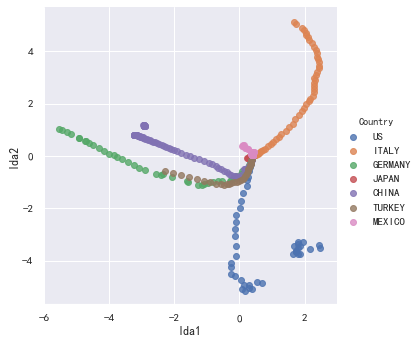
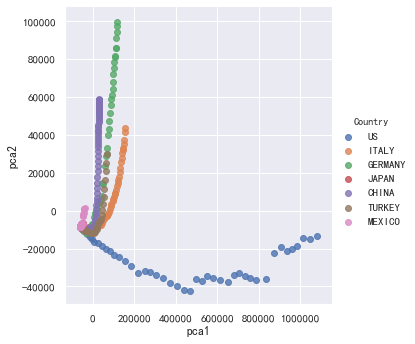
Yiming Ge

Personal report

At first, I tried to use what we learned in class to make a cluster and classification.

I use PCA and LDA separately to reduce the covid-19 [recovery, deaths and confirmed cases] from US, ITALY,GERMAN,JAPAN,CHINA,TURKEY,MEXICO 7 countries dimensions.



Then I tried to use both LDA and PCA to do the SVM classifier and RF (random forest)classifier.

No matter how I changed the C, the f1 score of SVM classifier is always very low: around 0.5 for LDA and 0.3 for PCA

FI score of RF classifier is not good as well: 0.6 for both LDA and PCA

With C = 0.8

train set shape = (504, 3)

test set shape = (217, 3)

lda1 lda2 Country

104 1.750188 5.043696 ITALY

114 2.380807 3.922120 ITALY

647 0.377287 0.057276 MEXICO

425 -2.918821 1.161719 CHINA

676 0.406681 0.063000 MEXICO

f1 score for SVM classifier = 0.500187

train set shape = (504, 3)

test set shape = (217, 3)

pca1 pca2 Country

104 153564.138634 41817.794408 ITALY

114 126004.554889 17935.880961 ITALY

647 -59331.015021 -7831.066875 MEXICO

425 30821.424798 57986.683415 CHINA

676 -60901.906642 -8272.429732 MEXICO

f1 score for SVM classifier = 0.274626

Then we tried to classify these countries manually according to the PCA, LDA. From above PCA diagram we can easily find USA is unique. From LDA diagram we found Italy and USA are special. Thus, we decide to make two different groups: EAST VS WEST; DEVELOPING VS DEVELOPED. We made the k means cluster between recovered and confirmed to find further contact.

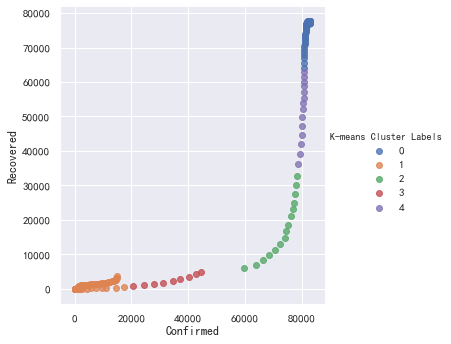
***EAST: Japan & China***

***WEST:USA MEXICO GERMANY ITALY***

EAST K MEANS CLUSTER

We combine Japan & china recovered vs confirmed data together, we can easily see till 5/2, china recovered has already covered all the confirmed case.

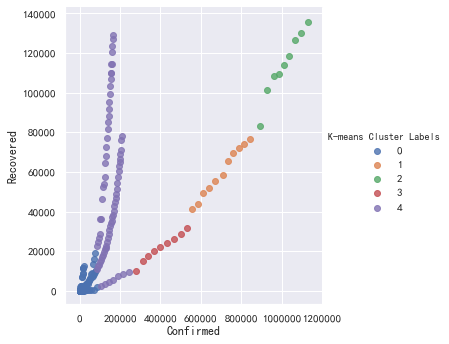
Japan is still in china’s period 1 recovered vs confirmed situation. (orange cluster)[slow confirmed growth rate and slow recovered growth rate]



WEST K MEANS CLUSTER:

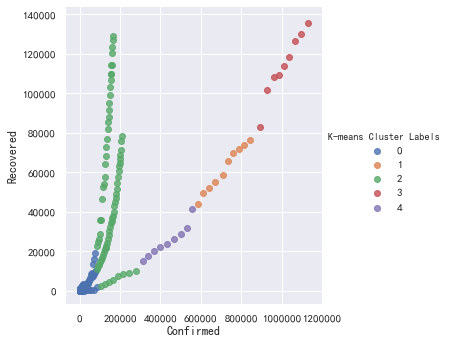
we can easily see till 5/2, US recovered vs confirmed situation is very different from other west countries. Most west countries recovered cases has almost covered all the confirmed case.[purple cluster: low confirmed cases growth rate and high recovered cases growth rate.

