Data Visualization Project - Dataset 7

The American University of Cairo

DSCI 2411

Dr. Seif Eldawlatly

9th of December 2022

Fall 2022

Farida Madkour

Zeina Bassiouny

Sama Amr

Introduction

In the Data Visualization final project, we choose Dataset 7, which includes information about technological devices and services, population, and economic state of different countries. In this report, we will summarize the main attributes of the dataset, identify relationships between attributes, and explore the possible occurrence of trends for countries that have similar economic states.

Description of the Dataset

This dataset contains 7 attributes, and they include GDP total, population, income per person, total broadband subscribers, total cell phones, internet users, and total personal computers.

GDP total, yearly growth: Based on Gapminder's GDP per capita, PPP

Population: Total population per year for each country.

Income per person (GDP/capita, PPP$ inflation-adjusted): Gross domestic product per person adjusted for differences in purchasing power (in international dollars, fixed 2017 prices, PPP based on 2017 ICP).

Broadband subscribers (total): Fixed broadband subscriptions refer to fixed subscriptions to high-speed access to the public Internet (a TCP/IP connection), at downstream speeds equal to, or greater than, 256 kbit/s. This includes cable modem, DSL, fiber-to-the-home/building, other fixed (wired)-broadband subscriptions, satellite broadband, and terrestrial fixed wireless broadband. This total is measured irrespective of the method of payment. It excludes subscriptions that have access to data communications (including the Internet) via mobile-cellular networks. It should include fixed WiMAX and any other fixed wireless technologies. It includes both residential subscriptions and subscriptions for organizations

Cell phones (total): Mobile cellular telephone subscriptions are subscriptions to a public mobile telephone service that provide access to the PSTN using cellular technology. The indicator includes (and is split into) the number of postpaid subscriptions, and the number of active prepaid accounts (i.e. that have been used during the last three months). The indicator applies to all mobile cellular subscriptions that offer voice communications. It excludes subscriptions via data cards or USB modems, subscriptions to public mobile data services, private trunked mobile radio, telepoint, radio paging, and telemetry services.

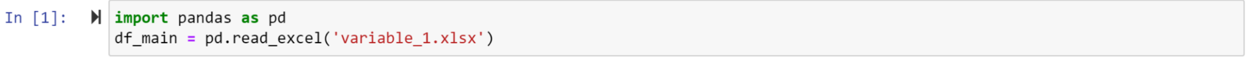
Individuals using the Internet (% of population): The percentage of the population that is using the internet.

Personal computers (total): Total number of computers per person.

Data Preparation

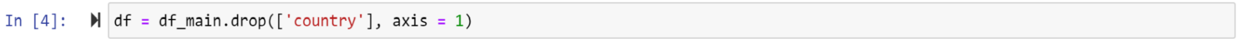
Each variable in our dataset was included in a separate spreadsheet, and we used Python in order to change characters like ‘K’, ‘M’, and ‘B’, which represent thousands, millions, and billions.

1. **We Imported the ‘pandas’ library and the excel sheet for each variable.**



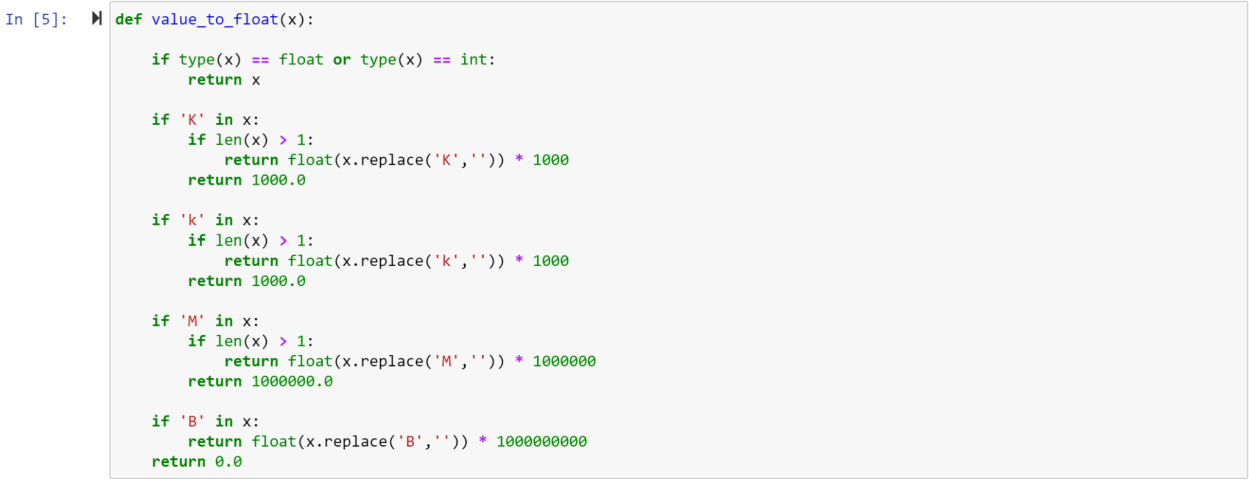
```
In [1]:  import pandas as pd
         df_main = pd.read_excel('variable_1.xlsx')
```

2. **In order to apply the changes to the values for each year, we dropped (removed) the ‘country’ column since it is of a type string. Then, we assigned this new modified data to ‘df’.**



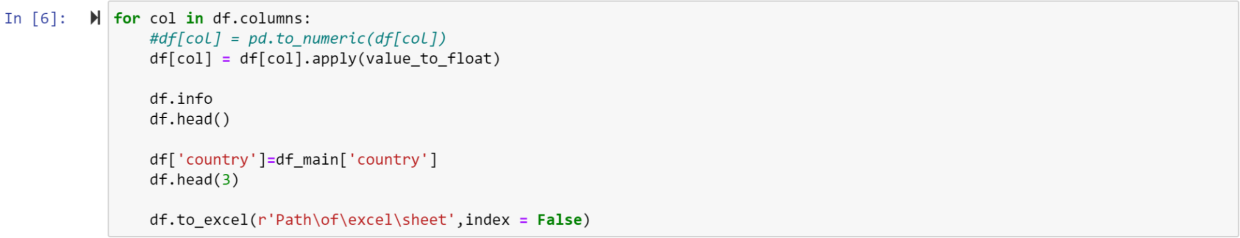
```
In [4]:  df = df_main.drop(['country'], axis = 1)
```

3. **We created a function “value_to_float”, which takes any float or integer from the excel sheet and replaces each character (‘K’, ‘M’, ‘B’) with its corresponding number of zeros.**



```
In [5]:  def value_to_float(x):

             if type(x) == float or type(x) == int:
                 return x

             if 'K' in x:
                 if len(x) > 1:
                     return float(x.replace('K','')) * 1000
                 return 1000.0

             if 'k' in x:
                 if len(x) > 1:
                     return float(x.replace('k','')) * 1000
                 return 1000.0

             if 'M' in x:
                 if len(x) > 1:
                     return float(x.replace('M','')) * 1000000
                 return 1000000.0

             if 'B' in x:
                 return float(x.replace('B','')) * 1000000000
             return 0.0
```

