Clustering COVID-19 Data

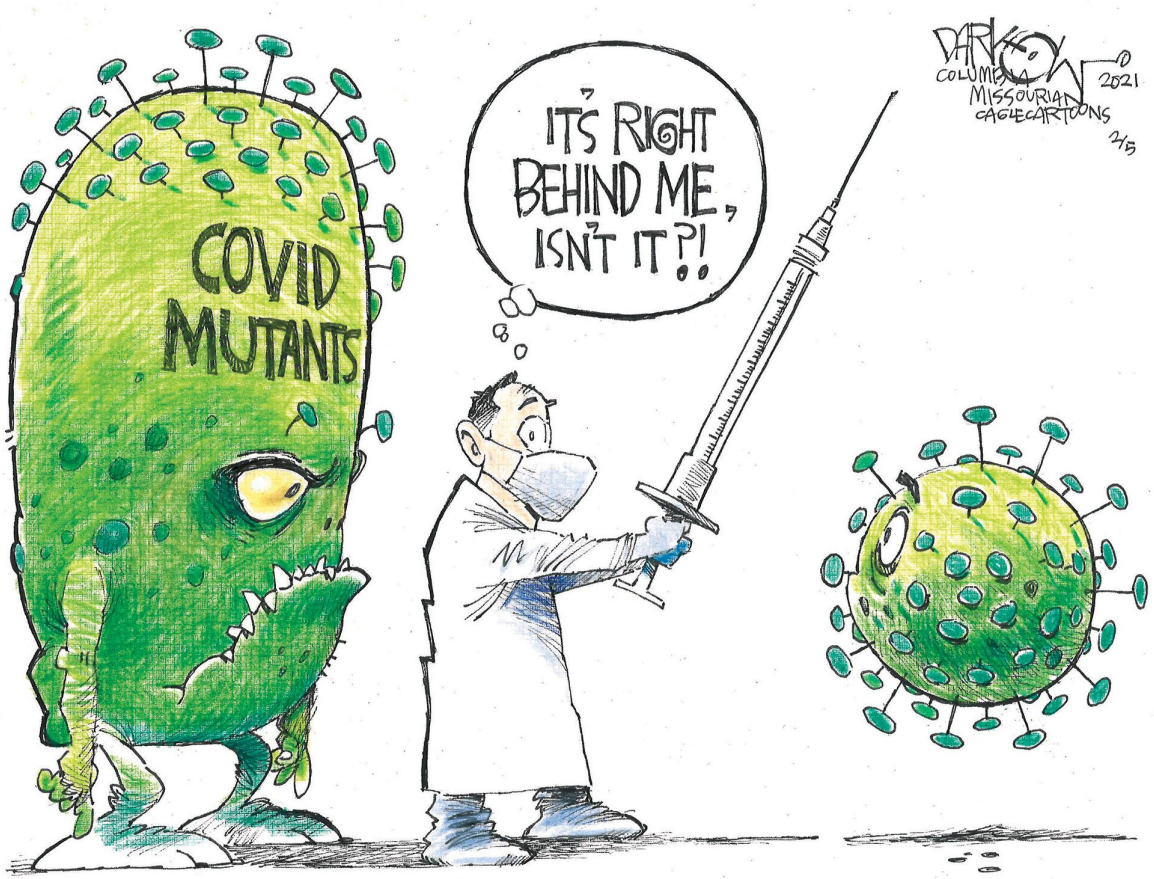
By Jenny Vo, Elisse Wright, Deb Kazos, and Christopher Mason

TOPIC

For this analysis we chose to look into recent COVID-19 data

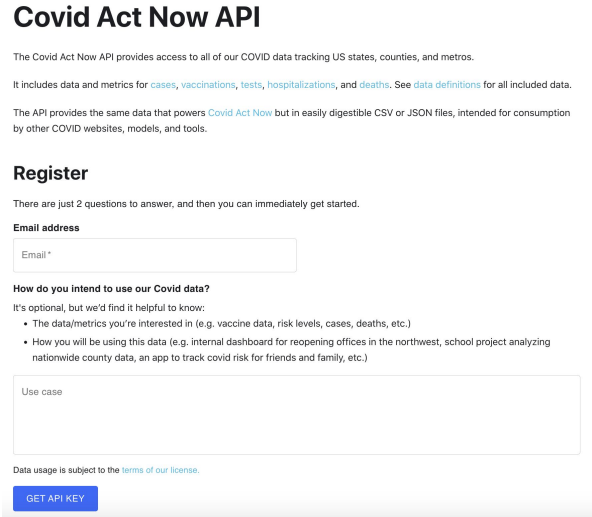
Overview:

1. Data selection
2. Goal
3. PCA Model
4. Data Analysis
5. Time Series Forecasting Models
6. Future Directions



DATA

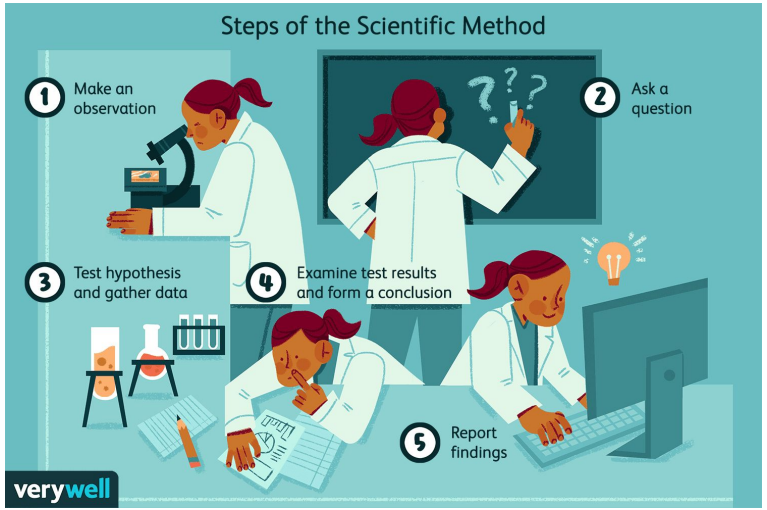
- **Selection**:
  - COVID Act Now API.
  - Time series data of all available states in the United States.
- **Data Cleanse**:
  - null rows
  - metrics data
  - empty columns
  - superfluous data



date	country	state	fips	locationId	actuals.cases	actuals.deaths	actuals.positiveTests	actuals.negativeTests	actuals.contactTracers
8/23/21	US	AK	2	iso1:us#iso2:us-ak	83581.0	397.0	131534.0	2663834.0	235.0
8/24/21	US	AK	2	iso1:us#iso2:us-ak	84174.0	410.0	132426.0	2673669.0	235.0
8/25/21	US	AK	2	iso1:us#iso2:us-ak	84794.0	421.0	133270.0	2682974.0	235.0
8/26/21	US	AK	2	iso1:us#iso2:us-ak	85521.0	421.0	134103.0	2693241.0	235.0
8/27/21	US	AK	2	iso1:us#iso2:us-ak	86055.0	421.0	134871.0	2702997.0	235.0
...	...	...	...	...	...	...	...	...	...
1/10/22	US	WY	56	iso1:us#iso2:us-wy	121519.0	1572.0	94793.0	1148706.0	50.0
1/11/22	US	WY	56	iso1:us#iso2:us-wy	122754.0	1588.0	95521.0	1150674.0	50.0
1/12/22	US	WY	56	iso1:us#iso2:us-wy	123743.0	1588.0	96401.0	1153618.0	50.0
1/13/22	US	WY	56	iso1:us#iso2:us-wy	124986.0	1588.0	97644.0	1158434.0	50.0
1/14/22	US	WY	56	iso1:us#iso2:us-wy	126468.0	1588.0	98656.0	1161656.0	50.0

GOALS

- Have a better understanding of the overall scope of COVID-19 in the United States using an unsupervised-machine learning model
- Determine if COVID-19 cases can be predicted using a time series forecasting model



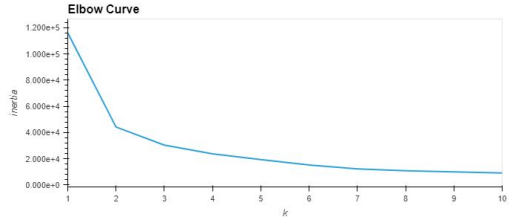
Our Primary Method: Principal Component Analysis

