PREDICTING DENGUE CASES

24th August 2023

Chloe, Nicole, Gloria

CONTENTS

01

CONTEXT & PROBLEM STATEMENT

CONTEXT

- Dengue fever - major health threat in tropical regions like Singapore
- Wolbachia project - to limit dengue virus transmission
- Complex factors influence dengue transmission - weather: rainfall and temperature
- Online search for dengue-related terms - indicative of disease prevalence

Singapore

Dengue surge fuelled by more mosquitoes, re-emergence of previously uncommon virus serotype: Experts

A worker wearing a face mask fumigates a construction site to prevent the spread of dengue fever in Singapore on Apr 17, 2020. (Photo: AFP/Roslan Rahman)

SINGAPORE: As dengue cases in Singapore continue to spike, the chance of a major outbreak this year looks to be an increasing possibility.

As of the week ending Apr 9, a total of 3,979 cases have been recorded this year. This is in contrast to a total of 5,258 cases in 2021.

PROBLEM STATEMENT

Part 1:
Short-term prediction model

Develop a reasonably accurate model to **predict dengue case numbers for the subsequent 3 months** by using Climate data and Google search trends.

Having an accurate forecast of upcoming dengue cases would **allow mitigating actions to be taken by NEA**

Part 2:
Cost-Benefit Analysis of Wolbachia Implementation

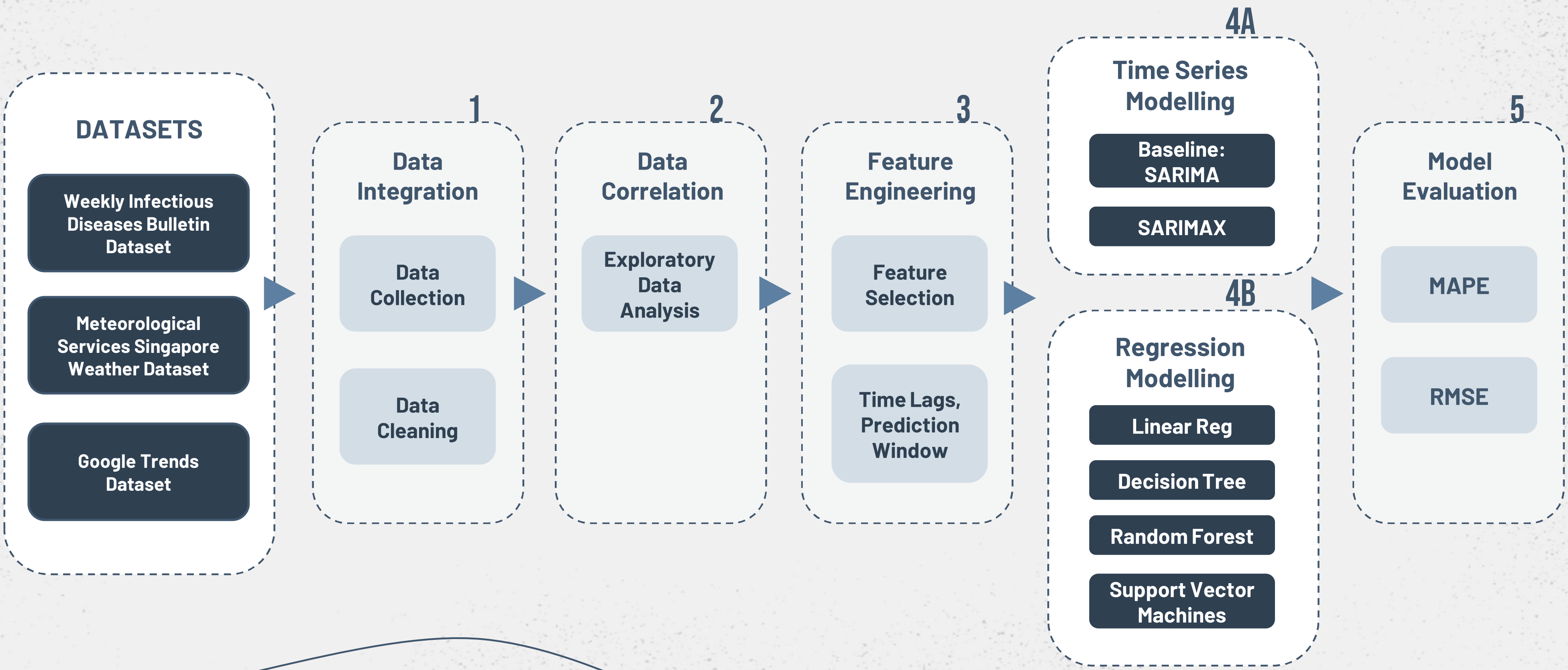
Perform a **cost-benefit analysis** of Project Wolbachia and determine the **decision threshold** for rolling out Project Wolbachia.

02

METHODOLOGY

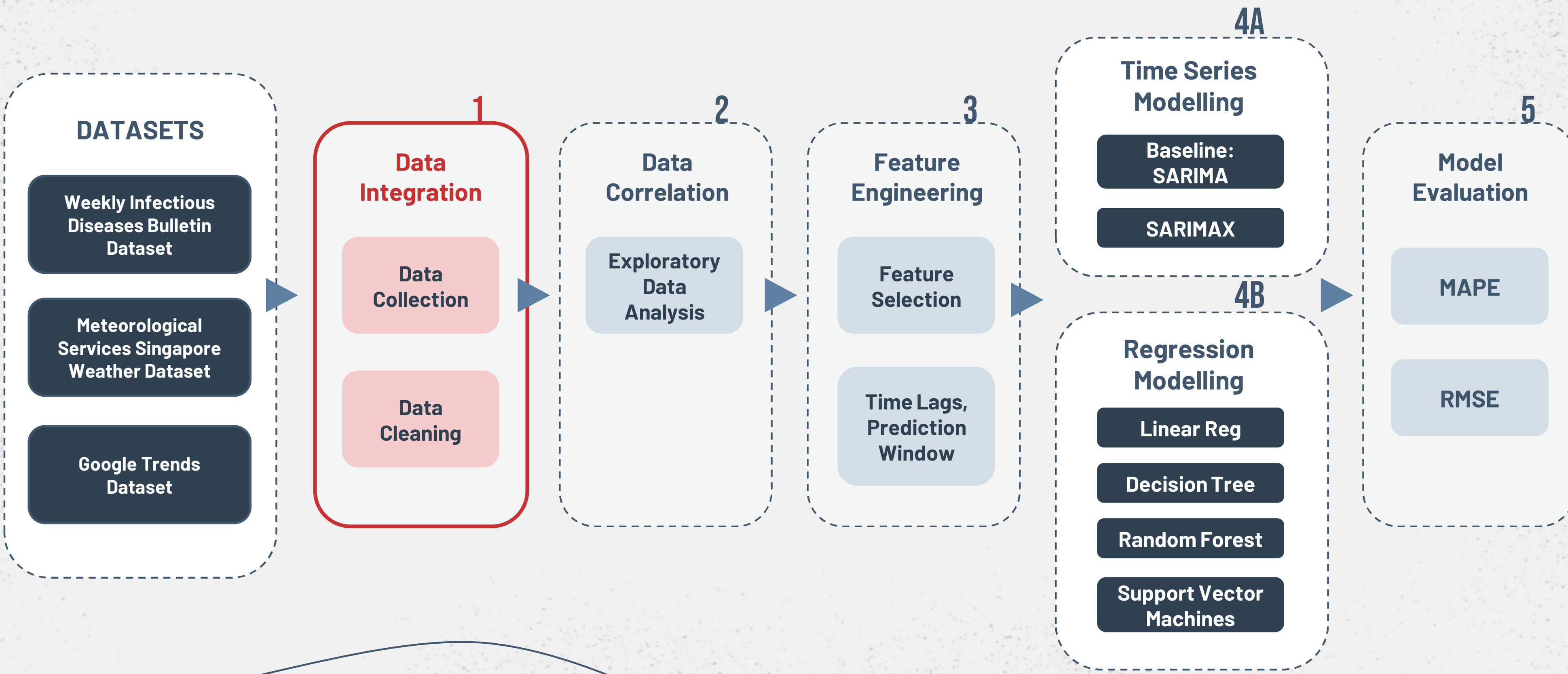
METHODOLOGY

Workflow, Models and Metrics



METHODOLOGY

Workflow, Models and Metrics



DATA COLLECTION, CLEANING

Data Collection

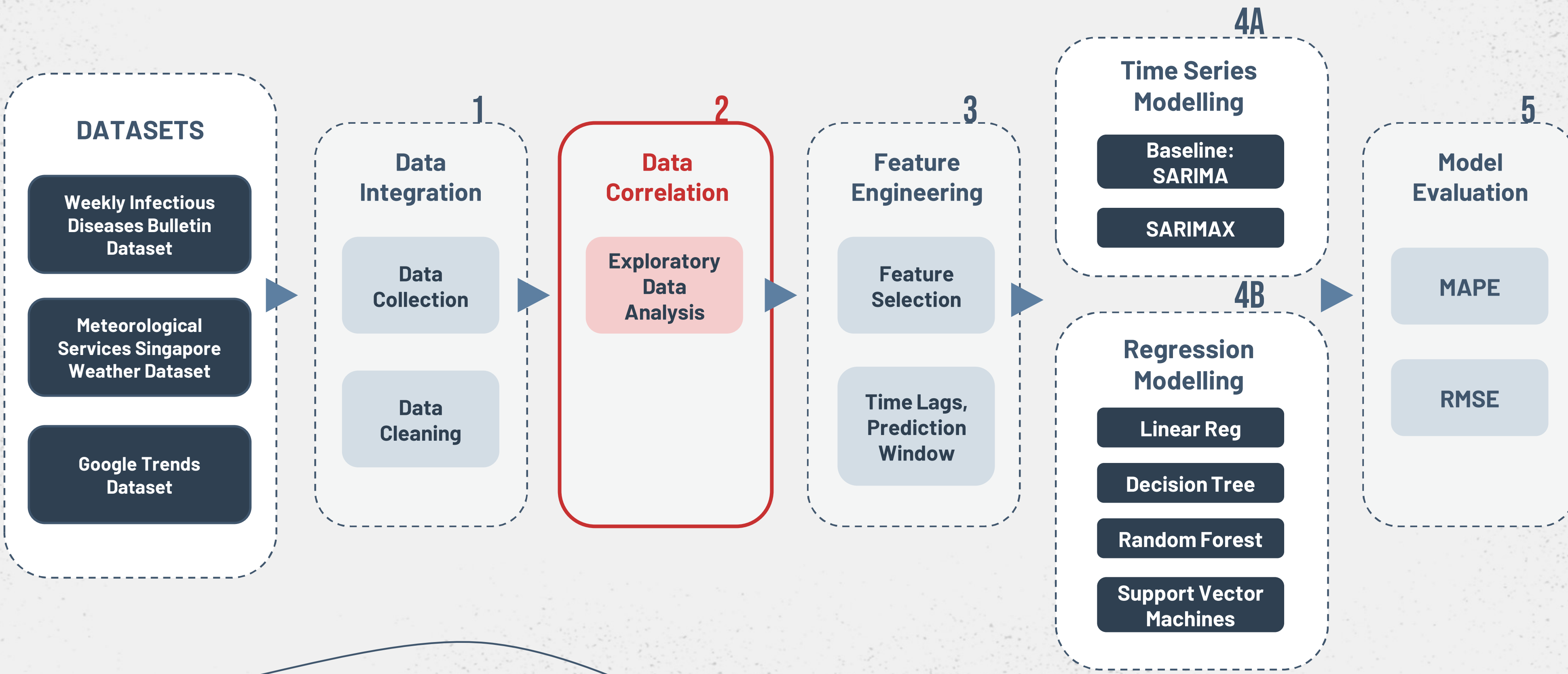
- Data download was straightforward for: **Diseases Dataset** and **Google Trends Dataset** (Weekly granularity)
- For **Weather Dataset**, a function was created to concatenate all the months of daily weather data into a single dataframe

Data Cleaning

- Conversion of date-time format, setting date as index, resampling to weekly granularity
- **Weather dataset:**
Imputing nulls using iterative imputer
- **Dengue Cases dataset:**
Imputing nulls with mean of week before and after