4. **Lastly, we applied the function to each column in the original excel sheet.**

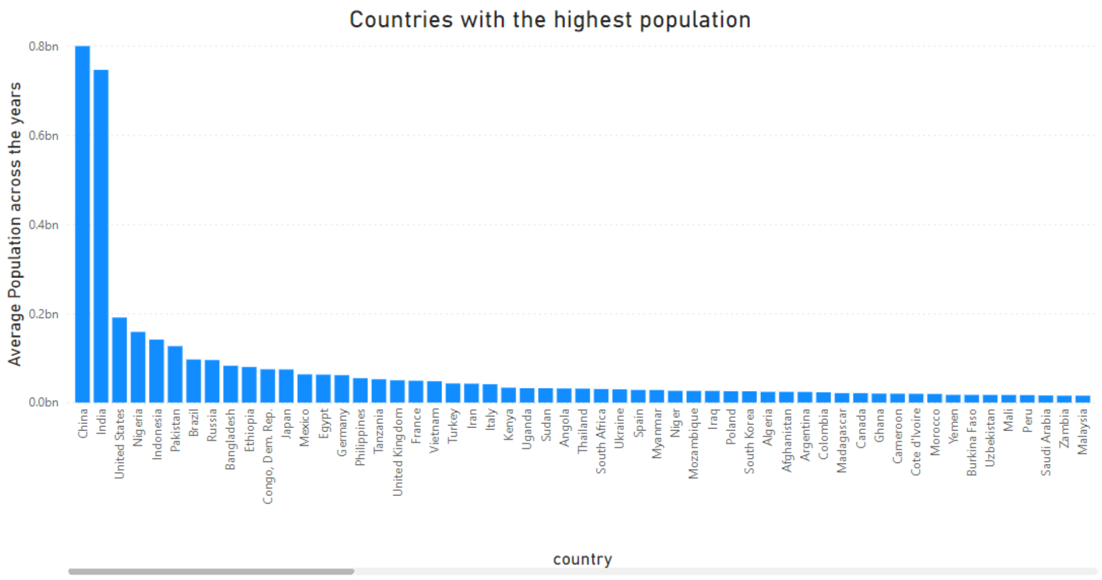


```
In [6]:  for col in df.columns:
             #df[col] = pd.to_numeric(df[col])
             df[col] = df[col].apply(value_to_float)

             df.info
             df.head()

             df['country']=df_main['country']
             df.head(3)

             df.to_excel(r'Path\of\excel\sheet',index = False)
```

5. **After applying these changes to the excel sheets, we had to prepare the data in both Power Bi and Tableau.**
    A. We first needed to unpivot our datasets to be able to turn our years into a column accompanied by their values.
    B. Then we needed to merge every 2 datasets that we were to figure between the relationships so that we could find the intersection between the countries and years since they're not all the same.
    C. For some of the datasets, we needed to group them according to the country to avoid the double counting of countries.

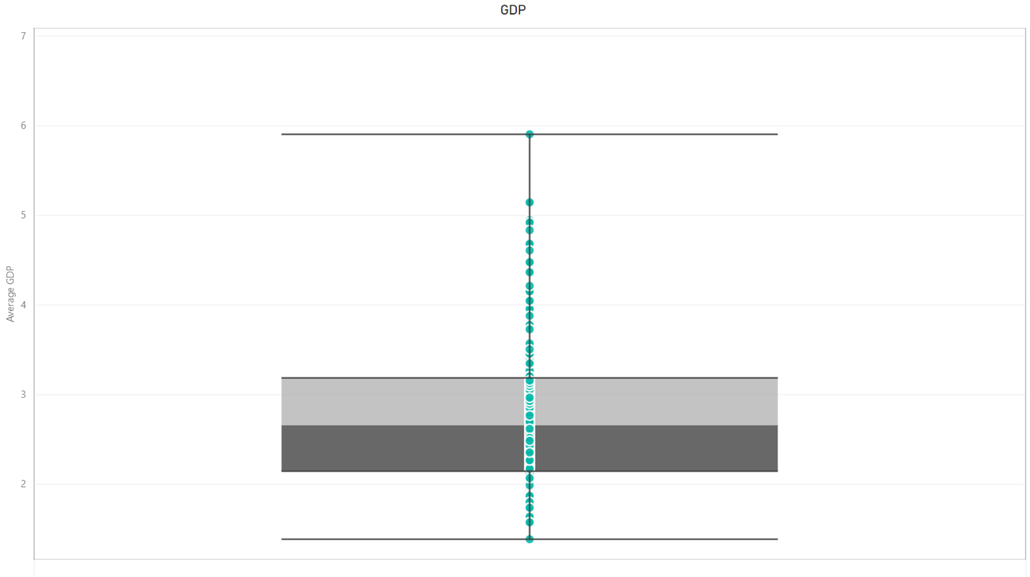
Data Summarization

Population Total



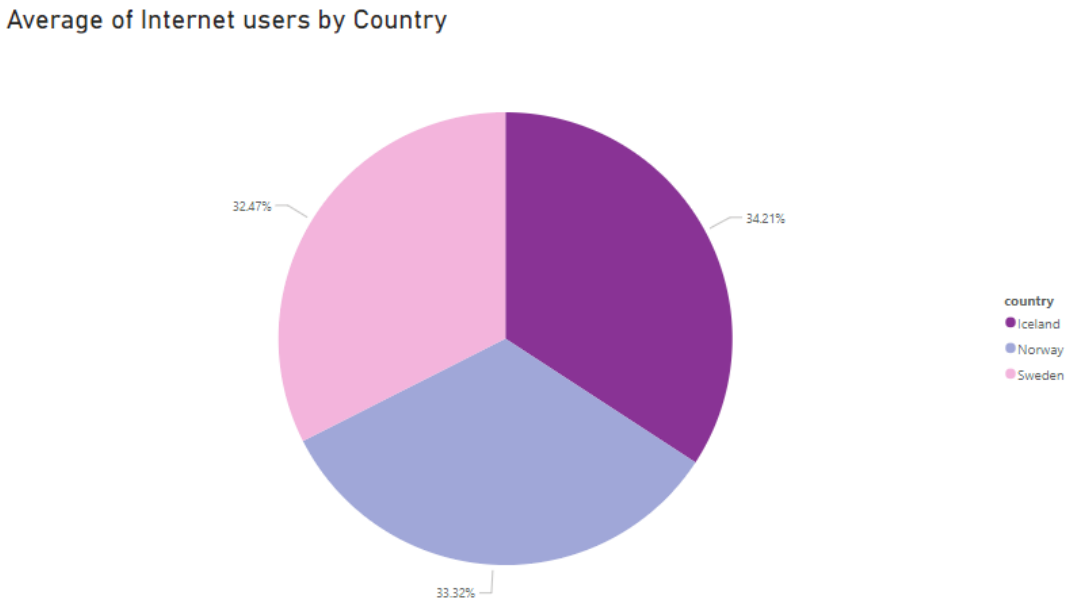
This bar chart is a summary of the population attribute across the years. It is evident that the average population for both China and India is a lot greater than the rest of the countries in the dataset.

