Simultaneous Time Series Forecasting on the World's COVID-19 Daily Vaccinations
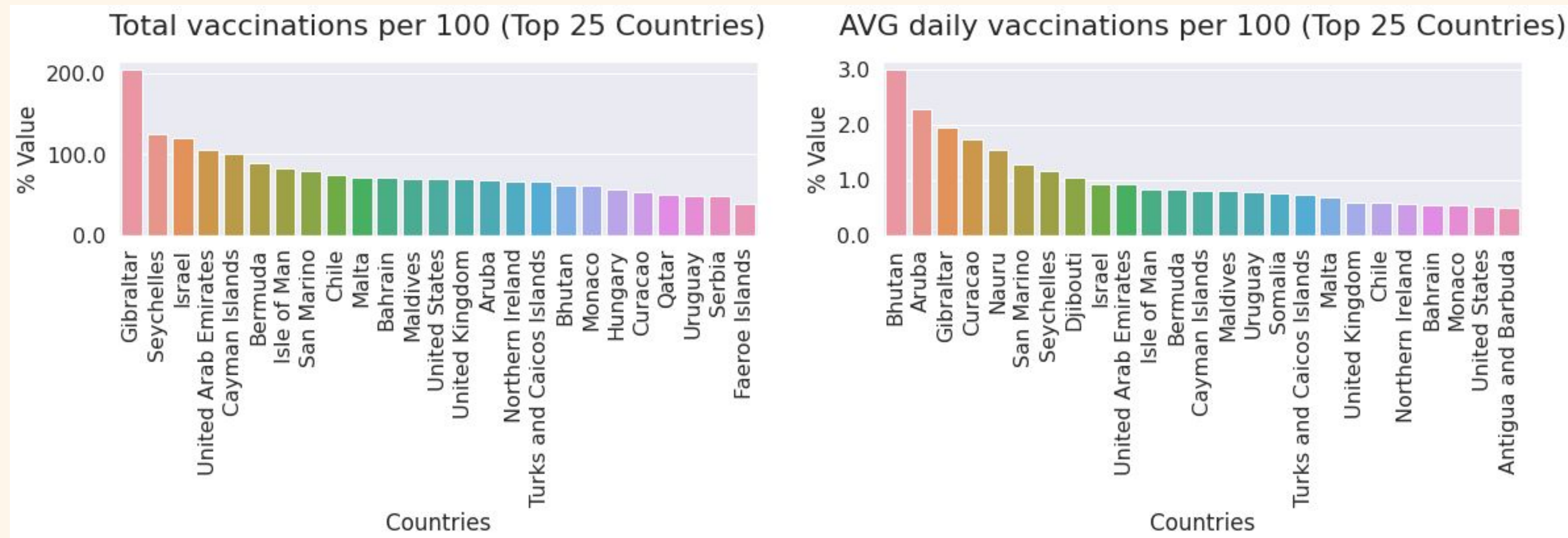
Konstantinos Georgiou

