COVID-19 Infection Rates and Analysis

Project by Ryan Li Jian Tang

Table of Contents:

Data Cleansing

Process of Data Cleansing.

In [1]:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
```

First Dataframe

In [2]:

```
df = pd.read_csv('Covid-data.csv')
df.shape       # to check how many columns and rows there are.
```

Out[2]:

```
(1575, 8)
```

Checking through the 'location' column:

Checking through the types of available countries in the dataset. Furthermore, if there are any noticeable errors, make ammends to them.

In [3]:

```
df['location'].unique()
```

Out[3]:

```
array(['Australia', 'Australia  ', 'China', ' China', 'France', 'Iran',
       'iran', 'Italy', 'Itly', 'Spain', 'United Kingdom',
       'UnitedKingdom', 'United States', 'United Stats'], dtype=object)
```

As seen above, there are data that have been mispelt or have the same name but contains whitespaces. However, Python treats them differently due to the whitespaces; thus, they need to be corrected to prevent any data from not being called in future commands.

In [4]:

```
df['location'] = df['location'].replace(['Australia  '], 'Australia')
df['location'] = df['location'].replace([' China'], 'China')
df['location'] = df['location'].replace(['iran'],'Iran')
df['location'] = df['location'].replace(['Itly'], 'Italy')
df['location'] = df['location'].replace(['UnitedKingdom'], 'United Kingdom')
df['location'] = df['location'].replace(['United Stats'], 'United States')
```

In [5]:

```
df['location'].unique()     # Double check to see if changes have been made
```

Out[5]:

```
array(['Australia', 'China', 'France', 'Iran', 'Italy', 'Spain',
       'United Kingdom', 'United States'], dtype=object)
```

Checking through 'date' column:

In [6]:

```
df['date'].unique()  # Looking through format of the dates, to see if there are any mistake
```

Out[6]:

```
array(['2019-12-31', '2020-01-01', '2020-01-02', '2020-01-03',
       '2020-01-04', '2020-01-05', '2020-01-06', '2020-01-07',
       '2020-01-08', '2020-01-09', '2020-01-10', '2020-01-11',
       '2020-01-12', '2020-01-13', '2020-01-14', '2020-01-15',
       '2020-01-16', '2020-01-17', '2020-01-18', '2020-19-01',
       '2020-01-20', '2020-01-21', '2020-01-22', '2020-01-23',
       '2020-01-24', '2020-01-25', '2020-01-26', '2020-01-27',
       '2020-01-28', '2020-01-29', '2020-01-30', '2020-31-01',
       '2020-02-01', '2020-02-02', '2020-02-03', '2020-02-04',
       '2020-02-05', '2020-02-06', '2020-02-07', '2020-02-08',
       '2020-02-09', '2020-02-10', '2020-02-11', '2020-02-12',
       '2020-02-13', '2020-02-14', '2020-02-15', '2020-02-16',
       '2020-02-17', '2020-18-02', '2020-02-19', '2020-20-02',
       '2020-02-21', '2020-02-22', '2020-02-23', '2020-02-24',
       '2020-02-25', '2020-02-26', '2020-02-27', '2020-02-28',
       '2020-02-29', '2020-03-01', '2020-03-02', '2020-03-03',
       '2020-03-04', '2020-03-05', '2020-03-06', '2020-03-07',
       '2020-03-08', '2020-03-09', '2020-03-10', '2020-03-11',
       '2020-03-12', '2020-03-13', '2020-03-14', '2020-03-15',
       '2020-03-16', '2020-03-17', '2020-03-18', '2020-19-03',
       '2020-03-20', '2020-21-03', '2020-22-03', '2020-03-23',
       '2020-03-24', '2020-03-25', '2020-26-03', '2020-03-27',
       '2020-28-03', '2020-03-29', '2020-03-30', '2020-03-31',
       '2020-04-01', '2020-04-02', '2020-04-03', '2020-04-04',
       '2020-04-05', '2020-04-06', '2020-04-07', '2020-04-08',
       '2020-04-09', '2020-04-10', '2020-04-11', '2020-04-12',
       '2020-04-13', '2020-04-14', '2020-04-15', '2020-04-16',
       '2020-04-17', '2020-04-18', '2020-04-19', '2020-04-20',
       '2020-04-21', '2020-04-22', '2020-04-23', '2020-04-24',
       '2020-04-25', '2020-04-26', '2020-27-04', '2020-04-28',
       '2020-04-29', '2020-30-04', '2020-05-01', '2020-05-02',
       '2020-05-03', '2020-05-04', '2020-05-05', '2020-05-06',
       '2020-05-07', '2020-05-08', '2020-05-09', '2020-05-10',
       '2020-05-11', '2020-05-12', '2020-13-05', '2020-14-05',
       '2020-05-15', '2020-05-16', '2020-05-17', '2020-05-18',
       '2020-05-19', '2020-05-20', '2020-05-21', '2020-05-22',
       '2020-23-05', '2020-05-24', '2020-05-25', '2020-05-26',
       '2020-05-27', '2020-28-05', '2020-05-29', '2020-30-05',
       '2020-05-31', '2020-06-01', '2020-06-02', '2020-06-03',
       '2020-06-04', '2020-06-05', '2020-06-06', '2020-06-07',
       '2020-06-08', '2020-06-09', '2020-06-10', '2020-06-11',
       '2020-06-12', '2020-06-13', '2020-06-14', '2020-06-15',
       '2020-16-06', '2020-06-17', '2020-18-06', '2020-06-19',
       '2020-06-20', '2020-06-21', '2020-06-22', '2020-06-23',
       '2020-24-06', '2020-06-25', '2020-06-26', '2020-06-27',
       '2020-06-28', '2020-06-29', '2020-06-30', '2020-07-01',
       '2020-07-02', '2020-07-03', '2020-07-04', '2020-07-05',
       '2020-07-06', '2020-07-07', '2020-07-08', '2020-07-09',
       '2020-07-10', '2020-07-11', '2020-07-12', '2020-07-13',
       '2020-07-14', '2020-14-01', '2020-18-01', '2020-01-19',
       '2020-21-01', '2020-27-01', '2020-01-31', '2020-02-18',
       '2020-02-20', '2020-27-02', '2020-14-03', '2020-16-03',
       '2020-03-19', '2020-03-22', '2020-25-03', '2020-03-26',
       '2020-03-28', '2020-31-03', '2020-04-27', '2020-04-30',
       '2020-05-13', '2020-05-14', '2020-05-28', '2020-15-06',
```

```
       '2020-06-16', '2020-17-06', '2020-23-06', '2020-06-24',
       '2020-25-06', '2020-13-07', '2020-14-07', '2020-23-01',
       '2020-26-01', '2020-30-01', '2020-17-02', '2020-21-02',
       '2020-25-02', '2020-29-02', '2020-20-03', '2020-03-21',
       '2020-24-03', '2020-21-04', '2020-24-04', '2020-19-05',
       '2020-24-05', '2020-05-30', '2020-21-06', '2020-30-06',
       '2020-13-01', '2020-16-01', '2020-13-02', '2020-16-02',
       '2020-26-02', '2020-23-03', '2020-17-04', '2020-18-05',
       '2020-06-18', '2020-19-06', '2020-20-06', '2020-22-06',
       '2020-26-06', '2020-27-06', '2020-29-06', '2020-17-01',
       '2020-20-01', '2020-28-01', '2020-29-01', '2020-19-02',
       '2020-22-02', '2020-27-03', '2020-13-04', '2020-25-04',
       '2020-15-05', '2020-05-23', '2020-26-05', '2020-29-05',
       '2020-15-01', '2020-25-01', '2020-14-02', '2020-15-03',
       '2020-29-03', '2020-15-04', '2020-18-04', '2020-23-04',
       '2020-26-04', '2020-20-05', '2020-23-02', '2020-17-03',
       '2020-14-06', '2020-28-02', '2020-14-04', '2020-29-04',
       '2020-22-05', '2020-28-06'], dtype=object)
```

There are errors in the 'date' column in which there is a change in format that occurs, this can be seen in the last few rows above. The change occurs in the Month and Day column, where the Month is now in the 3rd spot instead of Day and vice-versa. This results in the Month's location displaying numbers above 12. Thus changes needs to be made.

Function to change the format of the dates:

It checks the month's column and changes the format of the date if it's larger than 12. Then the respective date's format will be changed to the form of (DD-MM-YYYY).

In [7]:

```
def change_date_format(date):      #Changes format of date to one format (DD-MM-YYYY)
    date1 = date.split('-')
    slash = ('/')
    if date1[1] > '12':
        date1[1],date1[2] = date1[2],date1[1]
    date1.reverse()
    return slash.join(date1)

df['date'] = df.apply(lambda row: change_date_format(row['date']),axis = 1)  # Applies the
```

For the next few columns of data, I will check through the data to see if there are any negative numbers. If there are any, then they need to be replaced with the correct data. There should be no negative cases as it does not make sense to have less than 0 amount of cases.

Check through 'total_cases':

In [8]:

```
df[df['total_cases'] < 0]   # Checking for any negative numbers, there can't be any negativ
```

Out[8]:

location	date	total_cases	new_cases	total_deaths	new_deaths	gdp_per_capita	population

Check through 'new_cases'

In [9]:

```
df[df['new_cases'] < 0]    # Check to see if there are any negative new cases
```

Out[9]:

	location	date	total_cases	new_cases	total_deaths	new_deaths	gdp_per_capita
549	France	03/06/2020	151325.0	-766.0	28940.0	107.0	38605.671
960	Italy	20/06/2020	238011.0	-148.0	34561.0	47.0	35220.084
1095	Spain	19/04/2020	193252.0	-713.0	20453.0	410.0	34272.360
1131	Spain	25/05/2020	235400.0	-372.0	26834.0	-1918.0	34272.360
1323	United Kingdom	21/05/2020	248293.0	-525.0	35704.0	363.0	39753.244
1366	United Kingdom	03/07/2020	283757.0	-29726.0	43995.0	89.0	39753.244

Since there are negative new_cases, checking needs to be done to compare these numbers with other sources. If the data is suspected to be an outlier, look towards its position in the array and compare it to other data that is above or below it to see how it fares. Then if need be, remove or replace them with corresponding accurate data.

Normally when such outliers appears in such a large data set, it would be better to ignore them due to the vast multitiude of possible errors and the amount of time it would take to cleanse them. Furthermore,these data needs to be treated before any of the graphs can be ploted to prevent a weird pattern. However, in this case since there are only a handful of errors, I personally find it fine to check through and change these data by hand.

In [10]:

```
# citations to the corrected values can be found at the end of the pdf file

df.loc[df['new_cases'] == -766.0, 'new_cases'] = 352.0
df.loc[df['new_cases'] == -148.0, 'new_cases'] = 264.0
df.loc[df['new_cases'] == -713.0, 'new_cases'] = 1313.0
df.loc[df['new_cases'] == -372.0, 'new_cases'] = 475.0
df.loc[df['new_cases'] == -525.0, 'new_cases'] = 2615.0
df.loc[df['new_cases'] == -29726.0, 'new_cases'] = 544.0
df.loc[(df['location'] == 'Iran') &( df['date'] == '04/04/2020'), 'new_cases'] = 2560.0 # F
```

In [11]:

```
df[df['new_cases'] < 0]      # Check to see if changes have been made
```

Out[11]:

location	date	total_cases	new_cases	total_deaths	new_deaths	gdp_per_capita	population

Check through 'total_deaths'

In [12]:

```
df[df['total_deaths'] < 0]
```

Out[12]:

location	date	total_cases	new_cases	total_deaths	new_deaths	gdp_per_capita	population

Check through 'new_deaths'

In [13]:

```
df[df['new_deaths'] < 0]
```

Out[13]:

	location	date	total_cases	new_cases	total_deaths	new_deaths	gdp_per_capita
965	Italy	25/06/2020	239410.0	577.0	34644.0	-31.0	35220.084
1131	Spain	25/05/2020	235400.0	475.0	26834.0	-1918.0	34272.360

In [14]:

```
df.loc[df['new_deaths'] == -31.0,'new_deaths'] = 34.0
df.loc[df['new_deaths'] == -1918.0, 'new_deaths'] = 72.0
df[df['new_deaths'] < 0]     # Check to see if changes have been made
```

Out[14]:

location	date	total_cases	new_cases	total_deaths	new_deaths	gdp_per_capita	population

Check through 'gdp_per_capita'

In [15]:

```
df[df['gdp_per_capita'] < 0]
```

Out[15]:

location	date	total_cases	new_cases	total_deaths	new_deaths	gdp_per_capita	population

Check through 'population'

In [16]:

```
df[df['population'] < 0]
```

Out[16]:

location	date	total_cases	new_cases	total_deaths	new_deaths	gdp_per_capita	population

Second Dataframe

In [17]:

```
df1 = pd.read_csv('CountryLockdowndates.csv')
df1
```

Out[17]:

	Country/Region	Province	Date	Type	Reference
0	Afghanistan	NaN	24/03/2020	Full	https://www.thestatesman.com/world/afghan-govt...
1	Albania	NaN	08/03/2020	Full	https://en.wikipedia.org/wiki/2020_coronavirus...
2	Algeria	NaN	24/03/2020	Full	https://www.garda.com/crisis24/news-alerts/325...
3	Andorra	NaN	16/03/2020	Full	https://en.wikipedia.org/wiki/2020_coronavirus...
4	Angola	NaN	24/03/2020	Full	https://en.wikipedia.org/wiki/2020_coronavirus...
...	...	...	...	...	...
302	Venezuela	NaN	16/03/2020	Full	https://en.wikipedia.org/wiki/2020_coronavirus...
303	Vietnam	NaN	19/03/2020	Full	https://en.wikipedia.org/wiki/2020_coronavirus...
304	West Bank and Gaza	NaN	05/03/2020	Full	NaN
305	Zambia	NaN	NaN	None	NaN
306	Zimbabwe	NaN	27/03/2020	Full	https://en.wikipedia.org/wiki/2020_coronavirus...

307 rows × 5 columns

In [18]:

```
df1['Country/Region'].unique()    # Check to see the names of countries wanted, or if there
```

Out[18]:

```
array(['Afghanistan', 'Albania', 'Algeria', 'Andorra', 'Angola',
       'Antigua and Barbuda', 'Argentina', 'Armenia', 'Australia',
       'Austria', 'Azerbaijan', 'Bahamas', 'Bahrain', 'Bangladesh',
       'Barbados', 'Belarus', 'Belgium', 'Belize', 'Benin', 'Bhutan',
       'Bolivia', 'Bosnia and Herzegovina', 'Botswana', 'Brazil',
       'Brunei', 'Bulgaria', 'Burkina Faso', 'Burma', 'Cabo Verde',
       'Cambodia', 'Cameroon', 'Canada', 'Central African Republic',
       'Chad', 'Chile', 'China', 'Colombia', 'Congo (Brazzaville)',
       'Congo (Kinshasa)', 'Costa Rica', 'Cote dIvoire', 'Croatia',
       'Cuba', 'Cyprus', 'Czechia', 'Denmark', 'Diamond Princess',
       'Djibouti', 'Dominica', 'Dominican Republic', 'Ecuador', 'Egypt',
       'El Salvador', 'Equatorial Guinea', 'Eritrea', 'Estonia',
       'Eswatini', 'Ethiopia', 'Fiji', 'Finland', 'France', 'Gabon',
       'Gambia', 'Georgia', 'Germany', 'Ghana', 'Greece', 'Greenland',
       'Grenada', 'Guatemala', 'Guinea', 'Guinea-Bissau', 'Guyana',
       'Haiti', 'Honduras', 'Hungary', 'Iceland', 'India', 'Indonesia',
       'Iran', 'Iraq', 'Ireland', 'Israel', 'Italy', 'Jamaica', 'Japan',
       'Jersey', 'Jordan', 'Kazakhstan', 'Kenya', 'Korea, South',
       'Kosovo', 'Kuwait', 'Kyrgyzstan', 'Laos', 'Latvia', 'Lebanon',
       'Liberia', 'Libya', 'Liechtenstein', 'Lithuania', 'Luxembourg',
       'Madagascar', 'Malaysia', 'Maldives', 'Mali', 'Malta',
       'Mauritania', 'Mauritius', 'Mexico', 'Moldova', 'Monaco',
       'Mongolia', 'Montenegro', 'Morocco', 'Mozambique', 'MS Zaandam',
       'Namibia', 'Nepal', 'Netherlands', 'New Zealand', 'Nicaragua',
       'Niger', 'Nigeria', 'North Macedonia', 'Norway', 'Oman',
       'Pakistan', 'Palestine', 'Panama', 'Papua New Guinea', 'Paraguay',
       'Peru', 'Philippines', 'Poland', 'Portugal', 'Qatar', 'Romania',
       'Russia', 'Rwanda', 'Saint Kitts and Nevis', 'Saint Lucia',
       'Saint Vincent and the Grenadines', 'San Marino', 'Saudi Arabia',
       'Senegal', 'Serbia', 'Seychelles', 'Sierra Leone', 'Singapore',
       'Slovakia', 'Slovenia', 'Somalia', 'South Africa', 'Spain',
       'Sri Lanka', 'Sudan', 'Suriname', 'Sweden', 'Switzerland', 'Syria',
       'Taiwan*', 'Tanzania', 'Thailand', 'Timor-Leste', 'Togo',
       'Trinidad and Tobago', 'Tunisia', 'Turkey', 'Uganda', 'Ukraine',
       'United Arab Emirates', 'United Kingdom', 'Uruguay', 'US',
       'Uzbekistan', 'Vatican City', 'Venezuela', 'Vietnam',
       'West Bank and Gaza', 'Zambia', 'Zimbabwe'], dtype=object)
```

Function to find the Earliest Instance of Lockdown occuring in a Country:

In [19]:

```
def find_min_date(Country : str):
    Country = df1[df1['Country/Region'] == Country]
    Country = Country[pd.notnull(Country['Date'])]
    Country = Country.reset_index()
    Country = Country.drop('index',axis = 1)
    minimum = (Country.at[0, 'Date']).split('/')
    for idx, row in Country.iterrows():
        item = row[2].split('/')
        for x in range(len(item)-1,-1,-1):
            if item[x] > minimum[x]:
                break
            if item[x] < minimum[x]:
                minimum = item
    slash = '/'
    return slash.join(minimum)
```

In [20]:

```
# Find the 1st lockdown dates of respective countries wanted.
Australia = (find_min_date('Australia'))
China = (find_min_date('China'))
France = (find_min_date('France'))
Iran = (find_min_date('Iran'))
Italy = (find_min_date('Italy'))
Spain = (find_min_date('Spain'))
UK = (find_min_date('United Kingdom'))
US = (find_min_date('US'))
```

In [21]:

```
# Create new dataframe that contains the lockdown dates of the wanted Countries

dates = pd.DataFrame({

'location'  : ['Australia','China','France', 'Iran', 'Italy', 'Spain', 'United Kingdom', 'U
'lockdown_date' : [Australia, China, France, Iran, Italy, Spain, UK, US]
})

dates
```

Out[21]:

	location	lockdown_date
0	Australia	24/03/2020
1	China	23/01/2020
2	France	16/03/2020
3	Iran	15/03/2020
4	Italy	11/03/2020
5	Spain	14/03/2020
6	United Kingdom	18/03/2020
7	United States	13/03/2020

Forming new Dataframe

In [22]:

```
# Merge both data frames based on the same location to create a new one
result = pd.merge(df, dates, on=['location'])
result
```

Out[22]:

	location	date	total_cases	new_cases	total_deaths	new_deaths	gdp_per_capita
0	Australia	31/12/2019	0.0	0.0	0.0	0.0	44648.710
1	Australia	01/01/2020	0.0	0.0	0.0	0.0	44648.710
2	Australia	02/01/2020	0.0	0.0	0.0	0.0	44648.710
3	Australia	03/01/2020	0.0	0.0	0.0	0.0	44648.710
4	Australia	04/01/2020	0.0	0.0	0.0	0.0	44648.710
...	...	...	...	...	...	...	...
1570	United States	10/07/2020	3118008.0	63004.0	133291.0	982.0	54225.446
1571	United States	11/07/2020	3184633.0	66625.0	134097.0	806.0	54225.446
1572	United States	12/07/2020	3247684.0	63051.0	134814.0	717.0	54225.446
1573	United States	13/07/2020	3304942.0	57258.0	135205.0	391.0	54225.446
1574	United States	14/07/2020	3363056.0	58114.0	135605.0	400.0	54225.446

1575 rows × 9 columns

In [23]:

```
result.to_csv('31902626_Task1DataSet.csv',index=False)
```

Data Visualisation

Creating Line Charts to display Covid-19 daily Infections

Creation line charts to show the trend of the daily number of new cases for each country and exploring the result of visualisation

Plot Graph Function:

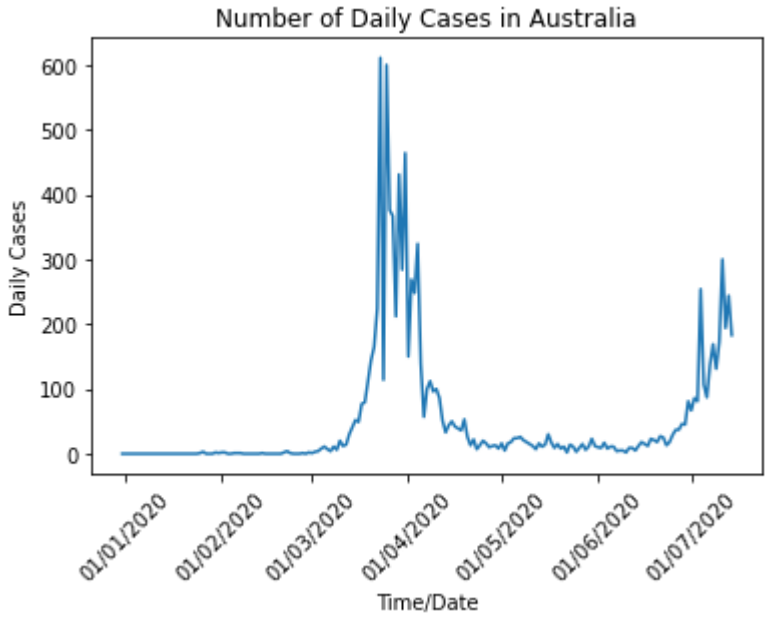
Plot a line graph based on a country given and an optional second input parameter if the lockdown date for the country is desired.

In [24]:

```
def plt_line1(location : str, show_lockdown_date = False):
    Country = result[result['location'] == location ]
    plt.plot(Country.date, Country.new_cases)
    plt.ylabel('Daily Cases')
    plt.xlabel('Time/Date')
    plt.title('Number of Daily Cases in ' + location)
    if show_lockdown_date:
        line = Country.iloc[0]['lockdown_date']
        plt.axvline(line, color = 'red')
    plt.xticks([])
    plt.xticks(['01/01/2020','01/02/2020' , '01/03/2020', '01/04/2020' , '01/05/2020', '01/
    plt.show()
    return
```

In [25]:

```
plt_line1('Australia')
```



Australia:

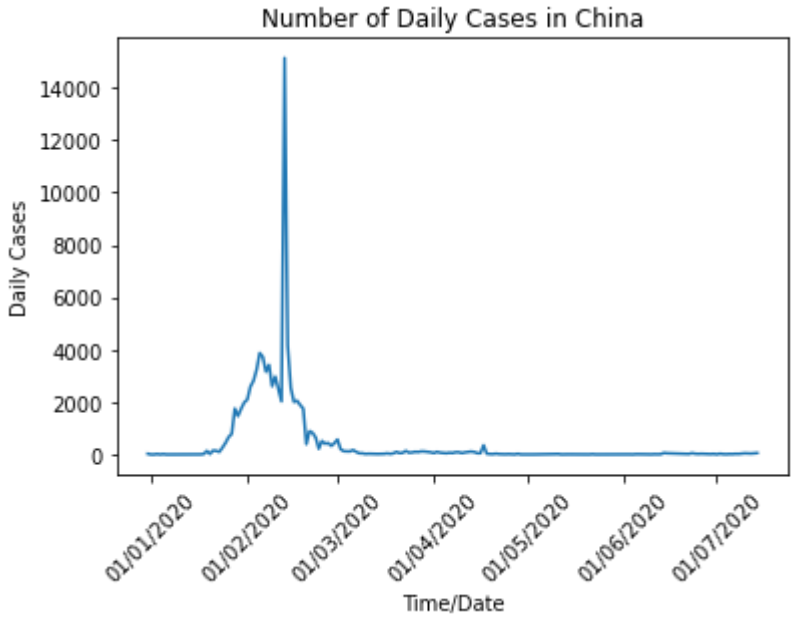
Looking at the graph above, it can be seen that the earliest outbreak of COVID19 started during the middle of March. It starts to increase exponentially to a peak of 600 daily cases soon after in late March. One of the possible reasons why the number of cases started to increase dramatically might be due to the fact that before the discovery of infected people, they were travelling around in public and spreading their infection to non-infected people. Furthermore, another reason as to why the number of cases increased dramatically might be due to the Government making people undergo mass testing for COVID19, thus for the few following weeks the number of reported daily cases exploded.

After the peak, the number of cases started to decrease slowly over the course of half a month due to a multitude of possibe infected people still getting tested. and the number of cases started to plataeu for the next few months until July. There was another insurgence of cases at the start of July. One of the reasons as to why

it began flaring up is as the Victoria's government stated; "many people are doing the wrong thing- such as going to work while sick" *[Mao, 2020, Are people breaking the rules?]* thus passing the infection to other people particularly in crowded areas.

In [26]:

```
plt_line1('China')
```



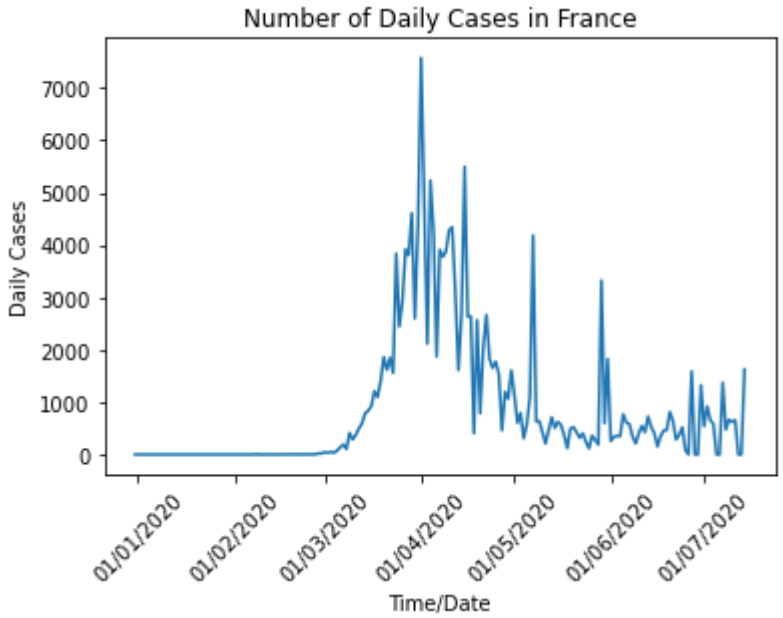
China:

China's first reported case of COVID19 appeared in the middle of January, the first of many countries to be hit with it. The number of daily cases across the board reaches much higher values compared to other Countries due to the much larger population that China has compared to other Countries. In Feburary, there is a slight dip in reported cases in China and was soon followed by a massive spike which peaked at 15000 cases in one day. One reason this might have occured may be due to the fact that some locations might not have reported those cases for the past few days due to possible technical errors, resulting in the slight dip in cases at the start of February, thus causing an astronomical amount in in the middle of February. Thus this particular day may possibly contain the cumulation of cases from the last few days.

Following after that, China's number of daily cases immediately dropped shortly after the peak and maintains this for the following months.

In [27]:

```
plt_line1('France')
```

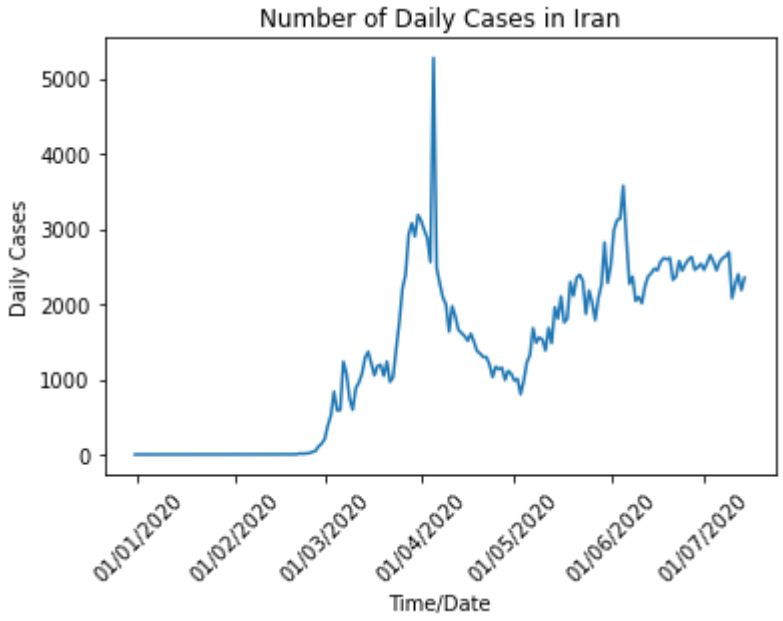


France

France's first reported case was from the start of March and increased steeply briefly afterwards reaching a peak at around 8000 cases at the start of April. Following suitedly, the number of cases started to decrease slowly across the months. However from time to time, there were occasional spikes and fluctuations that occured across the months and these continued on until the latest available data. There seems to be an occasional massive spike which happens once every month. This might be due to a cluster of infections being found and the health officers would begin mass testing, causing this spike in cases ocassionally.

In [28]:

```
plt_line1('Iran')
```

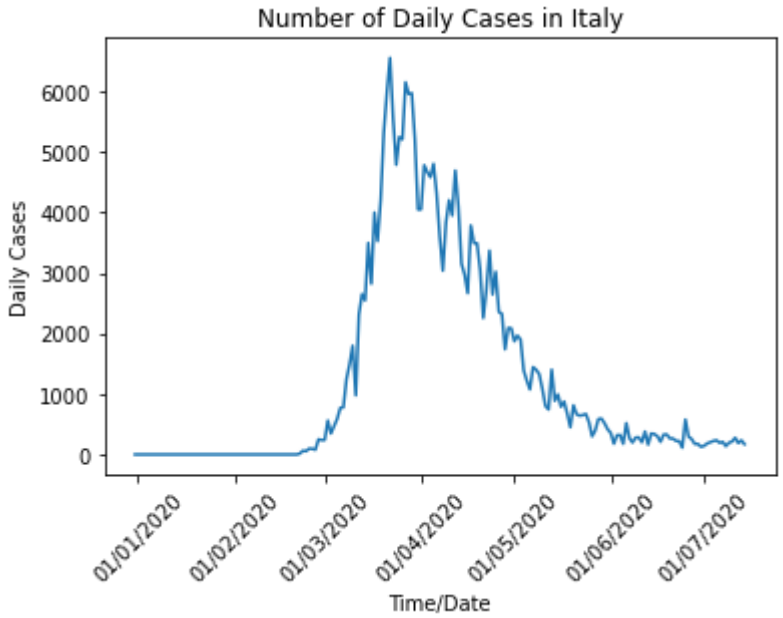


Iran

Iran had their first reported case at a similar time to the other countries at the beginning of March. The number of cases proceeds to peak at 5000 cases in April and drops immediately thereafter. It continues to dip until the start of May, where cases started to increase again. According to Authorities of Iran the reason for the sudden flare in cases is due to the increased number of testing and extending their reach to test for asymptomatic cases. Thus, effectively causing the increase in number of cases. *[Ali, 2020, Why cases have been rising?]*. This is helpful to the mass as it results in less amount of infected people walking around and helps to reduce the chances of people getting infected in the long term.

In [29]:

```
plt_line1('Italy')
```

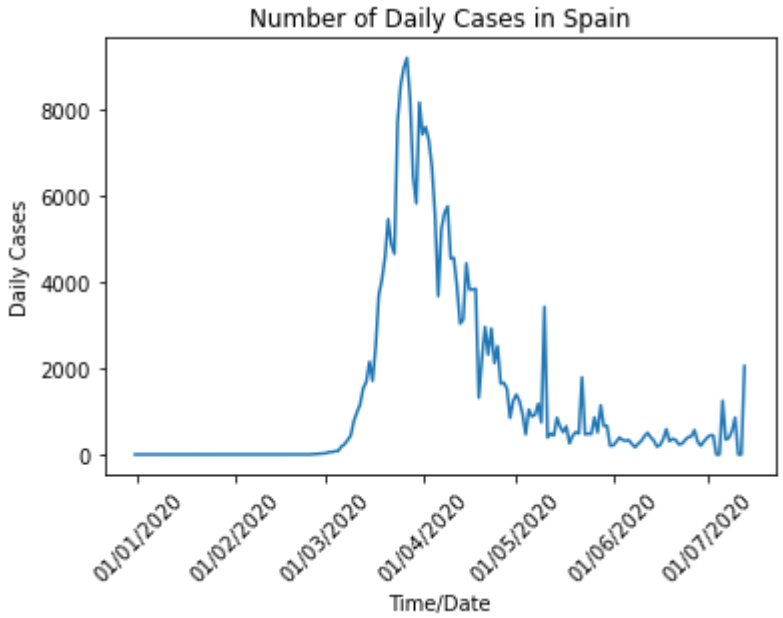


Italy

Italy experienced its first reported case a bit earlier than most other countries at the end of February. Following on from the first, the number of daily cases increased steeply reaching its highest peak in the middle of March at slightly over 6000 cases. Following on there was generally a constant decrease for the next 3 months until the start of June where the number of cases started to plateau out. During the decrease, there were constant fluctuations in the number of cases. This might be a result of either Italy increasing the amount of testing done on certain days of the week.

In [30]:

```
plt_line1('Spain')
```

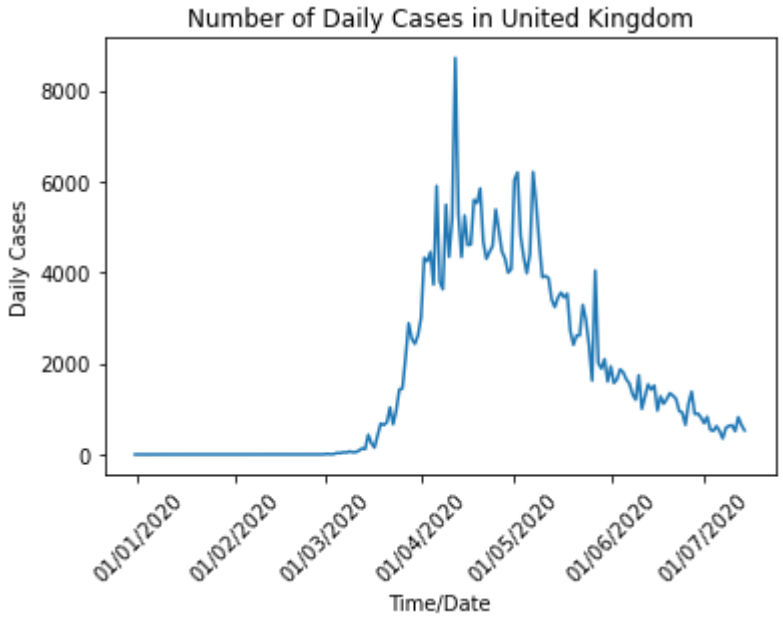


Spain

Spain's first reported case begun in the middle of March and increased exponentially reaching 9000 caes at the start of April. Shortly, it experiences a steady decline in number of cases with repetitive fluctuations that occurs, due to possible delay in official reports being made. There were occasional spikes that occured, one such time at the early part of May where it spiked up to nearly 4000 cases. This might be possibly due to a cluster that was infected being found and rapid mass testing was done in the vicinity.

In [31]:

```
plt_line1('United Kingdom')
```

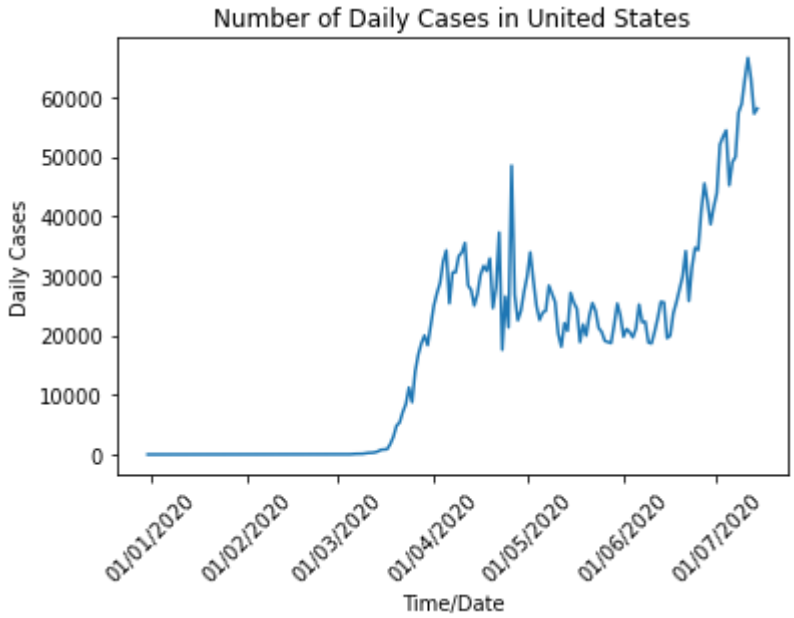


United Kingdom

The first reported case began in mid March climbing steeply up, due to a large increase in testing, to a peak at slightly more than 8000 cases in mid April. It starts to plateau with repetitive fluctuations occuring for the entirety of April. Then a decrease in the number of cases started in the middle of May. Up until the recent data that is available, there is still a few hundred cases being reported daily and will presumably decrease as time goes on.

In [32]:

```
plt_line1('United States')
```



United States

The United States first started their first reported case in late March, much later than most countries. This is due to lack of testing being done, thus resulting in no reported cases until then. Shortly afterwards the number of cases started to increase exponentially until early April, where the number of cases started to flatline with a side of fluctuations occuring. There was a sudden spike in number of cases in late April reaching at 50000 cases. However instead of decreasing in the number of cases afterwards, like the other countries, the number of cases started to increase more towards the beginning of June. This is contributed by many factors, one of which is the sudden increase in testing cases. Another factor is in the case where a lot clusters were beginning to be found by the health officers.

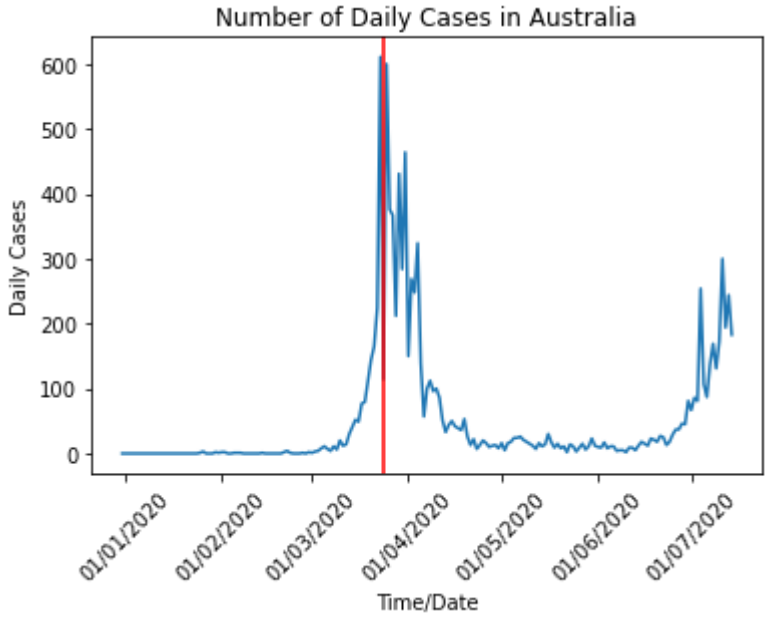
Out of all countries that were graphed, the United States has shown to have had the 'best' experience with the virus. Experiencing a staggering record at 60000 reported cases in one day and most likely to be beaten in the time to come ahead of it. This is partially due to the fact that the population of the United States is much larger than countries like Italy. However comparing the United States to China, where population is on a similar scale, they are doing much worse than their asian counterpart. This might be due to a disarray and lack of response from the Government causing countless to suffer from the pandemic.

Addition of Vertical Line for Lockdown Date

Exploring if lockdown affects the trend that was shown in the plots. Furthermore, to see if effects are similar to all countries.

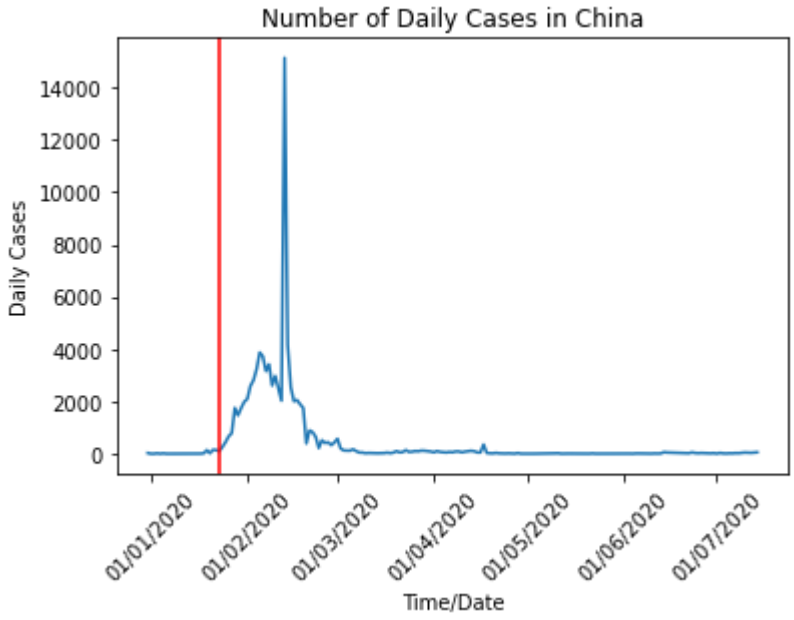
In [33]:

```
plt_line1('Australia', True)
```



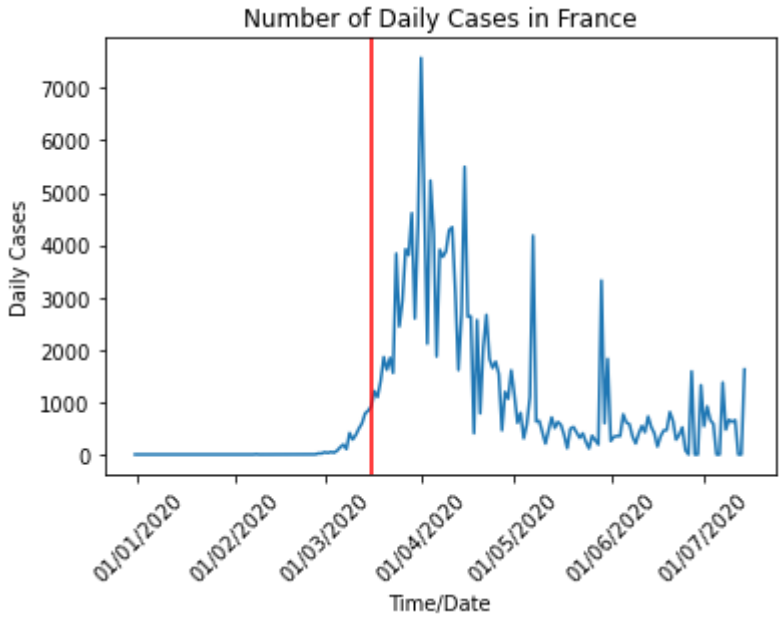
In [34]:

```
plt_line1('China',True)
```



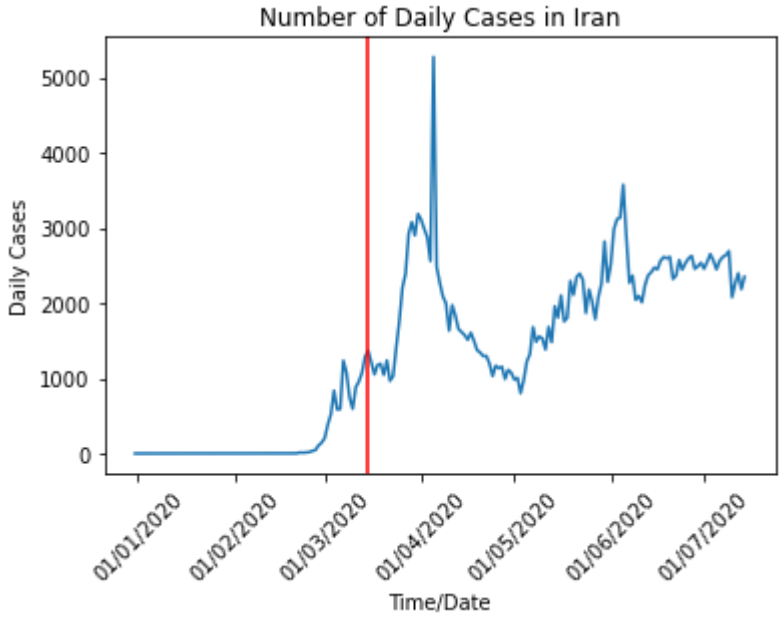
In [35]:

```
plt_line1('France', True)
```



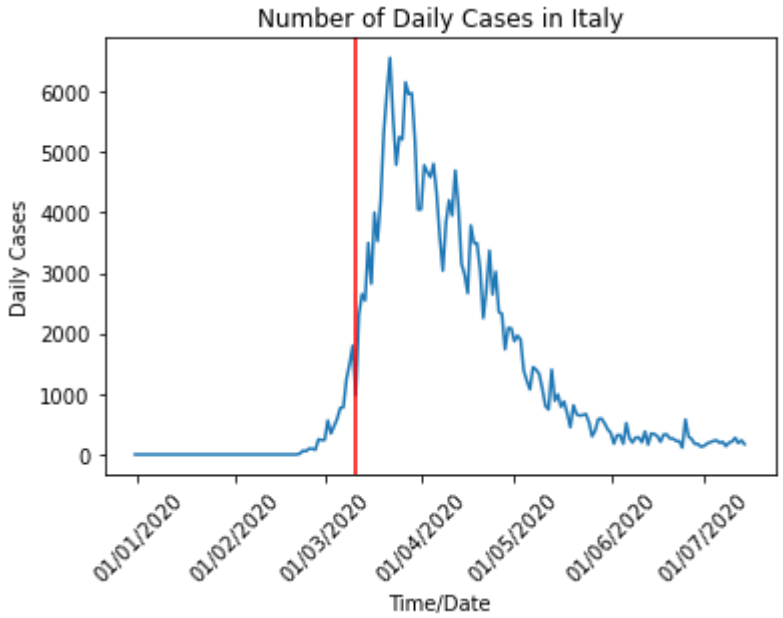
In [36]:

```
plt_line1('Iran', True)
```



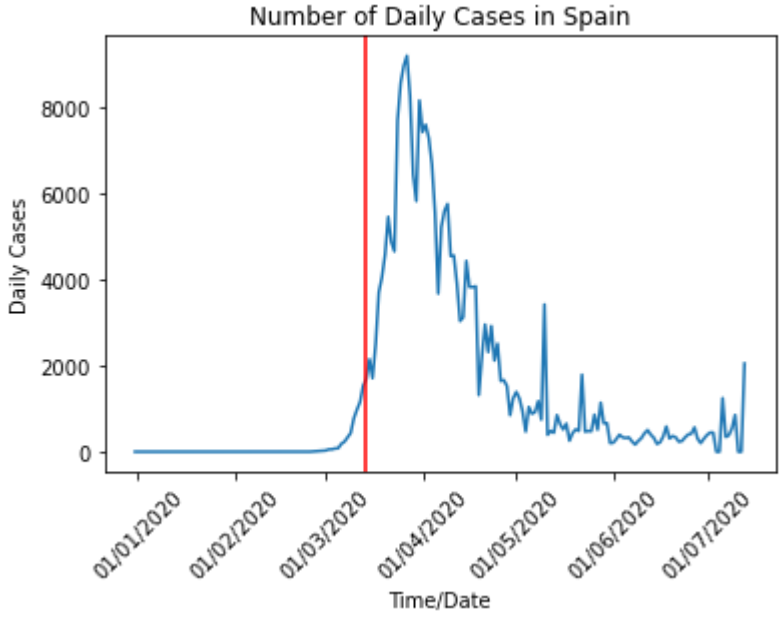
In [37]:

```
plt_line1('Italy', True)
```



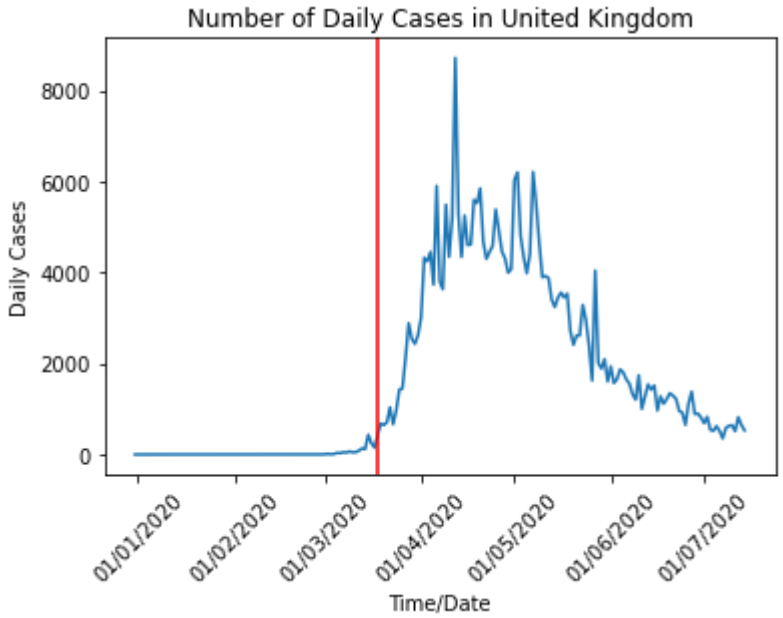
In [38]:

```
plt_line1('Spain', True)
```



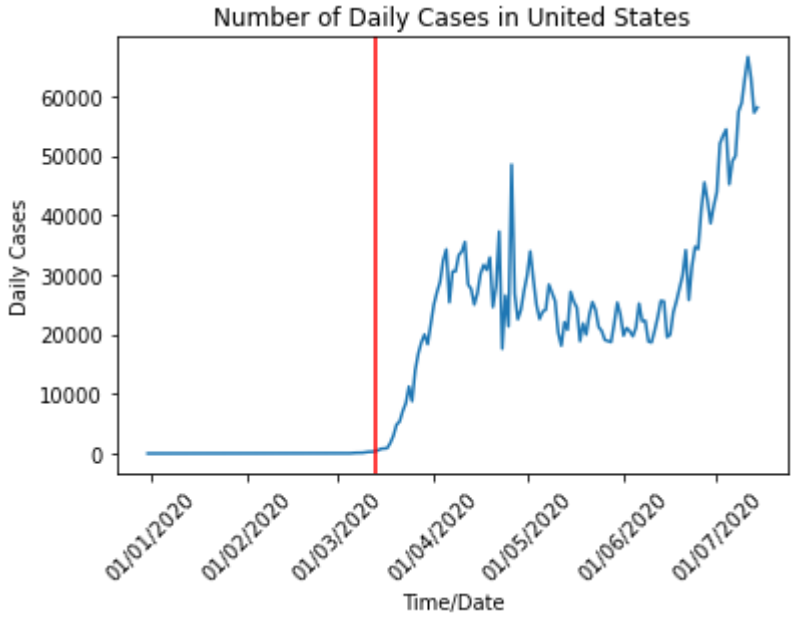
In [39]:

```
plt_line1('United Kingdom', True)
```



In [40]:

```
plt_line1('United States', True)
```



Analysis:

Looking at all the above graphs with their lockdown date inplace, we can see a general trend among most countries. Past the lockdown date, there is generally an increase in reported cases for another few weeks and starts to decrease steeply shortly afterwards. The reason behind why the number of cases still increase is due to the alertness of a country and the beginning of more rigorous testing for infected people. As more tests are done, more cases will be found. Thus, it is to be expected to have a steep increase in number of daily cases for a few weeks before experiencing a drop in cases. Now one of the underlying reasons as to why cases will start to experience a drop is due to the lockdown, where there are travel bans and plenty of restrictions being put on everyone. Restrictions like mandatory curfew and shutting down of all shops except necessities help to decrease the amount of people travelling around. This helps to decrease potential infected people from being in crowds and passing on the virus. With this in place, cases in these countries should expect a sharp decrease if they are to follow the lockdown procedures strictly and maintain proper hygiene.

Australia deviates from the trend slightly compared to the other countries. When their lockdown was in place, they experienced an immediate drop in cases compared to other countries with no apparent buffer period. This might be due to the fact that they could manage zones really well and had already begun testing earlier for cases and begun isolating zones that were infected, preventing an increase in cases. The lockdown further helped them in this regard and prevented any extra infections.

However there are some countries where this trend does not apply to them at all. From the above graphs, Iran and the United States. Instead of decreasing the number of cases to a minimal amount, it started to increase again not long after even through lockdown. There are plenty of reasons for the increase in cases. For Iran, as mentioned before, it is due to an increased amount of testing for asymptomatic cases. For the United States, it was the opposite. During the begining of their lockdown, there was not much active support from the government to begin testing. Resulting in the United States being in 'six weeks of complete blindness to the pandemic' *[Sullivan, 2020]*. In these several weeks, numerous numbers of infected people have passed the virus to countless of other lives. In conclusion, causing a still-ever-so increasing number of daily cases until the latest available data. Thus the lockdown wasn't as effective as it should have been.

Question 3.

Explore whether there is a relation between daily new case/death rate and the GDP of a country. To this aim, you need to calculate:

- The average of GDP of the countries, and then divide the countries into two groups, a group which its GDP is above the average GDP, and another group which its GDP is below the average GDP. We call the former group as "AboveGDP" and the later as "BelowGDP" from now onwards.
- The daily new cases rate (new cases divided by population) for each country
- The daily new death rate (new deaths divided by population) for each country

Then, you need to create two line charts, one which shows the new case rate of groups "AboveGDP" and "BelowGDP"; and, another line chart to show the death rate of the two groups ("AboveGDP" and "BelowGDP").

```
a) Which group ("AboveGDP" or "BelowGDP") usually had higher values of case rate?
b) Which group ("AboveGDP" or "BelowGDP”) usually had higher values of the death r
ate?
c) We would have expected that the case rate and death rate of group "AboveGDP" wi
ll be lower than group "BelowGDP". Does       the result of your visualisation is
 the same as the mentioned expectation? If no, why do you think the expectation is
different from the reality?
```

In [41]:

```
# Creating two new columns in result dataframe
result['daily_case_rate'] = (result['new_cases'] / result['population']) * 100
result['daily_death_rate'] = (result['new_deaths'] / result['population']) * 100
```

In [42]:

```
# Seperating the countries into 2 categories, AboveGDP and BelowGDP
average = result.gdp_per_capita.mean()
AboveGDP = result[result.gdp_per_capita > average]
BelowGDP = result[result.gdp_per_capita < average]
```

In [43]:

```
# A change in date format had to be done, as previously it would sort the dates in lexograp

def change_format(date):   # Change date format to YYYY-MM-DD from DD-MM-YYYY
    dash = '-'
    line = date.split('/')
    line.reverse()
    return dash.join(line)

AboveGDP['date'] = AboveGDP.apply(lambda row: change_format(row['date']),axis = 1 )
BelowGDP['date'] = BelowGDP.apply(lambda row: change_format(row['date']),axis = 1 )
```

```
<ipython-input-43-d5b86dd0f74b>:9: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pand
as.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-v
ersus-a-copy)
  AboveGDP['date'] = AboveGDP.apply(lambda row: change_format(row['date']),a
xis = 1 )
<ipython-input-43-d5b86dd0f74b>:10: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pand
as.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-v
ersus-a-copy)
  BelowGDP['date'] = BelowGDP.apply(lambda row: change_format(row['date']),a
xis = 1 )
```

Function to plot an Average Graph for Daily Case Rate:

In [44]:

```
def plot_avrg_case_rate(AboveGDP, BelowGDP):
    plt.plot(AboveGDP.date, AboveGDP.daily_case_rate, label = "AboveGDP")
    plt.plot(BelowGDP.date, BelowGDP.daily_case_rate, label = "BelowGDP")
    plt.ylabel('Daily Case Rate(%)')
    plt.xlabel('Time/Date')
    plt.title('Average Daily Case Rate')
    plt.legend()
    plt.xticks(['2020-01-01','2020-02-01','2020-03-01','2020-04-01','2020-05-01','2020-06-0
    plt.show()
    return
```

In [45]:

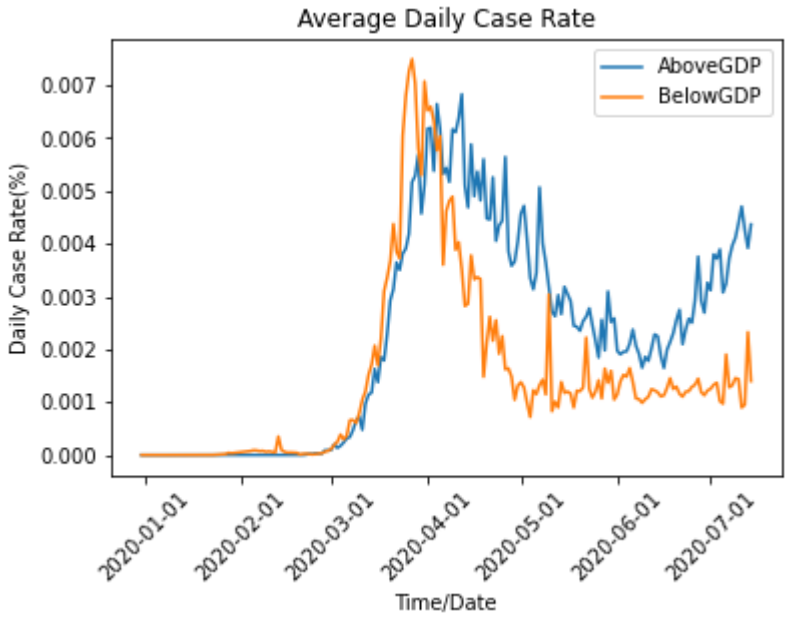
```
# Find the average of all columns that have the same date for AboveGDP
AboveGDPaverage = AboveGDP.groupby('date').mean()
AboveGDPaverage.reset_index(inplace = True)     # Return back the date column as the 1st co
```

In [46]:

```
# Find the average of all columns that have the same date for BelowGDP
BelowGDPaverage = BelowGDP.groupby('date').mean()
BelowGDPaverage.reset_index(inplace = True)     # Return back the date column as the 1st co
```

In [47]:

```
plot_avrg_case_rate(AboveGDPaverage, BelowGDPaverage)
```



3 a)

By looking at the graph, the average daily case rate for the AboveGDP group was mostly higher than the BelowGDP group most of the time, especially after April. However before April, it can be seen that the BelowGDP group was slightly higher than the AboveGDP ever so slightly. However the difference between the two groups was miniscule before April, compared to after April where the difference was extremely obvious at times where the AboveGDP reached double of the BelowGDP.

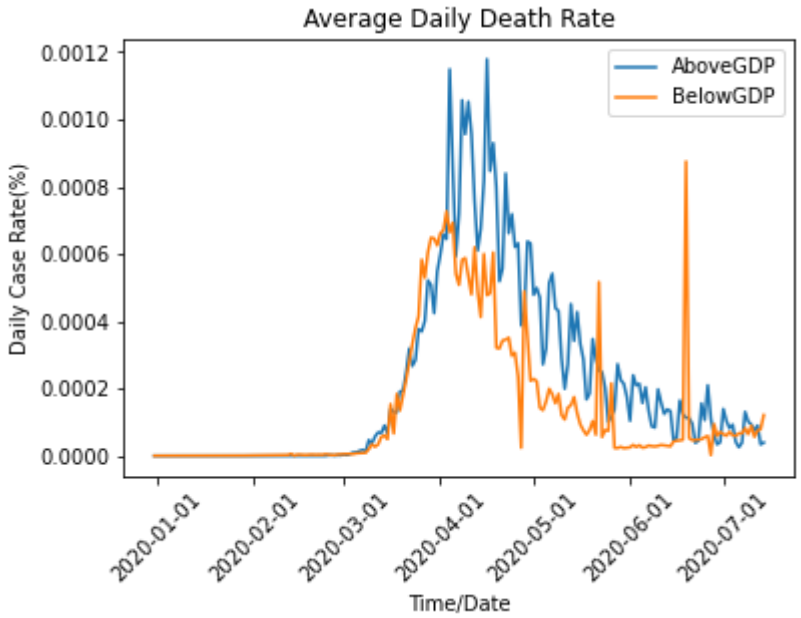
Function to create an Average Death Rate Graph

In [48]:

```
def plot_avrg_death_rate(AboveGDP, BelowGDP):
    plt.plot(AboveGDP.date, AboveGDP.daily_death_rate, label = "AboveGDP")
    plt.plot(BelowGDP.date, BelowGDP.daily_death_rate, label = "BelowGDP")
    plt.ylabel('Daily Case Rate(%)')
    plt.xlabel('Time/Date')
    plt.title('Average Daily Death Rate')
    plt.legend()
    plt.xticks([])
    plt.xticks(['2020-01-01','2020-02-01','2020-03-01','2020-04-01','2020-05-01','2020-06-0
    plt.show()
    return
```

In [49]:

```
plot_avrg_death_rate(AboveGDPaverage, BelowGDPaverage)
```



3 b)

Showing a similar pattern to the average daily case rate, the graph shows that the AboveGDP group is mostly higher than the BelowGDP group. However there are some spikes occasionaly from BelowGDP that overtakes AboveGDP. This might be due to the fact that as the data is collected as an average of a group of countries, the number of countries in a group will affect the data accordingly. The BelowGDP group only had 3 countries as opposed to the 5 countries in AboveGDP. Thus when one of the three countries in the BelowGDP experiences a sudden spike in death cases, it will affect the data much more compared to the AboveGDP group.

3 c)

No it is not the same as expectation. One of the main reasons why this might be so is that in higher GDP countries, they are generally more urbanised compared to countries that have lower GDP. This is important to note as countries that are more urbanised, more developed, tend to have more clustered spots of civilisation due to large amounts of tourism spots, trading ports and more. This creates an area dense with population

where infection can run rampart easily admist unsuspecting passerbys, whereas in less developed countries there would be less urban areas or less dense clusters thus resulting in less of an impact compared to the more developed countries.

One further possible note is that countries with higher GDP have better transportation systems that is readily made available within their areas, like trains. Therefore in the case where there is a infected person on a transportation system, it will cause a branching effect where the infection is now being passed on to many different locations at one go. This causes a potential rise in number of daily cases and death cases in the long term.

References:

Task 1:

Question 1:

Worldometer.(2020). Coronavirus cases in France. Retrieved from https://www.worldometers.info/coronavirus/country/france/ (https://www.worldometers.info/coronavirus/country/france/)

Worldometer.(2020). Coronavirus cases in Italy. Retrieved from https://www.worldometers.info/coronavirus/country/italy/ (https://www.worldometers.info/coronavirus/country/italy/)

Worldometer.(2020). Coronavirus cases in Spain. Retrieved from https://www.worldometers.info/coronavirus/country/spain/ (https://www.worldometers.info/coronavirus/country/spain/)

GOV.UK. (2020). Cases in United Kingdom. Retrieved from https://coronavirus.data.gov.uk/cases (https://coronavirus.data.gov.uk/cases)

Task 2:

Question 1:

Mao, F. (2020, July 31). Coronavirus: Why is Melbourne seeing more cases?. *BBC*. Retrieved from https://www.bbc.com/news/world-australia-53604751 (https://www.bbc.com/news/world-australia-53604751)

Ali, Z. (2020, August 20). Coronavirus: How Iran is battling a surge in cases. *BBC*. Retrieved from https://www.bbc.com/news/52959756 (https://www.bbc.com/news/52959756)

Question 2:

Sullivan, P. (2020, May 21). Why the US have the most reported coronavirus cases in the world. *The Hill*. Retrieved from https://thehill.com/policy/healthcare/498876-why-the-us-has-the-most-reported-coronavirus-cases-in-the-world (https://thehill.com/policy/healthcare/498876-why-the-us-has-the-most-reported-coronavirus-cases-in-the-world)