Total Annual GDP



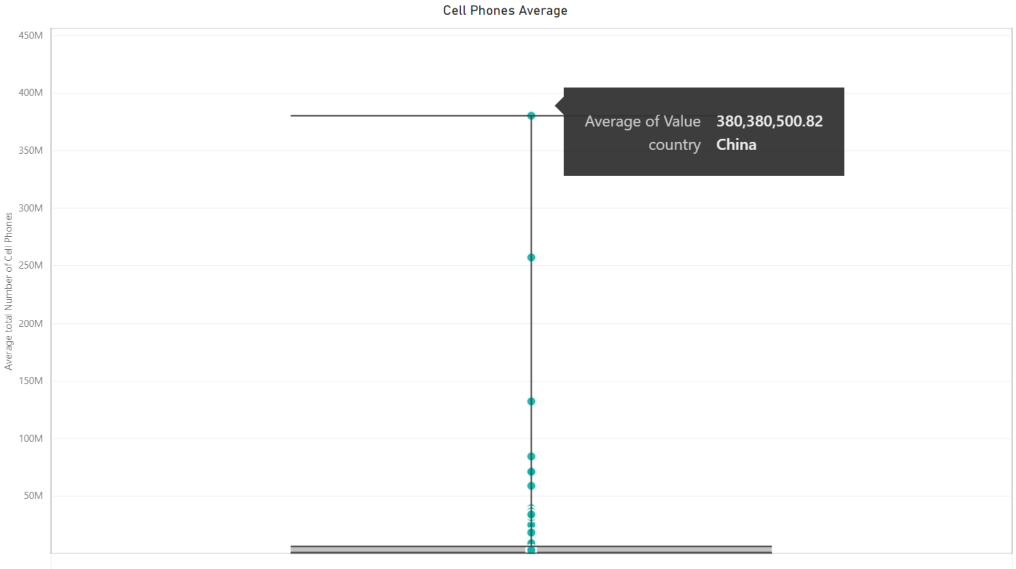
This box plot shows the general distribution of the world's GDP. The way this graph was created is that the average GDP for each country across the years was evaluated, and each dot represents each country's average GDP. The world's average GDP is around 2.79, and the lowest average GDP value is 1.39, while the highest is 5.91.

Internet Users

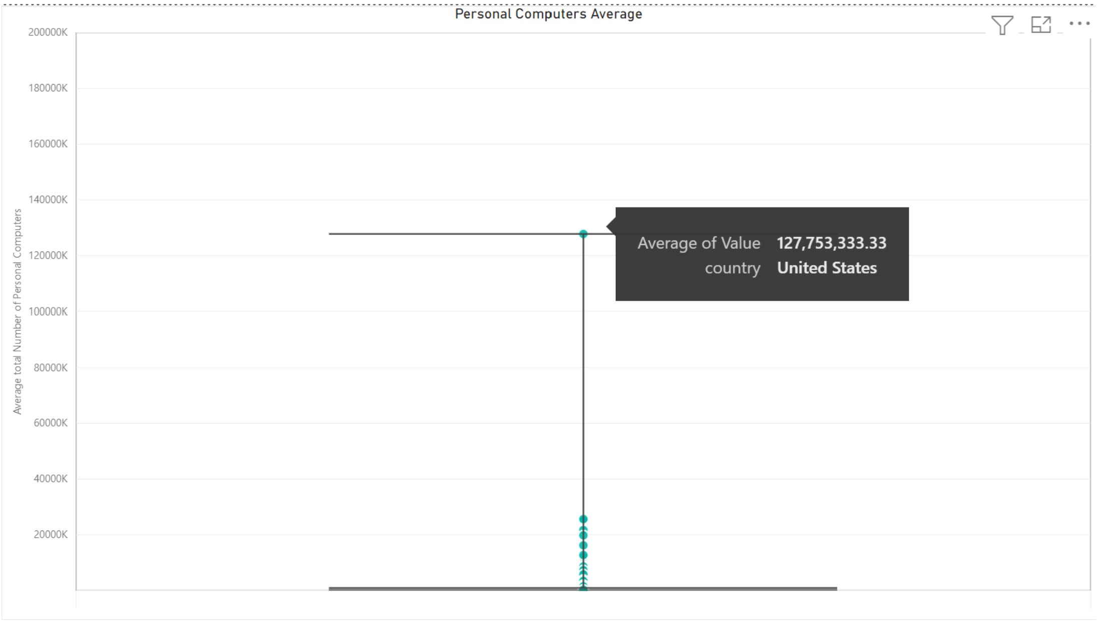


This pie chart shows the top three countries that have the highest average number of internet users. All three countries have almost similar percentages.

Cell Phones Total



Personal Computers Total



The two boxplots above represent the average total number of cell phones and personal computers for all countries across the years. It is clear that China has the highest total average of cell phones, but the United States has the highest total average of personal computers.

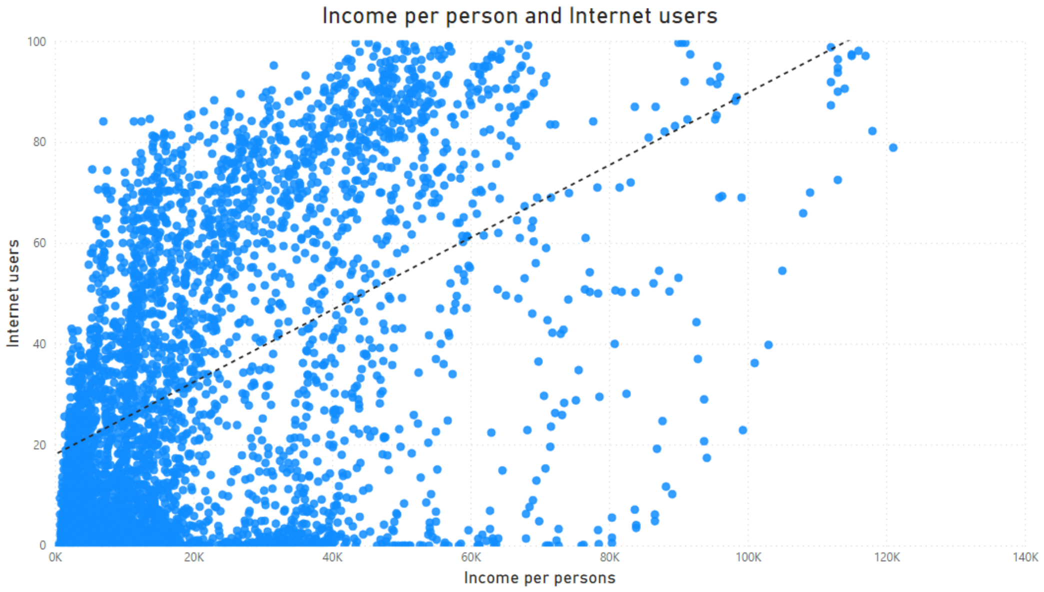
Identifying Relationships Between Attributes

A way by which the relationship between variables can be determined and visualized is through a scatter plot. A line of best fit (dashed line) is a line that minimizes the distance between it and the data points plotted; it is mainly used to express the relationship between variables to show whether there is a positive, negative, or no relationship.

