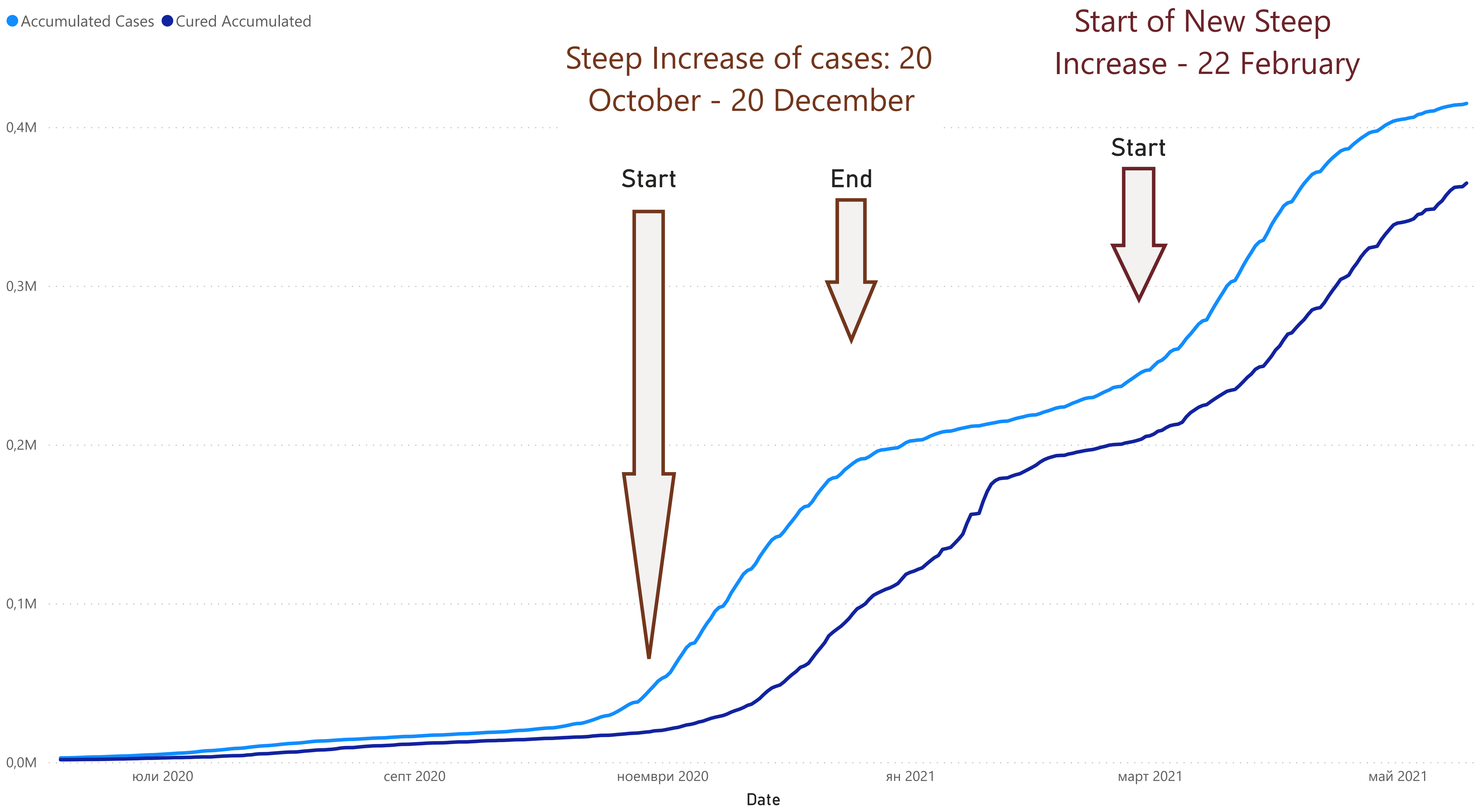
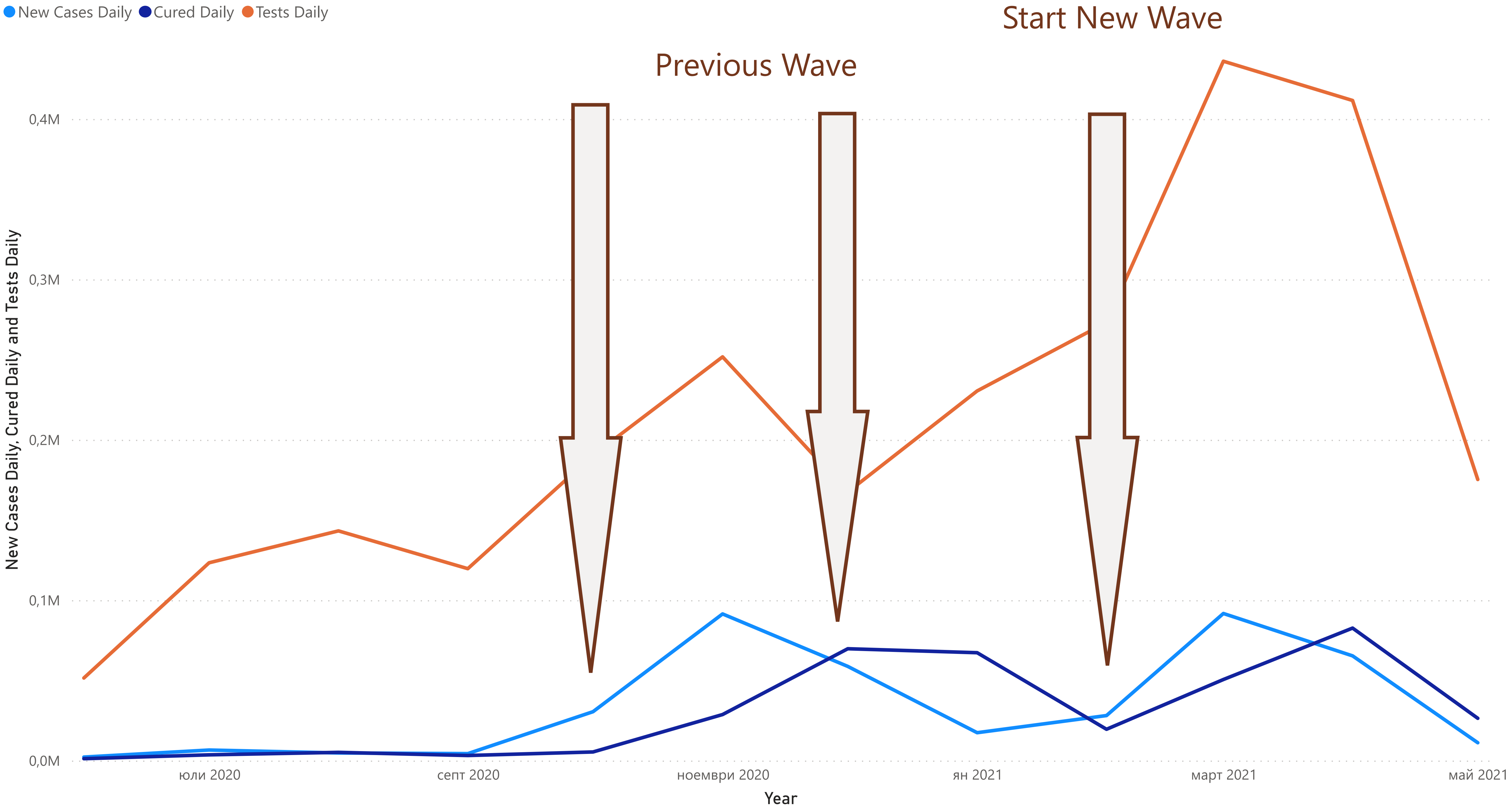
Accumulated Cases, Cured Accumulated and Active Cases



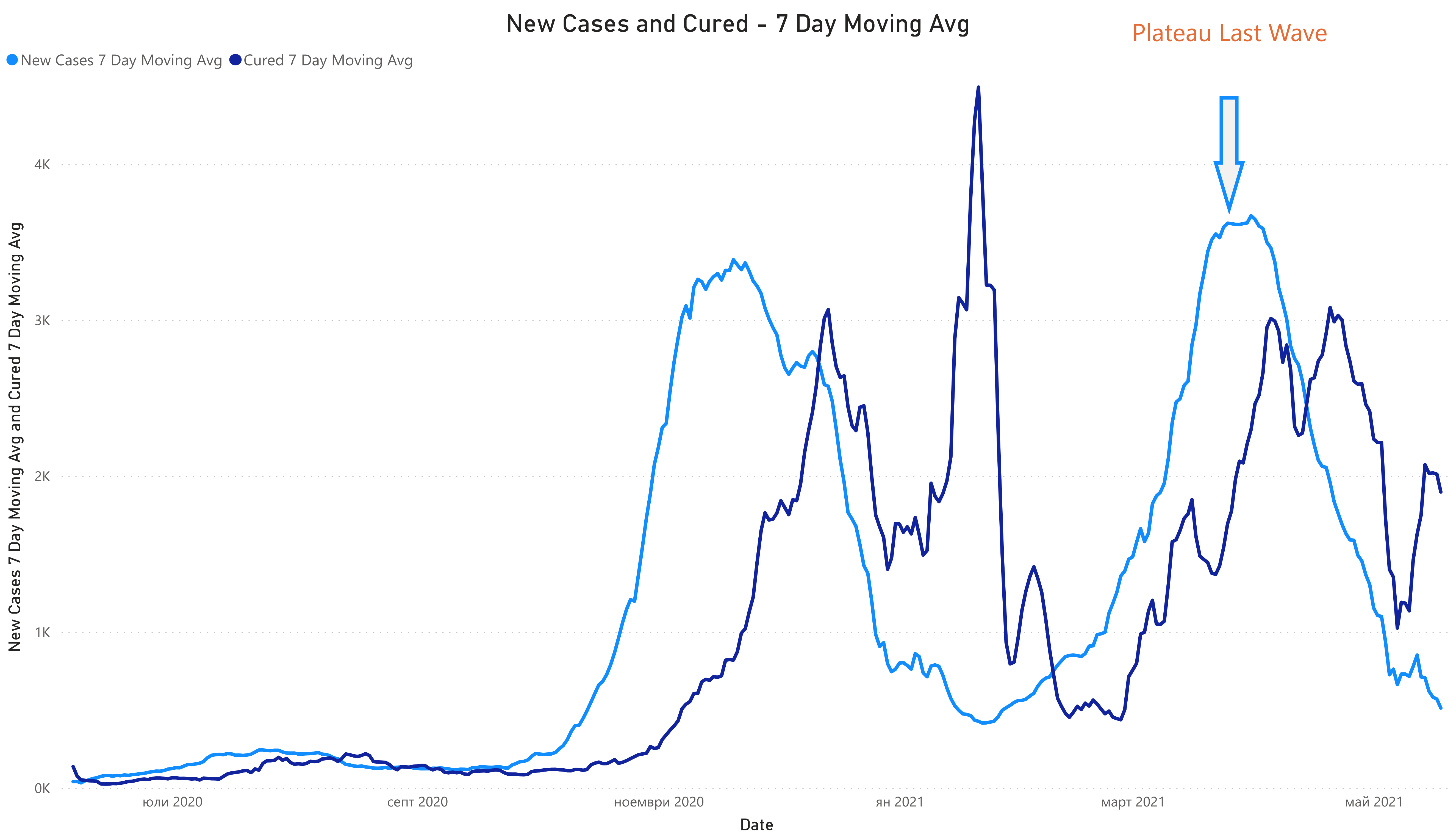
Steep Increase of cases: 20 October - 20 December

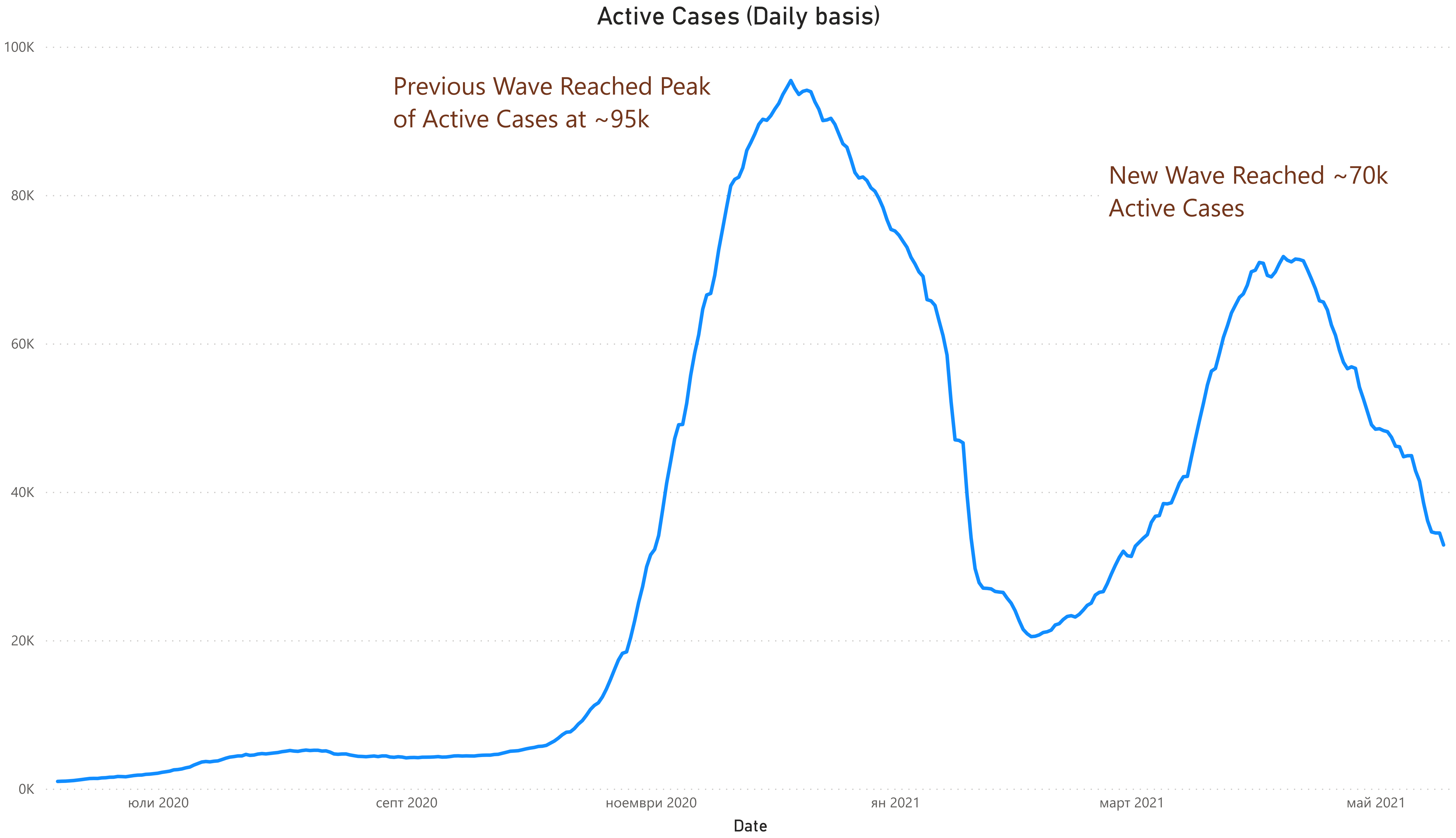
Start of New Steep Increase - 22 February

Tests, New Cases and Cured (Monthly basis)


Start New Wave

Previous Wave

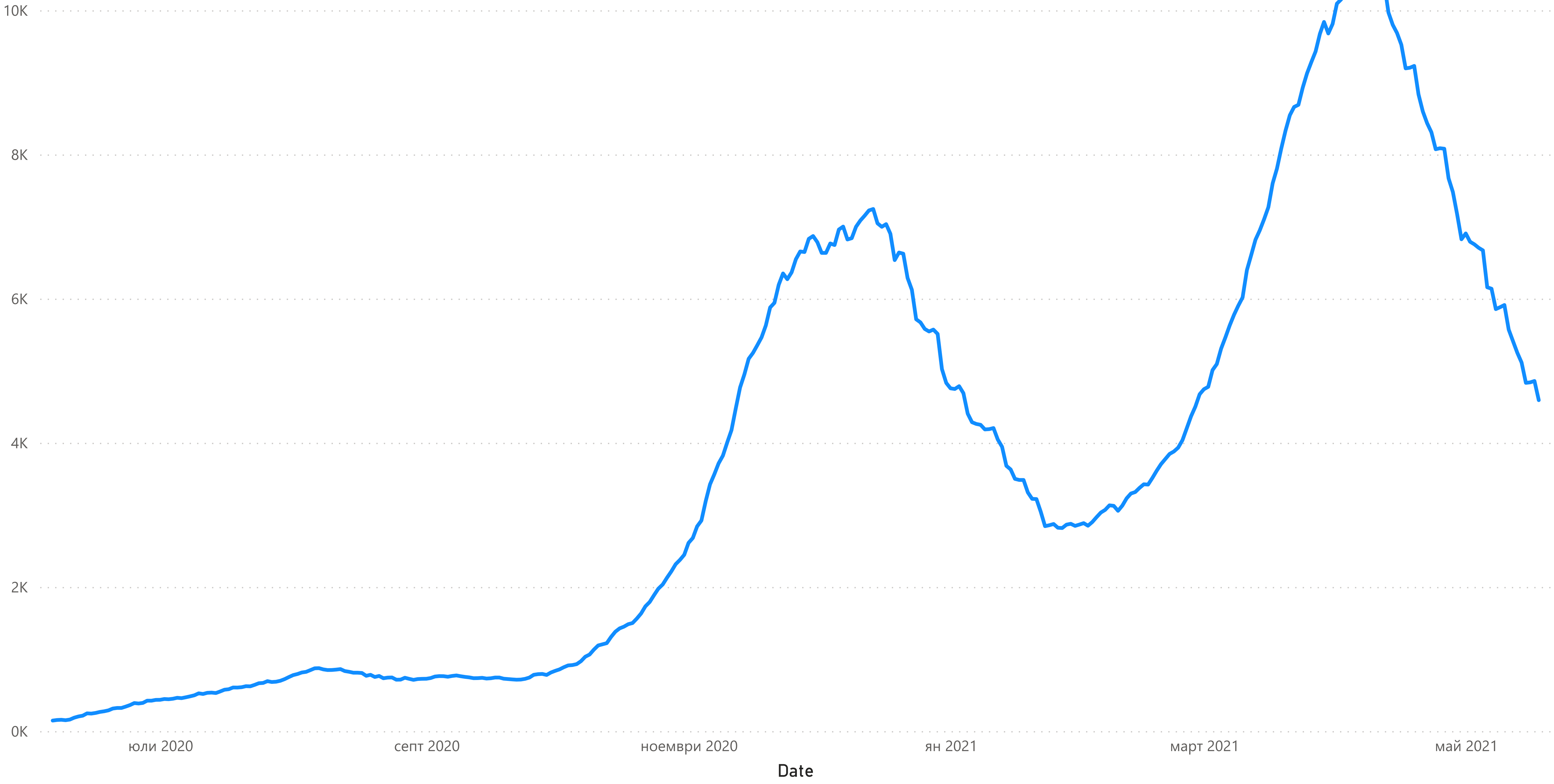

New Cases and Cured - 7 Day Moving Avg
Plateau Last Wave
● New Cases 7 Day Moving Avg
● Cured 7 Day Moving Avg
New Cases 7 Day Moving Avg and Cured 7 Day Moving Avg
4K
3K
2K
1K
0K
юли 2020
септ 2020
ноември 2020
ян 2021
март 2021
май 2021
Date



