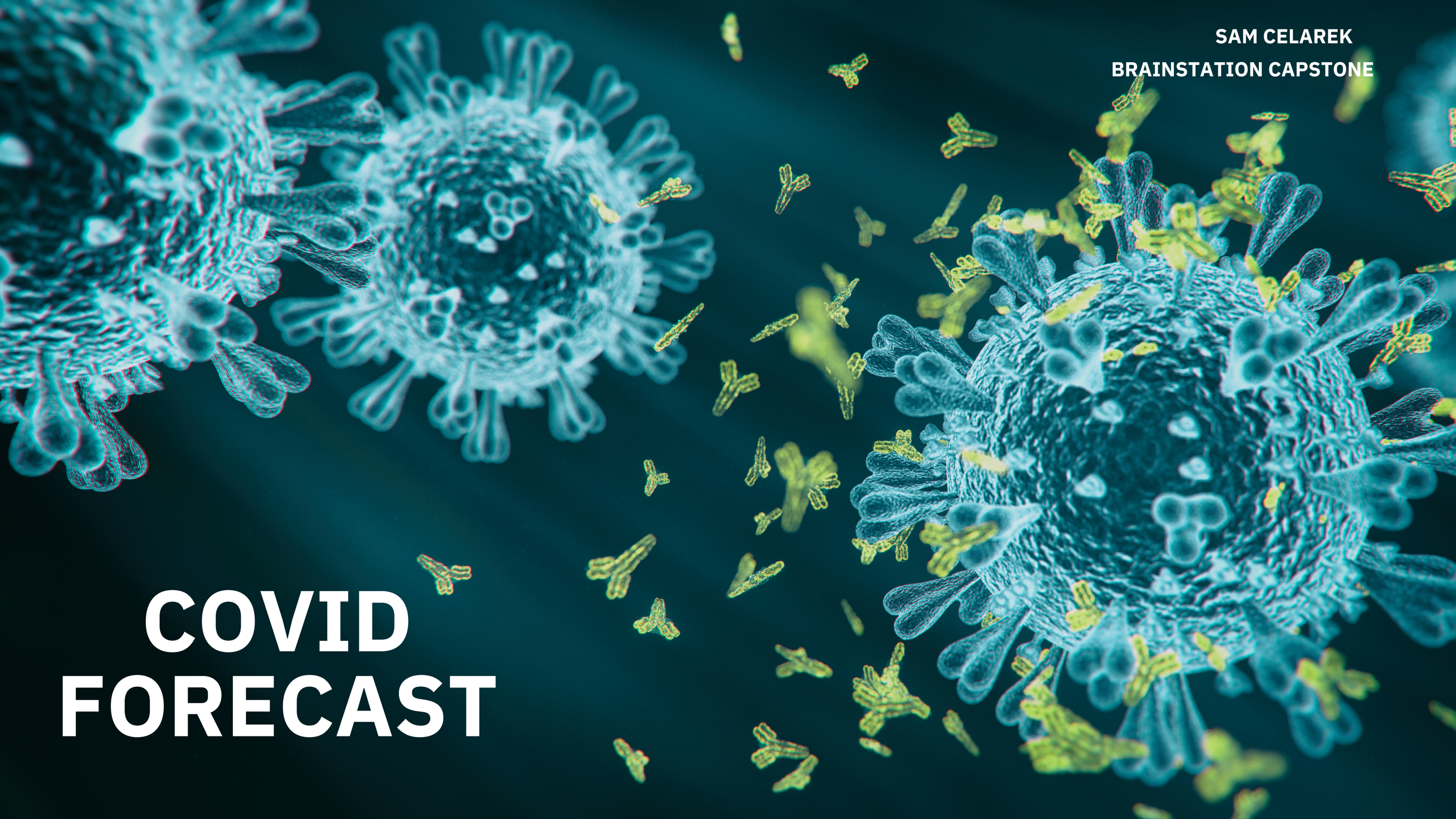


COVID FORECAST

SAM CELAREK
BRAINSTATION CAPSTONE

SAM'S COVID PREDICTION PROJECT

THE PROBLEM

3 years ago COVID shook the world's foundations, but it is only the tip of the existential iceberg that is the danger of pandemics. My project utilizes a wide array of features and epidemiological models to make a 7-14 day forecast for hospital admissions and new cases so that humanity might be better prepared for the unpreparable.

Time Line of Covid

59 Cases
1 Dead

WUHAN OUTBREAK

First death in January 2020. WHO declares global emergency

1m Cases
100k Dead

WORLDWIDE PANDEMIC

By March 2020, it has spread to every continent. 1 million cases by April.

30m Cases
1m Dead

LOCKDOWNS AND VARIANTS

By July, 1/3 of the world in lockdowns. Unemployment rose to 18%. Delta Variant emerged in October.

100m Cases
2m Dead

TESTING AND VACCINES

Remdesivir, and the first vaccines were rolled out in 2020 December.

