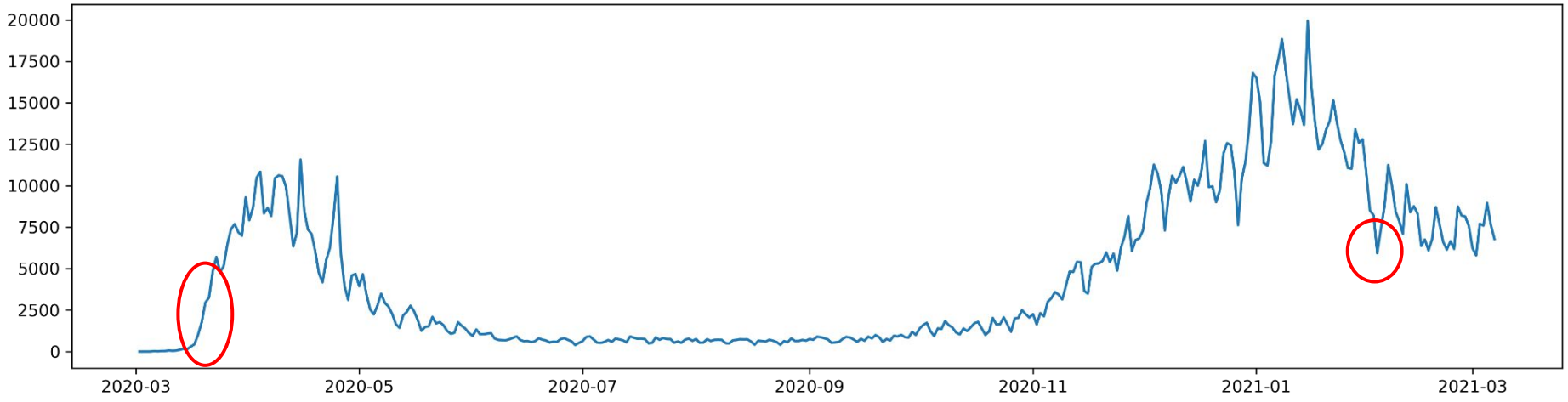
How can we identify dates of interest in the pandemic (COVID-19)?

Study done by:
Loh Zhi Heng
Wee Chang Han
Oh Zhi Hua

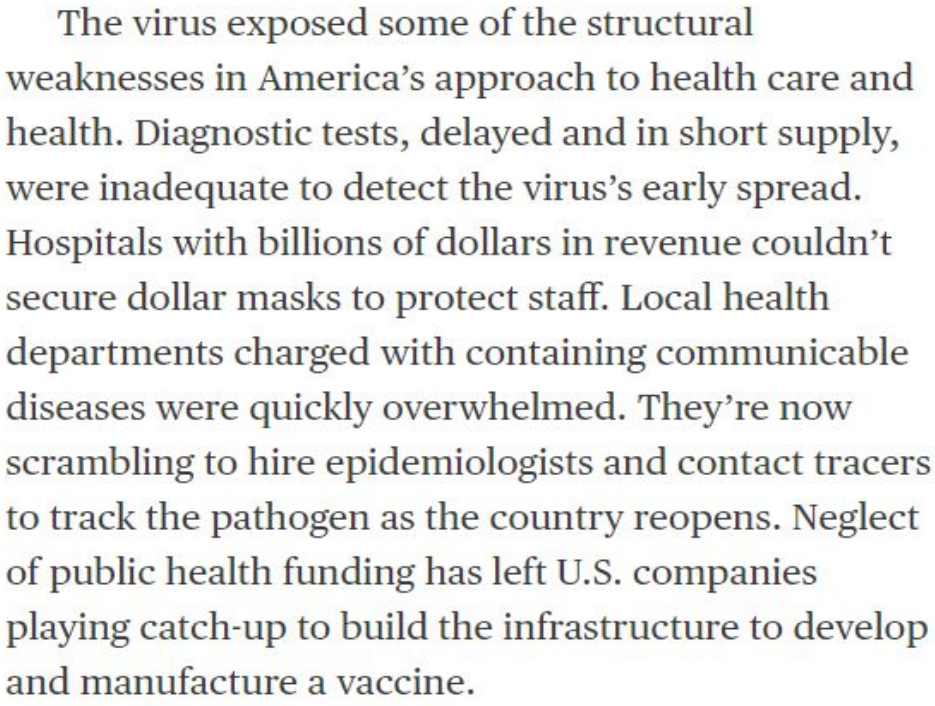
Dates of interest?

- Dates of interest could be days where there are surges or plunges in case numbers



Implementing a monitoring system to detect anomalies in the COVID-19 pandemic.

Authorities could use dates identified by the model to identify potential events that led to the anomaly.



The virus exposed some of the structural weaknesses in America's approach to health care and health. Diagnostic tests, delayed and in short supply, were inadequate to detect the virus's early spread. Hospitals with billions of dollars in revenue couldn't secure dollar masks to protect staff. Local health departments charged with containing communicable diseases were quickly overwhelmed. They're now scrambling to hire epidemiologists and contact tracers to track the pathogen as the country reopens. Neglect of public health funding has left U.S. companies playing catch-up to build the infrastructure to develop and manufacture a vaccine.

Figure 1: U.S. Health Care Puts $4 Trillion in All the Wrong Places, (Tozzi, 2020)

COVID-19

- Data collection
  - As of 7 March 2021, the Owners of the dataset has stopped collecting data.
  - Data is still accessible.
- Data curation / preparation
  - Deprecated columns are dropped.
  - Columns that only have values of ‘0’ are also dropped.
  - State column in New York Dataset has been also dropped.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 371 entries, 2021-03-07 to 2020-03-02
Data columns (total 8 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   deathIncrease                      371 non-null    int64
 1   hospitalizedIncrease               371 non-null    int64
 2   positive                           371 non-null    int64
 3   positiveIncrease                   371 non-null    int64
 4   totalTestEncountersViral           371 non-null    int64
 5   totalTestEncountersViralIncrease   371 non-null    int64
 6   totalTestResults                   371 non-null    int64
 7   totalTestResultsIncrease           371 non-null    int64
dtypes: int64(8)
memory usage: 26.1 KB
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 420 entries, 2021-03-07 to 2020-01-13
Data columns (total 16 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   death                      392 non-null    float64
 1   deathIncrease              420 non-null    int64
 2   inIcuCumulative            348 non-null    float64
 3   inIcuCurrently             347 non-null    float64
 4   hospitalizedIncrease       420 non-null    int64
 5   hospitalizedCurrently      356 non-null    float64
 6   hospitalizedCumulative     369 non-null    float64
 7   negative                   372 non-null    float64
 8   negativeIncrease           420 non-null    int64
 9   onVentilatorCumulative     341 non-null    float64
 10  onVentilatorCurrently      348 non-null    float64
 11  positive                   419 non-null    float64
 12  positiveIncrease           420 non-null    int64
 13  states                     420 non-null    int64
 14  totalTestResults           420 non-null    int64
 15  totalTestResultsIncrease   420 non-null    int64
dtypes: float64(9), int64(7)
memory usage: 55.8 KB
```

Figure 2: Data frame of New York (top) and National (bottom) Dataset after data preparation.