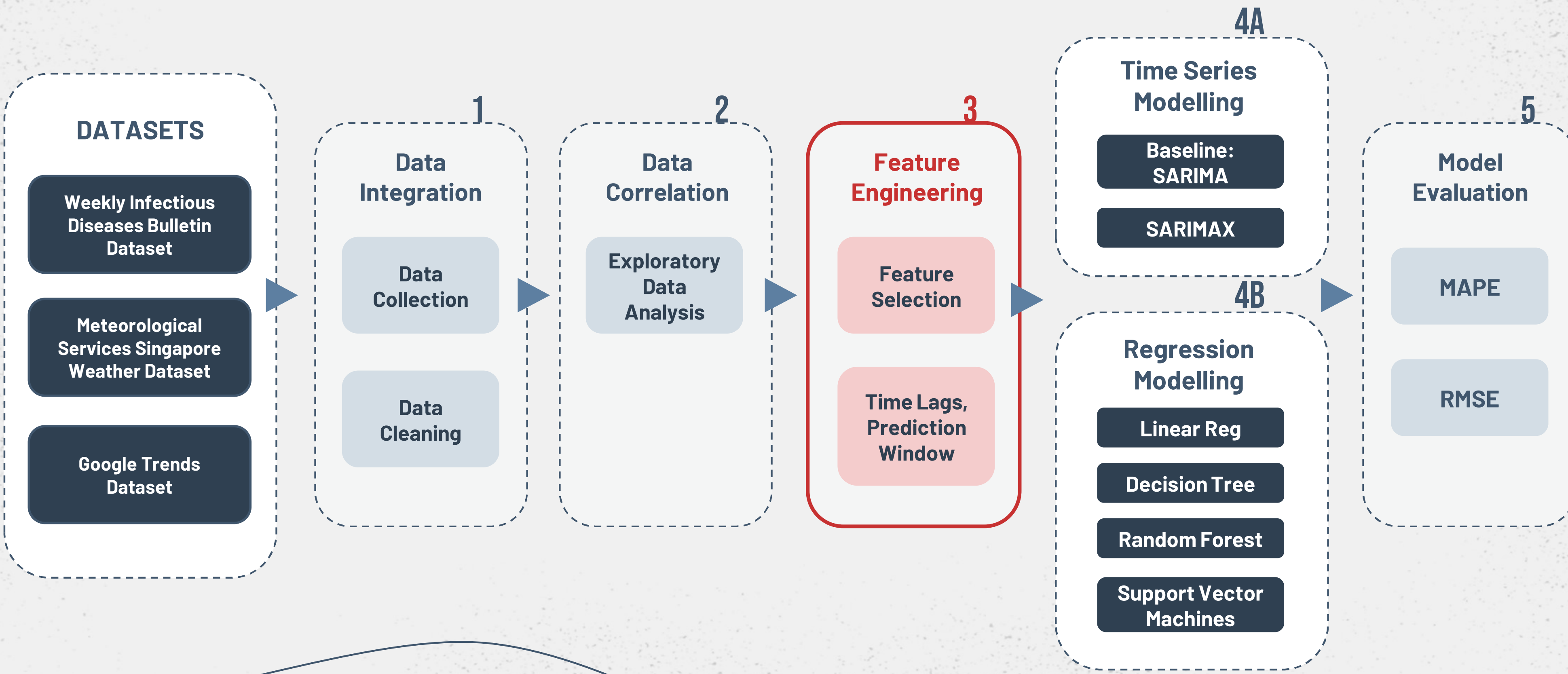
METHODOLOGY

Workflow, Models and Metrics



METHODOLOGY

Workflow, Models and Metrics



FEATURE ENGINEERING

Feature Selection

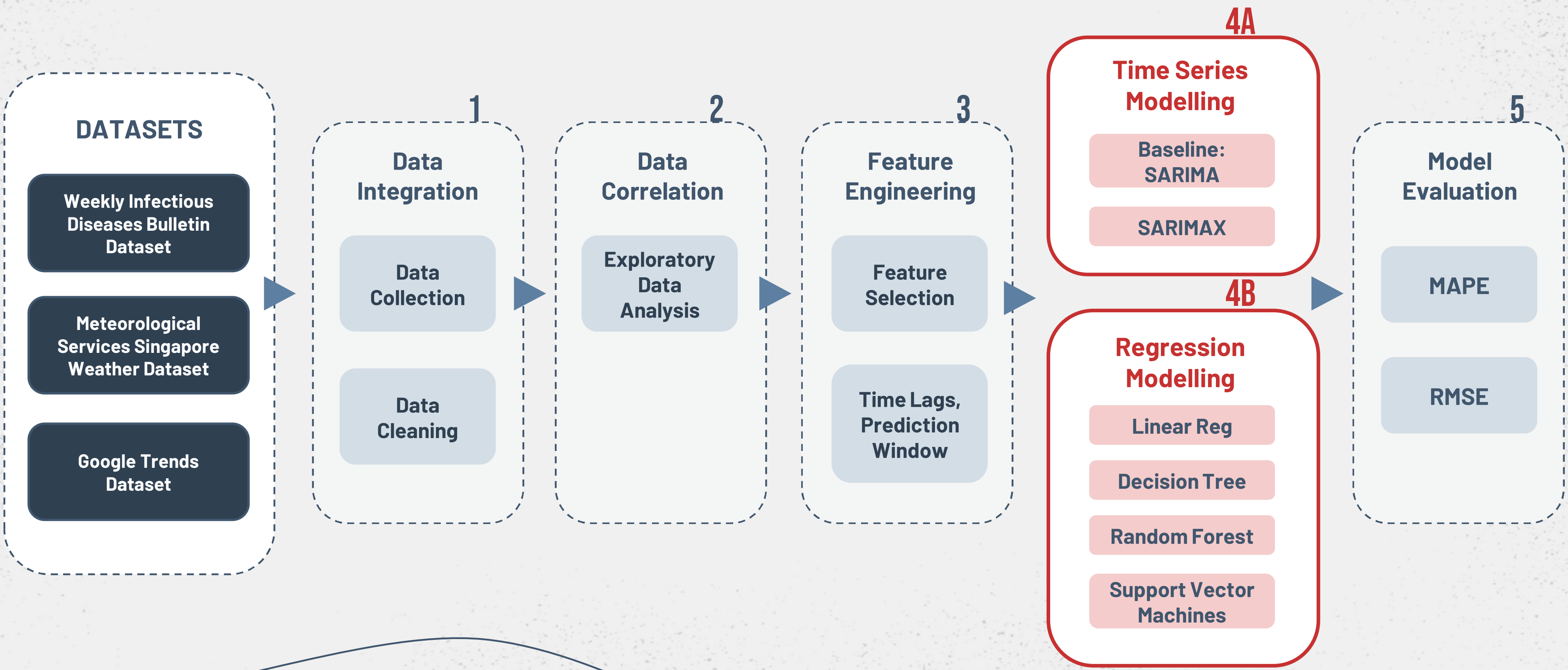
- **Feature engineering rainfall squared features**
  due to complex relationships observed between rainfall and dengue
- **Dropping features that display multicollinearity with other features**

Time Lags, Prediction Window

- EDA Findings: Weather features take about 3-8 weeks to impact dengue numbers
- **Prediction window should be minimally 12 weeks lead time**
- **Lagging of our exogenous X-features** for 2 prediction windows:
  - 52 weeks (1 year) and
  - 12 weeks (3 months)
  - Additional 3 weeks for weather features