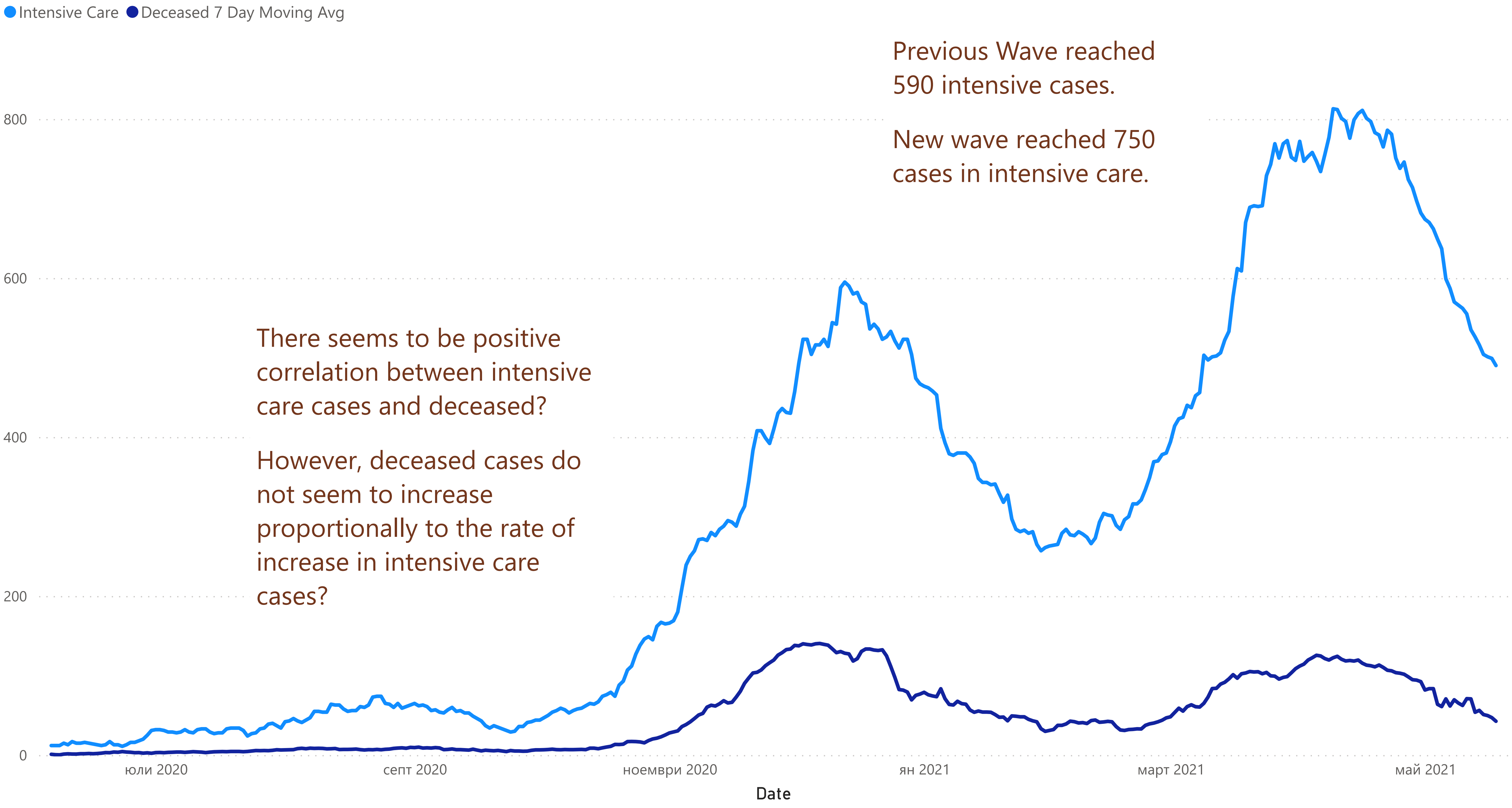
Previous Wave Reached Peak of Active Cases at ~95k

New Wave Reached ~70k Active Cases

In Hospital (Daily basis)



Intensive Care and Deceased (Daily basis)



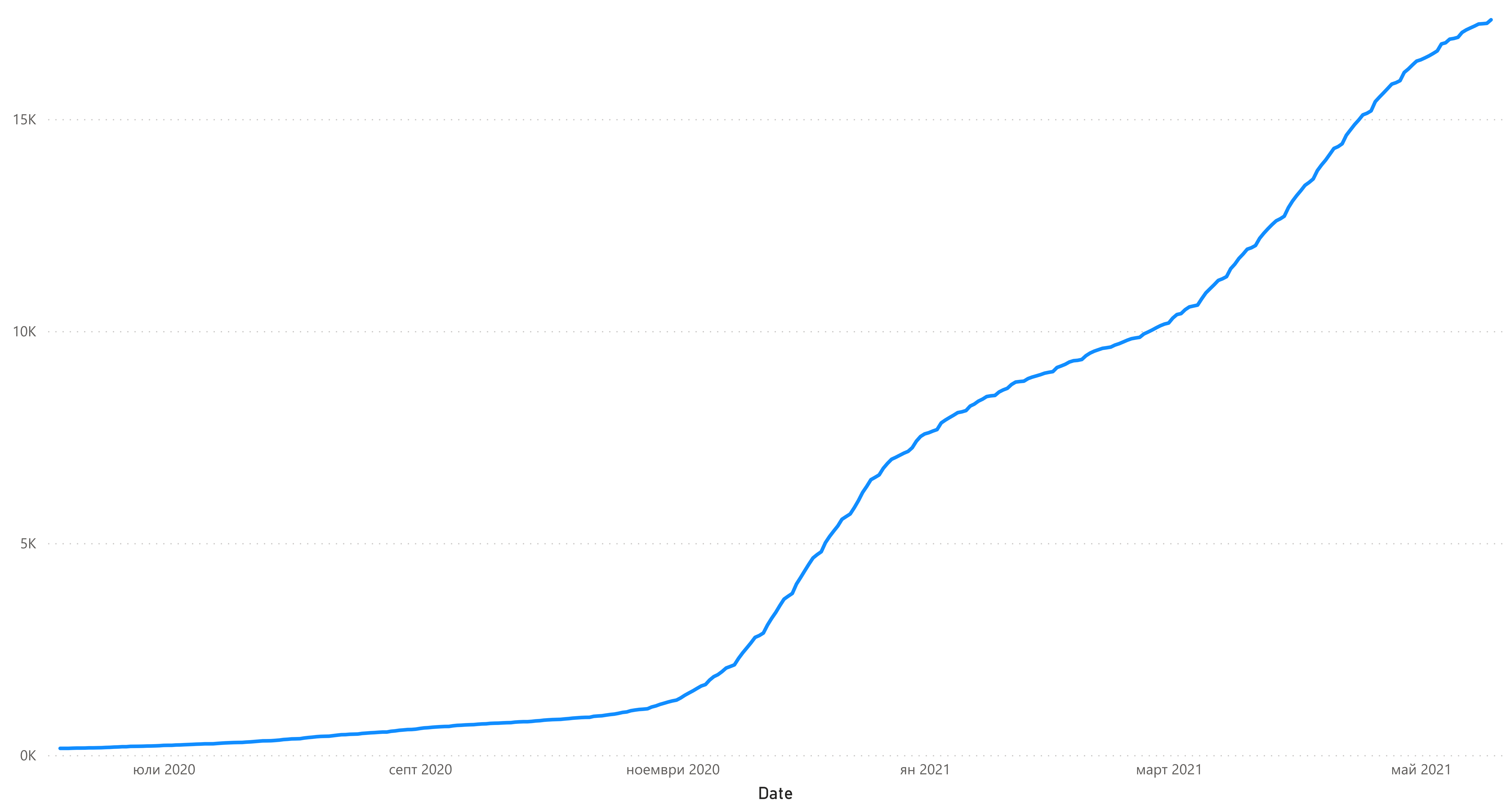
Previous Wave reached 590 intensive cases.

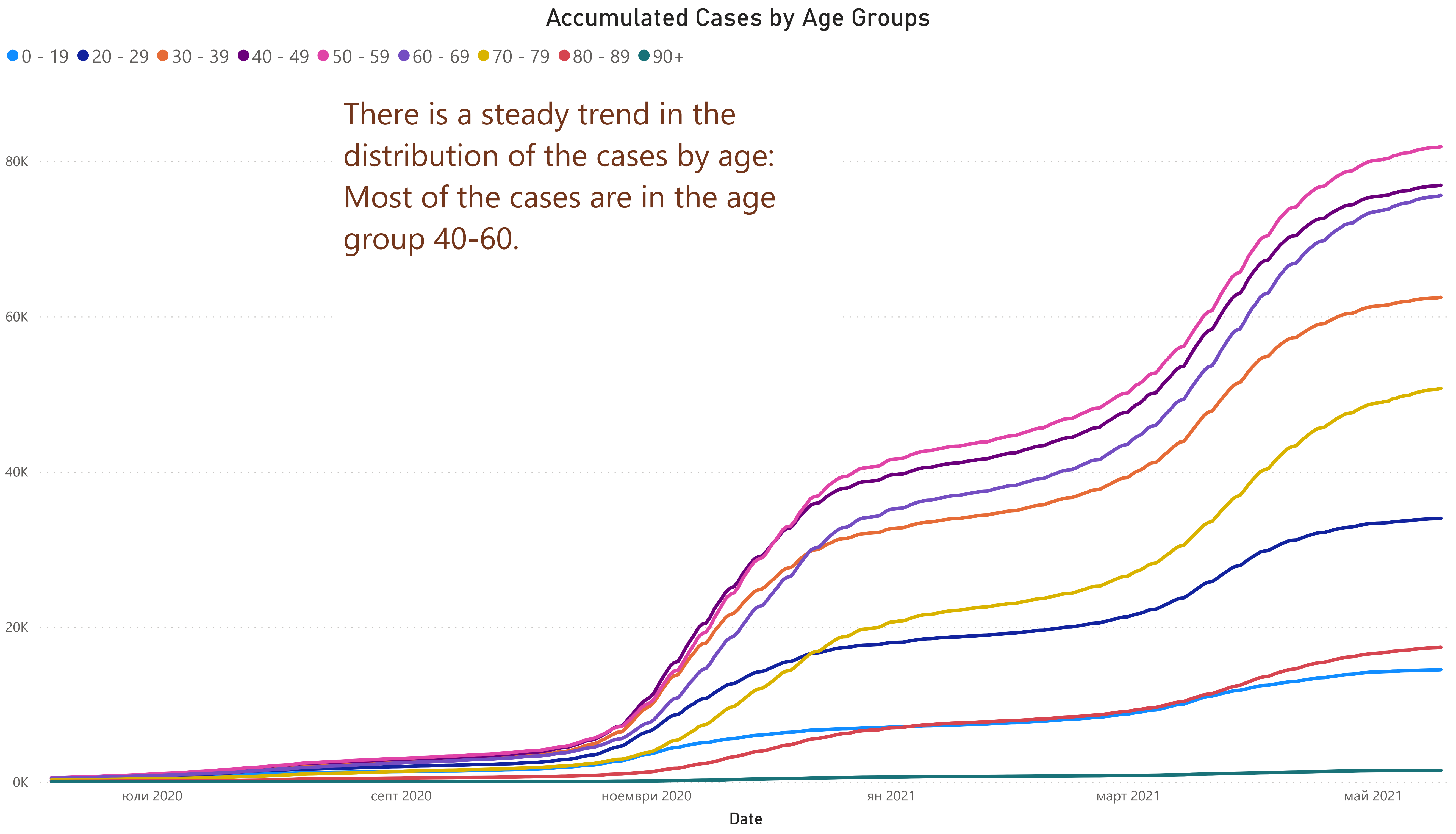
New wave reached 750 cases in intensive care.

There seems to be positive correlation between intensive care cases and deceased?

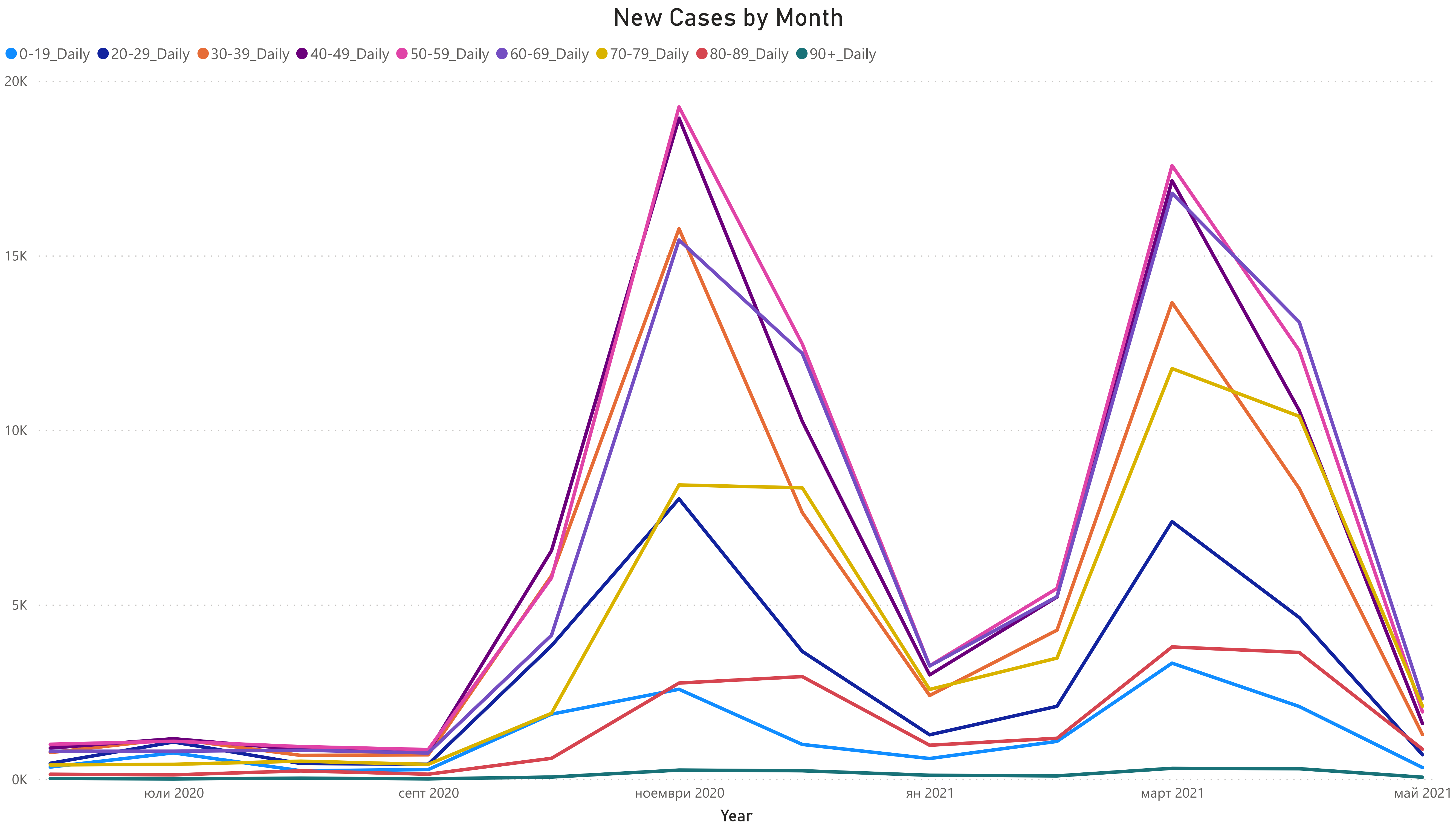
However, deceased cases do not seem to increase proportionally to the rate of increase in intensive care cases?

