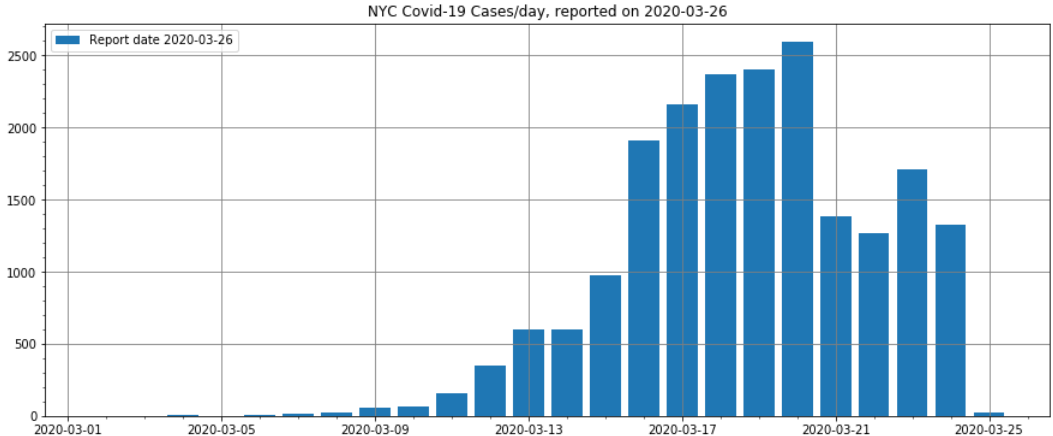
COVID-19: THE DATA ABUSE PANDEMIC

Harvey J. Stein
hjstein@bloomberg.net

Head, Quantitative Risk Analytics
Bloomberg L.P.

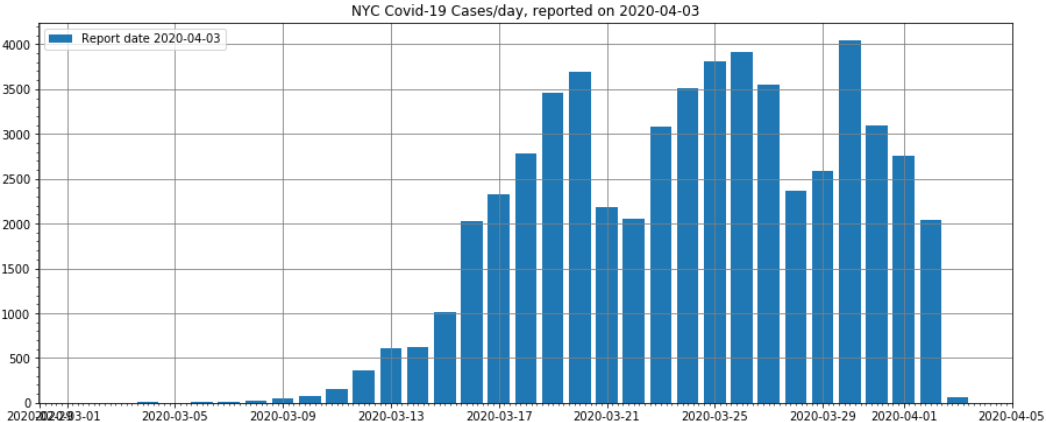
Bloomberg Quant Seminar
May, 2020
DRAFT SLIDES

WHAT THEY FIRST SHOWED IN NYC



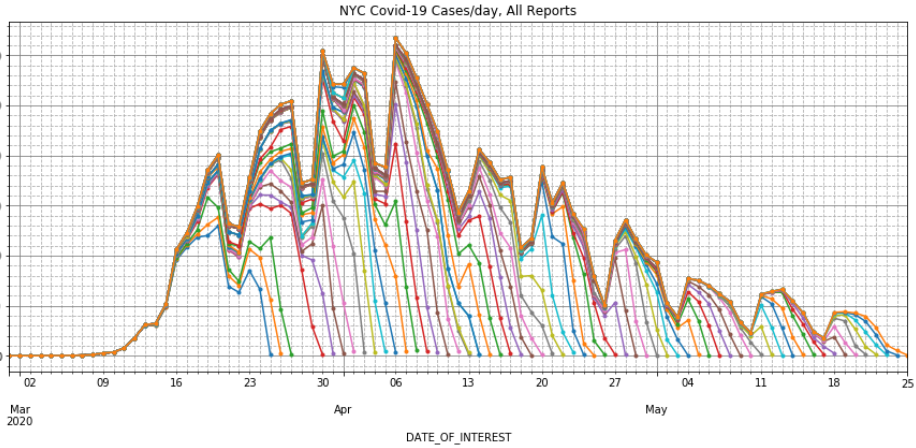
Looks like cases/day are dropping off after 3/20

