Exploratory Data Analysis

[Group6]
Carolyn Haythorn
YeEun Jeon
Haneul Kim
Satish Sneha

The big topic of our project is mitigating disruptions in the global supply chain caused by the continued COVID-19 pandemic, and even if another pandemic situation like COVID happens again in the future, I thought that 'vaccination' can be one good solution so that people can recover from pandemic situation faster and do the global trade like before the pandemic situation. So I wanted to see if the vaccination was actually related to global trade. I compared the number of vaccinations in Korea with trade dataset of Korea.

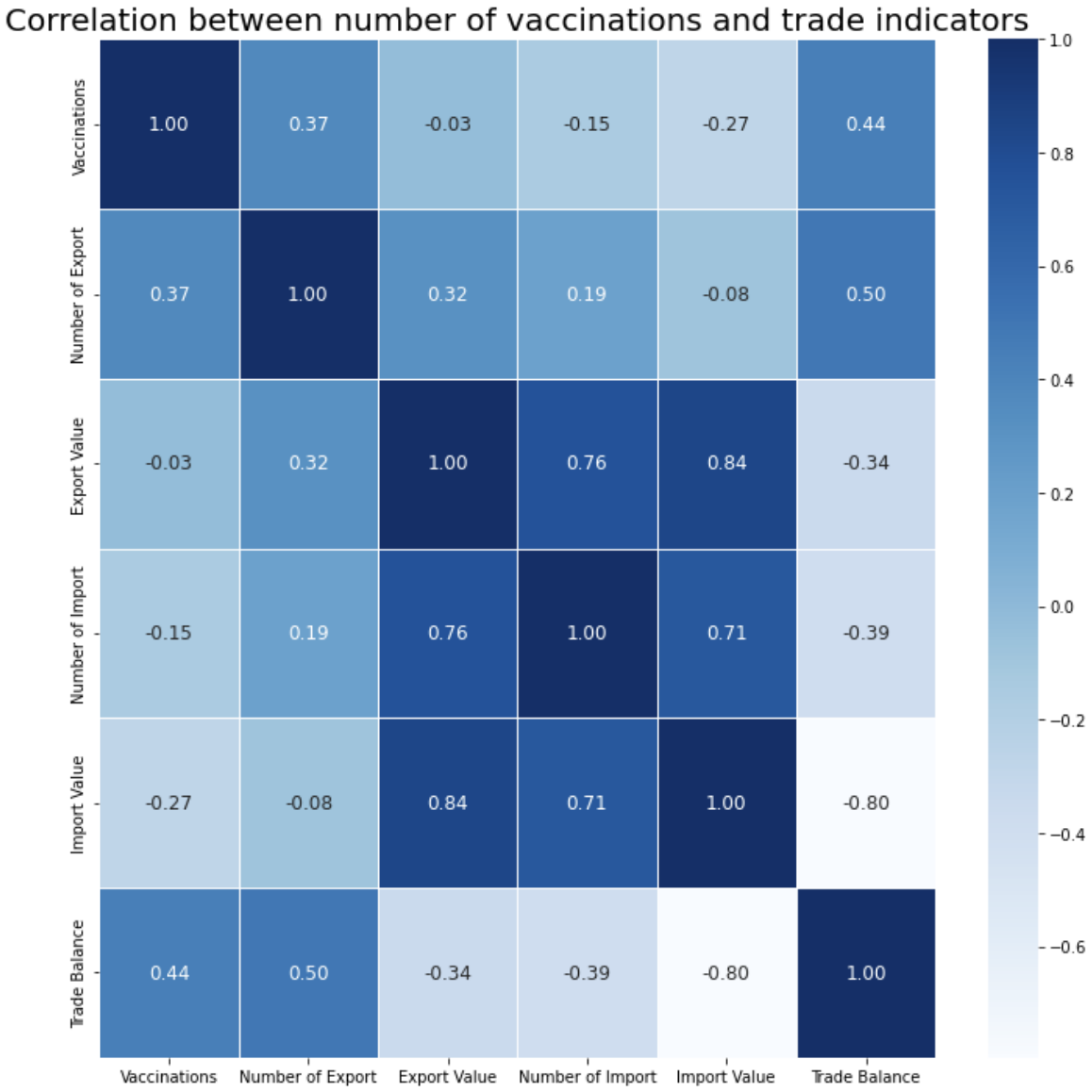
[Used datasets]

1) Vaccination datasets:
   https://www.kaggle.com/datasets/rsrishav/covid-vaccination-dataset
   Used the data of daily vaccinations in Korea
2) Trade datasets(published by the Korean Customs Service):
   https://unipass.customs.go.kr/ets/index_eng.do
   Used import/export data of Korea(Import value, number of import, export value, number of export, trade balance) by month

First, I merged two datasets by month from December 2020 to February 2023.

	Year	Month	Vaccinations	period	Number of Export	Export Value	Number of Import	Import Value	Trade Balance
0	2020	12	3.507162e+07	Dec-20	1275069.0	51332449.0	3274735.0	44638246.0	6694203.0
1	2021	1	3.699647e+08	Jan-21	1063782.0	48006974.0	3216535.0	44456822.0	3550152.0
2	2021	2	6.693220e+08	Feb-21	991831.0	44706907.0	2779591.0	42404636.0	2302271.0
3	2021	3	1.367415e+09	Mar-21	1439462.0	53690914.0	3298165.0	49742555.0	3948359.0
4	2021	4	2.168517e+09	Apr-21	1184249.0	51225991.0	3346730.0	50891152.0	334840.0
5	2021	5	3.218742e+09	May-21	1143835.0	50725161.0	3245257.0	47910522.0	2814640.0
6	2021	6	4.636108e+09	Jun-21	1513595.0	54778905.0	3101662.0	50429027.0	4349878.0
7	2021	7	4.351525e+09	Jul-21	1182251.0	55461518.0	3225209.0	53676256.0	1785262.0
8	2021	8	4.848919e+09	Aug-21	1122015.0	53165104.0	3180345.0	51581093.0	1584011.0
9	2021	9	3.914293e+09	Sep-21	1502399.0	55913862.0	3199940.0	51636341.0	4277521.0

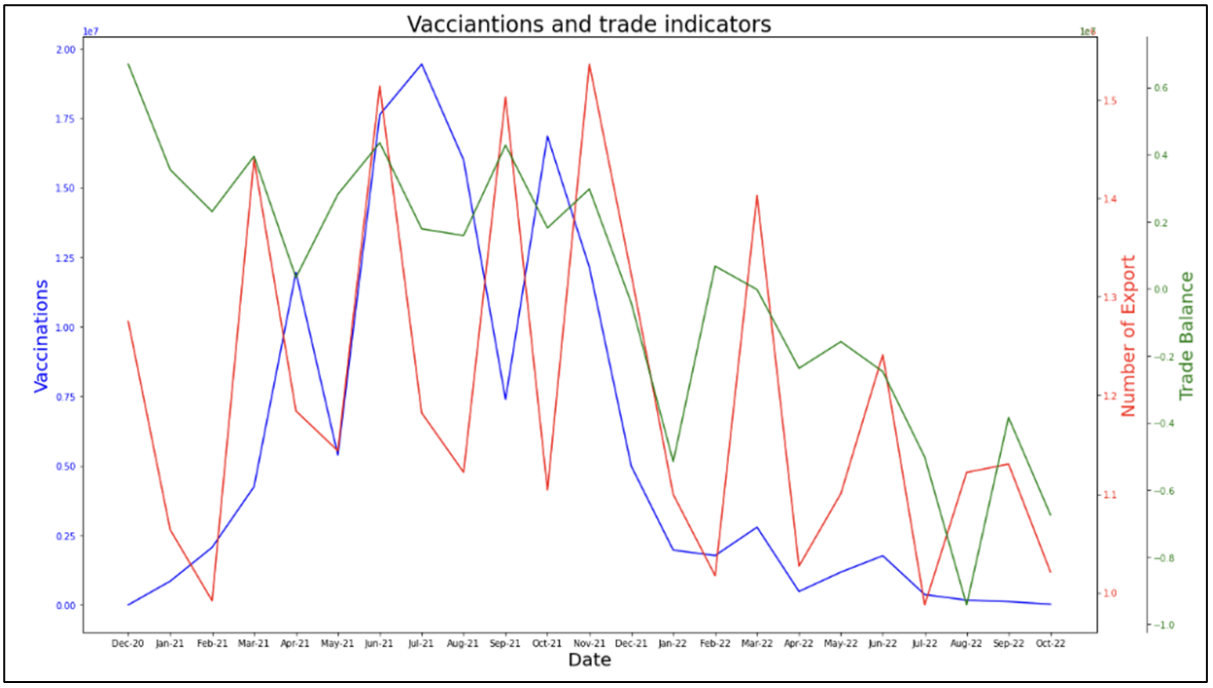
Then calculated the correlation between number of vaccinations and trade indicators(Import value, number of import, export value, number of export, trade balance). I showed the results by heatmap.



According to the scale of Pearson's correlation coefficient, I could find out that number of export has little correlation(0.37) and trade balance has moderate correlation(0.44) with number of vaccinations.

Scale of correlation coefficient	Value
$0 < r \leq 0.19$	Very Low Correlation
$0.2 \leq r \leq 0.39$	Low Correlation
$0.4 \leq r \leq 0.59$	Moderate Correlation
$0.6 \leq r \leq 0.79$	High Correlation
$0.8 \leq r \leq 1.0$	Very High Correlation

To see the specific relation between number of export and trade balance with number of vaccinations, I plotted the graph.



By plotting the line graph, I could see that the tendency of higher number of vaccinations lead to more number of exports in 2021. In 2022, as the number of vaccinations decreased, number of exports and trade balance and number of exports also decreased.
So, the results indicate that number of vaccinations are related to exports.

There was also paper saying like this: However, richer nations focused on being the first to develop and roll out COVID-19 vaccines to their own populations, rather than focus on what was best for all of humanity. The emergence of the VOC Omicron variant and its rapid spread reflects the legacy of wealthy nations' failure to equitably distribute COVID-19 vaccines globally. (https://www.ijidonline.com/action/showPdf?pii=S1201-9712%2821%2900888-2)

So I wanted to compare vaccine penetration between the top five countries' GDP(Gross domestic product) and the bottom five.

- GDP top ranking: USA, China, Japan, Germany, India
- GDP bottom ranking: Estonia, Nepal, Paraguay, Bahrain, Bolivia

I refers to the GDP ranking on Google, and it only ranks 100th. Meaning that the five countries in the lower ranking are 95th to 100th.