Deceased Accumulated (Daily basis)

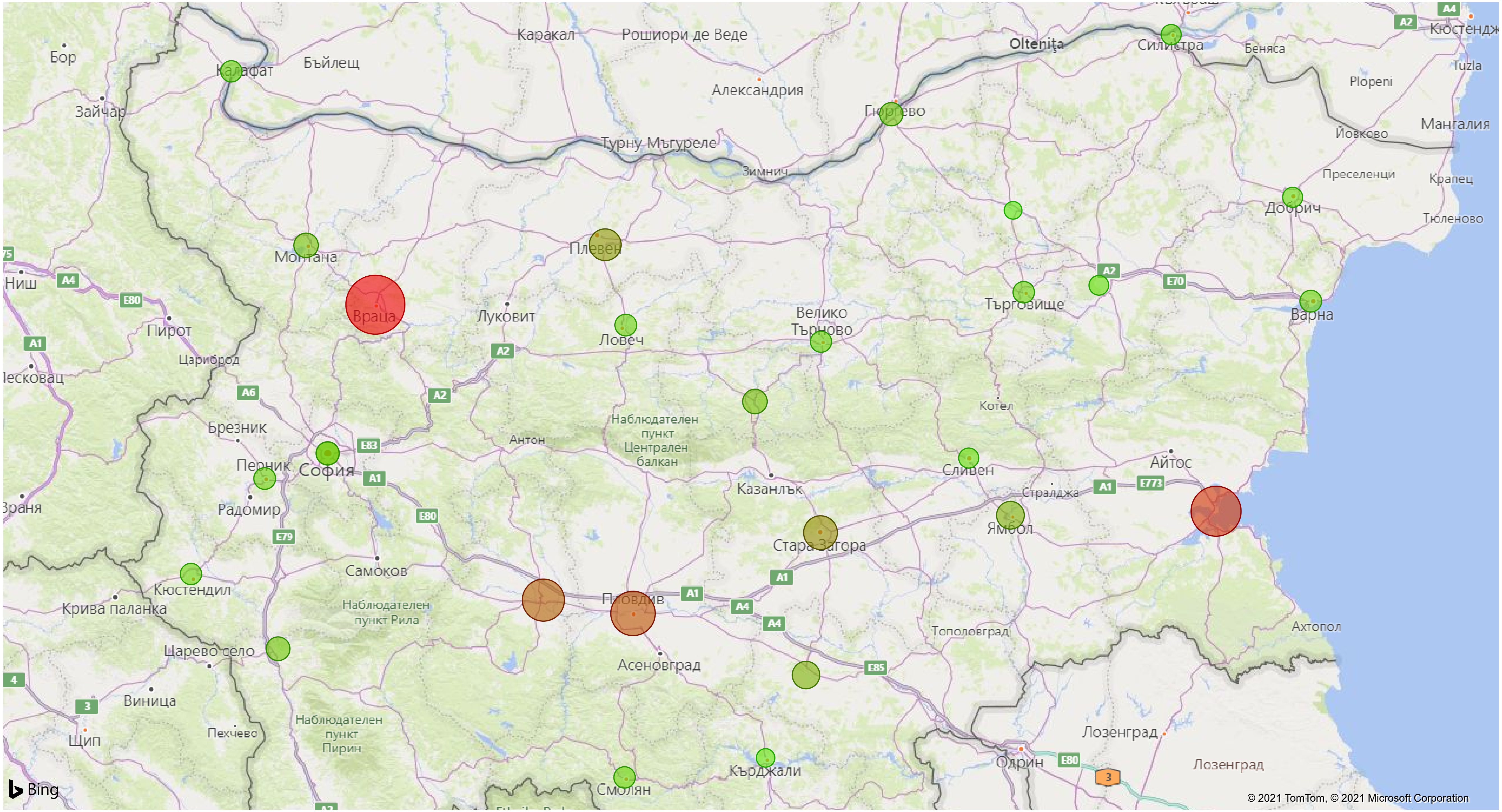


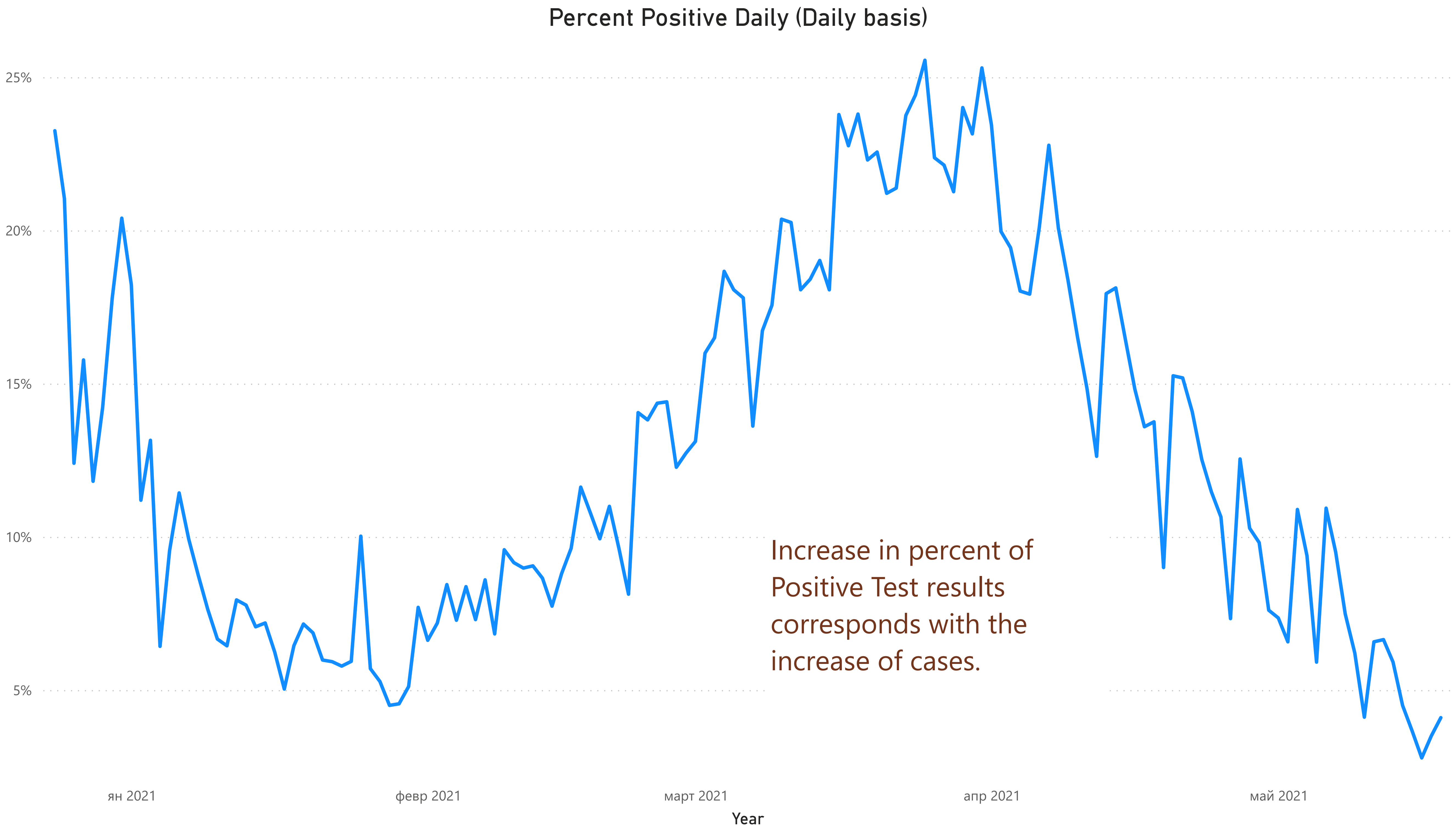


There is a steady trend in the distribution of the cases by age: Most of the cases are in the age group 40-60.

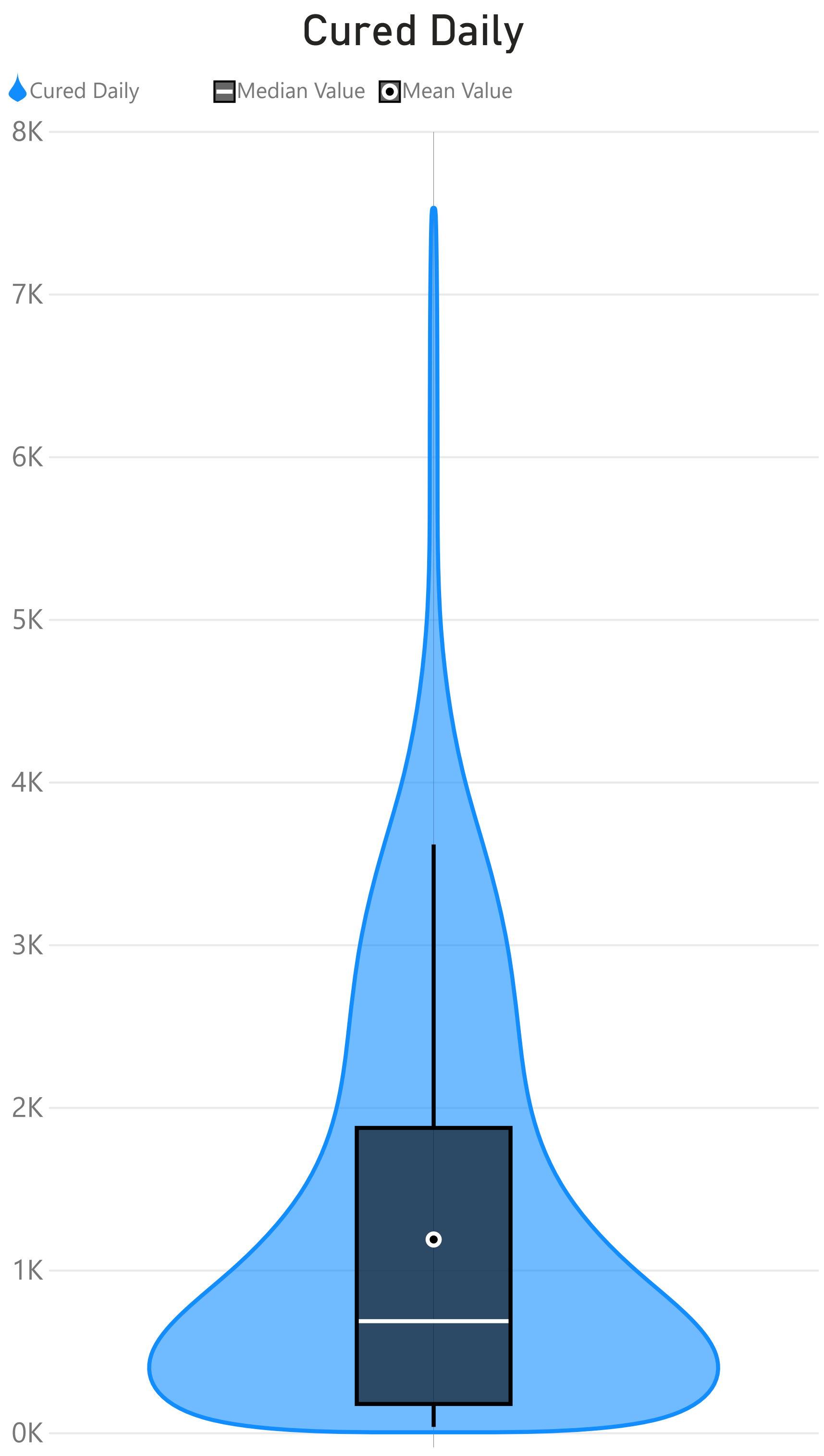

New Cases by Month
0-19_Daily
20-29_Daily
30-39_Daily
40-49_Daily
50-59_Daily
60-69_Daily
70-79_Daily
80-89_Daily
90+_Daily
20K
15K
10K
5K
0K
юли 2020
септ 2020
ноември 2020
ян 2021
март 2021
май 2021
Year

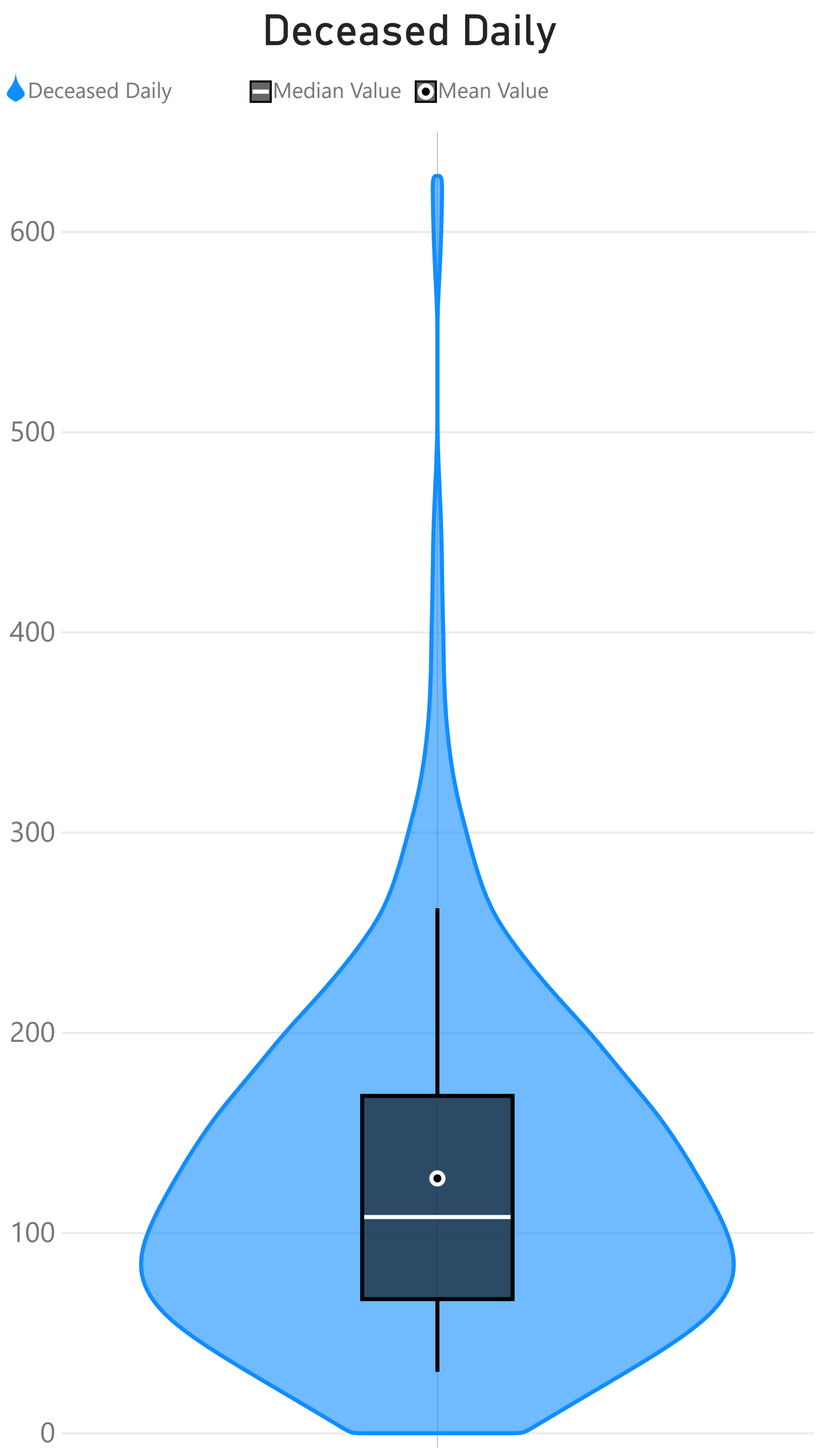
Latest Active Cases per 100k

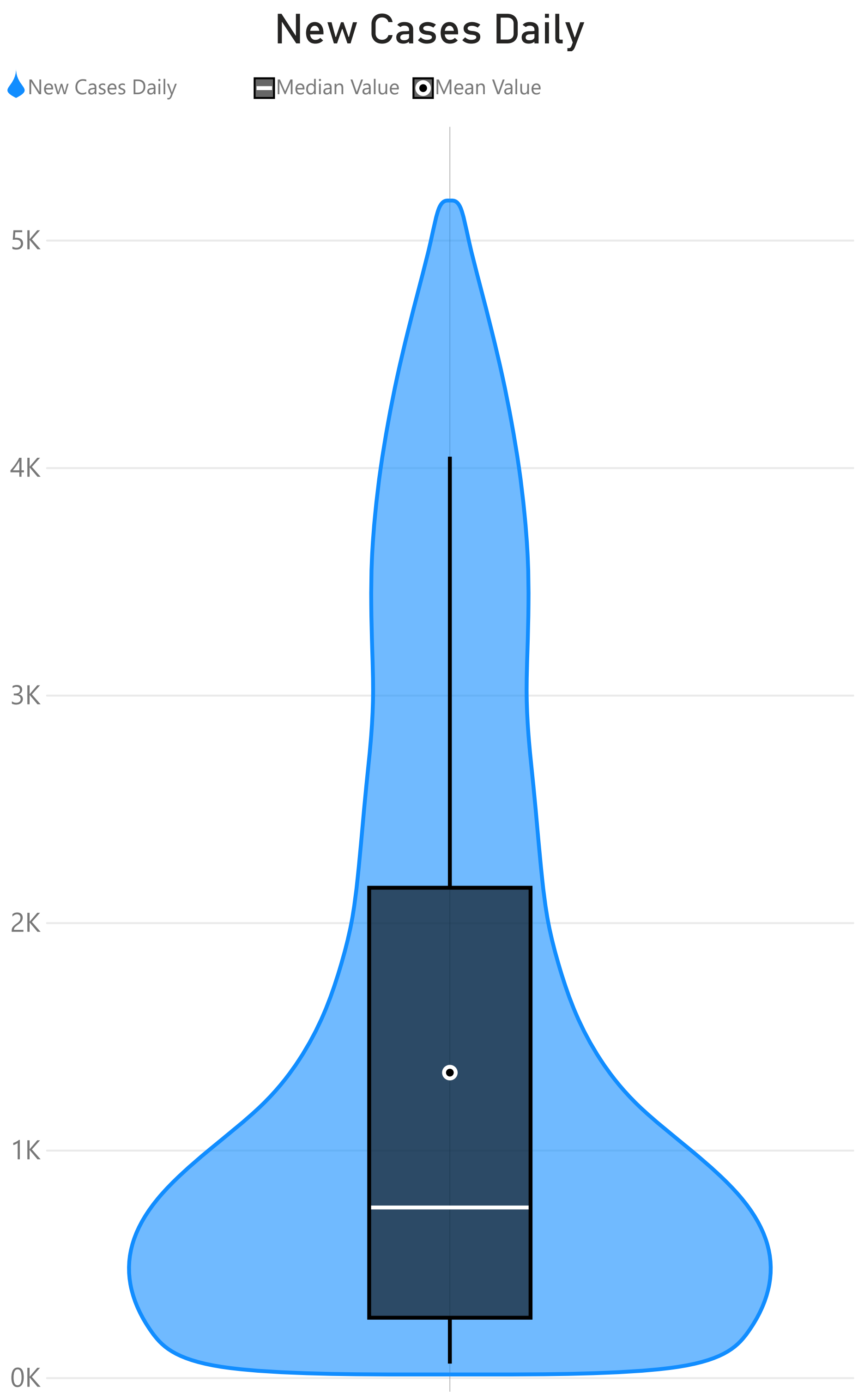




Increase in percent of Positive Test results corresponds with the increase of cases.