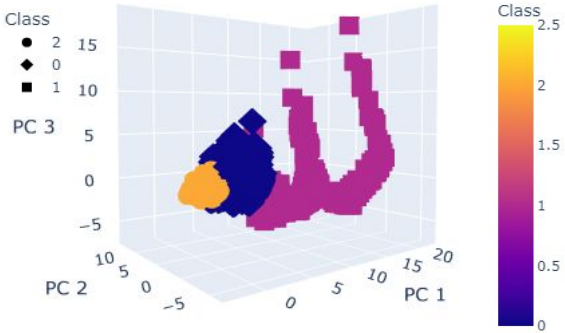
Principal Component Analysis (PCA) using **K-Means** was used to help group the most similar data points to help give us an overview of COVID-19 in the United States.

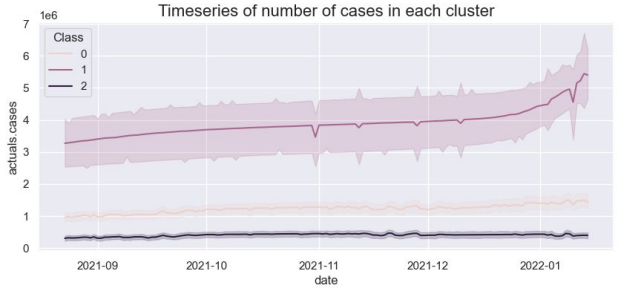
PCA was used to determine the most **significant features** that can separate the data to different clusters without the addition of location data.

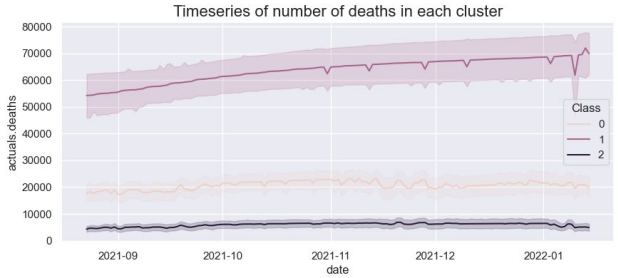
An elbow-curve was created with the PCA component data to determine the most appropriate k-means.

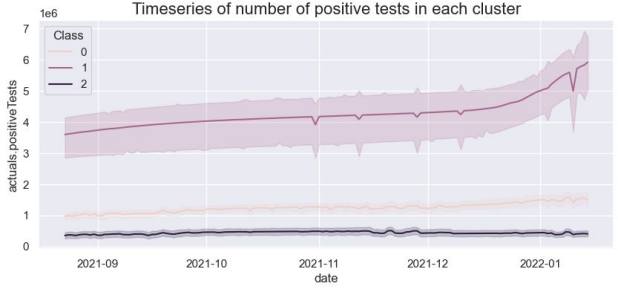
Three distinct clusters were created with the data