Course Advisor: Michela Taufer

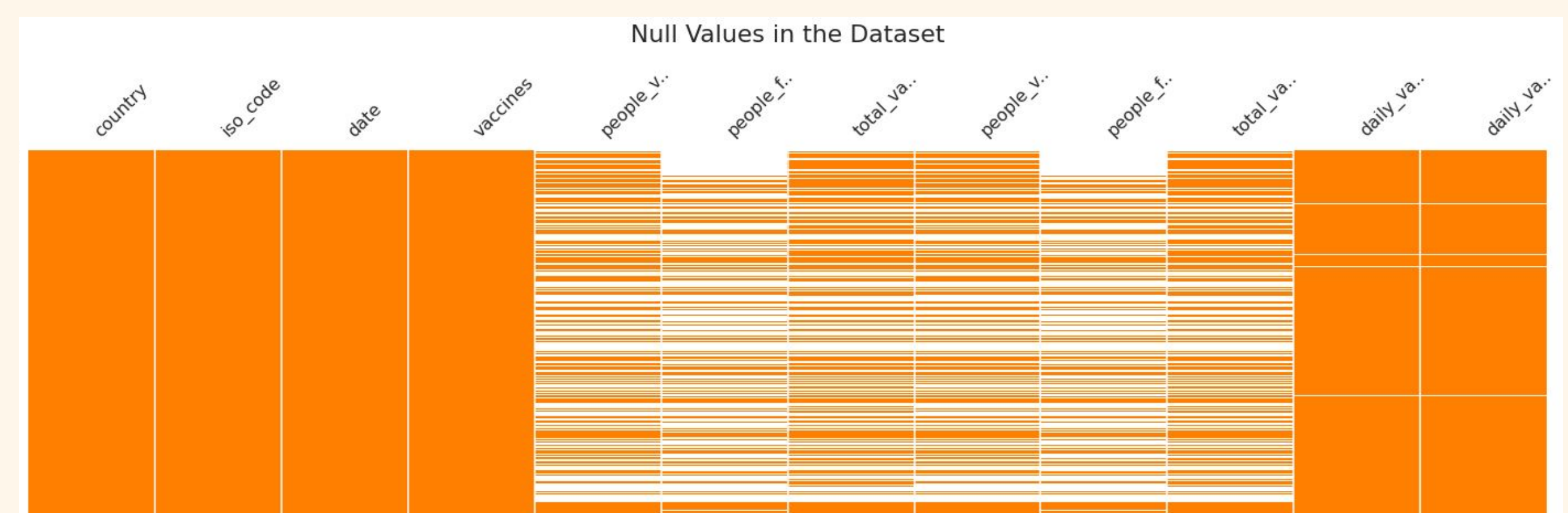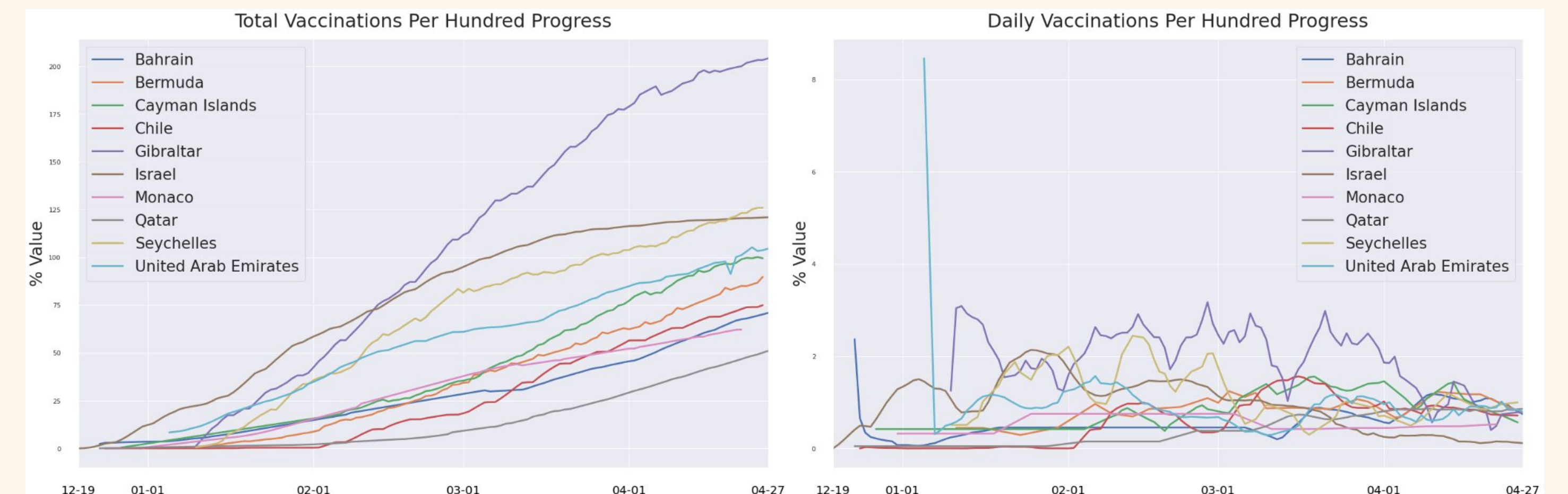The University of Tennessee, Knoxville

