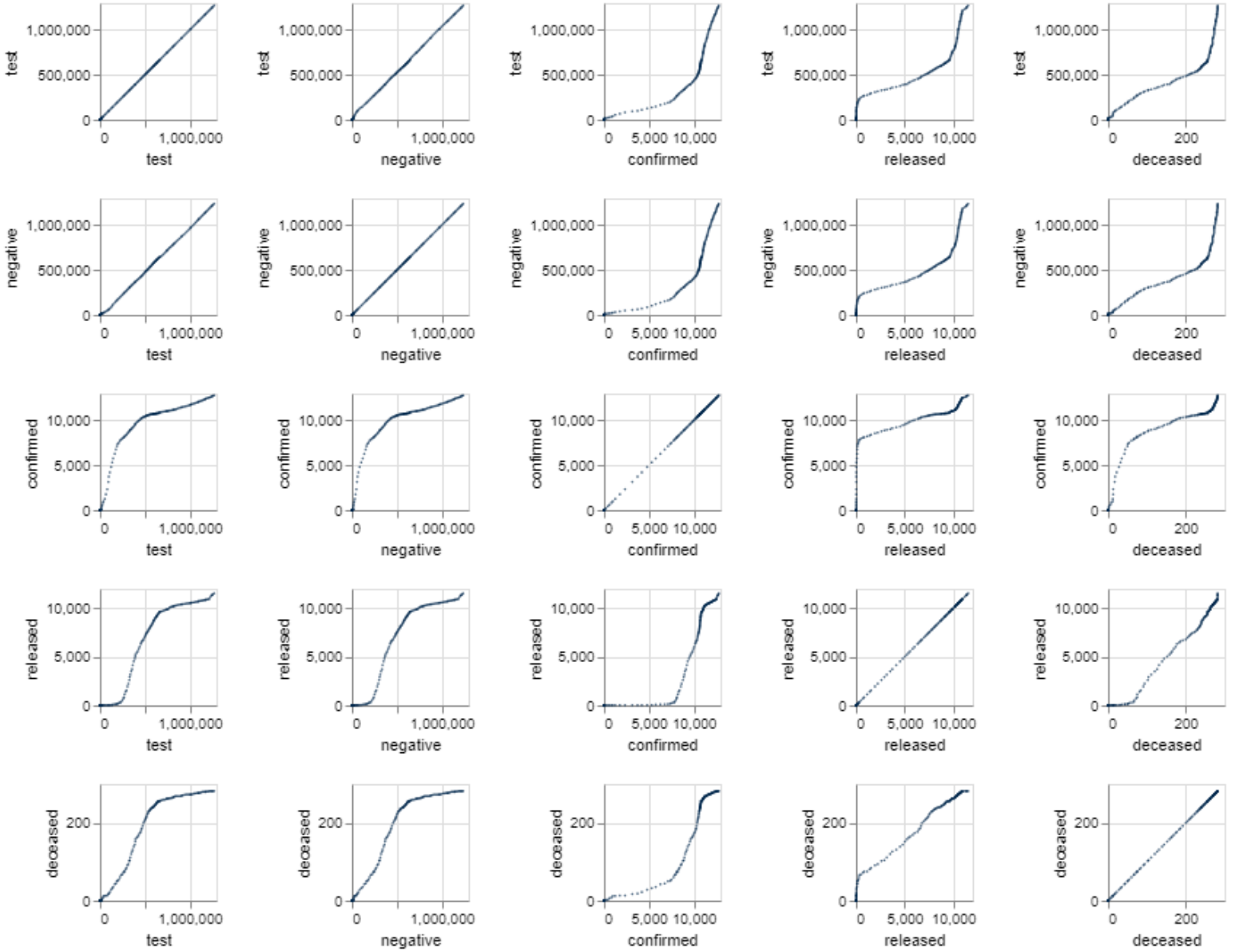
Live presentation

VISUAL DATA EXPLORATION

Dataset: Korea Center for Disease and Control data on Covid-19

Melissa Rebelo Kerezsi

Detect import variables

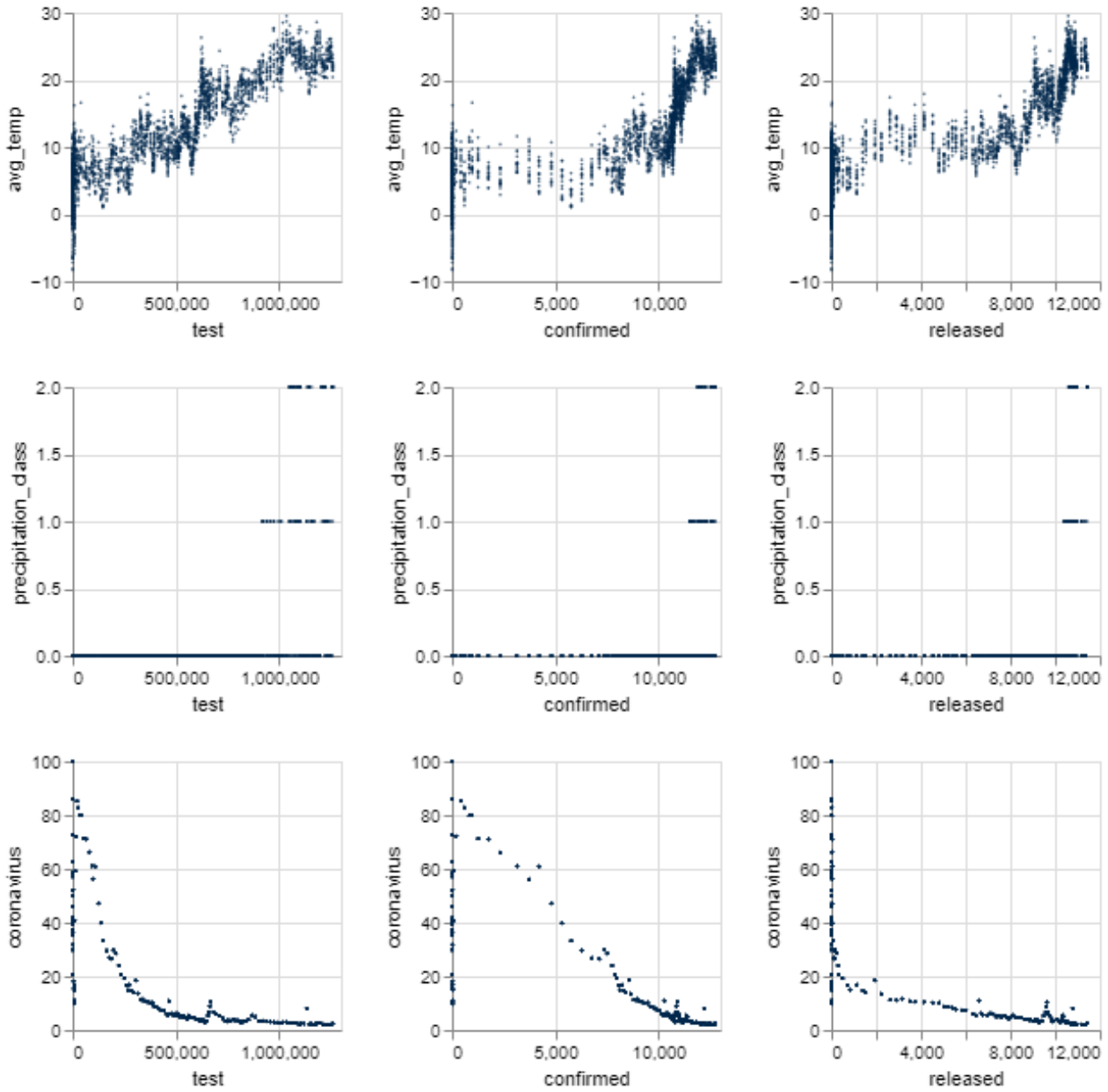


- Time dataset

Are there correlations we should be trying to understand?