Cured Daily
Cured Daily
Median Value
Mean Value
8K
7K
6K
5K
4K
3K
2K
1K
0K


Deceased Daily
Deceased Daily
Median Value
Mean Value
600
500
400
300
200
100
0


New Cases Daily
New Cases Daily
Median Value
Mean Value
5K
4K
3K
2K
1K
0K

Descriptive Statistics by Region and Daily Cases

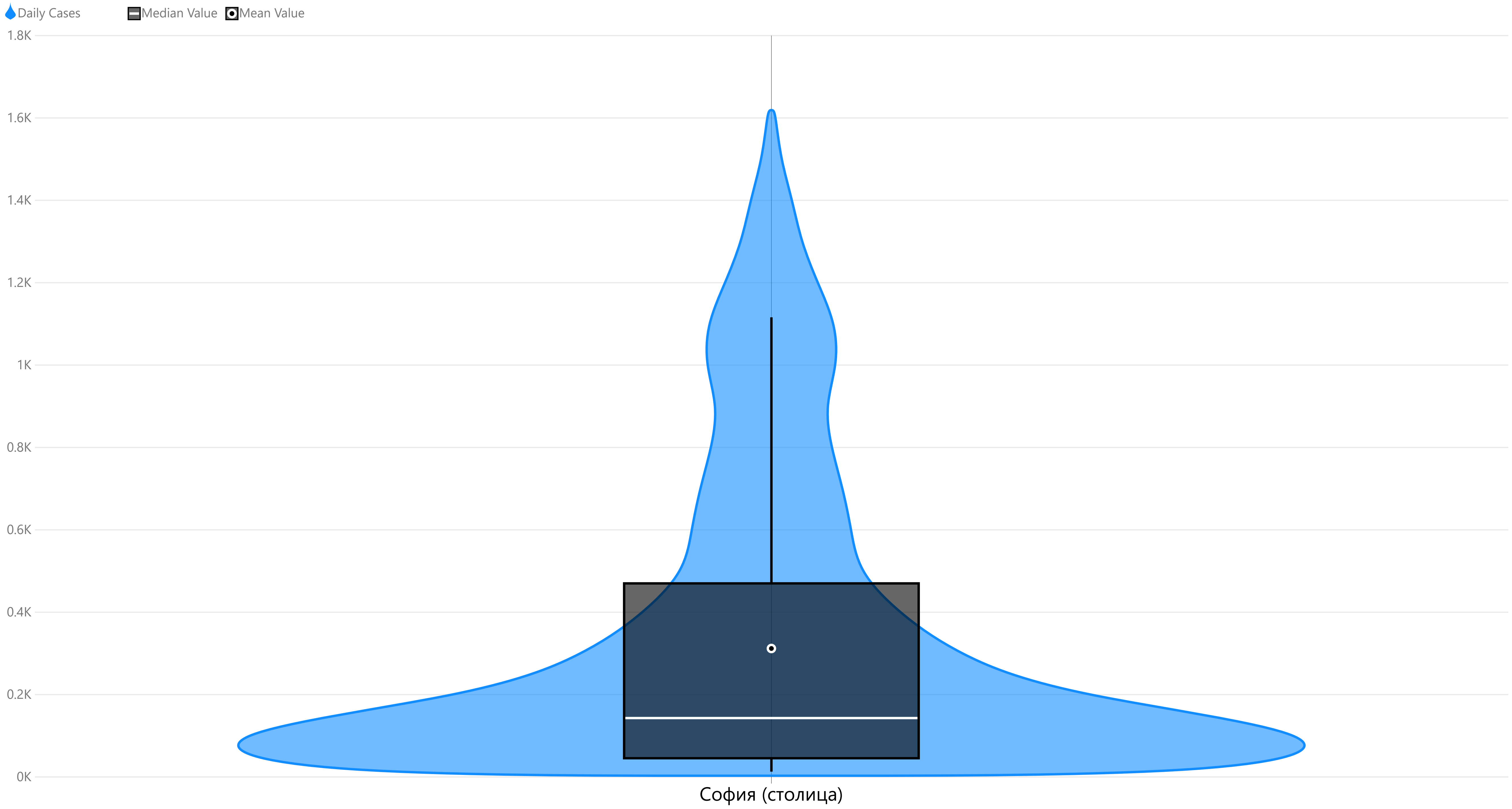
Region	Mean	Median	Standard Deviation
Благоевград	52,90	29	56,29
Бургас	78,20	35	100,96
Варна	89,33	37	106,44
Велико Търново	27,35	11	36,54
Видин	8,29	3	14,35
Враца	28,96	12	35,78
Габрово	19,30	8	28,90
Добрич	22,50	12	26,74
Кърджали	12,81	6	15,70
Кюстендил	25,60	13	29,70
Ловеч	16,15	7	20,21
Монтана	18,07	7	24,36
Пазарджик	32,12	18	33,34
Перник	22,94	11	26,98
Плевен	37,71	19	45,35
Пловдив	104,22	45	122,12
Разград	10,76	5	12,98
Русе	40,37	15	53,58
Силистра	16,97	5	25,49
Сливен	29,67	17	32,62
Смолян	13,67	7	16,05
София	38,39	17	45,28
София (столица)	311,94	143	369,15
Стара Загора	50,70	19	66,68
Търговище	11,29	5	14,25
Хасково	28,07	10	34,70
Шумен	27,83	10	34,75
Ямбол	19,50	8	25,47

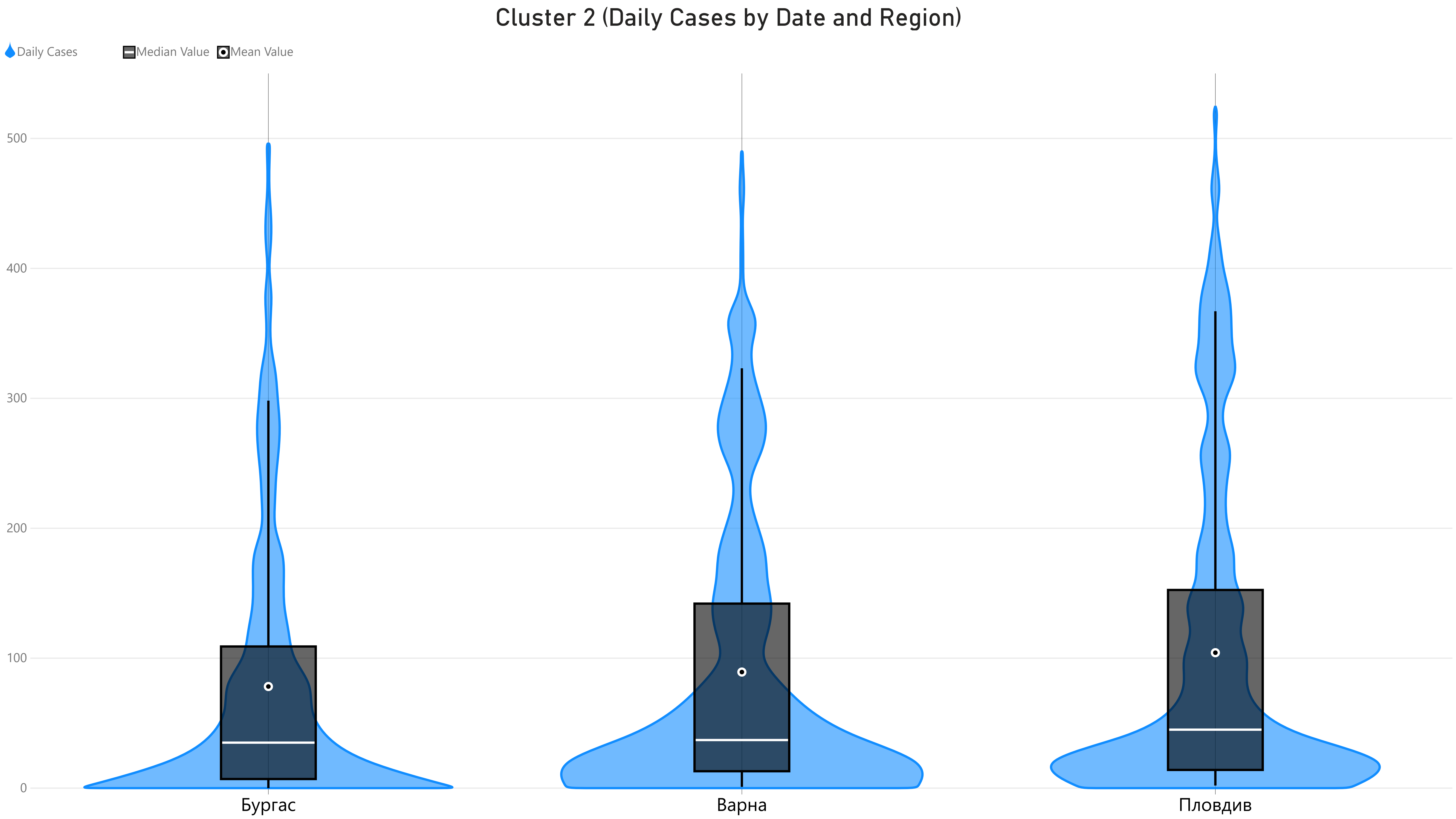
Distribution of daily cases seem to vary from one region to another.

Still in some regions, the distribution looks similar.

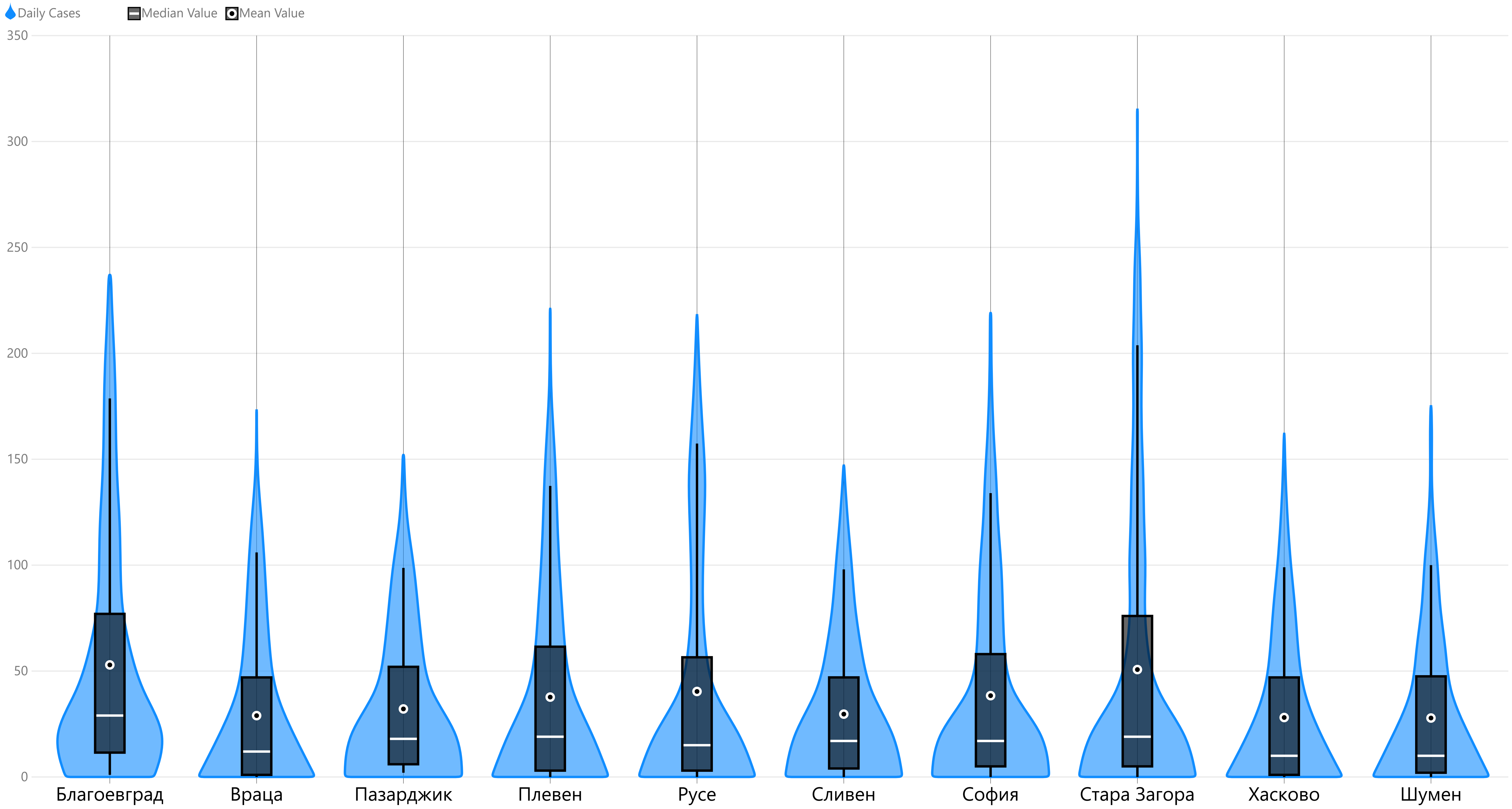
Using k-means, the regions are grouped in 4 clusters in the following pages.

Cluster 1 (Daily Cases by Date and Region)