760m Cases
7 m Dead

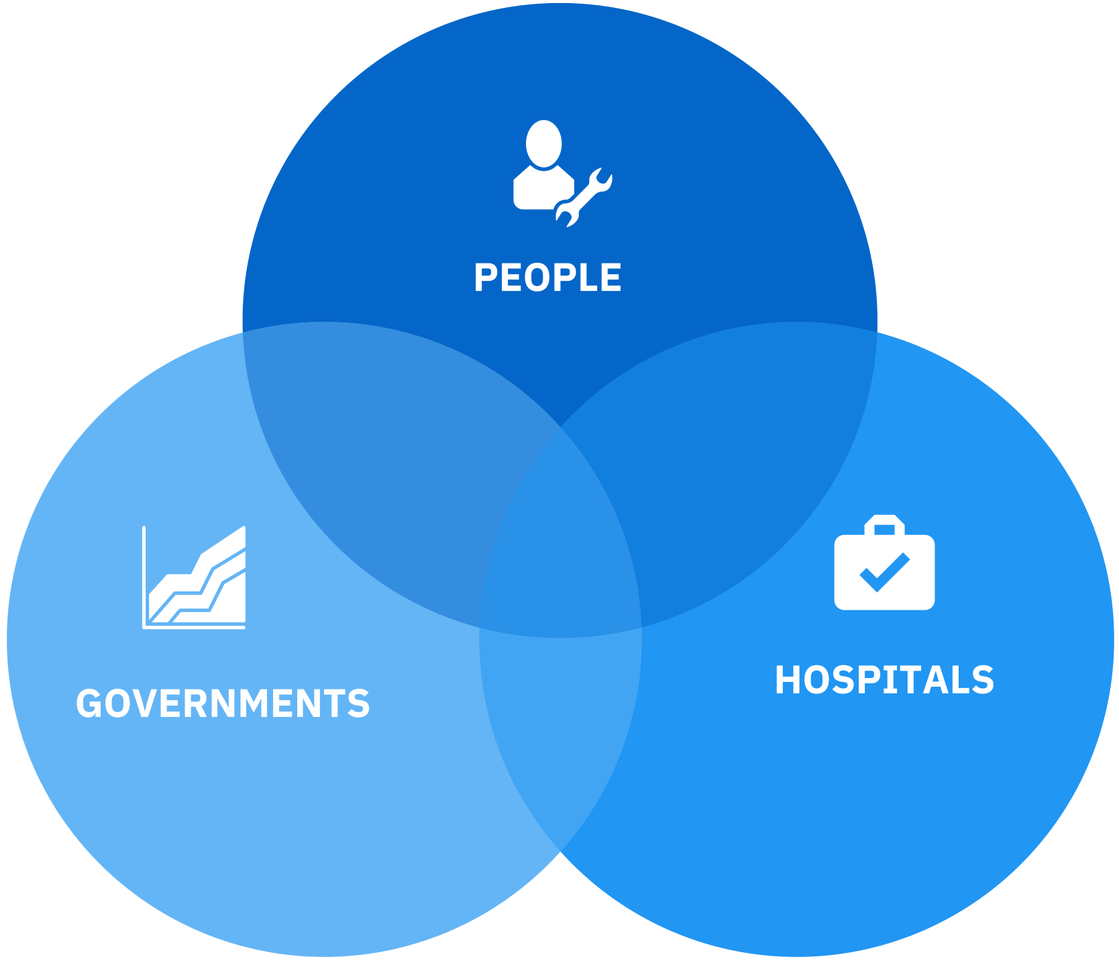
WHAT'S NEXT?

WHO COULD THIS HELP

HOSPITALS
Preparing for covid surges, increasing supplies.

GOVERNMENTS
Messaging, Regulations on Mobility, and Seasonal Adjustments.

PEOPLE
Adjusting your risk tolerance accordingly. Microcovid.

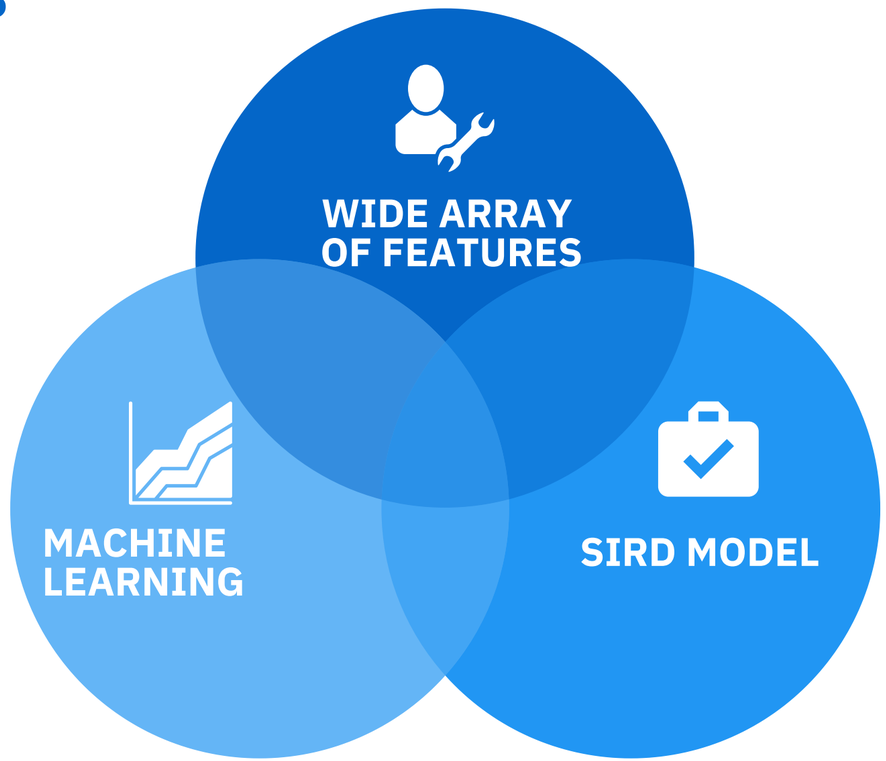


WHAT IS MY VALUE ADD

WIDE ARRAY OF FEATURES

Most projects use only autoregressive analysis.