Cell Phones Total with Internet Users

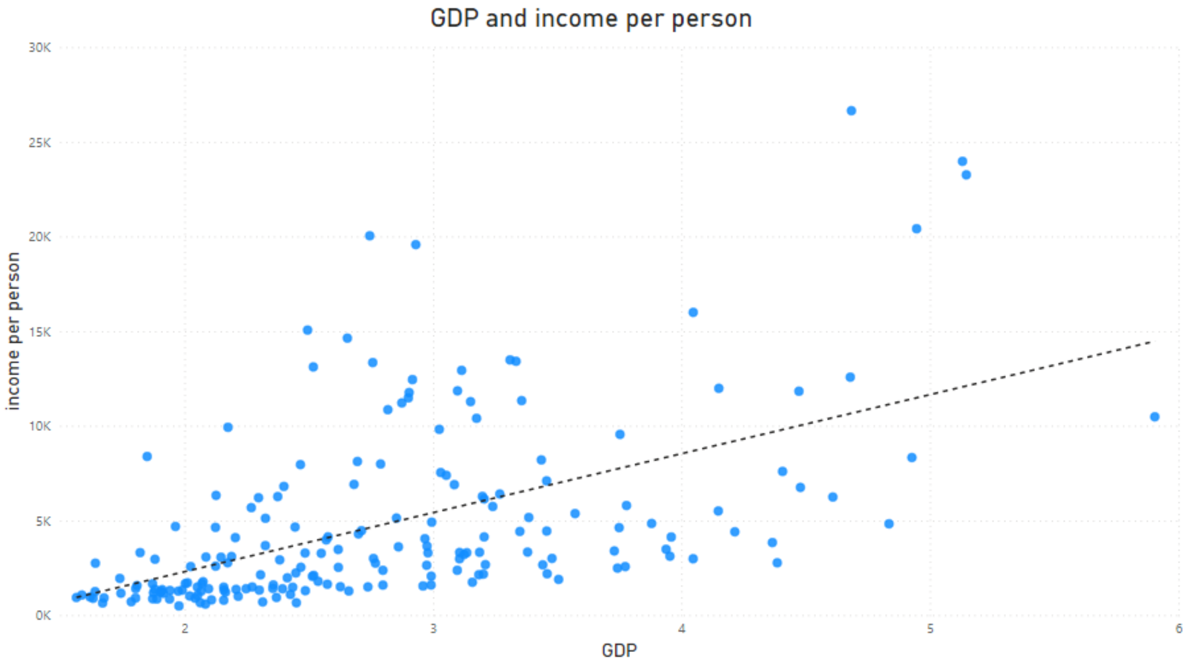
In this plot, there is a positive relationship between the increase in the number of internet users and the number of cell phones for all countries. The more cell phones people have, the more they would need to access the internet; Nowadays, people do not just use cell phones to call each other, but it is a way to be on social media platforms, which requires internet access.

The Income per Person with Internet Users



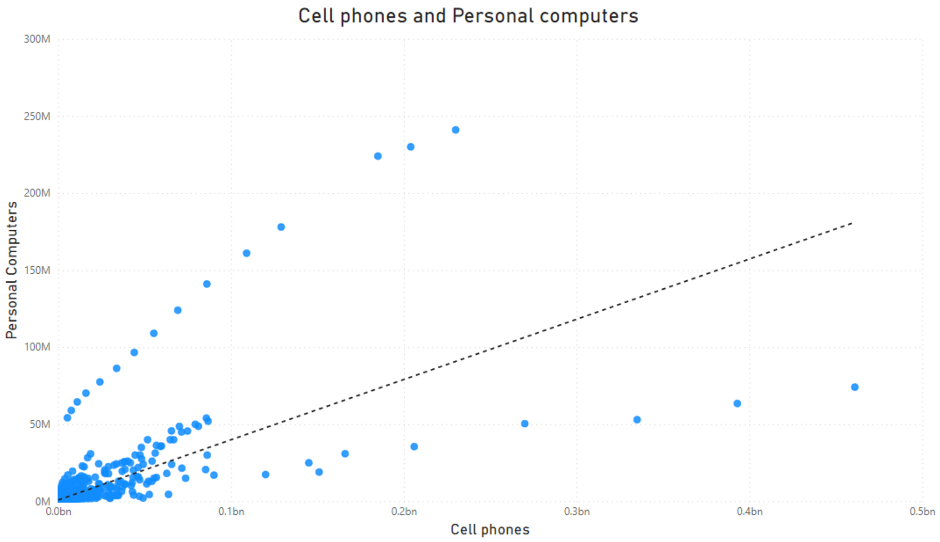
In the above plot, there is a strong positive relationship between the number of internet users and the income per person. Simply, this relationship states that the total number of internet users increases when every individual's income increases and vice versa. Internet services require people to pay in order to gain internet access, so if people individually earn more money, it is more probable that each person will be more willing to pay to use the internet.

Total Annual GDP with Income per Person



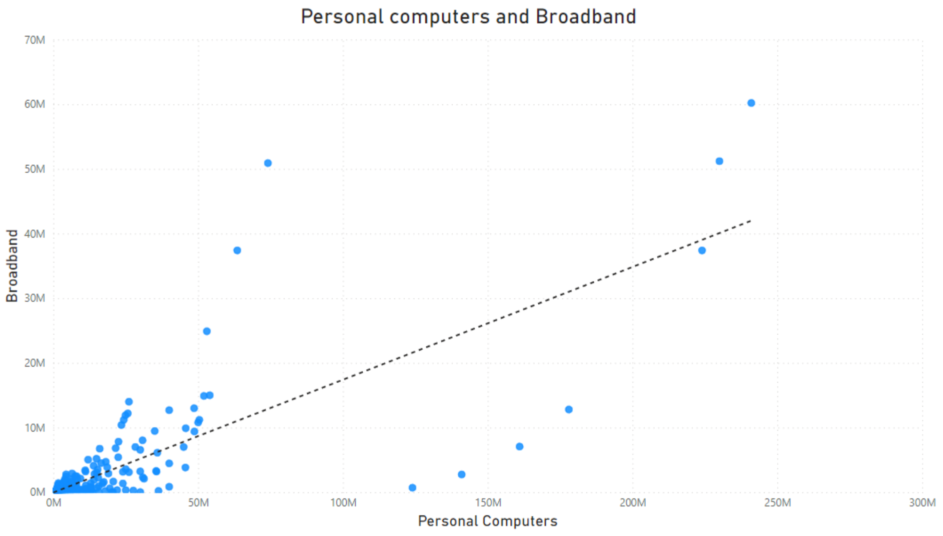
The scatterplot above represents a positive relationship between the total Annual GDP and Income per person.

