Searching for COVID: Using **Google Search Data** to **Predict COVID-19** Cases in **Philadelphia**

Rachel Sanderlin
May 2nd, 2023

Introduction and Goals

Stakeholder:

- City of **Philadelphia**

Goal:

- Use **Google's COVID-19 related search data** to improve predictions of **daily COVID-19 cases**

Data Sources

Google's Explore COVID-19 Symptoms Search Trends

- 01/01/2020 - 11/11/2022
- **65k+** rows and **400+** columns

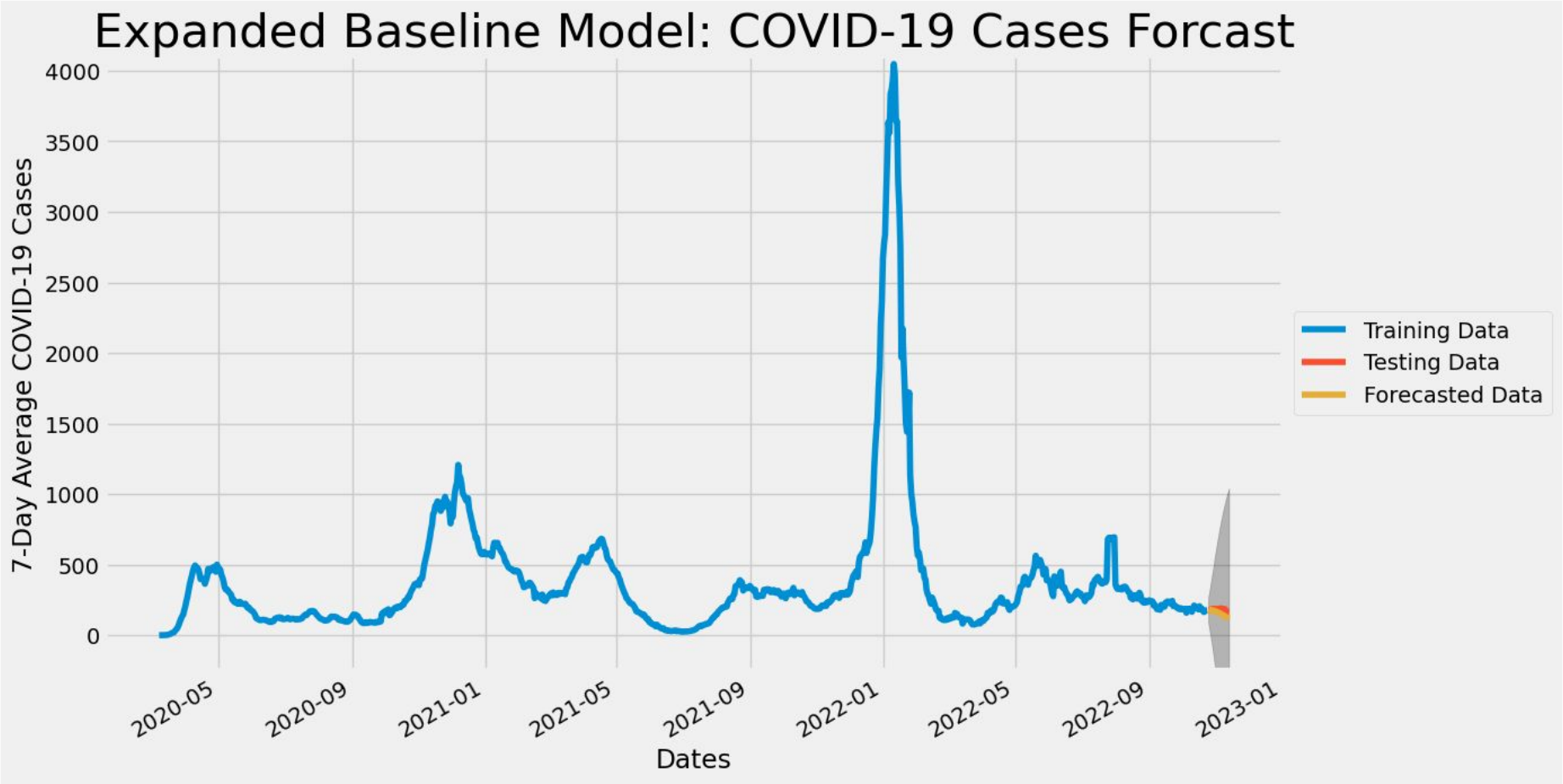
COVID-19 Data for Pennsylvania

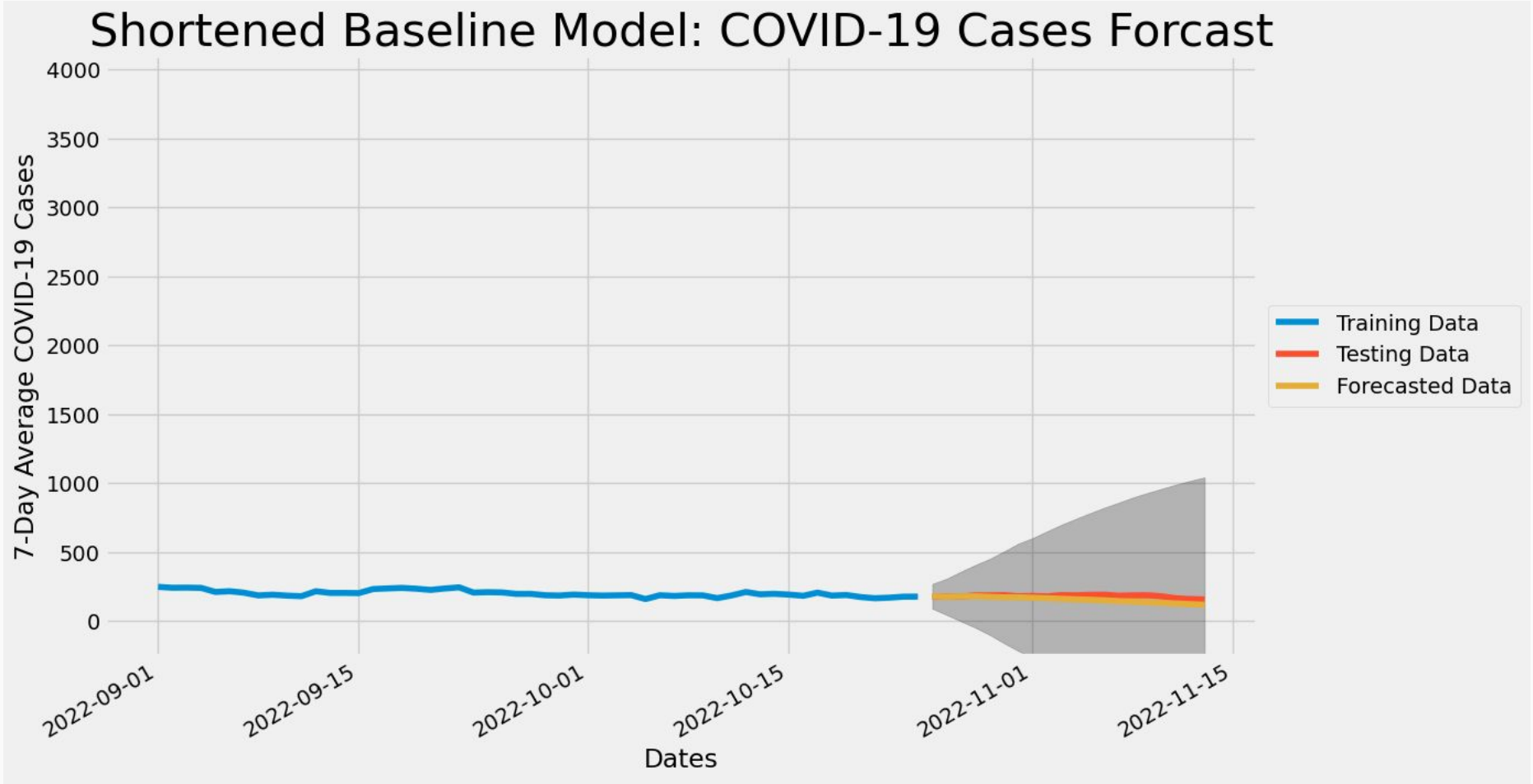
- 03/01/2020 - 03/14/2023
- **75k+** rows and **12** columns.

Final (Philly!) Data

- 03/08/2020 - 11/13/2022
- **~1k** rows and **400+** symptoms
- Target : **7-Day Average** COVID-19 Cases
- Prediction: **20 days**

Baseline Model: **SARIMA Model** of **7-Day Average COVID-19 Cases**


Expanded Baseline Model: COVID-19 Cases Forcast
7-Day Average COVID-19 Cases
4000
3500
3000
2500
2000
1500
1000
500
0
2020-05
2020-09
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
Dates
Training Data
Testing Data
Forecasted Data


Shortened Baseline Model: COVID-19 Cases Forcast
7-Day Average COVID-19 Cases
4000
3500
3000
2500
2000
1500
1000
500
0
2022-09-01
2022-09-15
2022-10-01
2022-10-15
2022-11-01
2022-11-15
Dates
Training Data
Testing Data
Forecasted Data

