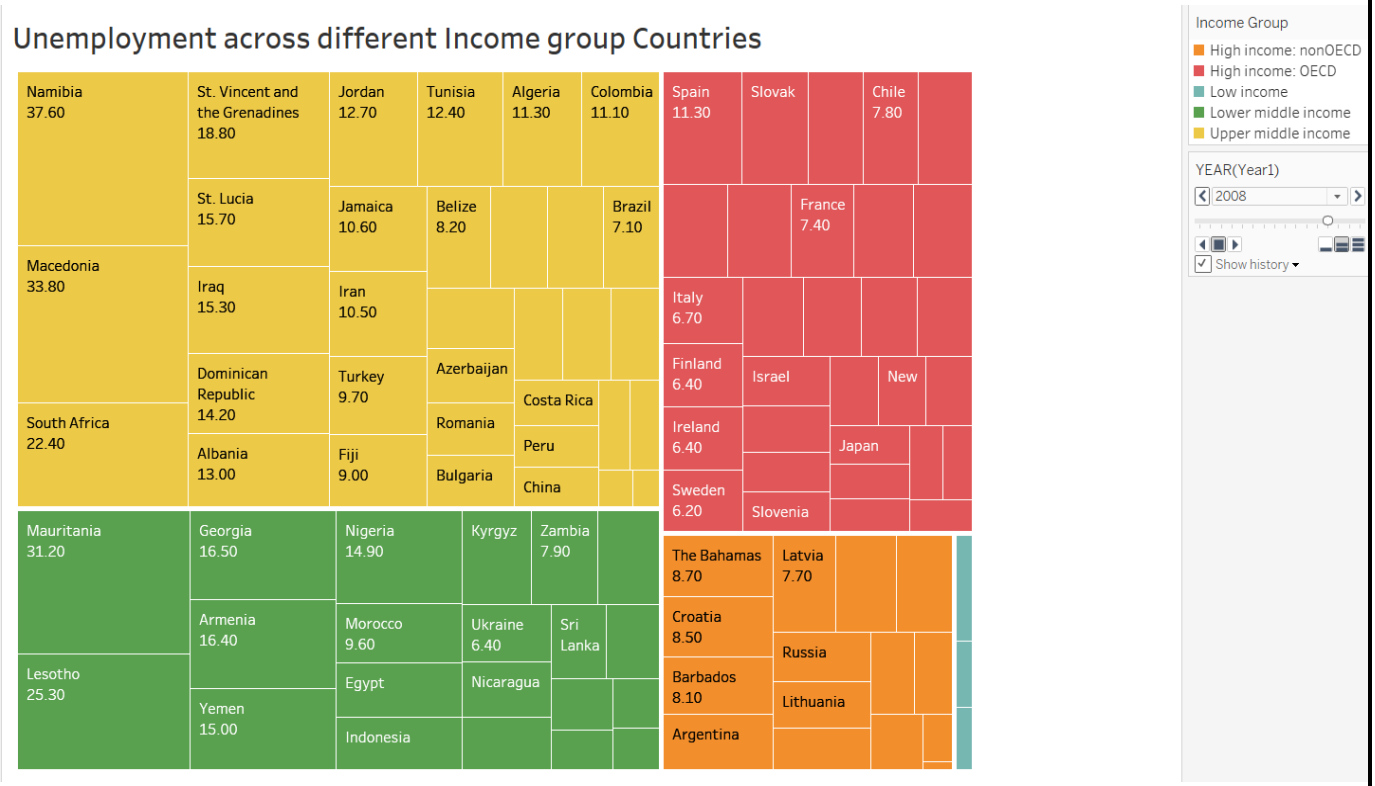
Unemployment Rate: The unemployment rate is defined as the percentage of unemployed workers in the total labor force. Workers are considered unemployed if they currently do not work, despite the fact that they are able and willing to do so. The total labor force consists of all employed and unemployed people within an economy.

This is a visualization of Unemployment across different regions in the world from the years 1996 to 2011. The Regions included in the visualization are East Asia & Pacific, European Union, Latin America, Middle East, North America.



The unemployment rate fluctuates throughout the years in every region but we could see a sharp spike in north America's unemployment rate between 2008 and 2009, at its high time the unemployment rate was 9.23%. Even the European union has a similar pattern but it was less severe though. Surprisingly, east Asia fared well and even its unemployment decreased right after the recession.

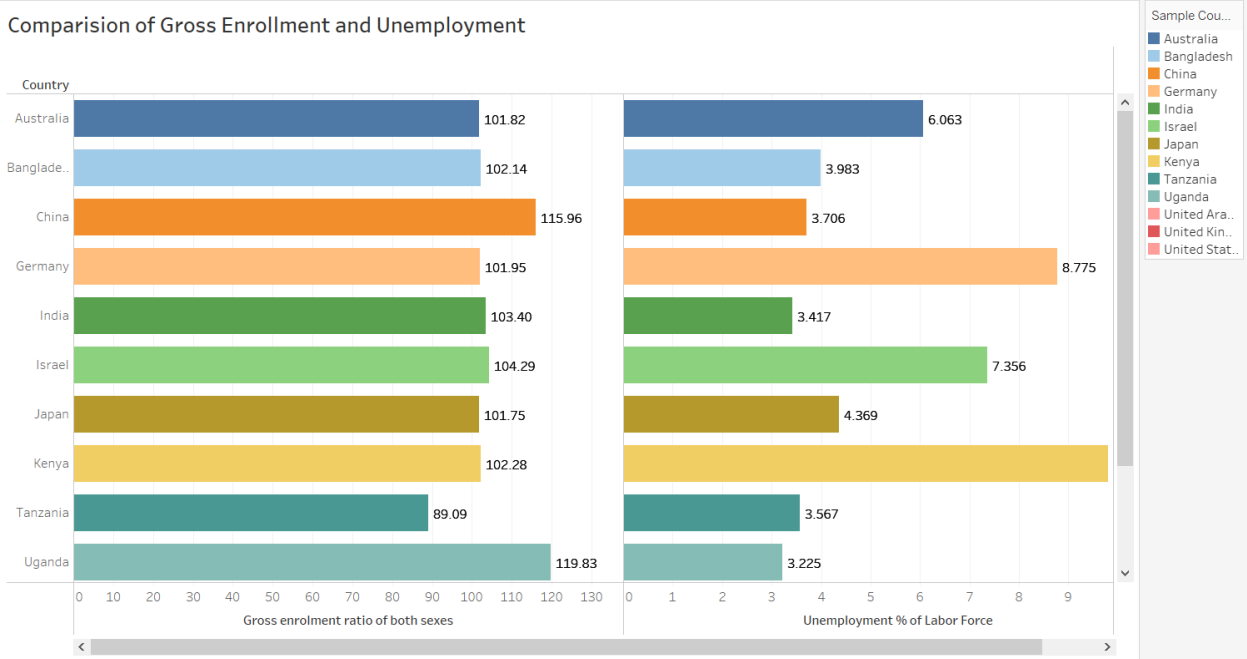
The visualization below depicts how different income groups have affected by the recession and what are the countries which were hit worst by the recession in their respective income groups during the year 2008. The below visualization is a treemap of different countries with size in correspondence to their unemployment rate and color defines their income region.