However US is in a unique path, we can see right now it is in green cluster with very high confirmed cases about 1,200,000 and about 9 times less than confirmed case only 140,000 recovered cases. But from the trend, we can see the US recovered growth speed is gradually getting faster.



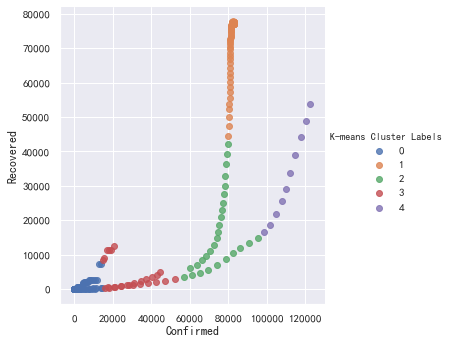
Developed cluster: USA GERMAN ITALY JAPAN

Almost same as west k means cluster, since we only add JAPAN in it.



Developing cluster: CHINA MEXICO TURKEY

Till 5/2, we can see MEXICO and TURKEY Recovered growth speed is gradually getting faster. Although they cannot act like china to nearly stop the confirmed cases growing, their confirmed cases growth speed is becoming slower than recovered rate( slope is obviously greater than 45 degree)



F

After group discussion, we only want to cluster Mexico, China, US, Germany, Turkey, Japan six countries.

So I used the PCA to reduce the newest 5.2 COVID-19 data [recovered, deaths, confirmed] dimension and made the hierarchical cluster to classify them.

Based on the following graph, we can either divide the countries into 2 cluster or 4 cluster