Though need to be careful not to engage in p-hacking

Detect import variables



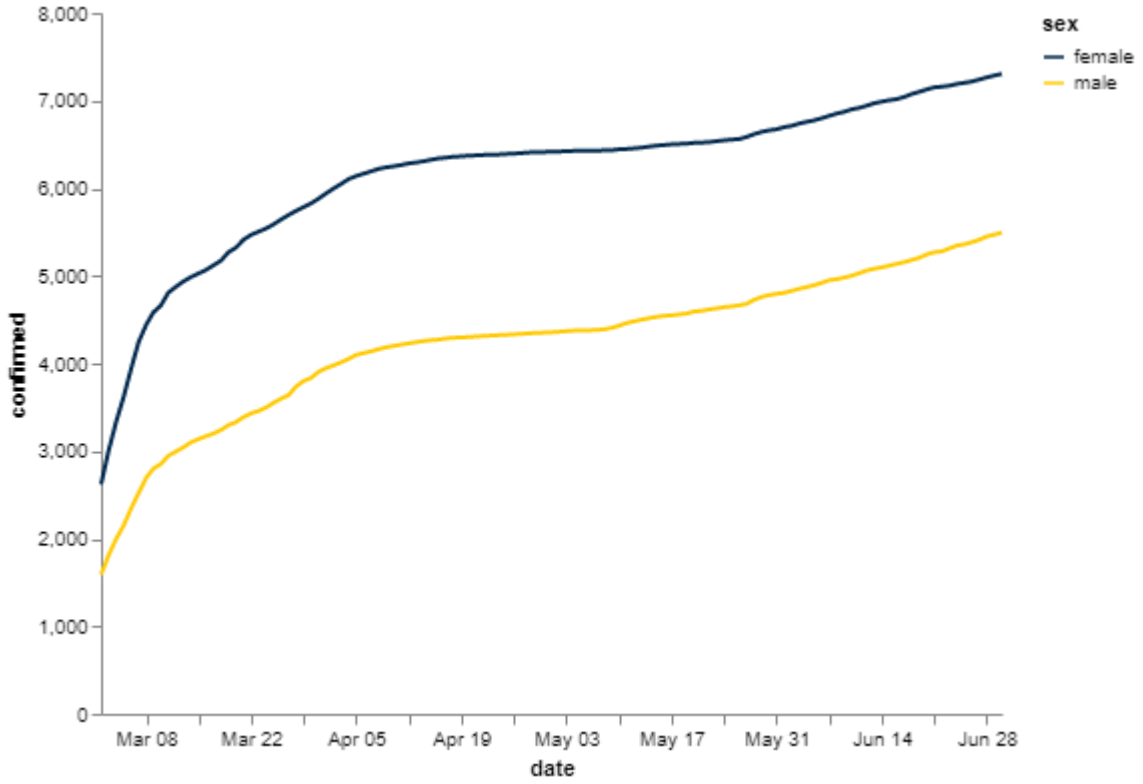
- Weather + Time + SearchTrend

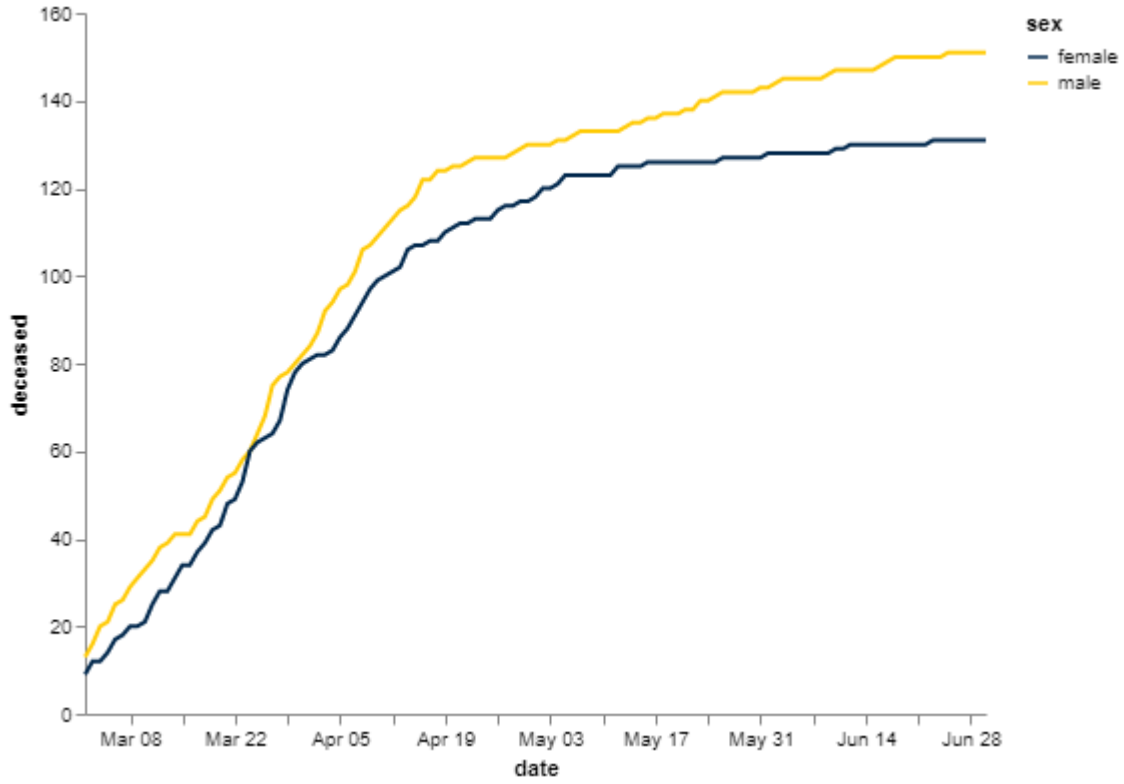
Avg_temp not very helpful unless we have data throughout seasons

Get insights

Time Gender

	date	time	sex	confirmed	deceased
0	2020-03-02	0	male	1591	13
1	2020-03-02	0	female	2621	9
2	2020-03-03	0	male	1810	16