COVID-19

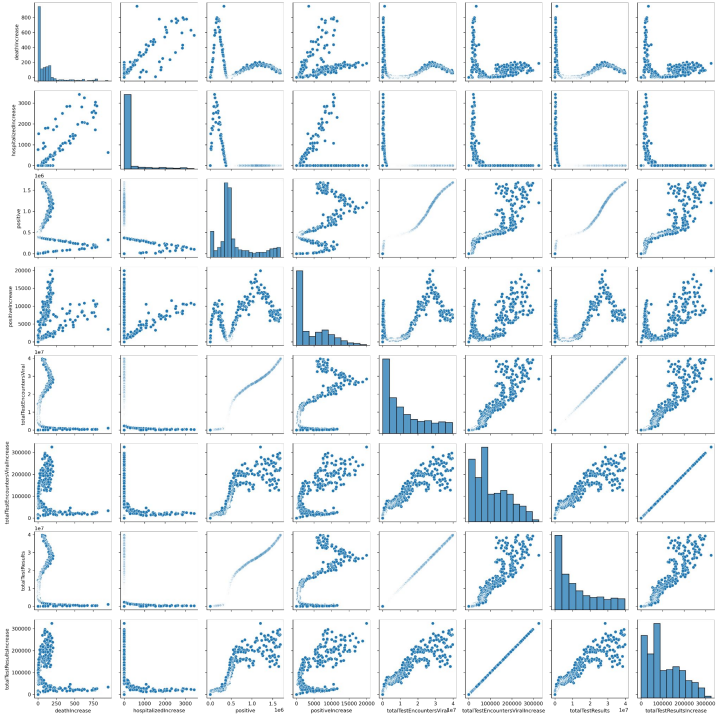


Figure 3: Pair plot of New York dataset.

Positive Increase

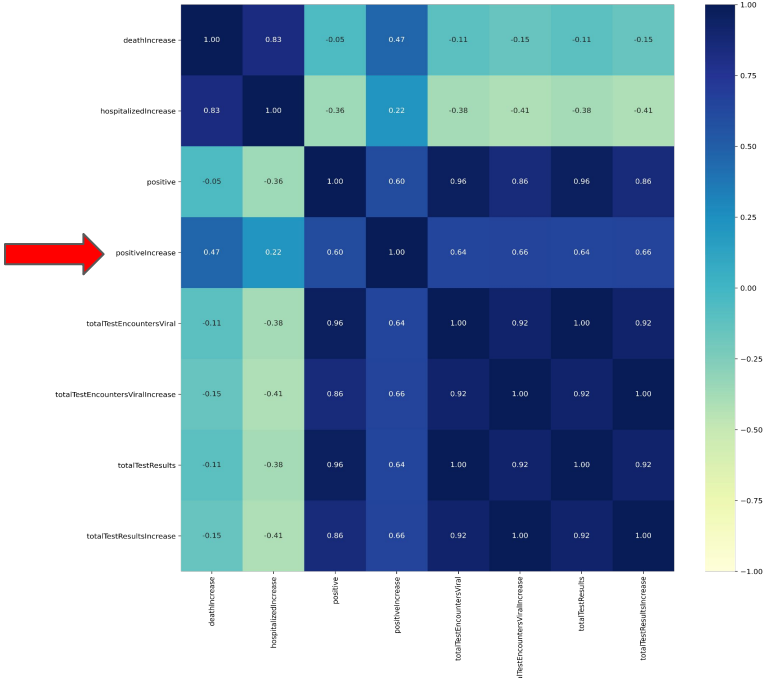
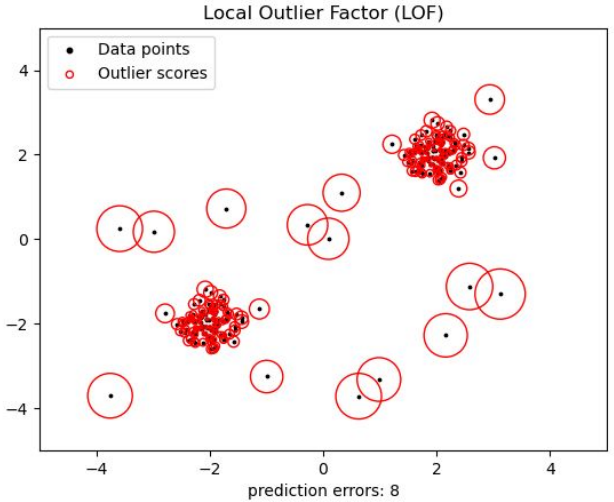


Figure 4: Correlation Matrix of New York dataset.

Local Outlier Factor

Local Outlier Factor

- Outlier scores based on density and distance to K nearest neighbours
- Data points with high outlier scores are identified as anomalies
- Usually used for “complete” data sets
- Usable in streaming data?



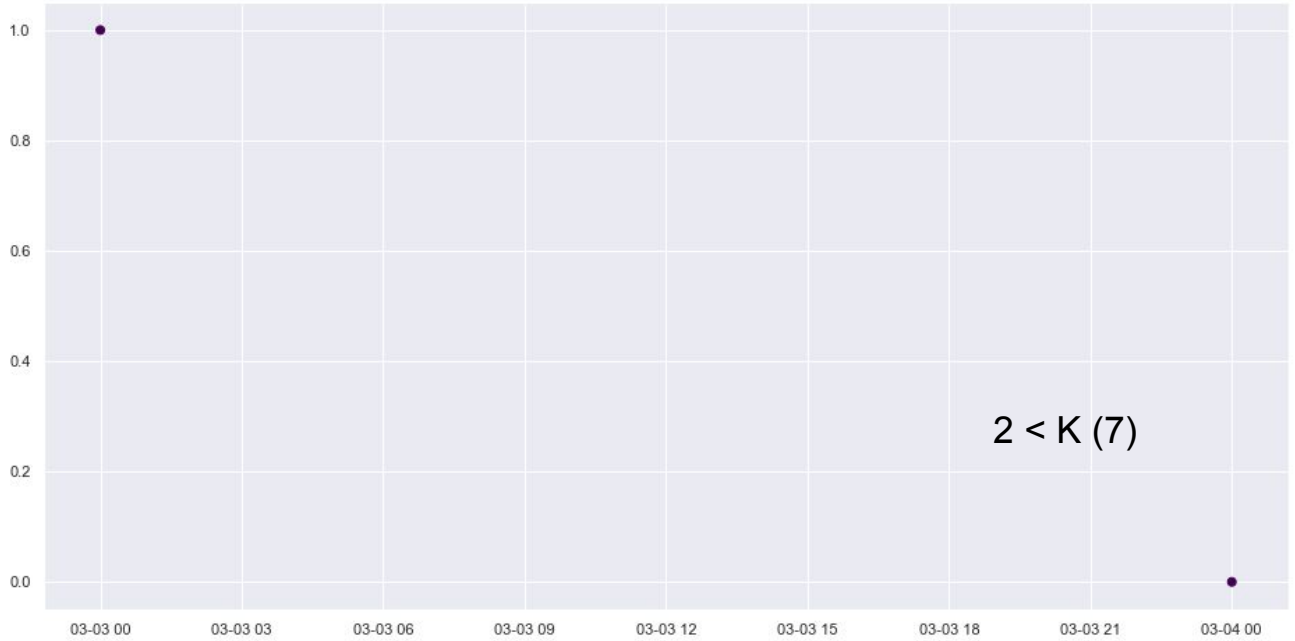
Hypothesis

Feeding last K days will result in reliable anomaly detection.