Country PC1 PC2 ClusterID

0 US 3.364531 -0.501806 0

1 GERMANY 0.280351 1.203510 1

2 JAPAN -1.369240 -0.656170 1

3 CHINA -0.413905 0.451763 1

4 TURKEY -0.628146 0.042960 1

5 MEXICO -1.233591 -0.540257 1

Country PC1 PC2 ClusterID

0 US 3.364531 -0.501806 0

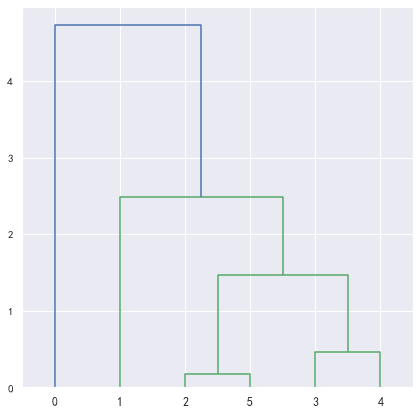
1 GERMANY 0.280351 1.203510 1

2 JAPAN -1.369240 -0.656170 2

3 CHINA -0.413905 0.451763 3

4 TURKEY -0.628146 0.042960 3

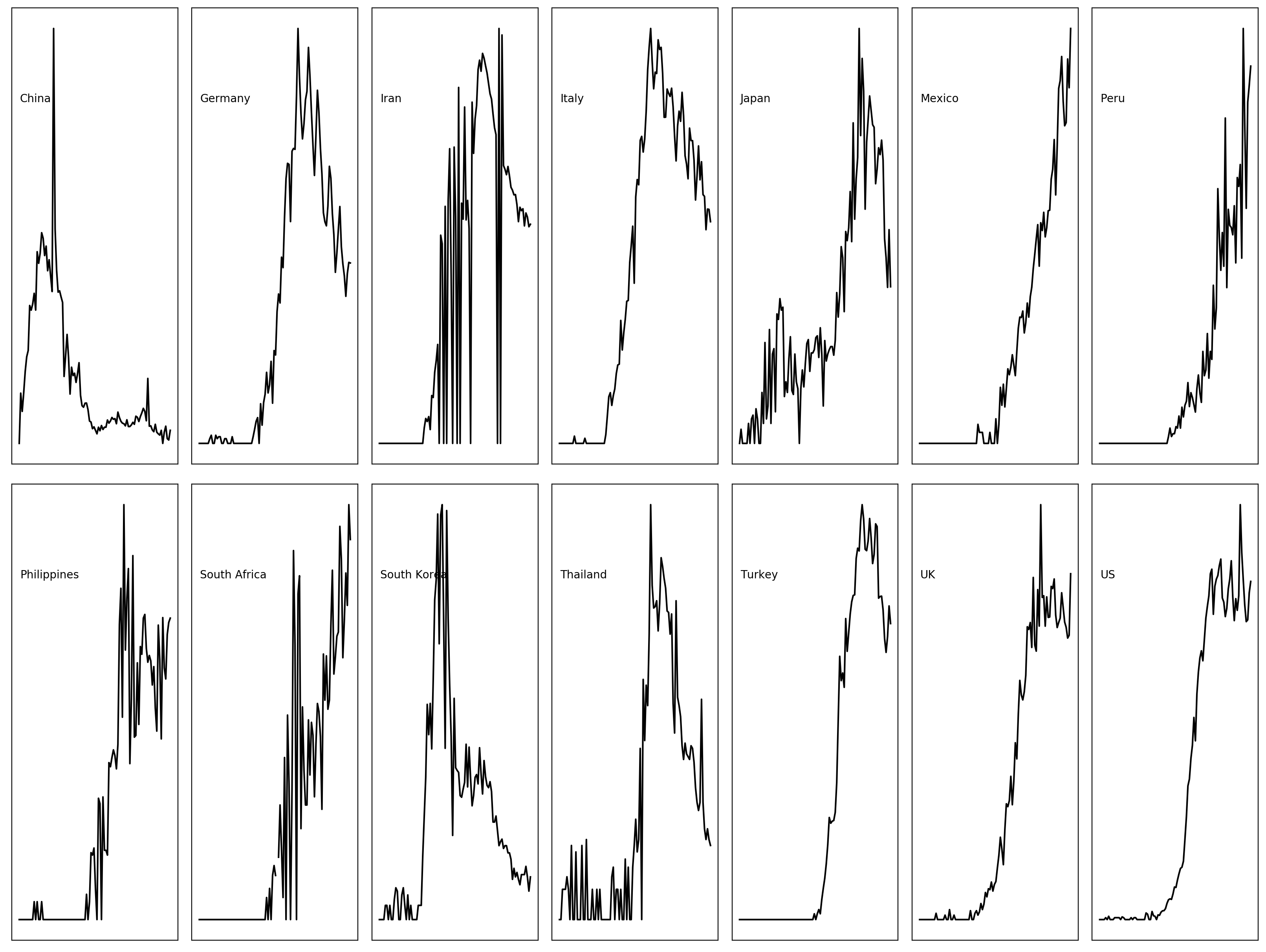
5 MEXICO -1.233591 -0.540257 2



Then I do the covid19 cases prediction part for some countries:

Firstly based on the daily new cases data I draw the new cases trend.

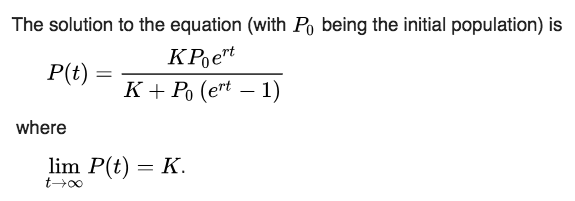
Some countries are still climbing(Mexico, Peru, South Africa), some countries are staying the peak(Phillippines, US), Some countries are downhill(China, Germany, Italy, Japan, South Korea, Turkey, UK)



I tried use logistic model to predict the confirmed cases trend in following 20 days.

Function is

P(t) = KP0e^rt/(K+P0(e^rt-1)



K => known MAX cases

P0 => initial cases

r => growth rate

K P0 r

CHINA:[8.16248877e+04 1.11762891e+03 2.20343072e-01]

IRAN: [9.71222683e+04 3.96190004e+01 1.08311810e-01]

JAPAN:[2.08286170e+04 7.72819951e+00 8.87734377e-02]

