**COVID-19 CASE ANALYSIS**

**OBJECTIVE**

The US COVID-19 Trends and Impact Survey (CTIS) has operated continuously since April 6, 2020, collecting over 20 million responses. As the largest public health survey conducted in the United States to date, CTIS was designed to facilitate detailed demographic and geographic analyses, track trends over time, and accommodate rapid revision to address emerging priorities. Using examples of CTIS results illuminating trends in symptoms, risks, mitigating behaviors, testing, and vaccination in relation to evolving high-priority policy questions over 12 mo of the pandemic, we illustrate the value of online surveys for tracking patterns and trends in COVID outcomes as an adjunct to official reporting, and showcase unique insights that would not be visible through traditional public health reporting.

**Abstract**

The US COVID-19 Trends and Impact Survey (CTIS) is a large, cross-sectional, internet-based survey that has operated continuously since April 6, 2020. By inviting a random sample of Facebook active users each day, CTIS collects information about COVID-19 symptoms, risks, mitigating behaviors, mental health, testing, vaccination, and other key priorities. The large scale of the survey—over 20 million responses in its first year of operation—allows tracking of trends over short timescales and allows comparisons at fine demographic and geographic detail. The survey has been repeatedly revised to respond to emerging public health priorities. In this paper, we describe the survey methods and content and give examples of CTIS results that illuminate key patterns and trends and help answer high-priority policy questions relevant to the COVID-19 epidemic and response. These results demonstrate how large online surveys can provide continuous, real-time indicators of important outcomes that are not subject to public health reporting delays and backlogs. The CTIS offers high value as a supplement to official reporting data by supplying essential information about behaviors, attitudes toward policy and preventive measures, economic impacts, and other topics not reported in public health surveillance systems.

**Sign up for PNAS alerts.**

During 2020, the coronavirus disease 2019 (COVID-19) pandemic precipitated the need for new public health surveillance to inform urgent policy decisions. Effective pandemic policy-making requires information on a broad array of indicators, including local morbidity and mortality, preventive behaviors, healthcare capacity, and economic impacts. Given the critical importance of COVID-19 trends for policy, health departments set up routine public reporting systems for tracking cases, deaths, testing, and hospitalizations ([1](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r1)). However, supplemental data can both augment official reporting, for example by providing additional indicators of COVID-19 prevalence not subject to reporting delays and backlogs, and supply complementary information about public behavior, attitudes toward policy and preventive measures, mental health, economic impacts, and other items not observed in public health surveillance systems.

A number of efforts have used surveys to provide supplemental surveillance data. For example, symptom-tracking smartphone apps invite users to self-report symptoms, in some cases encouraging repeated participation to enable longitudinal tracking ([2](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r2)–[4](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r4)). Other surveys have addressed broader impacts of the pandemic, such as economic consequences ([5](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r5)). In this paper, we present findings from the Delphi Group at Carnegie Mellon University (CMU) US COVID-19 Trends and Impact Survey (CTIS), in partnership with Facebook, which has operated continuously since April 6, 2020 and collected over 20 million responses. [An international version of the survey is described in a companion paper in this theme issue ([6](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r6)).]

As the largest public health survey conducted in the United States to date, CTIS is designed to facilitate detailed demographic and geographic analyses, to track trends over time, and to accommodate rapid response to emerging priorities ([7](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r7)). A random sample of Facebook active users are invited each day to complete a questionnaire comprising survey items on symptoms, COVID testing, social distancing, vaccination, schooling, mental health, and economic security. The survey instrument has been updated frequently to incorporate new policy-relevant topics. Results are aggregated and made publicly available, and microdata are available under institutional data use agreement, in both cases with fewer than 3 d of lag. These data provide information at a level of geographic and temporal detail that can supply essential inputs into short-term decision-making and longer-term strategic planning. These data also facilitate retrospective analysis of patterns, trends, and associations, supporting longer-term research on health policy decisions and the impacts of the pandemic.

In this study, we first compare COVID-19 indicators from CTIS with publicly reported case, hospitalization, and mortality data between April 2020 and April 2021. Despite potential limitations of our internet-based sample and the voluntary nature of the survey, we demonstrate high correspondence between the two, with CTIS less affected by holiday-related reporting anomalies. Second, we examine patterns and trends in symptoms, risks, mitigating behaviors, testing, and vaccination in US states and localities, in relation to evolving high-priority policy questions over 12 mo of the pandemic. The findings illustrate the value of online surveys for tracking patterns and trends in COVID-related outcomes as an adjunct to official reporting, while also showcasing insights that are only possible through a large-scale survey effort.