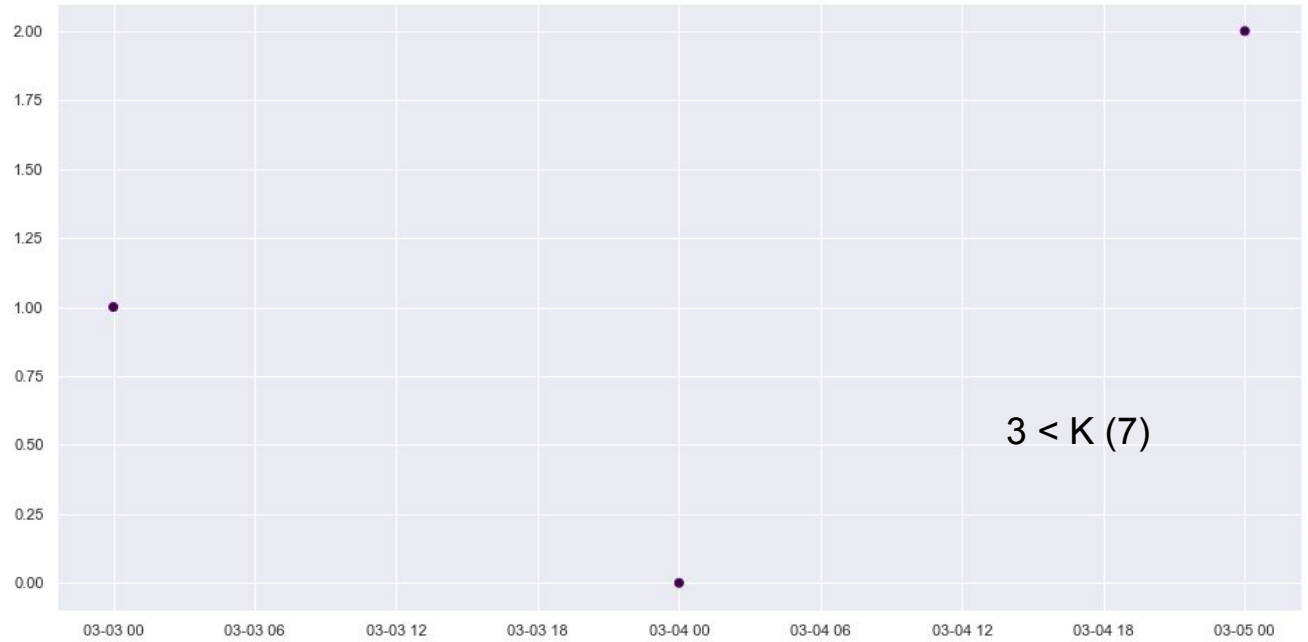
Test

Generate graph daily to observe if the anomaly detection for last day is reliable.

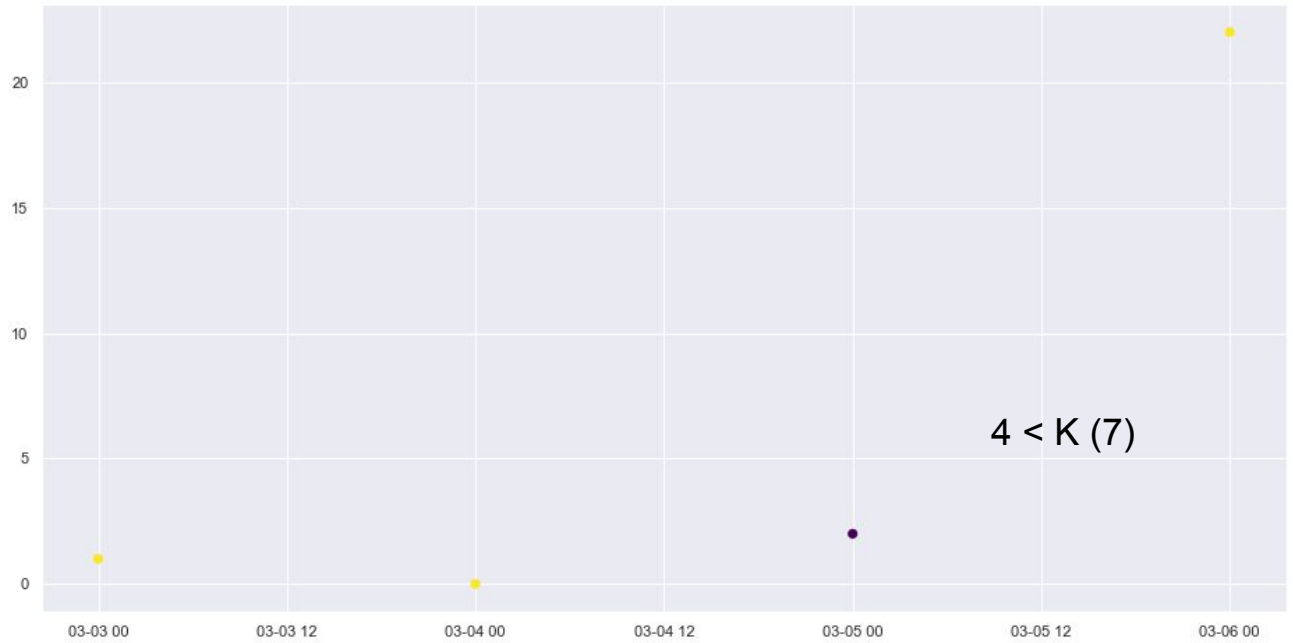
Local Outlier Factor