Through the process of cleaning the original data, the data was stored in a Kaggle. I used 'COVID Vaccination in World', and COVID Vaccination Dataset which gets updated daily. It is collected from OWID (Our World in Data) GitHub repository, which is updated on a daily basis. This data contains the records of vaccination received by people from all the countries. This dataset follows:

- location: name of the country
- iso_code: ISO 3166-1 alpha-3-three-letter-country-codes
- date: date of observation

- total_vaccinations: total number of doses administered. This is counted as a single dose.
- total_vaccinations_per_hundred: total_vaccinations per 100 people in the total population of the country.
- daily_vaccinations_raw: daily change in the total number of doses administered.
- daily_vaccinations: new doses administered per day.
- daily_vaccinations_per_million: daily_vaccinations per 1,000,000 people in the total population.
- people_vaccinated: total number of people who received at least one vaccine dose.
- people_vaccinated_per_hundred: people_vaccinated per 100 people in the total population of the country
- people_fully_vaccinated: total number of people who received all doses prescribed by the vaccination protocol
- people_fully_vaccinated_per_hundred: people_fully_vaccinated per 100 people in the total population of the country.

With this Vaccination datasets (https://www.kaggle.com/datasets/rsrishav/covid-vaccination-dataset), I used [iso_code],[date], [total_vaccinations], [total_vaccinations_per_hundred], [people_fully_vaccinated], [people_fully_vaccinated_per_hundred] columns.

```
from autoviz.AutoViz_Class import AutoViz_Class

# AutoViz 인스턴스 생성
AV = AutoViz_Class()
```

```
Imported v0.1.58. After importing, execute '%matplotlib inline' to display charts in Jupyter.
    AV = AutoViz_Class()
    dfte = AV.AutoViz(filename, sep=',', depVar='', dfte=None, header=0, verbose=1, lowess=False,
            chart_format='svg',max_rows_analyzed=150000,max_cols_analyzed=30, save_plot_dir=None)
Update: verbose=0 displays charts in your local Jupyter notebook.
        verbose=1 additionally provides EDA data cleaning suggestions. It also displays charts.
        verbose=2 does not display charts but saves them in AutoViz_Plots folder in local machine.
        chart_format='bokeh' displays charts in your local Jupyter notebook.
        chart_format='server' displays charts in your browser: one tab for each chart type
        chart_format='html' silently saves interactive HTML files in your local machine
```

First, I wanted to used ‘AutoViz’ for data visualization.

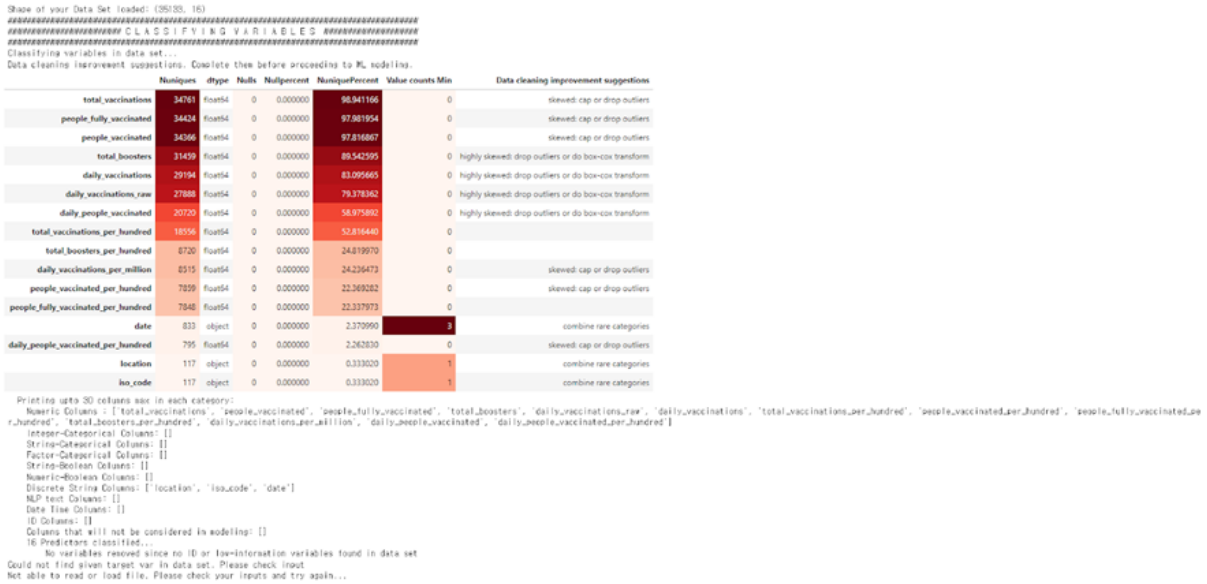
```
    Since nrows is smaller than dataset, loading random sample of 150000 rows into pandas...
Shape of your Data Set loaded: (150000, 16)
############################################################################################
##################### C L A S S I F Y I N G  V A R I A B L E S ######################
############################################################################################
Classifying variables in data set...
Data cleaning improvement suggestions. Complete them before proceeding to ML modeling.
```

	Nuniques	dtype	Nulls	Nullpercent	NuniquePercent	Value counts Min	Data cleaning improvement suggestions
daily_vaccinations	6466	float64	77	0.770000	64.660000	0	fill missing, highly skewed: drop outliers or do box-cox transform
daily_people_vaccinated	5144	float64	56	0.560000	51.440000	0	fill missing, highly skewed: drop outliers or do box-cox transform
total_vaccinations	4445	float64	5542	55.420000	44.450000	0	fill missing, highly skewed: drop outliers or do box-cox transform
people_vaccinated	4246	float64	5729	57.290000	42.460000	0	fill missing, highly skewed: drop outliers or do box-cox transform
daily_vaccinations_per_million	4127	float64	77	0.770000	41.270000	0	fill missing, skewed: cap or drop outliers
people_fully_vaccinated	4111	float64	5869	58.690000	41.110000	0	fill missing, highly skewed: drop outliers or do box-cox transform
total_vaccinations_per_hundred	3953	float64	5542	55.420000	39.530000	0	fill missing
daily_vaccinations_raw	3422	float64	6340	63.400000	34.220000	0	fill missing, highly skewed: drop outliers or do box-cox transform
people_vaccinated_per_hundred	3135	float64	5729	57.290000	31.350000	0	fill missing
people_fully_vaccinated_per_hundred	3024	float64	5869	58.690000	30.240000	0	fill missing
total_boosters	2476	float64	7435	74.350000	24.760000	0	fill missing, highly skewed: drop outliers or do box-cox transform
total_boosters_per_hundred	1888	float64	7435	74.350000	18.880000	0	fill missing
date	843	object	0	0.000000	8.430000	1	combine rare categories
daily_people_vaccinated_per_hundred	679	float64	56	0.560000	6.790000	0	fill missing, highly skewed: drop outliers or do box-cox transform
location	235	object	0	0.000000	2.350000	4	combine rare categories
iso_code	235	object	0	0.000000	2.350000	4	combine rare categories

```
    Printing upto 30 columns max in each category:
    Numeric Columns : ['total_vaccinations', 'people_vaccinated', 'people_fully_vaccinated', 'total_boosters', 'daily_vaccinations_raw', 'daily_vaccinations', 'total_vaccinations_per_hundred', 'people_vaccinated_per_hundred', 'people_fully_vaccinated_per_hundred', 'total_boosters_per_hundred', 'daily_vaccinations_per_million', 'daily_people_vaccinated', 'daily_people_vaccinated_per_hundred']
    Integer-Categorical Columns: []
    String-Categorical Columns: []
    Factor-Categorical Columns: []
    String-Boolean Columns: []
    Numeric-Boolean Columns: []
    Discrete String Columns: ['location', 'iso_code', 'date']
    NLP text Columns: []
    Date Time Columns: []
    ID Columns: []
    Columns that will not be considered in modeling: []
    16 Predictors classified...
        No variables removed since no ID or low-information variables found in data set
Since Number of Rows in data 150000 exceeds maximum, randomly sampling 150000 rows for EDA...
Could not find given target var in data set. Please check input
Not able to read or load file. Please check your inputs and try again...
```

As you can see, there was too much missing data. So that the model was not able to read/load this data set.

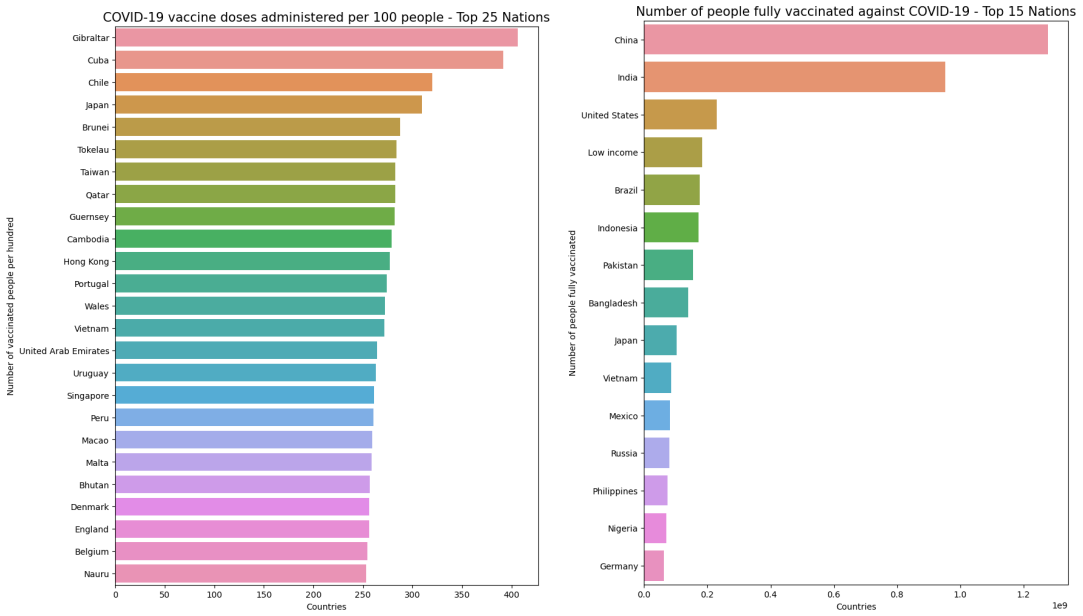
Therefore, I used ‘dropna()’ for data processing of null values. However I would like to see the first line ‘ Shape of your Data Set loaded’, before dealing the null value (150000, 16) to (35144, 16). It can be seen that the amount of data has been reduced due to too much missing data. For this reason, I decided to use the original data to combine the data.