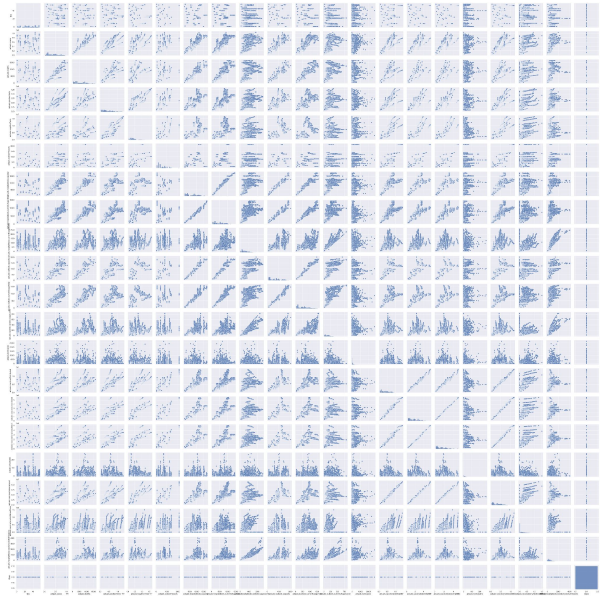


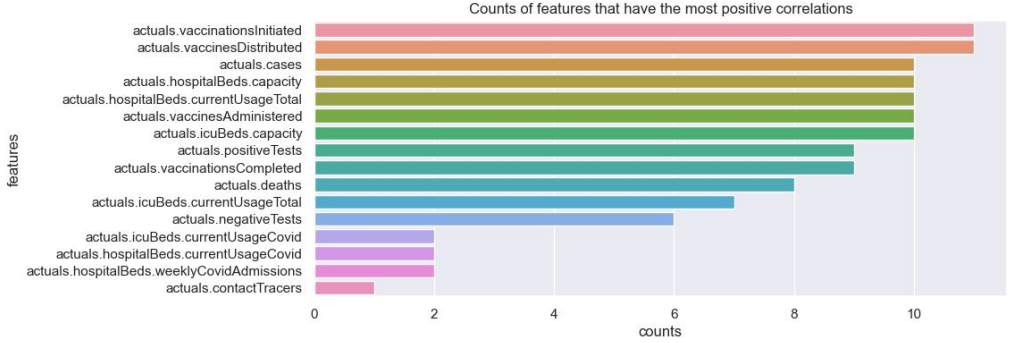


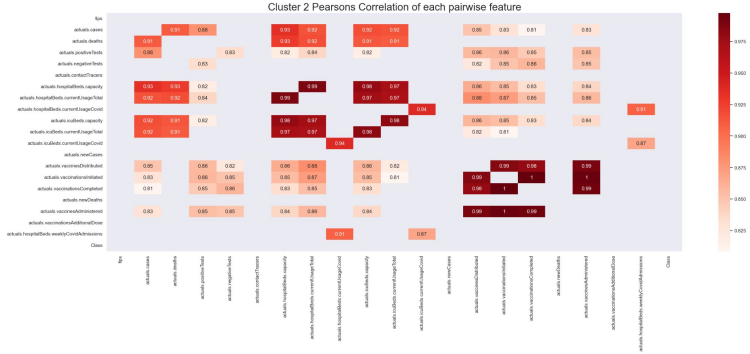


Determining significant features most likely causing clustering

To determine the features that are most likely causing the clustering, **the pearson's correlation coefficient** was determined for each pairwise feature, and **filtered** for the pairs that are significantly present in each cluster.

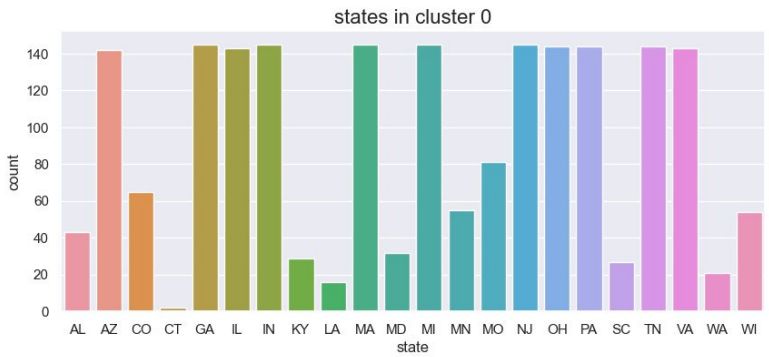


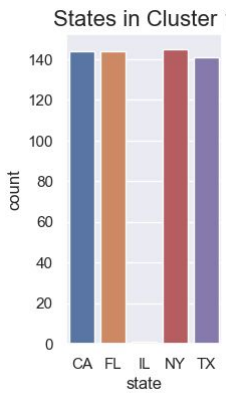


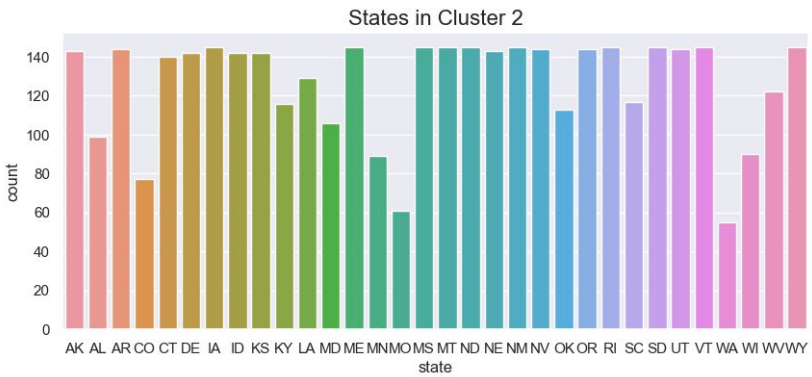


States in each cluster

Independent of location data, the PCA was able to cluster states with little separation of data-points.