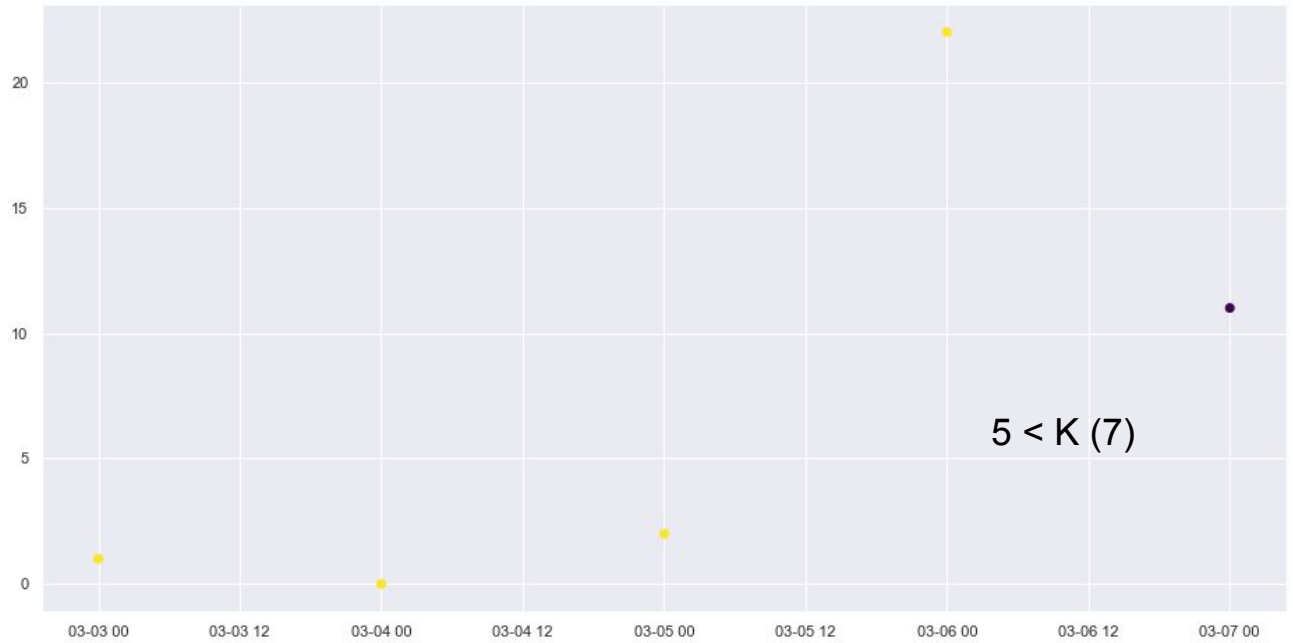
Local Outlier Factor



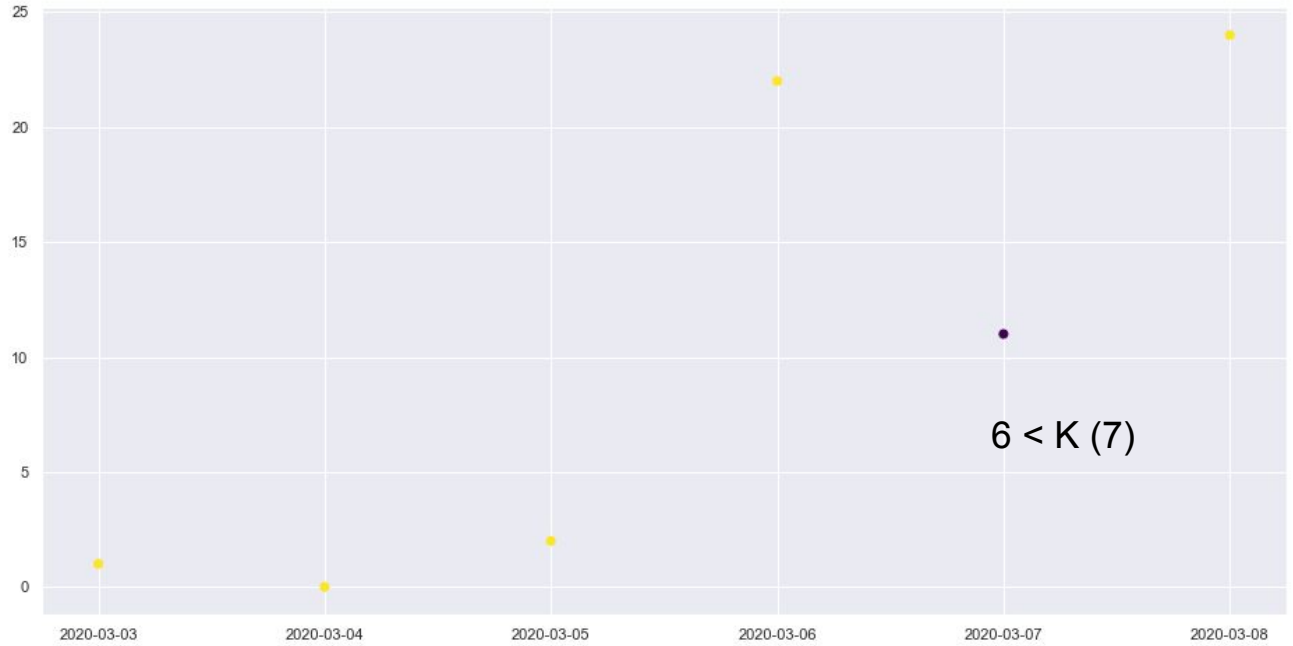
Local Outlier Factor



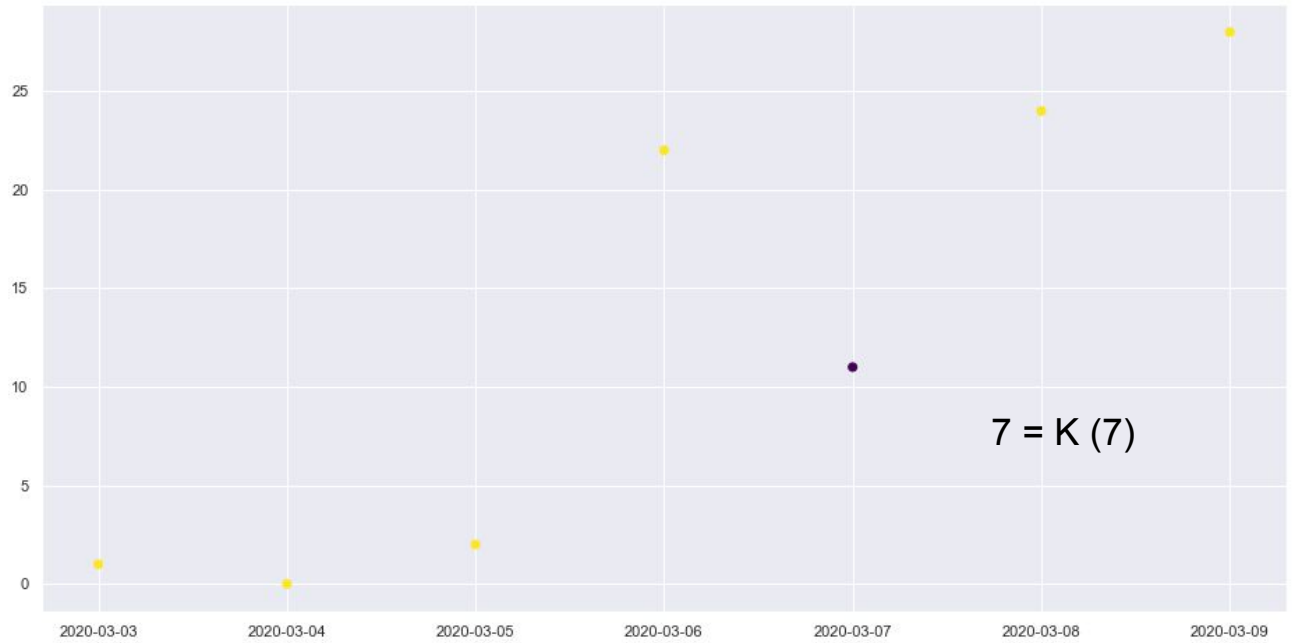
Local Outlier Factor



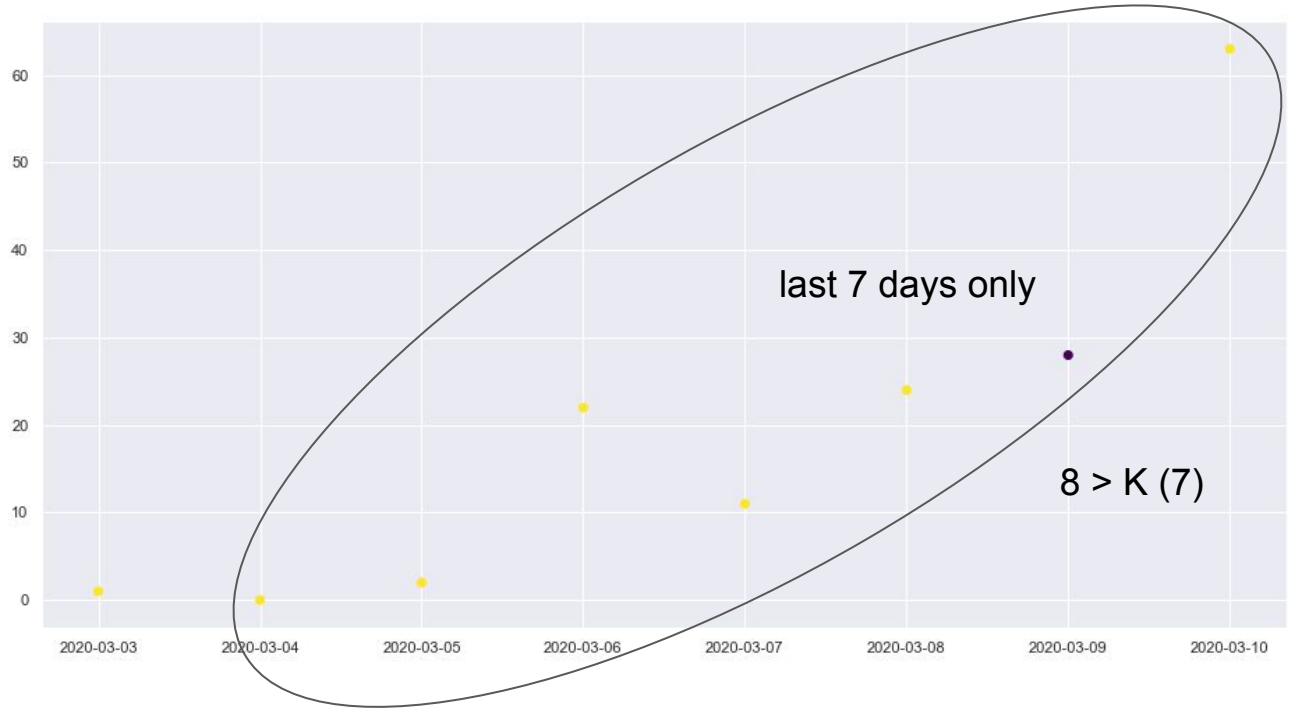
