gdr

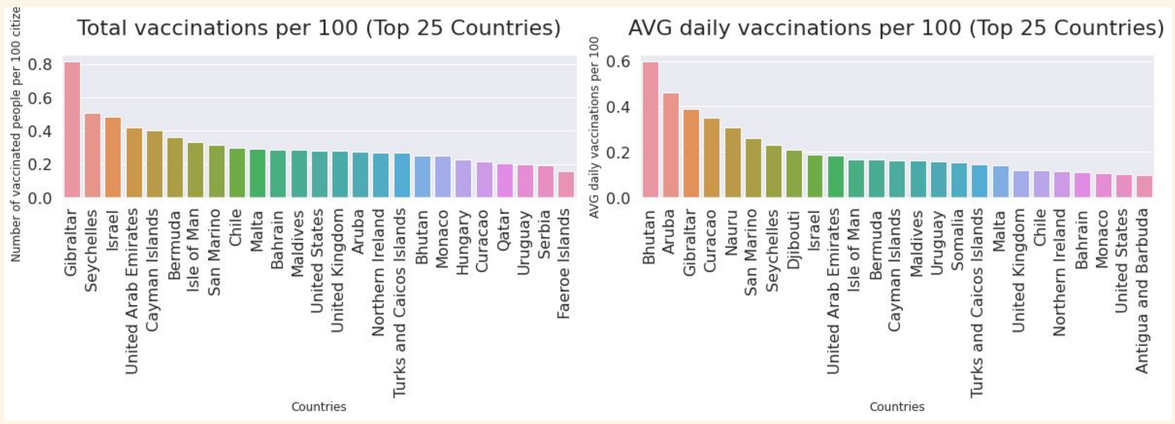
Data columns (total 15 columns): # Column Non-Null Count Dtype --- ------ -------------- ----- 0 country 14230 non-null object 1 iso_code 14230 non-null object 2 date 14230 non-null object 3 total_vaccinations 8486 non-null float64 4 people_vaccinated 7850 non-null float64 5 people_fully_vaccinated 5732 non-null float64 6 daily_vaccinations_raw 7101 non-null float64 7 daily_vaccinations 14033 non-null float64 8 total_vaccinations_per_hundred 8486 non-null float64 9 people_vaccinated_per_hundred 7850 non-null float64 10 people_fully_vaccinated_per_hundred 5732 non-null float64 11 daily_vaccinations_per_million 14033 non-null float64 12 vaccines 14230 non-null object 13 source_name 14230 non-null object 14 source_website 14230 non-null object
(126, 1409)