African countries are worst hit by the recession and on whole upper-middle income, the group has the most number of countries with unemployment rates in double digits. Namibia has the worst unemployment rate at 37.50, low-Income group countries are not much affected by the recession in 2008, right next to the high-Income group countries most of their countries had single-digit unemployment rates.

The comparison of gross enrollment ratios of countries and their unemployment rates to check does Gross enrollment ratio have any effect on the unemployment rate.

Gross Enrollment ratio: Gross Enrollment Ratio (GER) or Gross Enrollment Index (GEI) is a statistical measure used in the Education sector, and formerly by the UN, to determine the number of students enrolled in school at several different grade levels (like elementary, middle school and high school),The GER can be over 100% as it includes students who may be older or younger than the official age group.

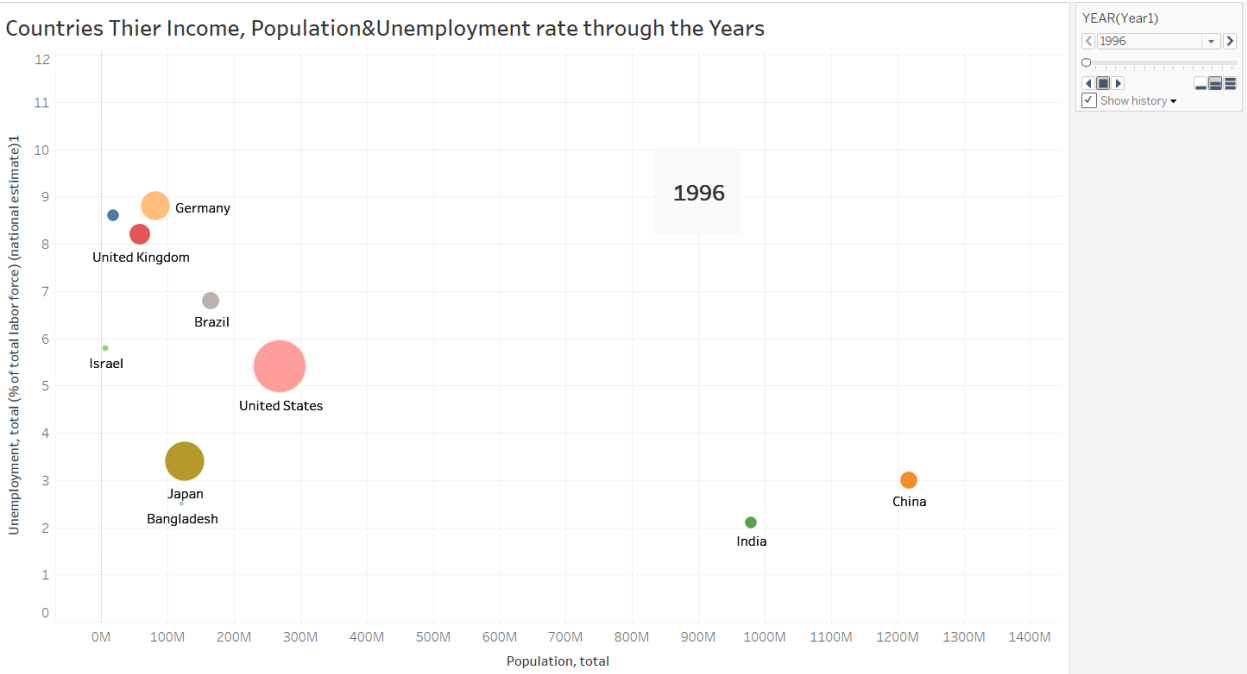


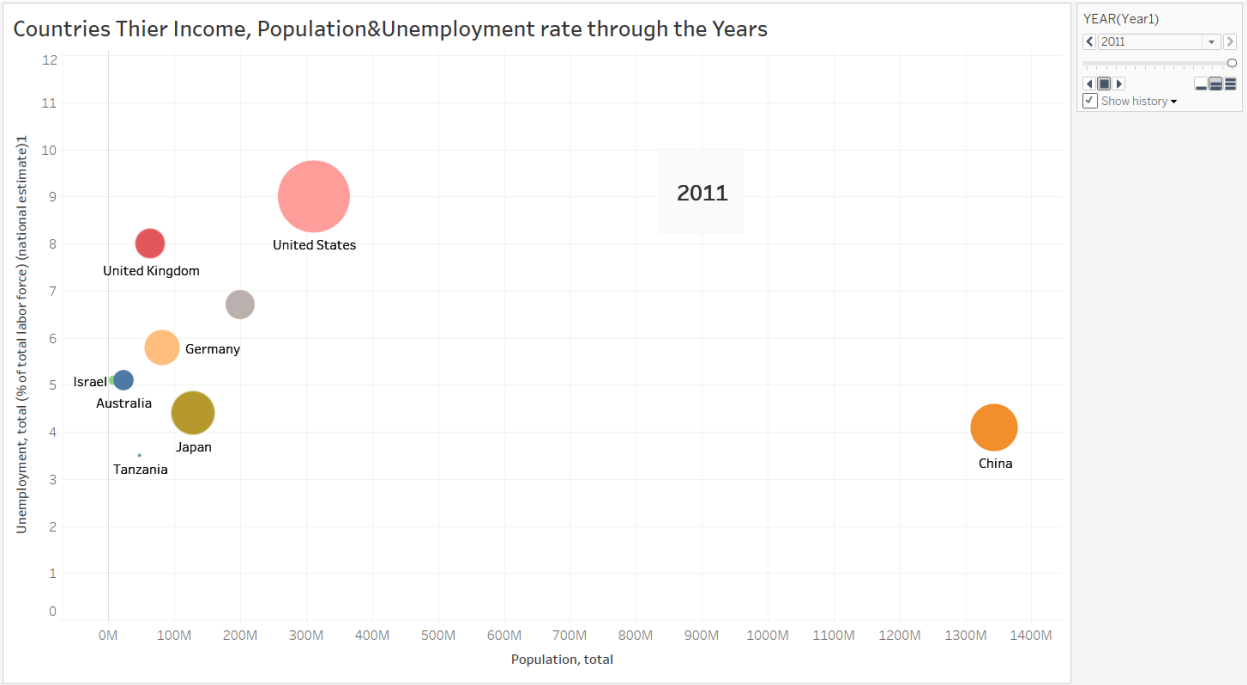
As one could see in the visualization most of the counties had somewhat similar gross enrollment ratios among our sample countries but their unemployment rates are different. For example, India and Kenya have the same gross enrollment ratios but the unemployment rate of India is 3.417 whereas unemployment rate of Kenya is 9.8.

