I am a data analyst working for the UK government and my task is to analyse COVID-19 data recorded by the government to help find solution to increase the number of fully vaccinated individuals. Please follow through my Jupyter Notebook where I will guide through all the decision makings and analysis I made.

The first step I took is to import all the necessary libraries among Python to work through my analysis, including the most common ones like Pandas and Numpy for my exploratory analysis; and Matploitlib and Seaborn for my visualisations. The data sets I used are the 5 CSV files provided, and I imported all them into my Notebook then assigning each to a data I created.

My initial steps to tackle the business question is to explore all the data sets I imported. First, I viewed all the values inside each data frame to have a clear understanding of the types of data and the amount of data I'm dealing with. One question that came up during the process is the number of rows in many of the data frames I went through, and how can I efficiently use them during my analysis. Next, I determined any missing values from the data sets, which when I was exploring the data sets, I know because of the nature of the data, in terms of statistical records for Covid-19 cases, it is expected to have 0 as a value for certain columns. In order to answer my previous question, I narrowed down the Covid-19 cases data frame by subsetting it to only view values for Gibraltar, and by focusing on the deaths, cases and hospitalised columns I got the vital information on how the pandemic has affected region, and whether the peak has passed yet.

My next step towards my analysis is to merge the cases and vaccination data sets to get a more detailed understanding on how the pandemic has affect each region within UK and how the vaccination campaign has progressed. The purpose of this is to combine together all the relevant columns into one data frame, which then I can use vaccination and cases, deaths, recovered and hospitalised data for further analysis and visualisation. After sense checking the data sets have been merged correctly, I decided to convert the 'Date' column data type from string value to Python's datatime value. Then I filtered out all the relevant columns using the pivot function to make comparison, from which I noticed that the 'Deaths', 'Cases', 'Recovered' and 'Hospitalised' columns contain cumulative data and the three vaccination columns contain daily value data. However, I'm taking assumption that the vaccination data contain errors during data entry to explain the high figures compare to actual population size for all regions except 'Others'; thus, I will continue my analysis based upon the assumption.

I aggregated the 'Cases' column from the data frame to find out the regions latest infection cases, and region 'Others', by excluding all other regions, represent UK mainland and has the highest cases, which I will be focusing on during my future analysis. Then I did the same to find out the total vaccination number for 'Others' and also the number of individuals who has yet to receive their second dose. In contrast, there were over 8 million cases across mainland UK and only 2.4 million people have been fully vaccinated, but most have received their second dose, with around 116 thousand who still haven't. Overall, all regions have around 95.5% of population who received both doses, but Gibraltar has the highest number of people that haven't received second dose, around 260 thousand. Refer to appendix a and b for comparison in percentage of vaccinations received. Next, I will use visualisations to demonstrate and discuss the statistical data from my descriptive analysis. I will separate the analysis into two parts, one focusing on UK mainland, and the other focusing on the rest of the regions. In terms of answering the question of identifying the region with greatest number of recoveries, Channel Island is the answer, and therefore can be ignored for the time being. One point I want to emphasise is the number of deaths across all regions is still rising, and the peak has not yet been reached, the trend is particularly rising in UK mainland and I will suggest we should focus our vaccination campaign towards this region first. Please refer to appendix d and f for visualisation.