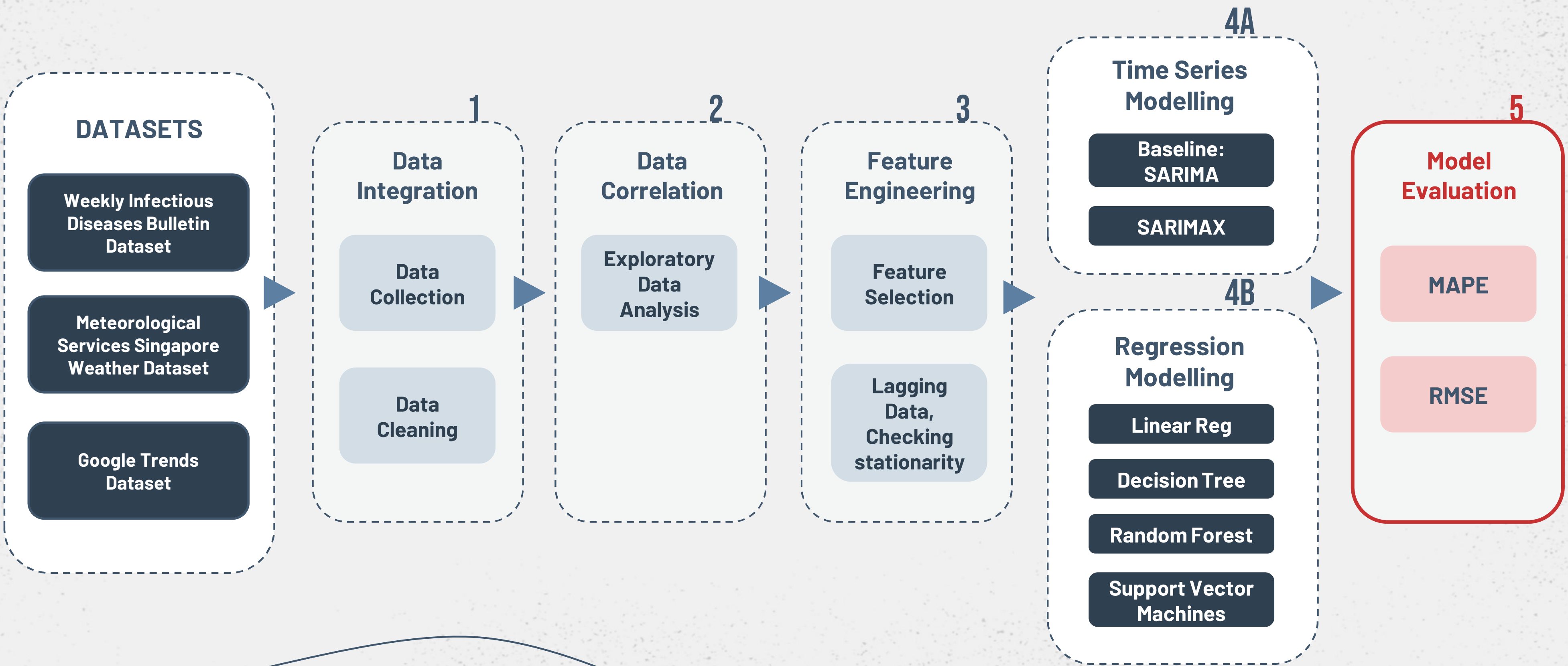
METHODOLOGY

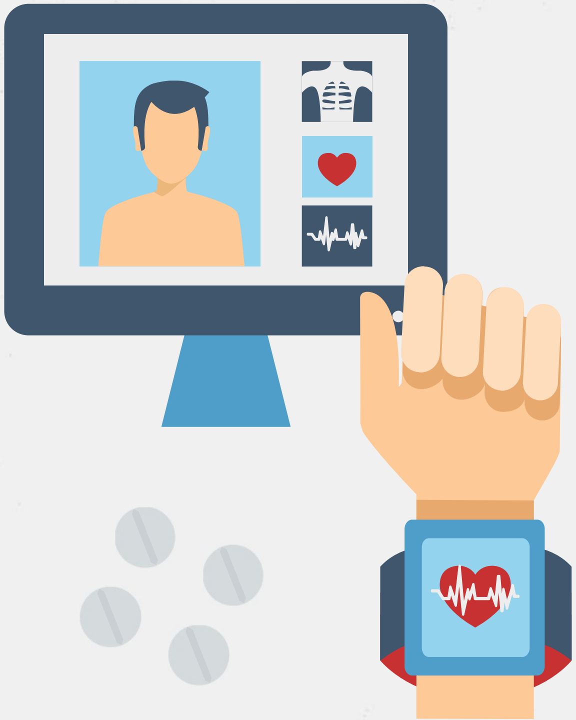
Workflow, Models and Metrics



METHODOLOGY

Workflow, Models and Metrics

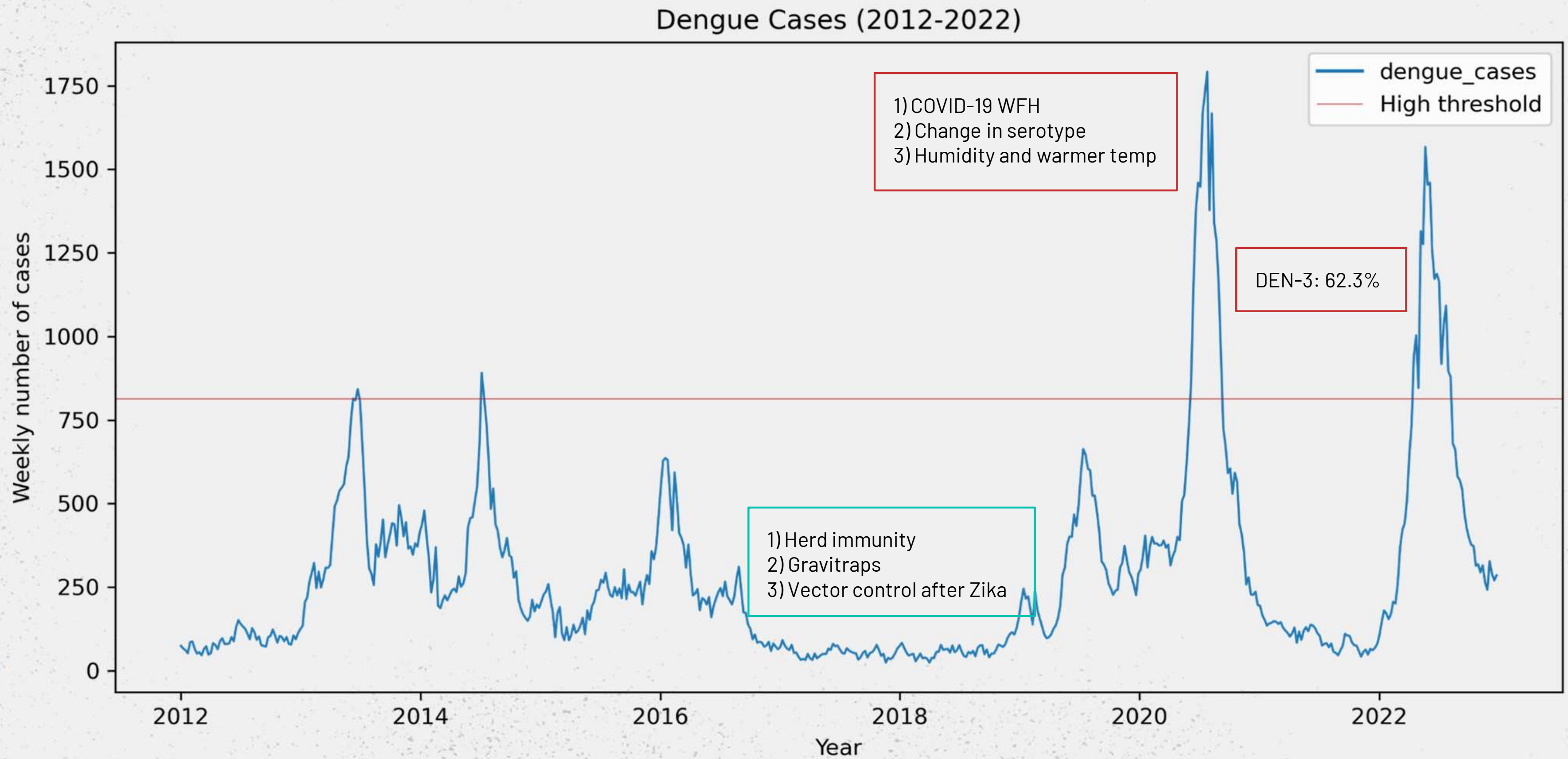




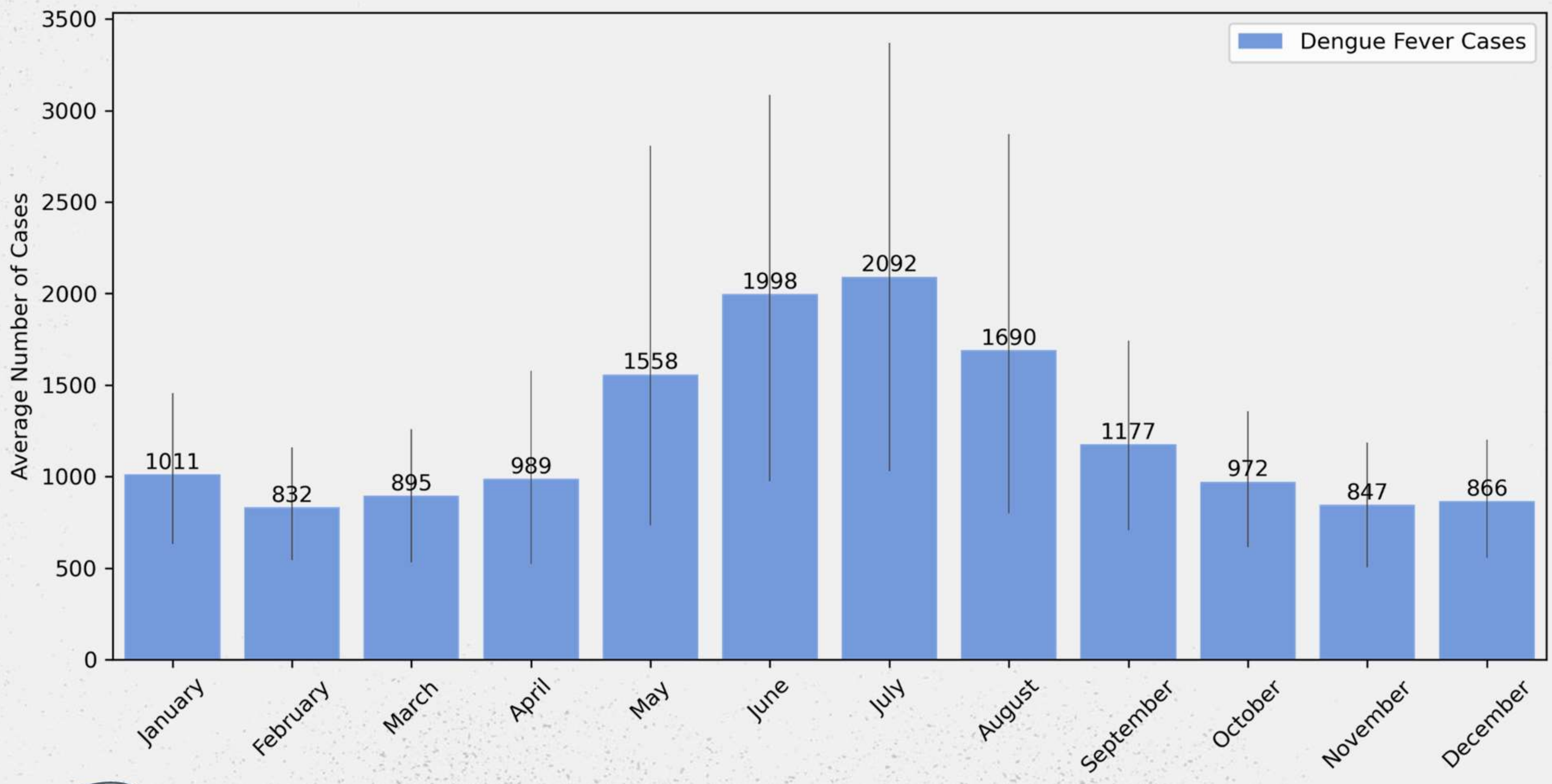
03

EXPLORATORY DATA ANALYSIS (EDA)

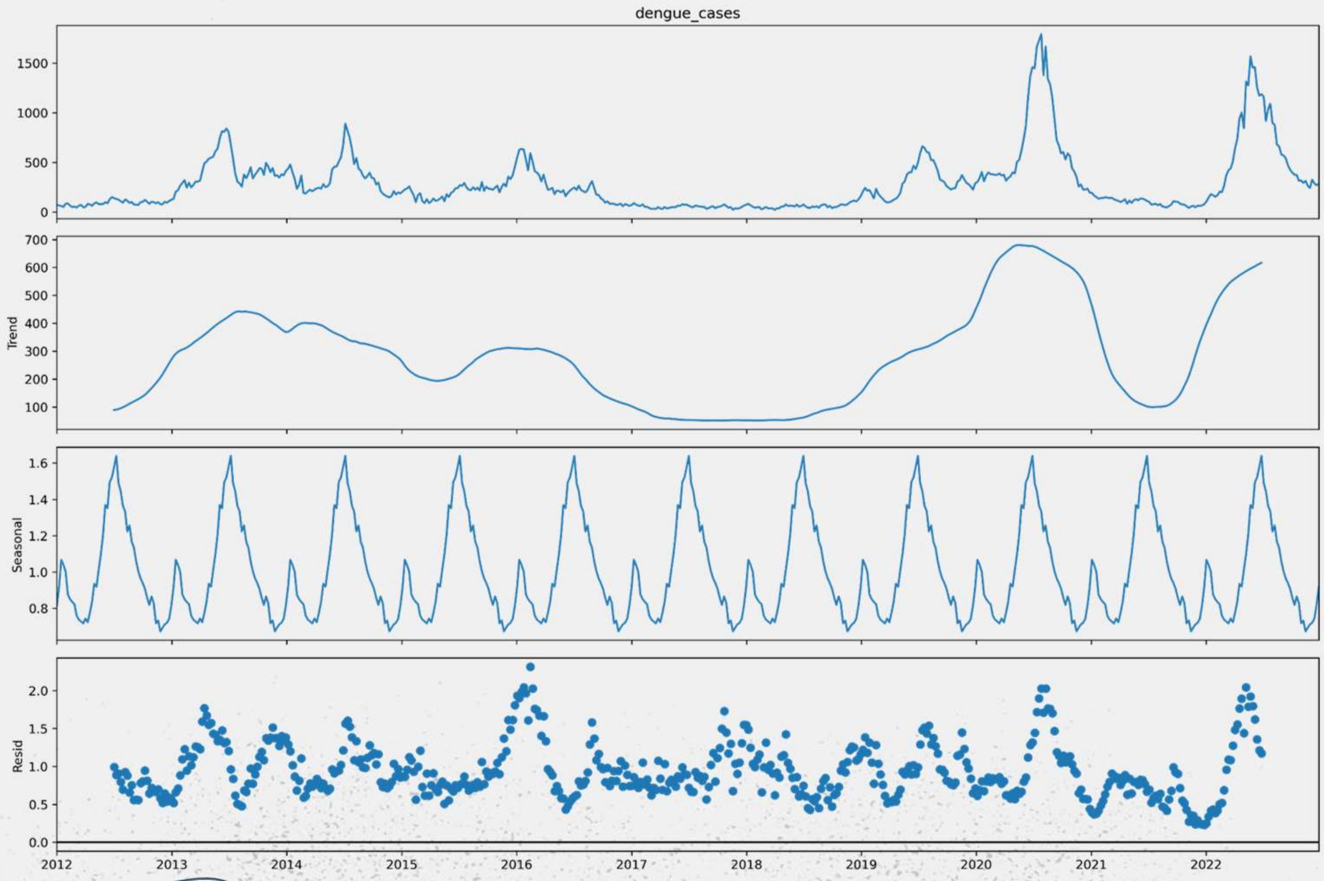
Dengue cases (2012-2022)