It seems like the Gross Enrollment Ratio does not have any effect on unemployment rates of a country. Let's explore several other factors that might contribute to change in unemployment rates.

Analyze of how population & adjusted national Income has an effect on the country's population. An animation was built by adding countries to the color marks, the adjusted national income to the size marks, unemployment rates on Y-axis and population on x-Axis, The year is added to the pages shelf.

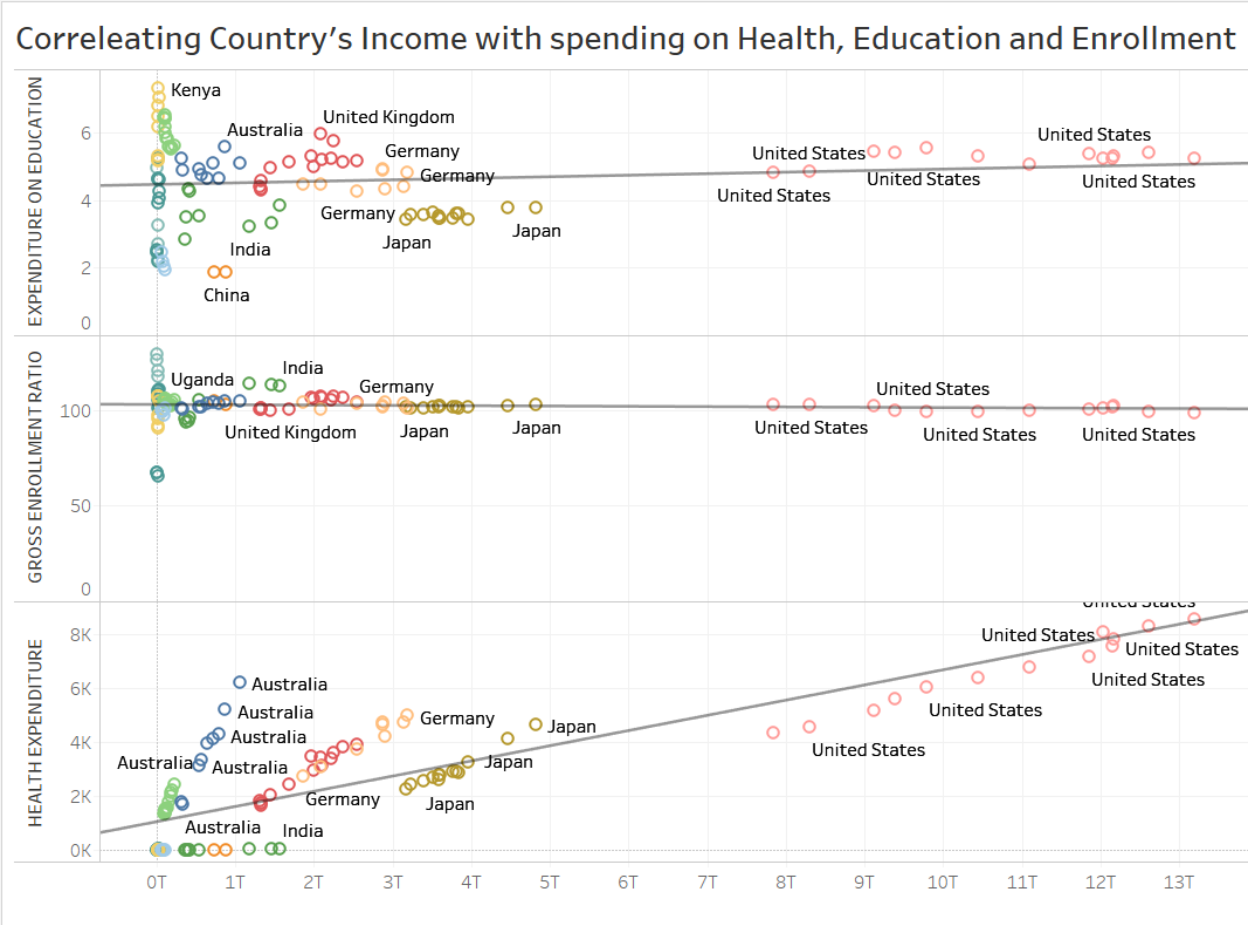
Below are the two visualizations one in 1996 and one in 2011, this gives us a glimpse of what happened in 15 years in our sample countries population, unemployment rate and income.





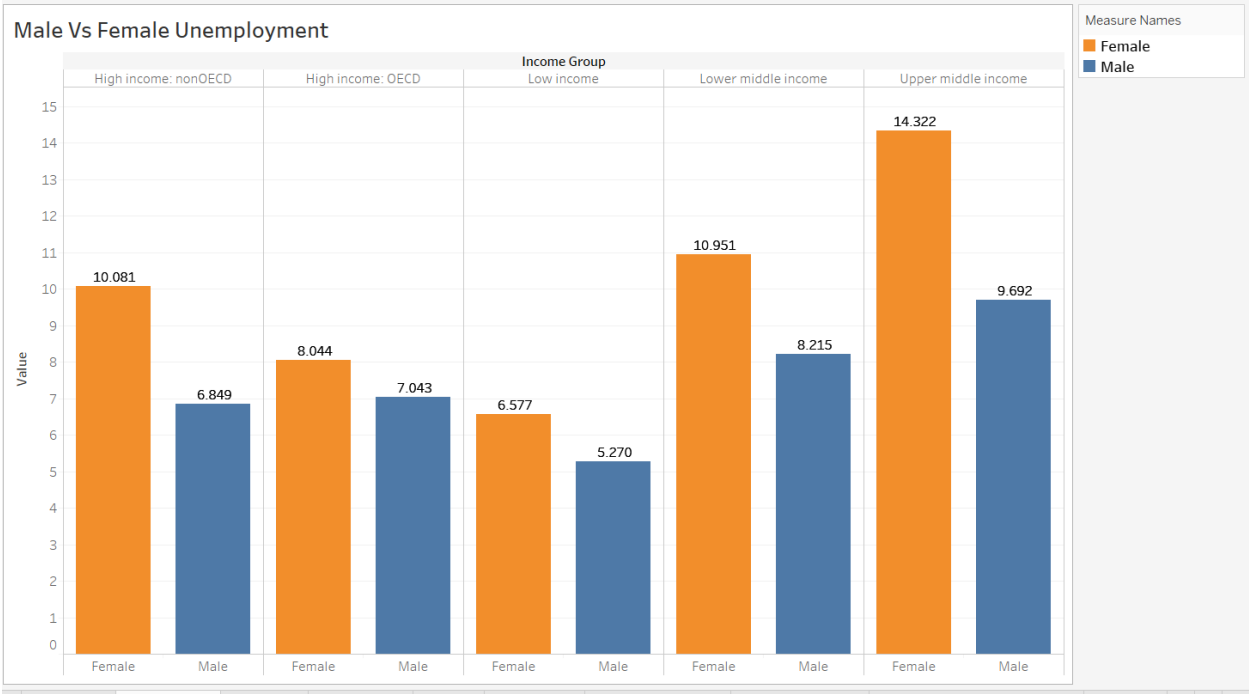
Almost every country's population has increased and unemployment rate fluctuated across the years, even the adjusted income of every country has increased over the years, United States unemployment rate has peaked right after the recession and stayed high for a long time whereas china’s and India's didn’t change much.

