Multivariate Time Series Clustering of US States Using COVID-19 Data

Brandon Vittetoe

2023-12-12

In this study, we conducted a hierarchical clustering analysis of COVID-19 data across US states from 2021 to 2022, utilizing the dataset from the COVID-19 Data Hub. We methodically cleaned the data, preprocessed it, and performed gap statistic analysis. Our analysis identified distinct regional clusters that reflect the varied impact of the pandemic.

Introduction

The COVID-19 pandemic has profoundly impacted public health systems worldwide, with the United States facing significant challenges in controlling the virus's spread. This paper presents a clustering analysis of US states based on COVID-19 data trends, focusing on cases, deaths, tests, vaccines, hospitalizations, and ICU admissions. Our methodology includes data cleaning, preprocessing, and clustering.

The dataset, sourced from the COVID-19 Data Hub (Guidotti & Ardia, 2020), provides daily summaries of epidemiological data, policy measures, mobility data, and geospatial identifiers from government sources globally (Guidotti, 2022). We accessed the dataset through the `covid19` function from the `COVID19` package.

```
library(COVID19)

covid19(country = c("US"), level = 2)
```

This subset comprises approximately 75,000 rows and 50 columns. Our analysis focused on the 50 US states (excluding territories like Puerto Rico and Washington DC) between 2021 and 2022. Furthermore, we focued on the epidemiological variables, defined in Table 1.

Table 1: Definitions of Epidemiological Variables

Variable	Description
confirmed	Cumulative number of confirmed cases.
deaths	Cumulative number of deaths.
recovered	Cumulative number of patients released from hospitals or reported recovered.
tests	Cumulative number of tests.
vaccines	Cumulative number of total doses administered.
hosp	Number of hospitalized patients on date.
icu	Number of hospitalized patients in intensive therapy on date.
vent	Number of patients requiring invasive ventilation on date.

However, we excluded the `recovered` and `vent` variables from the analysis due to significant missing data, with about 95% of values missing. Table 2 summarizes the missing values per variable.

Table 2: Missing Values per Variable

Variable	Missing Values
confirmed	0
deaths	0
recovered	34586
tests	2
vaccines	0
hosp	0
icu	0
vent	34586

During the cleaning process, we will address the remaining missing values in the `tests` variable.

Data Cleaning

To ensure the integrity of our analysis, we performed the following data cleaning steps:

1. **Column Selection**: We selected these columns for our analysis: `date`, `confirmed`, `deaths`, `tests`, `vaccines`, `hosp`, `icu`, and `administrative_area_level_2`.
2. **Row Filtering**: We included only the data for the 50 US states from the years 2021 and 2022.
3. **Missing Value Interpolation**: We employed linear interpolation to fill in missing values in the `tests` variable.
4. **Cumulative Variable Transformation**: We converted the cumulative `confirmed`, `deaths`, `tests`, and `vaccines` data into daily increments.

Table 3 presents the first 10 rows of the cleaned data.

Table 3: First 10 Rows of Cleaned Data.

date	confirmed	deaths	tests	vaccines	hosp	icu	administrative_area_level_2
2021-01-01	4521	45	13414	1487	3013	790	Alabama
2021-01-01	0	0	5002	206	69	11	Alaska
2021-01-01	6438	136	33266	10265	4661	1017	Arizona
2021-01-01	4304	35	8359	961	1211	353	Arkansas
2021-01-01	37951	271	240083	5807	21121	4556	California
2021-01-01	2778	57	26709	1063	1101	313	Colorado

The cleaned dataset now contains 36,500 rows, which corresponds to 365 days per year over 2 years for each of the 50 states, and it is free of missing values. Figure 1 illustrates the cleaned data for four representative states—California, Michigan, New York, and Texas—showcasing the West, Midwest, Northeast, and South regions.