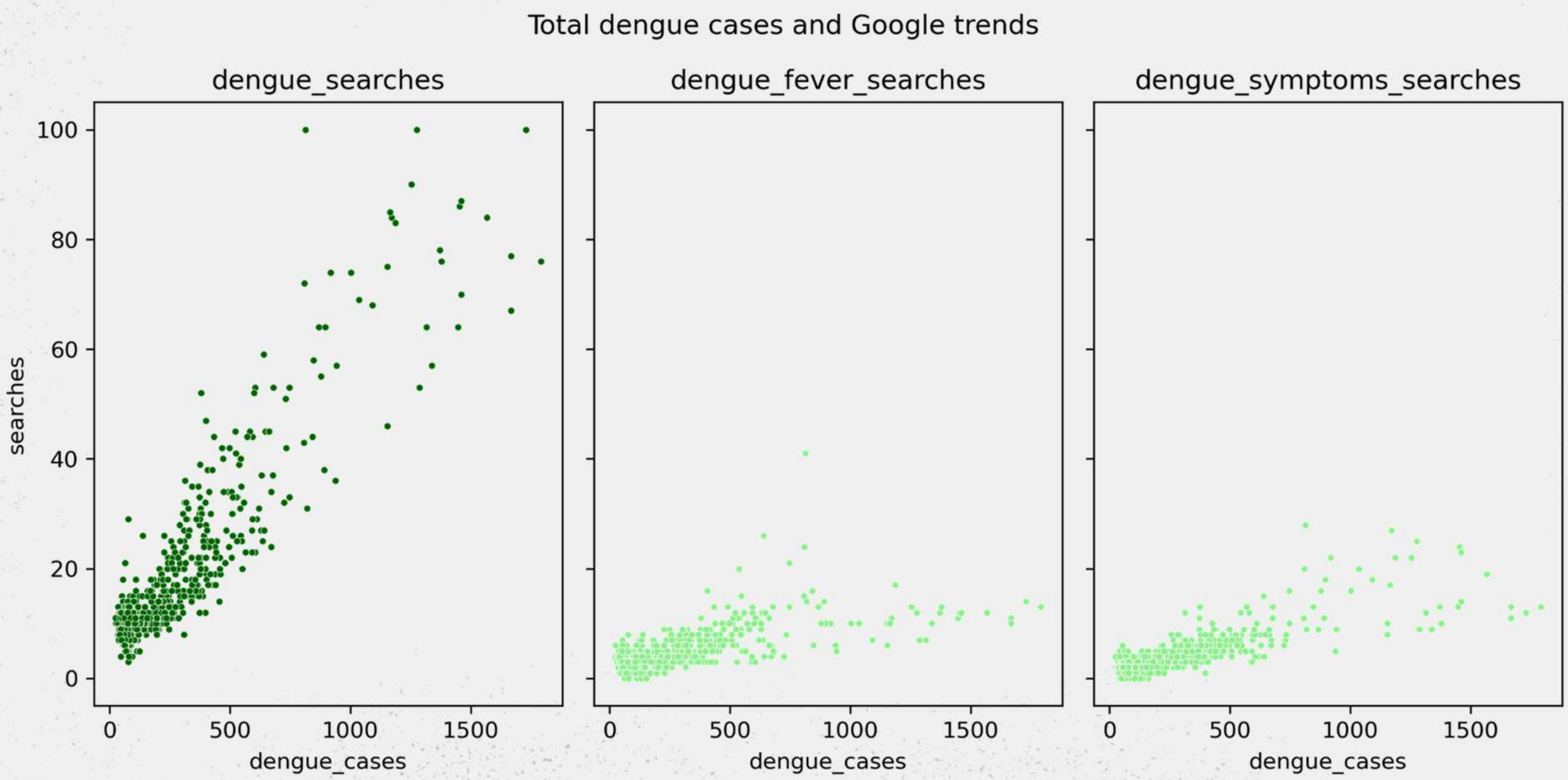
Dengue annual trend



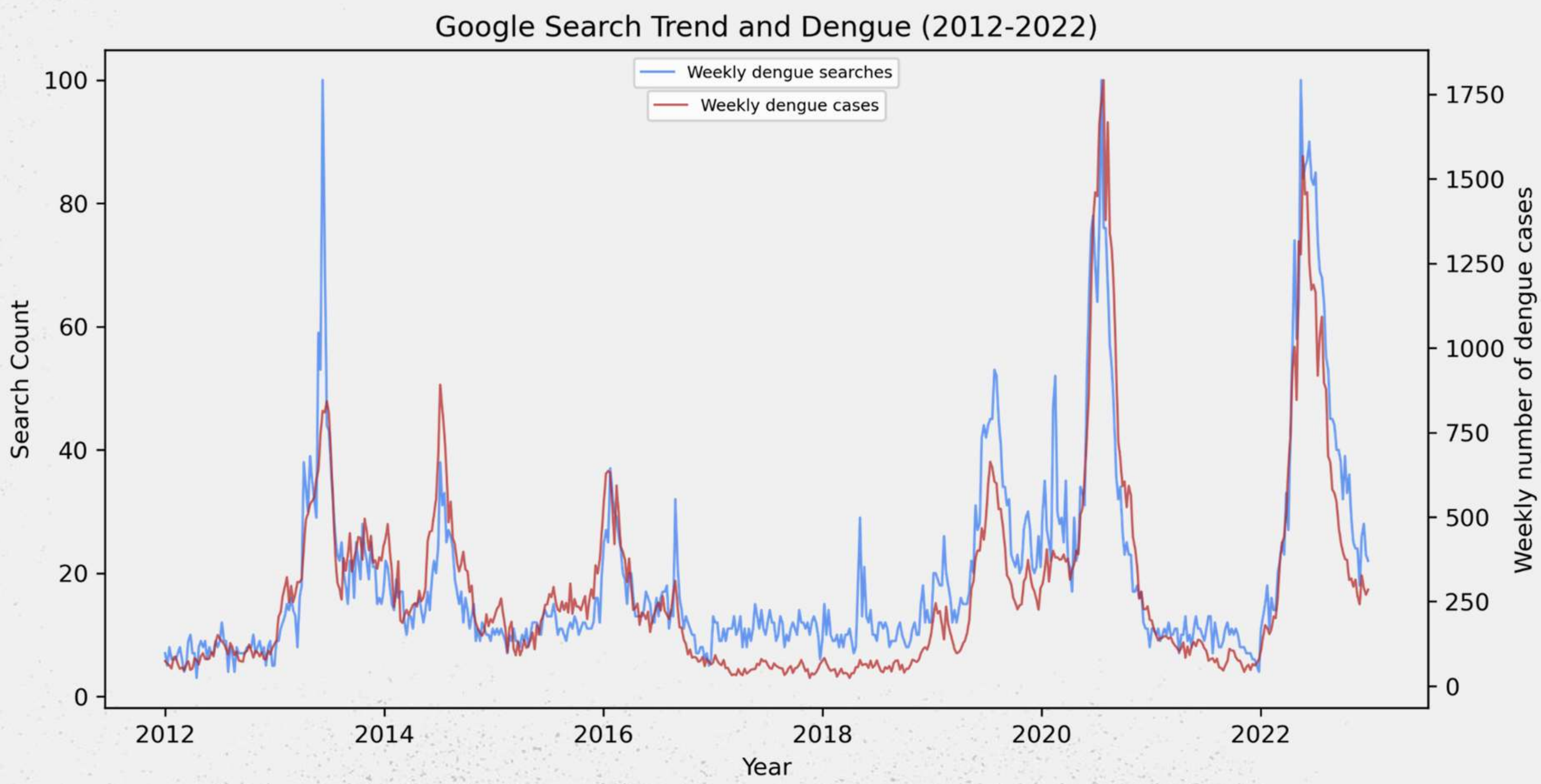
Dengue: Time-series decomposition



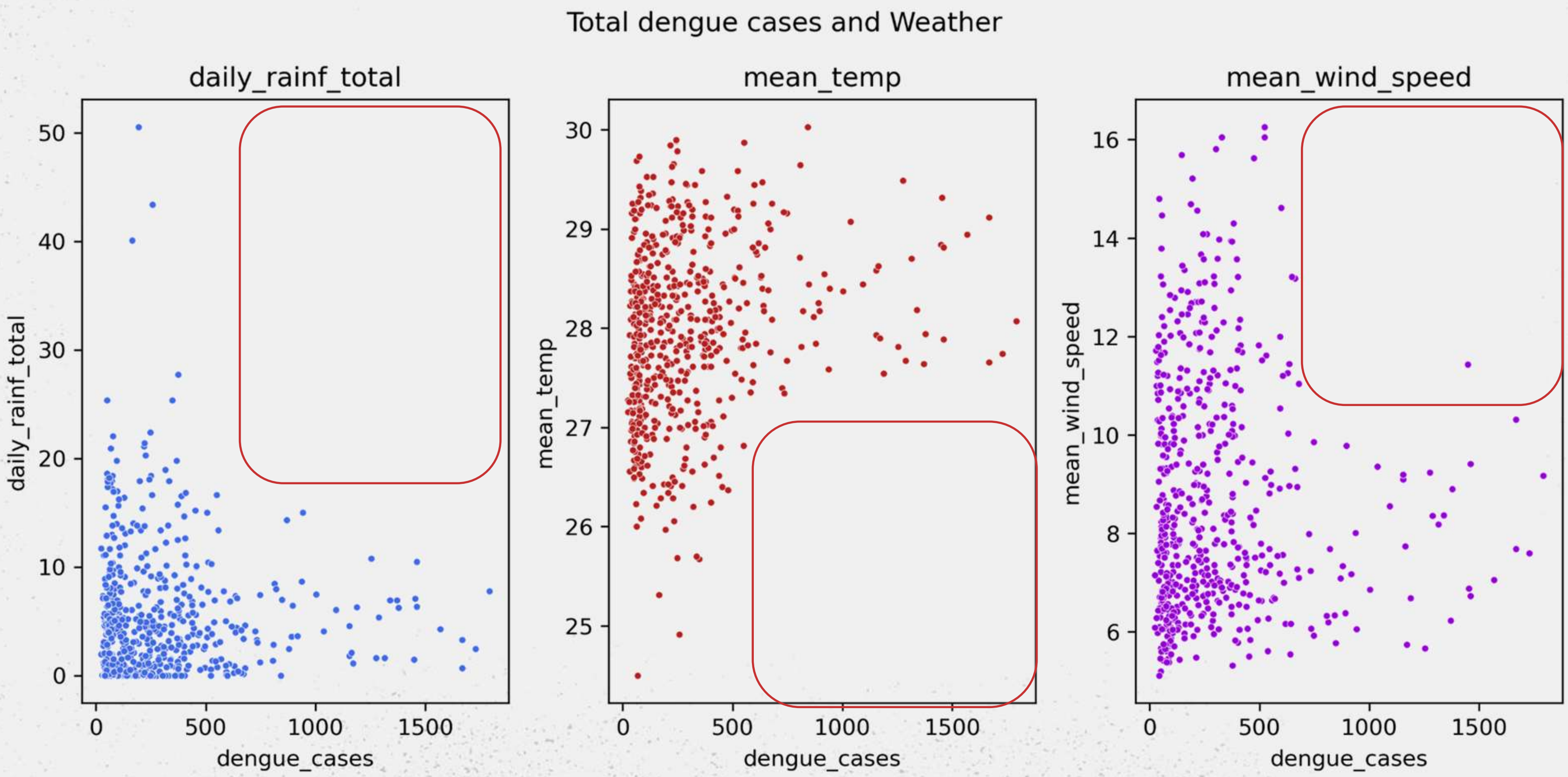
Total dengue cases vs google trends



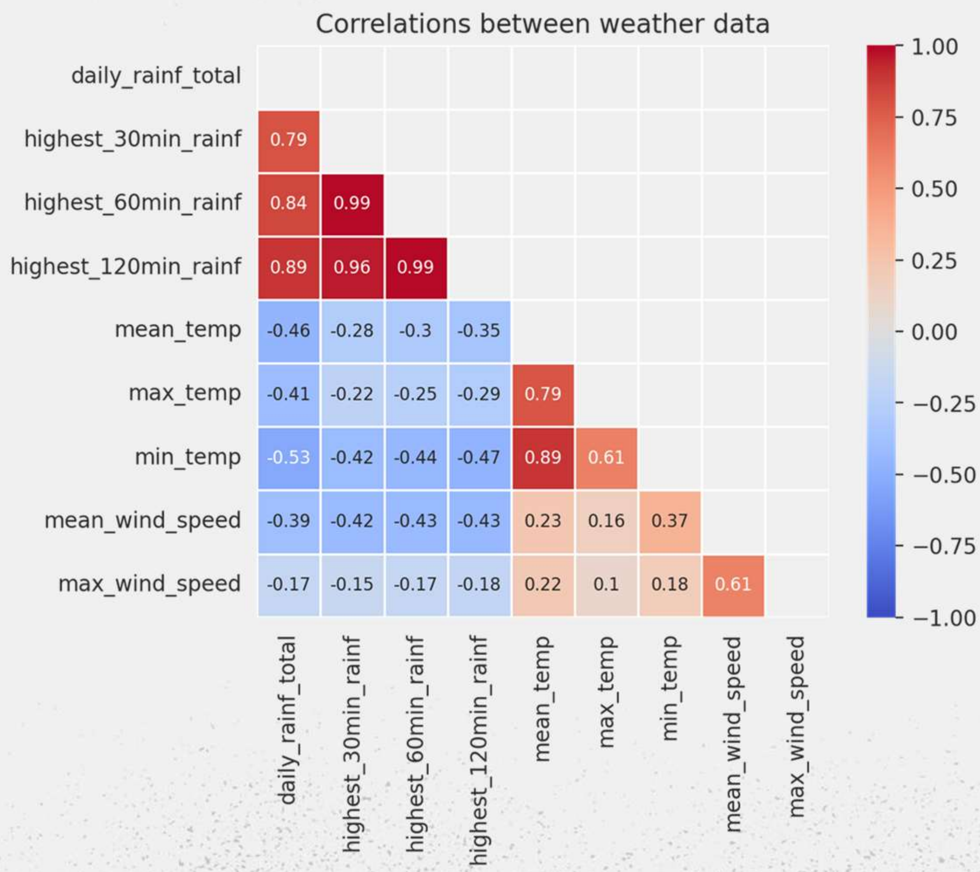
Dengue and Google Search trends



Total dengue cases vs rainfall, temperature and windspeeds



Correlations between weather data



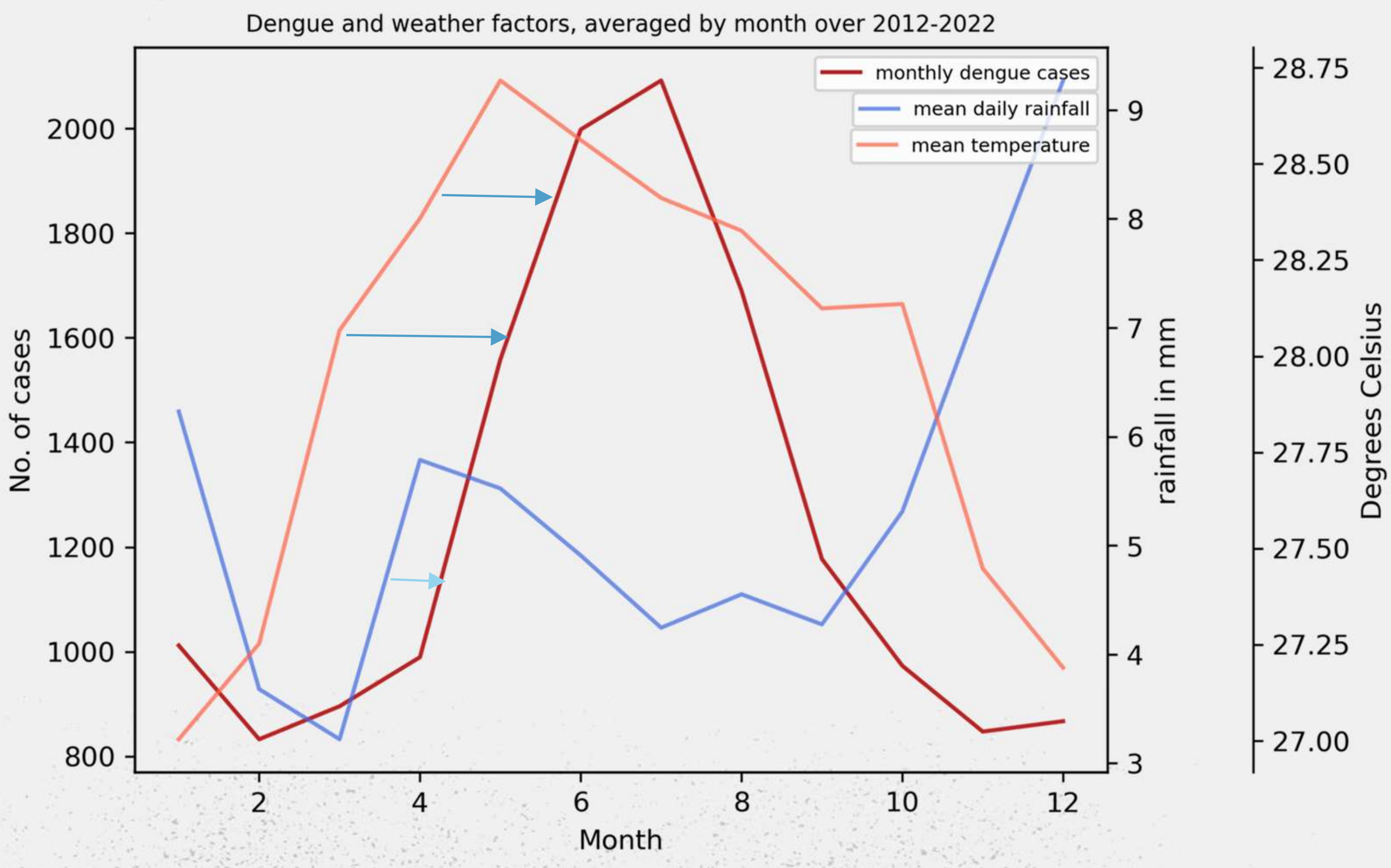
Total dengue cases vs rainfall and temperature

Time lag

1) Aedes life cycle (<6 days)

1) Extrinsic incubation period (7-14 days)

1) Time taken for a bitten individual to show symptoms (5-7 days)



Temperature and its effects on the Aedes mosquito

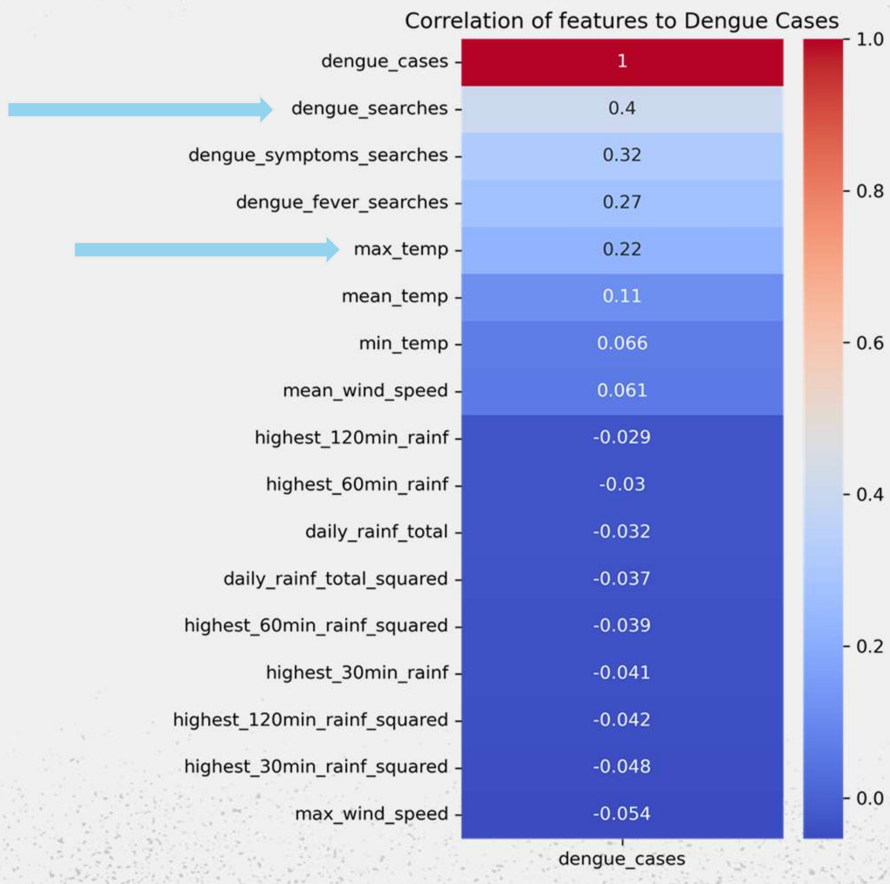
Survival, life cycle, feeding characteristics