Local Outlier Factor



Local Outlier Factor



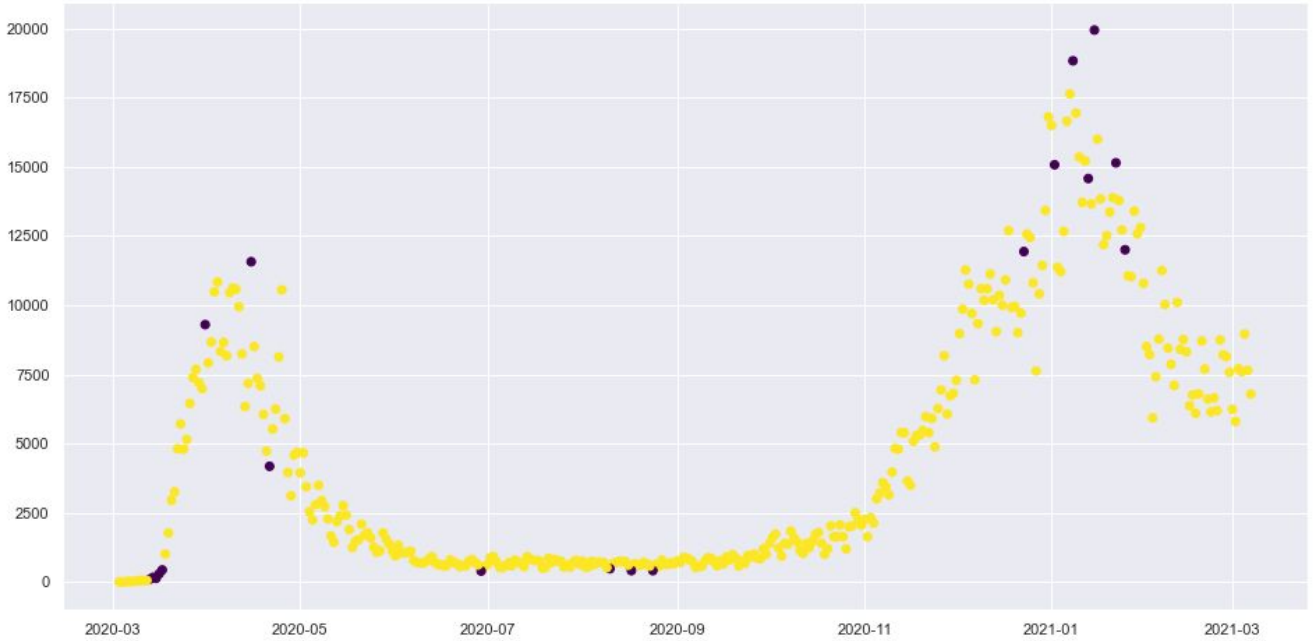
Local Outlier Factor



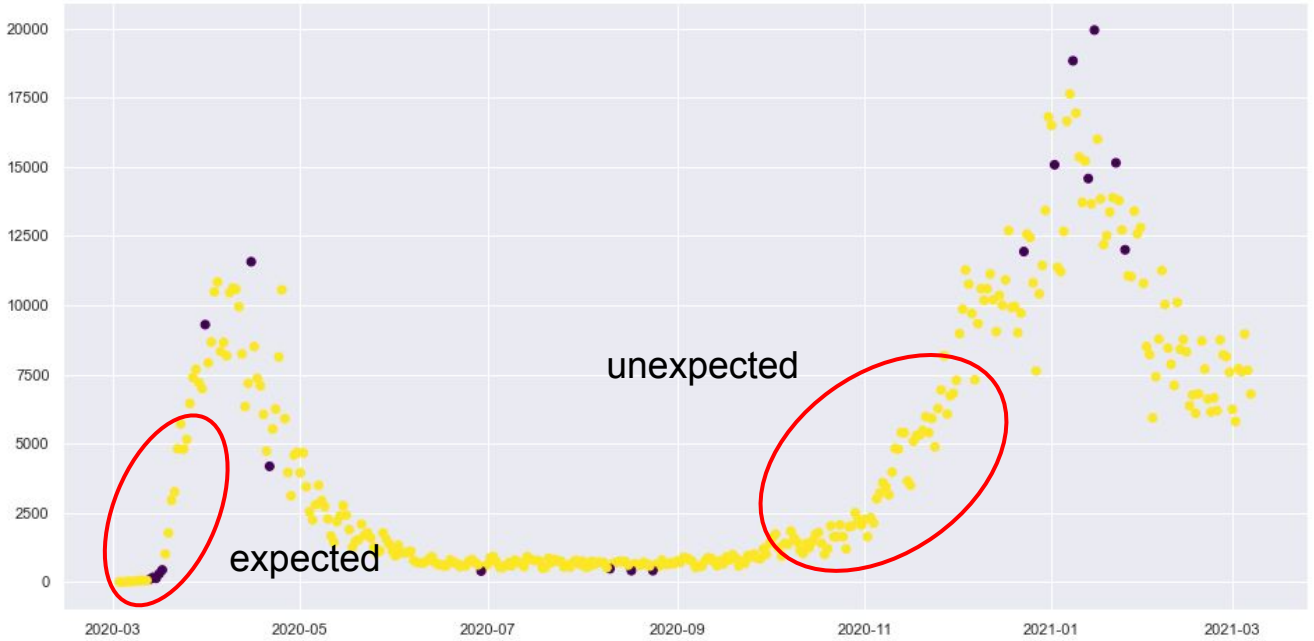
Testing...

many simulated days later...