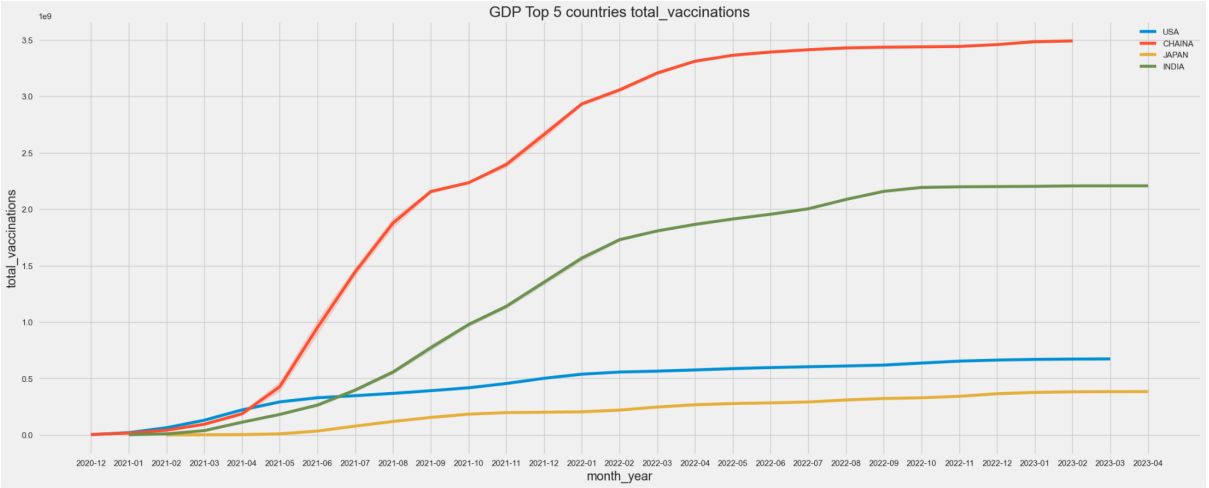
	location	iso_code	date
1585	Africa	OWID_AFR	2023-04-04
2015	Albania	ALB	2022-03-15
6676	Argentina	ARG	2023-04-09
8961	Asia	OWID_ASI	2023-04-04
9378	Australia	AUS	2022-04-07
...	...	...	...
157183	Uruguay	URY	2023-04-05
159620	Vietnam	VNM	2022-05-18
160724	Wales	OWID_WLS	2023-03-29
162250	World	OWID_WRL	2023-04-04
164246	Zimbabwe	ZWE	2022-10-03

117 rows × 16 columns

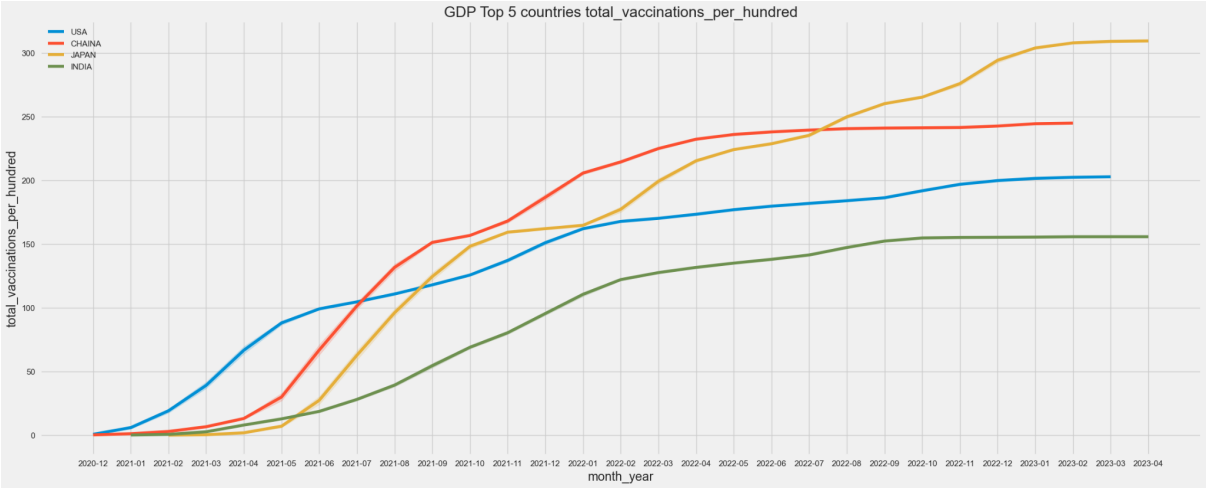
This data set is a data set that accumulates day by day, so I left the last day of month for comparison only on the last day of each month. Unfortunately, since each country has different last date of the month, i decided to use the original data.



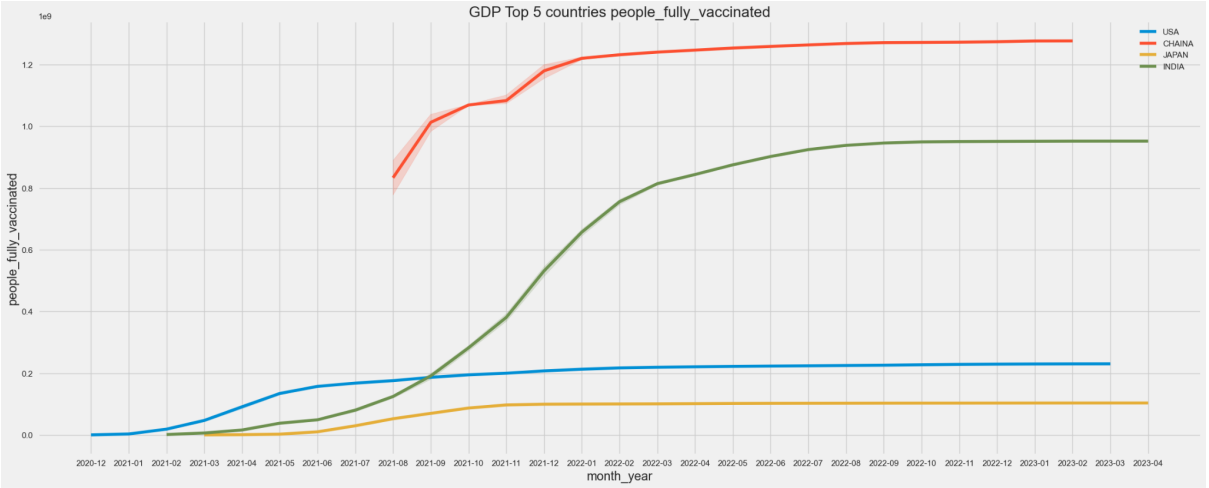
Using the matplotlib, I found that COVID-19 vaccines administered per 100 people To 25 countries and Number of people fully vaccinated against COVID-19 top nations.



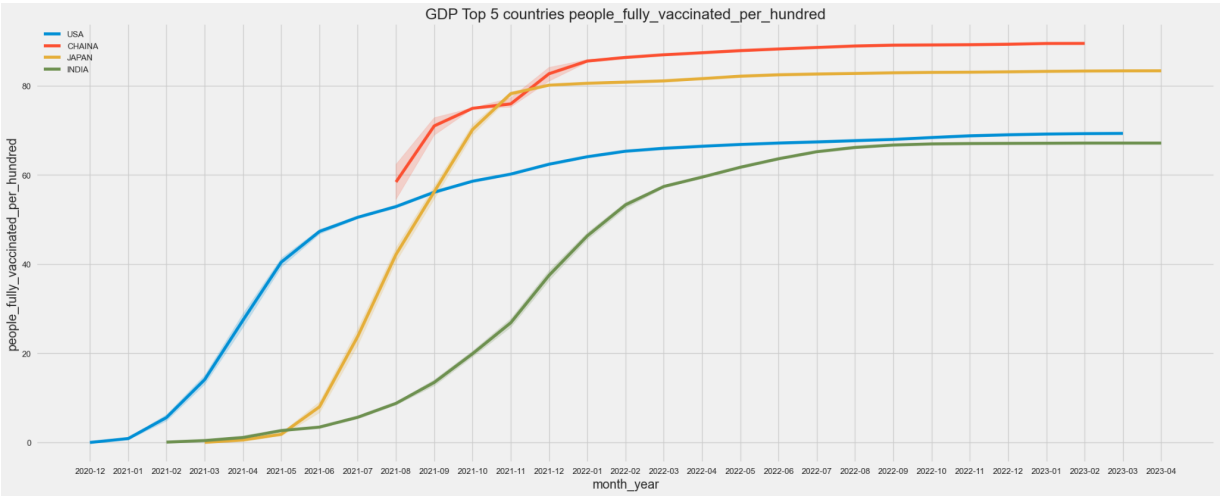
This is a GDP top ranking of total_vaccinations. USA, China, Japan, Germany, India were the top 5 countries of GPD.
China is the highest, and India is seen following in the second. This is graph showing only the number of people because the data of the number if vaccinated people and the number of people are not combined. The result was likely to be different if checked by the radio of the of the number of people.



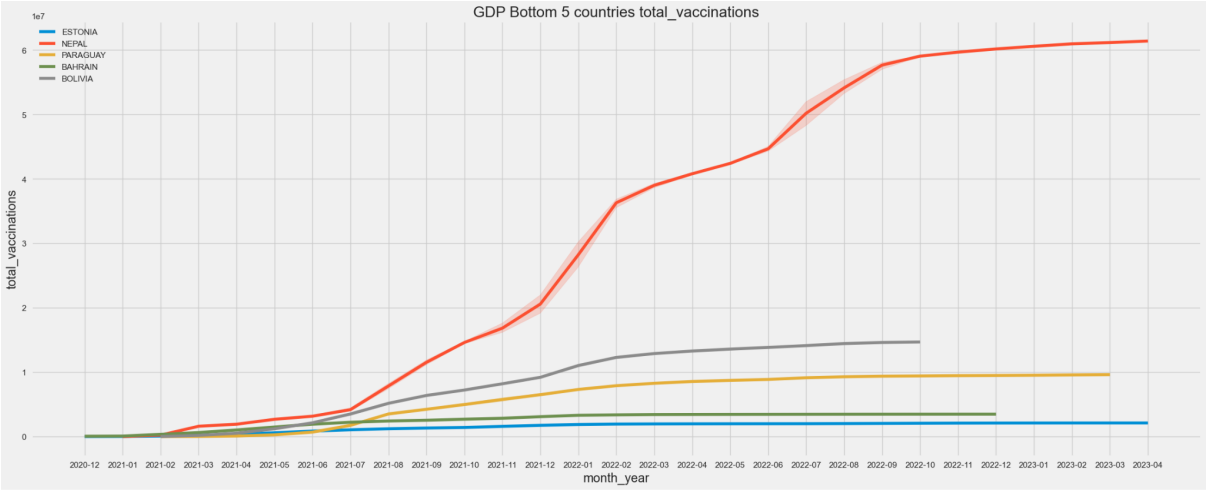
Earlier, the graph that did not consider total_vaccinations was followed by Chain, India, Japan, America, while the graph that considered vaccinations_per_hundred, the number of people was filled by Japan, China, the United States, and Germany.



