1. EDA

ID 131090 non-null object

age 5116 non-null object

sex 5400 non-null object

city 59680 non-null object

province 117623 non-null object

country 131230 non-null object

wuhan(0)\_not\_wuhan(1) 64744 non-null float64

latitude 129279 non-null float64

longitude 129279 non-null float64

geo\_resolution 129279 non-null object

date\_onset\_symptoms 1913 non-null object

date\_admission\_hospital 961 non-null object

date\_confirmation 129877 non-null object

symptoms 803 non-null object

lives\_in\_Wuhan 3629 non-null object

travel\_history\_dates 1163 non-null object

travel\_history\_location 2636 non-null object

reported\_market\_exposure 1154 non-null object

additional\_information 4021 non-null object

chronic\_disease\_binary 435 non-null object

chronic\_disease 43 non-null object

source 114800 non-null object

sequence\_available 1111 non-null object

outcome 315 non-null object

date\_death\_or\_discharge 210 non-null object

notes\_for\_discussion 206 non-null object

location 2856 non-null object

admin3 5347 non-null object

admin2 40679 non-null object

admin1 59861 non-null object

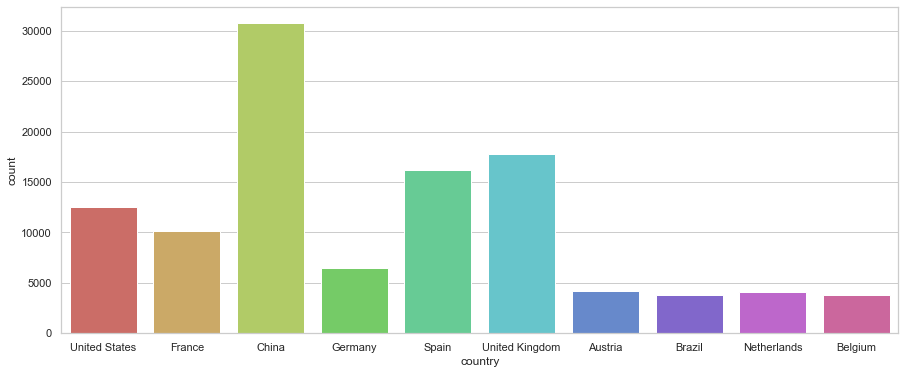
country\_new 70899 non-null object

admin\_id 62660 non-null object

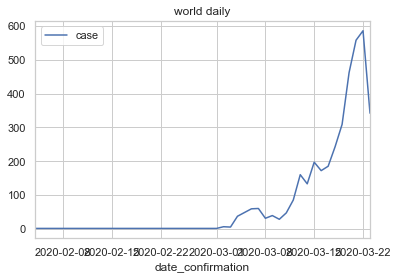
data\_moderator\_initials 35117 non-null object

travel\_history\_binary 1287 non-null float64

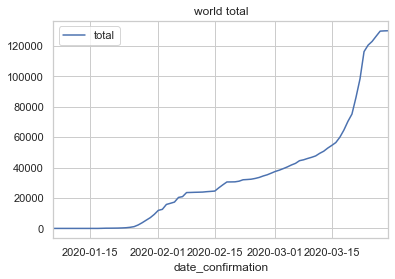
The data is very uncompleted. I chose country, date\_confirmation as the features



First, a distribution of total cases of top-ten countries



World-wide daily increased cases



World-wide total cases

Then I did national curves for top-ten countries. Please see them in the code.

The problem is that I want to predict daily increase by historical data.

1. ML prediction for daily increased cases

I extracted the data for each country, calculated the days after the first case, build time series(5 days daily cases), then concat them up, changed the country into dummy variables. In that way, I can have nearly 10 times of the data. Then build two regression model (linear regression, random forest regressor). The results on testing set are like that:

LR:

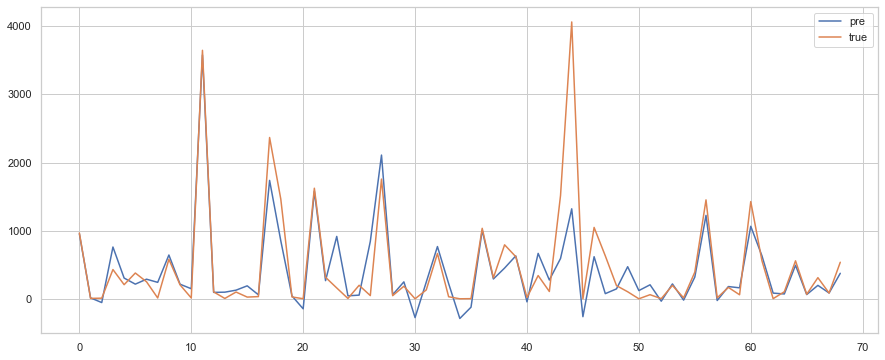
explained\_variance: 0.7112

r2: 0.7054

MAE: 207.0418

MSE: 177860.1039

RMSE: 421.7346



RF:

explained\_variance: 0.7753

r2: 0.7749

MAE: 180.2814

MSE: 135869.1704

RMSE: 368.6044