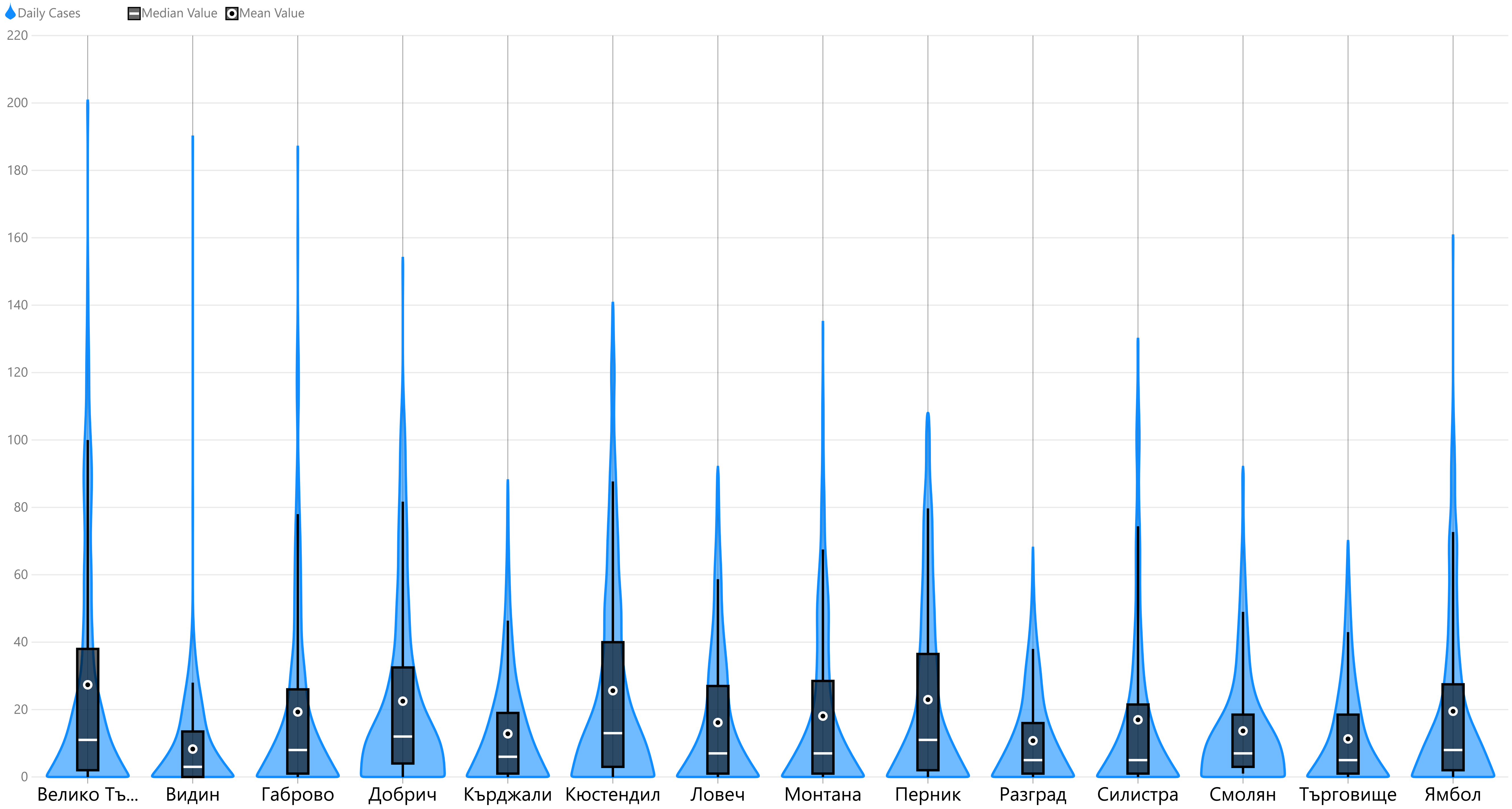


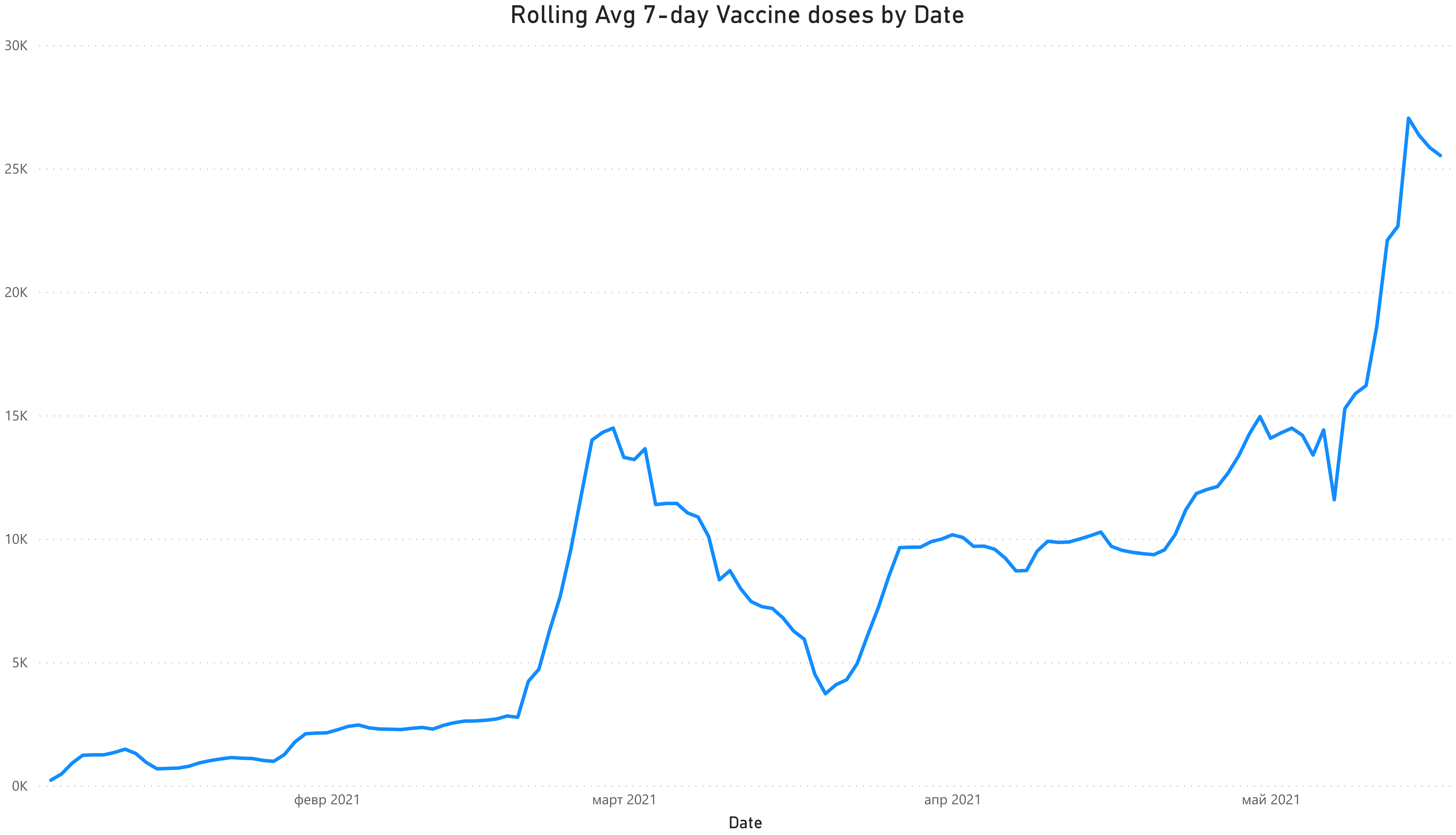

Cluster 2 (Daily Cases by Date and Region)
Daily Cases
Median Value
Mean Value
500
400
300
200
100
0
Бургас
Варна
Пловдив

Cluster 3 (Daily Cases by Date and Region)

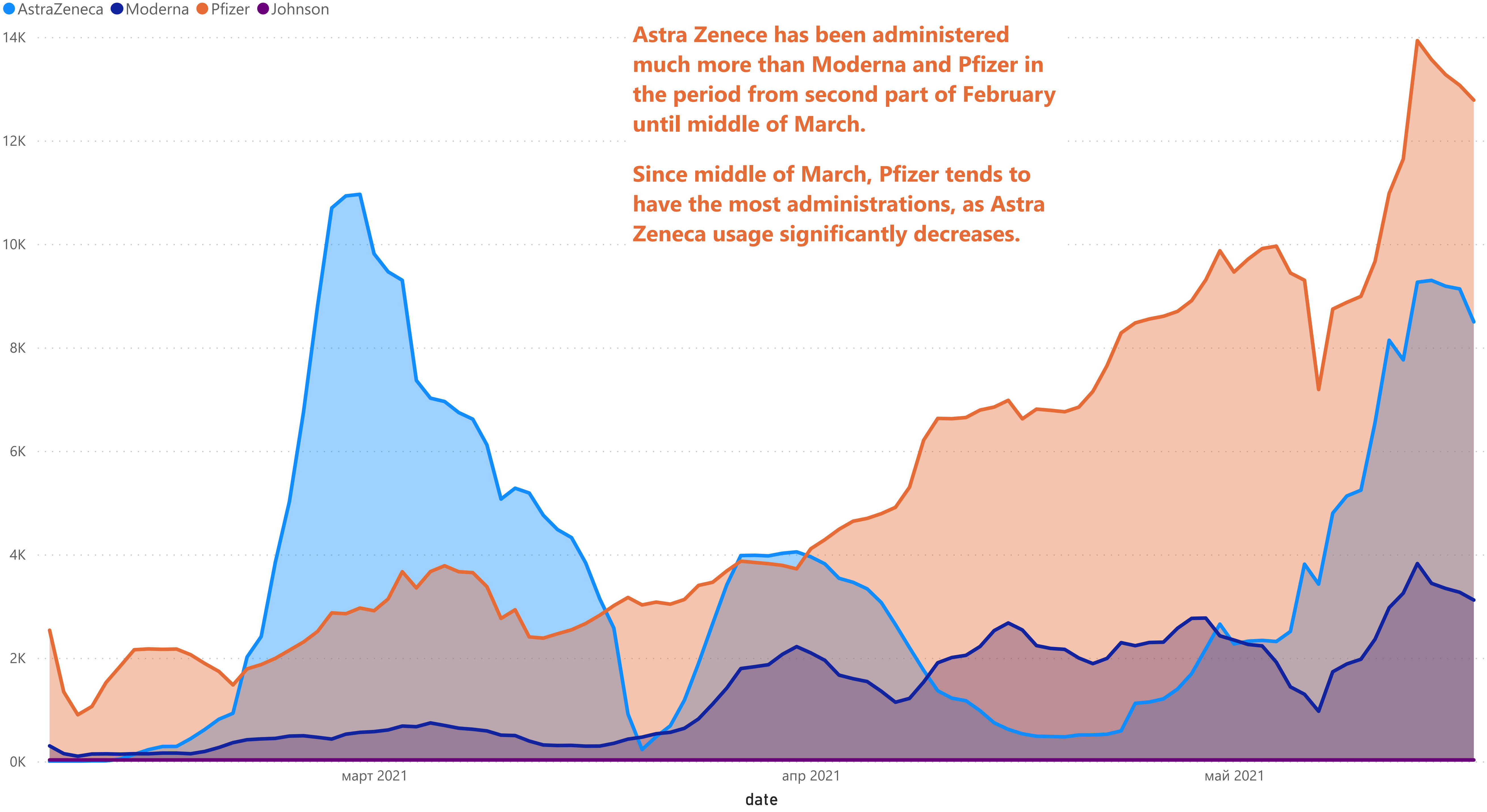


Cluster 4 (Daily Cases by Date and Region)




Rolling Avg 7-day Vaccine doses by Date
30K
25K
20K
15K
10K
5K
0K
февр 2021
март 2021
апр 2021
май 2021
Date

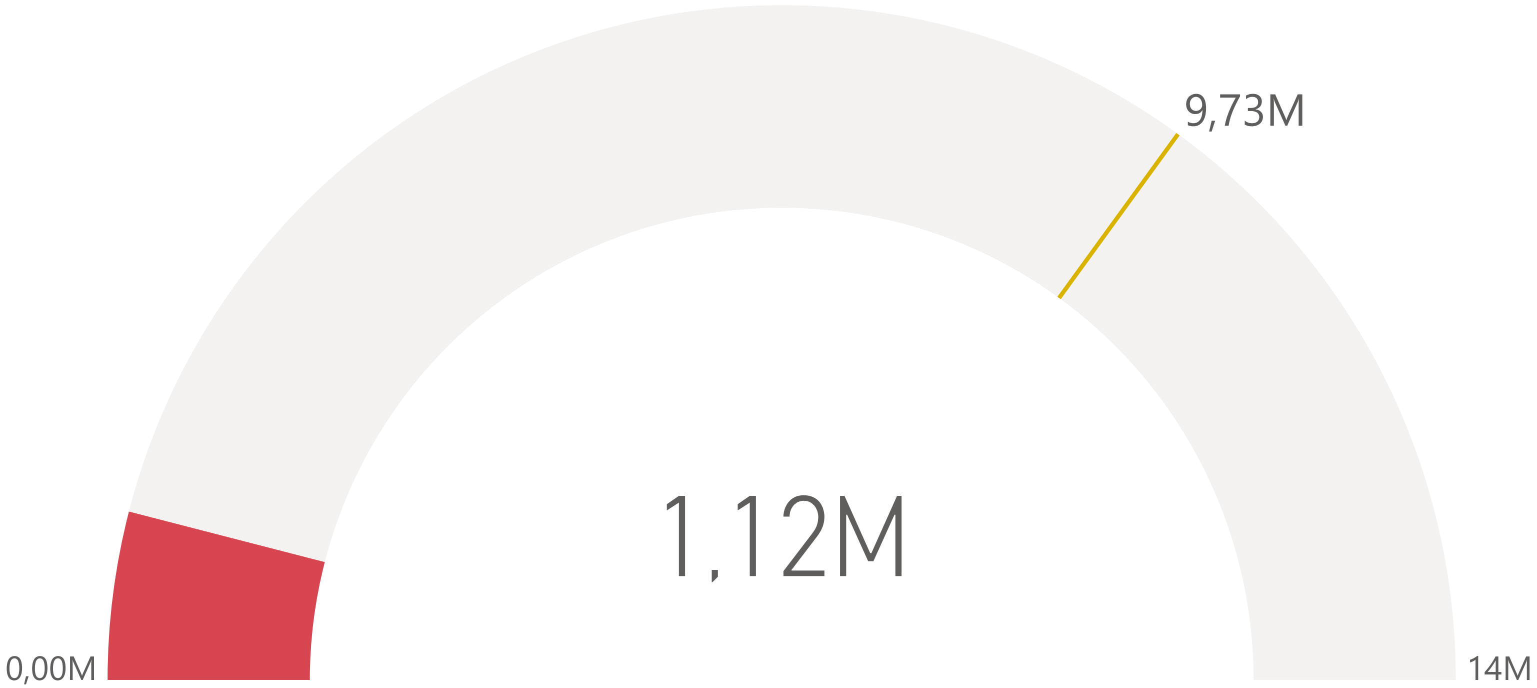
Astra Zeneca, Moderna and Pfizer doses by date (7-day moving average)



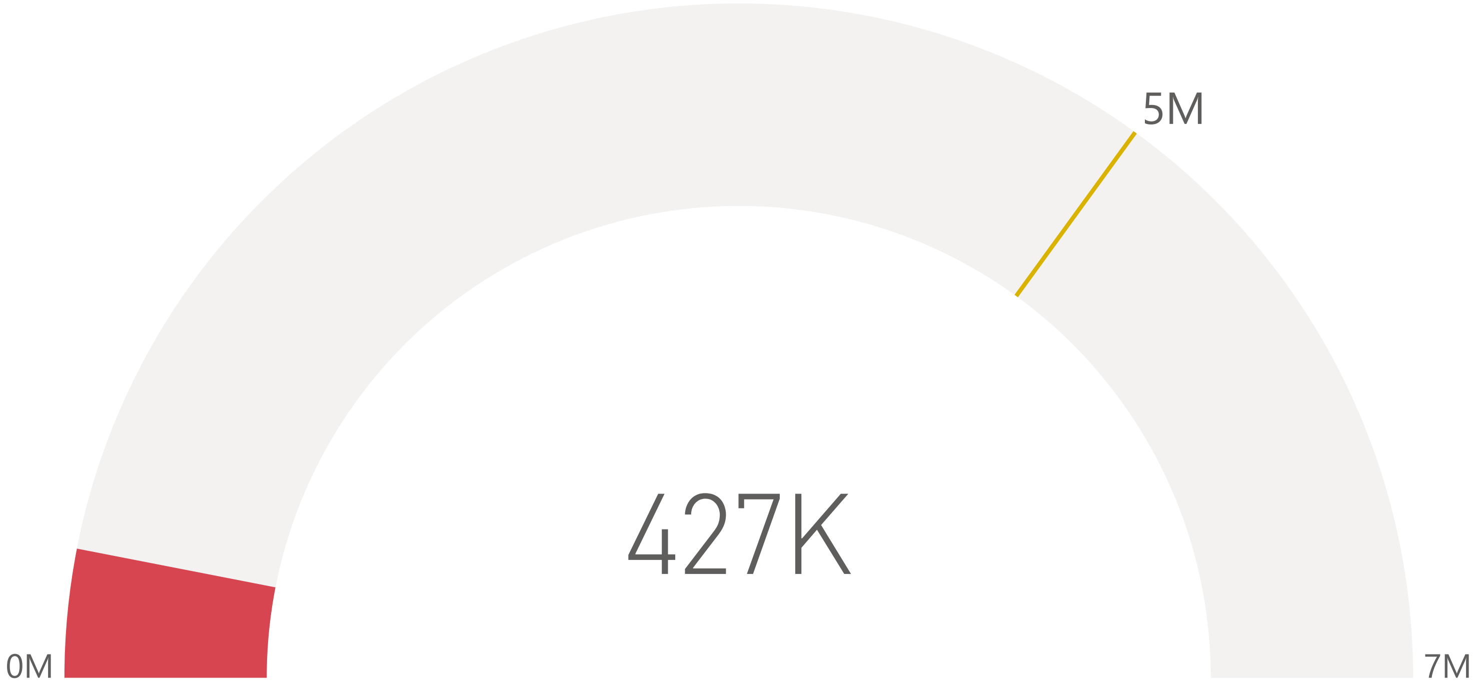
Astra Zenece has been administered much more than Moderna and Pfizer in the period from second part of February until middle of March.

Since middle of March, Pfizer tends to have the most administrations, as Astra Zeneca usage significantly decreases.

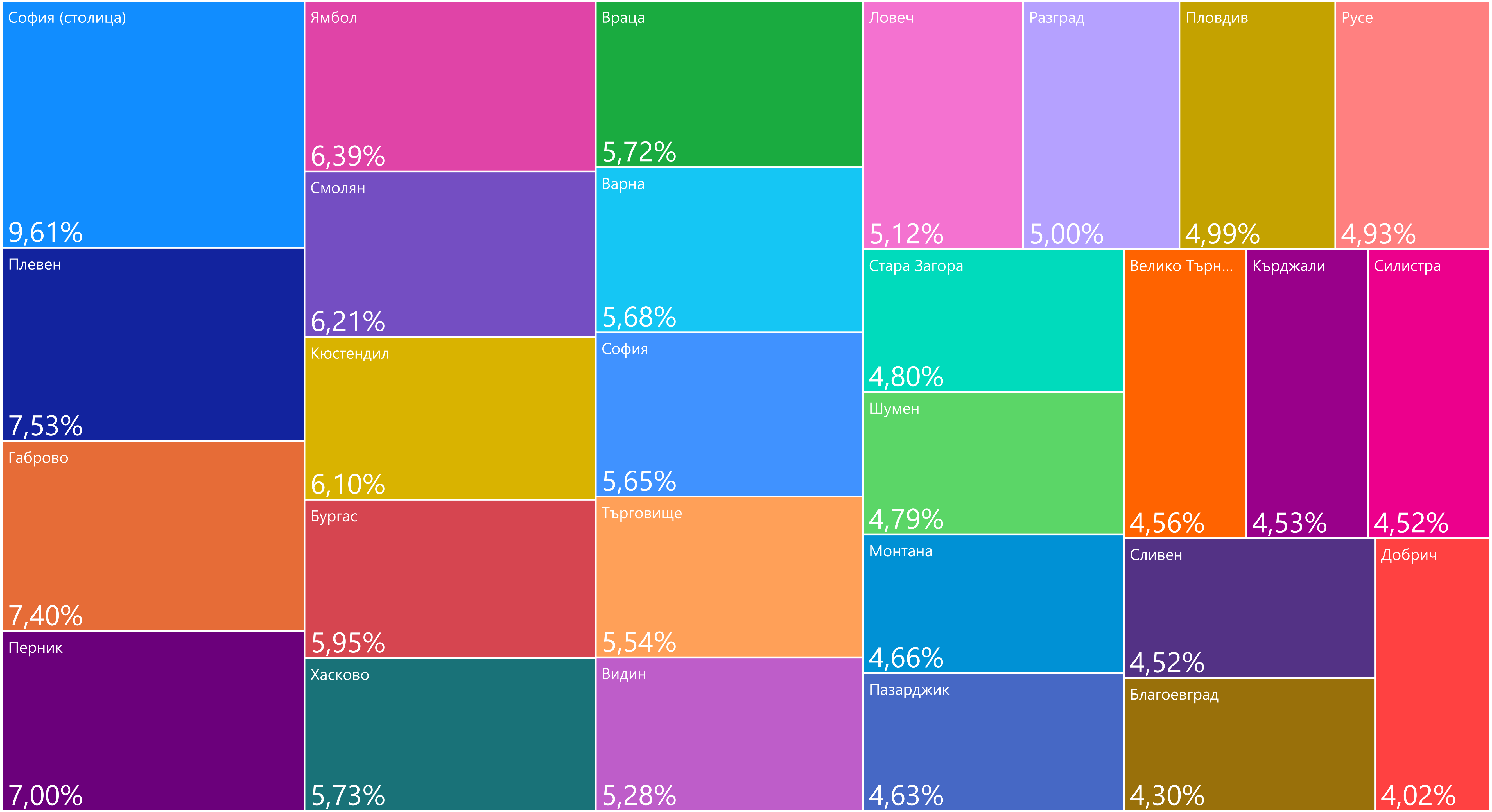
Total Vaccine Doses



Number of People with Two Doses



Percent Region Population with Second Dose



Sources

1. https://data.egov.bg/
2. https://coronavirus.bg/
3. https://www.nsi.bg/
4. https://github.com/svilens/covid-bulgaria
5. https://covid19bg.github.io/