Time Series Analysis on World COVID-19 Vaccination Data

Konstantinos Georgiou
The University of Tennessee, Knoxville

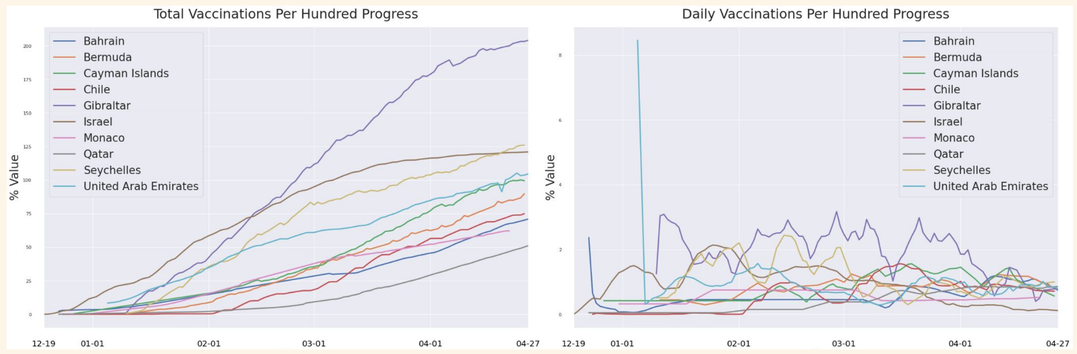
Motivation

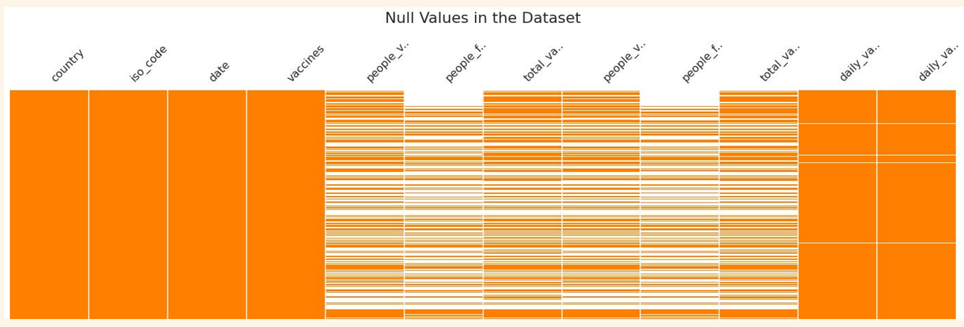
- The distribution of COVID-19 vaccines affects the health of billions of people as well as the state of the world's economies
- Many efforts have been made to extract useful insights from these data, but most of them are comparative analysis between two or more countries
- As of today, no method attempted to predict the number of daily vaccinations of all the countries by utilizing the correlations between them
- We introduce a method that uses Encoder-Decoder Long Short-Term Memory Networks With Multivariate Inputs and Walk-Forward Validation of ten days.



Dataset

- Contains daily vaccination data for 193 different countries and 135 dates
- 14230 15-dimensional data from which 8 dimensions where used
- Many NULLs most of which can be inferred from other existing values





Workflow

New Data every day → Dataset → Preprocessing → Training → Evaluation

Data columns (total 15 columns, 14230 Entries):

Column	Non-Null Count
country	14230 non-null
iso_code	14230 non-null
date	14230 non-null
total_vaccinations	8486 non-null
people_vaccinated	7850 non-null
people_fully_vaccinated	5732 non-null
daily_vaccinations_raw	7101 non-null
daily_vaccinations	14033 non-null
total_vaccinations_per_hundred	8486 non-null
people_vaccinated_per_hundred	7850 non-null
people_fully_vaccinated_per_hundred	5732 non-null
daily_vaccinations_per_million	14033 non-null
vaccines	14230 non-null
source_name	14230 non-null
source_website	14230 non-null

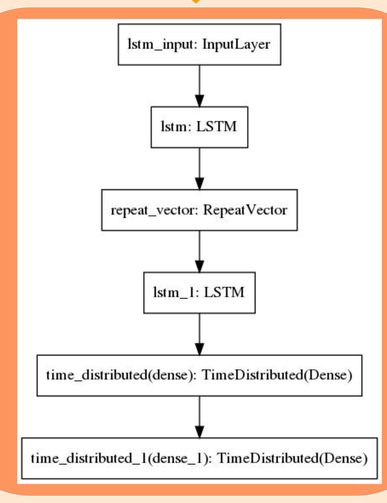
Countries metadata Dataset → Join the two datasets and drop rows with missing countries → Recalculate the per hundred values using the population of the countries → Drop recent dates when many countries are missing → Normalize → Infer nulls for the non- percentage float columns from their relationships

Columns kept:
- **iso_code**
- **Date**
- daily_vaccinations
- total_vaccinations
- people_vaccinated
- people_fully_vaccinated
- total_vaccinations_per_hundred
- people_vaccinated_per_hundred
- people_fully_vaccinated_per_hundred
- daily_vaccinations_per_million

Pivot the Countries in the Dataset by grouping rows by date
New Columns (1409 Columns, 126 Entries):
- Date
- daily_vaccinations_country_1
- …
- daily_vaccinations_country_N
- people_vaccinated_country_1
- …
- people_vaccinated_country_N
- …

- Group the Dataset in 10-day windows and discard the excess days
  - 126 days → 12 10-day groups + 6 days discarded
- Split in Train (~80%) and Test (~20%) set

- Group the Dataset in 10-day windows and discard the excess days
  - 126 days → 12 10-day groups + 6 days discarded
- Split in Train (~80%) and Test (~20%) set



- Predict the Daily Vaccinations per hundred people on the Test set
- Each time, use the 10 previous days to predict the 10 next

- Gather Predictions
- Recreate the test set and the predictions in the original tabular format