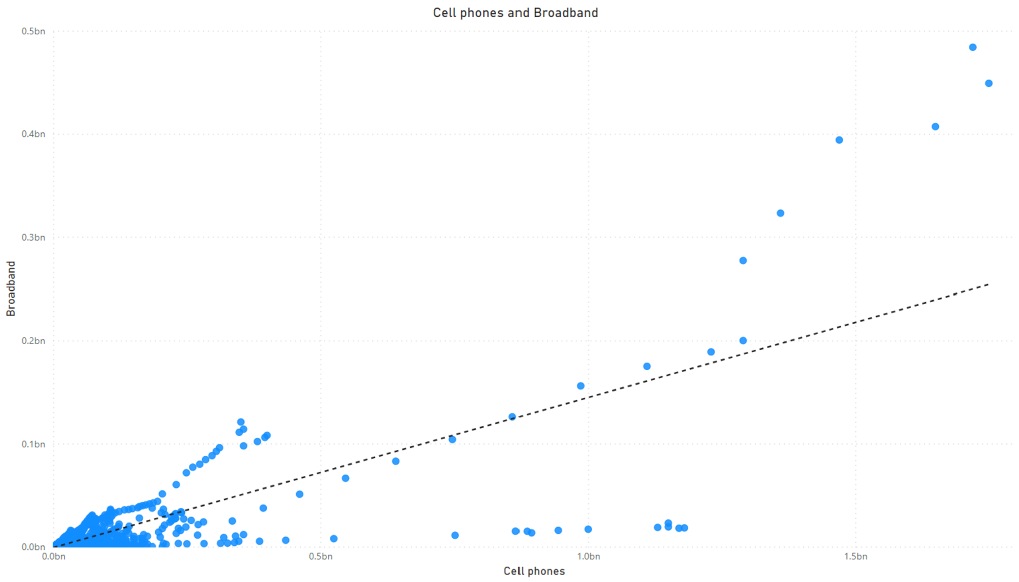
Cell Phones Total with Personal Computers Total



The scatterplot above represents a strong positive relationship between the number of cell phones and the number of computers. In other words, when someone has a cell phone, it is more likely that they will have a personal computer as well.

Personal Computers and Cell Phones Total with Broadband Subscribers





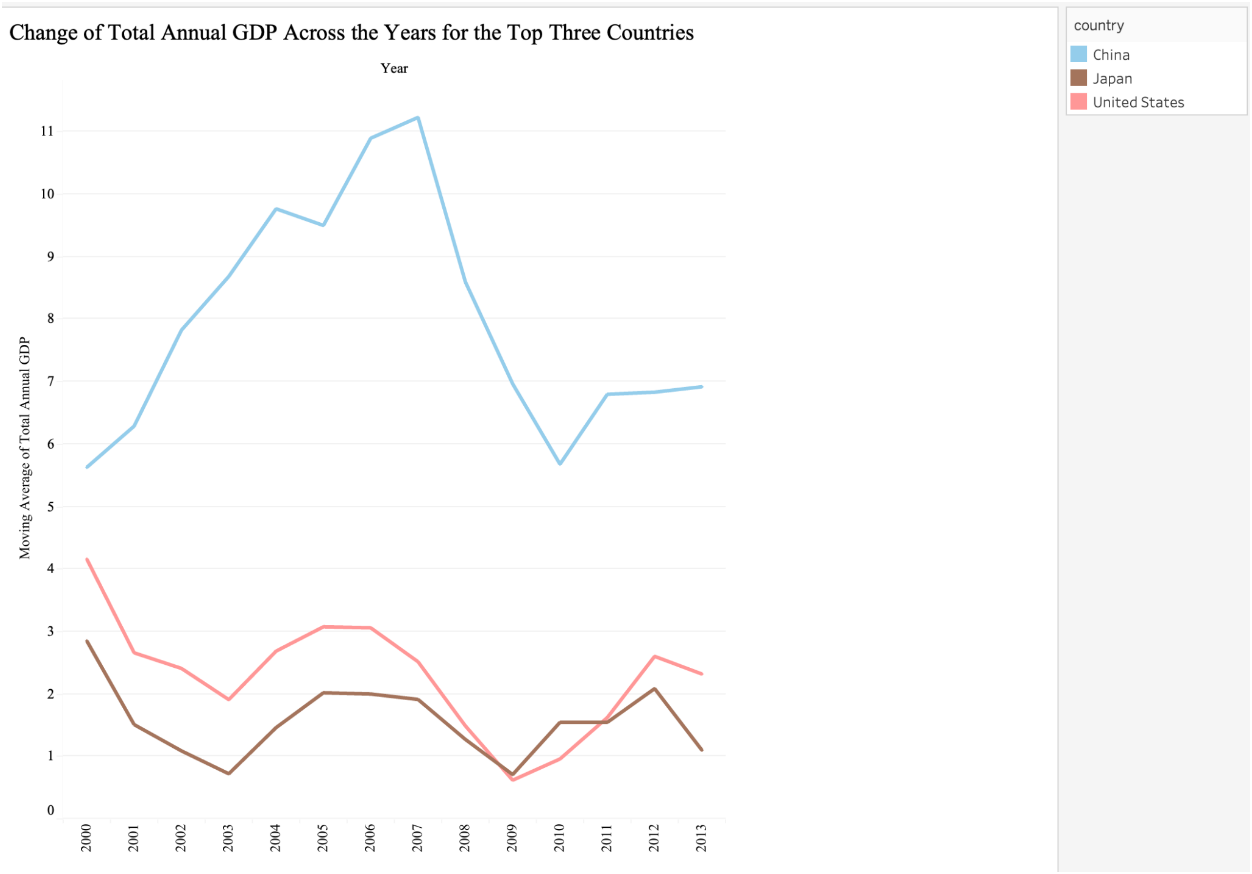
The two plots above show that there is a positive relationship between the number of broadband subscribers and both the number of cell phone users and the number of computer users. This indicates that when more people have cell phones and personal computers, they are more likely to want to subscribe to a high-speed internet service.

