COVID-19 Total Cases Prediction on Country-based Multivariable Time Series

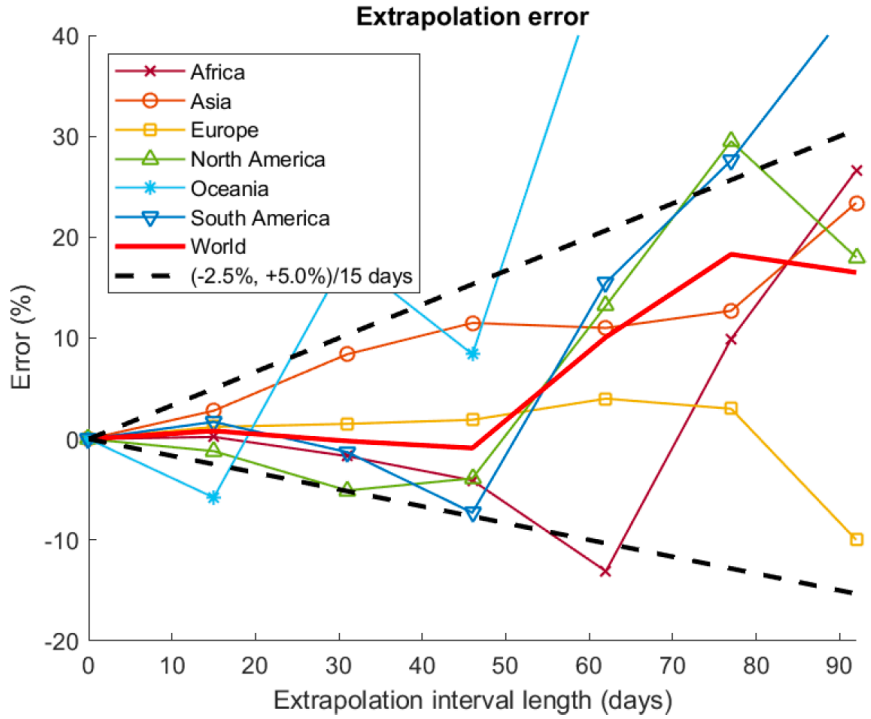
Nange Li, DSI Brown University

Oct. 15th 2021

GitHub Repository: https://github.com/7ericany/1030Project/tree/master

Introduction

- COVID-19 Pandemic
  - Target: total cases
  - Regression
- Dataset
  - OWID[1] (*Our World in Data*)
  - Previous works
    - Bimodel distribution[2] – deaths
    - Cloud Computing & Curve Fitting[3] – new_cases



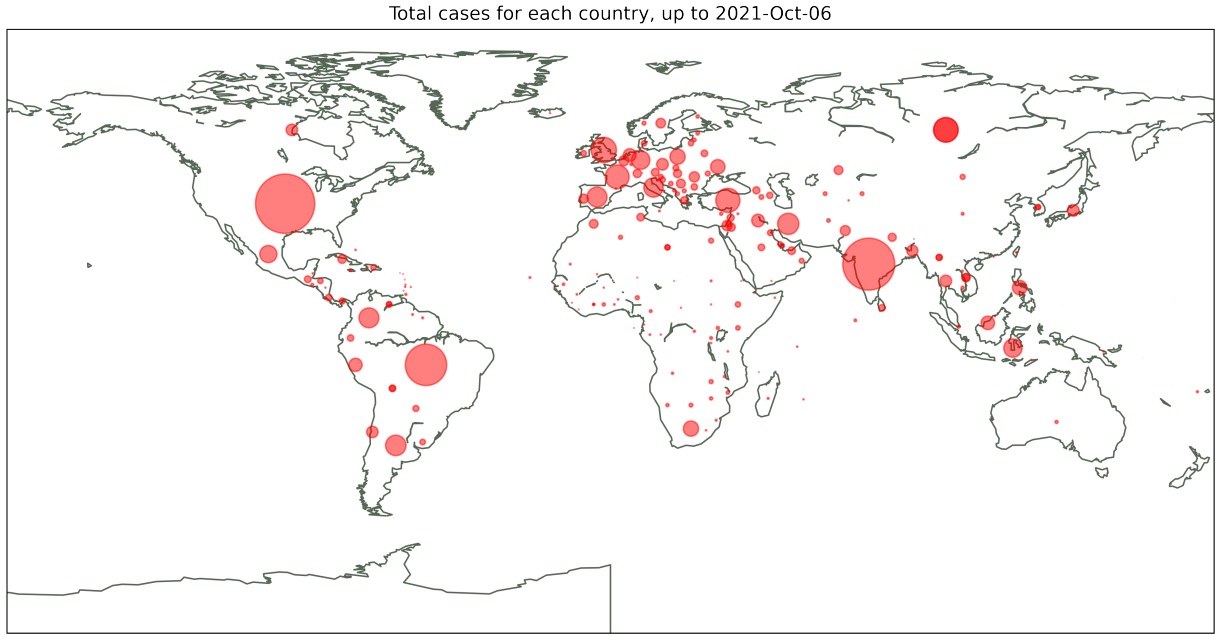
[1] Ritchie, H., Mathieu, E., et al. 2020. "Coronavirus Pandemic (COVID-19)". Published online at *OurWorldInData.org*. Retrieved from: https://ourworldindata.org/coronavirus