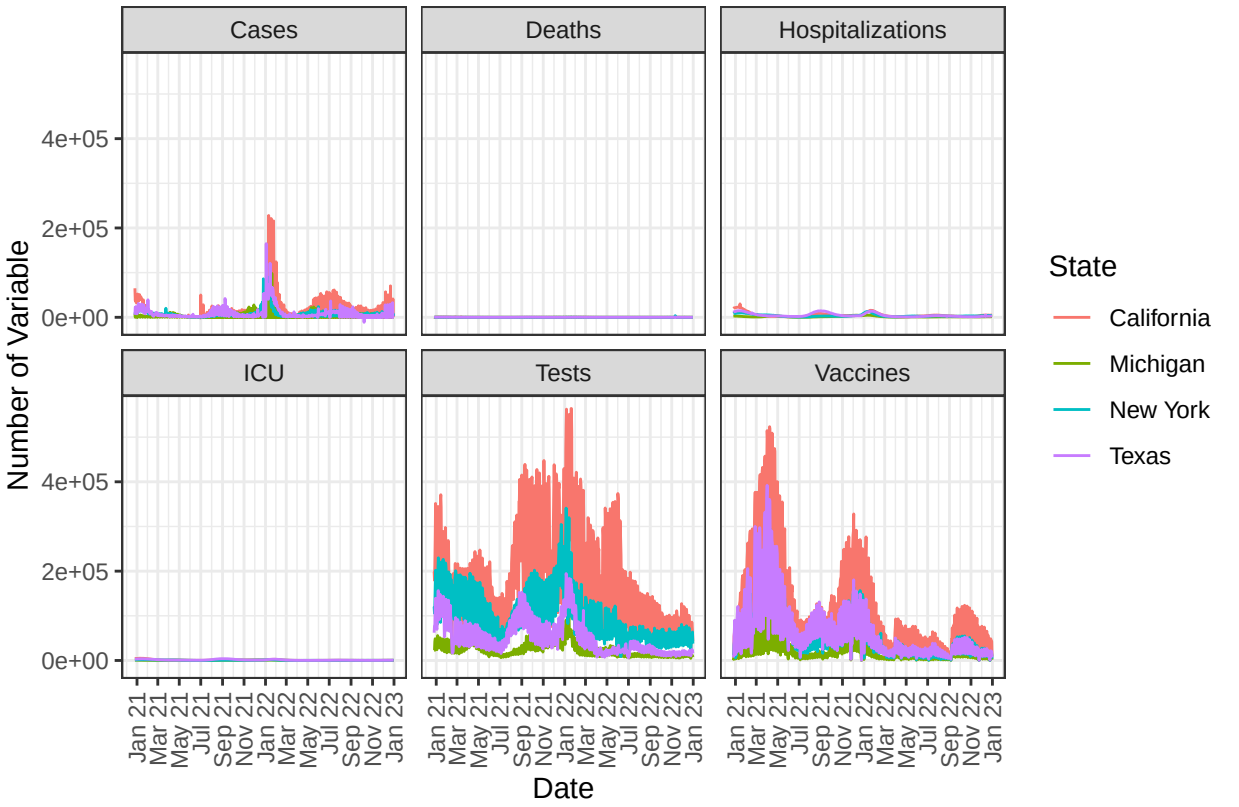

Figure 1: COVID−19 Variables vs Time (Cleaned)

However, Figure 1 reveals three issues with the data. Firstly, the jagged nature of the data could negatively impact the distance calculations in our clustering analysis. Secondly, the lack of scaling within variables could introduce bias towards the `population` variable during clustering. Thirdly, the absence of scaling between variables could skew the clustering towards the `tests` and `vaccines` variables. We will rectify these issues during preprocessing.

Data Preprocessing

In the preprocessing stage, we focused on two key steps: smoothing and scaling.

For smoothing, we considered two techniques: the 7-day moving average, which averages data over a week to mitigate short-term fluctuations, and LOESS (Locally Estimated Scatterplot Smoothing), which employs local polynomial regression to fit the data more accurately. Figure 2 illustrates a comparison between these two methods.