15 to 30 deg: Lower mortality rates

32 degrees:

- Pupae development reduces to <1 day (from 4 days at 22 deg)
- Feeding frequency increases 2-fold (compared to at 24 deg)
- Extrinsic incubation period shortens to 7 days (from 12 days at 30 degrees)

Correlation of features to Dengue Cases (3mths)



04

MODELS, MODEL EVALUATION

MODELLING

Time Series Modelling

12 month predictions

- **Baseline: SARIMA**
- **SARIMAX**

3 month predictions

- **SARIMAX**

Regression Modelling

3 month predictions

- **Linear Reg**
- **Decision Tree**
- **Random Forest**
- **Support Vector Machines**

MODELLING

Time Series Modelling

12 month predictions

- **Baseline: SARIMA**
- **SARIMAX**

3 month predictions

- **SARIMAX**

Regression Modelling

3 month predictions

- **Linear Reg**
- **Decision Tree**
- **Random Forest**
- **Support Vector Machines**

STATIONARITY - “d”

```
# Check stationarity:
ad_fuller_result = adfuller(y_train['dengue_cases'])
print(f"p-value: {str(ad_fuller_result[1])}")
```

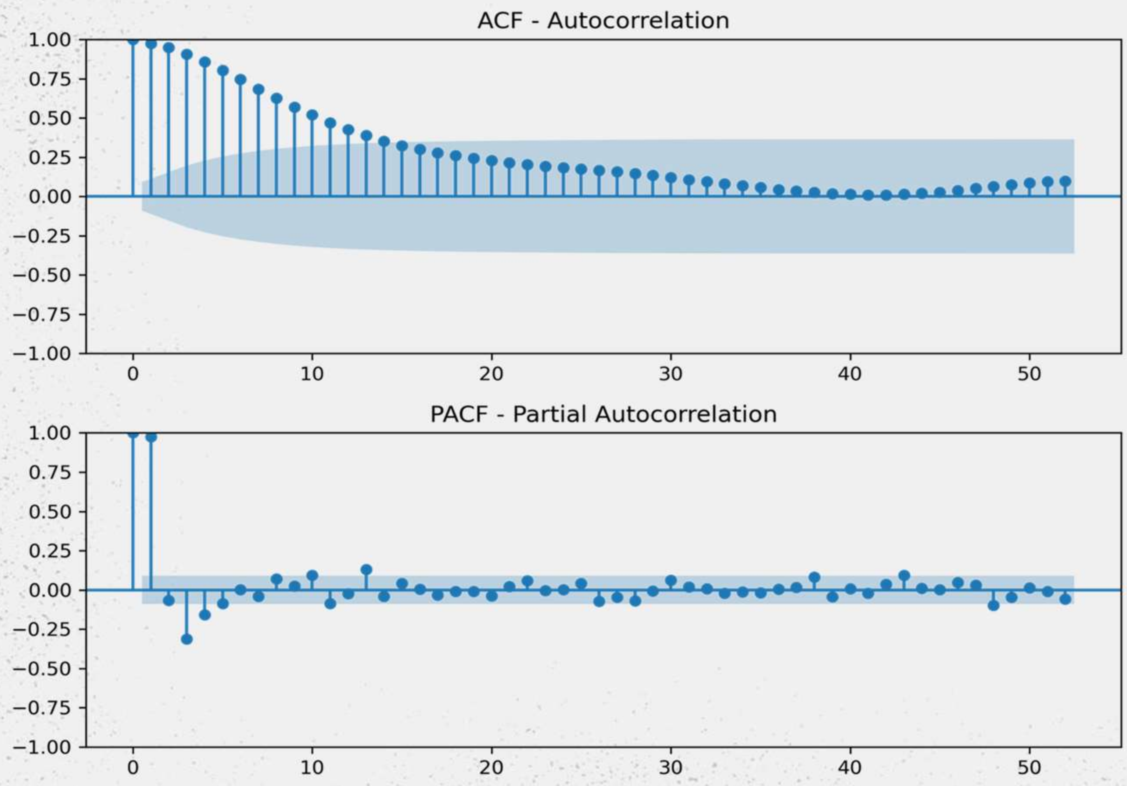
p-value: 0.026

- From the ADF test, p-value < 0.05

- The null hypothesis can be rejected
- Data can be deemed as stationary

“d” = 0

Autoregression, Moving Average - “p”, “q”



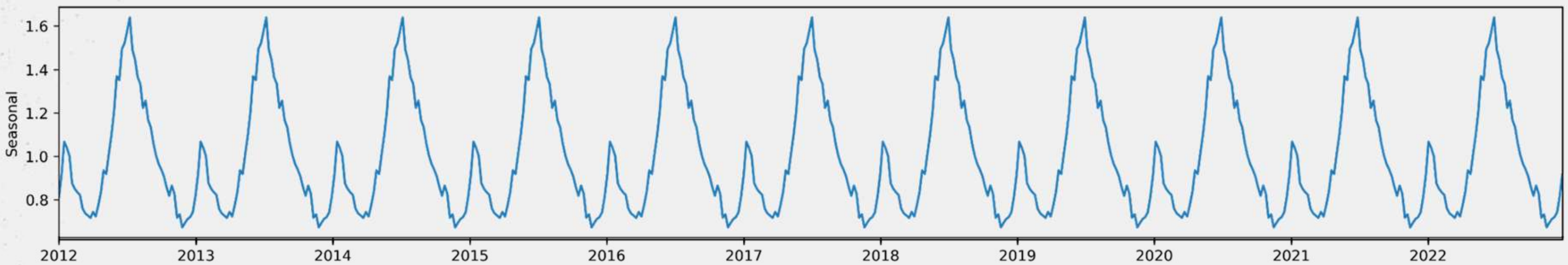
`“p”: start_p=2, max_p=3`

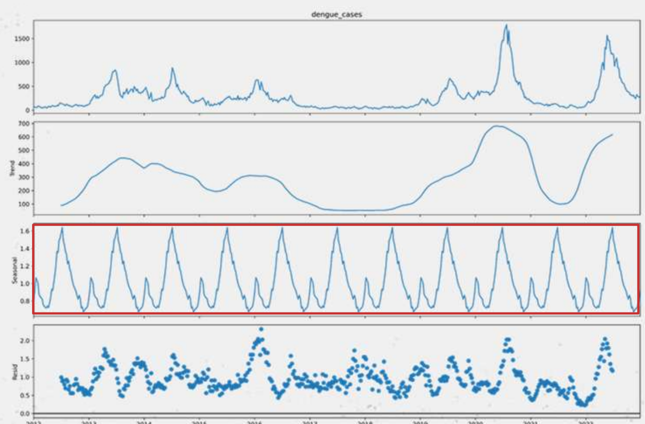
- geometric decay suggesting an AR model
- PACF shows **lag 2** dropping to fall within significance threshold

`“q”: start_q=1, max_q=10`

- ACF shows initial lags as spikes that exceed significance threshold
- Test a range of 1-10 for order of ‘q’

SEASONALITY - “m”





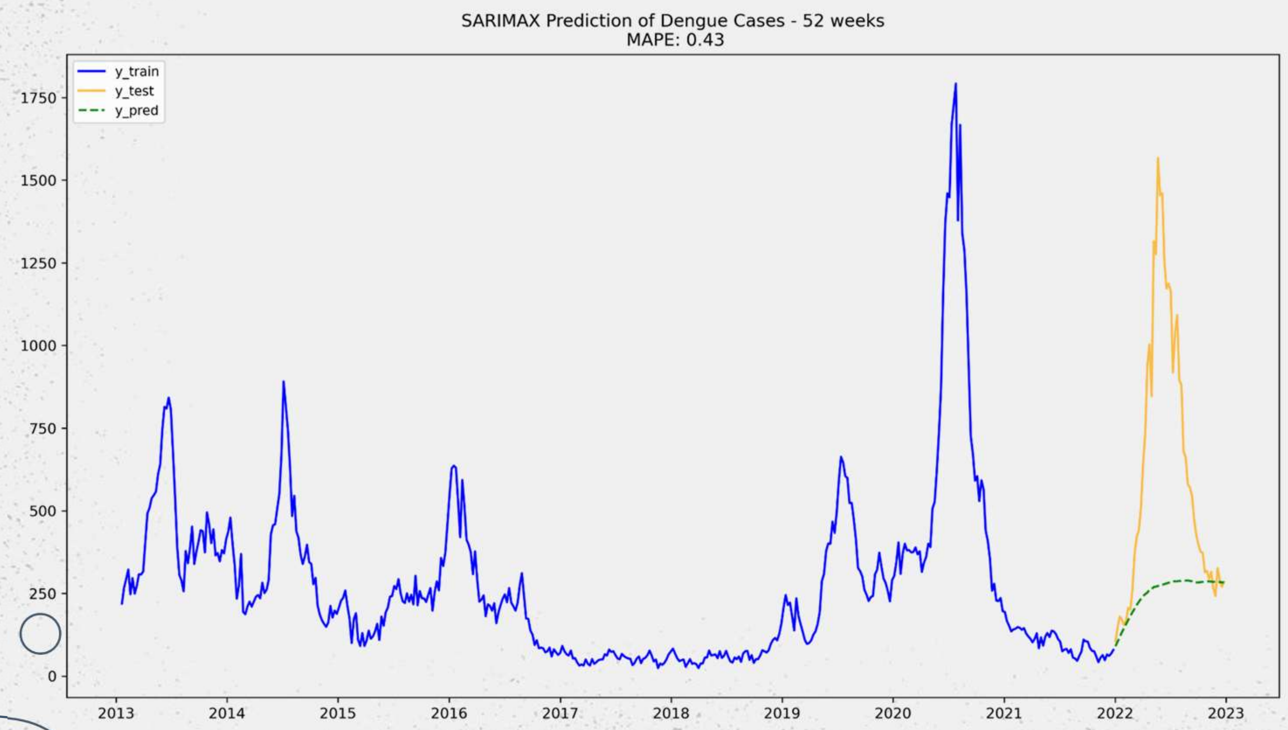
- Presence of seasonality

- Period of each seasonal cycle is 12 months / 52 weeks (peak-to-peak, trough-to-trough)

“m” = 52

BASELINE SARIMA, SARIMAX - 52WKS

52-week prediction for 2022



Best model:
(3,0,1)(1,0,0)[52]

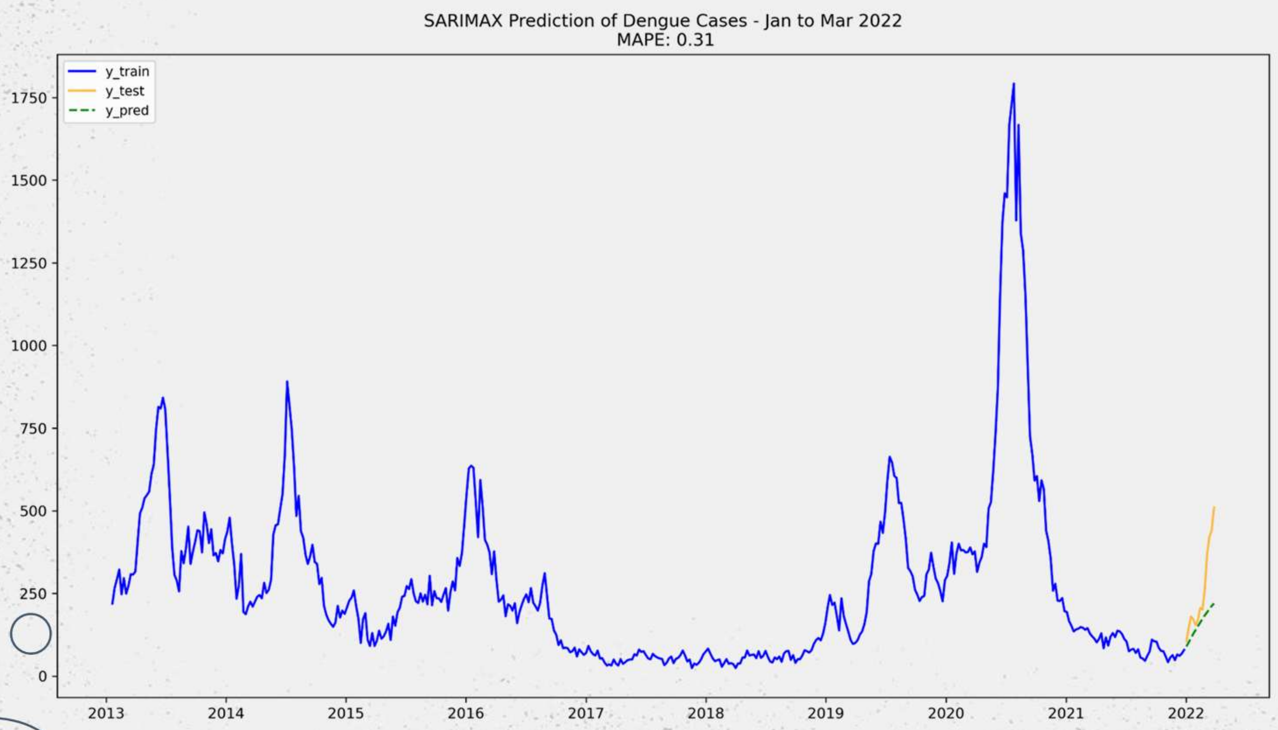
- Model predicts initial spike, before graduating to the mean no of cases