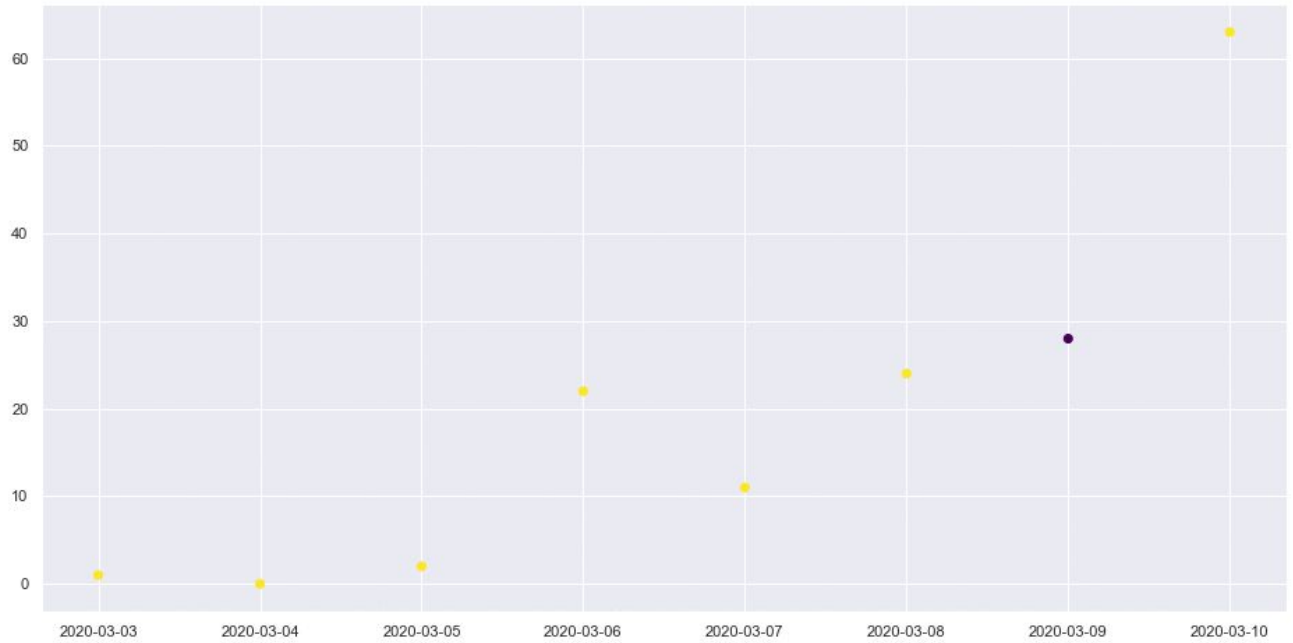
Local Outlier Factor



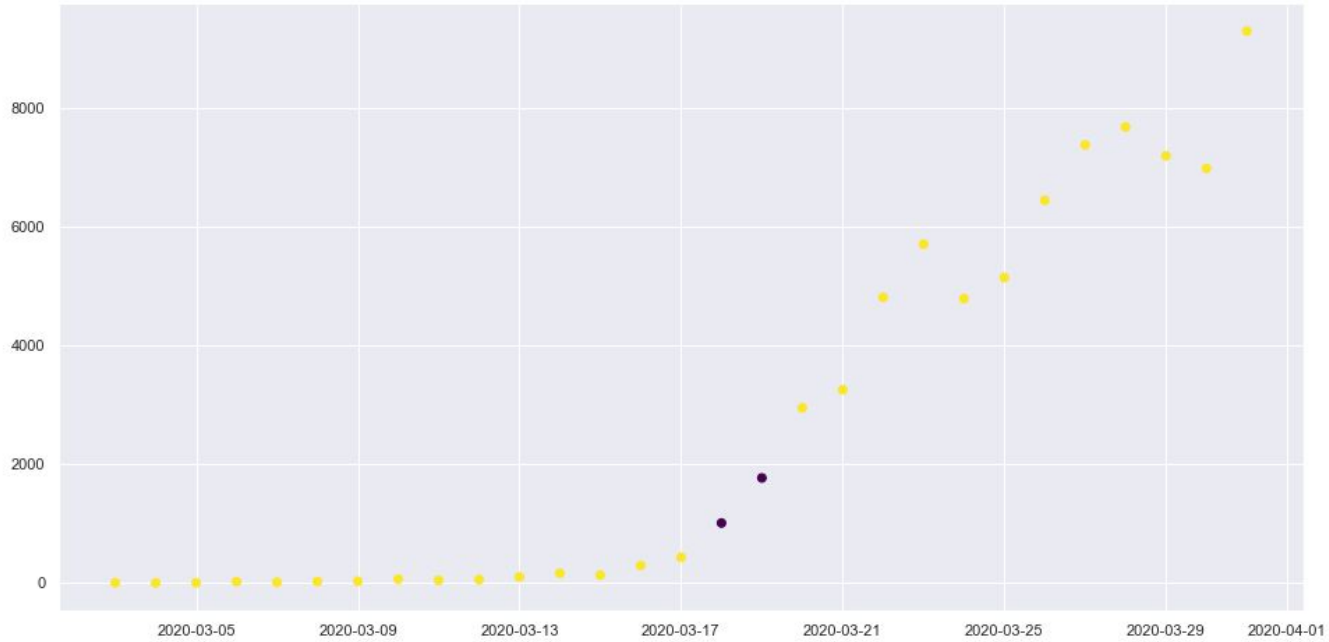
Local Outlier Factor



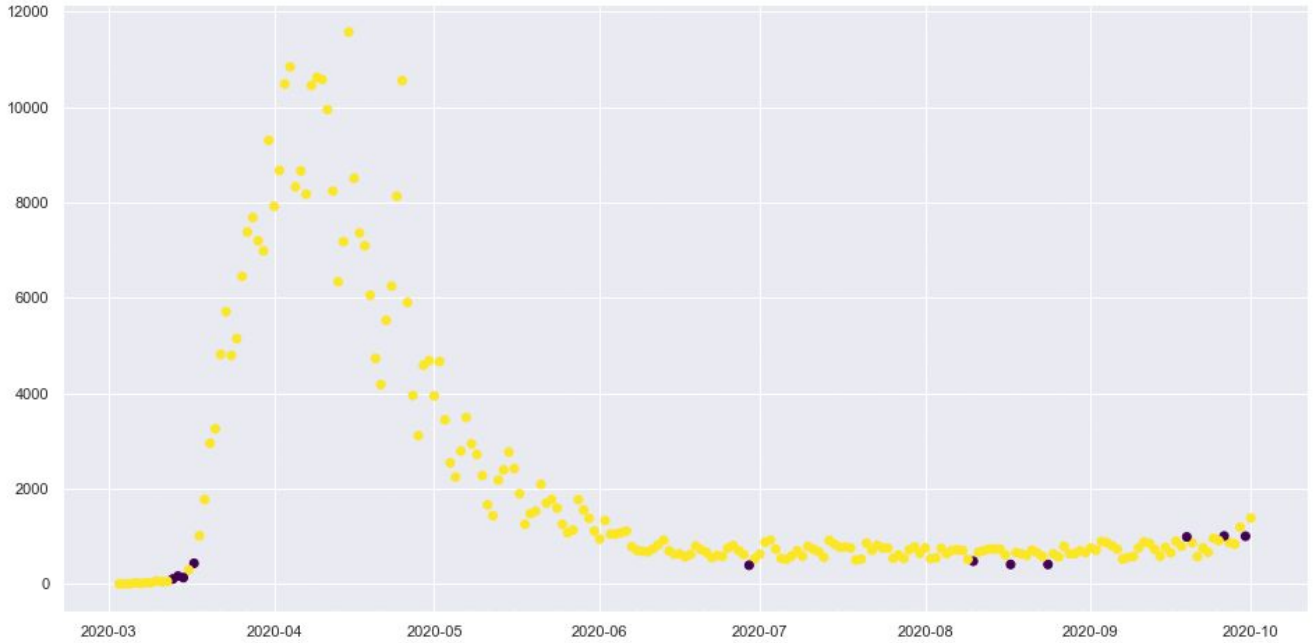
Local Outlier Factor



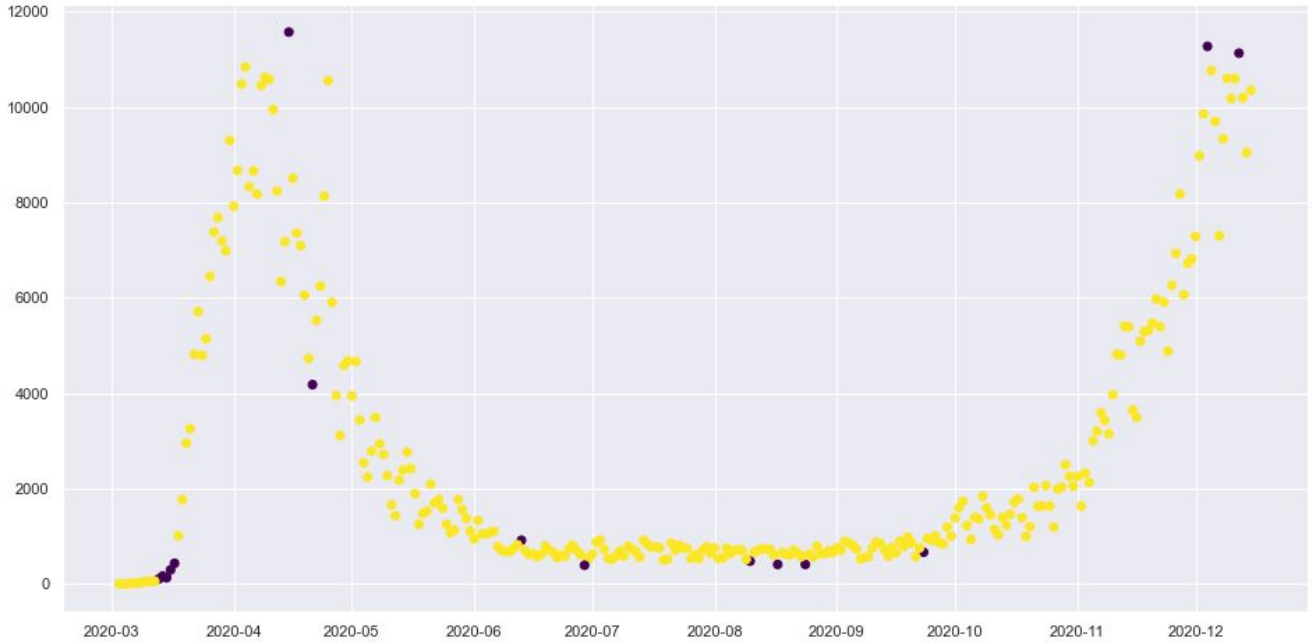
Local Outlier Factor



Local Outlier Factor



Local Outlier Factor



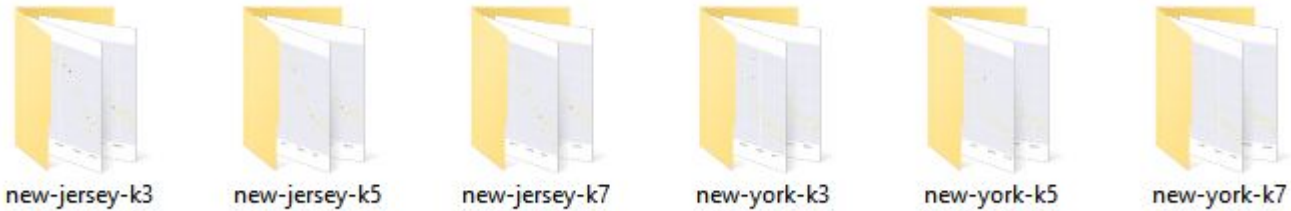
Somewhat reliable?

Requires more testing with different data sets.

Local Outlier Factor

More tests, different:

- K values