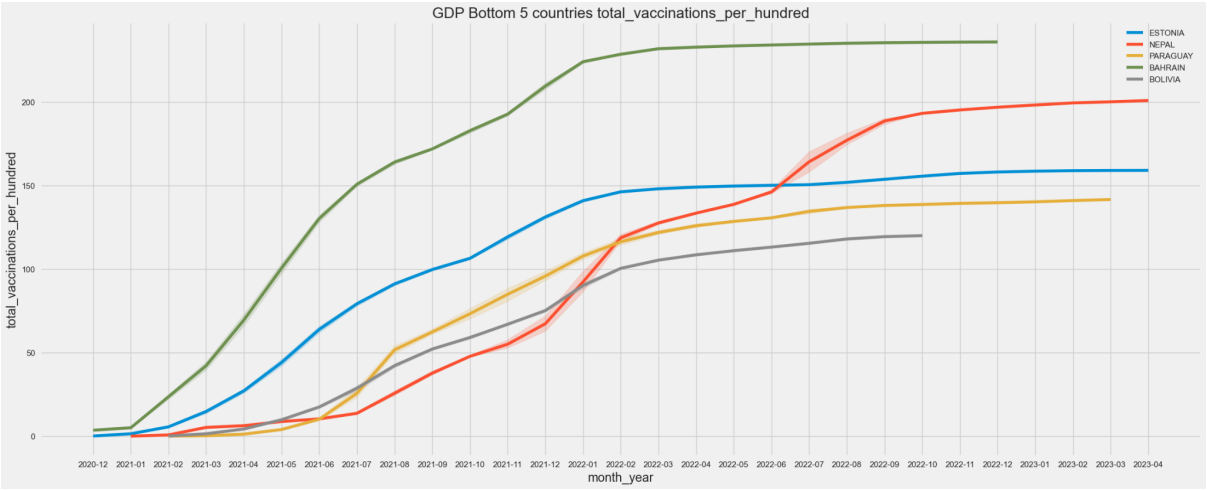
The graph above shows that China's data will be cut off. It is followed by India, USA, Germany. (In the future, in design sketch, I plan to process and use the China data.) Total_vaccination refers to a person who has been vaccinated at least once, and Fully_vacination refers to a person who has been vaccinated more than twice.



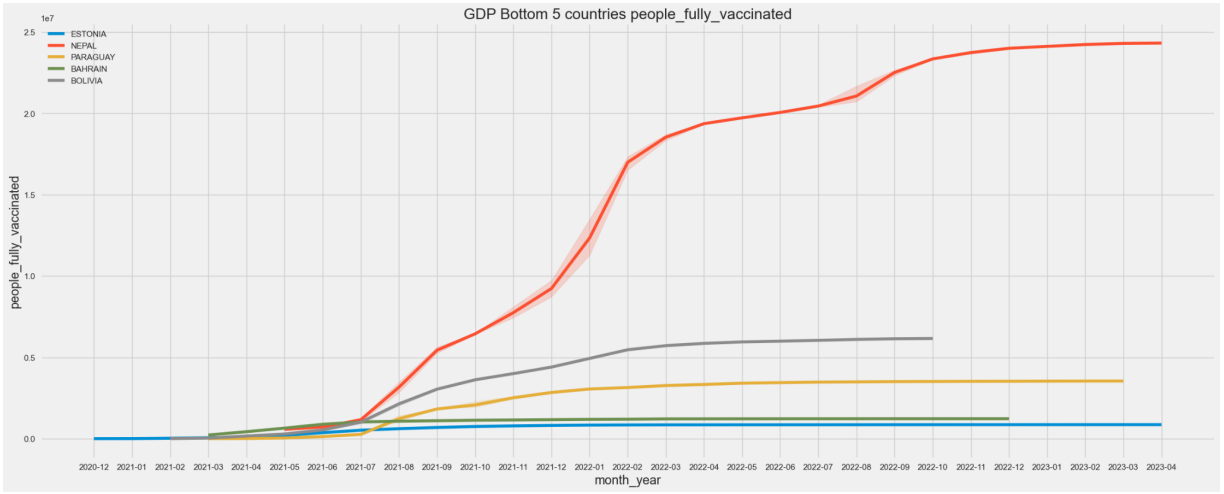
As expected, China is the highest in the graph reflecting the number of people, followed by Japan, the United States, and Germany. It can seen that most people in japan been vaccinated more than twice.



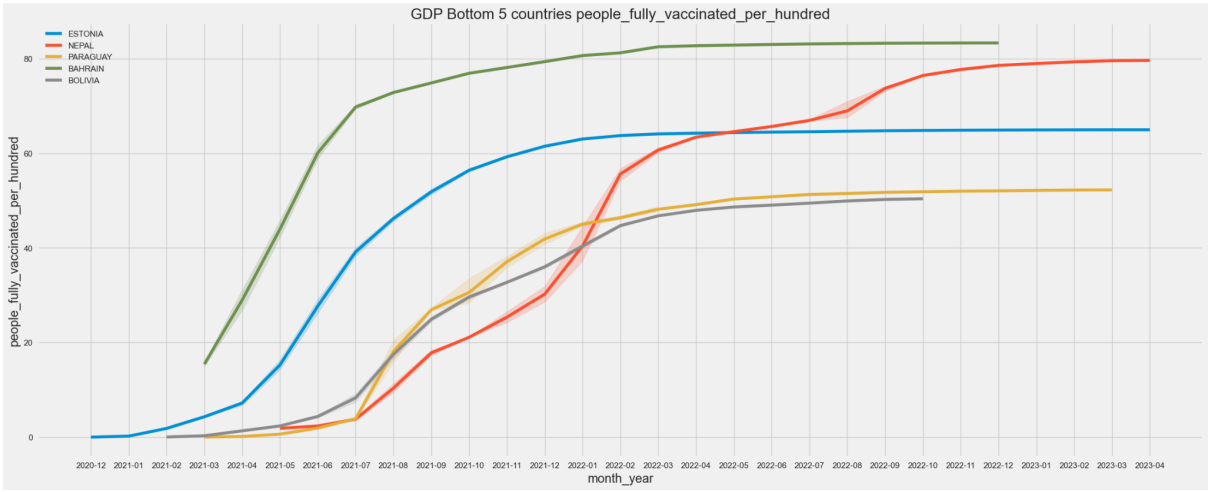
Next, I looked for and analyzed the bottom five countries for calculating the vaccine rate of the top and bottom GDP.
GDP bottom 5 countries are Estonia, Nepal, Paraguay, Bahrain, Bolivia. Nepal has the most total _vaccine, and Bolivia, Paraguay, Bahrain,Estonia follow it.



In the graph reflecting the number of people, Nepal, which was first place above, has decreased, and Bahrain seems to be the country with the most vaccines.



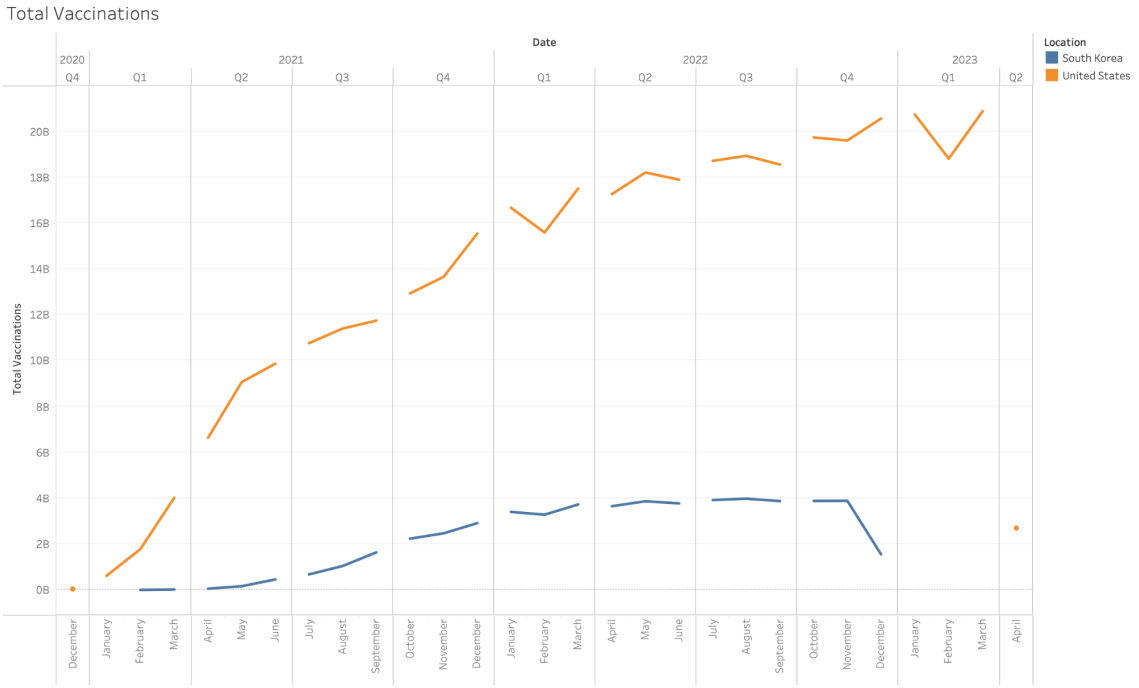
This following is fully_vaccinated in the bottom five countries of GDP that have been vaccinated more than twice. Nepal, which has a larger population than other countries, is also the highest in this area.



It is a graph that reflects the number of people. Before reflecting the population number, Nepal seemed to have the highest vaccination rate, but after reflecting the population number, Bahrain had the highest vaccination rate.

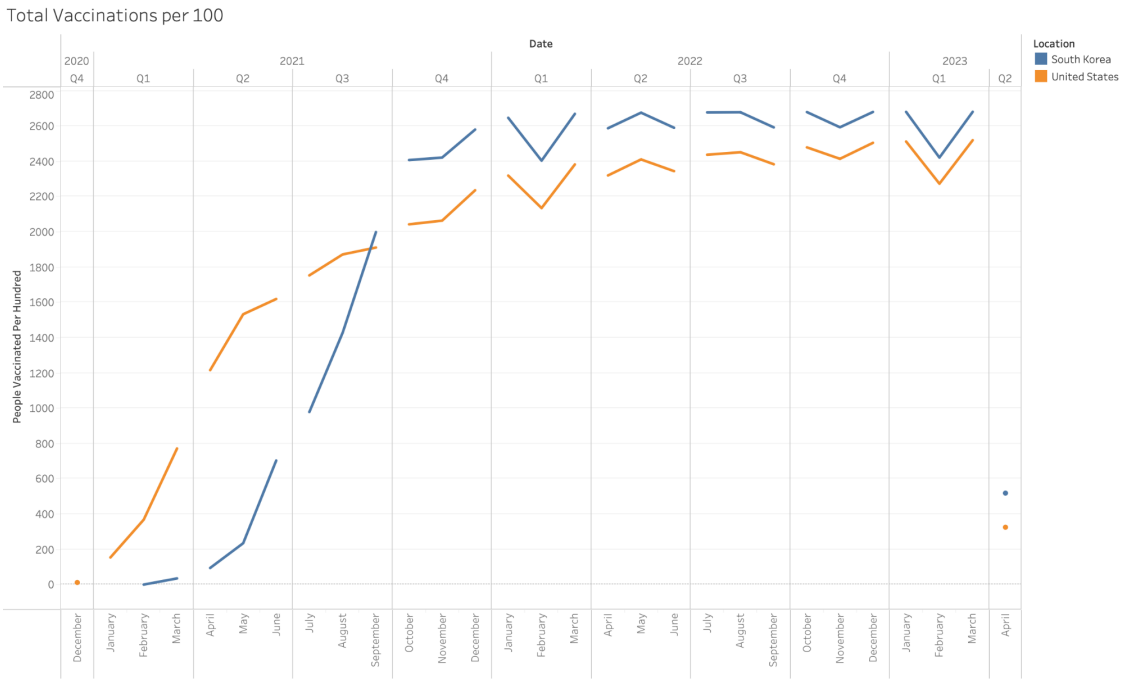
We did additional visualizations about the vaccination dataset.
The following visualizations have been done from the Covid Vaccination in the World dataset, which is updated daily, collected from Our world in Data (OWID) Github repository.