- Forecasts are 57% accurate

	MAPE	RMSE
Train	1.47	**268**
Test	0.43	**533**

SARIMAX - 12WKS

12-week prediction - Jan to Mar 2022



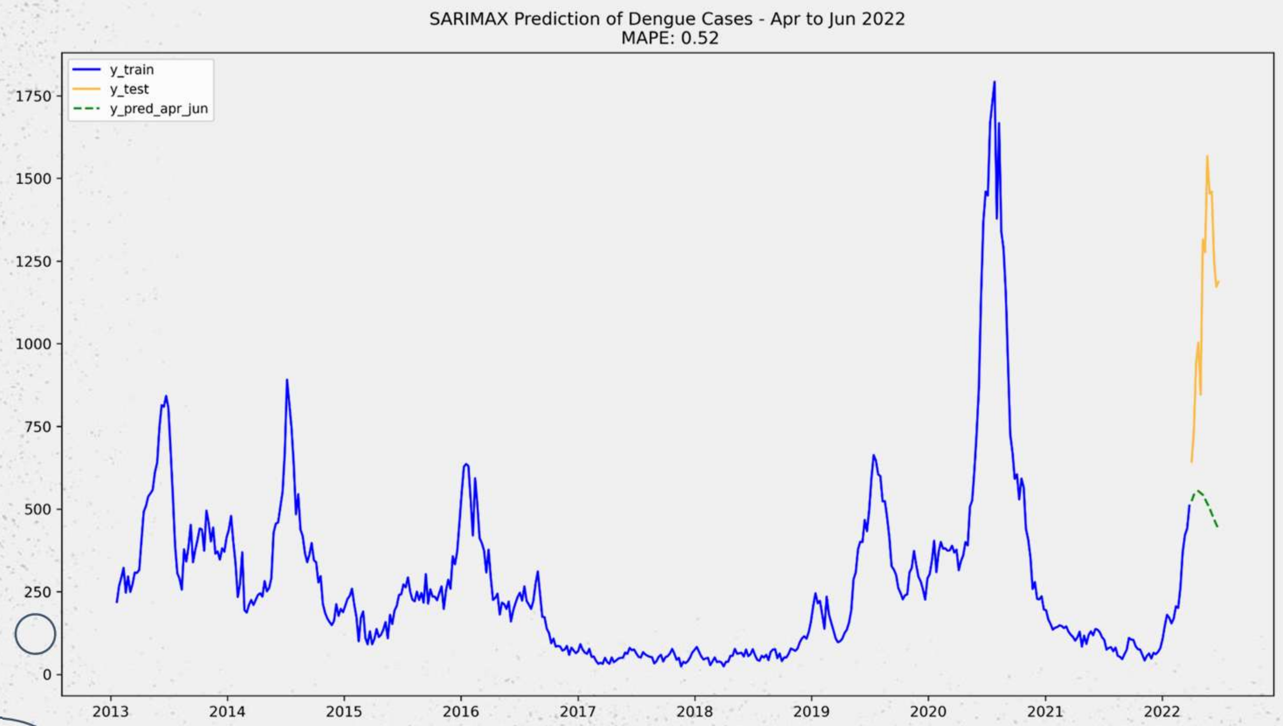
- Model able to predict early upward trend of the spike in the first quarter of 2022

- Forecasts are 69% accurate for Q1

	MAPE	RMSE
Train	1.39	**269**
Test	0.31	**133**

SARIMAX - 12WKS

12-week prediction - Apr to Jun 2022



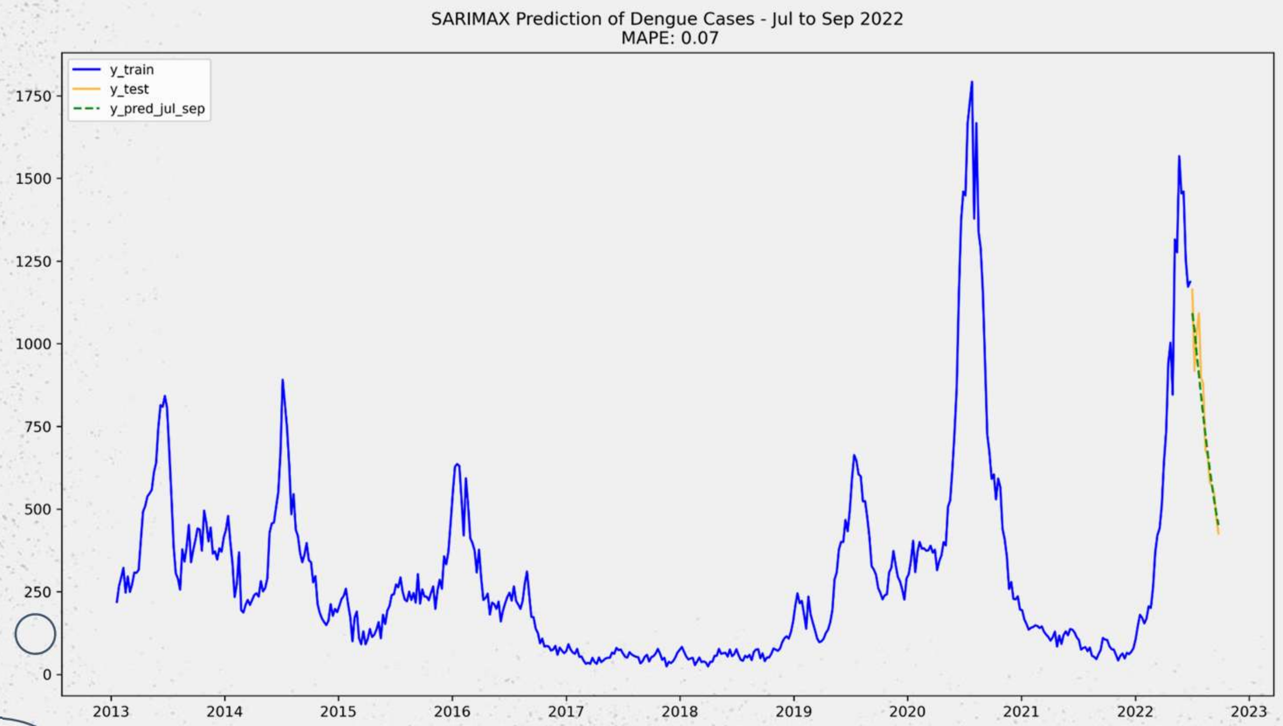
- Model able to predict a slight peak, but is ineffective in predicting the actual spike.

- Forecasts are 48% accurate for Q2

	MAPE	RMSE
Train	1.51	**265**
Test	0.52	**696**

SARIMAX - 12WKS

12-week prediction - Jul to Sep 2022



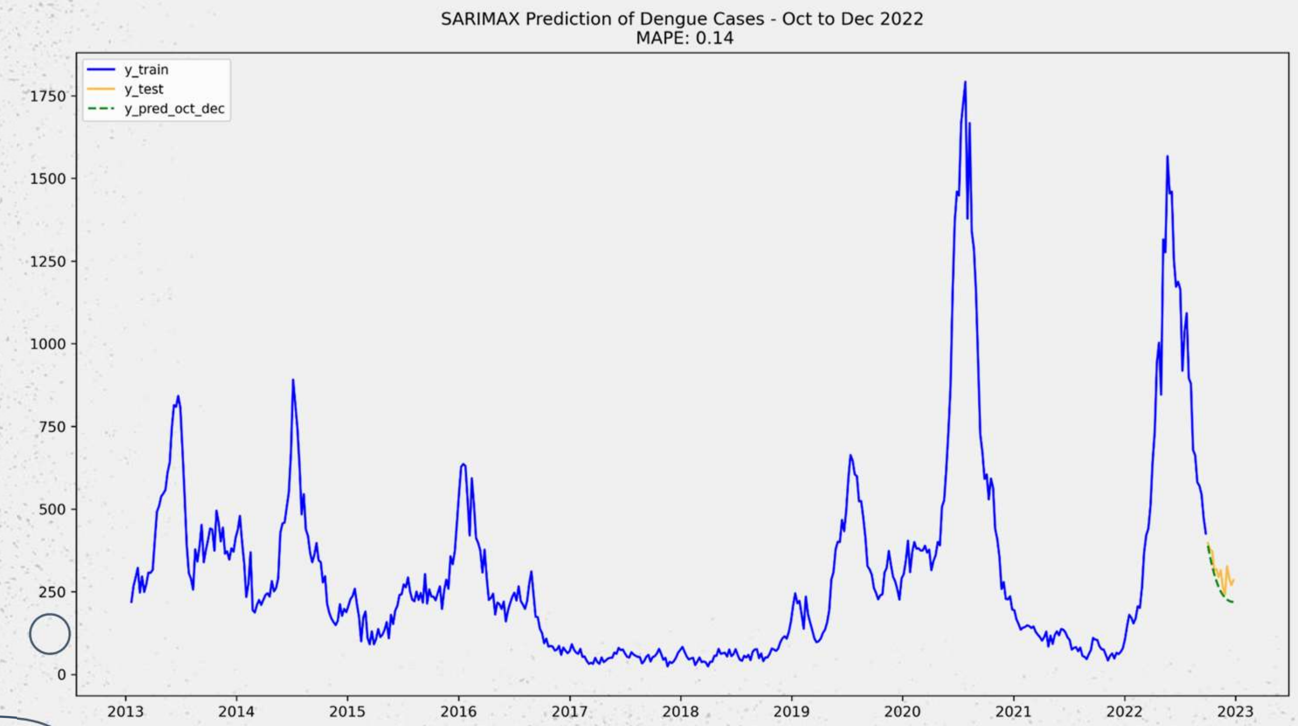
- Model is able to better capture downward trends with higher accuracy.

- Forecasts are 93% accurate for Q3

	MAPE	RMSE
Train	1.59	**311**
Test	0.07	**76**

SARIMAX - 12WKS

12-week prediction - Oct to Dec 2022



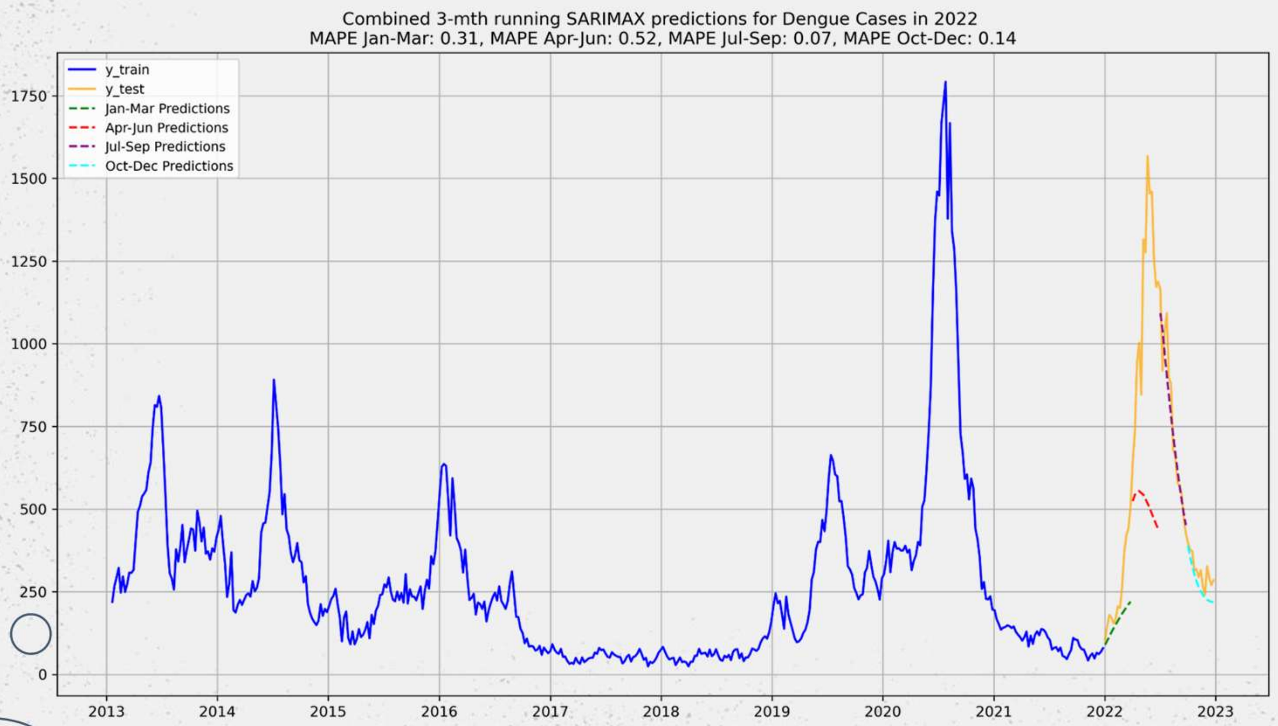
- Model able to predict the downward trend, although it may not be capturing the precise undulations within the trend.