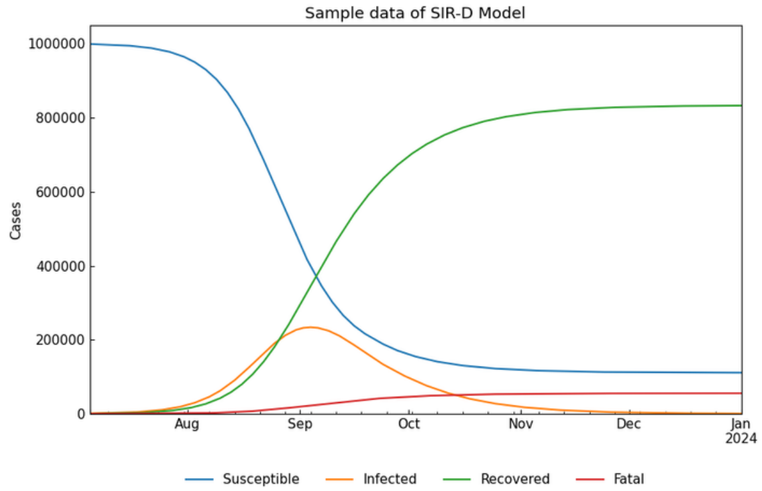
- **WEATHER**
- **MOBILITY**
- **GOVERNMENT**
- **SEVERITY**



SIRD MODEL

Epidemiologist use the SIRD model to extrapolate the spread of a disease in the population.

$$R_0 = \frac{\text{transmission rate}}{\text{recovery rate} + \text{mortality rate}}$$

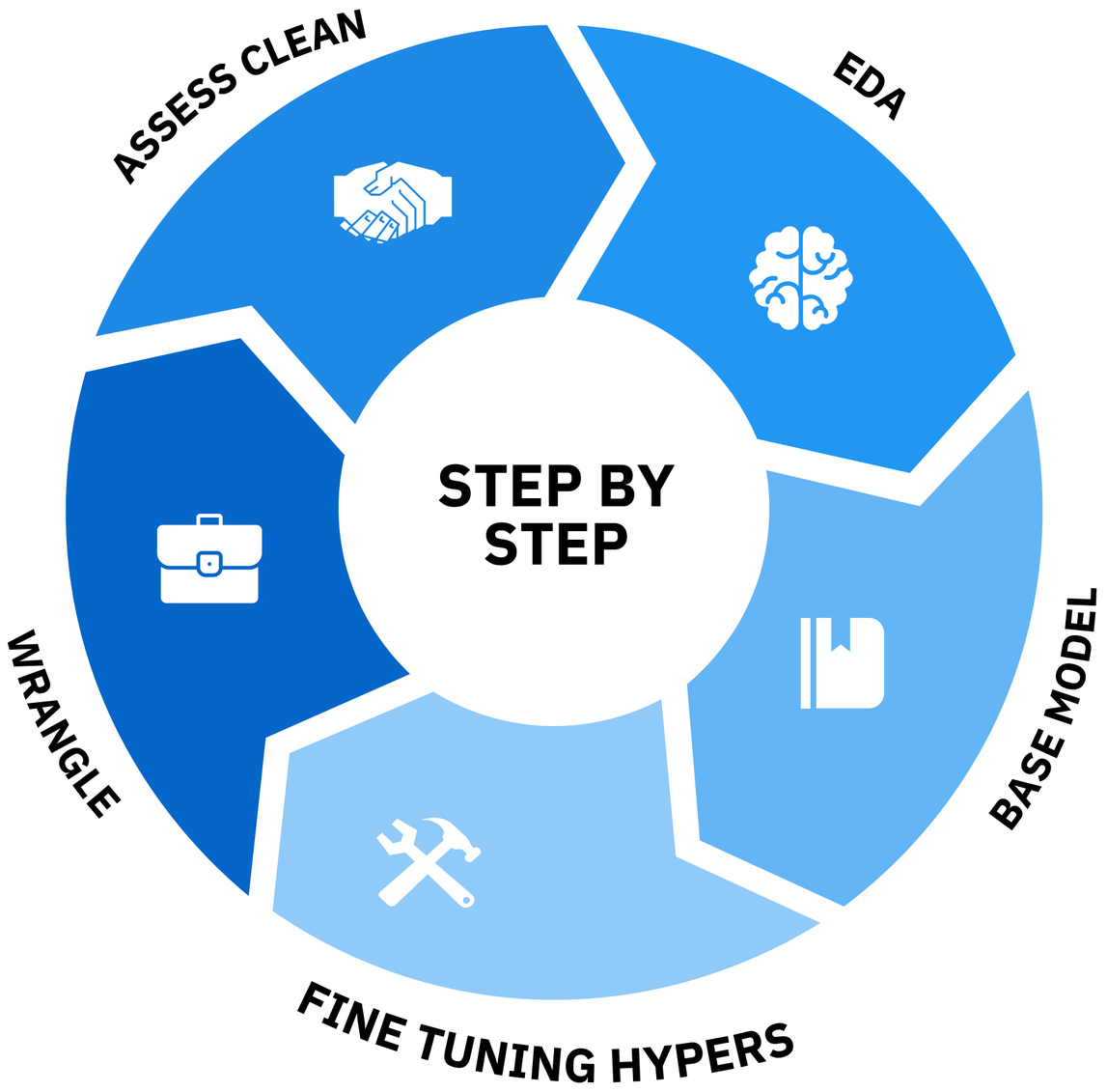


MACHINE LEARNING

Combining Epidemiological Parameters with Machine Learning Flexibility

CAPSTONE PROGRESSION

- **STEP ONE**
  Wrangle six different datasets
- **STEP TWO**
  Assess the datasets and clean. Data Interpolation.
- **STEP THREE**
  Exploratory Data Analysis. ACF and PACF.
- **STEP FOUR**
  SARIMA and SARIMAX model
- **STEP FIVE**
  Facebook Prophet Model and RNN's



DATA COLLECTION

COVSIRPHY

- Epidemiological Modeling.
- Real Time Global Covid Datasets.