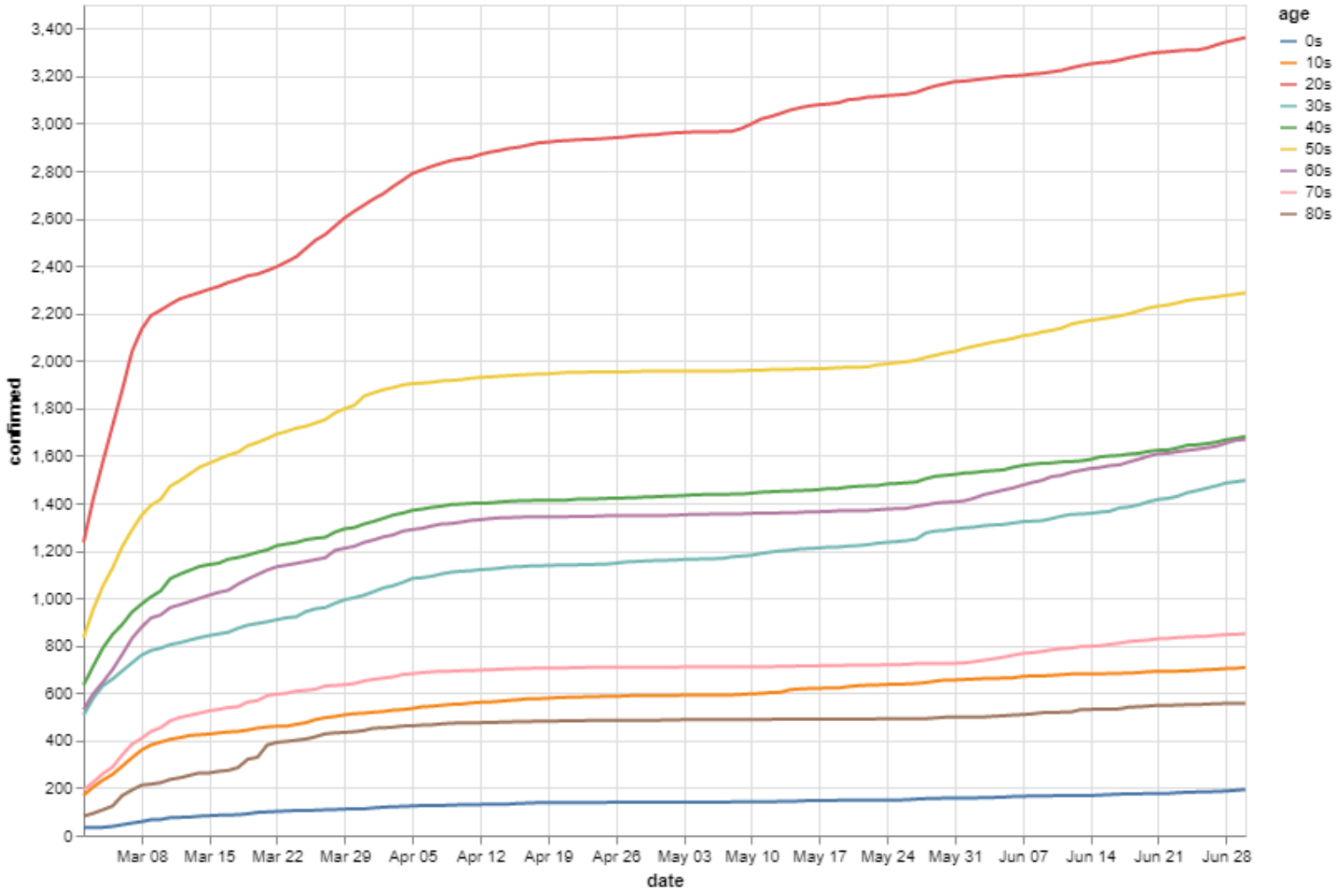




Test underlying assumptions

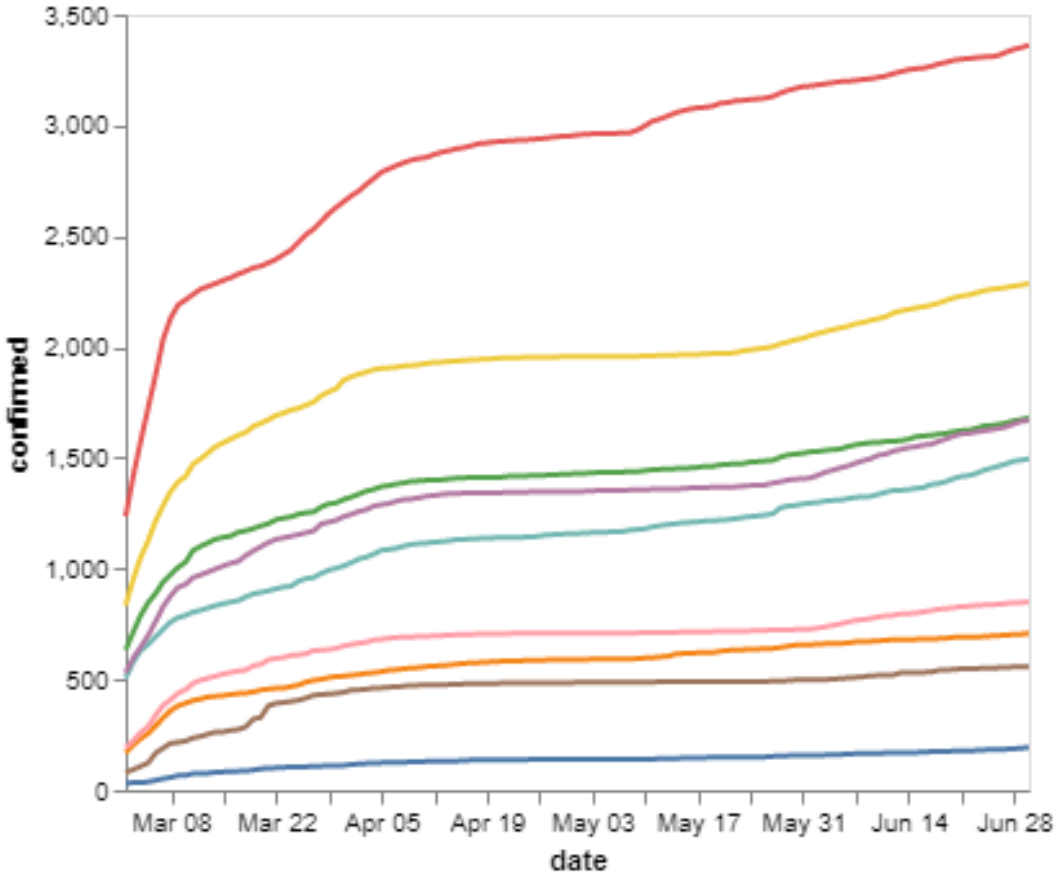
Time Age

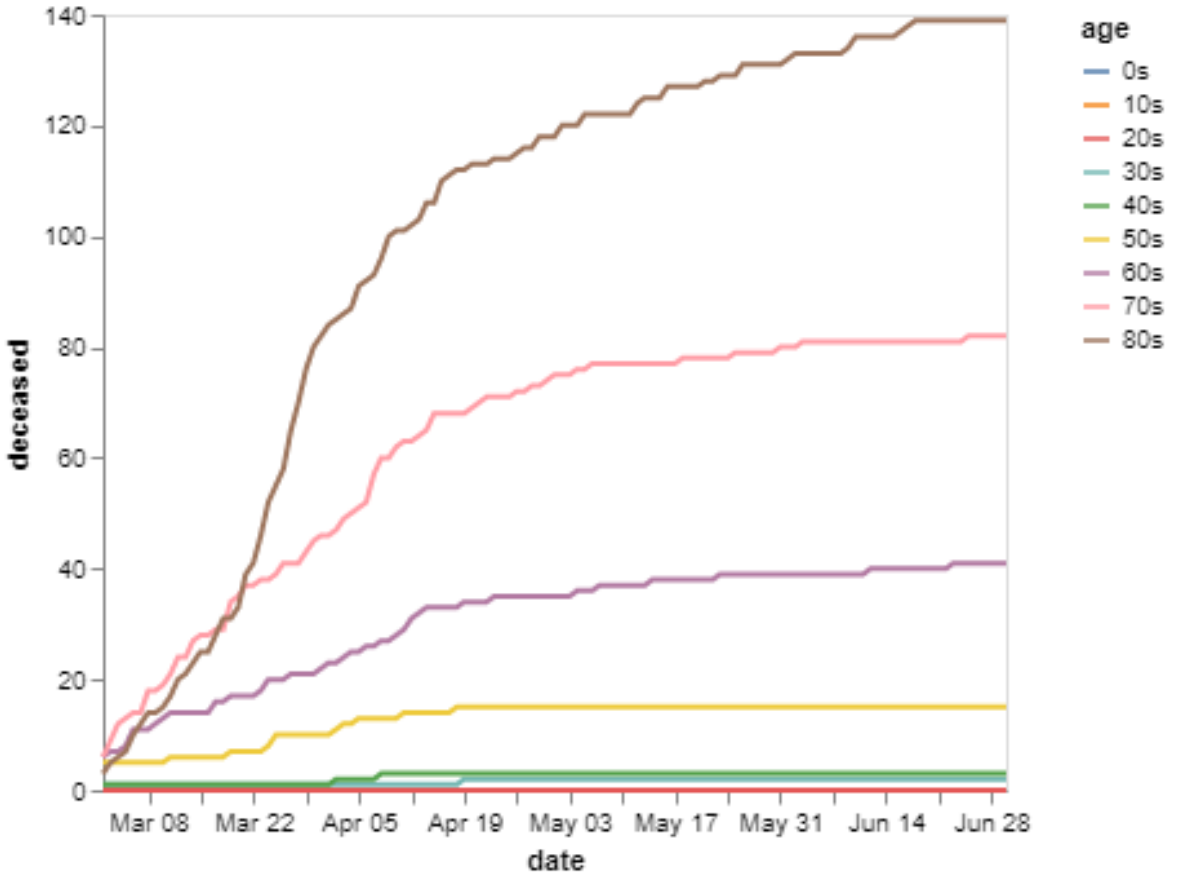
	date	time	age	confirmed	deceased
0	2020-03-02	0	0s	32	0
1	2020-03-02	0	10s	169	0
2	2020-03-02	0	20s	1235	0



Time Age

	date	time	age	confirmed	deceased
0	2020-03-02	0	0s	32	0
1	2020-03-02	0	10s	169	0
2	2020-03-02	0	20s	1235	0

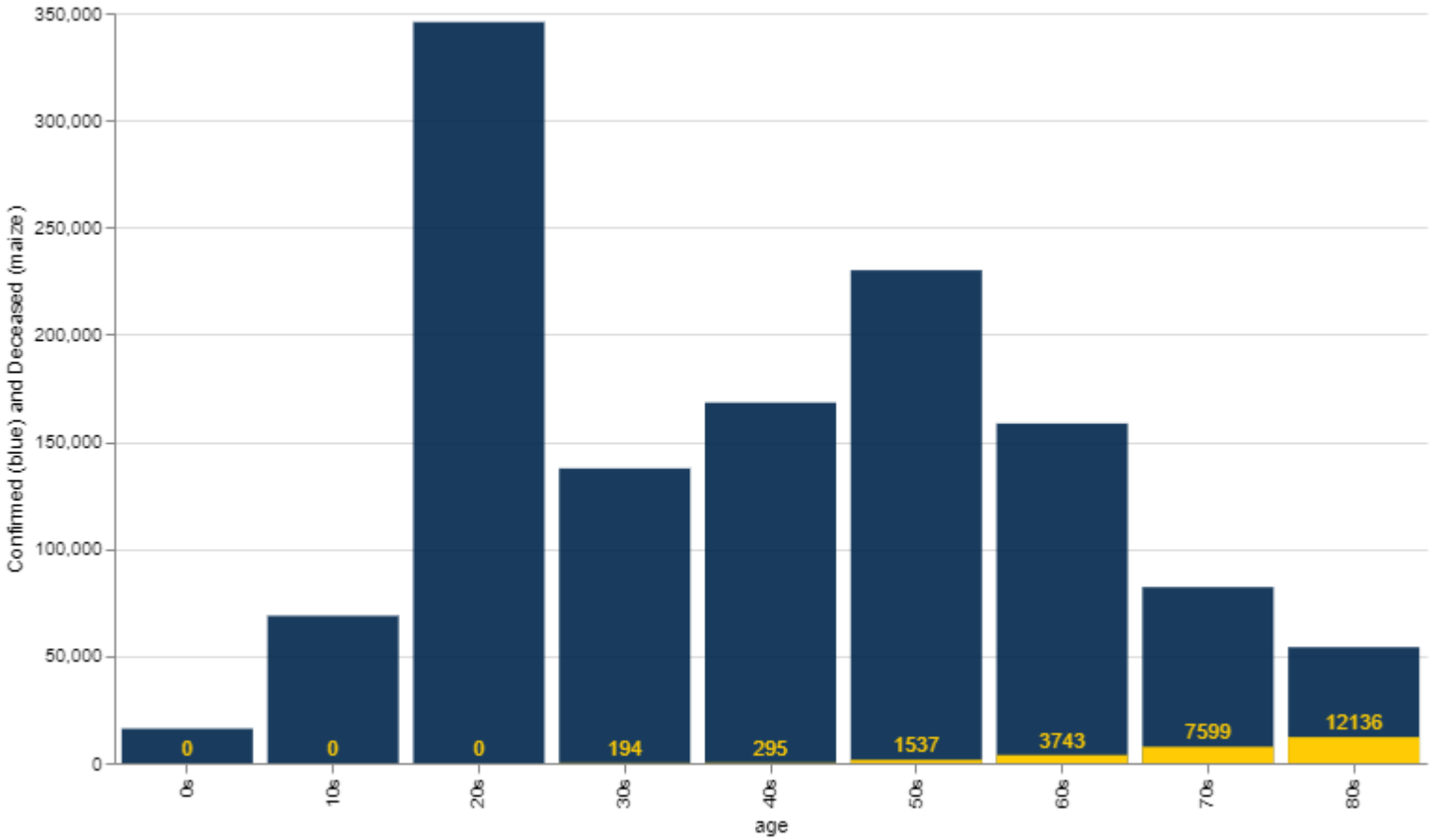




1. Basic visualization

Time Age

	date	time	age	confirmed	deceased
0	2020-03-02	0	0s	32	0
1	2020-03-02	0	10s	169	0
2	2020-03-02	0	20s	1235	0



Time_age.groupby('age').sum()

Time period:
[2020-03-02,
2020-06-30]