Predicting Accumulated Cases with ARIMA

(18.5.2021)

The procedure follows these steps:

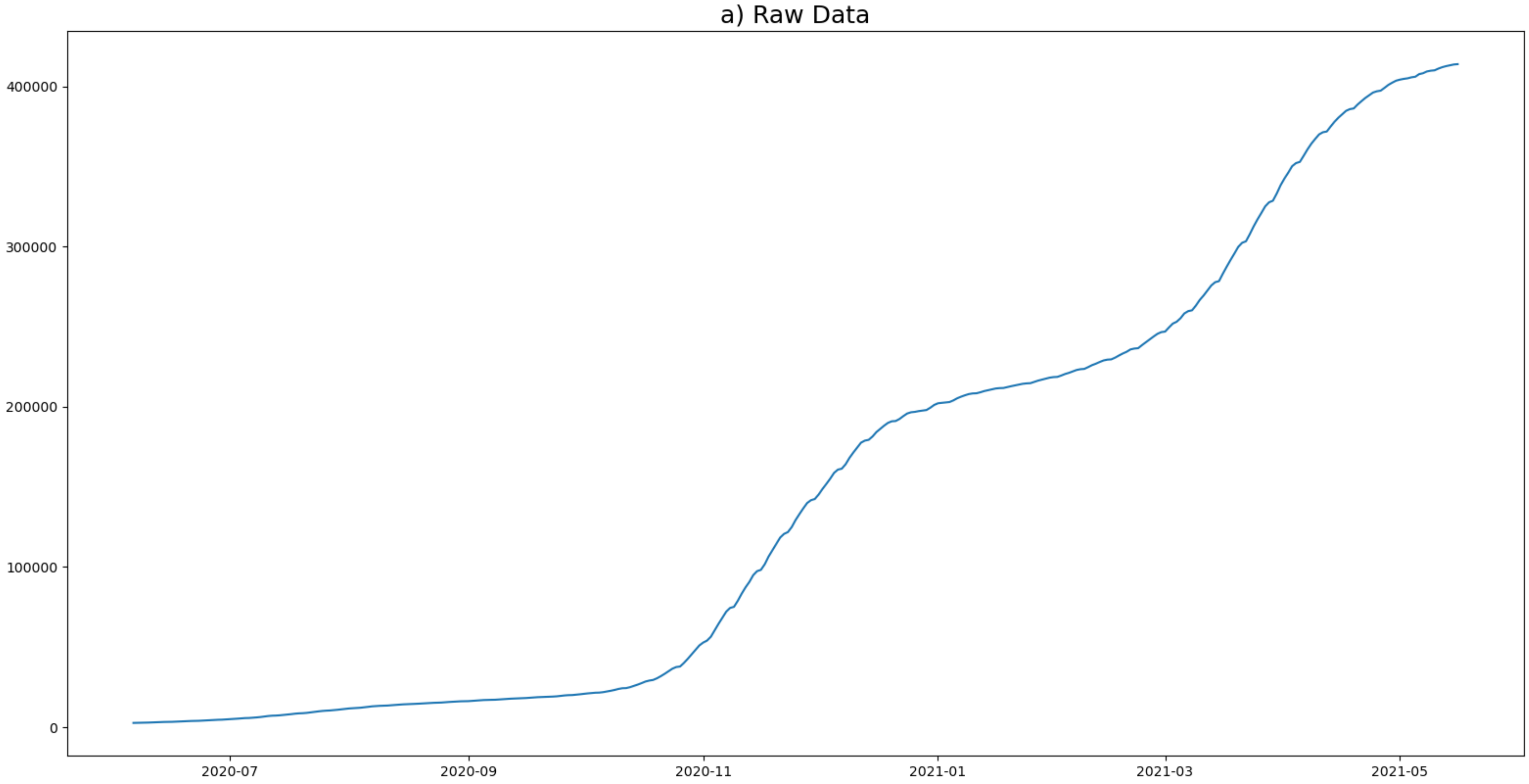
1. Exploration of the raw data, along with ACF and PACF plots, and Dickey-Fuller test to determine whether data is stationary.
2. If data is non-stationary, apply differencing (if needed also 2nd order differencing) and plot the data again. Use Dickey-Fuller test to confirm that data is stationary. Determine the presence of seasonality in the data.
3. Find the optimal parameters of ARIMA model (using auto_arima from pmdarima library in python)
4. Train and test the determined optimal ARIMA model.
5. Evaluate the results from the testing of the ARIMA model.
6. Build the ARIMA model and make a forecast with 30 day horizon.

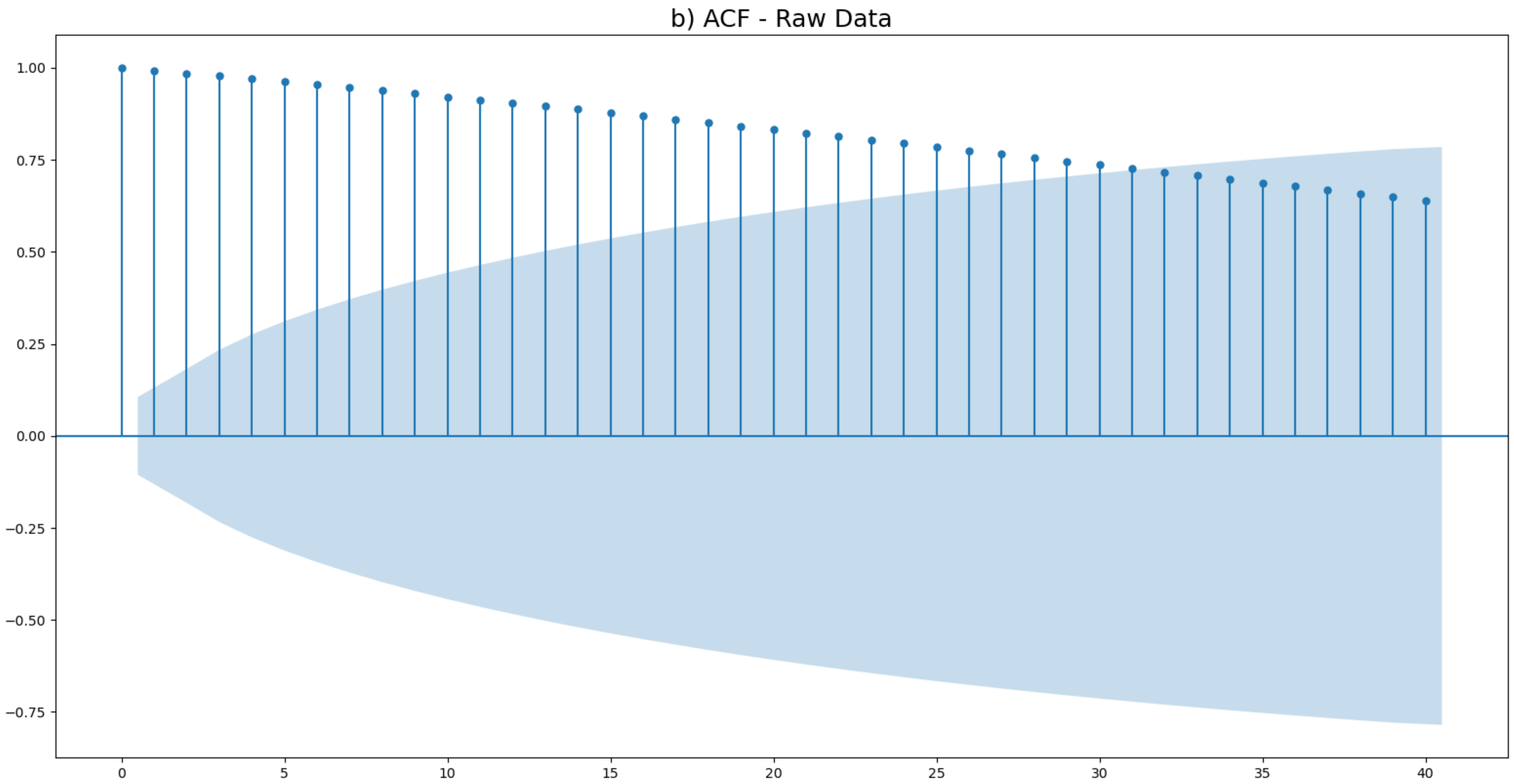
In the first three graphs are displayed:
a) The raw data (Accumulated Cases). The data appear to have a trend.

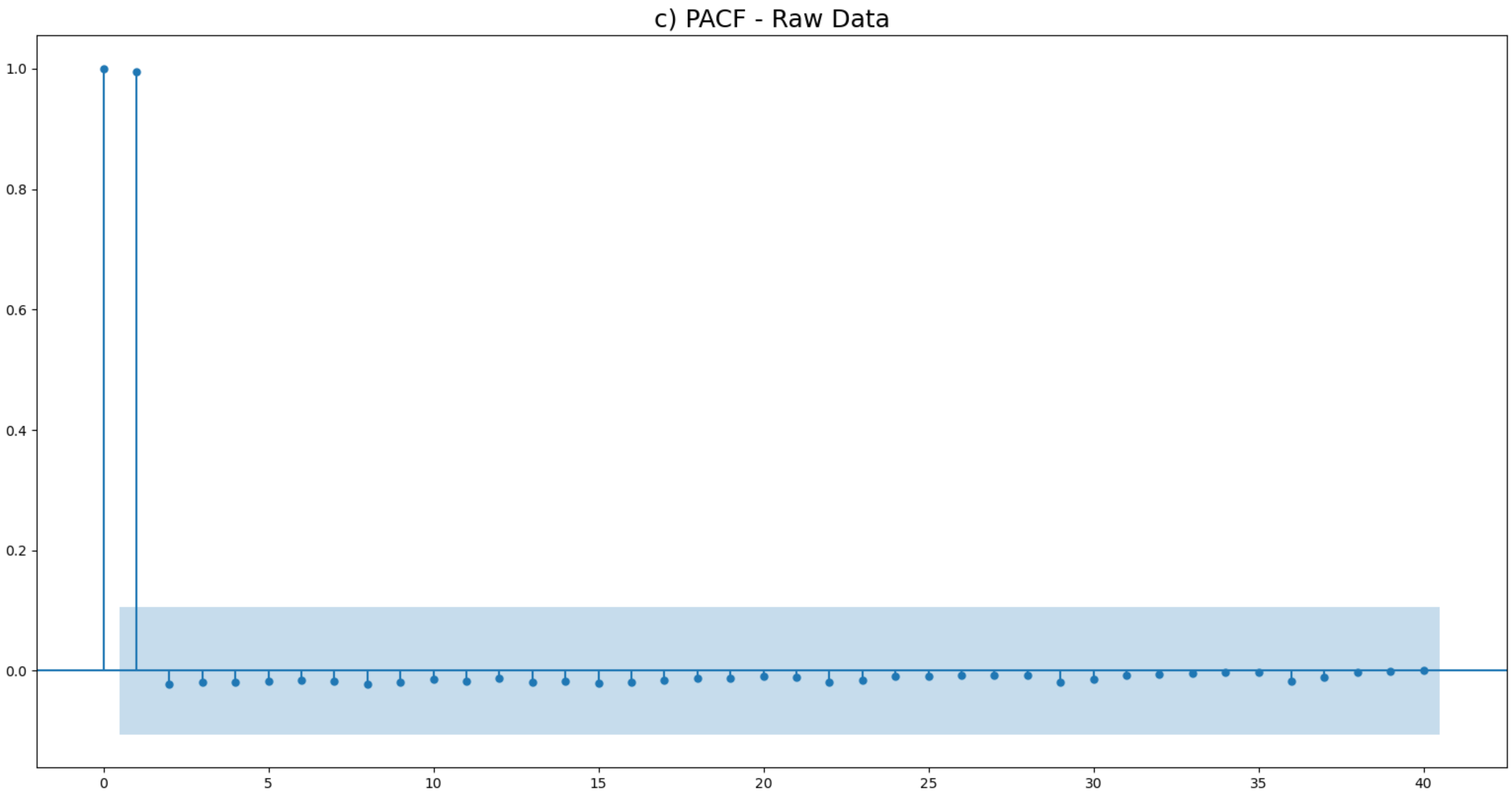
b) The auto-correlation function plot of the raw data. The auto-correlation coefficients decrease gradually with the increase of lags. This suggests non-stationarity.

c) The partial auto-correlation function plot of the raw data.

Additionally, Augmented Dickey-Fuller test is used as statistic for checking whether data is stationary. The result of the the statistic is 0.77 with p-value of 0.99. This is also a strong evidence that the data is non-stationary.


a) Raw Data
400000
300000
200000
100000
0
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05


b) ACF - Raw Data
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0
5
10
15
20
25
30
35
40



c) PACF - Raw Data
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35
40

First Order differencing is applied to the time series, in order to handle the stationary issue.

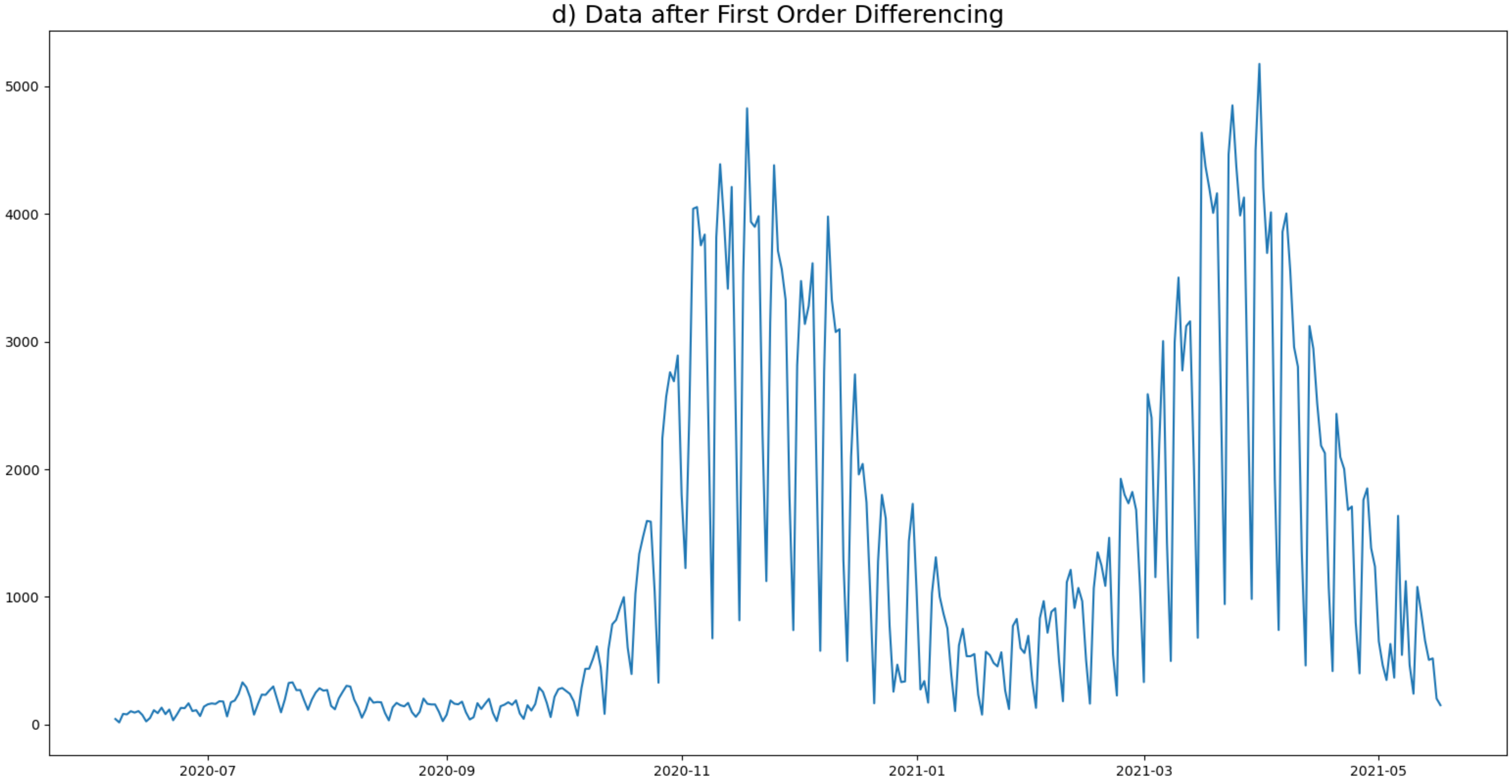
The next three figures (d, e and f) plot again the time series and its ACF and PACF plots.