[2] Valvo, & Paolo S. 2020. "A Bimodal Lognormal Distribution Model for the Prediction of COVID-19 Deaths" . *Applied Sciences 10*, no. 23: 8500. Doi: https://doi.org/10.3390/app10238500

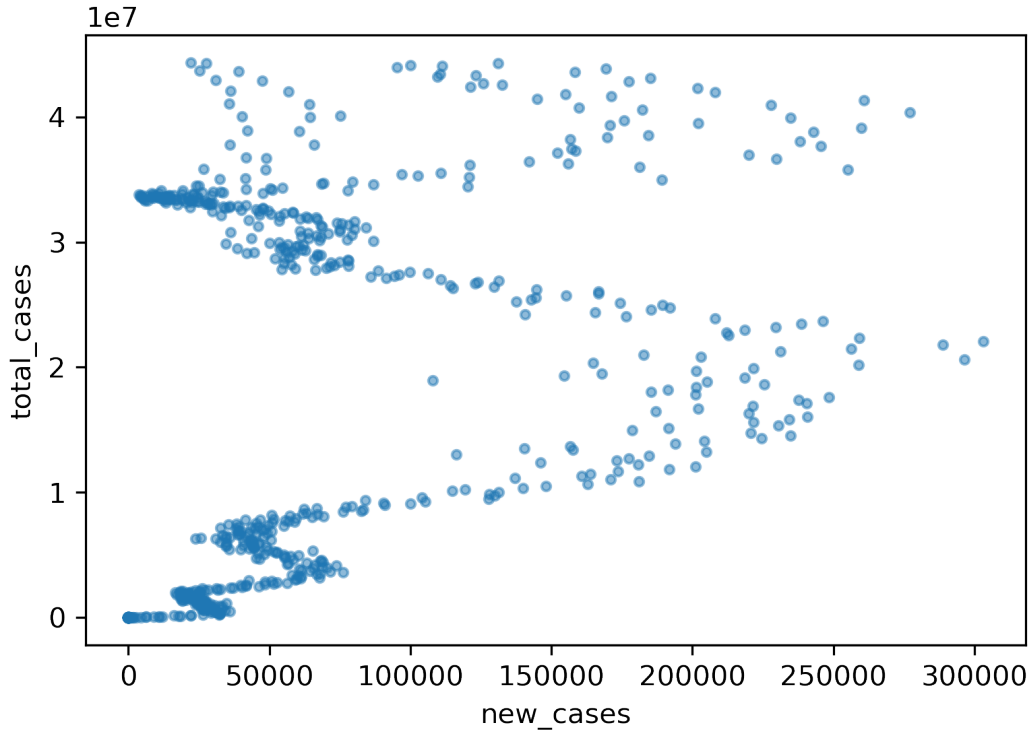
[3] Tuli, S., Tuli, R., Gil, S S., 2020. "Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing" . *Internet of Things*, Volume 11, 2020, 100222, ISSN 2542-6605, Doi: https://doi.org/10.1016/j.iot.2020.100222