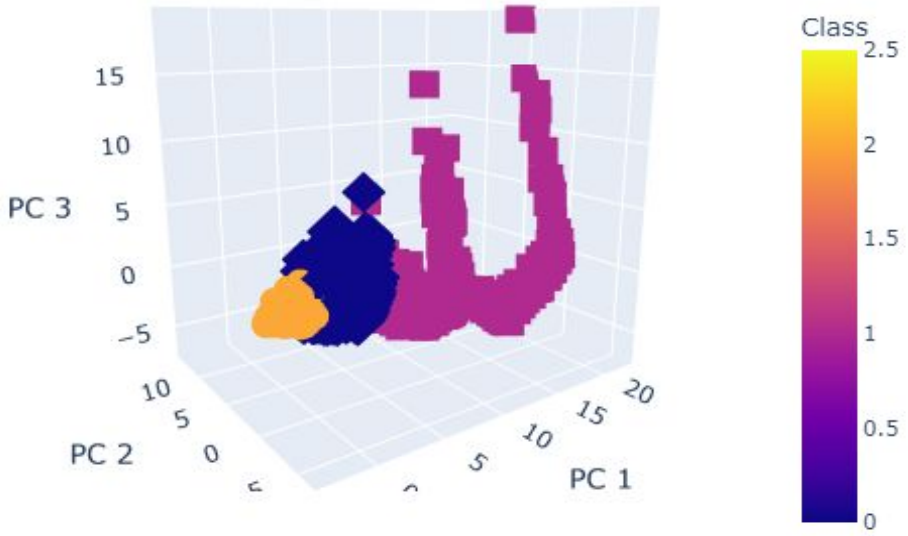




Optimizing PCA

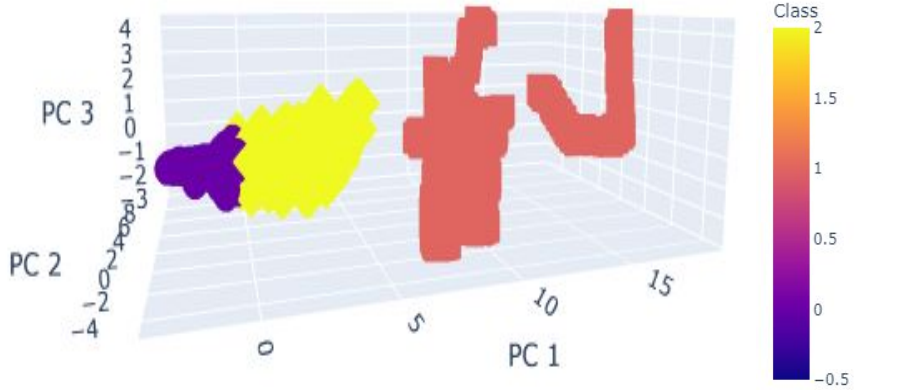
Adding Location Data:

PCA including location data does not affect the clustering



Significant features only:

PCA of only significant features separated the clusters a little bit better, but cluster 2 and 0 are still close.



To the Dashboard!

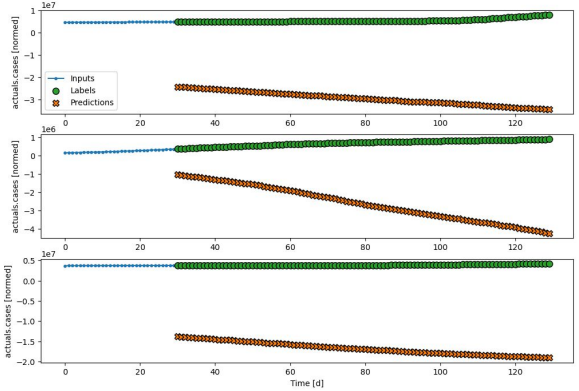
Time series forecasting

PCA being an unsupervised learning model, is not able to create predictions of data.