Modeling the Data

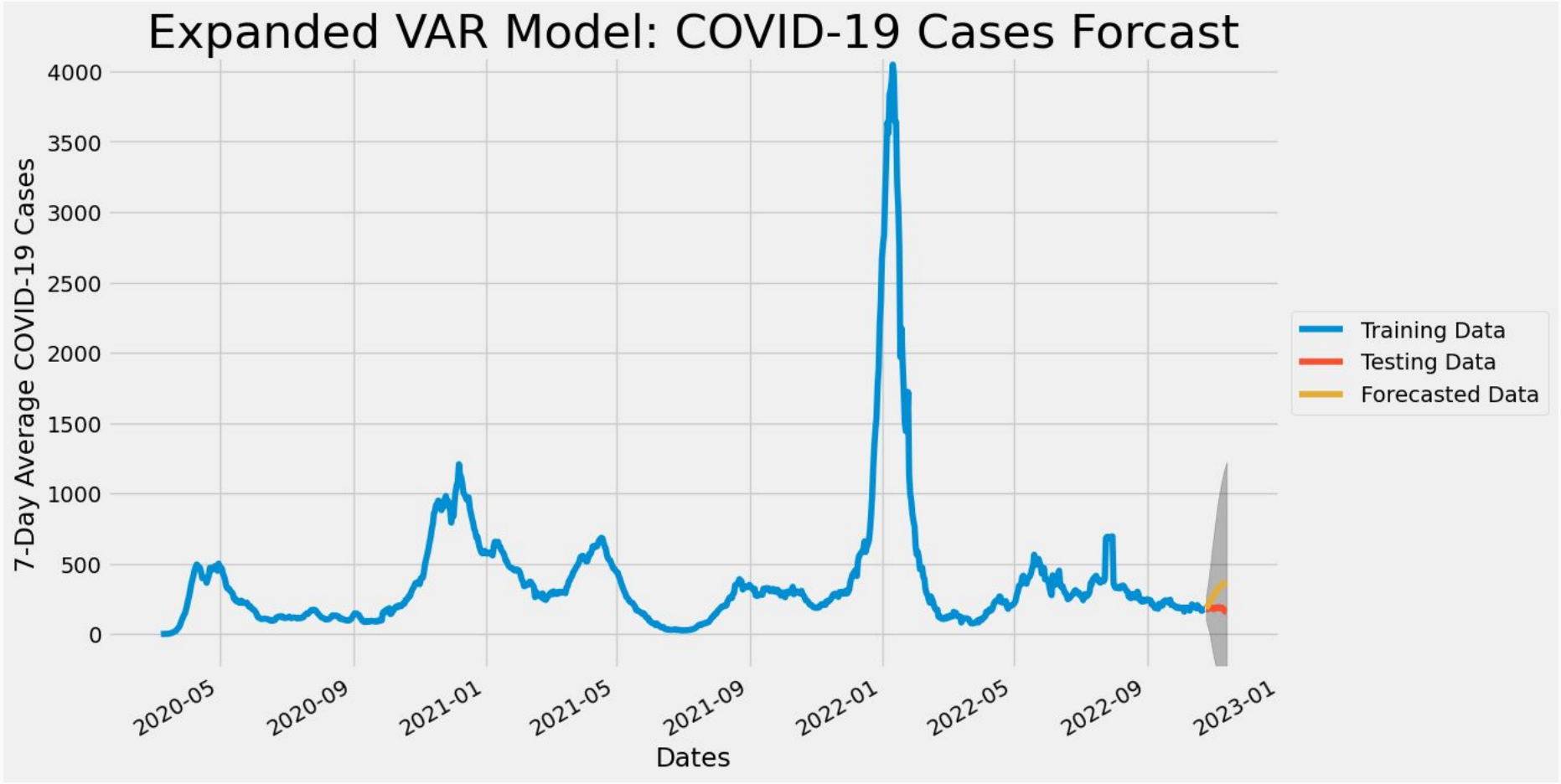
Used **PCA** (Principal Component Analysis) for dimensionality reduction