Time Age

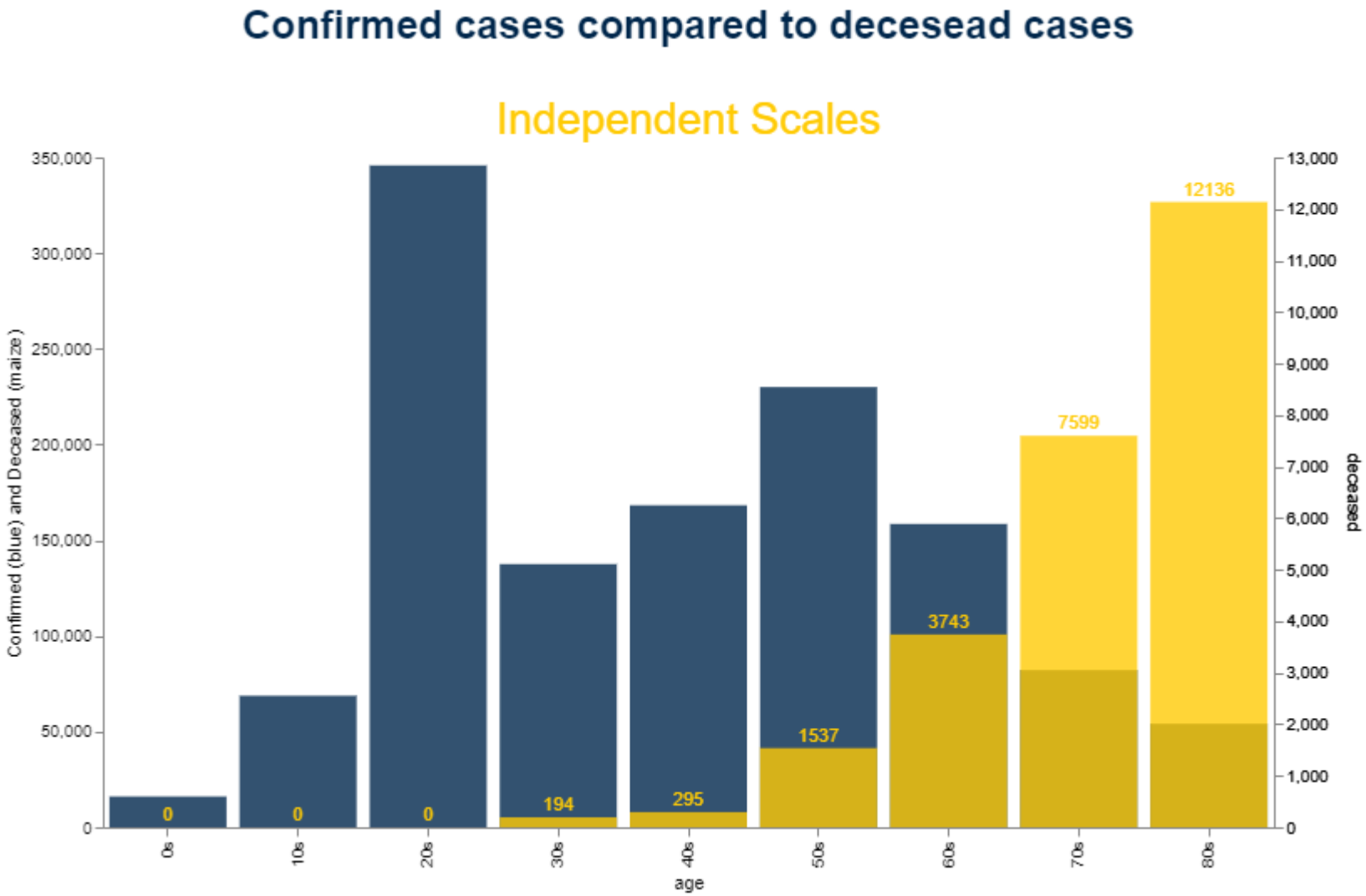
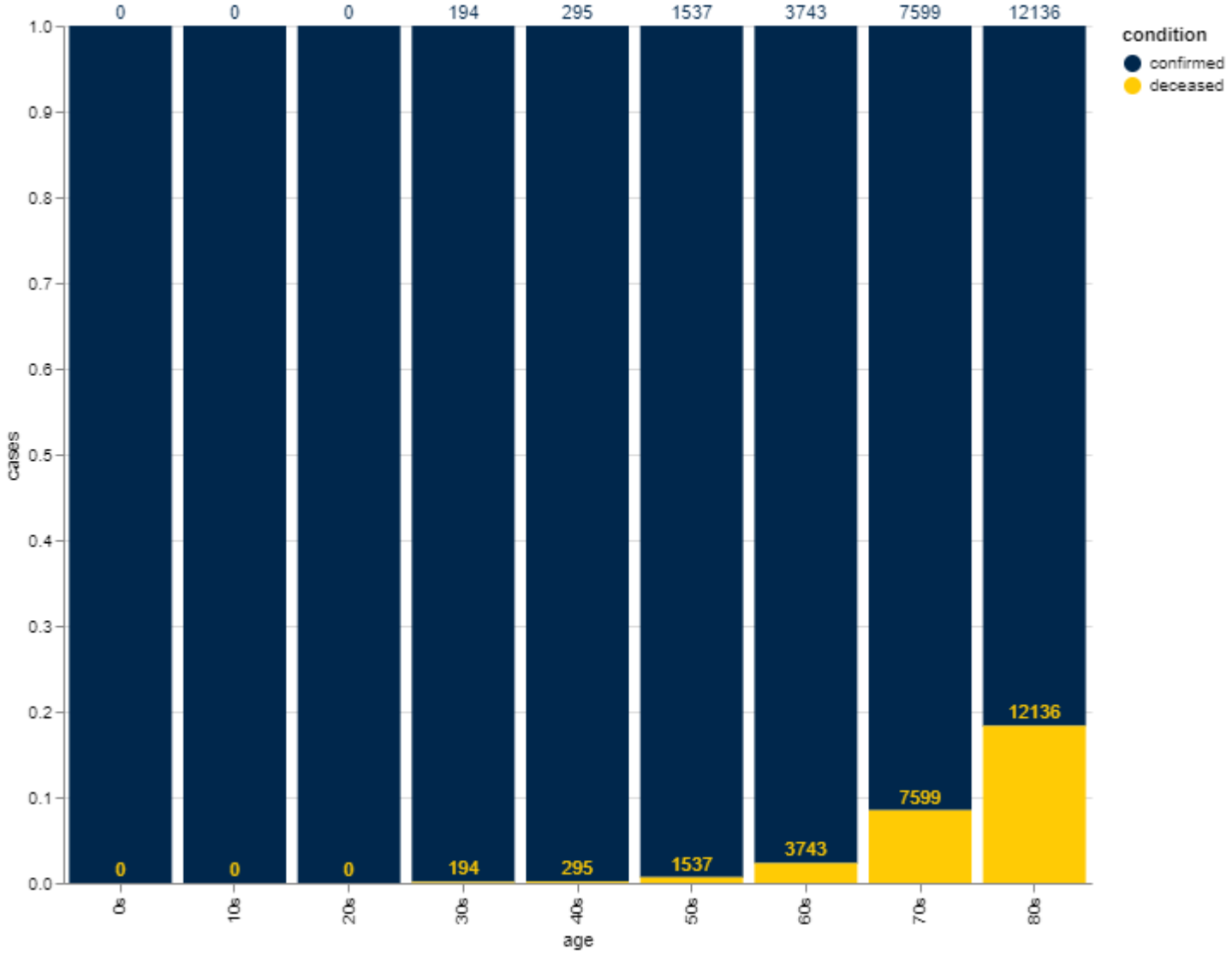


chart.resolve_scale(y='independent')

Time_age.groupby('age').sum()

Time period:
[2020-03-02,
2020-06-30]

Time Age




```
norm = alt.Chart(time_bar).mark_bar(color=blue_
).transform_fold(
    ['confirmed', 'deceased'],
    as_=['situation', 'cases']
).encode(
    alt.Y('cases:Q', stack='normalize'),
    alt.X('age'),
    alt.Color('situation:N',
    scale=alt.Scale(range=[blue_, maize_]))
)

dec_text = norm.mark_text(
    align='center', dy=-7, color=maize_
).encode(alt.Text('deceased'))

(norm+dec_text).configure_axis(grid=False
).configure_view(strokeWidth=0
).properties(width=500, height = 400
).configure_axis(titleFontWeight=100)
```

Time_age.groupby('age').sum()

Time period:
[2020-03-02,
2020-06-30]

2. Advanced visualization

choropleth

Case

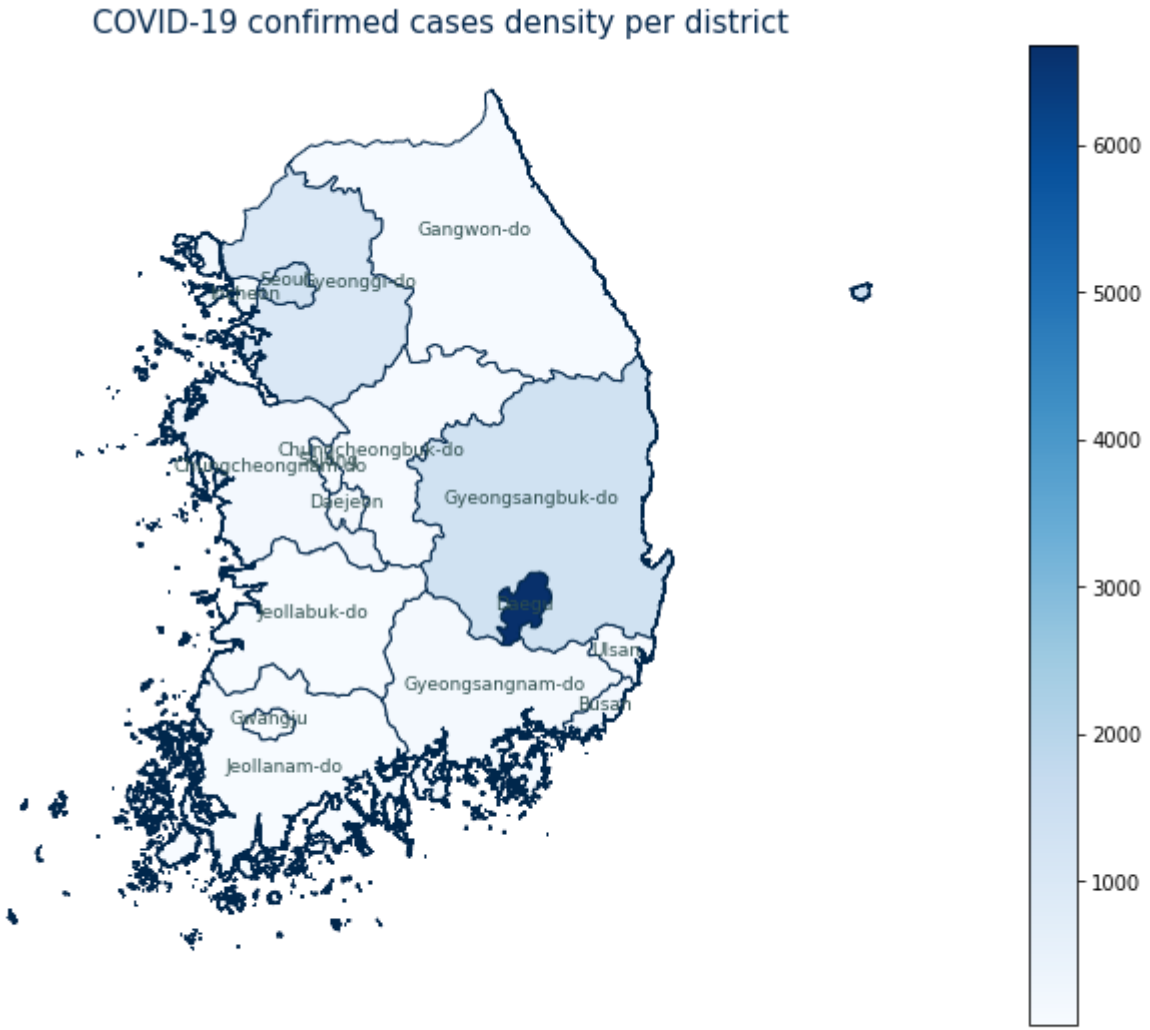
	case_id	province	city	group	infection_case	confirmed	latitude	longitude
0	1000001	Seoul	Yongsan-gu	True	Itaewon Clubs	139	37.538621	126.992652
1	1000002	Seoul	Gwanak-gu	True	Richway	119	37.48208	126.901384
2	1000003	Seoul	Guro-gu	True	Guro-gu Call Center	95	37.508163	126.884387