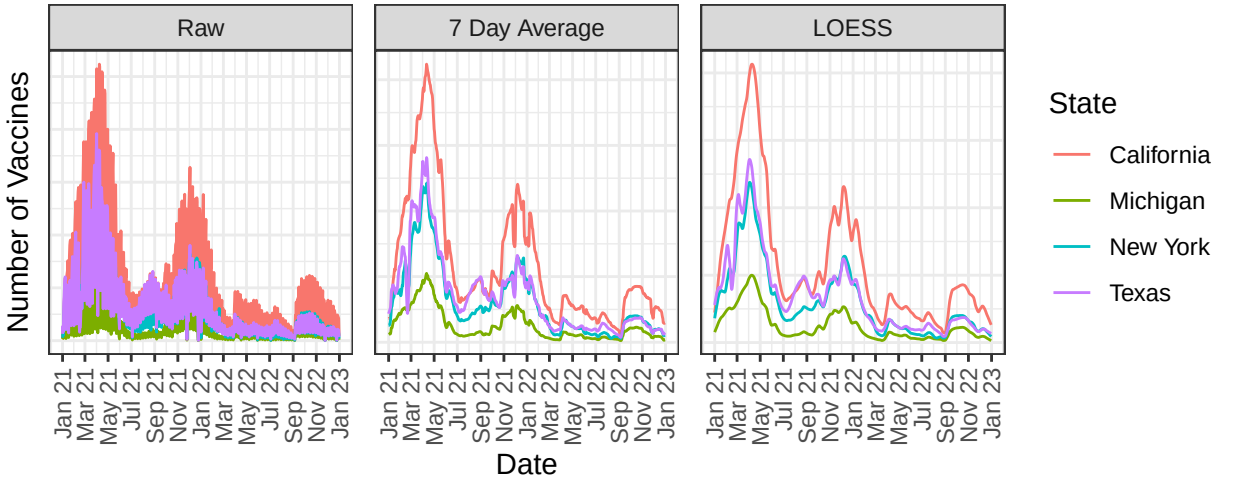

Figure 2: Comparison of Smoothing Methods

The results indicated that LOESS was superior to the 7-day moving average in terms of preserving the original trends of the time series without excessive smoothing. Following the smoothing process, we proceeded to scale the data to ensure comparability. Figure 3 displays the scaled data for the states of California, Michigan, New York, and Texas.