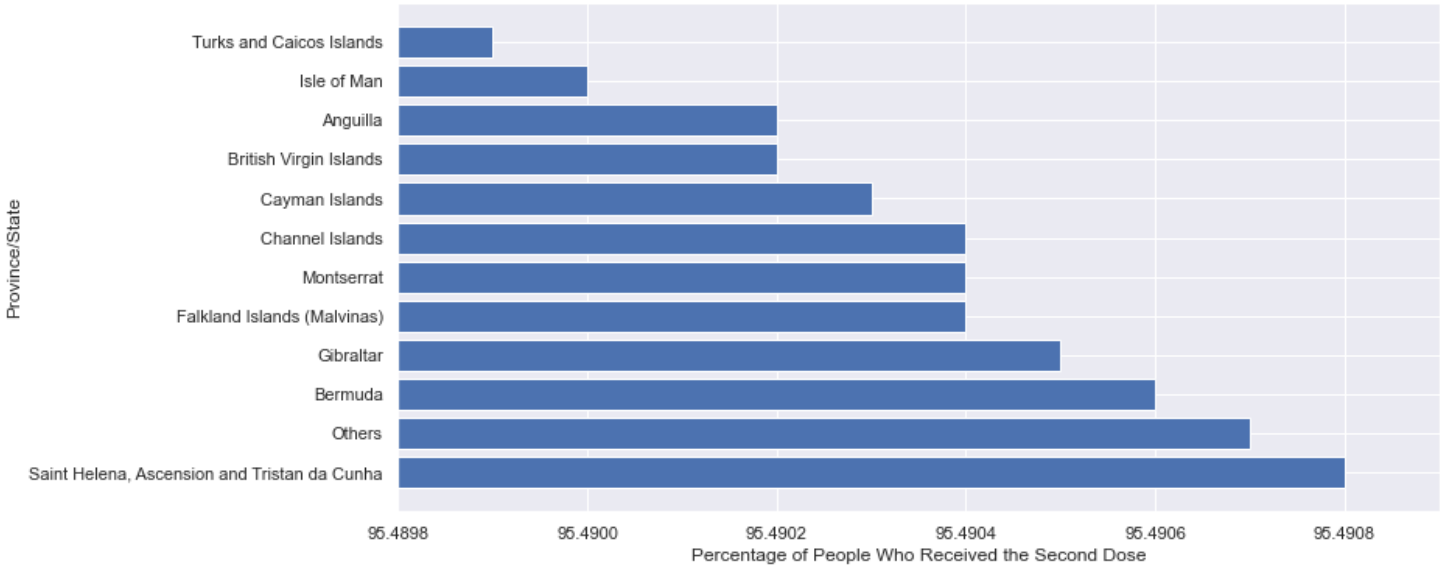
Another point I want to add is when focusing on UK mainland, I analysed the total cases compare to deaths, recoveries and vaccinations received, and the result is that the number of cases and deaths are all the highest among all regions, while the number of vaccinations is rather low, leveraging around 2.5 million doses. Deaths and cases rate are also still on the rise meaning the peak has yet been reached over time, therefore I suggest this should be the region to focus on the vaccination campaign, not limited to just increase second dose vaccine rate, but also first dose vaccine rate. In comparison, the remaining 11 regions have much lower cases and deaths, and also the deaths visualisation suggests only a few regions still have a sign showing an increase.

Moving on my analysis towards the Twitter data points and finding out the number of tweets contain hashtag posts mentioning Covid-19. From the data set, all main hashtags used are relating to Covid-19, and I believe it's a sign that the public are showing concern about the pandemic and it will be a good window to push up on the campaign. I believe the use of external data can help us to better picture the scenario but we shouldn't rely on it too heavily due the validity of data we will be working with.

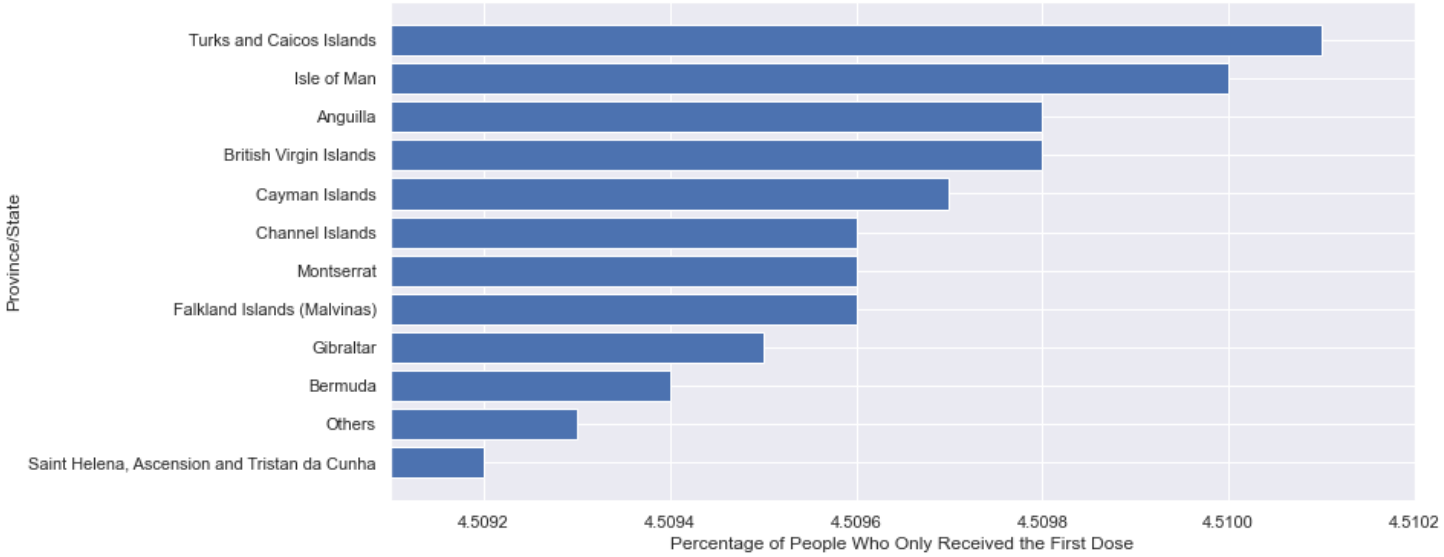
In the end, my suggestion from the analysis is that the government should focus on ‘Others’ region (UK mainland) for the future vaccination campaign. From the last section of my Jupyter Notebook I carried out predictive analysis where in my presentation I will talk about potential future outcomes based on the trends in hospitalisation rates based on my time-series analysis and how my suggestion will be effective for the scenario. and going back to my assumption on the data error in the vaccination data set, I believe data entry has been falsely adding people receiving their second dose to the first dose column, resulting in why there was such a high number of vaccinations given out daily.