Visualize confirmed cases across districts

Detecting outliers

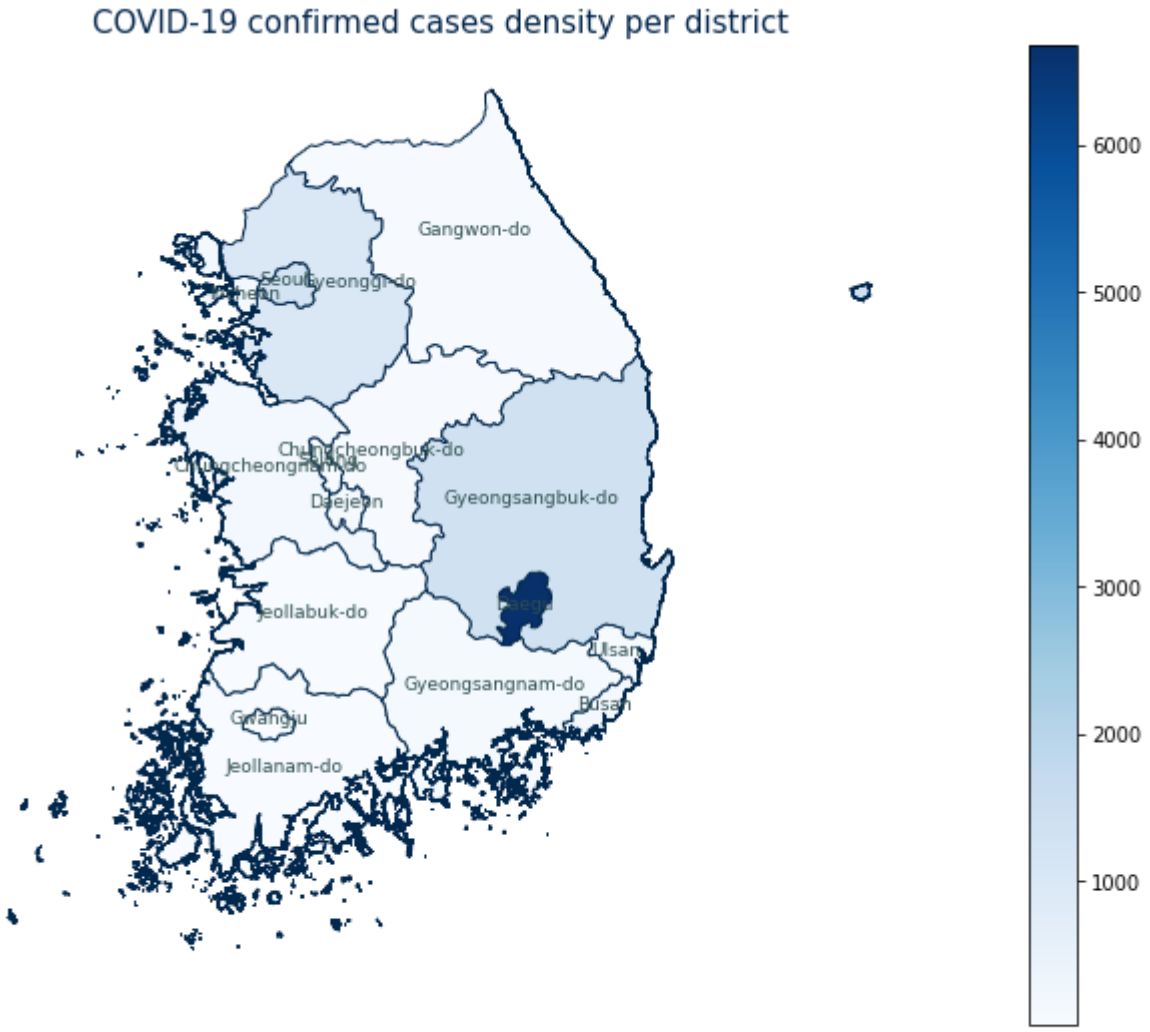
- Source a shapefile of districts of South Korea;
- Use geopandas to create choropleth

In this viz, we encode the number of confirmed cases as color saturation.
The darker the color, the more confirmed cases in a district.



- Source a shapefile of districts of South Korea;
- Use geopandas to create choropleth

In this viz, we encode the number of confirmed cases as color saturation.
The darker the color, the more confirmed cases in a district.

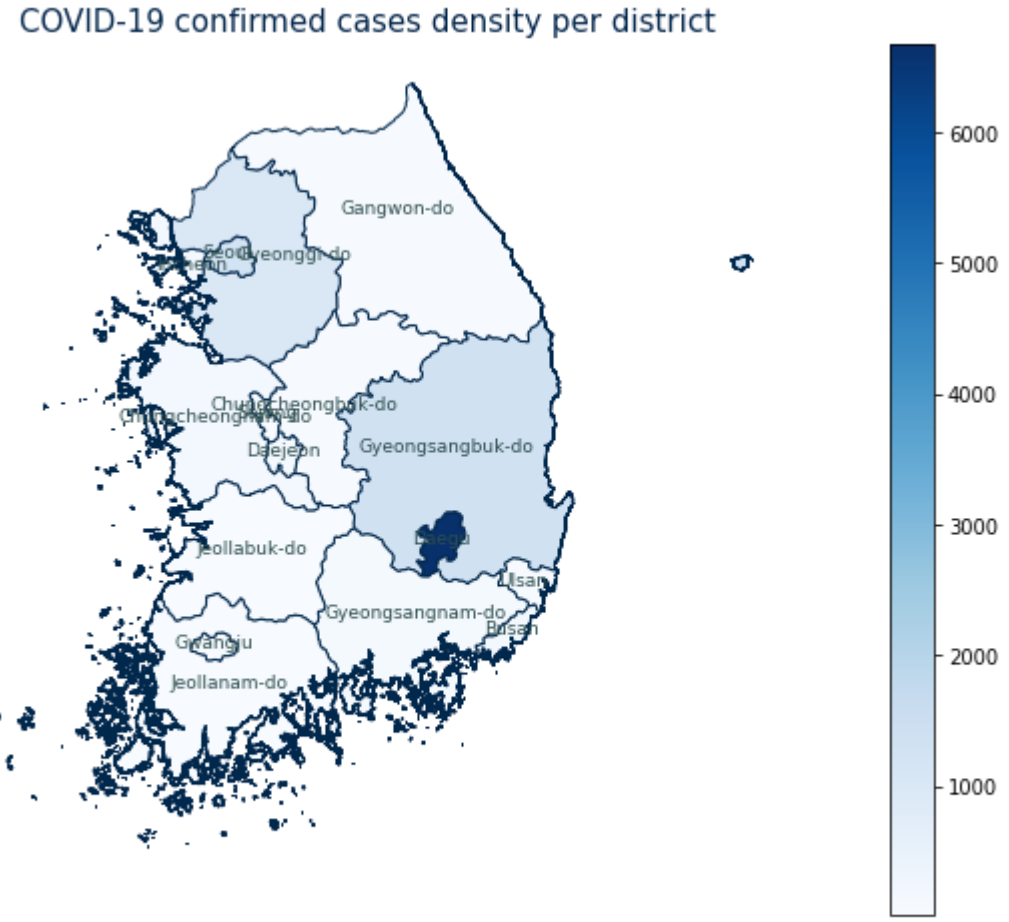



```
fp = '/datasets/capstone/591CE/Choropleth/data/KOR_adm1.shp'
map_df = gpd.GeoDataFrame.from_file(fp)
case_4merge = case.groupby('province')[['confirmed']].sum()
map_df = map_df.set_index('NAME_1').join(case_4merge).reset_index()
map_df.rename(columns={'index': 'province'}, inplace=True)
```

```
fig,ax=plt.subplots(1,figsize=(20,9))
ax.axis('off')

ax.annotate('Dataset: Case.csv',
            xy=(0.6, 0.05),
            xycoords='figure fraction',
            fontsize=12,
            color='#555555')

map_df['coords'] = (map_df['geometry']
                    .apply(lambda x: x.representative_point().coords[:]))
map_df['coords'] = [coords[0] for coords in map_df.coords]

for idx, row in map_df.iterrows():
    plt.annotate(text=row['NAME_1'], xy=row['coords'], fontsize=9,
                 color='darkslategrey', horizontalalignment='center')

map_df.plot(column = 'confirmed', cmap='Blues', ax=ax,
            linewidth=1, edgecolor=blue_, legend=True);
plt.title('COVID-19 confirmed cases density per district',
            fontdict={'fontsize':'15','fontweight':'3', 'color': blue_});
```