Total Vaccinations:



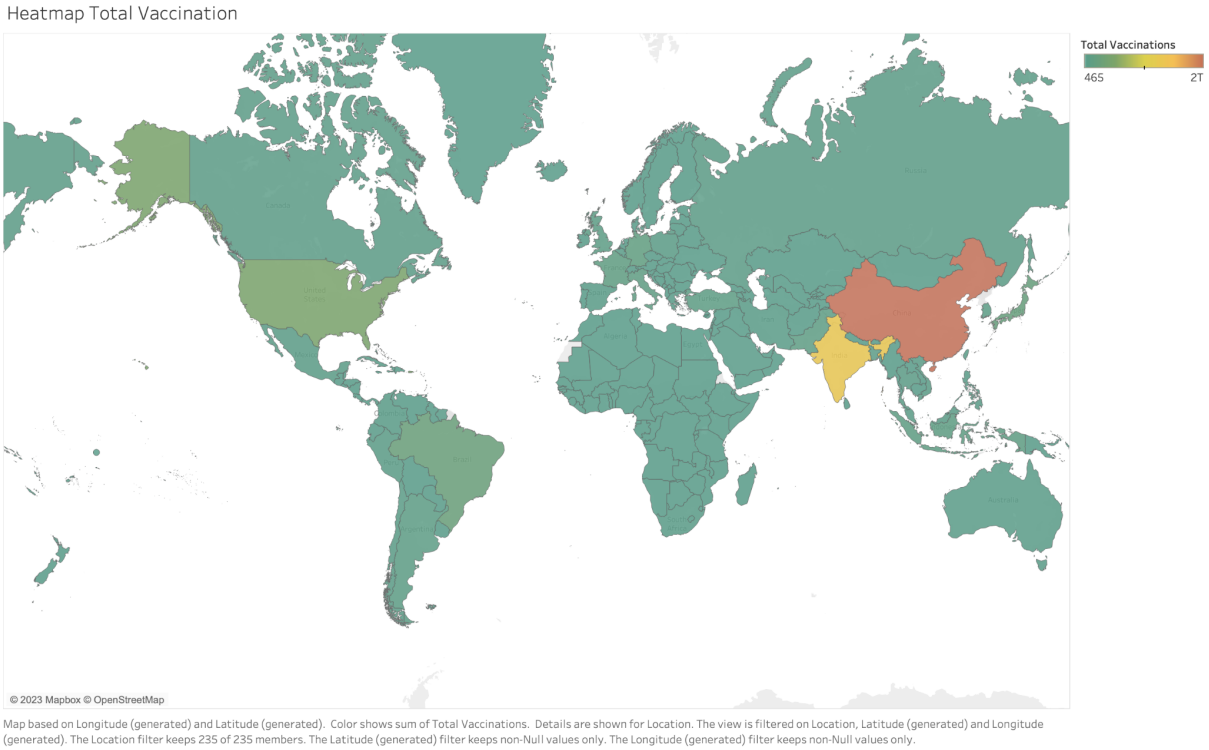
The trend of sum of Total Vaccinations for Date Month broken down by Date Year and Date Quarter. Color shows details about Location. The view is filtered on Location, which keeps South Korea and United States.

This is a temporal visualization which shows the total vaccinations received in South Korea (blue) and United States (Orange) from December 2020 to April 2023.

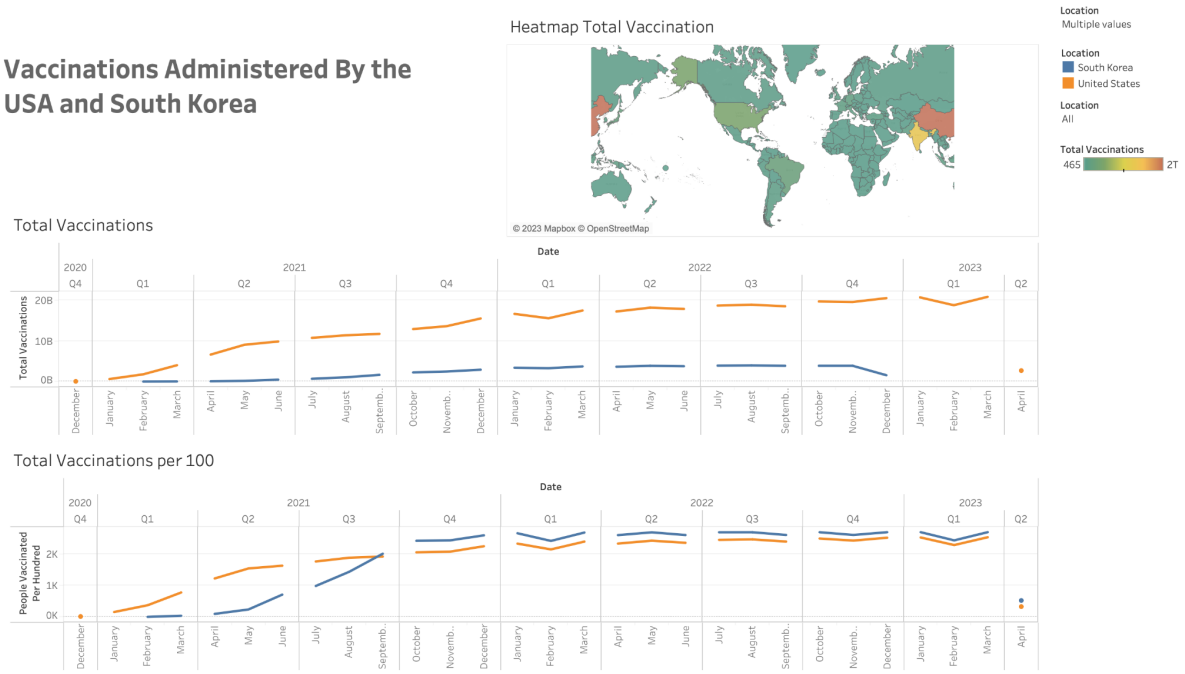


The trend of sum of People Vaccinated Per Hundred for Date Month broken down by Date Year and Date Quarter. Color shows details about Location. The view is filtered on Location, which keeps South Korea and United States.

This is a temporal visualization which shows the total vaccinations administered per 100 persons received in South Korea (blue) and United States (Orange) from December 2020 to April 2023.

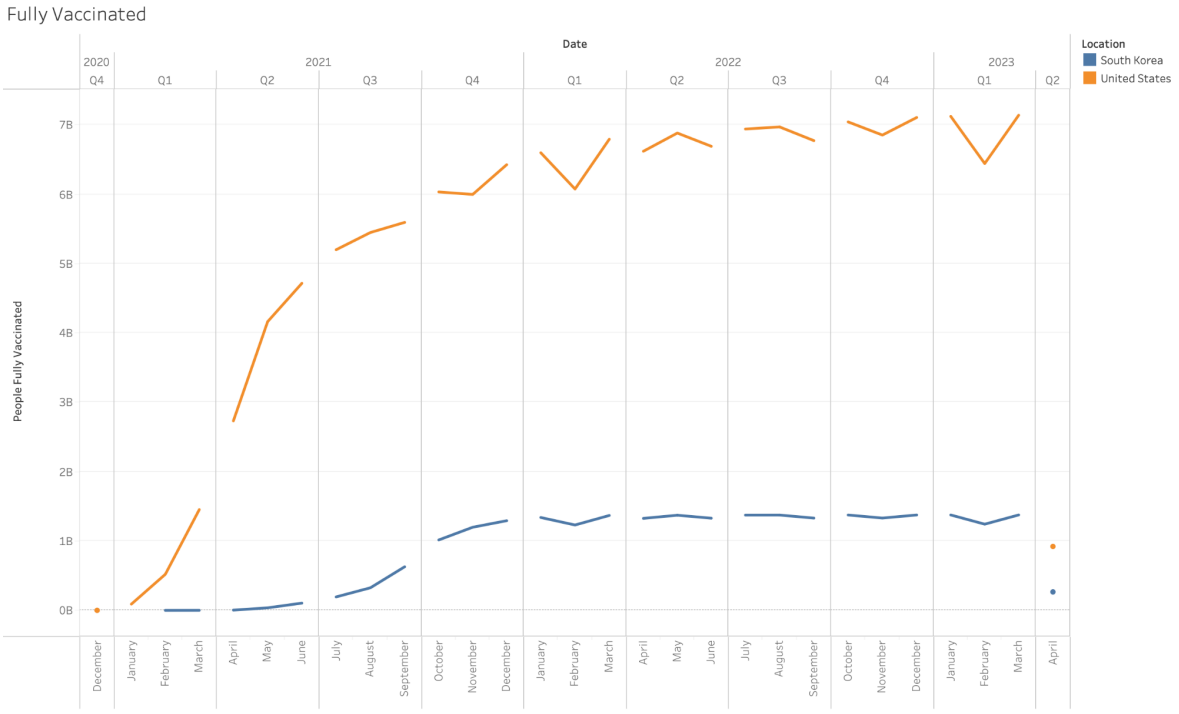


Geospatial heat map where the darker color represents higher number of vaccinations administered and the lighter represents lower number of vaccinations administered.