Monitoring Change of Variables with Time

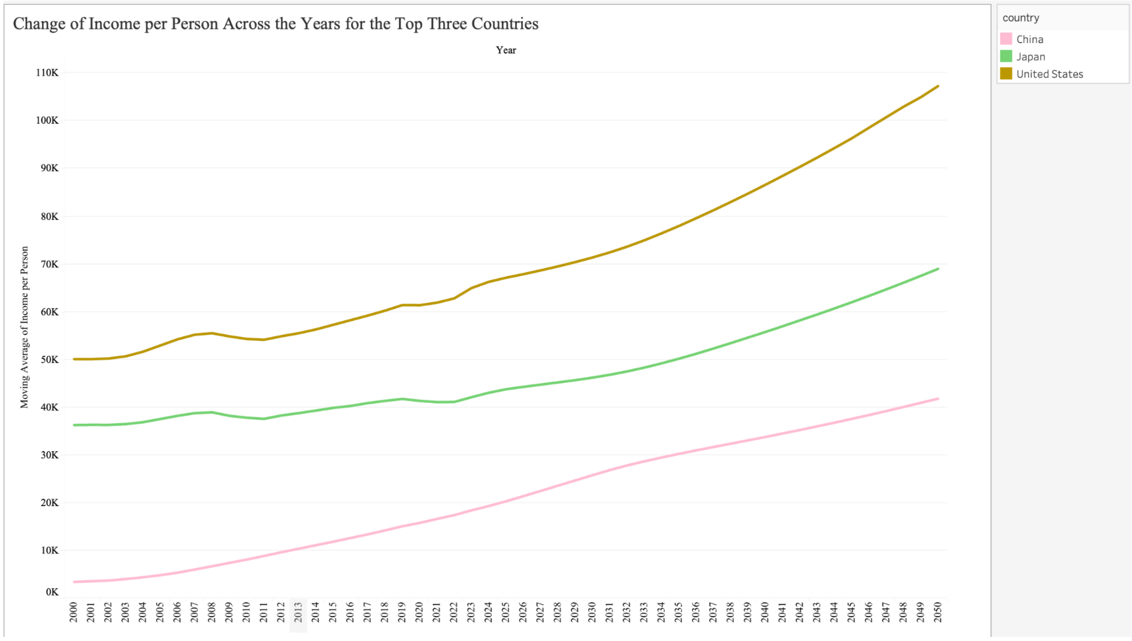
Since our main purpose was to find insights about countries that share common economic states, GDP is the most important variable of interest. Generally, GDP is an indicator of how well a country is doing economically; an increasing GDP signifies an increase in development, goods, and services. In our case, we want to focus on the top three countries that have the highest moving average for GDP across the years because we assume that technological advancements should have reached them earlier and faster than other undeveloped countries.

A moving average is a method by which the average changes in a data variable over time, and it is beneficial whenever we want to smooth out small fluctuations and only focus on the more extreme ones. In order to show the main trends in the data, we applied the moving average for some variables of interest.

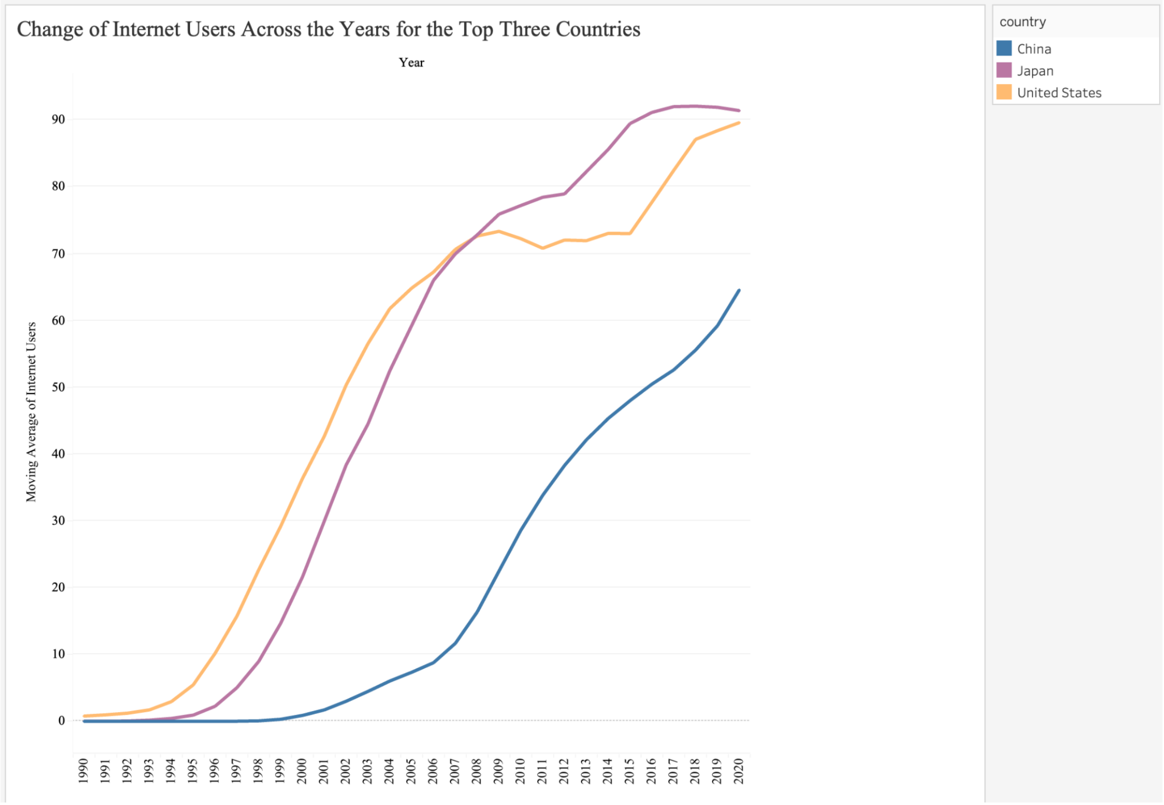
After eliminating the moving averages for all the countries and keeping the three countries that remained consistently higher than the others throughout the years, we found that they are China, the United States, and Japan.



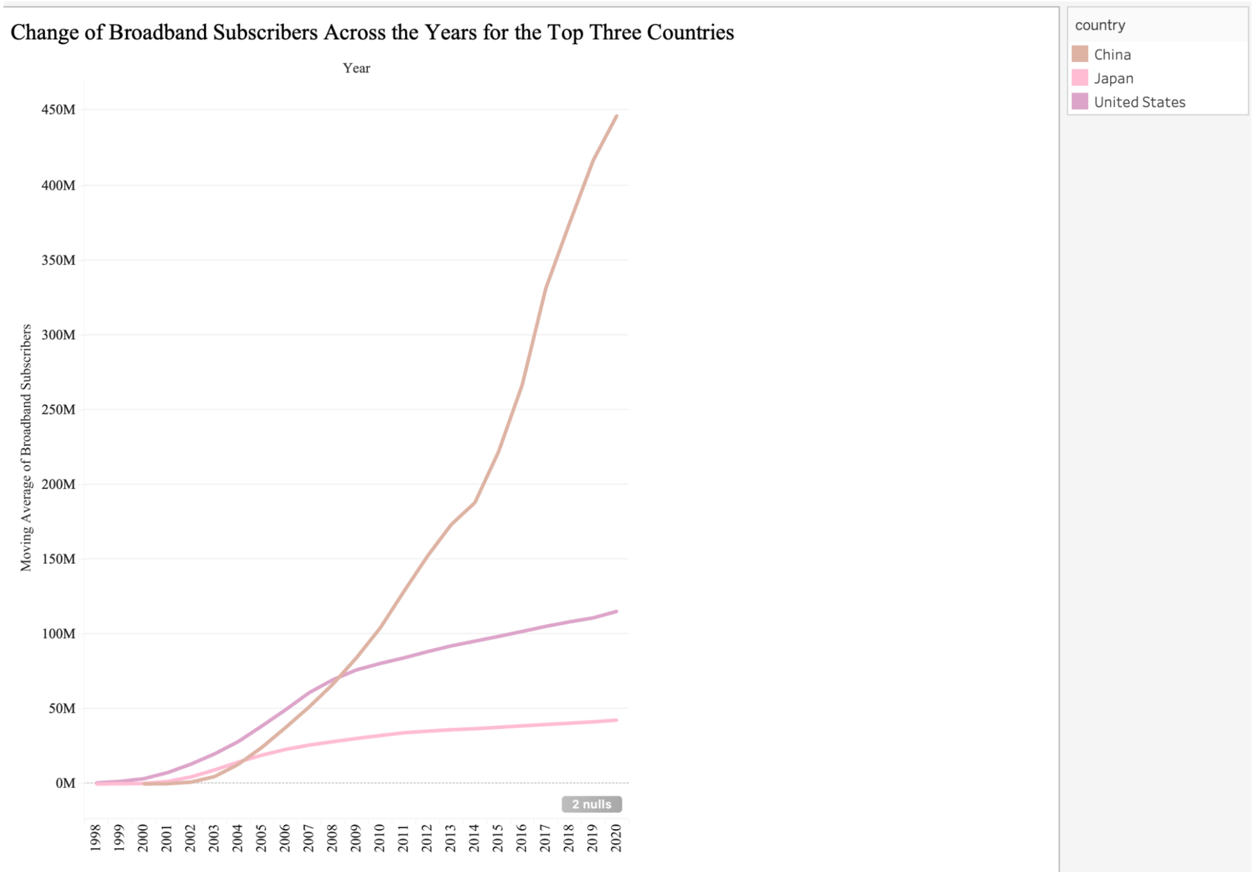
The graph above captures the moving averages of the GDP of the top three countries. From 2000 to 2003, the Total Annual GDP of the United States and Japan was decreasing while China's GDP was strongly increasing. Around 2009, the moving averages for all three countries decreased. Throughout the years, Japan's GDP was the lowest except from 2009 to 2011, when it was higher than that of the United States.