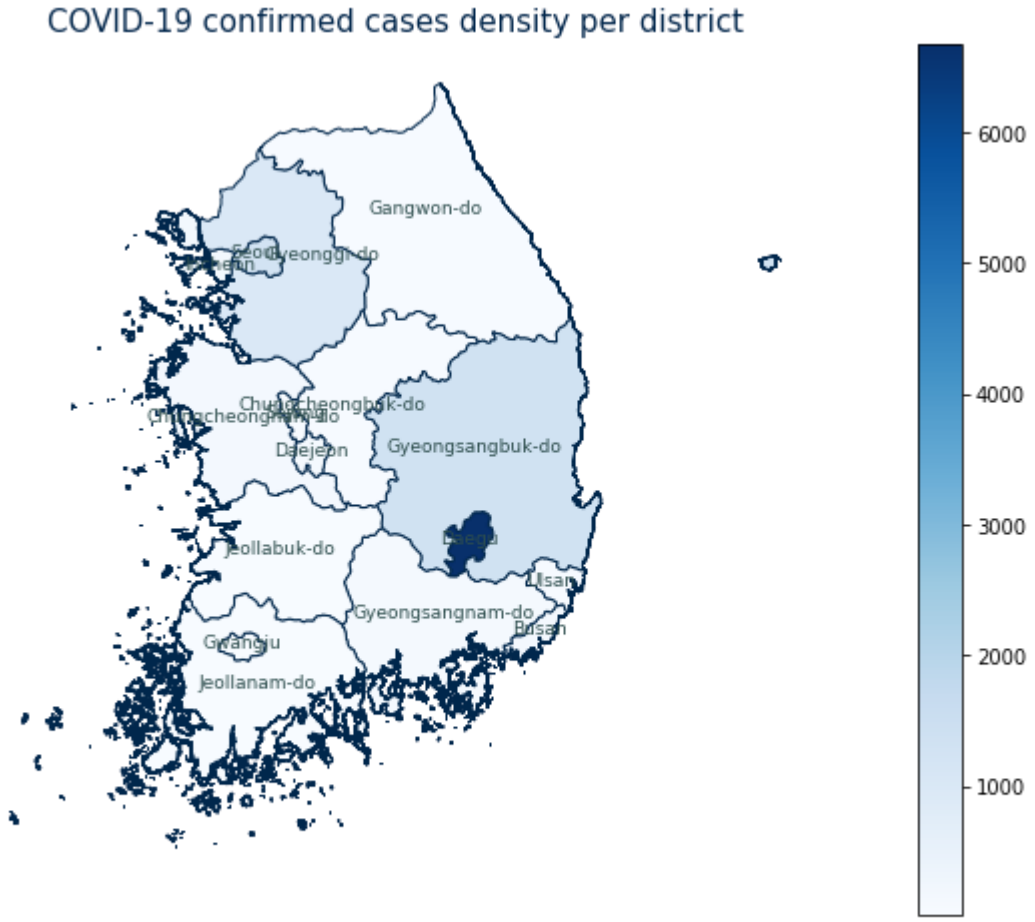
- **Expressiveness problem**

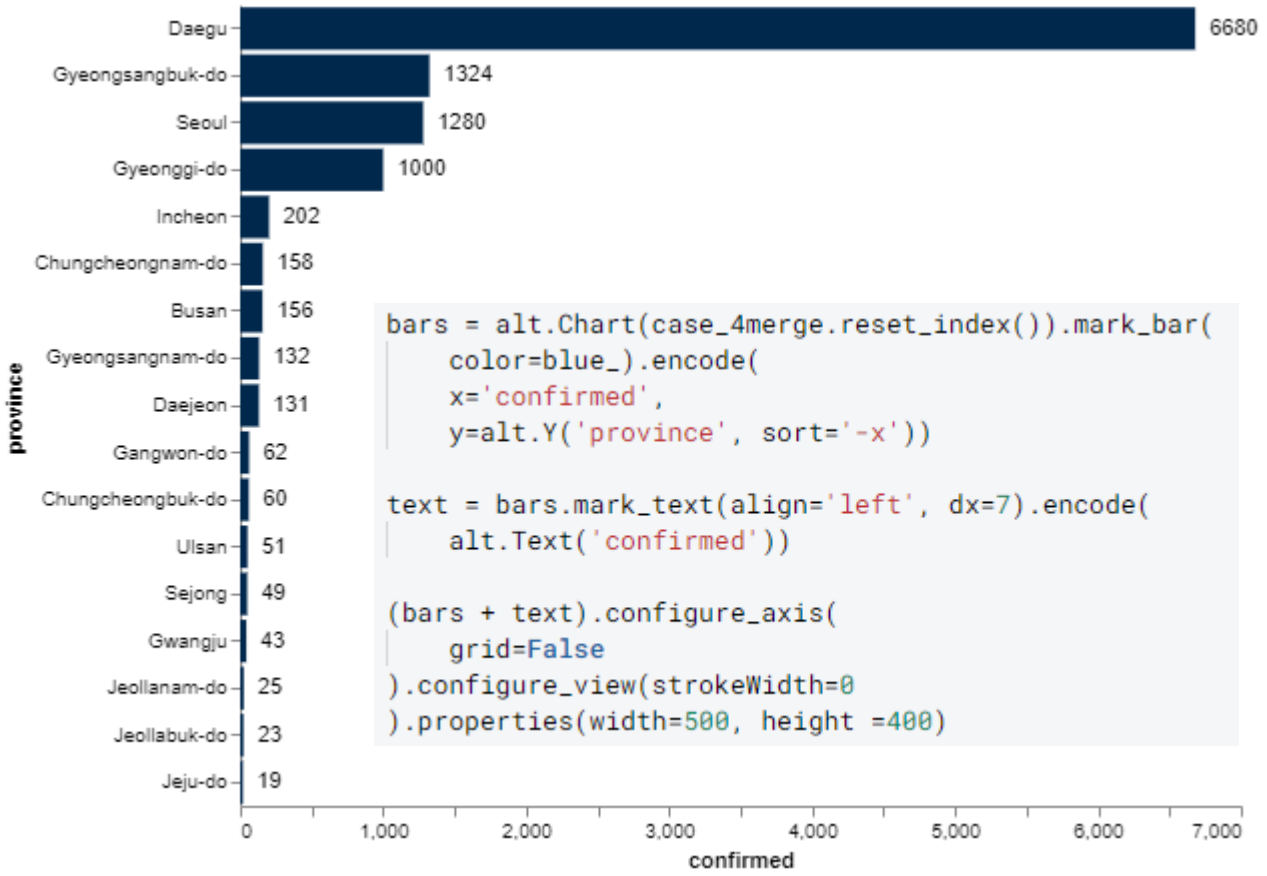
Range in number of confirmed cases is too large and resulting choropleth doesn't communicate much aside from the highest affected district.



Detect outliers and anomalies

A simple bar chart expresses the proportion of cases across districts.
It also shows the disparity in number of confirmed cases, explaining why a choropleth is not helpful in this situation.






```
bars = alt.Chart(case_4merge.reset_index()).mark_bar(
    color=blue_).encode(
    x='confirmed',
    y=alt.Y('province', sort='-x'))

text = bars.mark_text(align='left', dx=7).encode(
    alt.Text('confirmed'))

(bars + text).configure_axis(
    grid=False
).configure_view(strokeWidth=0
).properties(width=500, height =400)
```

2b. Advanced visualization

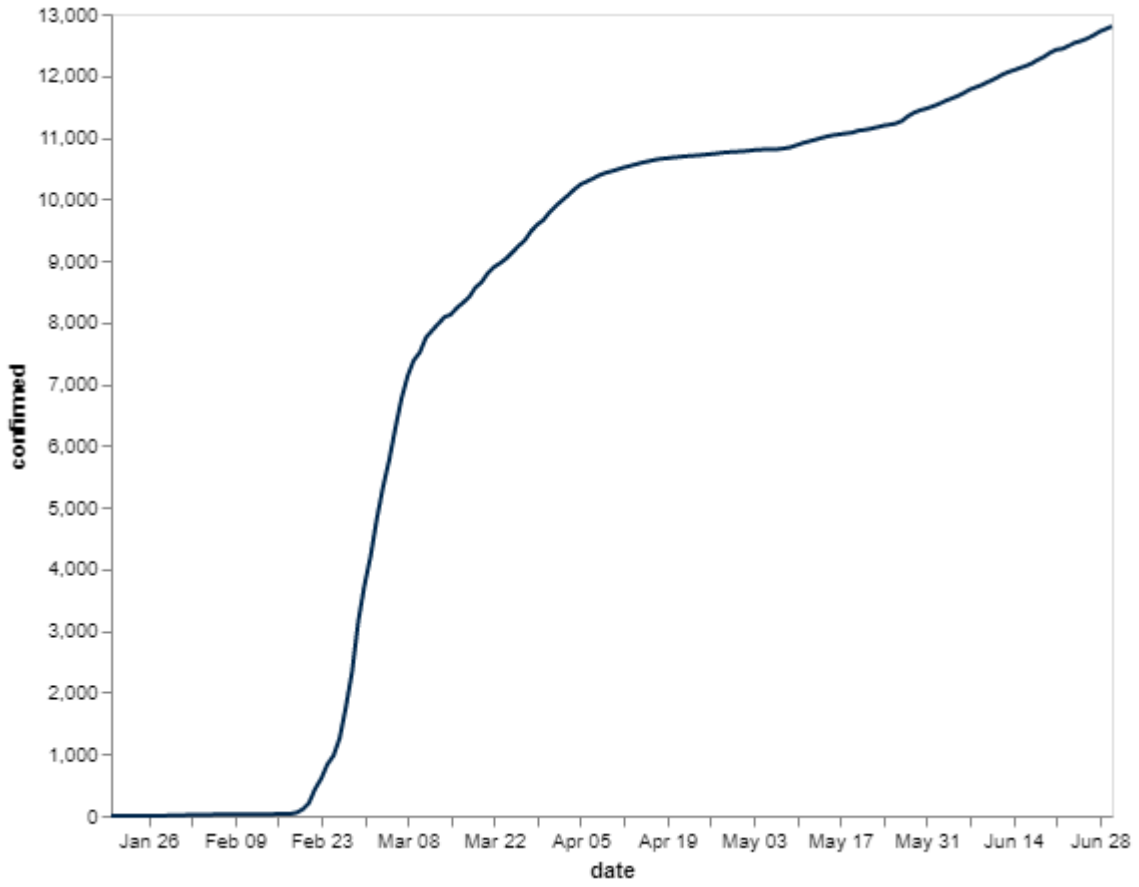
time series

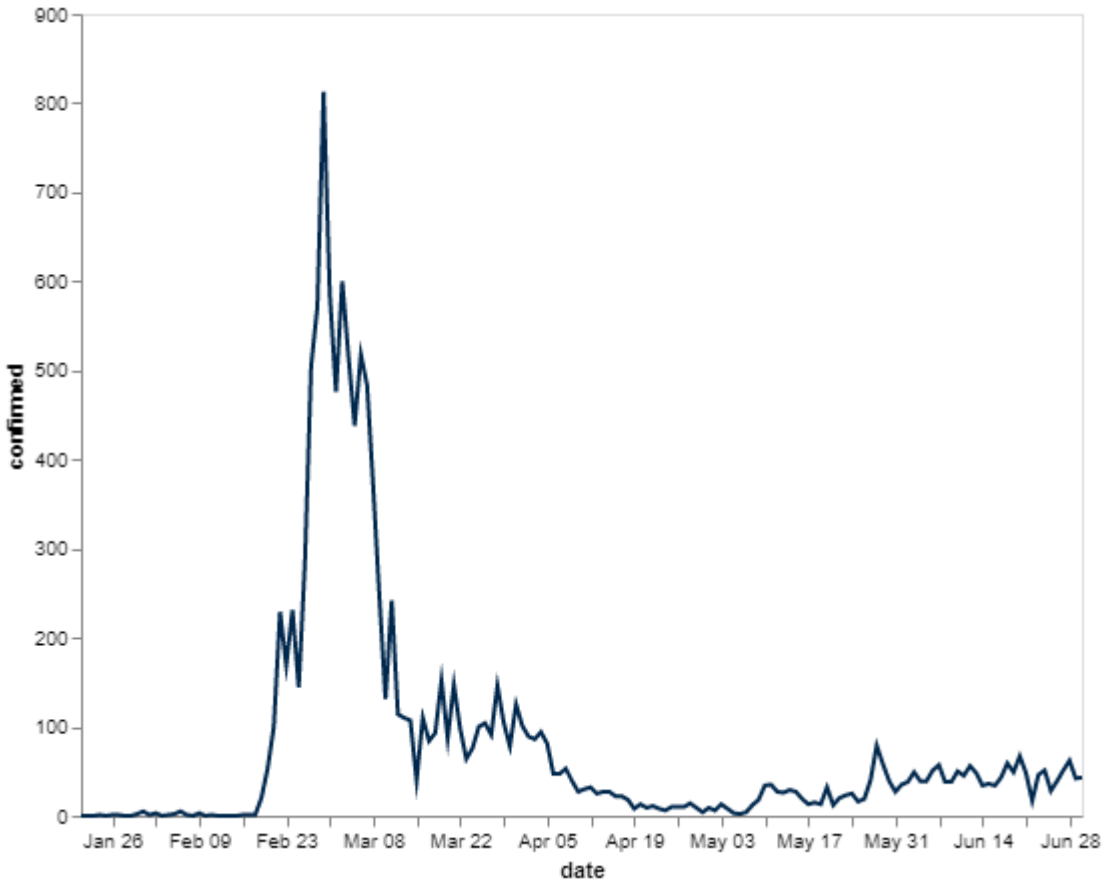
Time

	date	time	test	negative	confirmed	released	deceased
0	2020-01-20	16	1	0	1	0	0
1	2020-01-21	16	1	0	1	0	0
2	2020-01-22	16	4	3	1	0	0