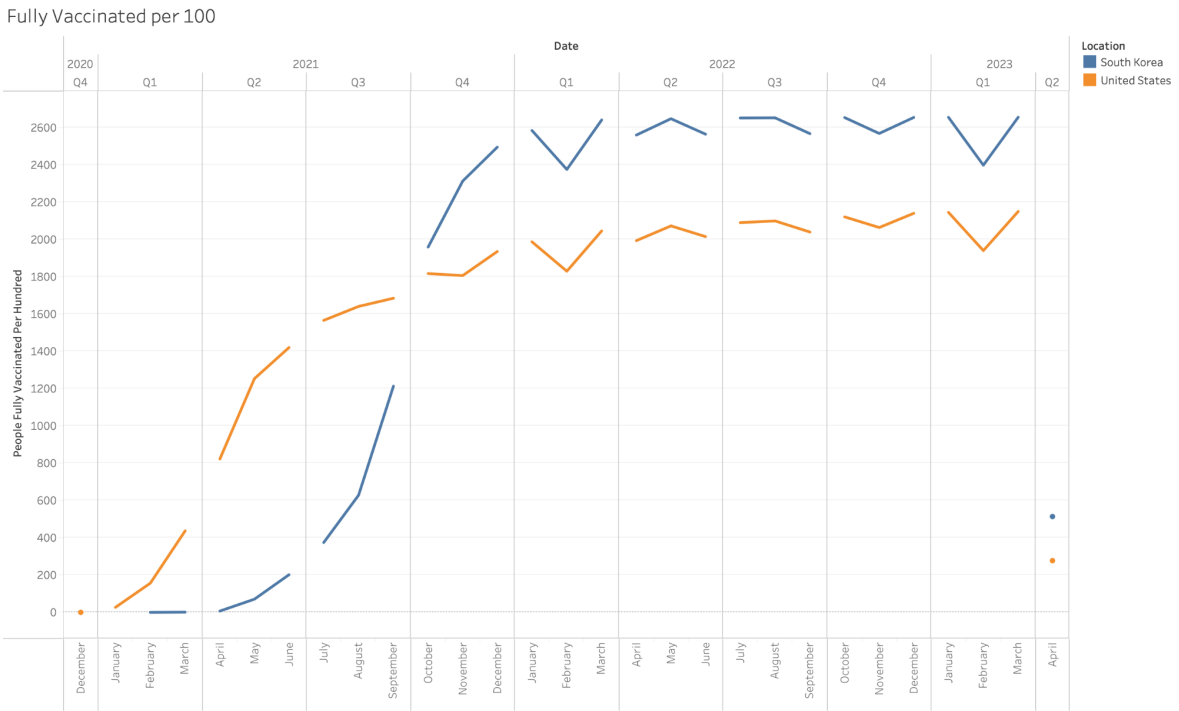
This is the dashboard for the above.

Fully Vaccinated:



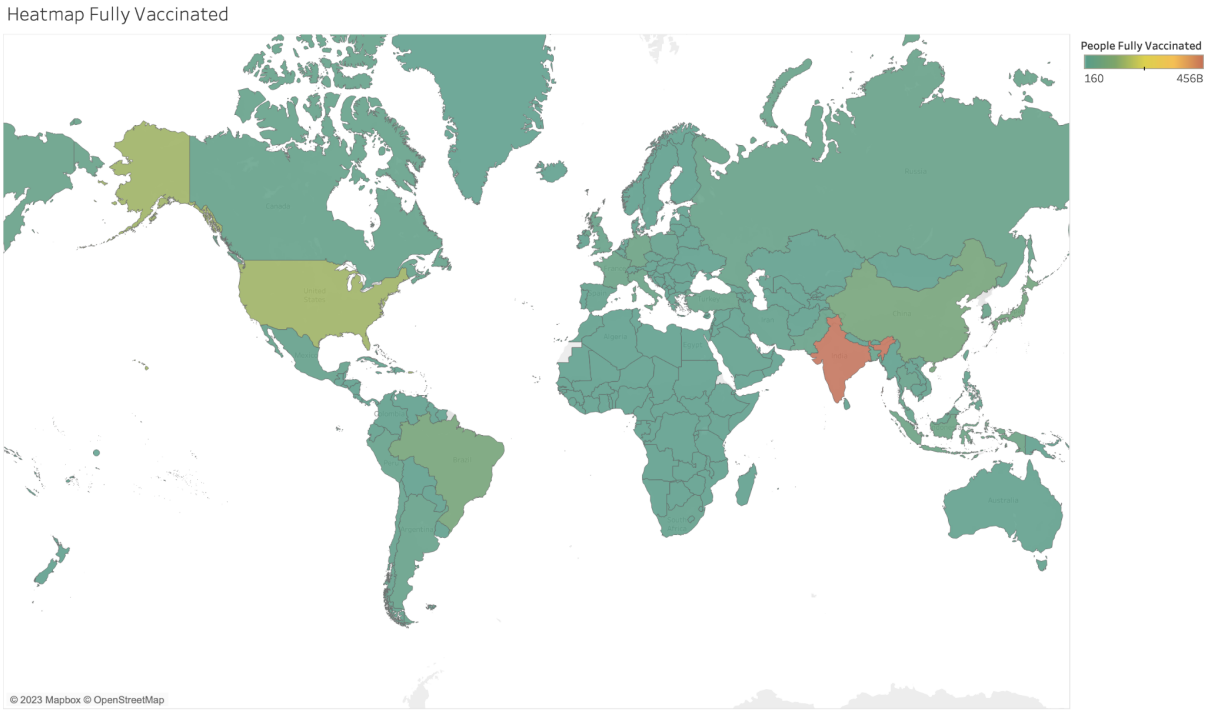
The trend of sum of People Fully Vaccinated for Date Month broken down by Date Year and Date Quarter. Color shows details about Location. The view is filtered on Location, which keeps South Korea and United States.

This is a temporal visualization which shows the people who are fully vaccinated in South Korea (blue) and United States (Orange) from December 2020 to April 2023.

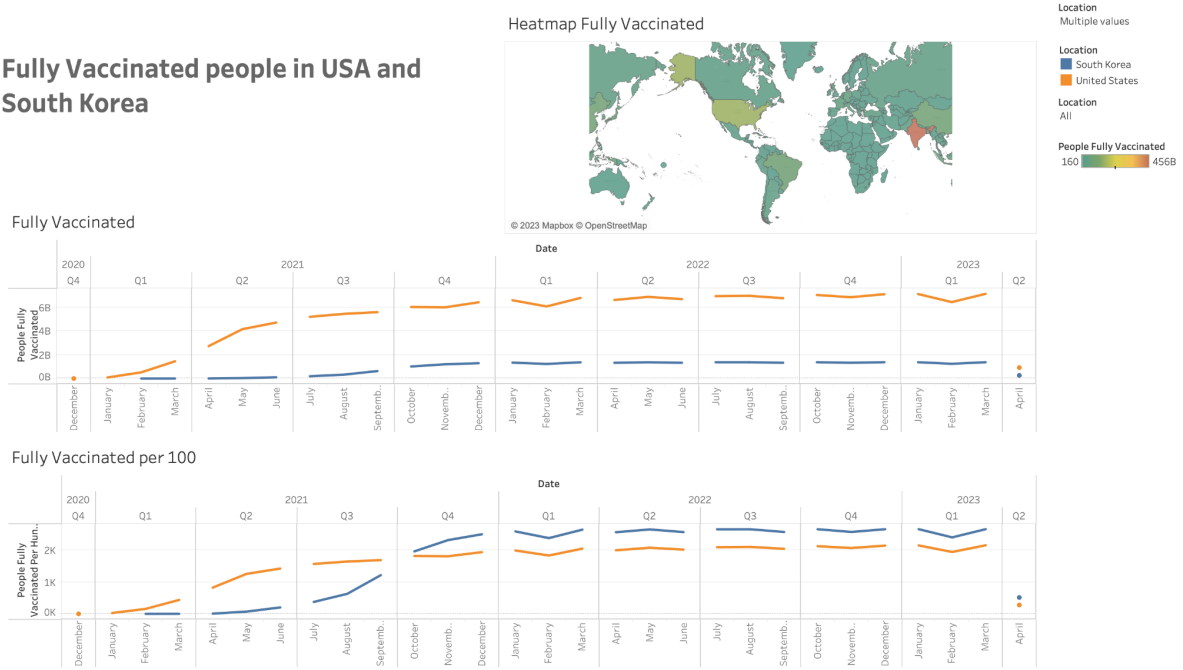


The trend of sum of People Fully Vaccinated Per Hundred for Date Month broken down by Date Year and Date Quarter. Color shows details about Location. The view is filtered on Location, which keeps South Korea and United States.

This is a temporal visualization which shows the people who are fully vaccinated per 100 persons in South Korea (blue) and United States (Orange) from December 2020 to April 2023.

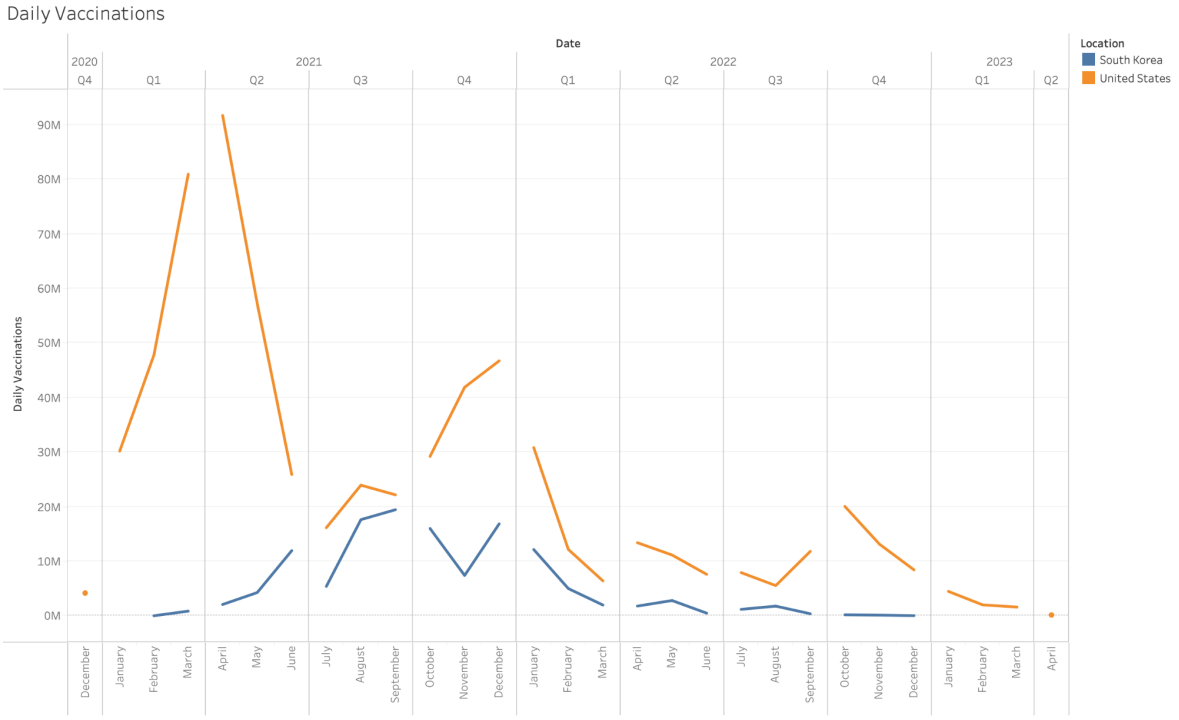


Geospatial heat map where the darker color represents a higher number of fully vaccinated people and the lighter represents a lower number of fully vaccinated people.