This visualization gives us information on how our sample countries spend their income and do that have anything do with unemployment rates across the years. A trend line was fit to emphasize the relationship.



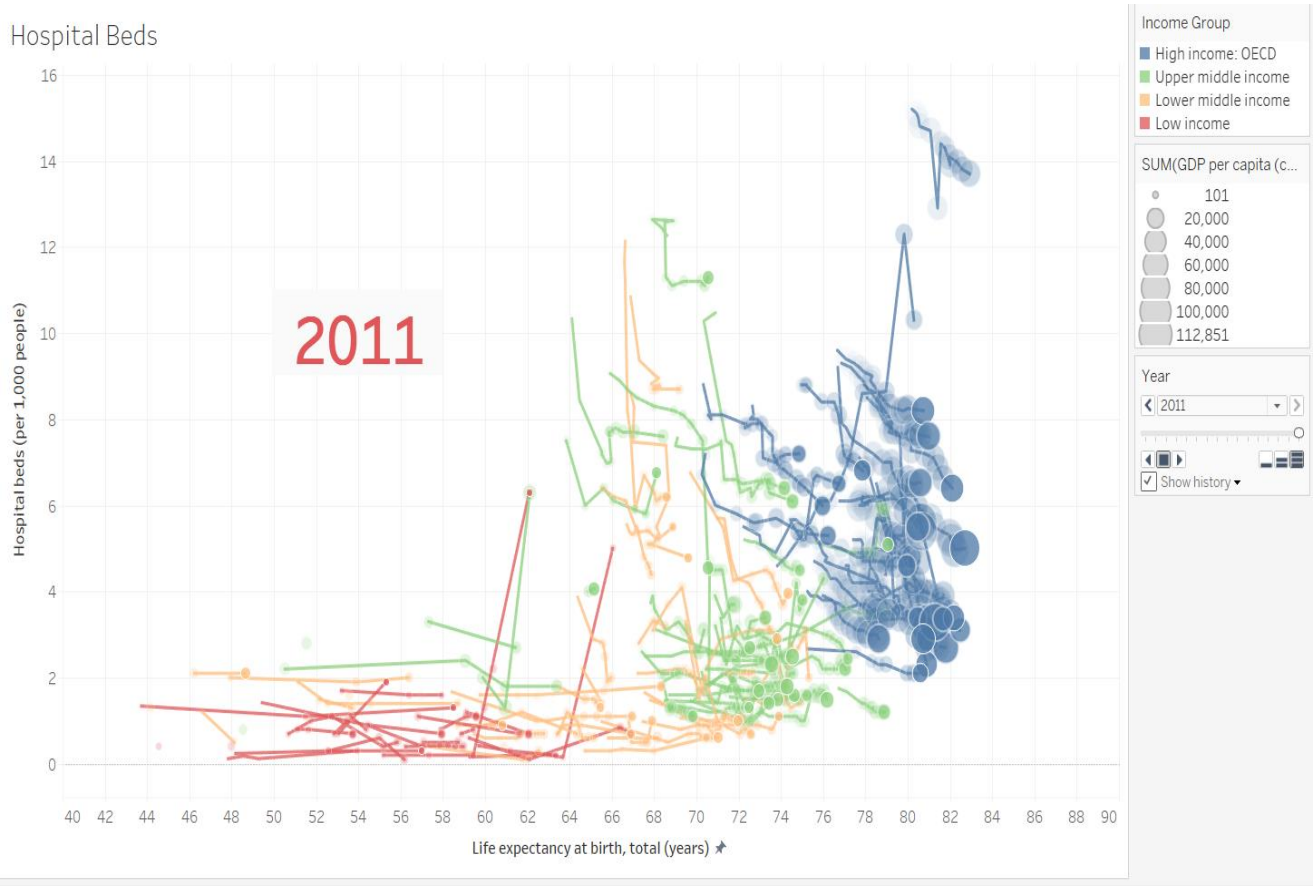
Almost every year each country's income has increased and their spending on health care and education have also increased, despite the increase in spending on education the gross enrollment ratio of countries haven't improved much, the trend line more or less flat.

Compared unemployment rates of aale and female across different Income groups to check whether there is a universal similarity in their difference or does other factors influence their disparity.



There is a difference between male and female unemployment rates is not that significant in high-income countries but in lower and upper middle-income group countries, the disparity is significant and clearly visible. The difference could be attributed to that fact that the majority of the countries in this region are patriarchal societies where females work less.