- Data sets
- Contamination factor



More testing...

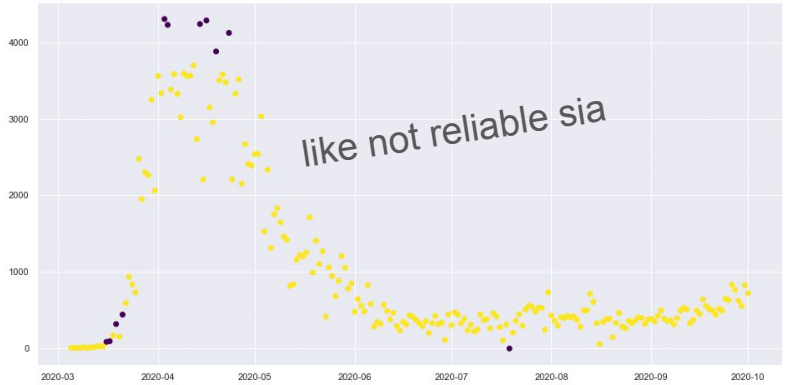
many more simulated days later...

more testing == more better

Local Outlier Factor

Findings

- Dates marked as outliers do not always reflect anomalies of the trend
- Require adjustments to Contamination Factor and K values for each data set to be accurate



Need something else!

kthxbye

hav a gud pm

ARIMA

Auto Regressive Integrated Moving Average

What is the ARIMA model?