Appendix:

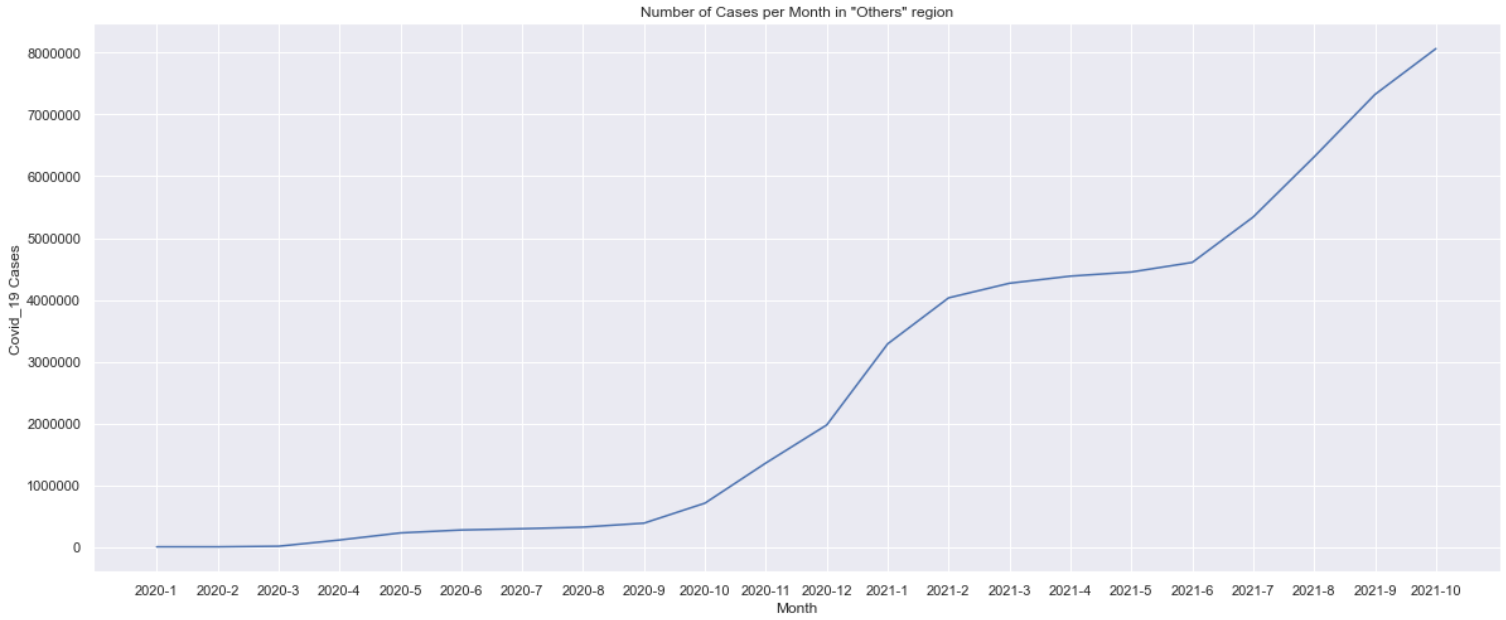
a:


Turks and Caicos Islands
Isle of Man
Anguilla
British Virgin Islands
Cayman Islands
Channel Islands
Montserrat
Falkland Islands (Malvinas)
Gibraltar
Bermuda
Others
Saint Helena, Ascension and Tristan da Cunha
Province/State
95.4898
95.4900
95.4902
95.4904
95.4906
95.4908
Percentage of People Who Received the Second Dose

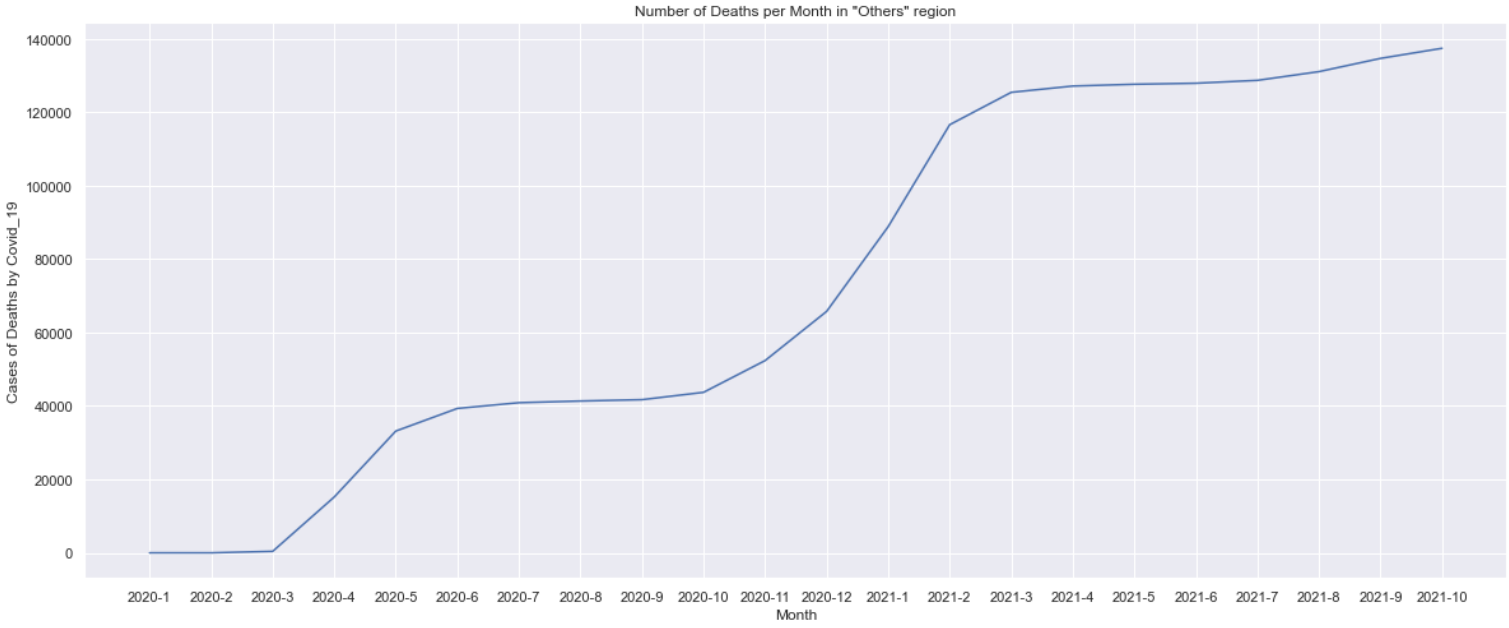
b:


Turks and Caicos Islands
Isle of Man
Anguilla
British Virgin Islands
Cayman Islands
Channel Islands
Montserrat
Falkland Islands (Malvinas)
Gibraltar
Bermuda
Others
Saint Helena, Ascension and Tristan da Cunha
Province/State
4.5092
4.5094
4.5096
4.5098
4.5100
4.5102
Percentage of People Who Only Received the First Dose