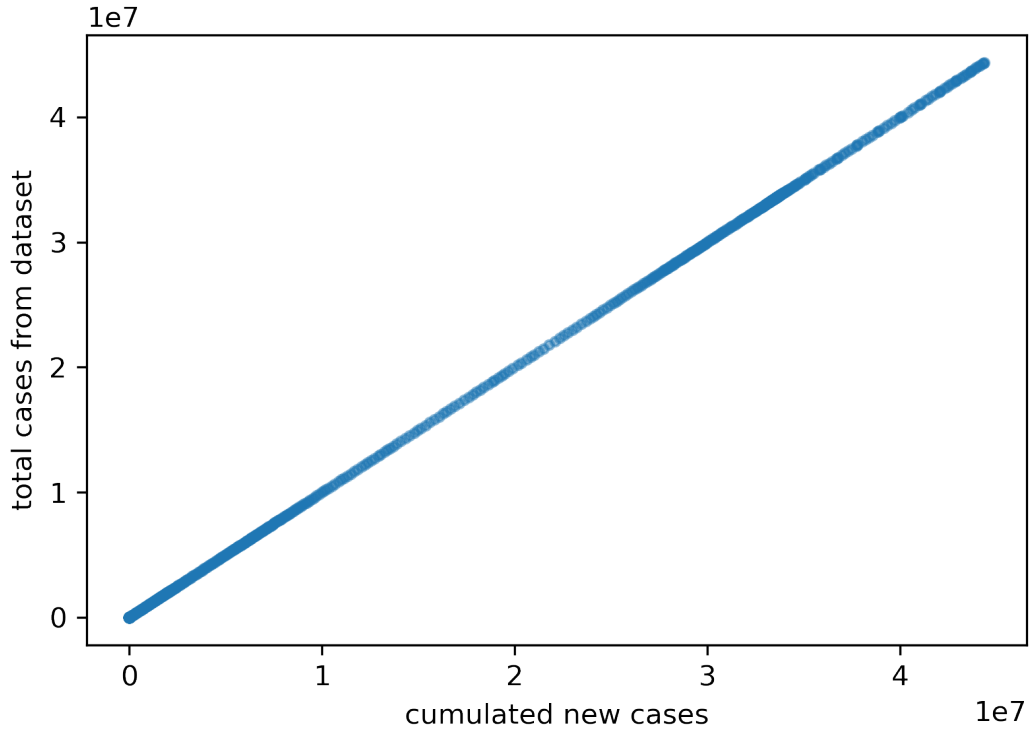
EDA

- **65** columns
- **122636** records
- **233** countries
- 2020-02-24 to 2021-10-06