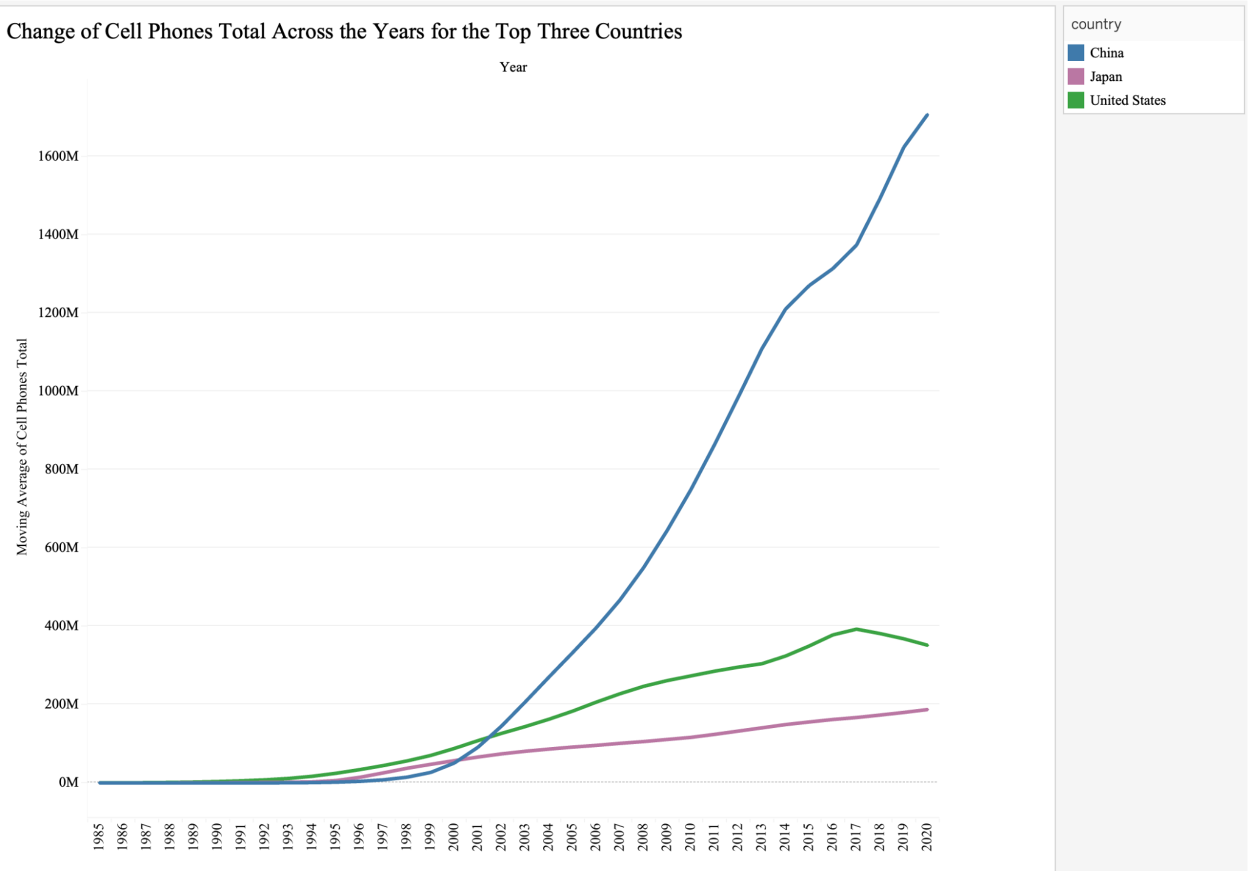
Although China's Annual GDP was consistently high, its moving average for income per person was the lowest when compared to Japan and the United States.