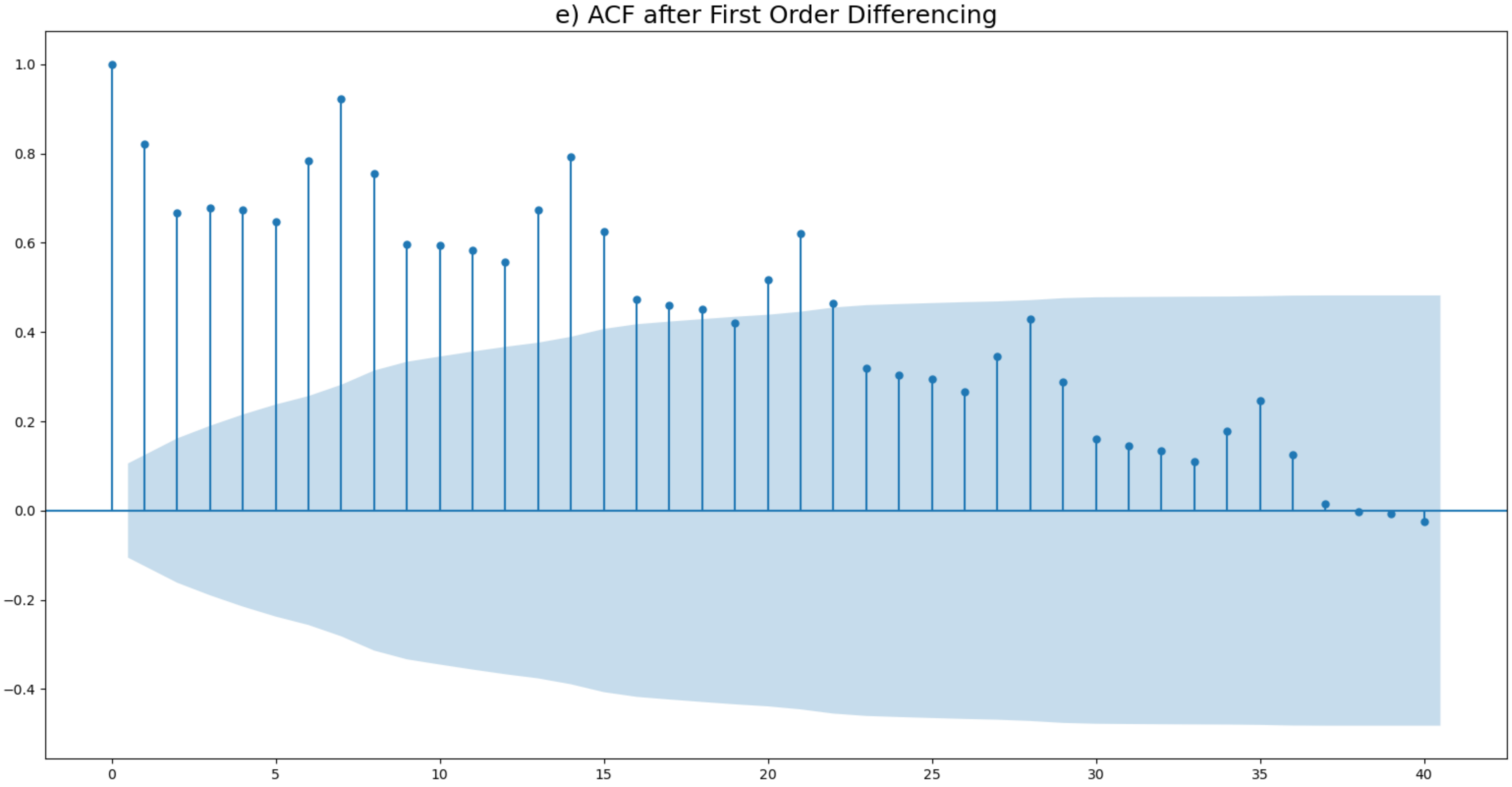
Although the data appears now to be better suited for ARIMA modelling, it still seems to be non-stationary.

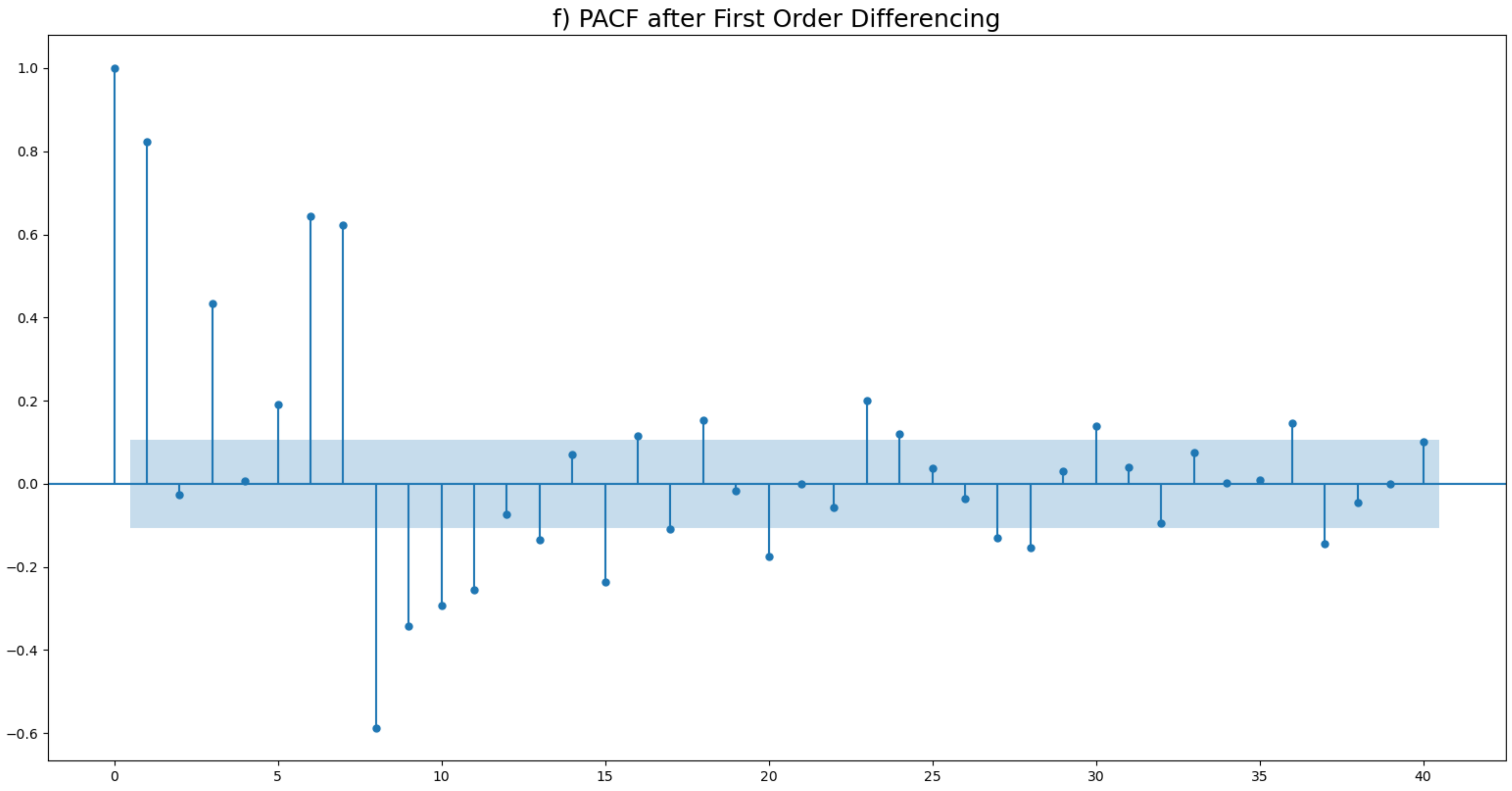
Dickey-Fuller test also has a value suggesting a possible stationary with statistics result of -2.61 (p-value = 0.09).

Second order differencing will be applied to be certain of stationary.

From the ACF plot, it is important to note that there seems to be seasonality in the data (7 observations per cycle), which suggests using SARIMA, in order to handle seasonality. For example, on Mondays confirmed cases tend to be less than cases reported on other days of the week.


d) Data after First Order Differencing
5000
4000
3000
2000
1000
0
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05

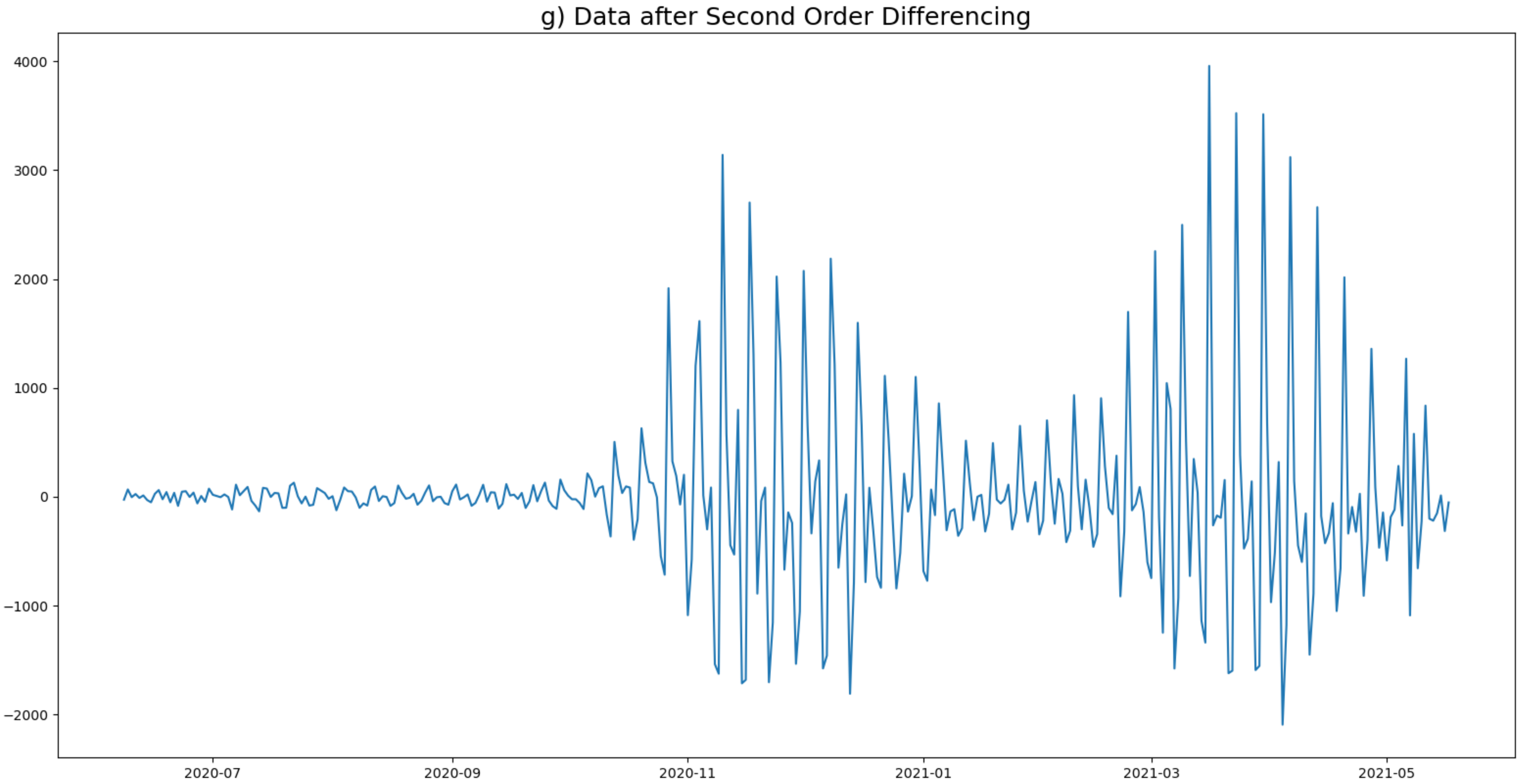

e) ACF after First Order Differencing
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
5
10
15
20
25
30
35
40

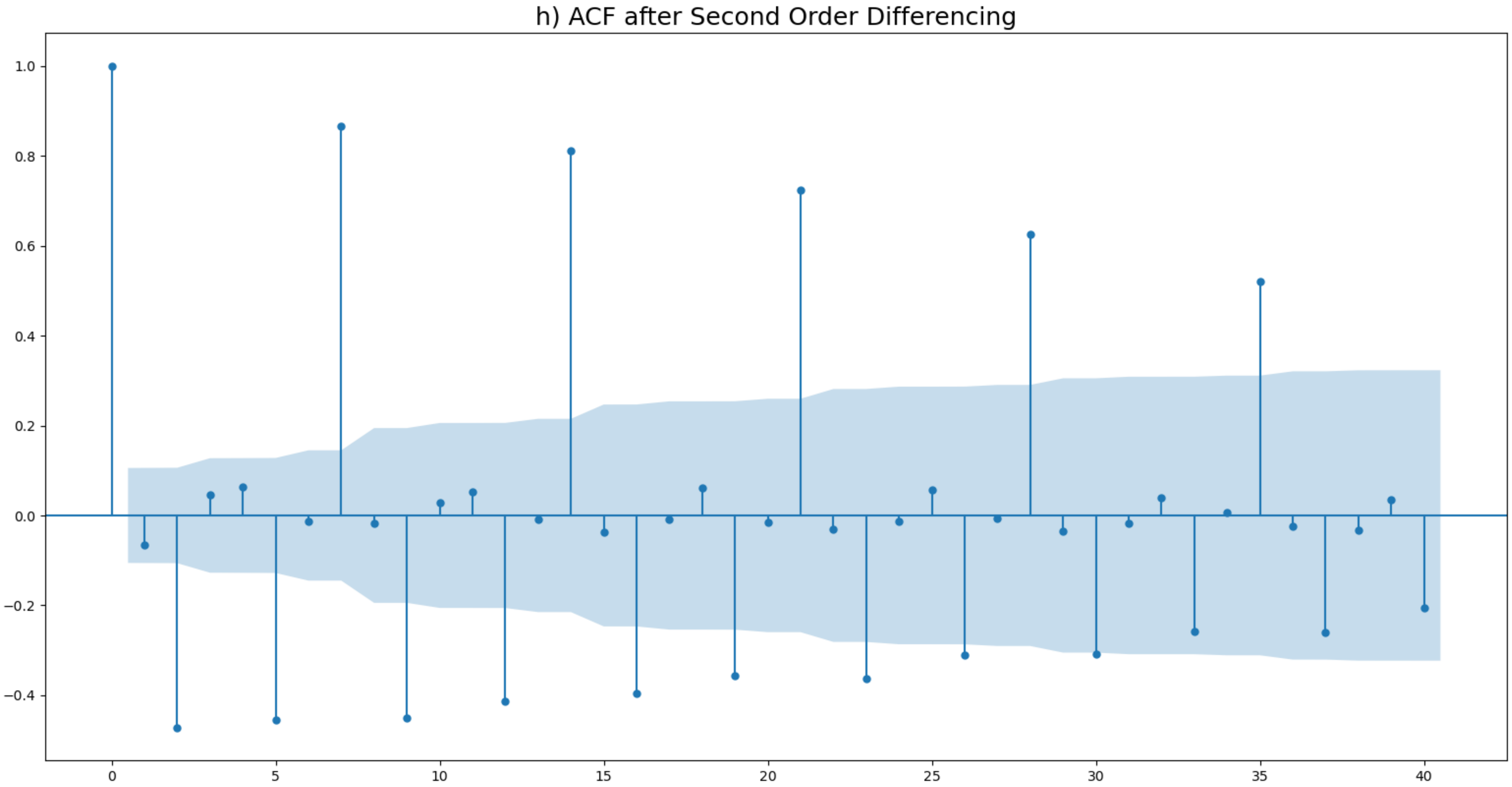

f) PACF after First Order Differencing
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
5
10
15
20
25
30
35
40

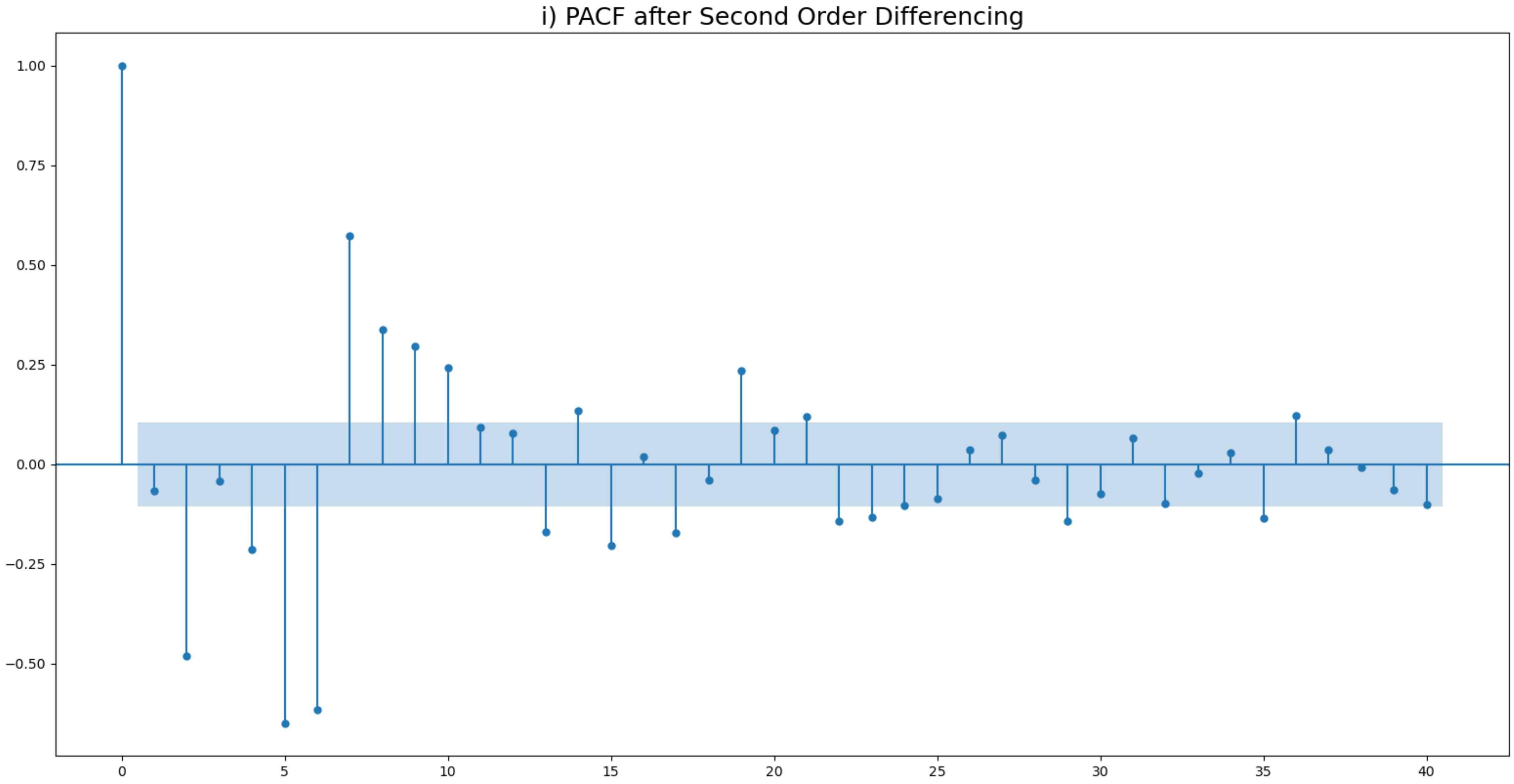
Graphs g), h), i) visualize the time series data, ACF and PACF after applying second order differencing.