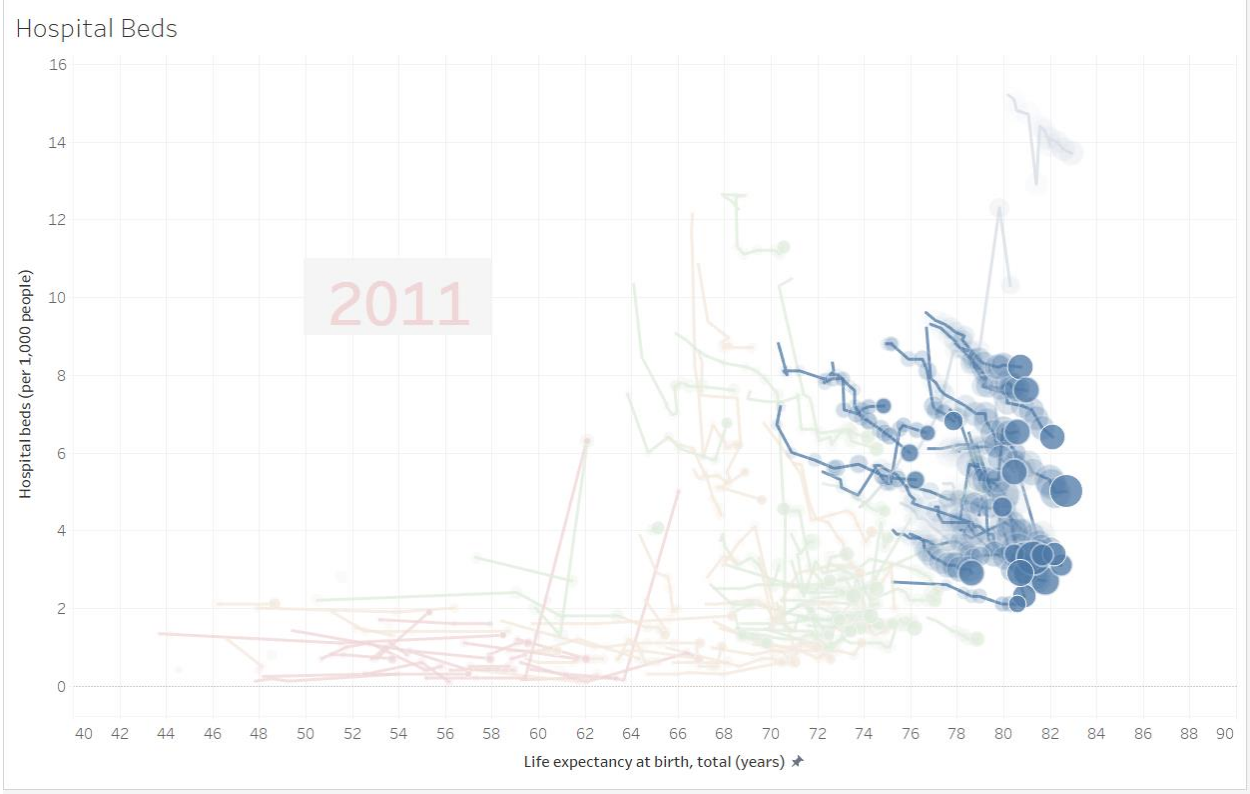
In this part, three indicators were evaluated: Hospital beds per 1,000 people in the Rows, Life expectancy at birth in the Columns, and GDP per capita is shown by the Size.



Income group was added to Color marks. Blue is for high income, green is for upper middle, yellow is for lower middle, and red is for low income.

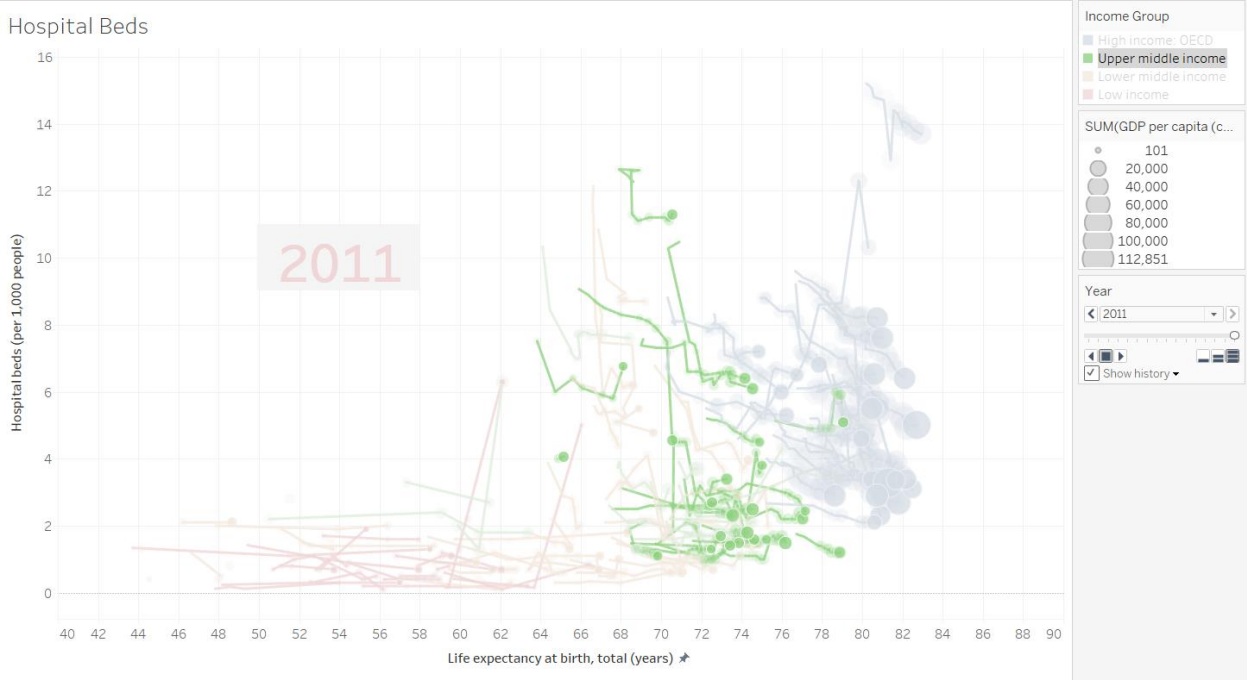
The above visualization is how those three indicators have changed in relation to one another during a period of 1996 to 2011. There is nothing much to infer based on that visualization

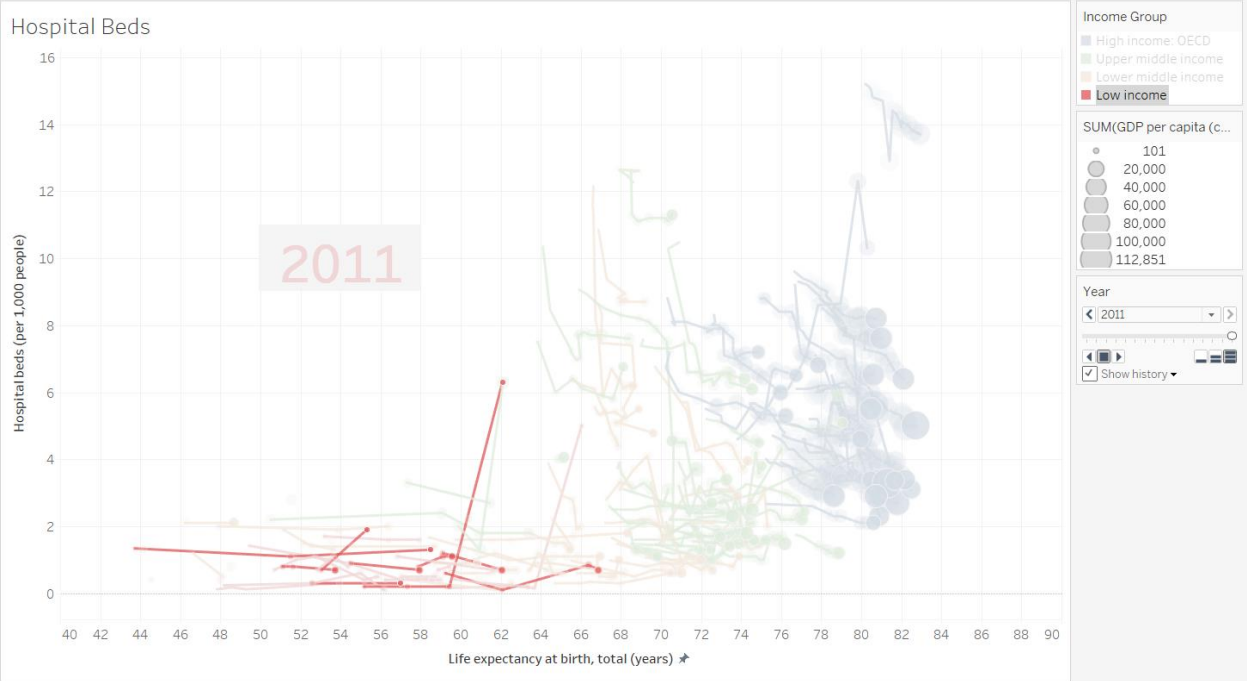
Analyzed each individual group and traced it's their changes in that period.