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 simultaneously 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.
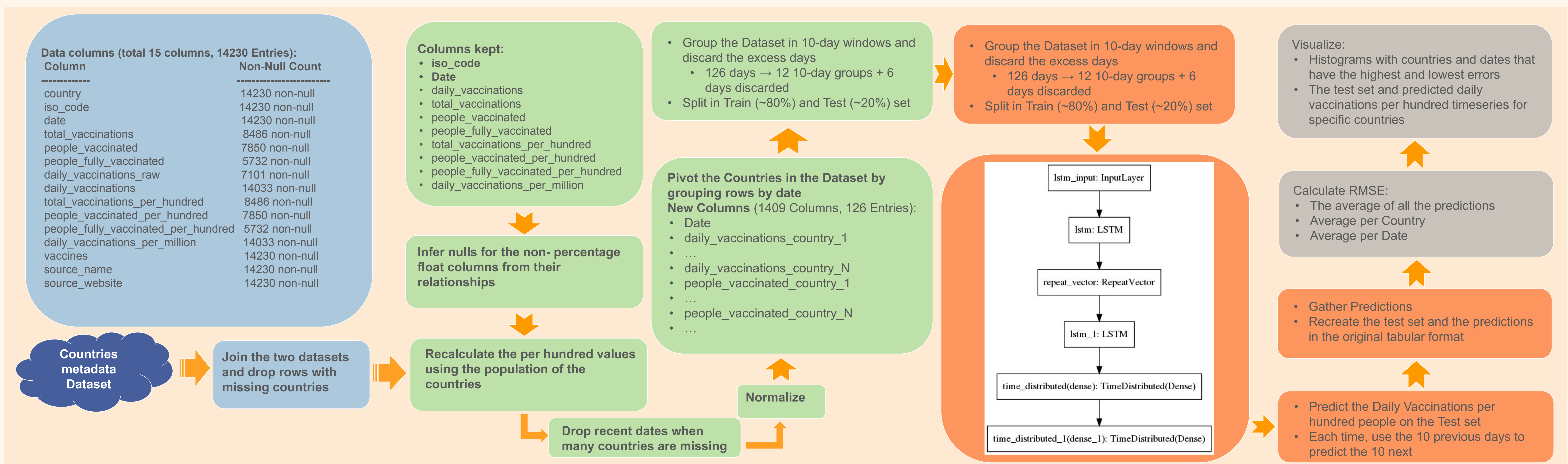

Dataset

- Contains daily vaccination data for 193 different countries and 135 dates
- 14230 15-dimensional data from which 8 dimensions where used
- The dataset has many null values, most of which can be inferred from other 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

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

Infer nulls for the non-percentage float columns from their relationships

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
- …

Normalize

- 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

lstm_input: InputLayer
lstm: LSTM
repeat_vector: RepeatVector
lstm_1: LSTM
time_distributed(dense): TimeDistributed(Dense)
time_distributed_1(dense_1): TimeDistributed(Dense)

- 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: *0.246174*

- Out of the **170** countries, **164** had mean RMSE less than **1.0**, **153** less than **0.5**, and **73** less than **0.1**
  - Min: **0.00023**, Max: **5.5133**
- Out of the **30** dates, **25** had mean RMSE less than **0.30**, **15** less than **0.25**, and **11** less than **0.20**
  - Min: **0.16928**, Max: **0.35180**