GOOGLE DATASET

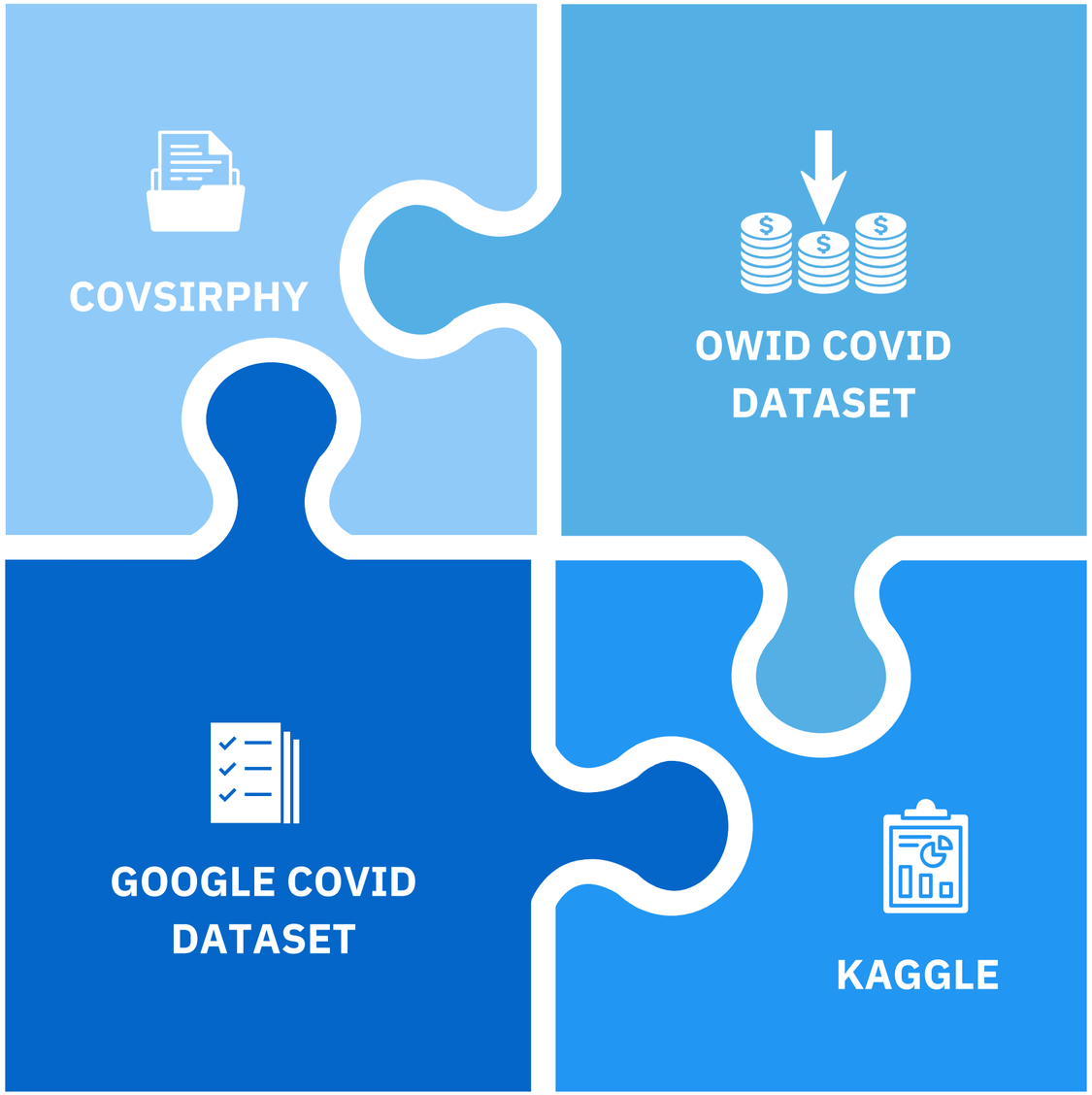
- Weather, Mobility, Regulation
- Stopped 9/15/2022

OWID COVID DATASET

- Hospitalizations, Testing, Excess Morbidity.