Above visualization is of how several high-income countries like the US, Germany, Japan, and Australia have changed,they all have a similar trail shape. They keep moving to the lower right corner throughout the years. Means their lives were going up, but the numbers of hospital beds per capita were shrinking.

On analyzing how middle-income group countries have It is clear that some middle-income countries like China, Brazil, and India didn't change much. There are some countries whose trails show very different shapes. They are basically flat or at least not as sloping as high-income Countries. Below Visualizations supports the observation





In the high-income group; countries are similar to each other. The second one is the upper-middle group; part of them are basically flat, but others are strangely dropping, even more, obvious than the high-income group. Then in the lower-middle group, they are similar to the upper-middle group that there were some countries whose trail

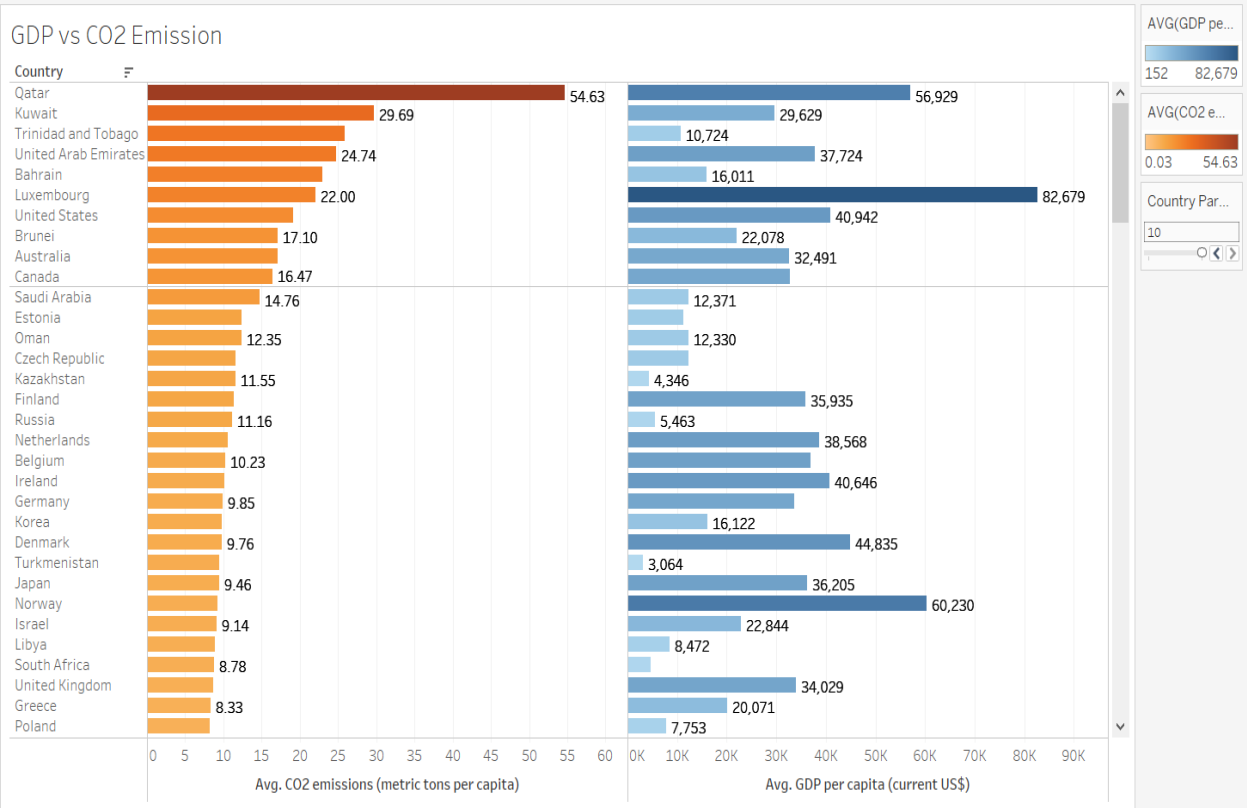
dropped strangely. The last group is low-income, Most of them are flat, but there are two countries: Ethiopia and Nepal are going up dramatically.

Based on research, the reason why hospital beds per capita were shrinking in high-income countries is probably the following: the progress in medical science; the higher efficiency on hospital management; More people choose to stay on other health facilities because their health problem was not so severe.

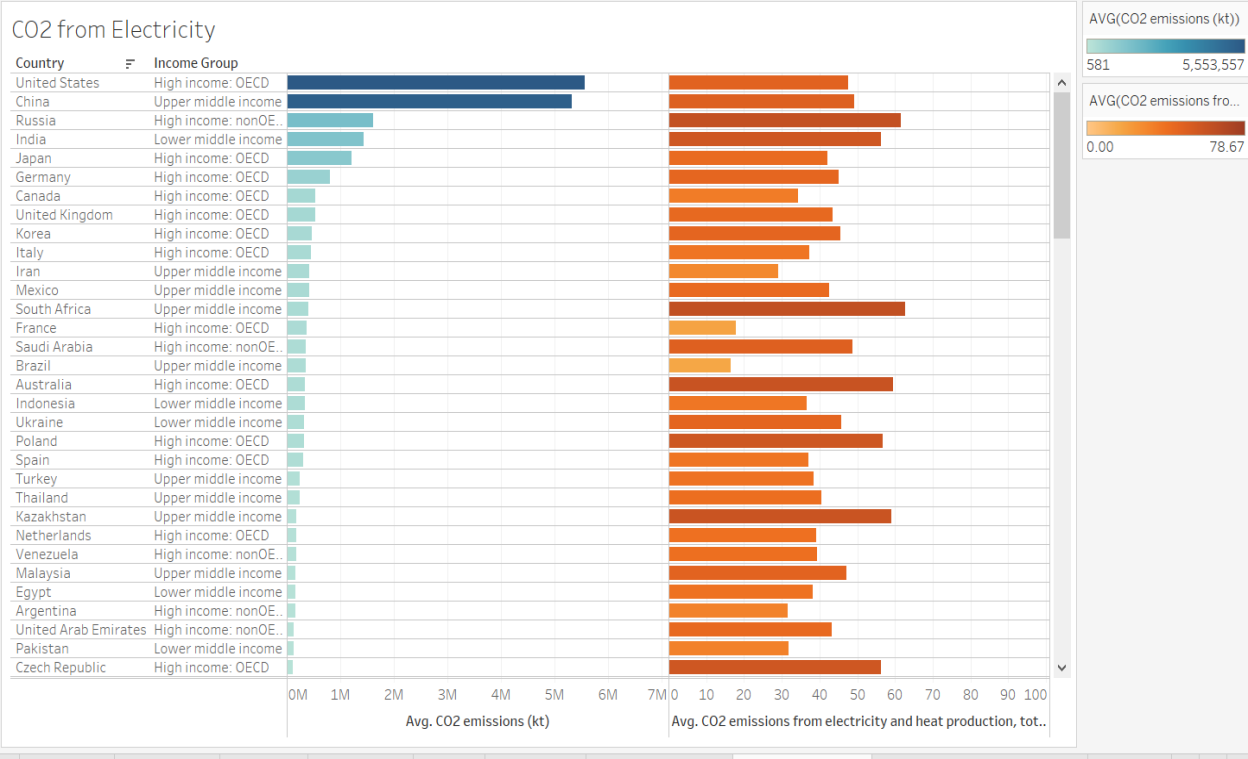
About other countries, their trails seem flat because their population and number of hospital beds were growing at the same time.