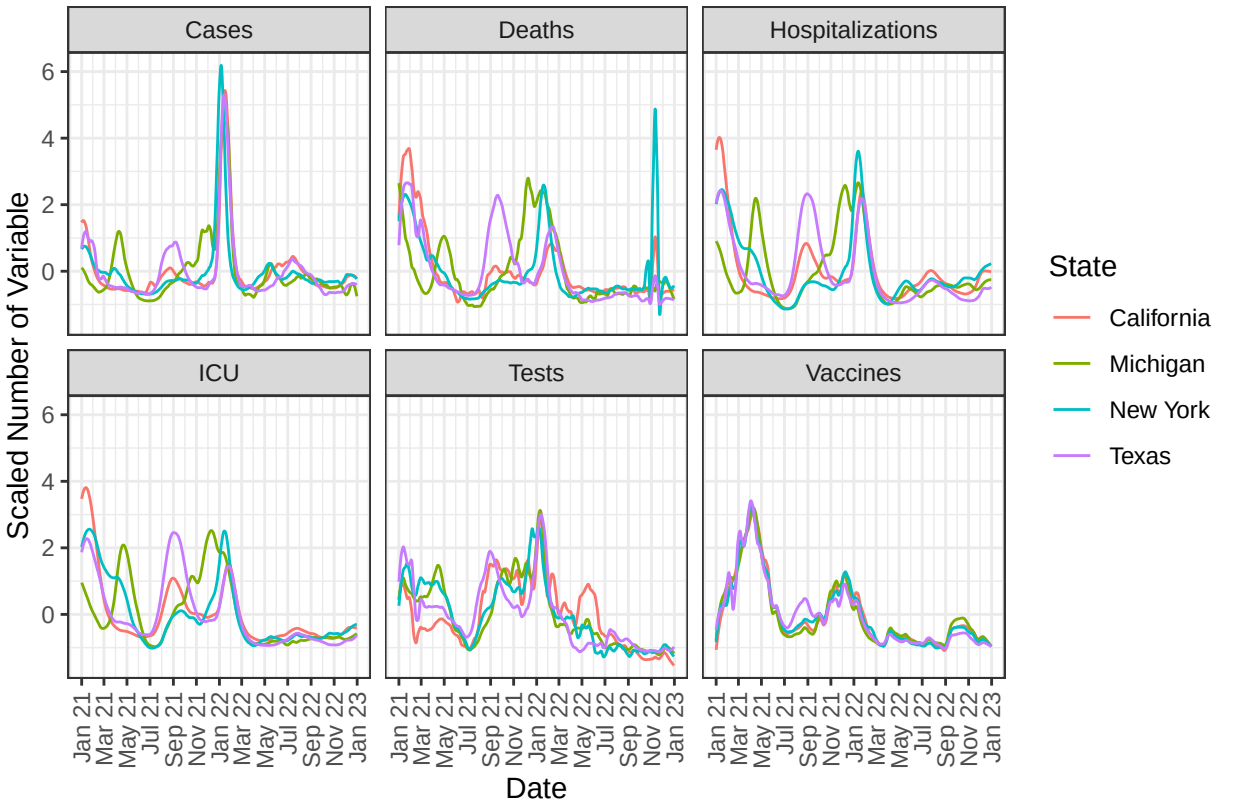

Figure 3: COVID−19 Variables vs Time (Preprocessed)

The application of both smoothing and scaling techniques ensures that the time series data can be compared equitably, both within individual variables and across different variables. For an in-depth discussion on these preprocessing methods, refer to "Local Regression Models" and "The New S Language" (Becker, Chambers, & Wilks, 1988; Cleveland, Grosse, & Shyu, 2017). With the data now preprocessed, we are ready to cluster the data.

Data Clustering

In our study, we employed hierarchical clustering to group the data. The initial step was to calculate the distances between the states. We generated a 50 x 50 distance matrix for each epidemiological variable using the Euclidean distance measure. We then aggregated these matrices to create a comprehensive distance matrix, ensuring equal weighting for each variable.

With the overall distance matrix prepared, we integrated it into the hierarchical clustering algorithm. For additional information, refer to "The New S Language" (Becker et al., 1988). Figure 4 depicts the dendrogram resulting from our analysis.