- Forecasts are 86% accurate in Q4

	MAPE	RMSE
Train	1.5	**300**
Test	0.14	**49**

SARIMAX - COMBINED PLOT

Running 12-week prediction - Full 2022



- Stronger predictive capability for the shorter term 3 month intervals

- Achieves strong accuracy in predicting downwards trends

- More complex data inputs are required

AVG	MAPE	RMSE
Train	1.50	**286**
Test	0.26	**137**

MODELLING

Time Series Modelling

12 month predictions

- Baseline: SARIMA
- SARIMAX

3 month predictions

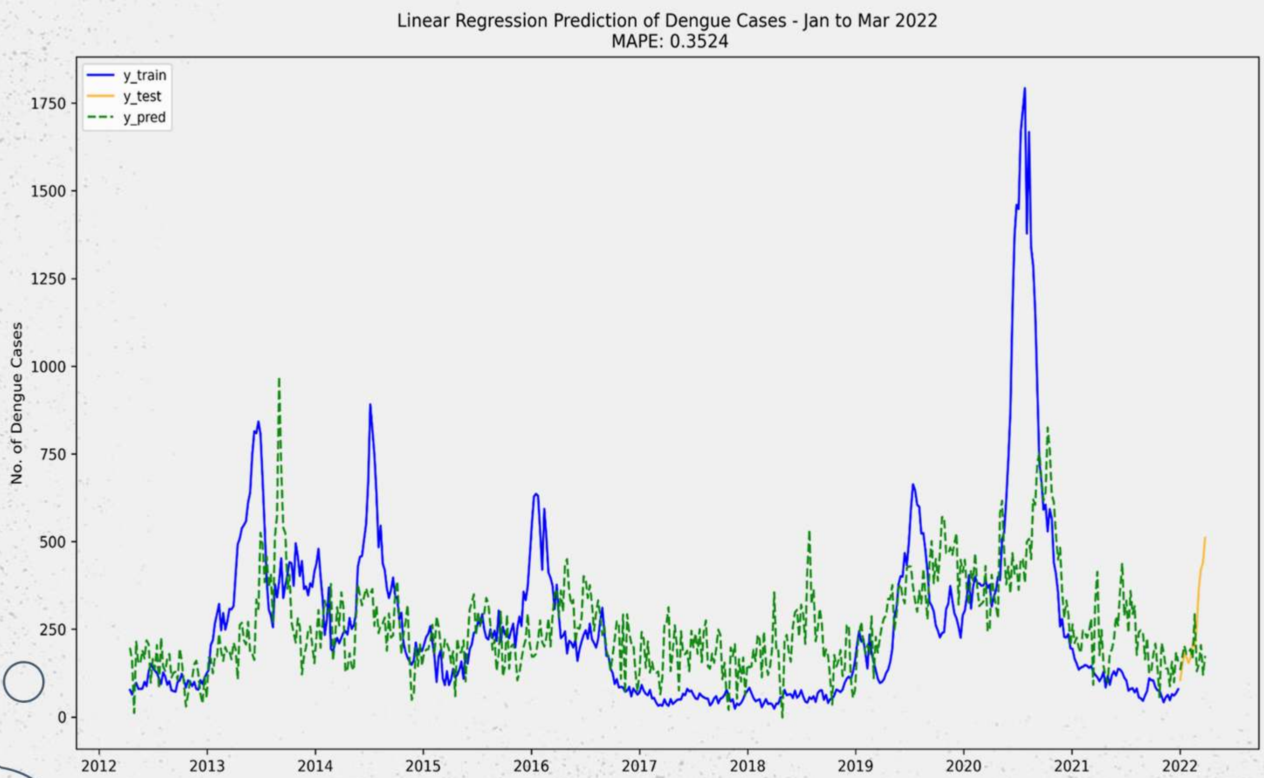
- SARIMAX

Regression Modelling

3 month predictions

- Linear Reg
- Decision Tree
- Random Forest
- Support Vector Machines

LINEAR REGRESSION



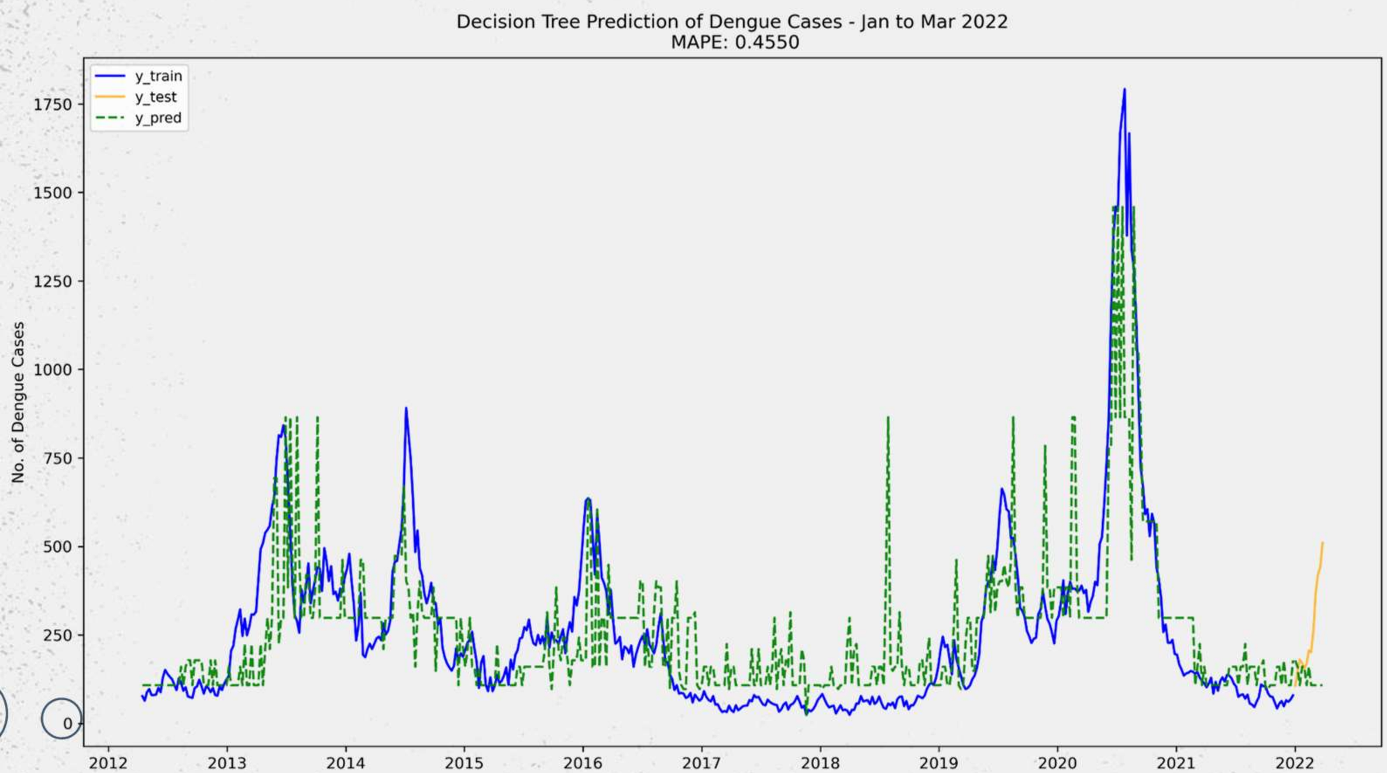
-Model predicted some historical uptrends correctly.

-Model managed to predict the initial uptrend in 2022.

-Forecast accuracy 65% in Q1 2022

	MAPE	RMSE
Train	1.03	**225**
Test	0.35	**162**

DECISION TREE



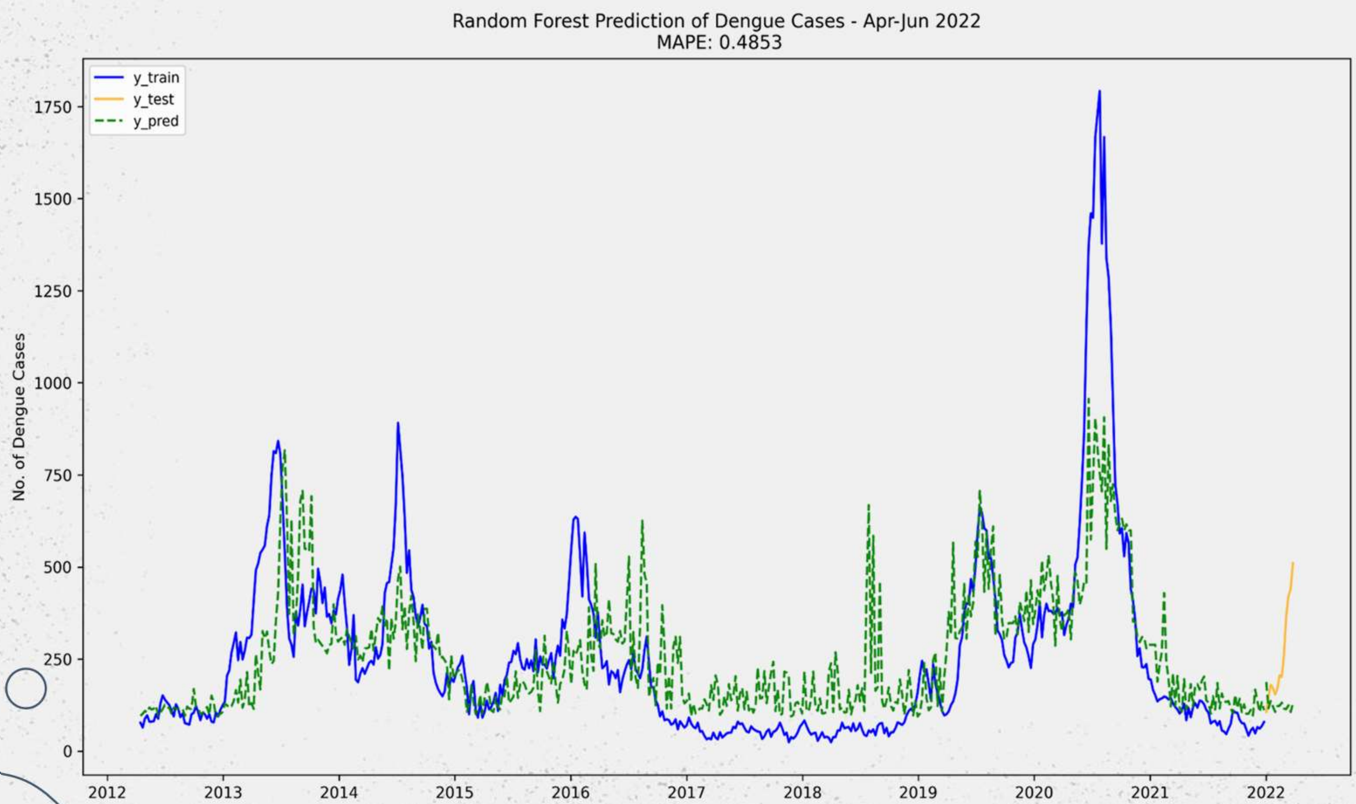
-Model predict some of historical trends well, especially the spike in 2020.

-But it fail to predict the uptrend in Q1 2022.

-Forecast accuracy 54% in Q1 2022

	MAPE	RMSE
Train	0.74	**166**
Test	0.46	**194**

RANDOM FOREST



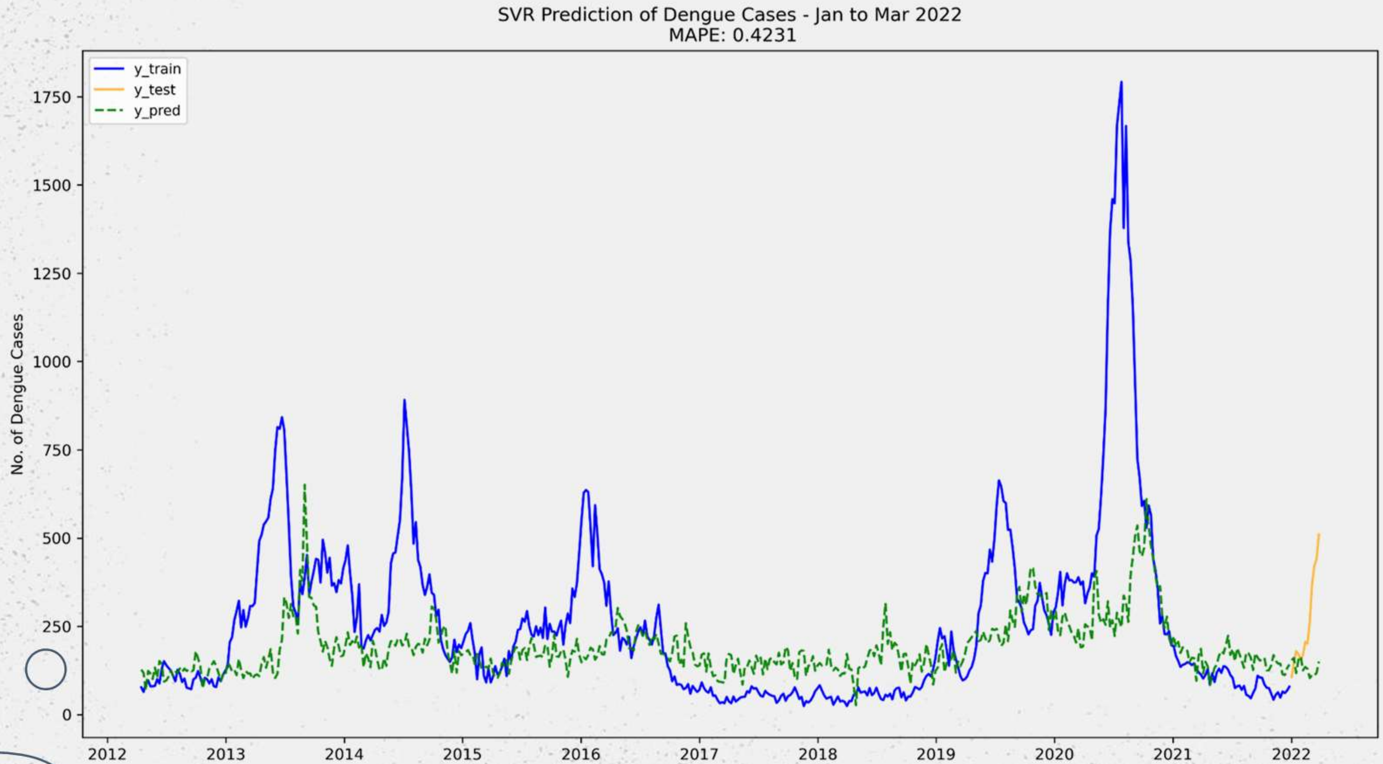
-Model didn't seem to predict trend well.