But what happened to the countries in this part? What are these countries? We found out these are mostly former Soviet socialist countries or Soviet's allies. They had high numbers of hospital beds but shrank dramatically. The major reasons for this change were likely due to Their political and economic system changes in the 90s.

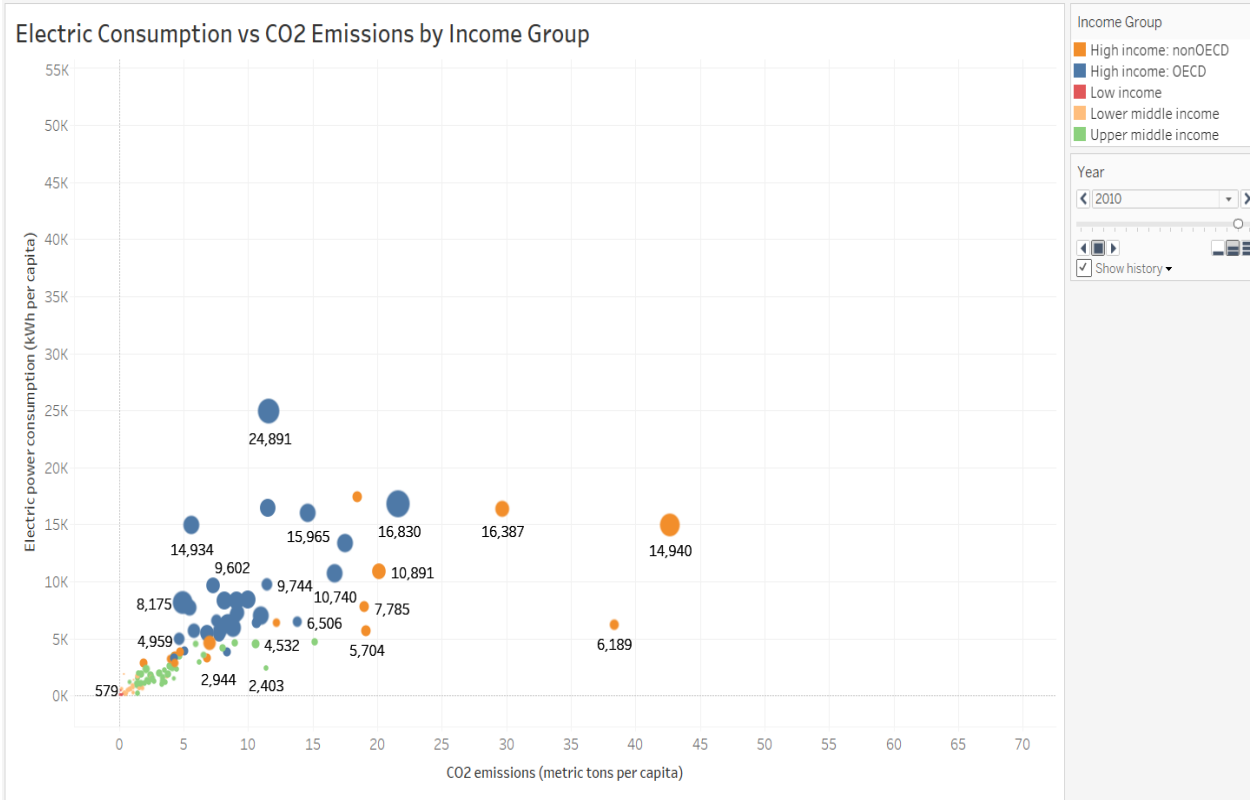
Achieving and maintaining such highly developed industry also bring drawbacks, especially to the environment. There are several indicators in the dataset which present the amount of carbon dioxide released to the air from various sources. The graph below shows the comparison between GDP and carbon dioxide emission per capita for every country.