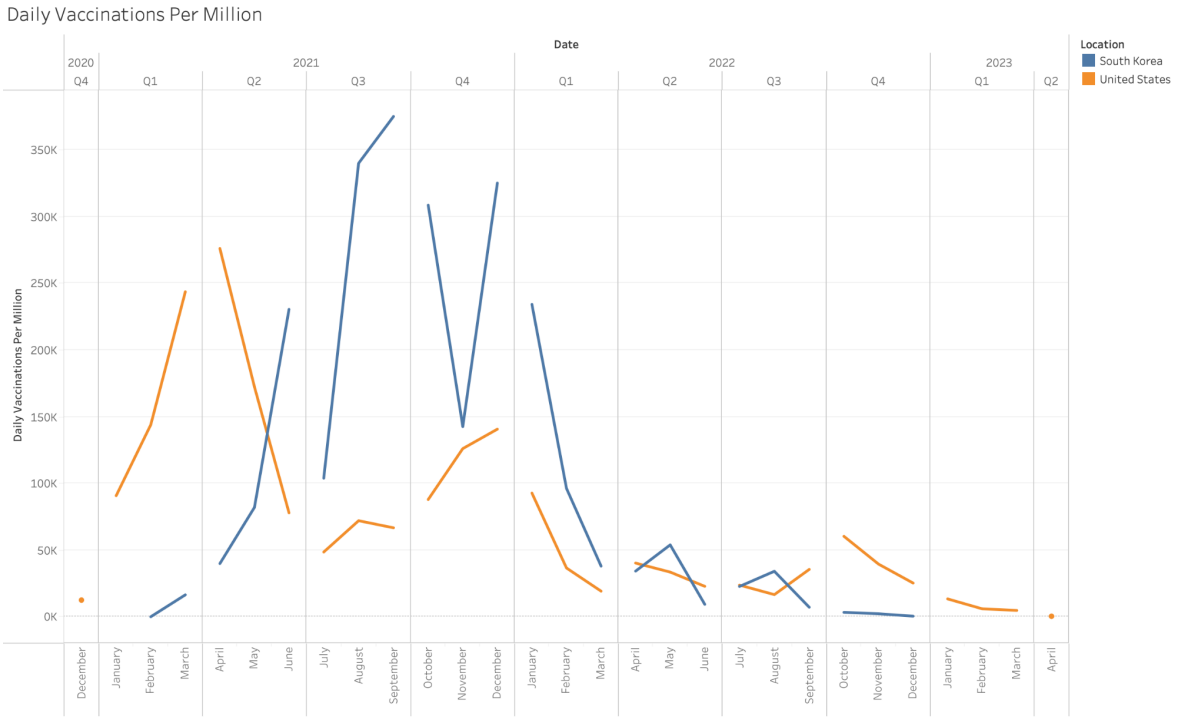
This is the dashboard for the above.

Daily Vaccinations:



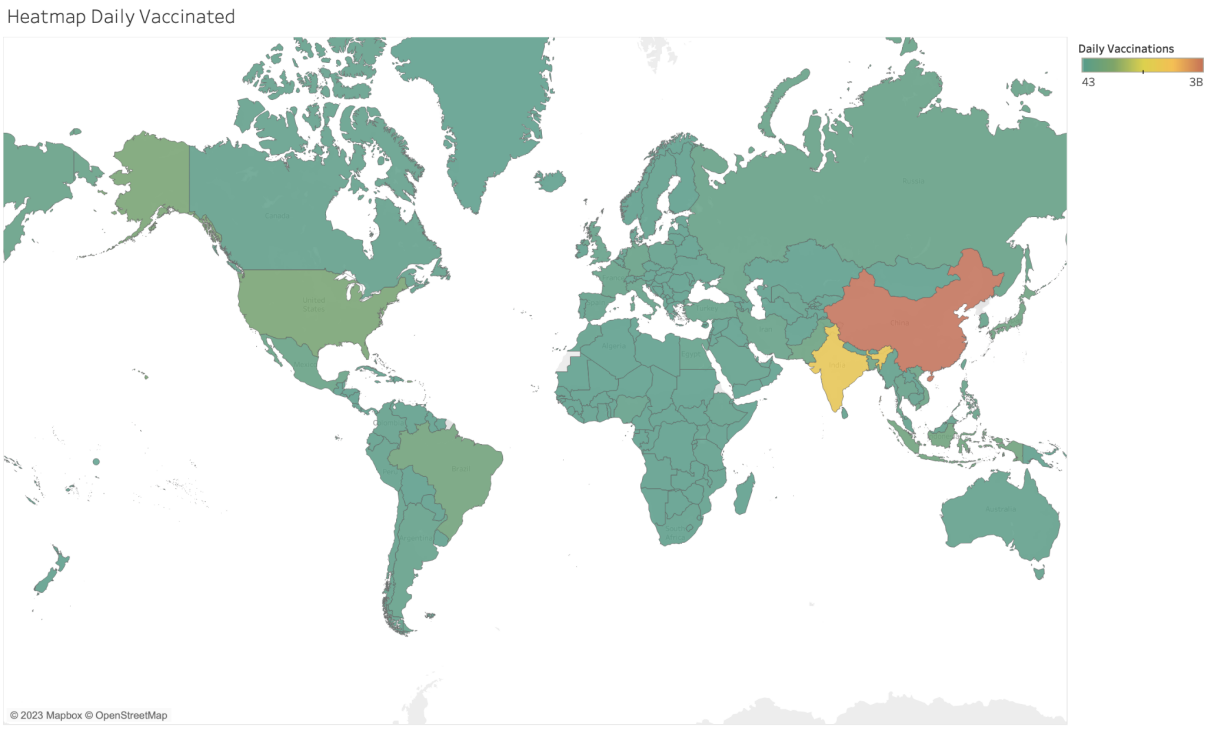
The trend of sum of Daily Vaccinations for Date Month broken down by Date Year and Date Quarter. Color shows details about Location. The view is filtered on Location and sum of Daily Vaccinations. The Location filter keeps South Korea and United States. The sum of Daily Vaccinations filter keeps non-Null values only.

This is a temporal visualization which shows the daily vaccinations administered in South Korea (blue) and United States (Orange) from December 2020 to April 2023.



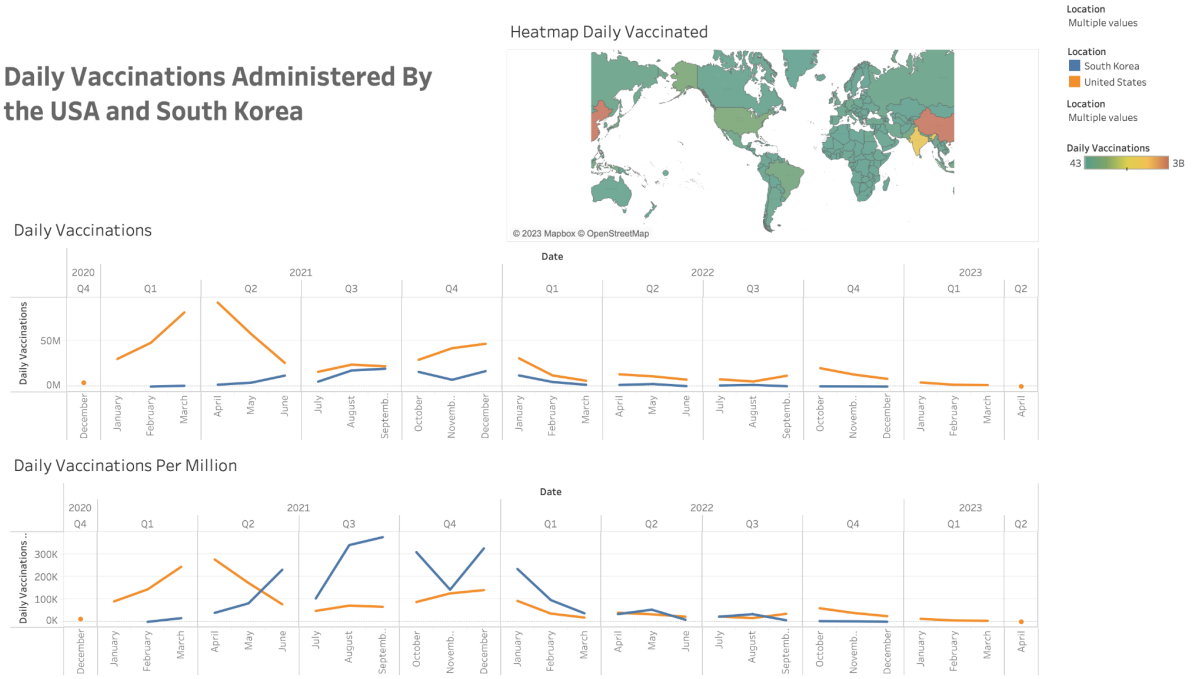
The trend of sum of Daily Vaccinations Per Million for Date Month broken down by Date Year and Date Quarter. Color shows details about Location. The view is filtered on Location, which keeps South Korea and United States.

This is a temporal visualization which shows the daily vaccinations administered by 1,000,000 people in South Korea (blue) and United States (Orange) from December 2020 to April 2023.



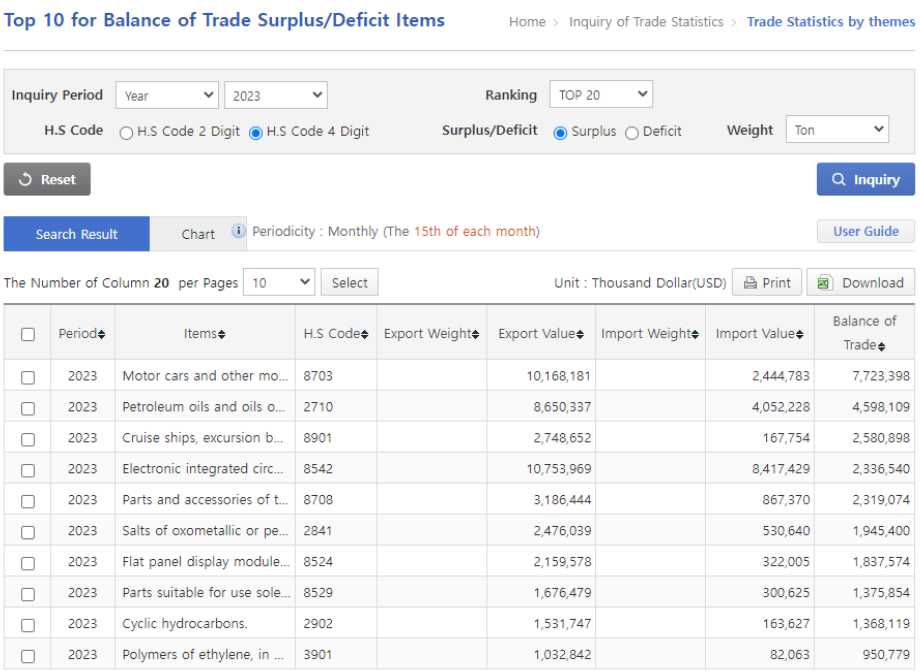
Map based on Longitude (generated) and Latitude (generated). Color shows sum of Daily Vaccinations. Details are shown for Location. The view is filtered on Location, Latitude (generated) and Longitude (generated). The Location filter keeps 235 of 235 members. The Latitude (generated) filter includes everything. The Longitude (generated) filter includes everything.