VSCODE WRANGLING

- Coding in VSCode, Downloaded csvs



TARGET FEATURES

NATIONAL US DATA

1142 rows x 59 columns

NEW CONFIRMED

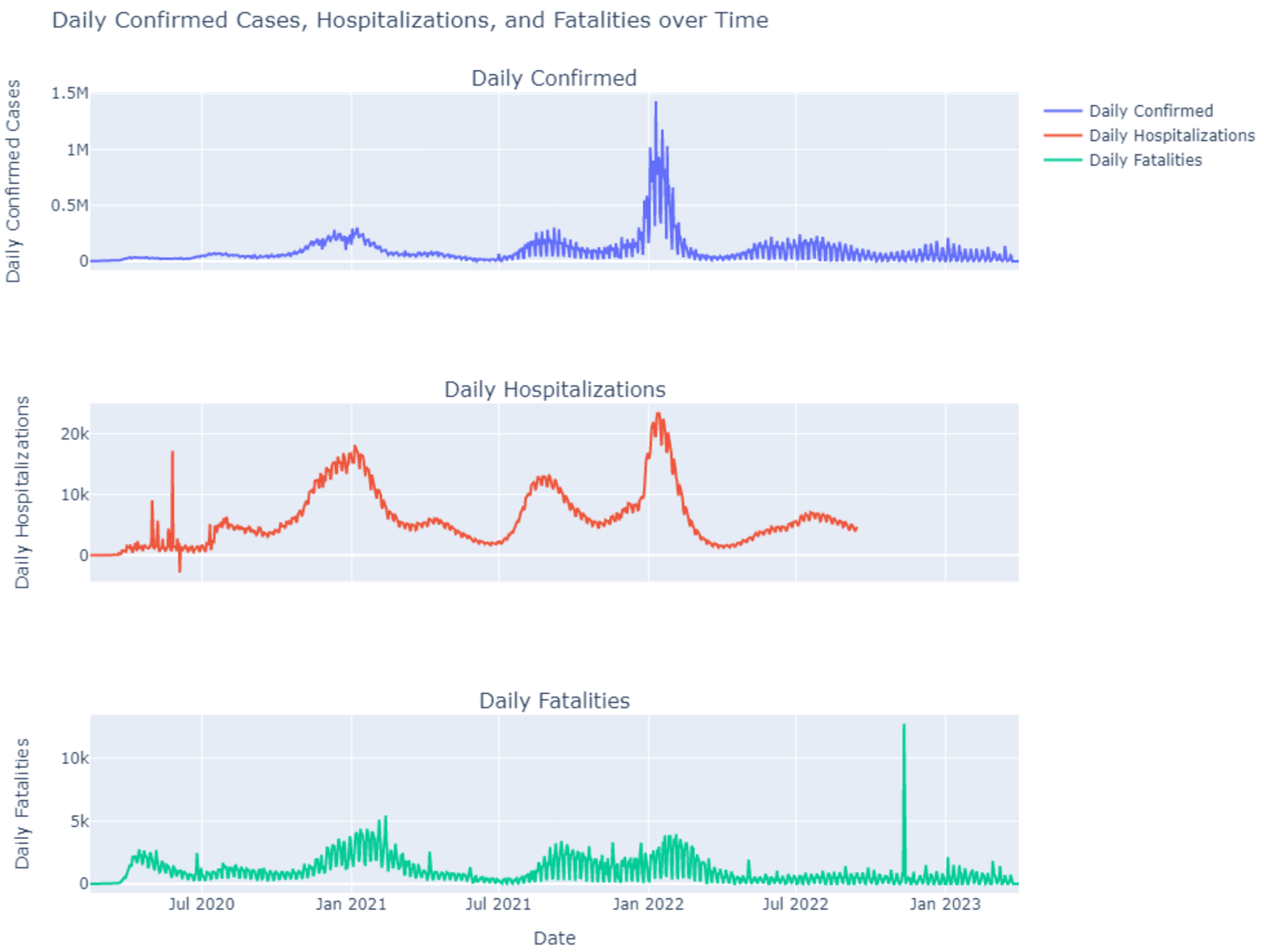
The number of new confirmed COVID-19 cases on the given date.

NEW HOSPITALIZED

The number of new hospitalizations due to COVID-19 on the given date.

NEW FATALITIES

The number of new COVID-19 related deaths on the given date.



PANDEMIC PARAMETERS

MORTALITY RATE

The rate at which infected people die from the disease.

TRANSMISSION RATE

Controls how fast infection spreads. Higher transmission rate leads to faster spread.

RECOVERY RATE

How quickly infected people recover.

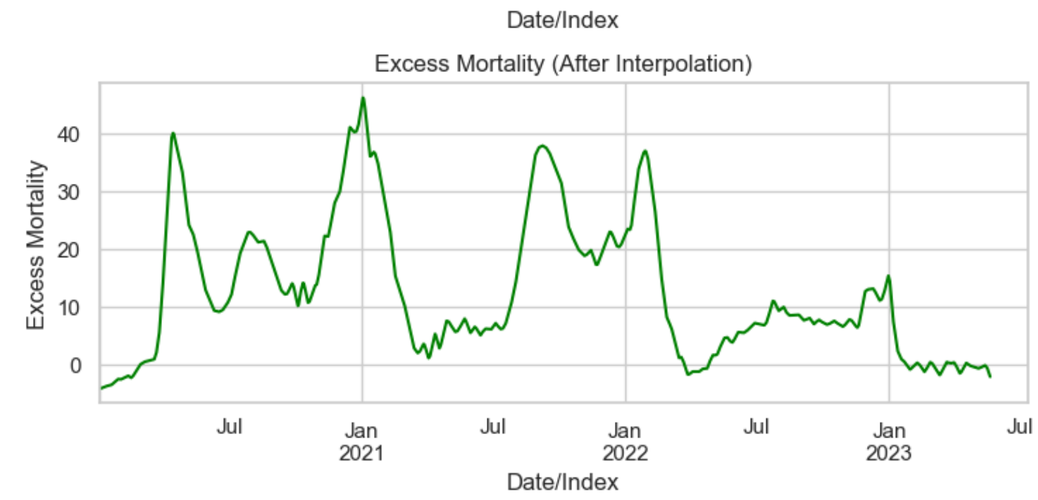
$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log_{10}(A_i+1) - log_{10}(P_i+1))^2}$$

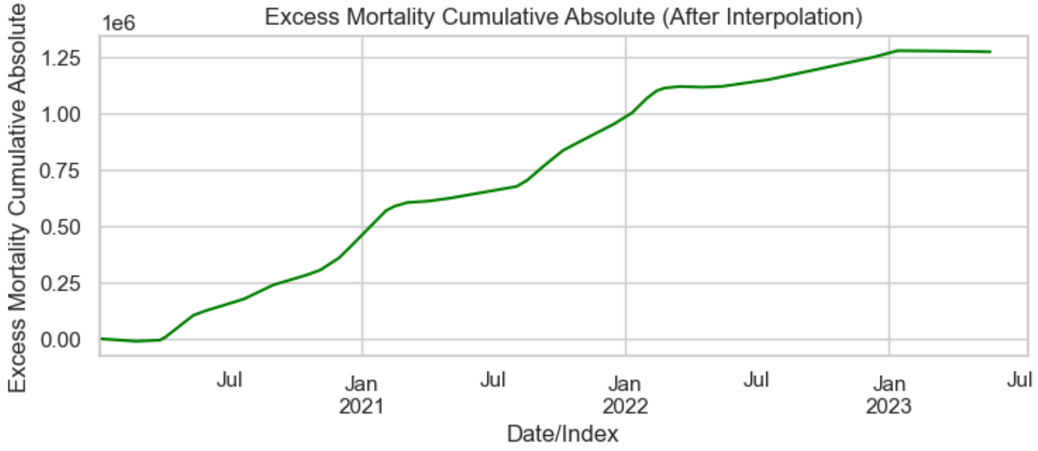
$$R_0 = \frac{\text{transmission rate}}{\text{recovery rate} + \text{mortality rate}}$$

REPRODUCTION RATE

Represents how many people an infected person spreads the virus to. Above 1 is an expanding epidemic.