The data seems to be always oscillating around 0 with different variance.

The result of Dickey Fuller statistic is -3.04 with p-value of 0.03. It can considered as a very strong evidence that the data is now stationary.


g) Data after Second Order Differencing
4000
3000
2000
1000
0
−1000
−2000
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05


h) ACF after Second Order Differencing
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
5
10
15
20
25
30
35
40


i) PACF after Second Order Differencing
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
0
5
10
15
20
25
30
35
40

The next step in the procedure is to determine the optimal parameters for the ARIMA model.

The pmdarima library available in python finds the parameters of ARIMA that would minimize the AIC (Akaike Information Criterion) in a given time series.

Since seasonality was detected during the data exploration step, SARIMA model is actually applied with 7 observations in a cycle.

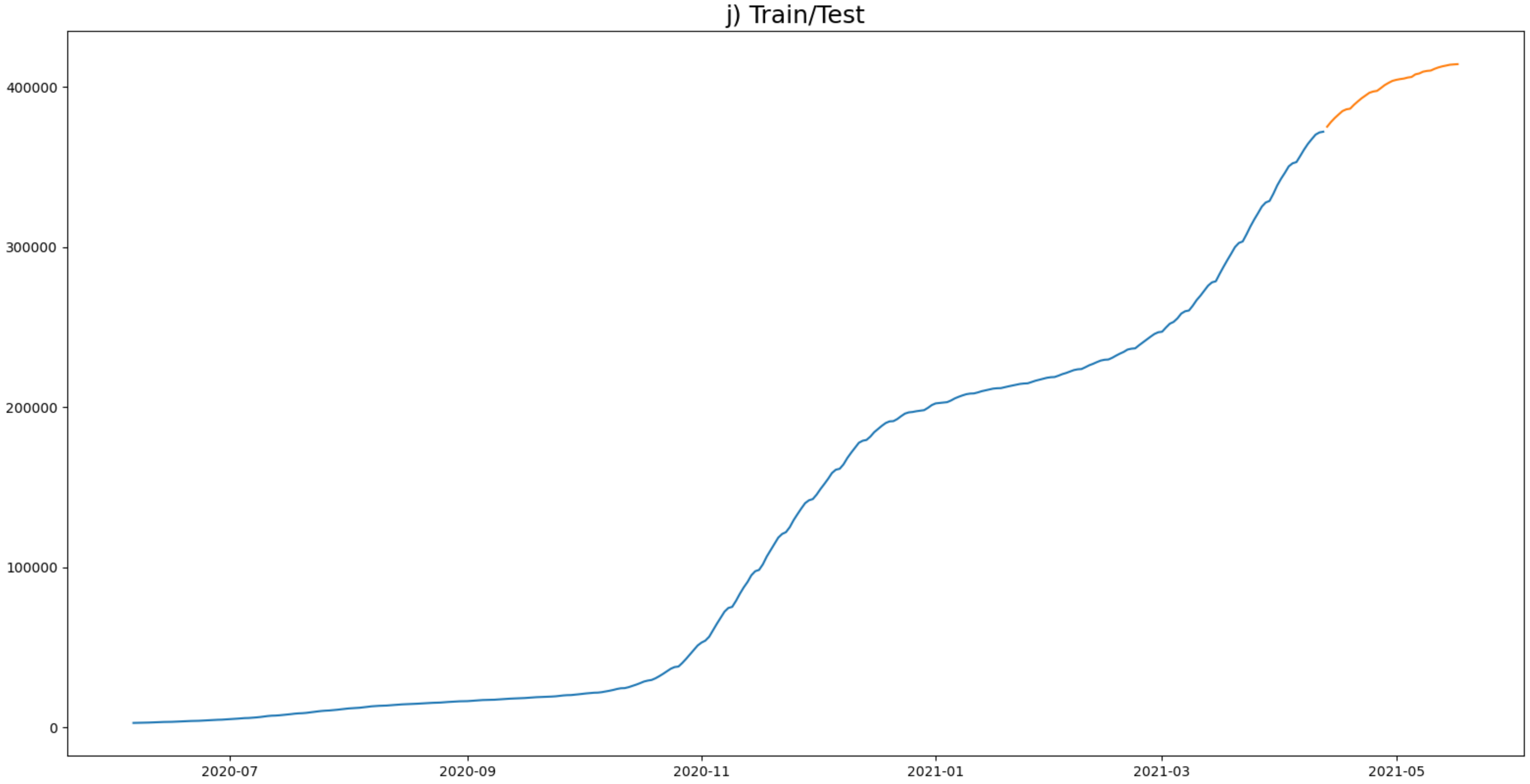
The results suggest using SARIMA with the following parameters: (p=3, d=2, q=0)(P=6, D=0, Q=1, m=7). The result for AIC is 4991.

This recommendation points to AR model with second order differencing (expected based on the exploration of the data).

According to the procedure, the data is split into train and test. The train subset is used to train the SARIMA model with determined parameters and the test subset is used to evaluate it. The data available spans from 6.6.2020 to 18.5.2021. These are in total 347 observations.

The observations up until 13.4.2021 are used for training and the remaining 10 percent (until 18.5.2021) are kept for testing.

Graph j) shows the split of the data.


j) Train/Test
400000
300000
200000
100000
0
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05

After the model is trained, it is used to make predictions for the period set aside for validation (14.4.2021 - 18.5.2021).

To evaluate the performance of the model three common metrics are used: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE)

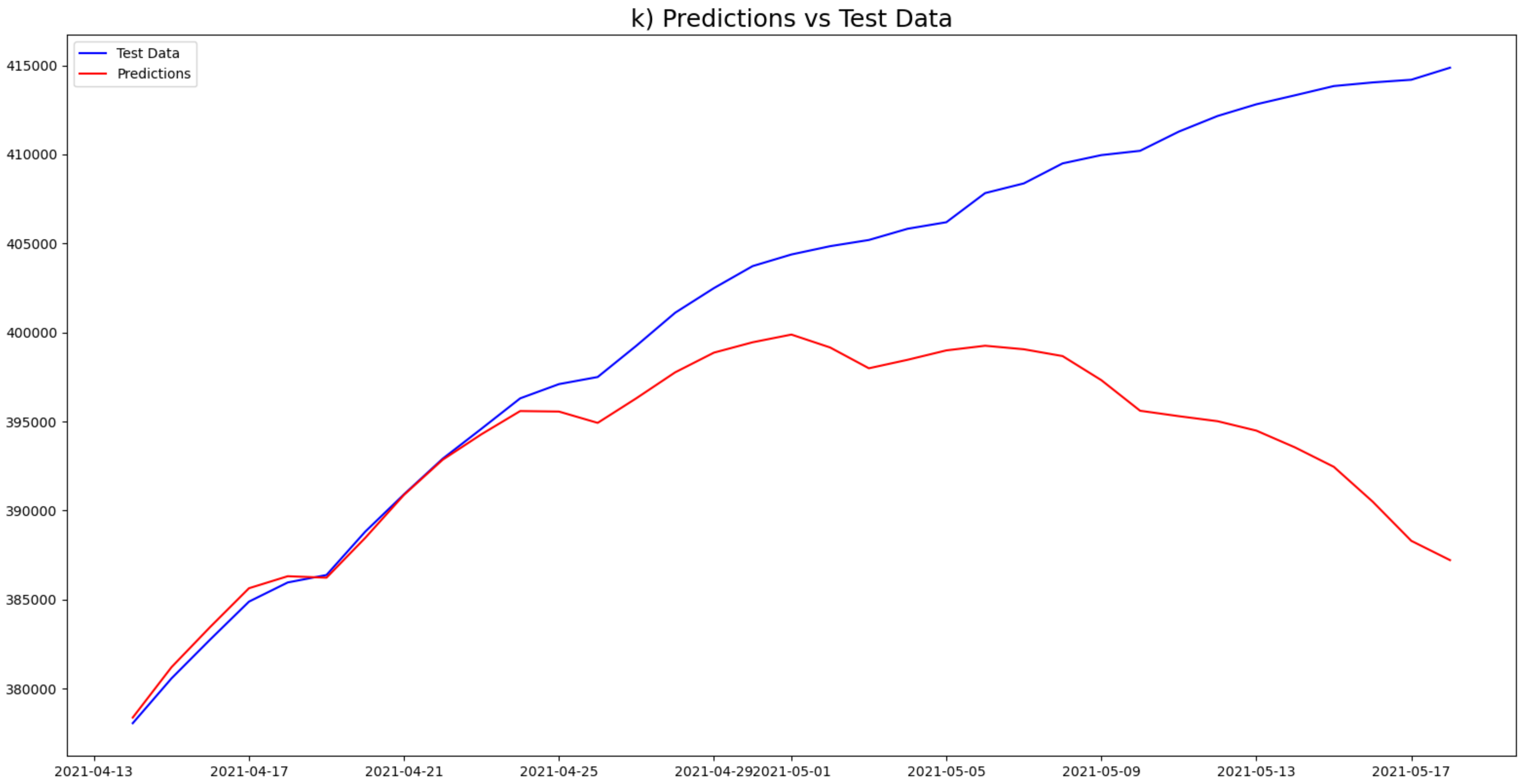
The results are as follows:

MSE = 132725755

RMSE = 11520

MAE = 8006

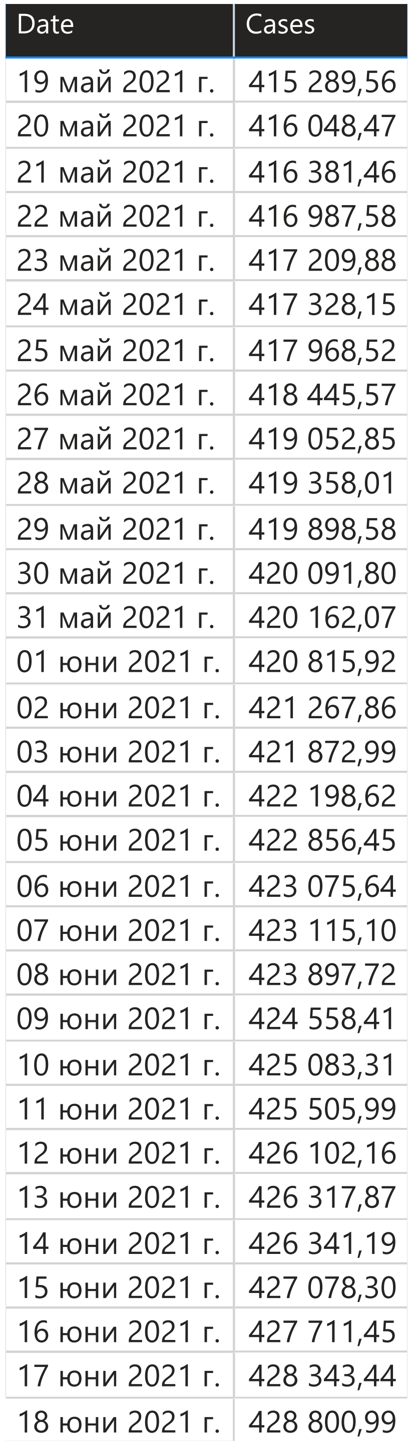
Graph k) plots the predictions against the actual test data. The model appears to make predictions closed to the expected in the first two weeks and after that error increases significantly.


k) Predictions vs Test Data
Test Data
Predictions
415000
410000
405000
400000
395000
390000
385000
380000
2021-04-13
2021-04-17
2021-04-21
2021-04-25
2021-04-29
2021-05-01
2021-05-05
2021-05-09
2021-05-13
2021-05-17

Finally, the model is built using all available data and forecast is made for the next 30 days. On the right is a table with the forecast of the accumulated confirmed cases for the next 30 days.

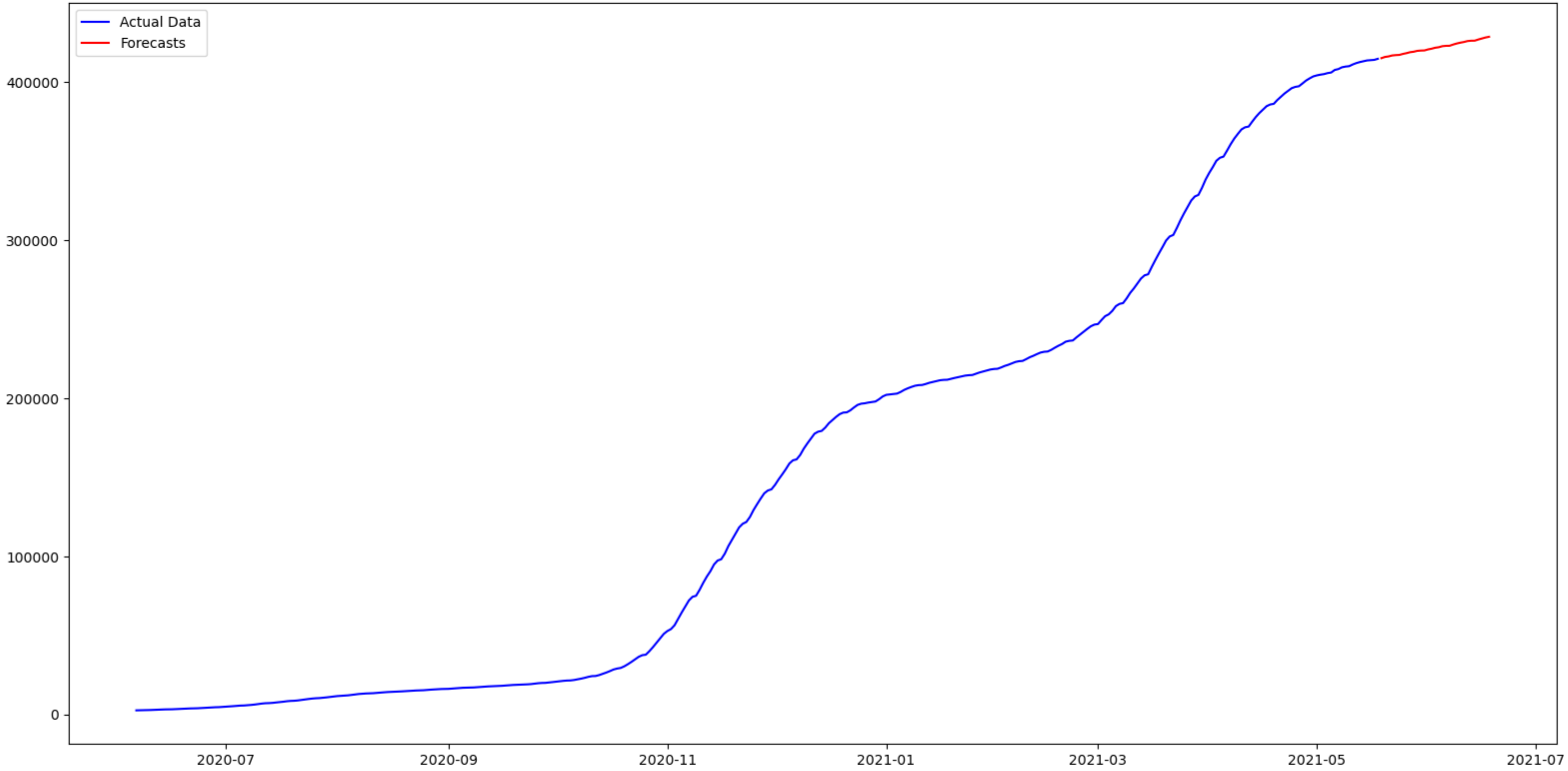
Graph l) visualizes the forecast.

Based on the predictions, it can be anticipated the cases to increases with around 13 000 by 18.6.2021.



Date	Cases
19 май 2021 г.	415 289,56
20 май 2021 г.	416 048,47
21 май 2021 г.	416 381,46
22 май 2021 г.	416 987,58
23 май 2021 г.	417 209,88
24 май 2021 г.	417 328,15
25 май 2021 г.	417 968,52
26 май 2021 г.	418 445,57
27 май 2021 г.	419 052,85
28 май 2021 г.	419 358,01
29 май 2021 г.	419 898,58
30 май 2021 г.	420 091,80
31 май 2021 г.	420 162,07
01 юни 2021 г.	420 815,92
02 юни 2021 г.	421 267,86
03 юни 2021 г.	421 872,99
04 юни 2021 г.	422 198,62
05 юни 2021 г.	422 856,45
06 юни 2021 г.	423 075,64
07 юни 2021 г.	423 115,10
08 юни 2021 г.	423 897,72
09 юни 2021 г.	424 558,41
10 юни 2021 г.	425 083,31
11 юни 2021 г.	425 505,99
12 юни 2021 г.	426 102,16
13 юни 2021 г.	426 317,87
14 юни 2021 г.	426 341,19
15 юни 2021 г.	427 078,30
16 юни 2021 г.	427 711,45
17 юни 2021 г.	428 343,44
18 юни 2021 г.	428 800,99

l) Forecast (30 days)


Actual Data
Forecasts