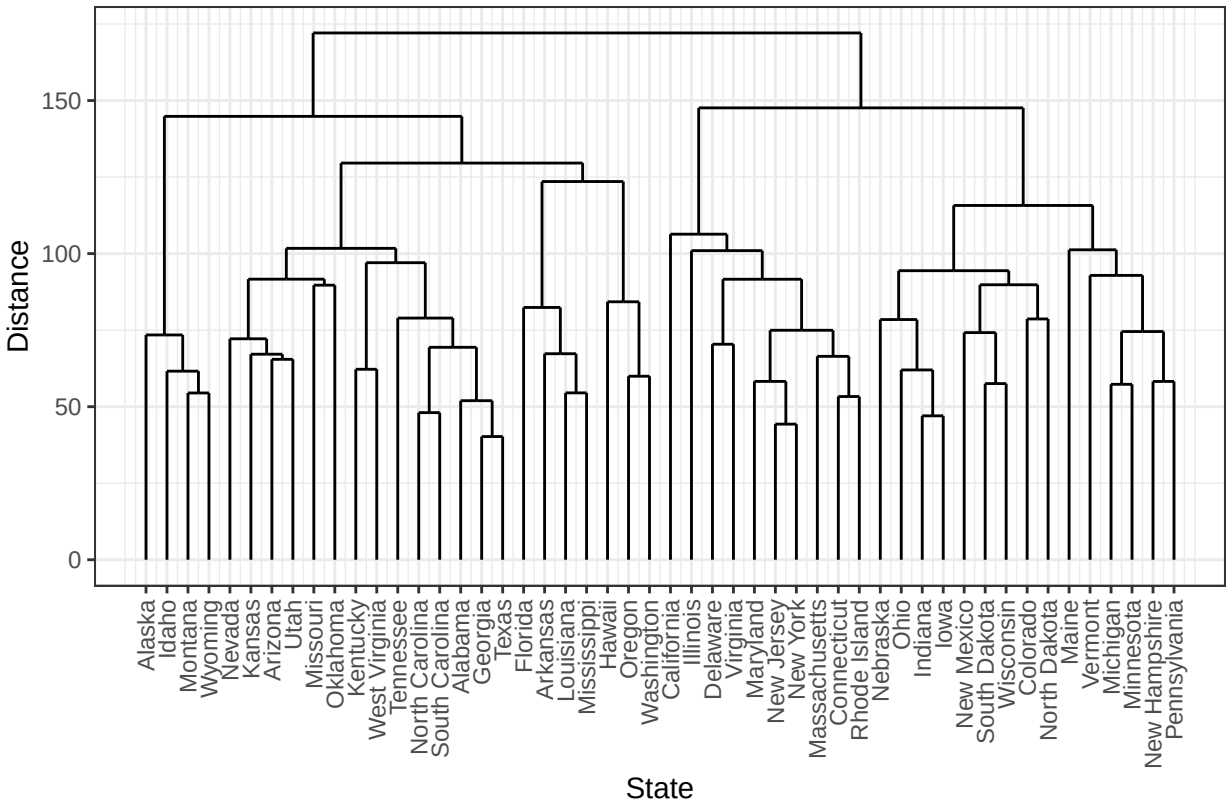

Figure 4: States Clustered by COVID−19 Variables

To determine the optimal number of clusters cluster, we computed the gap statistic for cluster numbers ranging from 1 to 10. The gap statistic compares the within-cluster variation with the expected variation under a reference null distribution, which we generated via bootstrapping. The greater the gap, the more distinct the clustering structure is from randomness (Tibshirani, Walther, & Hastie, 2001). Figure 5 illustrates the gap statistic, including the standard errors, for various cluster counts.

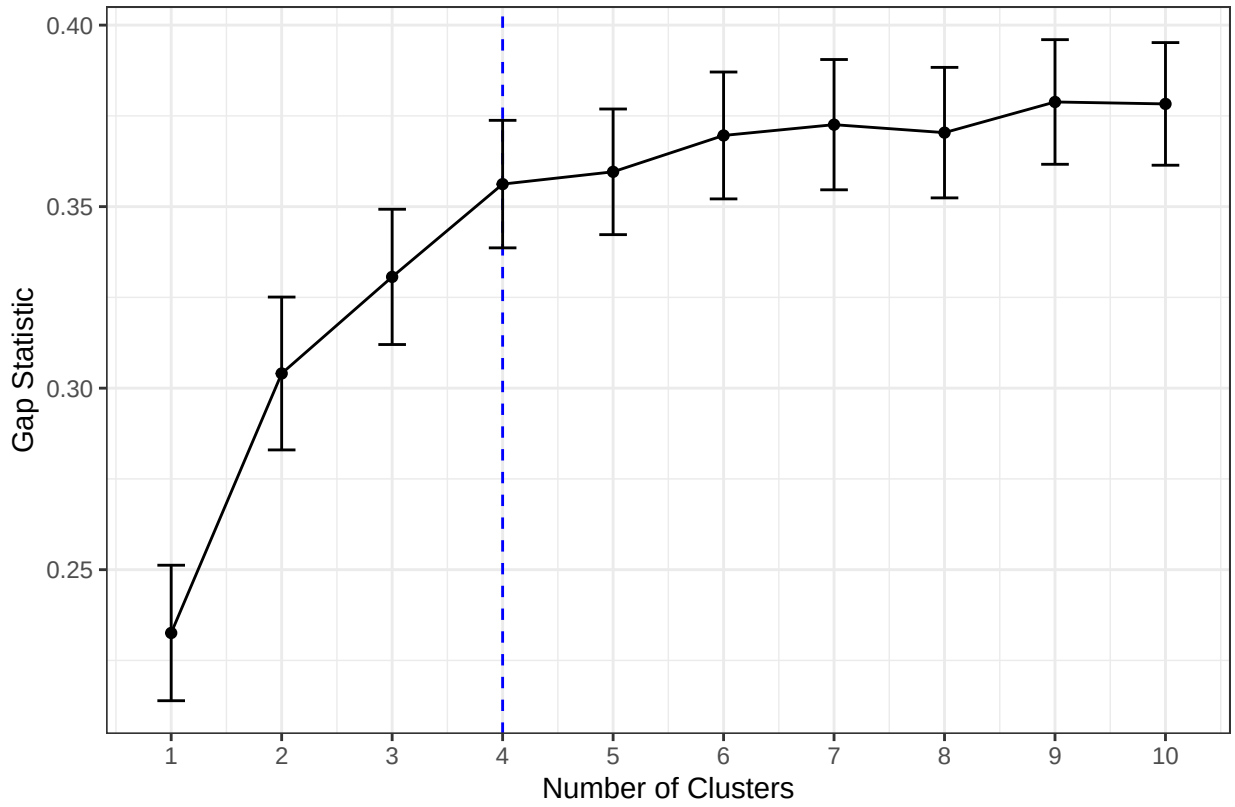

Figure 5: Gap Statistic vs Number of Clusters

The figure indicates the gap statistic with an error margin of plus or minus one standard error. We applied the `firstSEmax` method to identify the smallest cluster number whose gap statistic lies within one standard error of the first local maximum (Tibshirani et al., 2001). Here, the first local maximum is observed at $k = 7$, and the optimal cluster number is determined to be $k = 4$.