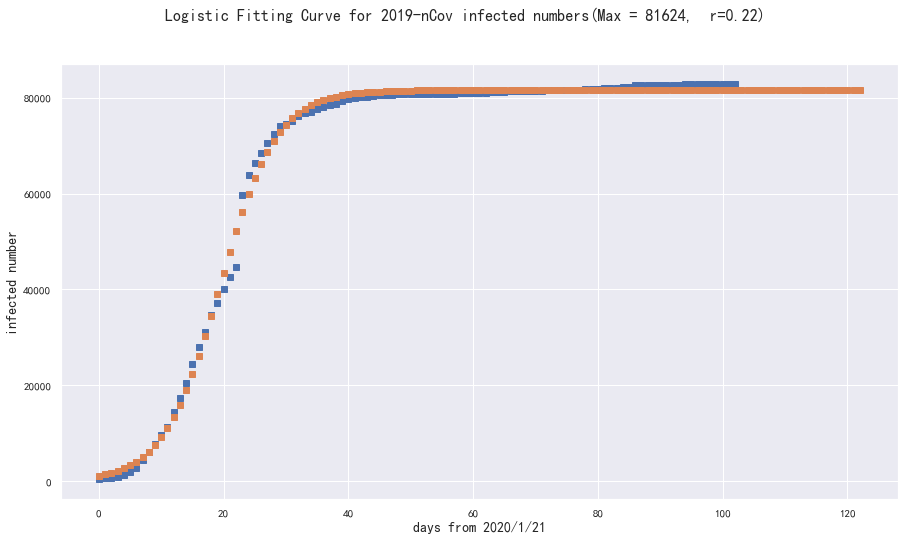
Philippines:[8.95578063e+03 4.25911571e-01 1.22736502e-01]

SOUTH KOREA:[1.02357612e+04 4.32311614e+00 1.76951532e-01]

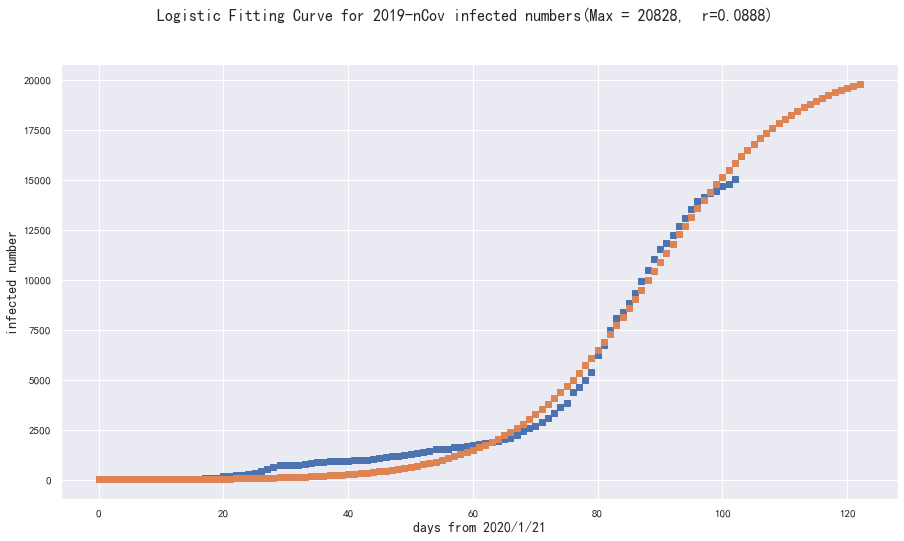
Thailand:[2.87928148e+03 1.52603380e-02 1.76574449e-01]

UK:[1.88877433e+05 3.44227064e+00 1.27943975e-01]

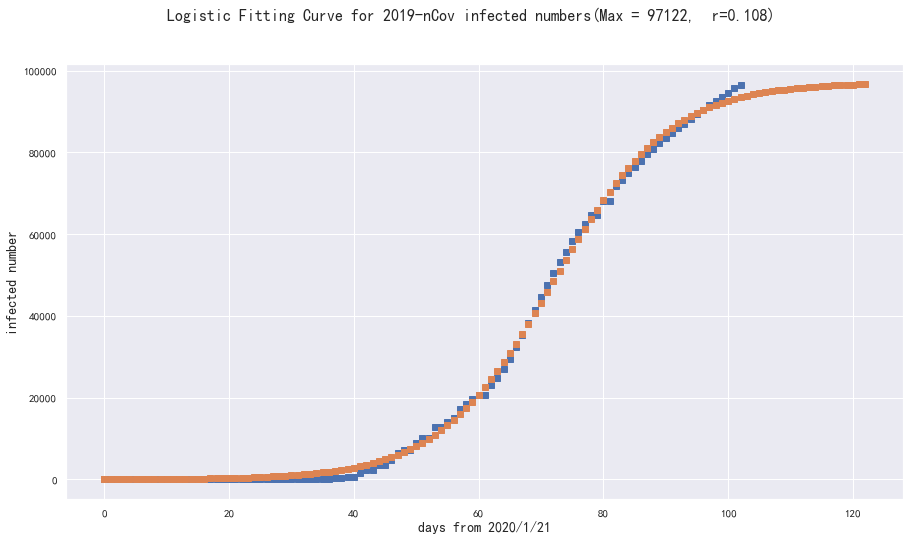
CHINA WAS STOPPED.