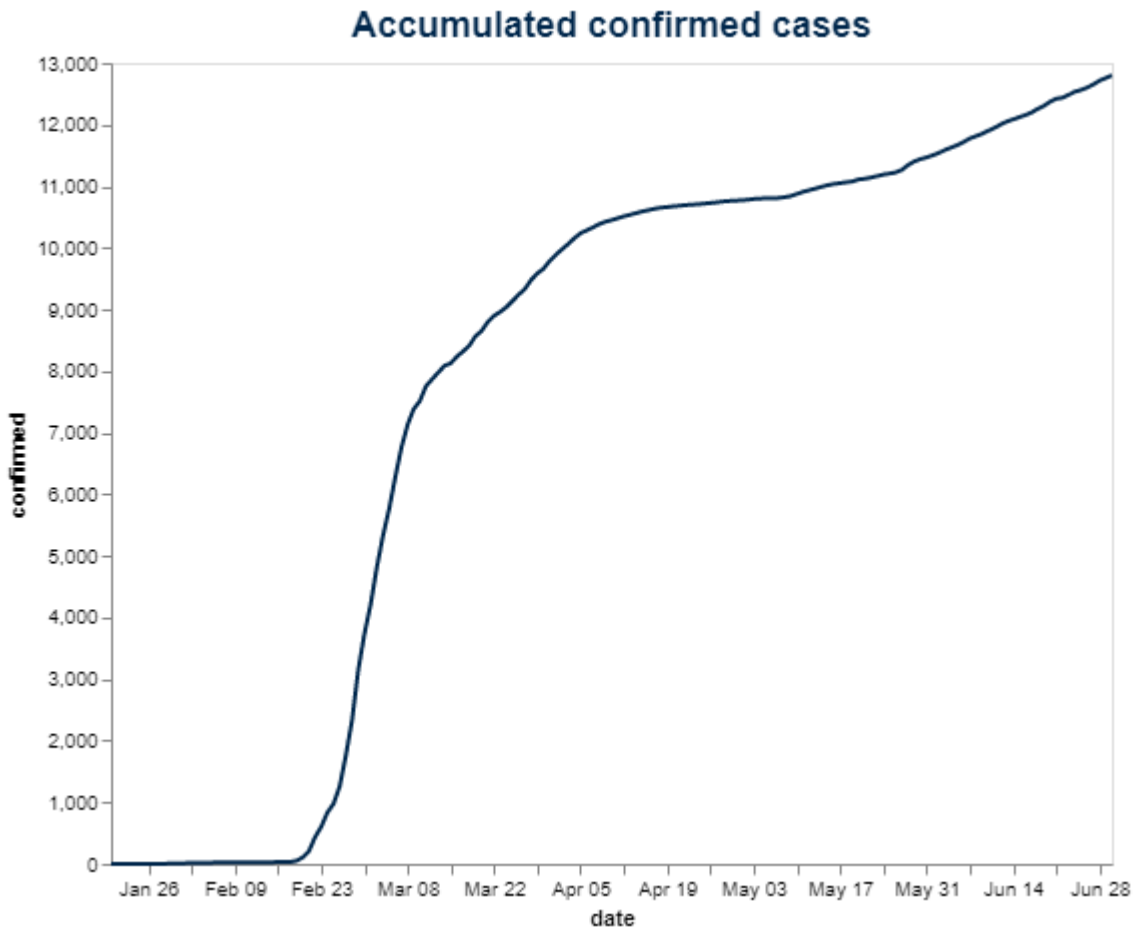
```
time_new = (
    time
    .set_index('date')
    .diff()
    .dropna()
    .reset_index()
    )
```

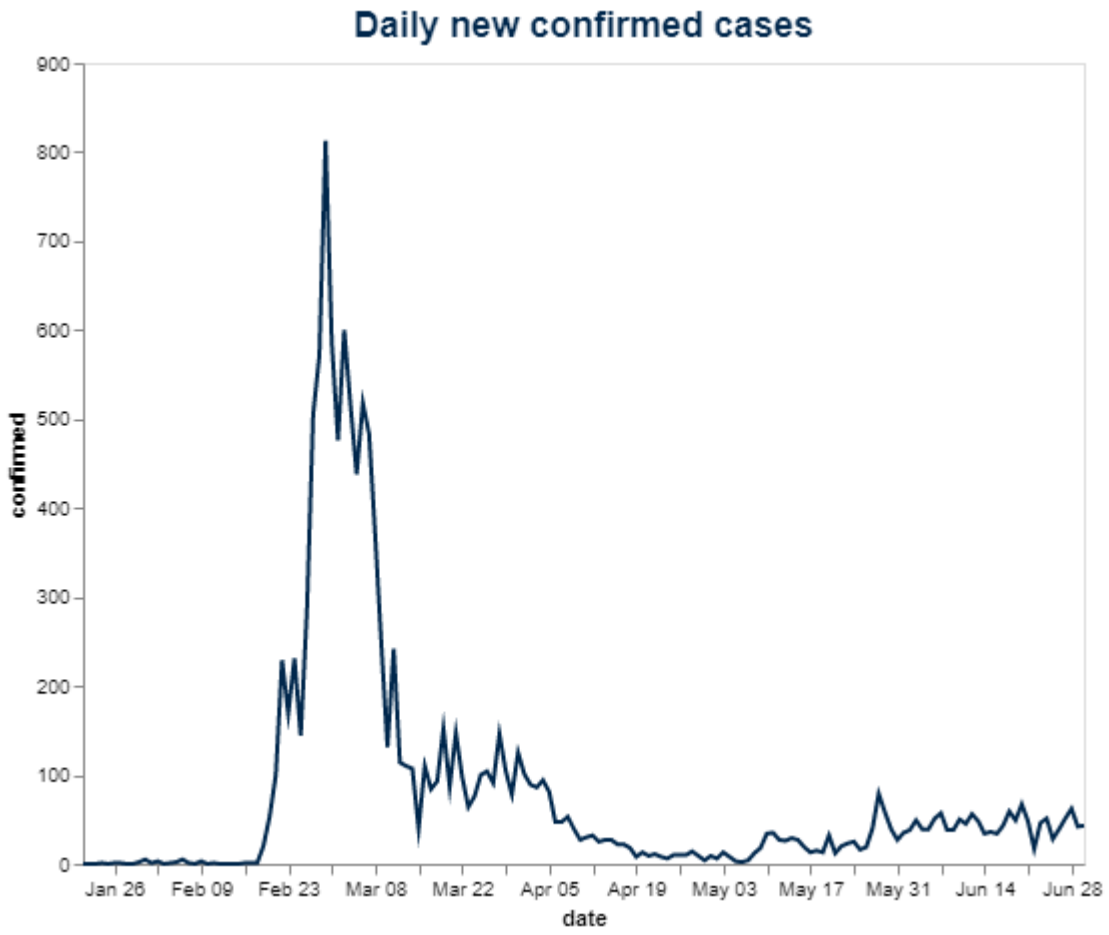
```
alt.Chart(time_new).mark_line(
    color=blue_
).encode(
    y='confirmed',
    x='date'
).properties(height=350, width=450
).configure_axis(grid=False)
```