-Feature importance: dengue searches, max temperature, max wind speed

-Forecast accuracy 51% in Q1 2022

	MAPE	RMSE
Train	0.78	**178**
Test	0.49	**188**

SUPPORT VECTOR REGRESSION

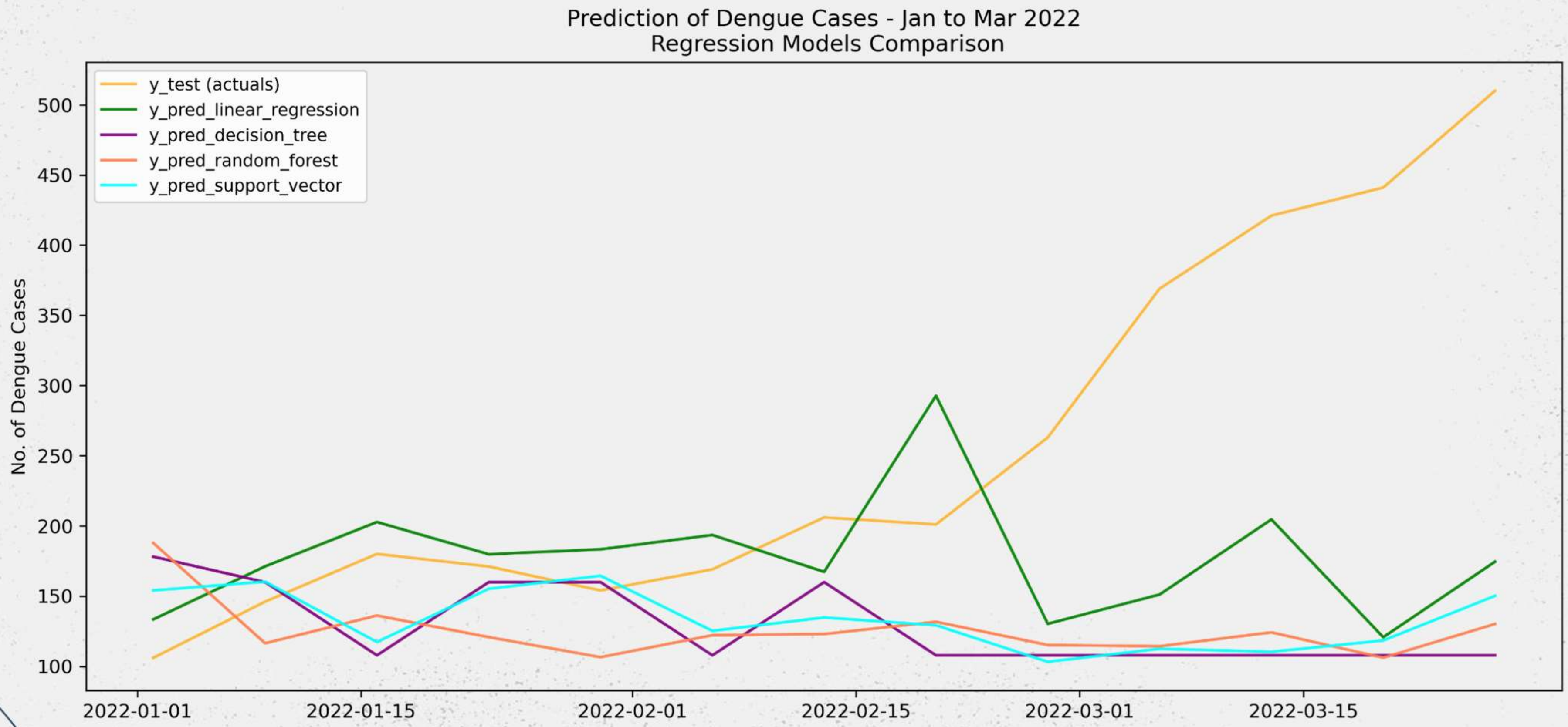


-Model didn't seem to predict trend well.

-Forecast accuracy 58% in Q1 2022

	MAPE	RMSE
Train	0.73	**243**
Test	0.42	**184**

MODEL COMPARISON



MODEL EVALUATION SUMMARY

Models	RMSE	MAPE
SARIMA	Train: 268 Test: 533	Train: 1.47 Test: 0.43
SARIMAX	Train: 269 Test: 133	Train: 1.39 Test: 0.31
Linear Regression	Train: 225 Test: 162	Train: 1.03 Test: 0.35
Decision Tree	Train: 166 Test: 194	Train: 0.74 Test: 0.46
Random Forest	Train: 178 Test: 188	Train: 0.78 Test: 0.49
Support Vector Regression	Train: 243 Test: 184	Train: 0.73 Test: 0.42

05

COST BENEFIT ANALYSIS

Cost of Dengue

Source: Stacy Soh et al

Lost time (DALYs)

Disability weights for different degrees of dengue (DF, DHF, hospitalisation vs outpatient).

Adjustment for age and local median life-expectancy

Economic costs

Costs of outpatient care and hospitalisation

Productivity loss (friction cost method)

Additional caregiving costs for children and elderly

Wolbachia costs

USD 22.7m (in 2010$) steady-state annual cost

Reno and eqmt cost for mosquito pdn facilities

Operating costs and manpower costs

Cost of community engagement initiatives

Wolbachia efficacy

Source: NEA

Rolling release (Tampines, YS)

Up to 88% reduction in dengue cases

70% less dengue compared to similar areas without Wolbachia in the 2022 outbreak

Targeted release (CCK, BB)

Up to 53% reduction in dengue cases

Use lower limit of 50% for CBA

Cost per Dengue Incidence

Year	Case prevention (40% efficacy)	Incidence prevention (40% efficacy)	Costs averted (40% efficacy)	Est'd DALYs (100%)	Total cases (100%)	Total incidences (100%)	Total economic costs (100%)	Expansion factor	Economic cost per incidence	DALYs per incidence	Lost-time in Days	Economic loss per lost day
2010	1712	4362	7.760	250	4280.0	10905.0	19.4000	2.547897	1779.000459	0.022925	8.367721	212.602740
2011	1708	4529	9.060	252	4270.0	11322.5	22.6500	2.651639	2000.441599	0.022257	8.123648	246.249185
2012	1508	4387	9.400	242	3770.0	10967.5	23.5000	2.909151	2142.694324	0.022065	8.053795	266.047775
2013	7203	21844	51.677	1282	18007.5	54610.0	129.1925	3.032625	2365.729720	0.023476	8.568577	276.093646
2014	5753	19793	48.550	1139	14382.5	49482.5	121.3750	3.440466	2452.887384	0.023018	8.401657	291.952806
2015	3618	11910	27.900	667	9045.0	29775.0	69.7500	3.291874	2342.569270	0.022401	8.176490	286.500585
2016	4197	15413	36.400	842	10492.5	38532.5	91.0000	3.672385	2361.642769	0.021852	7.975865	296.098656
2017	889	3466	7.470	168	2222.5	8665.0	18.6750	3.898763	2155.222158	0.019388	7.076746	304.549902
2018	1041	3441	7.680	173	2602.5	8602.5	19.2000	3.305476	2231.909329	0.020110	7.340308	304.062079
2019	5130	17392	38.910	831	12825.0	43480.0	97.2750	3.390253	2237.235511	0.019112	6.975966	320.706197
2020	11282	38255	84.110	1851	28205.0	95637.5	210.2750	3.390800	2198.666841	0.019354	7.064331	311.234949

Original findings (Stacy Soh et al)

Lost time: 7 days

Econ loss:
USD 311 (2010$)

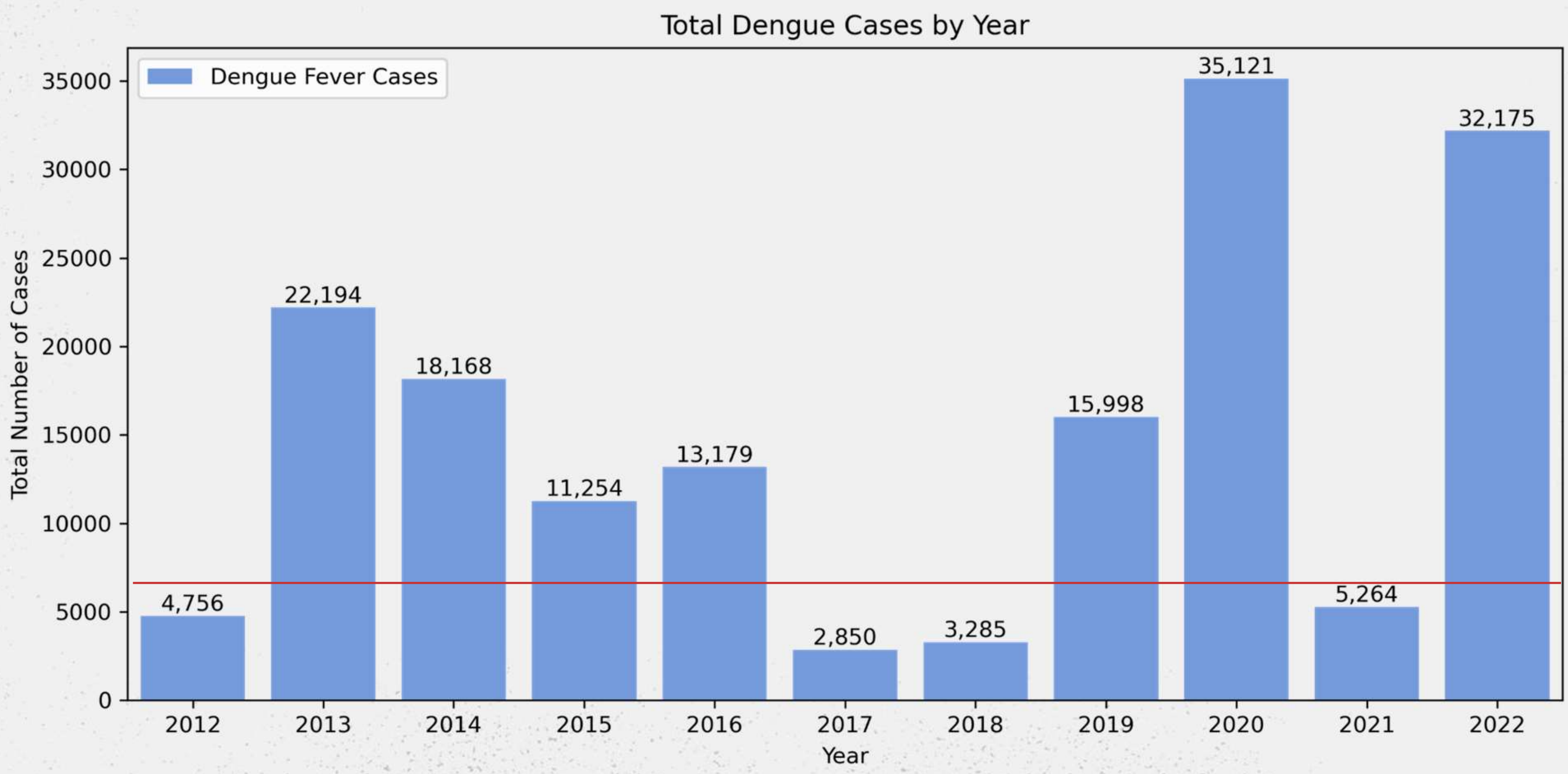
Expn factor:
3x

Assumptions

Lost time: 7 days
Econ loss per lost day: 300 (USD 2010$)
Expansion factor: 3
Wolbachia annual cost: 22.7m (USD 2010$)
Wolbachia efficacy: 0.5

Min annual dengue cases for Wolbachia to be cost effective: **7,200 cases**

Total Dengue Cases by Year



Decision Threshold

predictions	actual_if_overpred	actual_if_underpred	overpred_cost	underpred_cost
5000	3846.153846	7142.857143	0.000000e+00	0.0
5250	4038.461538	7500.000000	0.000000e+00	925000.0
5500	4230.769231	7857.142857	0.000000e+00	2050000.0
5750	4423.076923	8214.285714	0.000000e+00	3175000.0
6000	4615.384615	8571.428571	0.000000e+00	4300000.0
6250	4807.692308	8928.571429	0.000000e+00	5425000.0
6500	5000.000000	9285.714286	0.000000e+00	6550000.0
6750	5192.307692	9642.857143	0.000000e+00	7675000.0
7000	5384.615385	10000.000000	0.000000e+00	8800000.0
7250	5576.923077	10357.142857	5.132692e+06	0.0
7500	5769.230769	10714.285714	4.526923e+06	0.0
7750	5961.538462	11071.428571	3.921154e+06	0.0
8000	6153.846154	11428.571429	3.315385e+06	0.0
8250	6346.153846	11785.714286	2.709615e+06	0.0
8500	6538.461538	12142.857143	2.103846e+06	0.0
8750	6730.769231	12500.000000	1.498077e+06	0.0
9000	6923.076923	12857.142857	8.923077e+05	0.0
9250	7115.384615	13214.285714	2.865385e+05	0.0
9500	7307.692308	13571.428571	0.000000e+00	0.0
9750	7500.000000	13928.571429	0.000000e+00	0.0

MAPE: 0.3

Overprediction:

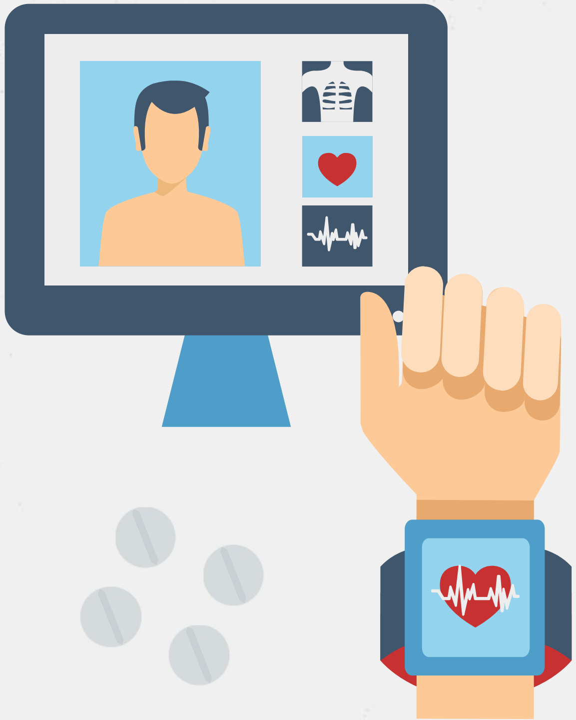
- Roll out Wolbachia when it is not cost-effective
- Net loss:
  Wolbachia cost - Econ benefit

Underprediction:

- Did not roll out Wolbachia when economic benefits would have exceeded costs
- Net loss:
  Econ benefit - Wolbachia cost

Roll out Wolbachia when model prediction > 6000 cases.

Est'd net loss: USD 5.1m (2010$)



06

CONCLUSION

CONCLUSION

1) Narrowed down an effective prediction window of 3 months, in order to strike a balance between:
   - Model accuracy
   - Enabling sufficient lead time to act adapt to an outbreak prediction
2) Amongst the models tested, SARIMAX yields the lowest RMSE and MAPE on test data hence performed the best

3) Running 3-mth predictions highlighted our models effectiveness in predicting downward trends, and weaknesses in capturing upward trends.

Further steps

1) Improve model with more features
    a) Circulating serotype
    b) Serological information
    c) Premises index

1) Improve Wolbachia CBA with more detailed cost breakdown
    a) Fixed costs vs Variable costs
    b) Choose between Rolling release / Targeted release
    c) Compare Suppression and Replacement method

THANK YOU

CREDITS: This presentation template was created by **Slidesgo,** including icons by **Flaticon** and infographics & images by **Freepik**