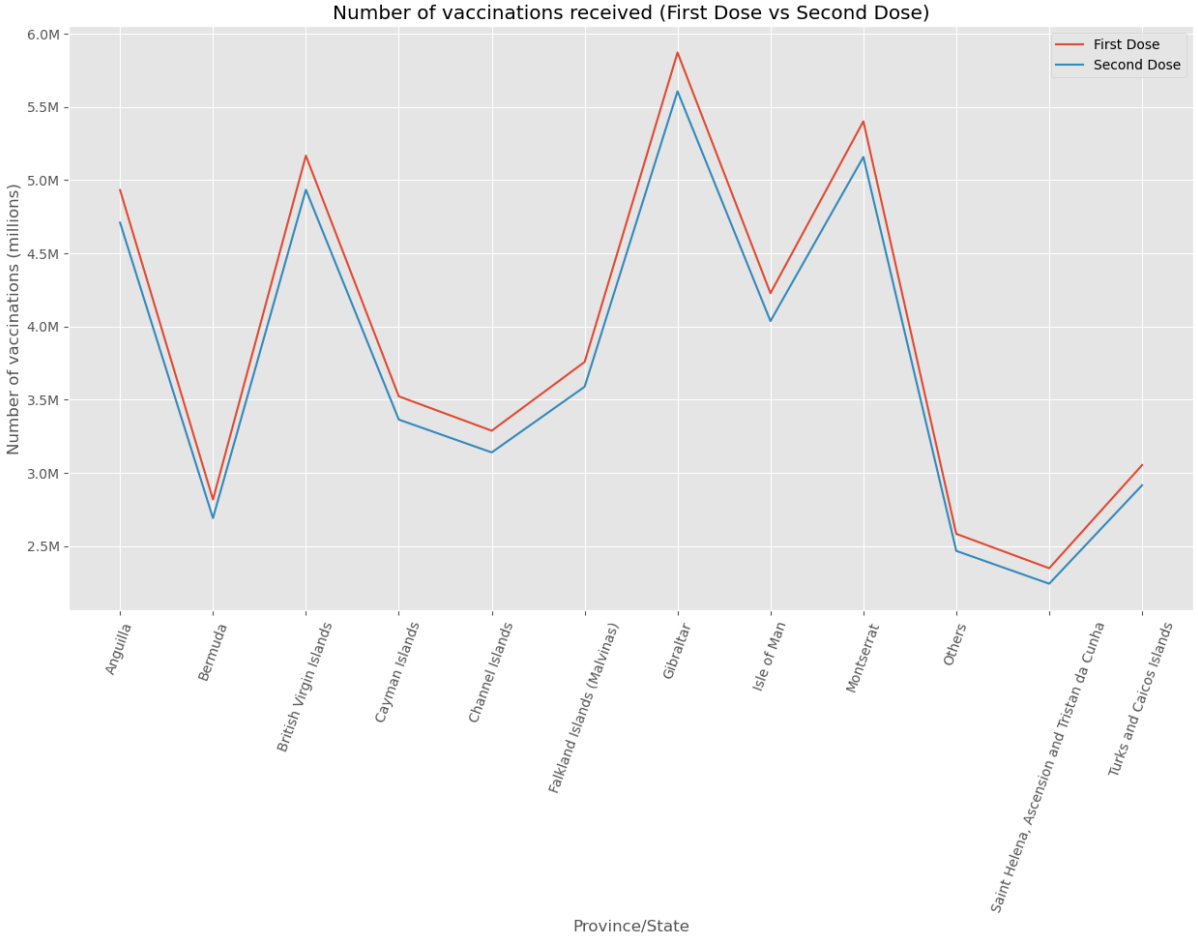
c:


Number of Cases per Month in "Others" region
Covid_19 Cases
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
2020-1 2020-2 2020-3 2020-4 2020-5 2020-6 2020-7 2020-8 2020-9 2020-10 2020-11 2020-12 2021-1 2021-2 2021-3 2021-4 2021-5 2021-6 2021-7 2021-8 2021-9 2021-10
Month