OTHER FEATURES





EXCESS MORTALITY

Measures amount of people who died in excess of last year

RETAIL AND RECREATION MOBILITY

Google's mobility trends in retail and recreation spaces.

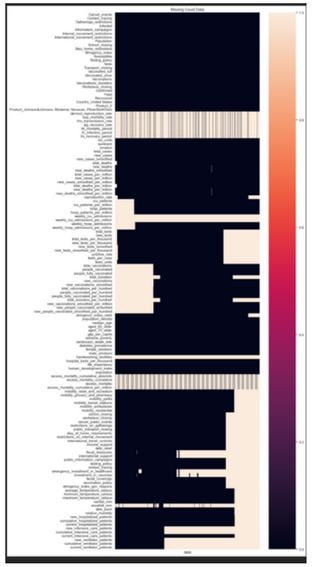
GOVERNMENT RESTRICTIONS

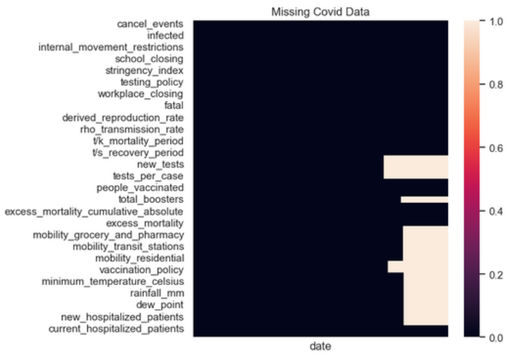
A measure indicating whether schools, parks, or businesses were closed.

WEATHER CONDITIONS

The average temperature, humidity, rainfall on the given date

The Stages of Cleaning



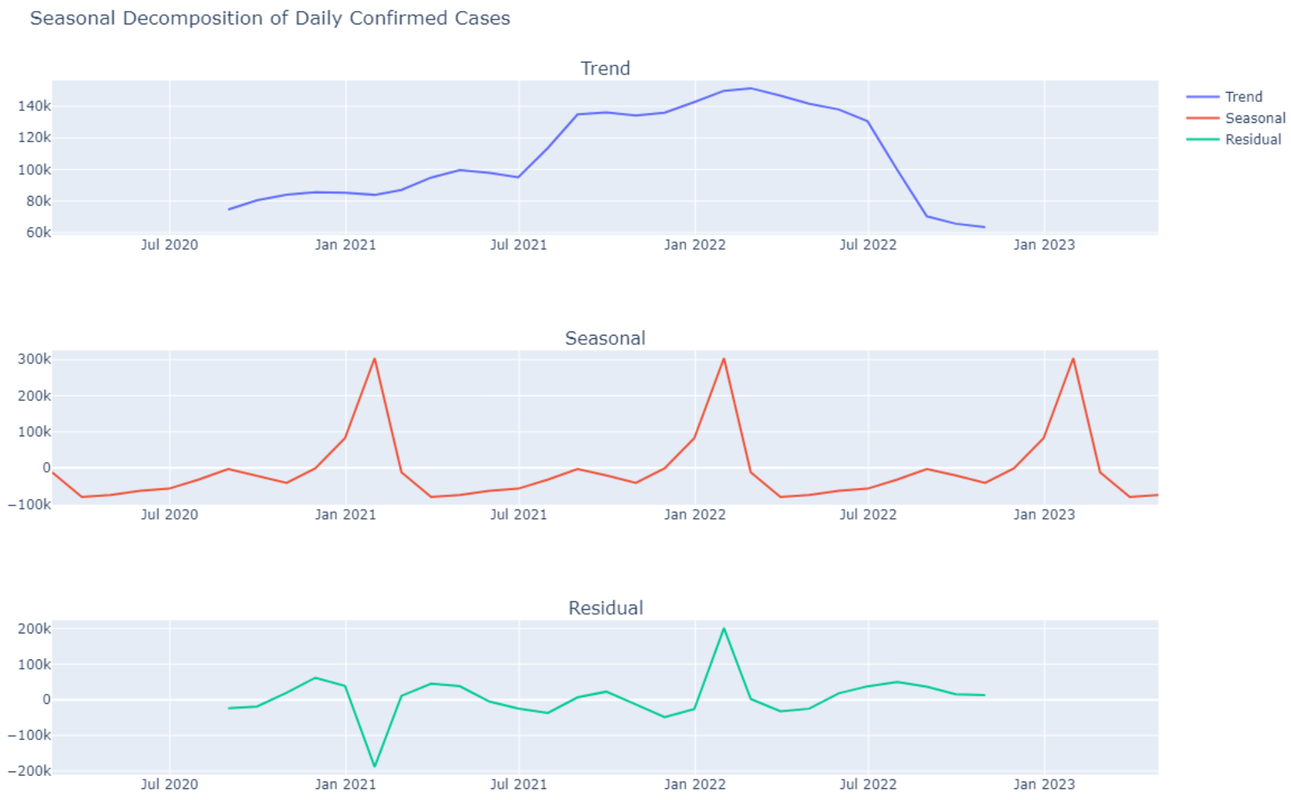


1142 x 141 columns, 30% nulls

1	**PROBLEM**	Lots of missing columns from different datasets.
2	**BRING IN MORE DATA**	Used OWID and Google COVID Hub
3	**INTERPOLATE**	Used Polynomial order 2 to interpolate. Fit the underlying data distribution, 'low error'.
4	**REMOVE COLUMNS, NARROW RANGE**	Remove Transformation Columns, Limit Date Range to 2020-02-15 to 2022-06-18 or 854days
5	**COMBINE**	Put into one master_df. Save to Parquet.