EDA



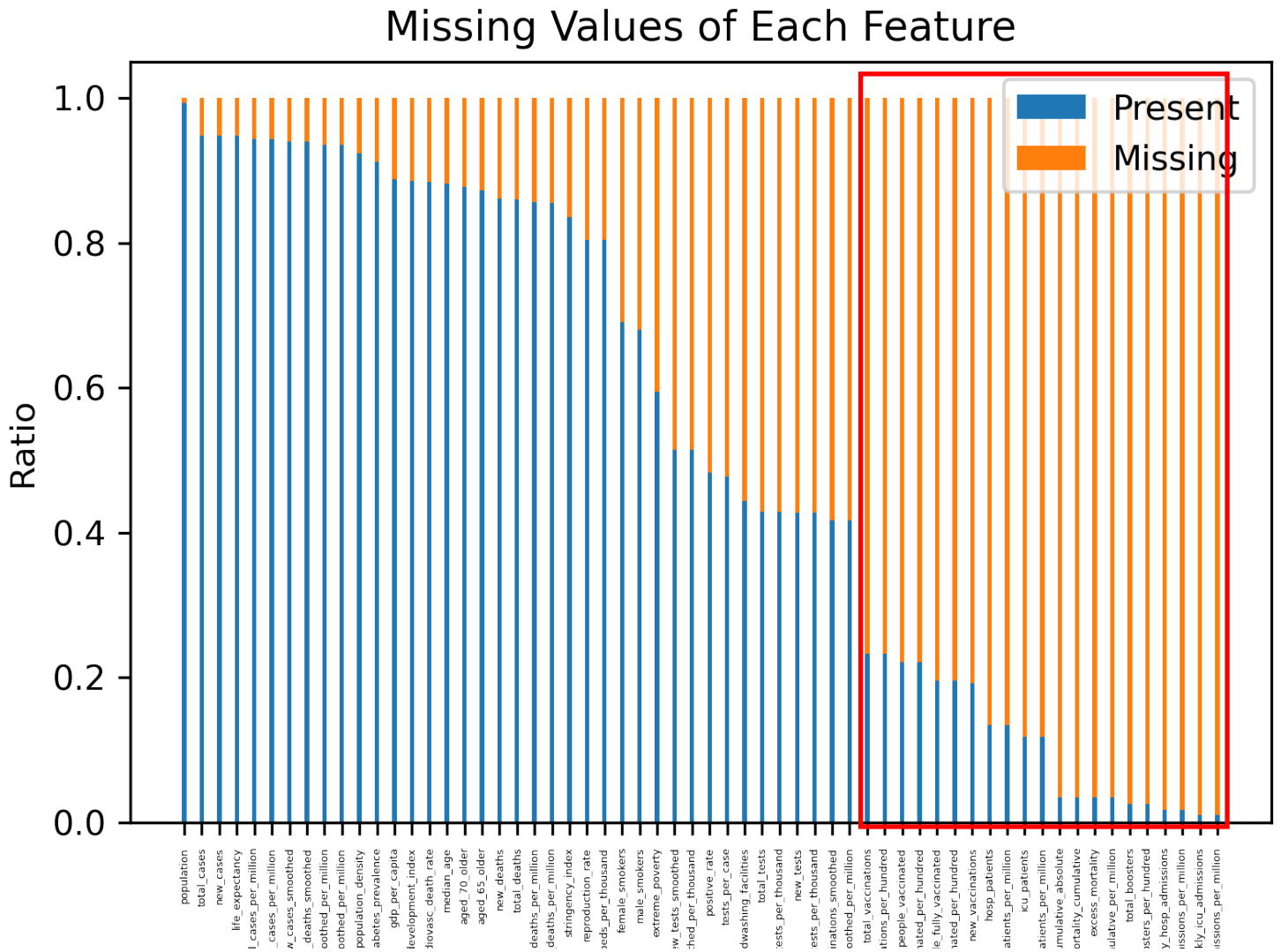


Some columns cannot be directedly applied to prediction

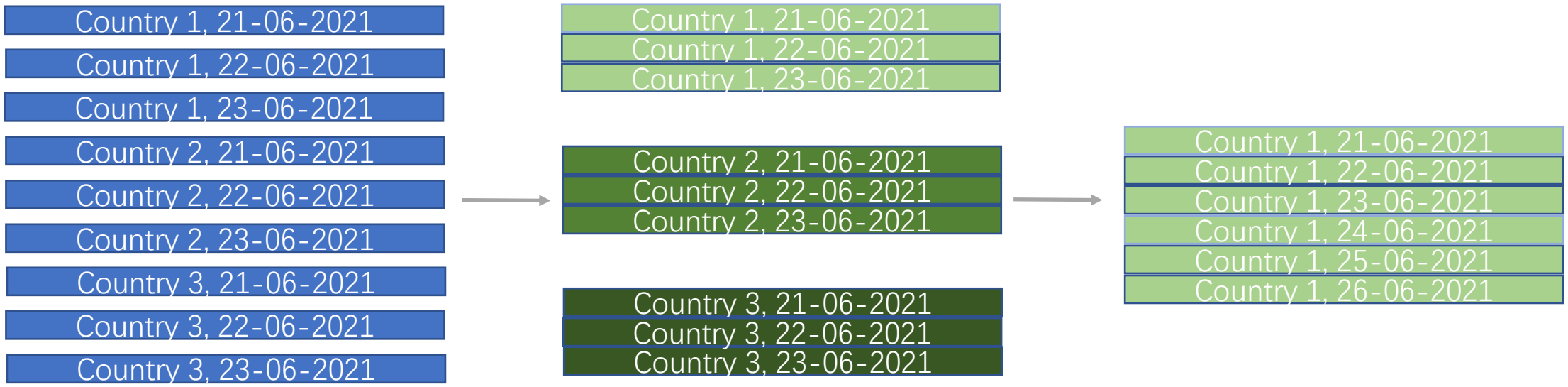
- Country indices, date
- Columns regarding new_cases