Geospatial heat map where the darker color represents a higher number of daily vaccinations administered and the lighter represents a lower number of daily vaccinations administered.



This is the dashboard for the above.

We were also curious about how the COVID-19 pandemic impacted specific types of exports. Using the Trade Statistics published by the Korean Customs Service, I compiled a dataset of the top 20 trade surplus items each month, from January 2020 - December 2022. Once the dataset was compiled, the only data cleaning required was to create a simplified name for each category. For example, “Beauty or make-up preparations and preparations for the care of the skin (other than medicaments), including sunscreen or sun tan preparations; manicure or pedicure preparations” became “Beauty and Skincare Products.”



Top 10 for Balance of Trade Surplus/Deficit Items

Home > Inquiry of Trade Statistics > Trade Statistics by themes

Inquiry Period: Year 2023 | Ranking: TOP 20

H.S Code: H.S Code 2 Digit / H.S Code 4 Digit | Surplus/Deficit: Surplus / Deficit | Weight: Ton

Reset | Inquiry

Search Result | Chart | Periodicity : Monthly (The 15th of each month) | User Guide

The Number of Column 20 per Pages 10 Select | Unit : Thousand Dollar(USD) | Print | Download

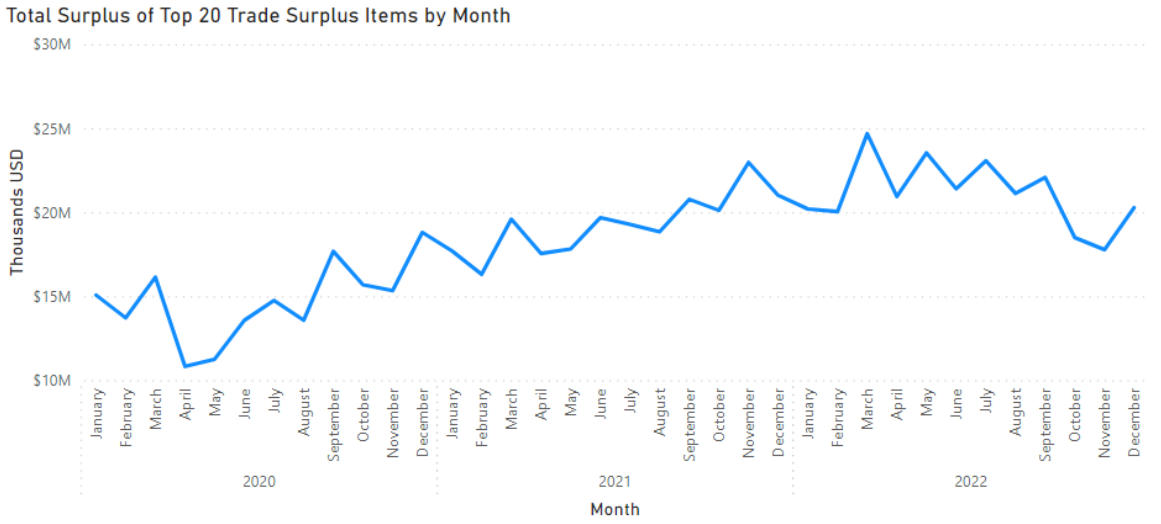
Period	Items	H.S Code	Export Weight	Export Value	Import Weight	Import Value	Balance of Trade
2023	Motor cars and other mo...	8703		10,168,181		2,444,783	7,723,398
2023	Petroleum oils and oils o...	2710		8,650,337		4,052,228	4,598,109
2023	Cruise ships, excursion b...	8901		2,748,652		167,754	2,580,898
2023	Electronic integrated circ...	8542		10,753,969		8,417,429	2,336,540
2023	Parts and accessories of t...	8708		3,186,444		867,370	2,319,074
2023	Salts of oxometallic or pe...	2841		2,476,039		530,640	1,945,400
2023	Flat panel display module...	8524		2,159,578		322,005	1,837,574
2023	Parts suitable for use sole...	8529		1,676,479		300,625	1,375,854
2023	Cyclic hydrocarbons.	2902		1,531,747		163,627	1,368,119
2023	Polymers of ethylene, in ...	3901		1,032,842		82,063	950,779

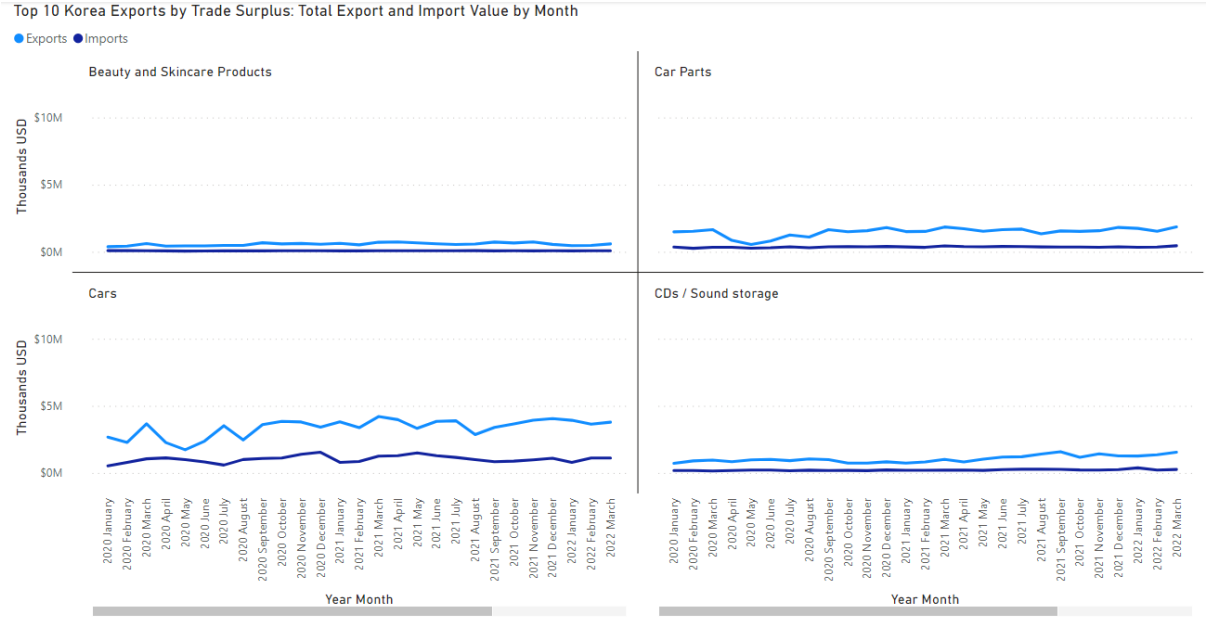
I visualized the data using PowerBI. For the time period examine, electronic integrated circuits, cars, and petroleum oil had the largest overall trade surplus (called positive trade balance). Then I plotted the total surplus of the items by month. We can see that April and May 2020 had the lowest overall trade surplus. This aligns with the time period when the lockdowns were first put in place and were most severe. After that, the trade surplus gradually increases, to a peak in March 2022.

Positive Trade Balance: Top 10 Items

Item	Total Trade Balance (Thousands USD)
Electronic Integrated Circuits	$152,011,018
Cars	$92,588,241
Petroluem Oil	$59,179,931
Large Ships	$41,469,331
Car Parts	$41,437,733
CDs / Sound storage	$30,059,038
Machine Parts	$24,326,535
Cyclic Hydrocarbons	$20,100,052
Beauty and Skincare Products	$17,743,667
Flat Panel Displays	$16,658,028

Next, I looked at specific items to see whether they matched the overall trend. I found that although many items dip in April and May 2020, their multi-year trend does not look the same as the overall trade surplus trend. This suggests that specific items are impacted differently by different events – they are not a monolith. I’ve clipped a few examples below.





Finally, I was curious whether the same items had a trade surplus month over month. I found that 11 items were in the Top 20 trade surplus items every single month. 12 items were in the top 20 surplus list less than 5 times. Further research could include looking at the specific months that these items were a top surplus item, to understand if there is any pattern in why they appeared in the top 20 list when they did.

Most Frequent Trade Surplus Items

Item	Average of Balance of Trade	Standard deviation of Balance of Trade	Count of Item
Beauty and Skincare Products	$492,879.6389	$96,187.6861	36
Car Parts	$1,151,048.1389	$257,053.0664	36
Cars	$2,571,895.5833	$723,743.9655	36
CDs / Sound storage	$834,973.2778	$252,486.679	36
Coated Flat Metal	$423,883.0278	$115,516.0464	36
Cyclic Hydrocarbons	$558,334.7778	$161,124.4095	36
Electronic Integrated Circuits	$4,222,528.2778	$997,117.5435	36
Large Ships	$1,151,925.8611	$415,748.1521	36
Petroluem Oil	$1,643,886.9722	$1,058,114.8989	36
Polyacetals	$419,109.5	$108,023.33	36
Propylene Polymers	$387,729.3056	$89,349.6048	36

Least Frequent Trade Surplus Items

Item	Average of Balance of Trade	Standard deviation of Balance of Trade	Count of Item
Acyclic Hydrocarbons	$261,744.25	$35,421.9931	4
Gold	$195,756	$0	1
Measuring items	$314,229	$0	1
Optical Fibres	$246,896	$10,290	2
Ovens	$357,061.5	$56,708.9446	4
Pre-fab Building Materials	$294,192.5	$278.5	2
Printed Circuits	$200,433.3333	$13,764.9334	3
Refrigerators / Freezers	$282,175.5	$5,566.5	2
Ships, misc.	$297,498	$0	1
Tyres	$218,305	$0	1
Uncoated Flat Metal, Cold-Rolled	$264,647	$12,548.227	3
Unwrought Zinc	$217,987.5	$13,408.5	2