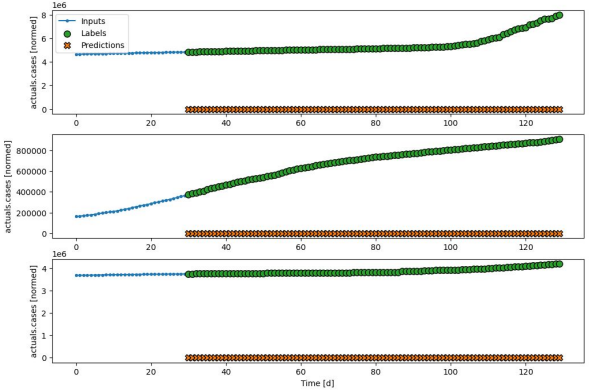
For determining forecasting, **California** datapoints were isolated from the dataset and cleaned of null values

A **(70%, 20%, 10%)** split for the **training, validation, and test sets** were created and **normalized** before being tested in different prediction forecasting models.

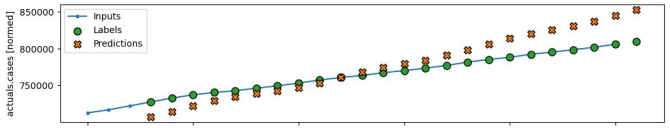
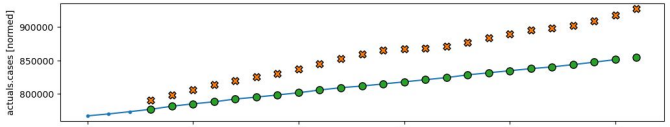
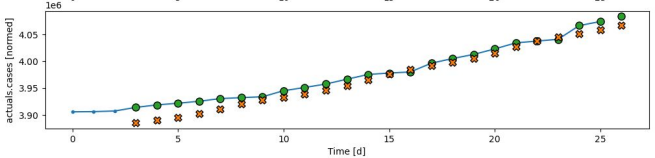
Linear Model



Recurrent Neural Network



Convolution Neural Network

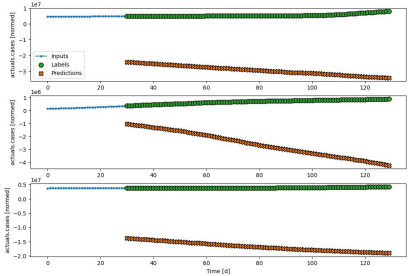
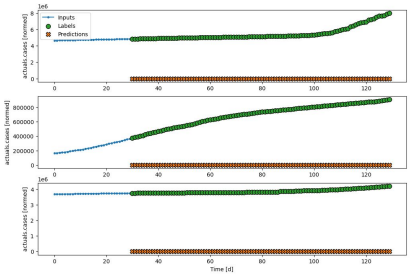
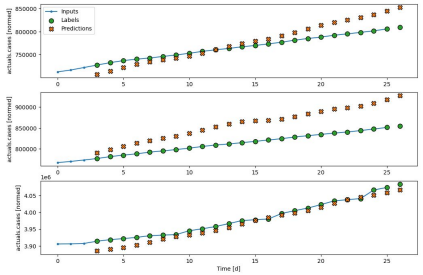
Sets		Data Split:
Training		70%
Validation		20%
Test		10%

Forecasting Prediction Model Accuracy

The accuracy of each model is 0% because no points had a perfect prediction.

For forecasting models, a mean absolute error is most appropriate for comparing forecasts with their eventual outcomes.

Mean absolute error is a measure of errors between paired observations expressing the same phenomenon. (predicted vs. observed)

	Accuracy	Mean Absolute Error
Linear Model 	0 %	9348830.0000
RNN Model 	0 %	10302069.0000
CNN Model 	0 %	435025.7812

Looking into the Future

- What features cause Illinois to be partially included in Cluster 1?
- For PCA analysis, normalize the data according to population.
- Do spearman’s correlation analysis of the clusters.
- As the rate of vaccination increases, does the death number decrease?
- What features caused Wyoming, Alabama, and Vermont to have lower case levels?