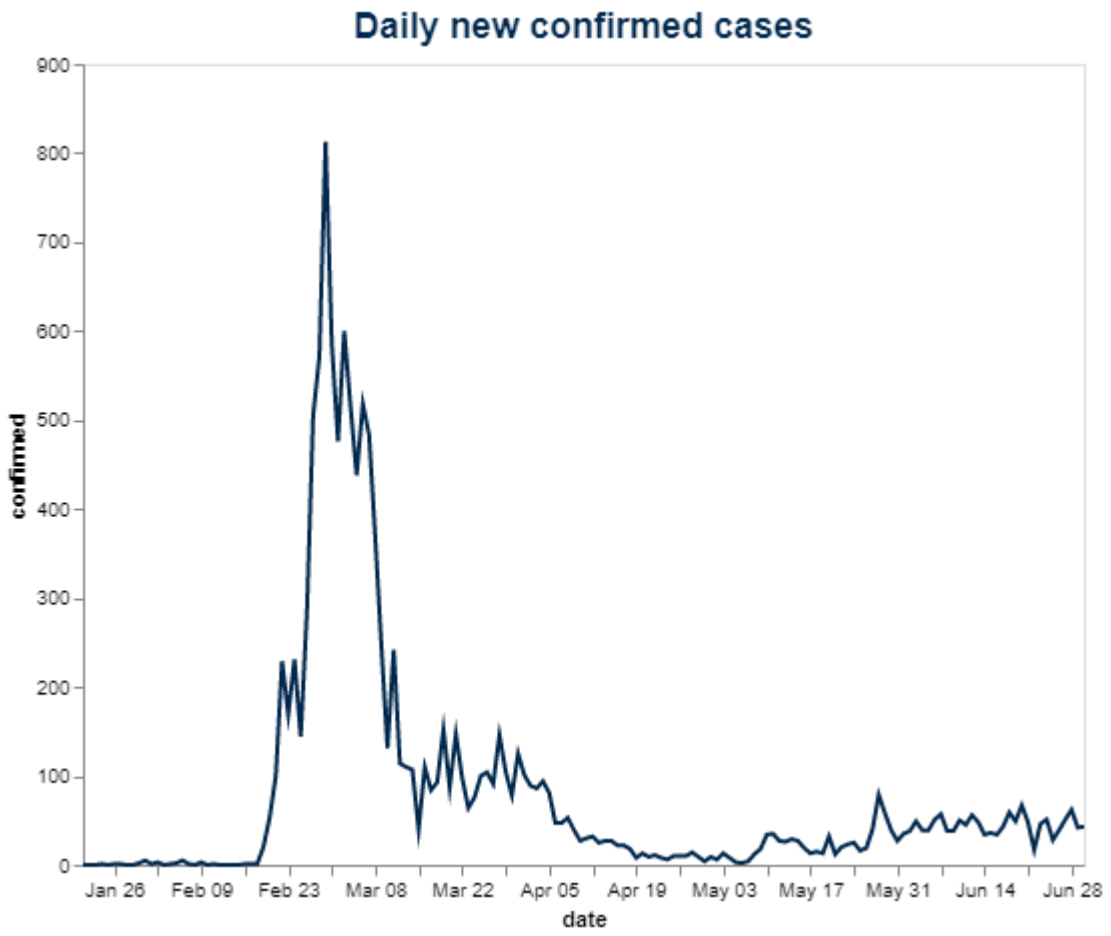
Time

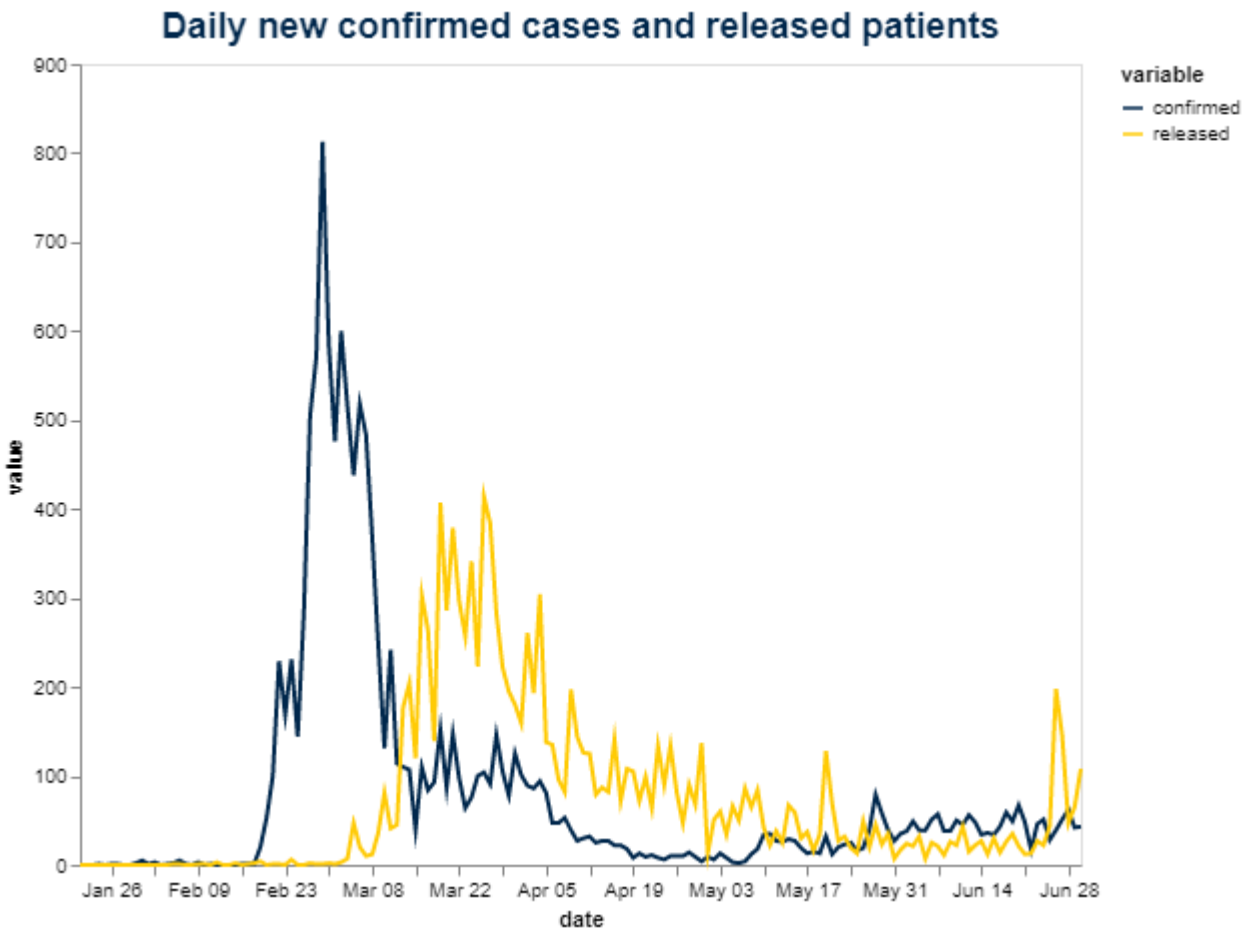

Accumulated confirmed cases
13,000
12,000
11,000
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
confirmed
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31
Jun 14
Jun 28
date


Daily new confirmed cases
900
800
700
600
500
400
300
200
100
0
confirmed
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31
Jun 14
Jun 28
date

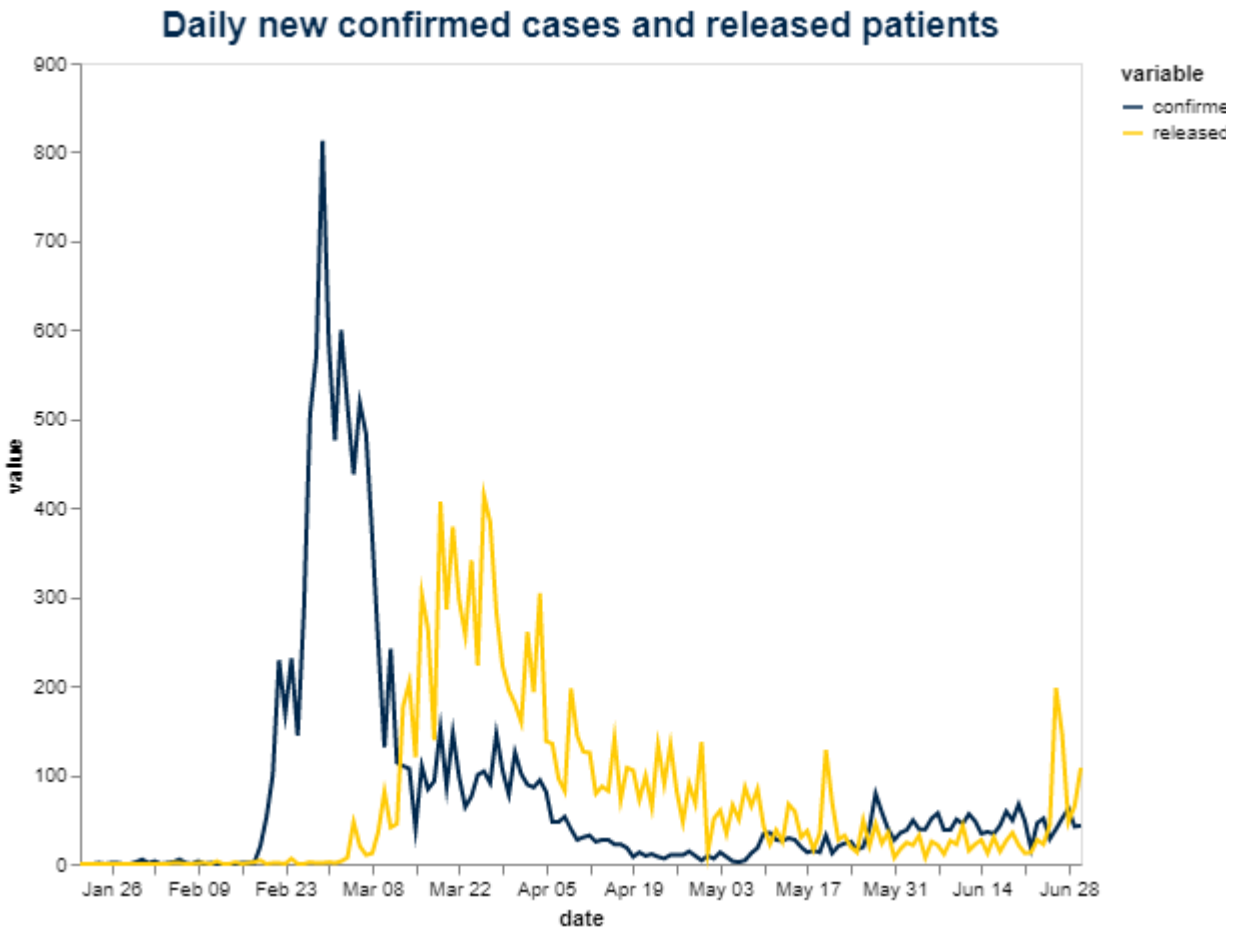
Time

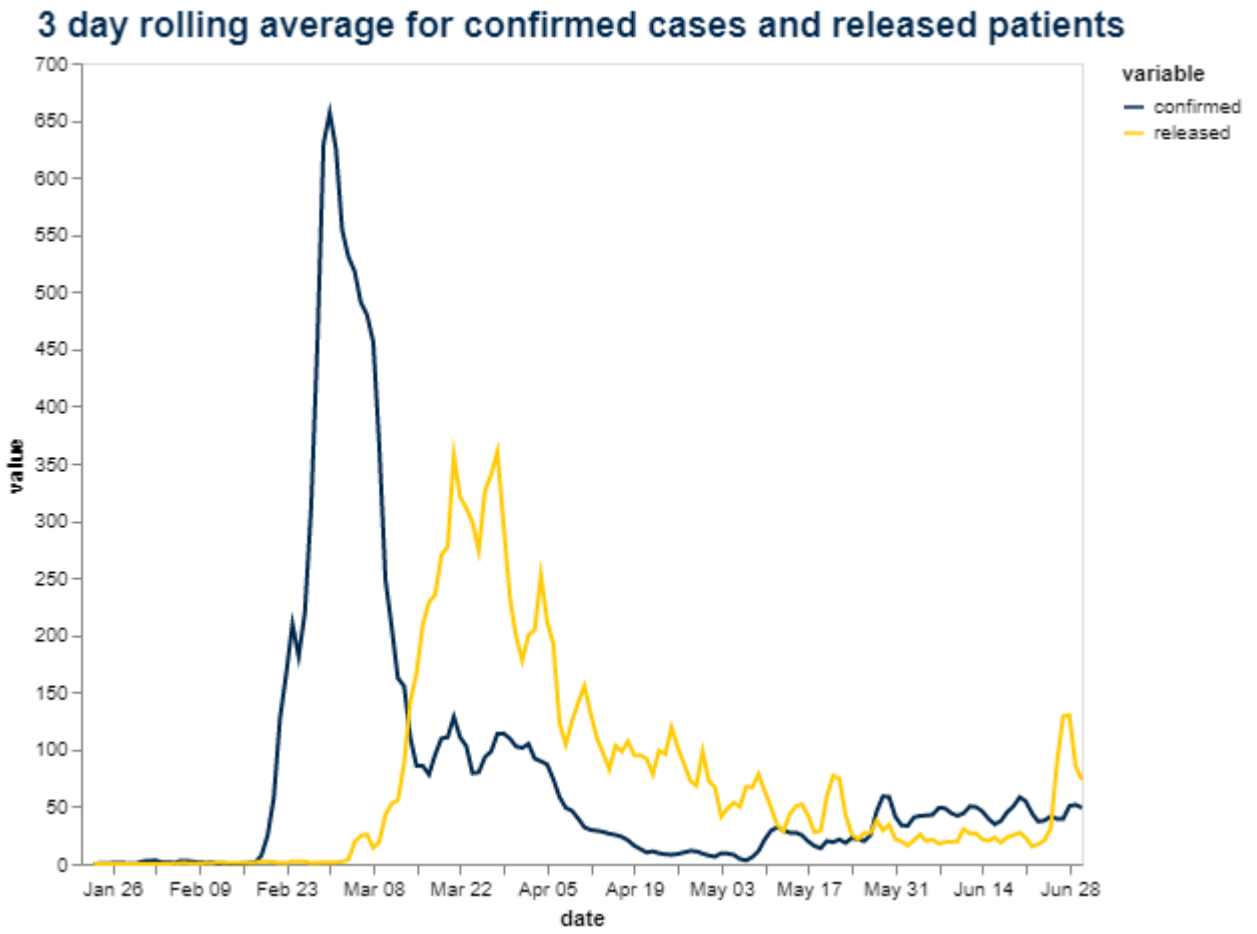
Uncover underlying structures





Time


Daily new confirmed cases and released patients
value
900
800
700
600
500
400
300
200
100
0
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31
Jun 14
Jun 28
date
variable
confirme
released


3 day rolling average for confirmed cases and released patients
value
700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31
Jun 14
Jun 28
date
variable
confirmed
released

Thank you!