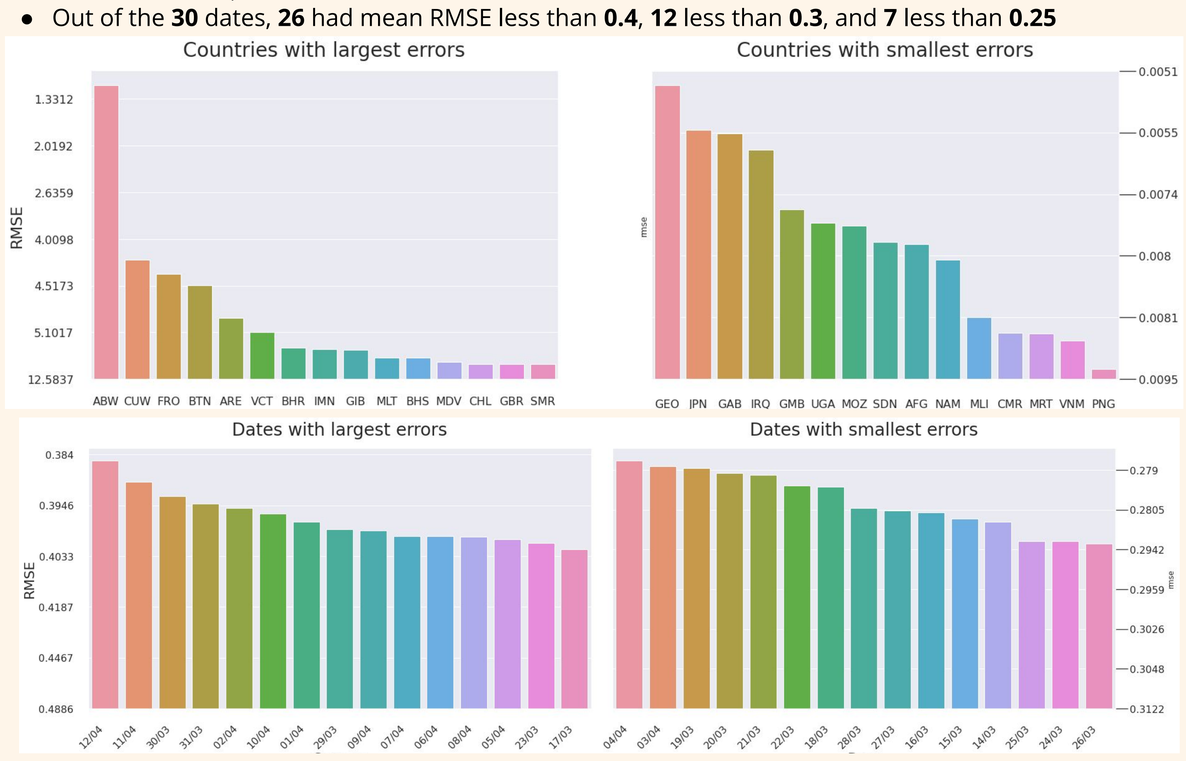
Calculate RMSE:
- The average of all the predictions
- Average per Country
- Average per Date

Visualize:
- Histograms with countries and dates that have the highest and lowest errors
- The test set and predicted daily vaccinations per hundred timeseries for specific countries

Evaluation

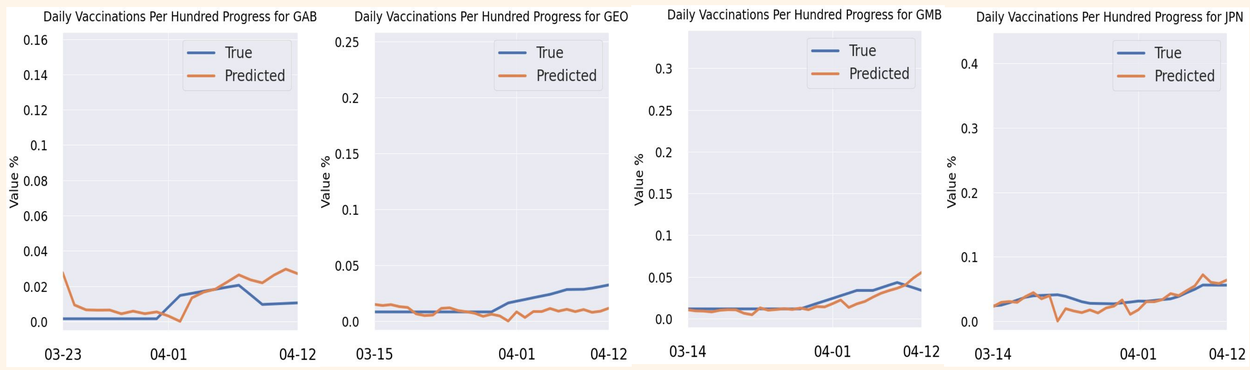
Average RMSE: ***0.31888512707360384***

- Out of the **170** countries, **161** had mean RMSE less than **1.0**, **149** less than **0.5**, and **77** less than **0.1**
  - Min: **0.00032**, Max: **12.5837**
- Out of the **30** dates, **26** had mean RMSE less than **0.4**, **12** less than **0.3**, and **7** less than **0.25**



Results

True VS Predicted Daily Vaccinations per million for 8 countries



Conclusions

- Proved that it's possible to predict the number of daily vaccinations of all the countries by finding correlations between their historical data

Future Work

Incorporate static features such as the type of vaccines, health expenditure per gdp and the number of physicians per million all of which are included in the metadata Dataset