Results

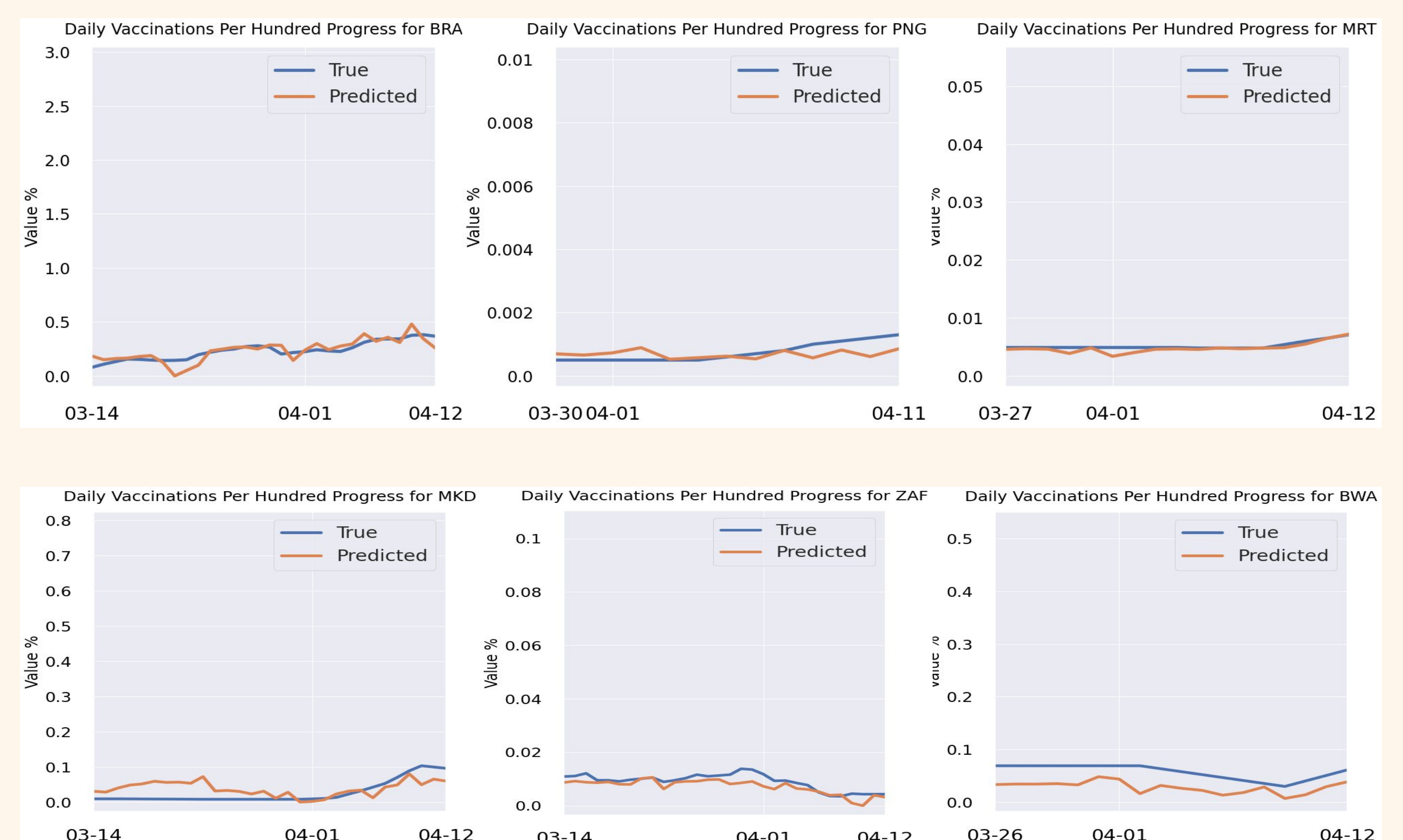**True VS Predicted Daily Vaccinations per million for 6 countries**

Conclusions

Proved that it's possible to simultaneously 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 **Vaccines** used, **Health Expenditure per GDP** and the **Number of Physicians per Million**, all of which are included in the metadata Dataset