EDA

- 54 features are selected
- 19 features – almost 80% missed

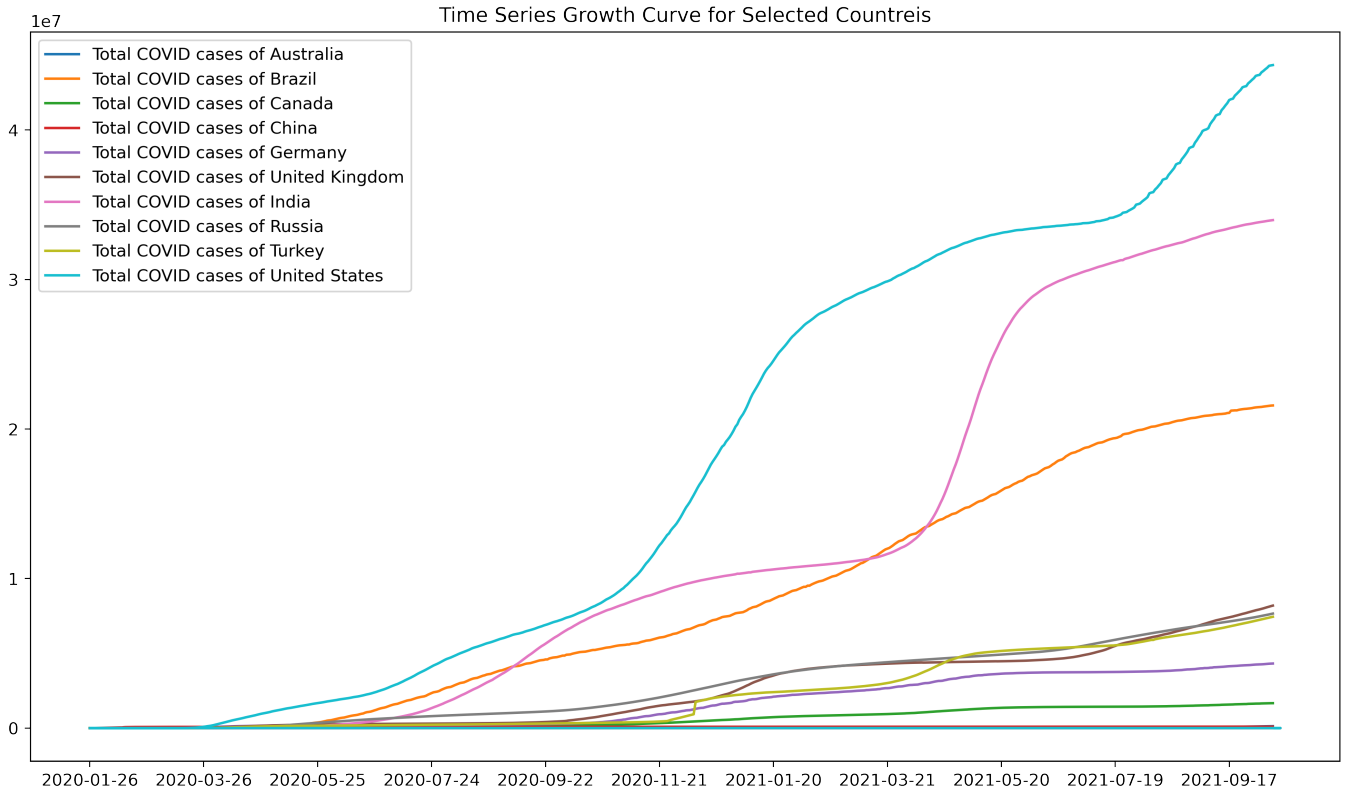


EDA



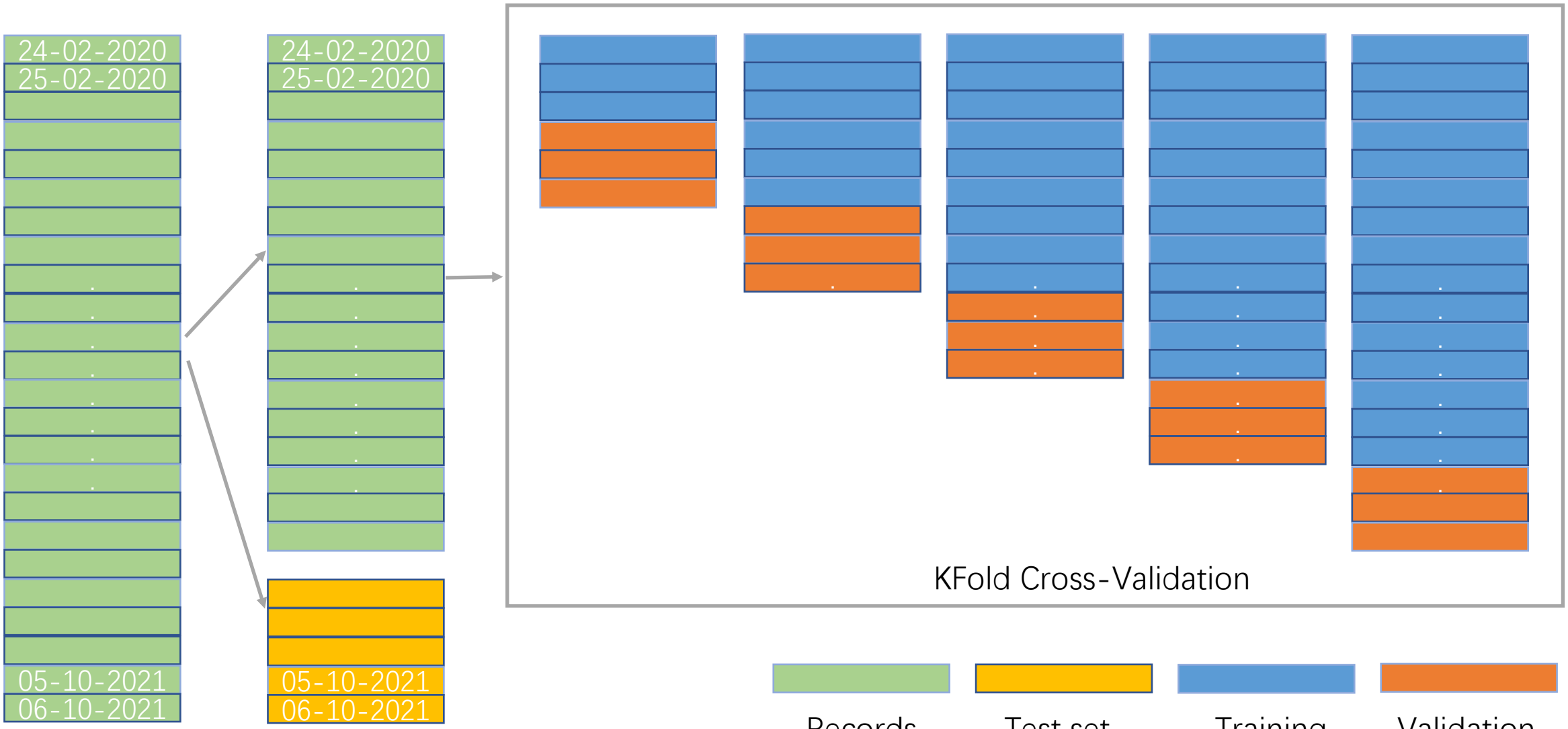
Raw Data → **Group** on country → **Time series** ordered by date

EDA



- Data: Time series of different countries

Data Splitting



Preprocessing

- Continuous
  - Some feature types: int
    - new_tests, aged_65_older, etc.
  - Standard Scaler

- Scaled within KFold splitting

Q&A

Thank you for your listening!