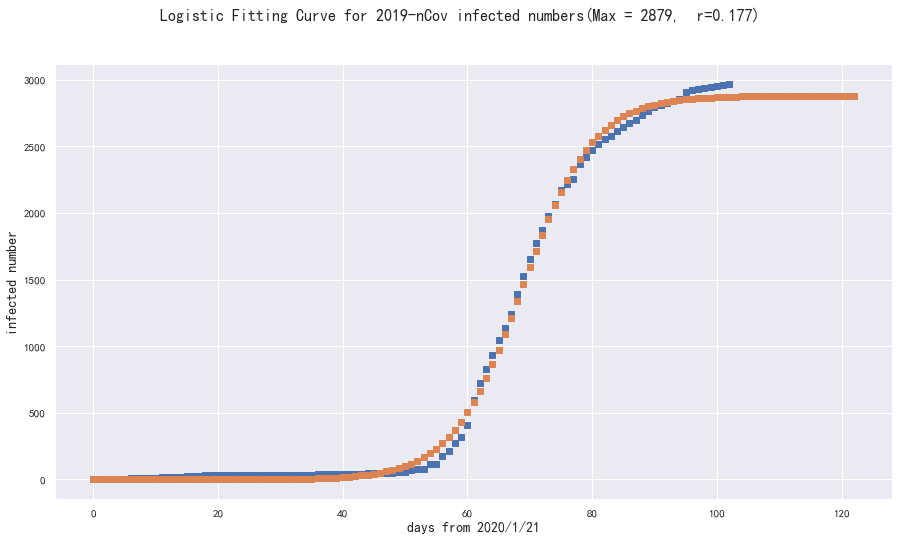
Iran



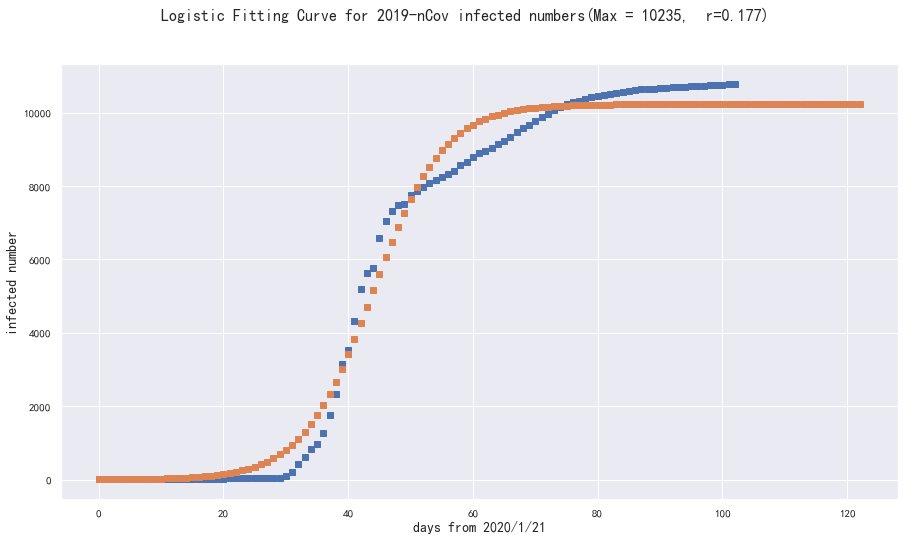
JAPAN



South Korea



Thailand



The above selected countries will almost stop the confirmed cases growth in the future 20 days.

However, I tried to use US data into this model. Python could not fit this curve and R report the error data are not appropriate for this logistic model.

So I tried polynomial model to predict US cases

The output is kind of scary. It seems like US still not hit the peak and will not hit the peak in the end of May as well.

A\*x^3+ B\*x^2 + C\*x + D

A B C D

US:[ 2.72398141 -183.09245568 2384.4934822 700.21119776]