EIGHT DAYS LATER



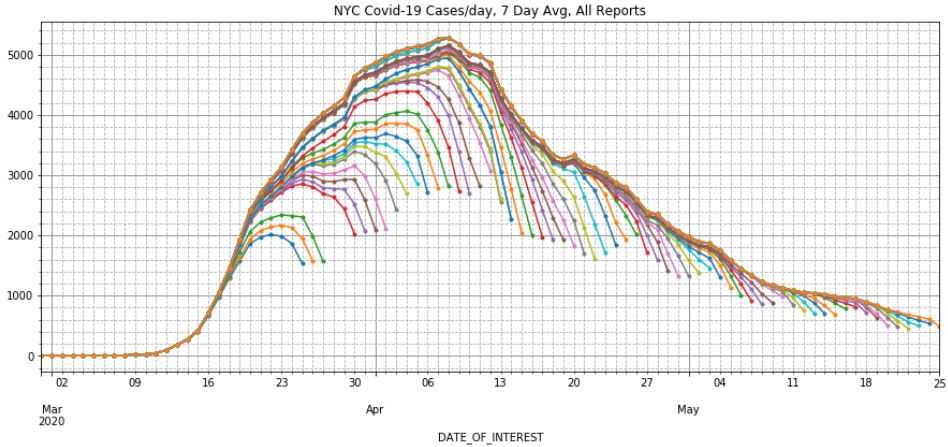
Now it looks like cases/day are rising through 3/30!

WHAT THEY DIDN'T SAY



Since April 12th, every report has peaked on April 6th

AFTER APPROPRIATE SMOOTHING



Since April 13th, every smoothed report has peaked on April 8th

ABUSES

What's going on?

- Different sites report cases at different times
- Large dip in confirmed cases every weekend

Impact:

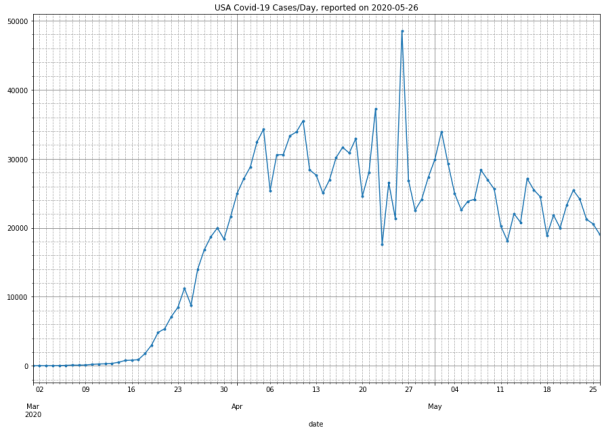
- Reporting delays can account for 30% changes as much as 3 weeks later
- Latest counts for each date are reported – history is harder to extract
- Low weekend counts obscure trends – need to use 7 day rolling average
- Reporting aggregates instead of per site data makes analysis difficult

But this is the good case!

See:

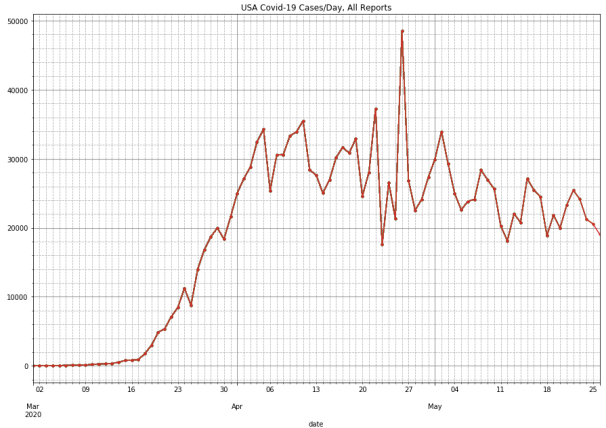
- *COVID-19 NYC Stats – A Ray Of Hope* [Ste20d]
- *COVID-19 NYC Stats – Not What They Seem* [Ste20e]

WHAT THEY SAID FOR THE USA



What happened to the weekend dip?

WHAT THEY DIDN'T SAY



Where are the data updates?

ABUSES

What's going on?

- Recording by report date instead of by incident date!

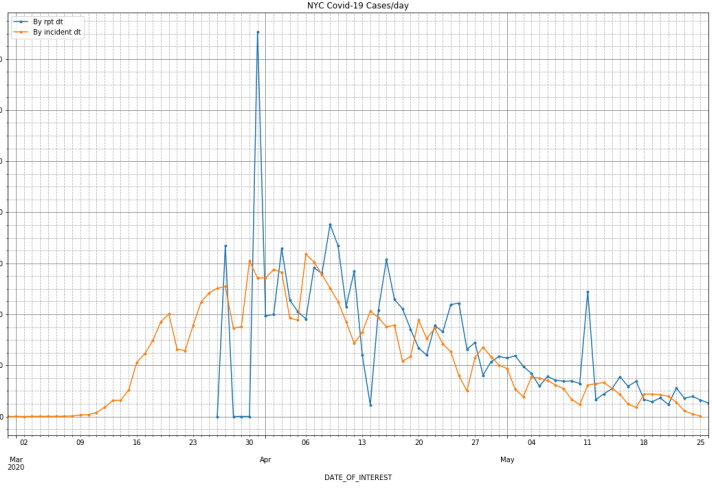
Impact:

- No record of when incidents actually occurred
- Misstates ramp up to peak, delays peak and overstates post-peak incidents
- Obscures periodicity and adds noise
- Makes modeling harder and more error prone
- News reports of yesterday's numbers are deceptive

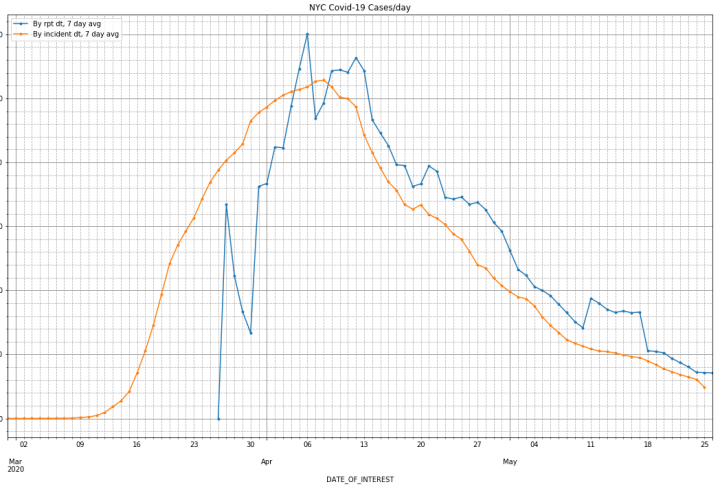
Recording by report date is rampant – all major sources do it:

- Our World in Data repository (data from the European CDC)
- Johns Hopkins University COVID-19 repository
- New York Times COVID-19 repository

IMPACT


NYC Covid-19 Cases/day
By rpt dt
By incident dt
14000
12000
10000
8000
6000
4000
2000
0
02
09
16
23
30
06
13
20
27
04
11
18
25
Mar
2020
Apr
May
DATE_OF_INTEREST

IMPACT


NYC Covid-19 Cases/day
By rpt dt, 7 day avg
By incident dt, 7 day avg
6000
5000
4000
3000
2000
1000
0
02
09
16
23
30
06
13
20
27
04
11
18
25
Mar
2020
Apr
May
DATE_OF_INTEREST

SUMMARY

What now?

- Know your data!
- Be aware that recent reports are inaccurate – either missing data or misrepresenting counts
- Account for this in your modeling
- Ask data sources to collect by incident date, publish full history and publish per site data instead of aggregating
- Investigate actual meaning of data – classification of COVID-19 hospitalizations, deaths due to COVID-19, ...

See COVID-19 Data Collection – Garbage In, Garbage Out [Ste20c])

References

[OWI20] OWID. *Data on COVID-19 (coronavirus) confirmed cases, deaths, and tests, All countries, Updated daily by Our World in Data*. 2020. URL: https://github.com/owid/covid-19-data.

[Ste20a] Harvey Stein. *Analysis of NYC COVID-19 infection rate*. Fork of https://github.com/nychealth/coronavirus-data. 2020. URL: https://github.com/hjstein/coronavirus-data.

[Ste20b] Harvey Stein. *Analysis of Our World of Data's COVID-19 dataset*. Fork of https://github.com/owid/covid-19-data. 2020. URL: https://github.com/hjstein/covid-19-data.

REFERENCES

[Ste20c] Harvey Stein. *COVID-19 Data Collection – Garbage In, Garbage Out*. Blog, Harvey J. Stein, Essays and commentary from a member of the quantitative community. Apr. 2020. URL: http://hjstein.blogspot.com/2020/05/covid-19-data-collection-garbage-in_33.html.

[Ste20d] Harvey Stein. *COVID-19 NYC Stats – A Ray Of Hope*. Blog, Harvey J. Stein, Essays and commentary from a member of the quantitative community. Apr. 2020. URL: https://hjstein.blogspot.com/2020/04/covid-19-nyc-stats-ray-of-hope.html.

[Ste20e] Harvey Stein. *COVID-19 NYC Stats – Not What They Seem*. Blog, Harvey J. Stein, Essays and commentary from a member of the quantitative community. Apr. 2020. URL: https://hjstein.blogspot.com/2020/04/covid-19-nyc-stats-not-what-they-seem.html.

REFERENCES

[Joh20] Johns Hopkins University. *Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE*. 2020. URL: https://github.com/CSSEGISandData/COVID-19.

[New20a] New York City. *NYC Health Coronavirus Data*. 2020. URL: https://github.com/nychealth/coronavirus-data.

[New20b] New York Times. *An ongoing repository of data on coronavirus cases and deaths in the U.S.*. 2020. URL: https://github.com/nytimes/covid-19-data.