Upon establishing the optimal cluster number, we segmented the dendrogram accordingly. Figure 6 presents the clusters superimposed on a map of the United States.

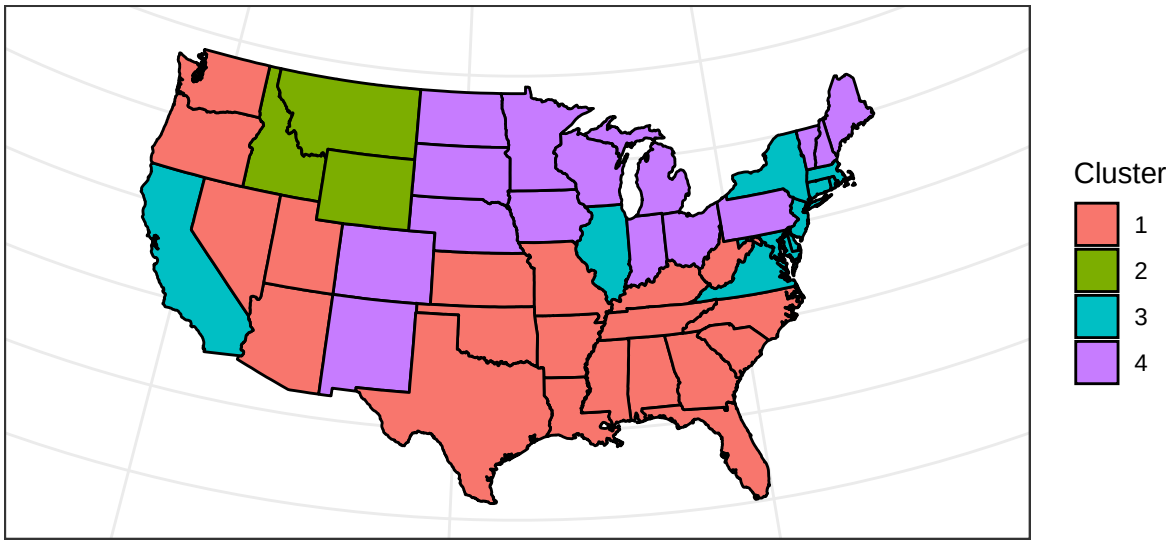

Figure 6: States Clustered by COVID−19 Variables

The map highlights regional trends within the COVID-19 data. A substantial cluster encompasses the South and extends into the West, while another significant cluster spans the Midwest, reaching into the Northeast and Southwest, indicating regional similarities in pandemic patterns. Moreover, a smaller cluster in the Northwest and a distinct, non-contiguous cluster comprising several Northeastern states, Illinois, and California, reflect diverse responses and outcomes to the pandemic. Notably, Alaska and Hawaii are part of clusters 2 and 1, respectively.

Conclusion

In our research, we utilized hierarchical clustering on multivariate time series data of COVID-19 metrics to uncover distinct regional patterns of the pandemic's impact throughout the US. Our methodical approach, which included data cleaning, preprocessing, and employing the gap statistic to determine the optimal number of clusters, led to the identification of pronounced clusters in the South and Midwest. Additionally, we discovered a smaller cluster in the Northwest and a distinctive grouping that included several Northeastern states, Illinois, and California. These results offer critical insights for shaping targeted public health policies and underscore the pivotal role of data-driven analysis in comprehending and addressing the challenges of the pandemic.

References

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). The new s language. Wadsworth & brooks. *Cole.[Google Scholar]*.

Cleveland, W. S., Grosse, E., & Shyu, W. M. (2017). Local regression models. In *Statistical models in s* (pp. 309–376). Routledge.

Guidotti, E. (2022). A worldwide epidemiological database for COVID-19 at fine-grained spatial resolution. *Scientific Data*, *9*(1), 112. doi:10.1038/s41597-022-01245-1

Guidotti, E., & Ardia, D. (2020). COVID-19 data hub. *Journal of Open Source Software*, *5*(51), 2376. doi:10.21105/joss.02376

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.