- A class of model that ‘explains’ a given time series
- Many uses, mostly involving a time series

    1. Identify Trends
    2. Forecast Future Values

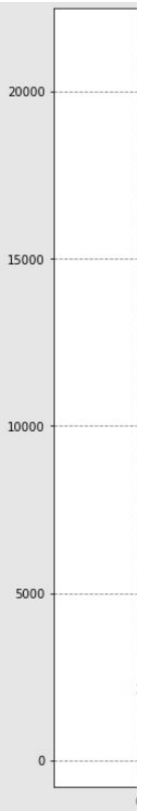
- Repurpose the ARIMA model to suit our problem of detecting anomalies

Auto Regressive Integrated Moving Average

P	D	Q
Number of lags to be used as predictors	Minimum number of differencing needed to make the series stationary	Number of lagged forecast errors

(4,1,1)

ARIMA(4,1,1) Confidence Interval of 85%

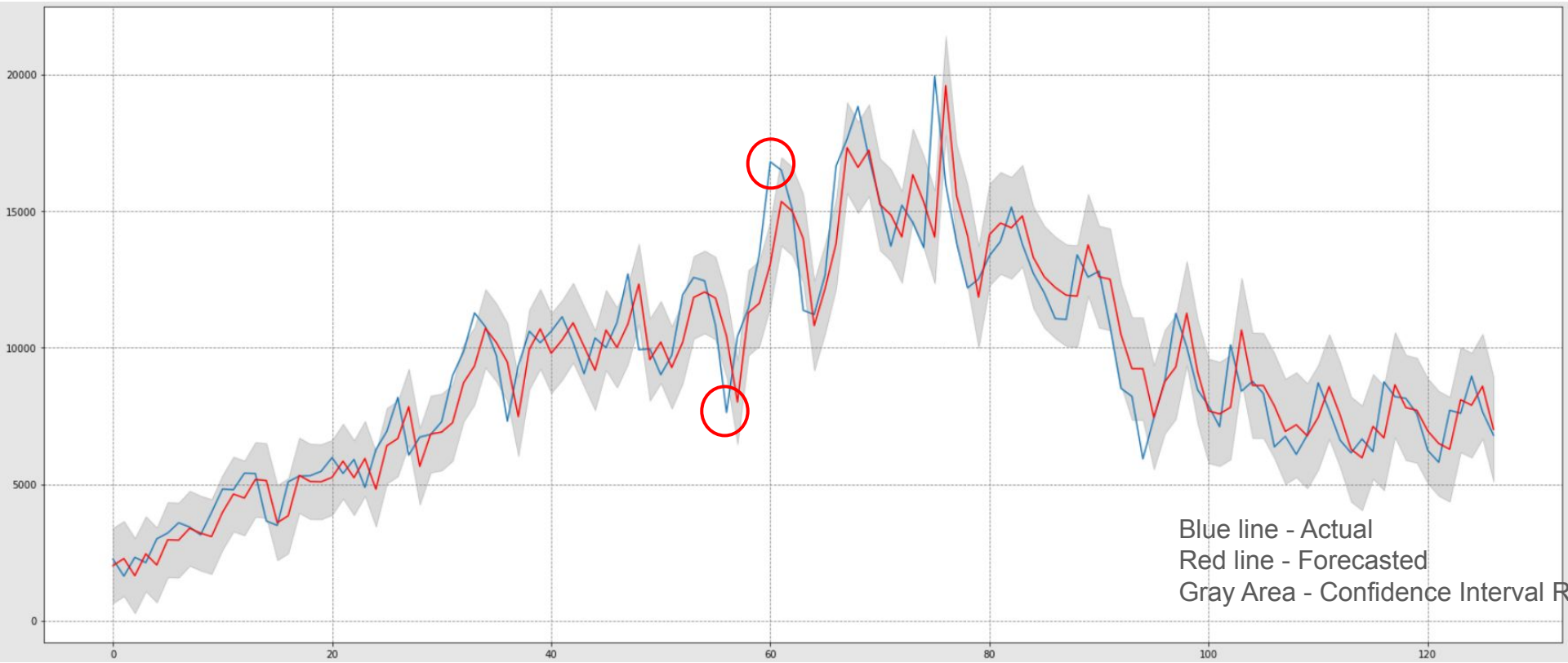


- Partition data into
  - Train (66%)
  - Test (34%)

- get_forecast()
- Add actual test values into the train set
- Remodel

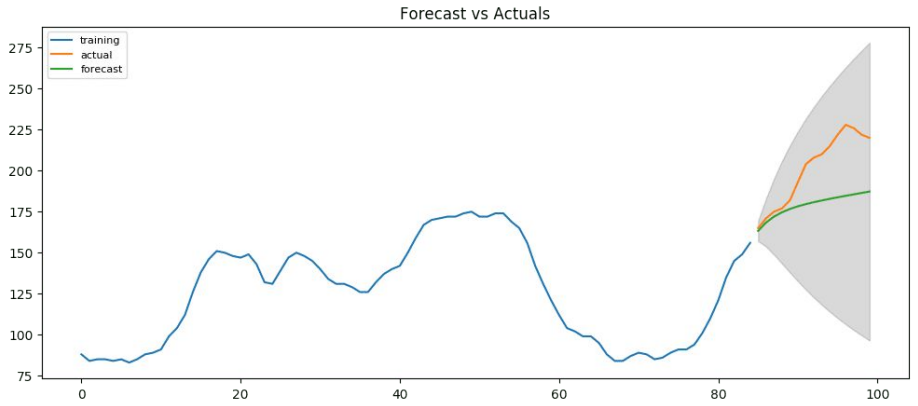
Blue line - Actual
Red line - Forecasted
Gray Area - Confidence Interval R

Here are some anomalies that we can identify!



ARIMA

- List of dates can be compiled, acting as outliers predicted by our model
- Our model can now be used as a tool to monitor anomalies
- However, it should be noted that if we want to forecast more than a day ahead, our range of confidence interval significantly widens (grey area).



Conclusion

jk im back

ARIMA

is what we need

Insights

- LOF is not recommended for streaming data
- ARIMA is more accurate because it is designed for time series

Authorities could use dates identified by the model to identify potential events that led to the anomaly.

	positiveIncrease
2020-11-15	3649
2020-11-25	6265
2020-11-27	8176
2020-11-28	6063
2020-12-02	8973
2020-12-04	11271
2020-12-07	7302
2020-12-08	9335
2020-12-18	12697
2020-12-19	9919
2020-12-23	11937
2020-12-27	7623
2020-12-28	10407
2020-12-30	13422
2020-12-31	16802
2021-01-03	11368
2021-01-06	16648
2021-01-08	18832
2021-01-13	14577
2021-01-15	19942
2021-01-16	15998
2021-01-18	12185
2021-02-01	8508
2021-02-03	5925
2021-02-06	11252
2021-02-11	10099
2021-02-12	8404
2021-02-25	8746

Conclusion

Since we want to build a program that will alert authorities to investigate suspicious surges and/or plunges in the increment of positive cases, we need a reliable algorithm that can reliably detect anomalies so that potential rectifications can be made early.

Therefore,

ARIMA is the way to go.

Contributions

- Loh Zhi Heng - Data Exploration (data/variables selection)
- Oh Zhi Hua - LOF testing
- Wee Chang Han - ARIMA testing
- Equal - problem formulation, exploring possible solutions, and conclusion

Thank you!

References

[1] Tozzi, J. (2020, June 11). U.S. Health Care Puts $4 Trillion in All the Wrong Places. *Bloomberg*. https://www.bloomberg.com/news/articles/2020-06-11/u-s-health-care-system-was-totally-overwhelmed-by-coronavirus.

[2]Thompson, C. N., Baumgartner, J., Pichardo, C., Toro, B., Li, L., Arciuolo, R., Chan, P. Y., Chen, J., Culp, G., Davidson, A., Devinney, K., Dorsinville, A., Eddy, M., English, M., Fireteanu, A. M., Graf, L., Geevarughese, A., Greene, S. K., Guerra, K., . . . Fine, A. (2020). COVID-19 Outbreak — New York City, February 29–June 1, 2020. MMWR. Morbidity and Mortality Weekly Report, 69(46), 1725–1729. https://doi.org/10.15585/mmwr.mm6946a2

[3] Kuchler, H., & Edgecliffe-Johnson, A. (2020, October 22). How New York's missteps let Covid-19 overwhelm the US. *Financial Times.* https://www.ft.com/content/a52198f6-0d20-4607-b12a-05110bc48723

[4] Shrivastava, S. (2019, July 26). Anomaly detection on a categorical and continuous dataset. *Medium.* Anomaly detection on a categorical and continuous dataset | by Shreyash Shrivastava | Medium.

[5] Scikit Learn. (n.d.). Outlier detection with Local Outlier Factor (LOF). *Scikit Learn.* Outlier detection with Local Outlier Factor (LOF) — scikit-learn 0.24.1 documentation (scikit-learn.org).

[6] Prabhakaran, S. (n.d.). ARIMA Model – Complete Guide to Time Series Forecasting in Python. *Machine Learning Plus.* https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/#:~:text=ARIMA%2C%20short%20for%20'Auto%20Regressive,used%20to%20forecast%20future%20values.