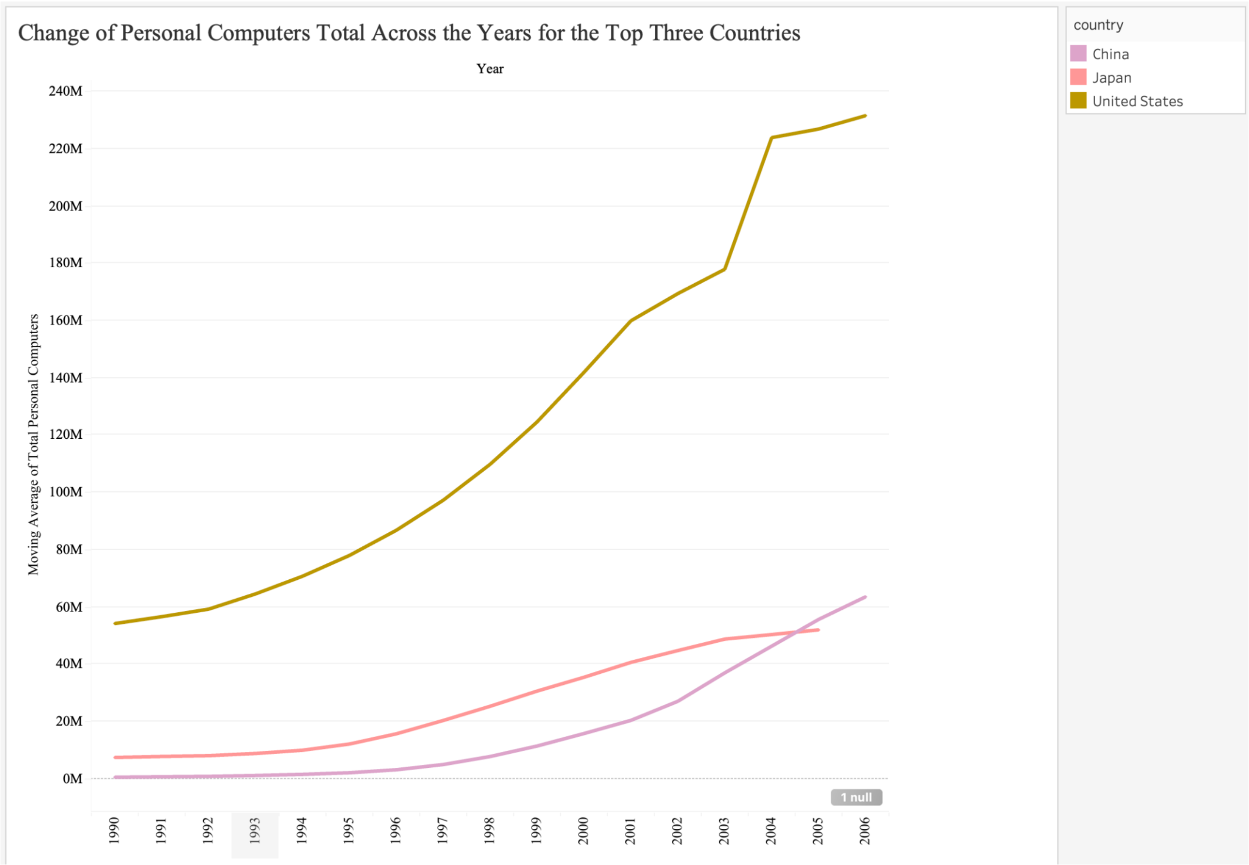
In this line graph, China had the lowest moving average for the number of internet users over the years, while the United States had, for the most part, the highest moving average. Around 2008, the United States and Japan had about the same moving averages, but after this year, the number of internet users in Japan was higher than that in the United States.



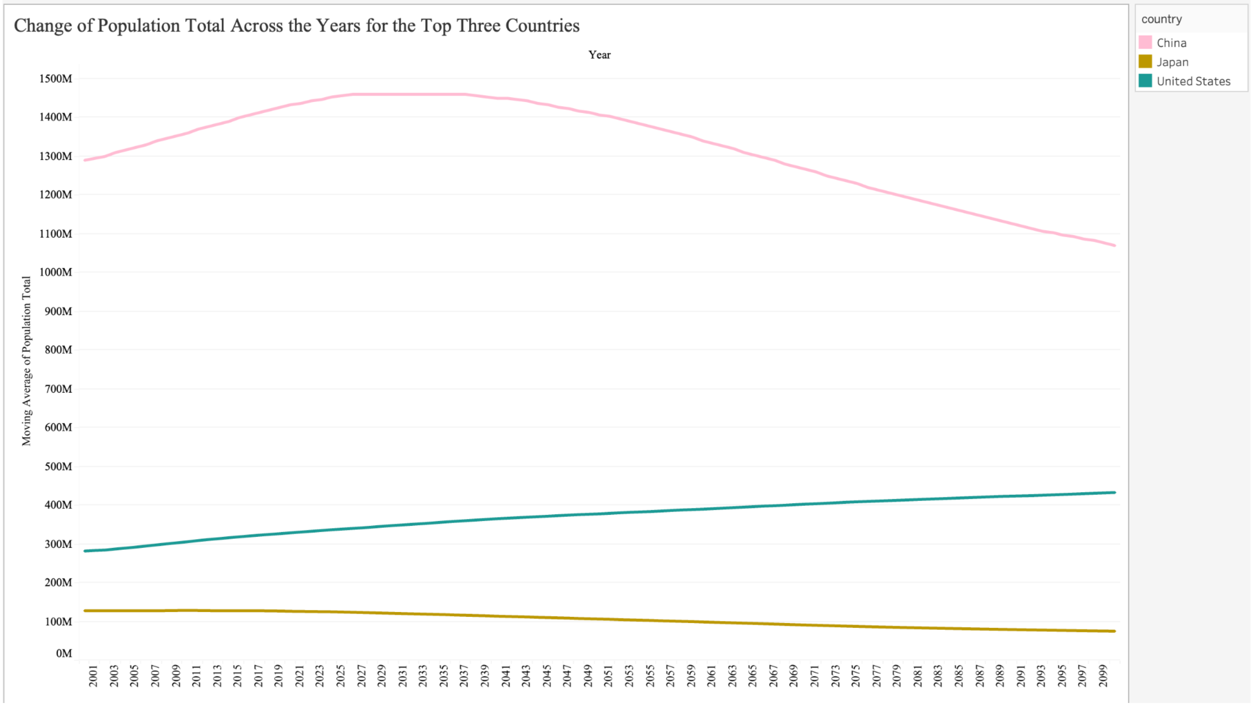
Out of the top three countries, the moving average of the number of broadband subscribers in China had the highest rate of increase when compared to Japan and the United States. However, this increase only started around 2003. Between 2000 and 2008, China's moving average was lower than that of the United States. Also, starting from 1998, it seemed like the United States was going to have an exponential increase, but after 2008, the rate of increase in the number of broadband subscribers has been affected.



This graph represents the moving average of the total number of cell phones across the years for China, the United States, and Japan. In the beginning, all three countries had around the same total number of cell phones, but after 2000, China had a strong exponential increase. The United States has a higher moving average than that of Japan, but still, both countries are relatively similar when compared to China.



This graph represents the moving average for the number of personal computers for the three top countries. The United States had the highest number of personal computers from 1990 to 2006, while Japan and China had almost similar moving averages. Up till 2004, the number of personal computers in Japan was greater than that in China. From the graph, it can be seen that after 2005 China's moving average is more likely to continue increasing at a higher rate than Japan's moving average.



The population average in China is a lot higher than in Japan and the United States. Currently, China's population is about a billion and a half, but from the graph, it is predicted that there will be a decrease in the population, reaching one billion in 2099. The population of the United States and Japan remained relatively the same across the years, and it is expected that the population of the United States will slightly increase while Japan's population will slightly decrease.

Conclusion

The relationships between the attributes and how they relate to the moving averages for the top three countries: China, the United States, and Japan. Furthermore, the previous graphs that were plotted to identify specific relationships were confirmed by all of the graphs that were used to monitor changes in these attributes over time. Nonetheless, we had an exception, which was China; the relationship between income per person and GDP is a positive relationship. However, China's moving average was less than both the United States and Japan, unlike its GDP, which is the highest out of all of them.

It is worth mentioning a specific drawback in the dataset, which is that we do not know whether the outliers were excluded while the attribute "income per person" was being created. However, throughout our data analysis, we assumed that the outliers were removed.