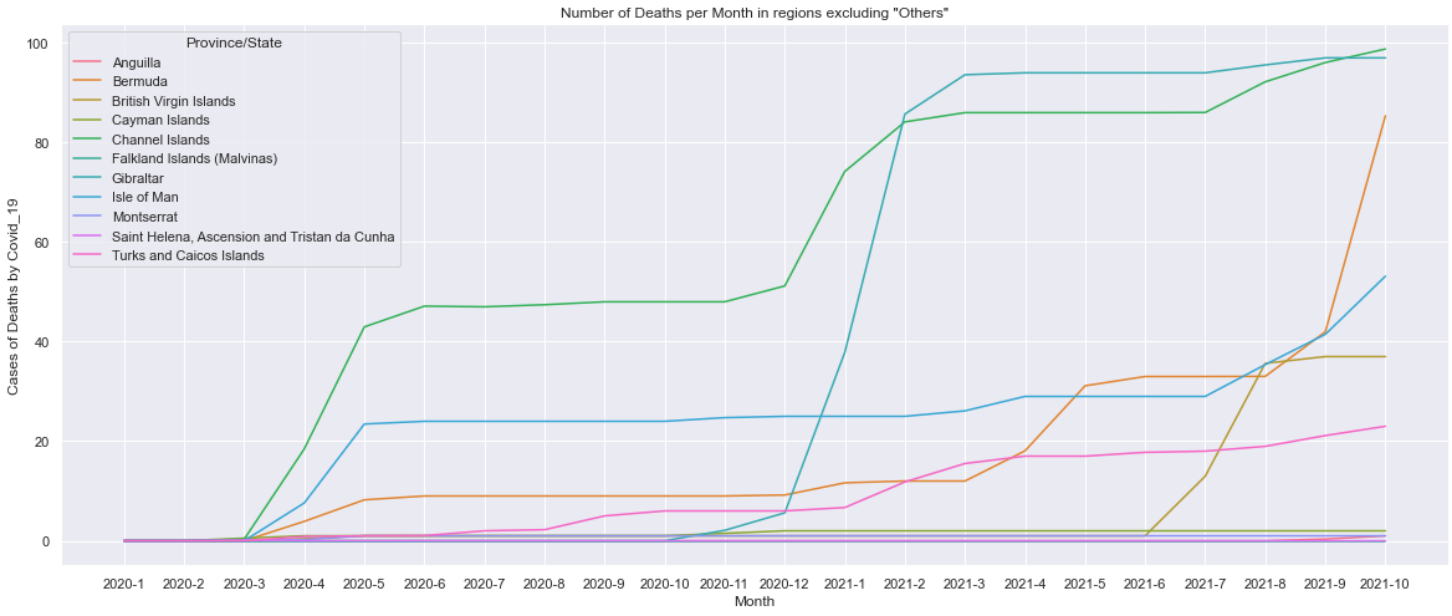
d:


Number of Deaths per Month in "Others" region
Cases of Deaths by Covid_19
140000
120000
100000
80000
60000
40000
20000
0
2020-1 2020-2 2020-3 2020-4 2020-5 2020-6 2020-7 2020-8 2020-9 2020-10 2020-11 2020-12 2021-1 2021-2 2021-3 2021-4 2021-5 2021-6 2021-7 2021-8 2021-9 2021-10
Month

e:


Number of vaccinations received (First Dose vs Second Dose)
First Dose
Second Dose
Number of vaccinations (millions)
6.0M
5.5M
5.0M
4.5M
4.0M
3.5M
3.0M
2.5M
Anguilla
Bermuda
British Virgin Islands
Cayman Islands
Channel Islands
Falkland Islands (Malvinas)
Gibraltar
Isle of Man
Montserrat
Others
Saint Helena, Ascension and Tristan da Cunha
Turks and Caicos Islands
Province/State

f:


Number of Deaths per Month in regions excluding "Others"
Province/State
Anguilla
Bermuda
British Virgin Islands
Cayman Islands
Channel Islands
Falkland Islands (Malvinas)
Gibraltar
Isle of Man
Montserrat
Saint Helena, Ascension and Tristan da Cunha
Turks and Caicos Islands
Cases of Deaths by Covid_19
100
80
60
40
20
0
2020-1
2020-2
2020-3
2020-4
2020-5
2020-6
2020-7
2020-8
2020-9
2020-10
2020-11
2020-12
2021-1
2021-2
2021-3
2021-4
2021-5
2021-6
2021-7
2021-8
2021-9
2021-10
Month