(854 Full) 1142 x 59 columns, 7% nulls

FEATURE ENGINEERING



SIRD PARAMETERS

LAG

ROLLING AVERAGE

SEASONALITY

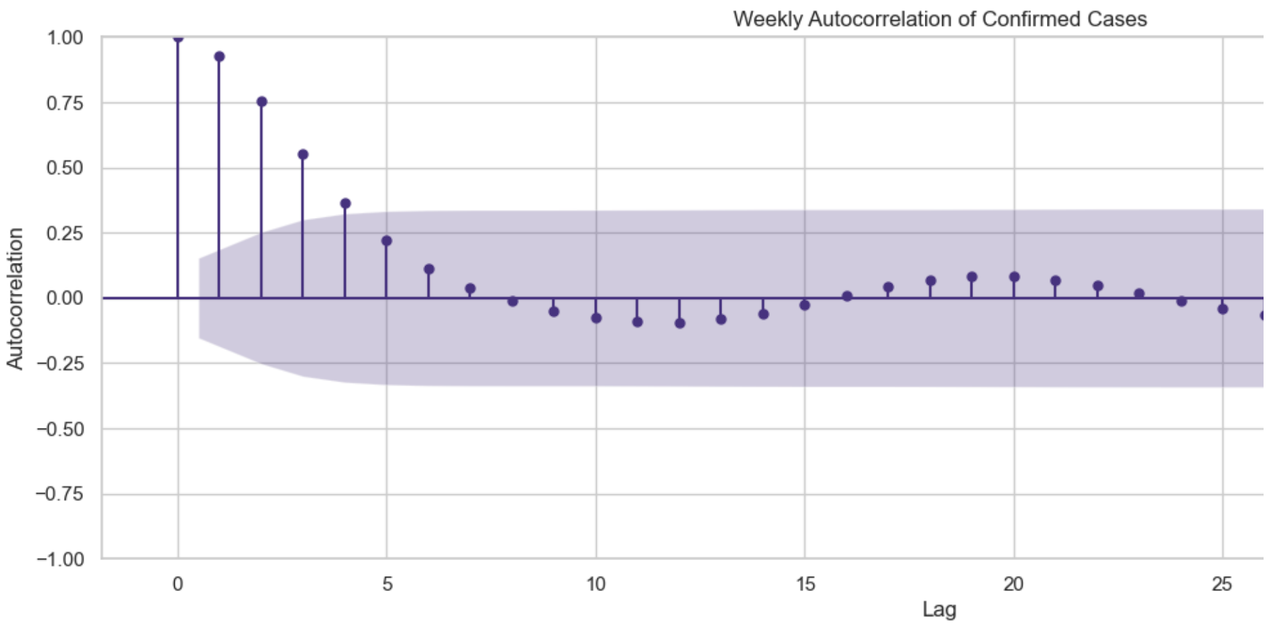
GROUP BY

ACF

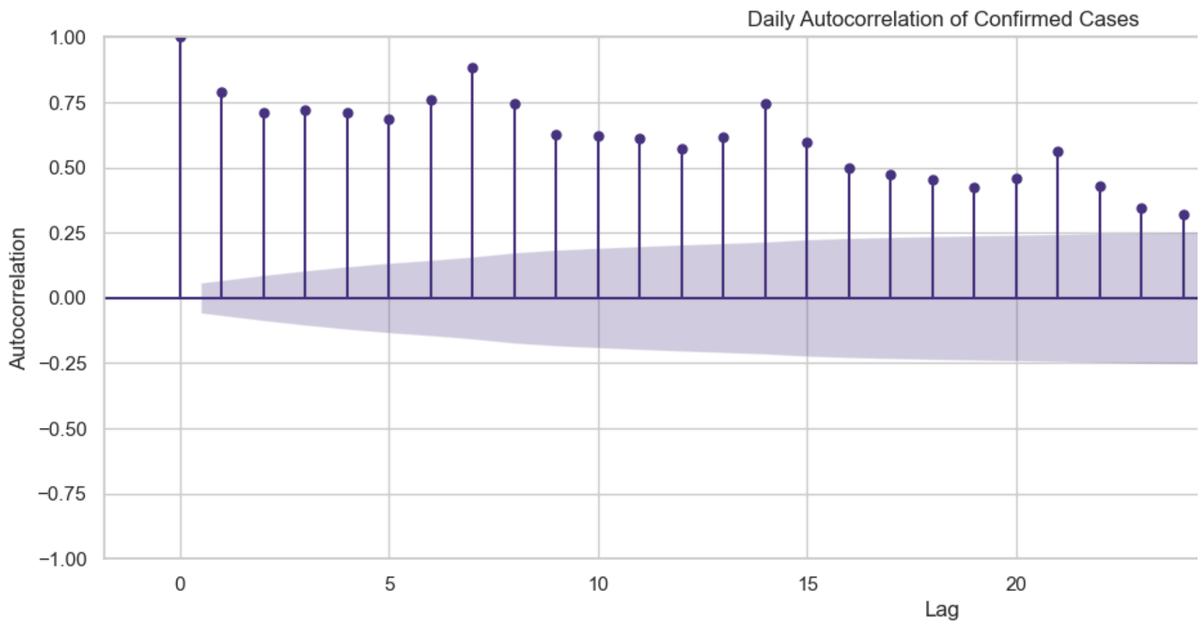
CHART SUMMARY

Some autocorrelation between weeks and days

WEEKLY



DAILY



PACF

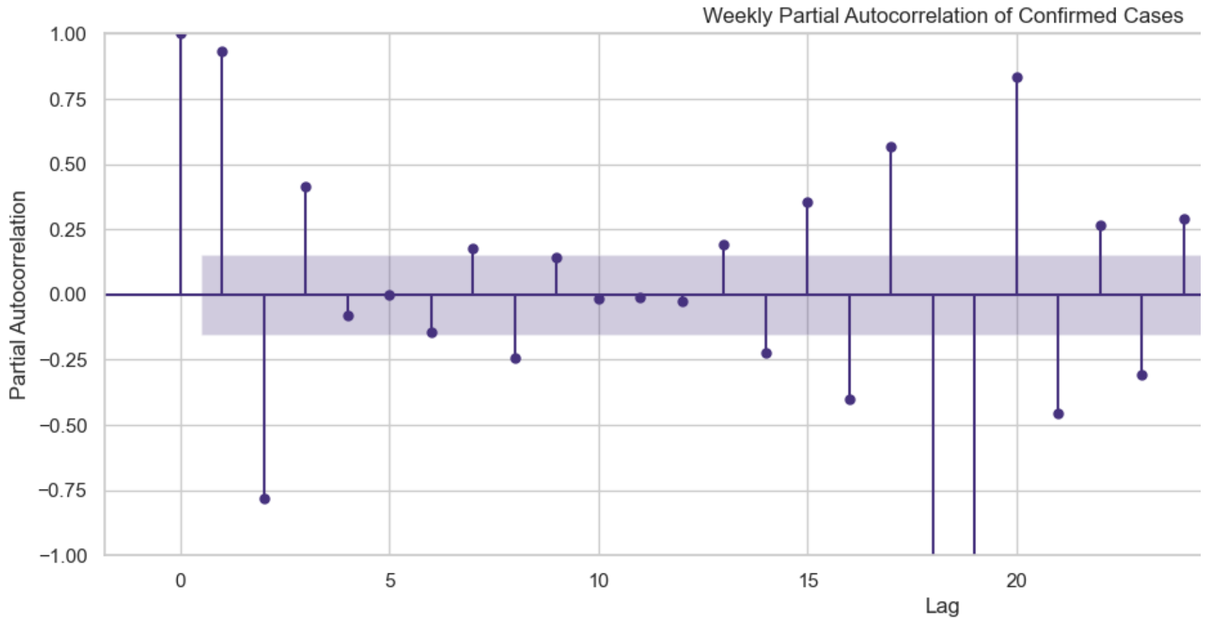
CHART SUMMARY

Some direct correlation between weeks and days

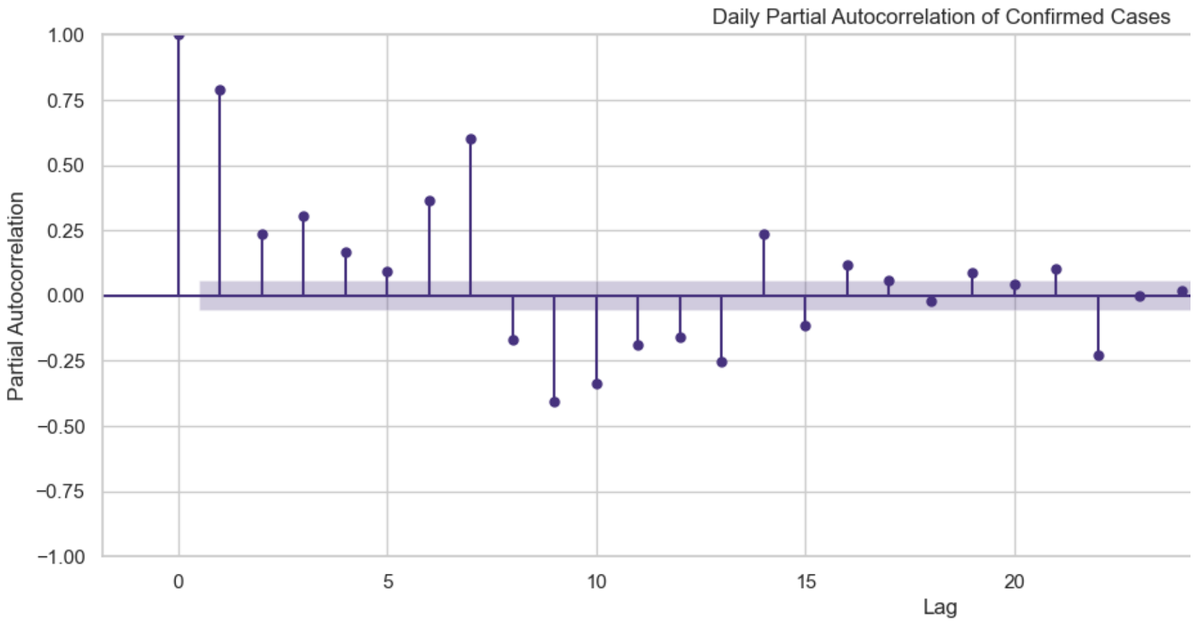
Daily ACF & PACF Heatmap

Metrics	Daily ACF	Daily PACF
0	1.00	1.00
1	0.79	0.79
2	0.71	0.23
3	0.72	0.30
4	0.71	
5	0.68	
6	0.76	0.36
7	0.88	0.60

WEEKLY



DAILY



The Stages of Modeling

1	**SARIMA**	The seasonal variation of ARIMA.
2	**SARIMAX**	Incorporate Exogenous Variables
3	**FACEBOOK PROPHET MODEL**	Incorporate Exogenous Variables
4	**RNNS**	Might not have enough data for this

SAM CELAREK

BRAINSTATION

THANKS FOR LISTENING