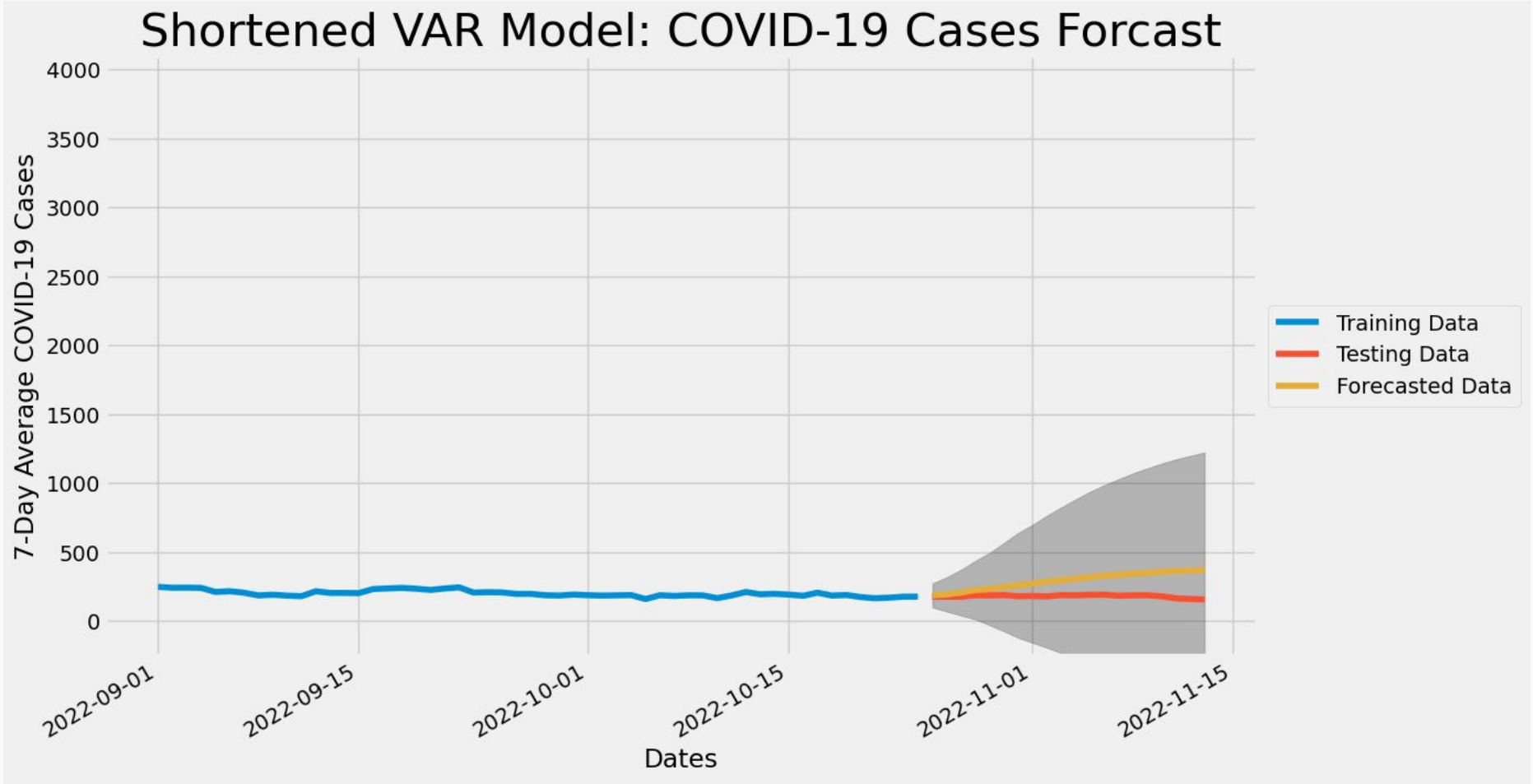
- **422** symptoms ➡ **2** PC's

~73%

Amount of **explained variance** found in the two Principal Components (PC's)

Vector Autoregressive (**VAR**) Model


Expanded VAR Model: COVID-19 Cases Forcast
7-Day Average COVID-19 Cases
4000
3500
3000
2500
2000
1500
1000
500
0
2020-05
2020-09
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
Dates
Training Data
Testing Data
Forecasted Data


Shortened VAR Model: COVID-19 Cases Forcast
7-Day Average COVID-19 Cases
4000
3500
3000
2500
2000
1500
1000
500
0
2022-09-01
2022-09-15
2022-10-01
2022-10-15
2022-11-01
2022-11-15
Dates
Training Data
Testing Data
Forecasted Data

Comparing the Models

	Baseline Model	VAR Model
MAE	24.86	111.87
RMSE	29.80	128.19
MAPE	14%	63%

Conclusion: Google search trends are **not helpful** (in our VAR model) at **predicting** COVID-19 cases

Next Steps

- Look into **other relevant** COVID-19 data

- Try **alternative models**
  - Recursive/Crossvalidation
  - VARMA, VARMAX

Thank you!

Email: sanderlin2013@gmail.com

Github: github.com/sanderlin2013

LinkedIn: linkedin.com/in/rachel-sanderlin