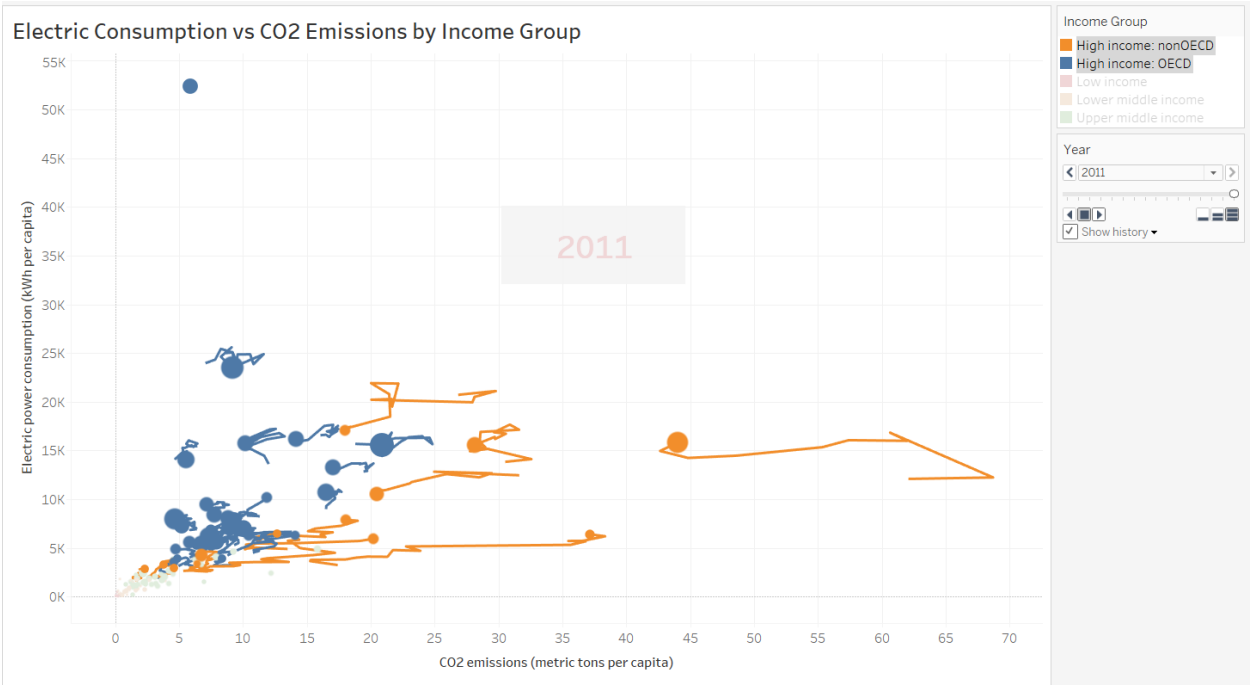
Out of the top 10 countries that have the most amount of CO2 emission per capita, four countries are high-income countries from Middle East regions. This make sense since they are well-known for their availability in fossil fuel resources. Other countries in the list are also in high-income group, but they are not from Middle East region. The reason to have such large amount of CO2 emission could come from their industry activities.



The chart is presenting the percentage of CO2 emission from electric and heat production to the total CO2 amount. This factor clearly contributes to an enormous amount of CO2 to the total. For the majority of developed countries, these activities produce almost to more than half of the total CO2 amount, no matter how much the total amount of CO2 each country releasing to the air. Therefore, if the CO2 emission from this sector could be reduced, the total amount of CO2 released could be reduced significantly.

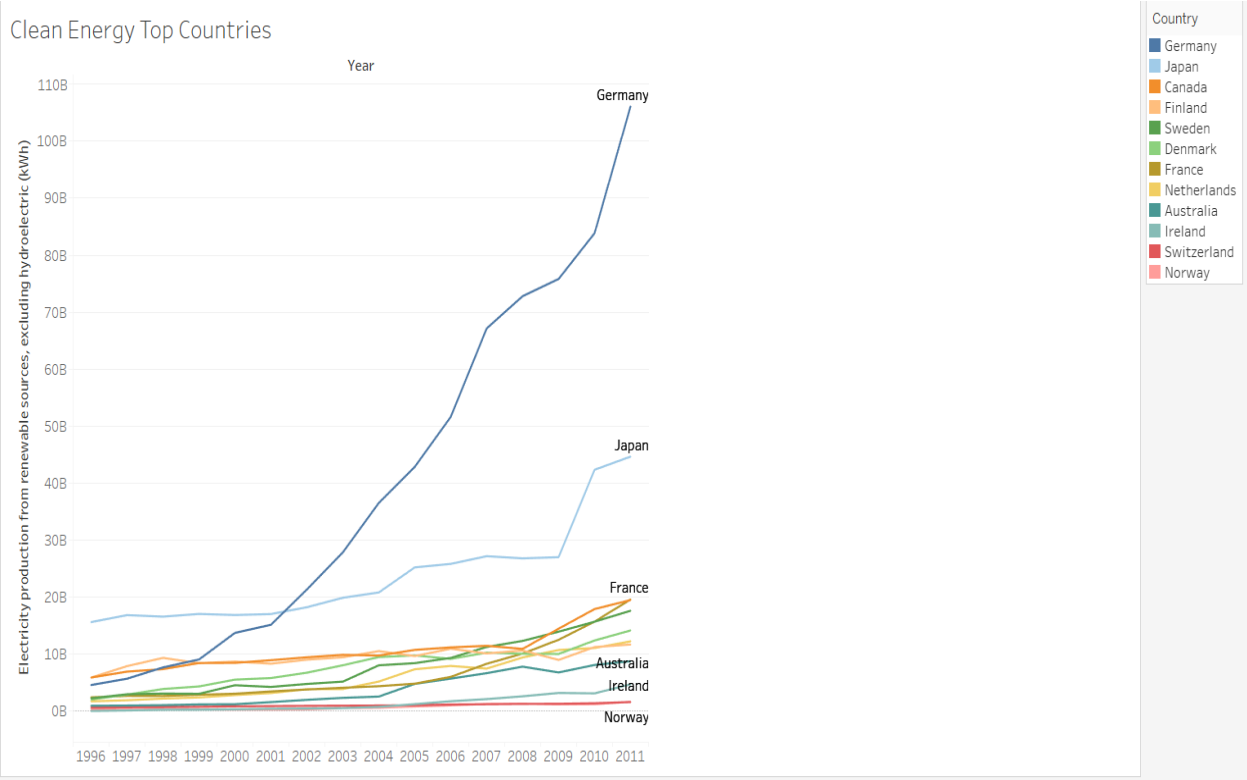


The graph is presenting the correlation between the amount of CO2 emission with the electric power consumption among countries. Colors are decided by income groups.



By the movement of the visualization, It is clearly that the blue marks which are the countries consume a lot of electric power but also maintaining very low CO2 emission production. Comparing with the orange group, they have very similar figures on electric power consumption but producing 4-5 times CO2 more than the greens do. Eventhough both blue and orange groups are countries with high level income, Blue countries are members of OECD (Organization of Economic Coporration and Development) which focus on both improving economic and social-well being. The orange countries are not members of this group.

According to visualization, blue countries must substitute their electric power sources of fossil fuels by other types of energy which could provide power without producing as much CO2 to the environment. We call it “clean energy”.



Among blue countries in the previous visualization, the chart is presenting top countries which are the leadings in electricity production from renewable sources rather than natural resources. In the past 15 years, this aspect is being received much more focus. The amount of electricity power produced from renewable sources in those countries has been constantly increased in the past fifteen years, especially in Germany and Japan where this figure raised from 2 to 10 times. Clean energy would be considered as

an alternative power source for fossil fuel in the future. With the high demand for energy and electric power for manufacturing activities and construction, developed countries are being required to find a better solution than using natural fuels. The graphs and charts show that they are leaning towards in substituting fossil fuel by renewable energy which is expected to cause less harm to people and the environment.

Conclusion :

Analyzed & visualized a sample group of countries from different regions and different Income groups on Indicators such as GDP, Labor Force, unemployment rate, Gross Enrollment ratio, Adjusted National Income, Hospital beds per 1000 people, Co2 emissions etc.. Discovered patterns and trends and researched supporting theories which form a basis for the trends. Created a Narrative on how country's different Indicators changed across the years and visualized it.