**Methods**

**Sampling and Recruitment.**

The US CTIS launched on April 6, 2020 and has run continuously since that time, with an average of more than 350,000 people participating each week over the first year of operation. The survey is implemented by the Delphi Group at CMU, with participants recruited via the Facebook platform. Every day, Facebook invites a new sample of active users ages 18 y or older to participate in the survey. Facebook uses stratified random sampling within US states to randomly select a sample of its users to see the survey invitation at the top of their News Feed. Users who click on the invitation are taken to the CMU-administered survey hosted on Qualtrics. To ensure privacy, Facebook does not see any individual survey response during or after the data collection. The survey is available in English, Spanish, Brazilian Portuguese, Vietnamese, French, and simplified Chinese.

**Survey Design.**

The survey instrument was deployed in multiple waves from launch through April 5, 2021, with contents of each survey version summarized in [Table 1](https://www.pnas.org/doi/10.1073/pnas.2111454118#t01). Revisions are ongoing as new public health needs arise. A number of core items have been included consistently across all survey versions, including questions about symptoms, contacts, and demographics. Key additions include items on mask wearing and occupation, added in September 2020, seasonal flu vaccination and schooling, added in November 2020, and COVID-19 vaccination, added in December 2020. As of April 5, 2021, the range of survey items spanned the following broad categories: household and individual symptoms, common comorbidities, contact patterns and mitigating behaviors, testing and diagnosis, worry and financial impact, schooling, vaccination, and demographics.

**Table 1.**

Summary of survey waves deployed between April 6, 2020 and April 5, 2021

| **Wave**[**\***](https://www.pnas.org/doi/10.1073/pnas.2111454118#tfn1) | **Contents** | **Start date** | ***n* (in millions)** |
| --- | --- | --- | --- |
| 1 | Household and individual symptoms | April 6, 2020 | 1.1 |
| Common comorbidities |
| Contacts with others |
| Anxiety, depression |
| Financial impact |
| Demographics (age, gender) |
| 2 | New question: Symptoms among “people in your local community that you know personally” | April 15, 2020 | 2.6 |
| Minor textual revisions |
| 3 | Translated into simplified Chinese, Spanish, French, Brazilian Portuguese, Vietnamese | May 21, 2020 | 7.4 |
| Minor textual revisions |
| 4 | New questions: Medical care sought, COVID testing and results, mask wearing, social isolation | September 8, 2020 | 3.0 |
| Additional demographics, including race, ethnicity, occupation, education |
| Textual revisions |
| Some unused items removed |
| 5 | New questions: Seasonal flu vaccination, schooling, and school precautions | November 24, 2020 | 1.3 |
| Textual revisions |
| 6 | New questions: Vaccine intent. Vaccine status item enabled on January 6, 2021 | December 19, 2020 | 1.2 |
| 7 | Textual revisions to vaccine intent items | January 12, 2021 | 1.3 |
| 8 | New questions: Reasons for vaccine hesitancy, vaccine dosing | February 8, 2021 | 0.9 |
| Minor textual revisions |
| 10 | New questions: Appointments for COVID vaccines, information about getting vaccinated | March 2, 2021 | 1.4 |
| Textual revisions |

EXPAND FOR MORE

\*

There was no Wave 9 survey. The numbering of waves skipped from 8 to 10 to synchronize numbering conventions with the international version of the survey.

[**OPEN IN VIEWER**](https://www.pnas.org/doi/10.1073/pnas.2111454118#t01)

Full versions of all survey instruments can be found at <https://cmu-delphi.github.io/delphi-epidata/symptom-survey/coding.html>.

**Weighting.**

Analytic weights have been developed to adjust for differences between Facebook users and the United States population, and to adjust for biases related to coverage and nonresponse ([8](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r8)). When Facebook links users to the survey, it generates a random unique identifier that is passed to CMU. For users who complete the survey, CMU returns the corresponding identifiers to Facebook, which then calculates analytic weights in two steps:

**1.**

To adjust for nonresponse bias, Facebook calculates the inverse probability that sampled users complete the survey using their age, gender, and geographical variables, as reported on their Facebook profiles, as well as other characteristics known to correlate with nonresponse. The inverse probabilities are then used to create weights for responses, after which the survey sample reflects the active adult user population on Facebook.

**2.**

To adjust for coverage bias, Facebook poststratifies the weights created in the first step so that the distribution of age, gender, and state or territory of residence in the survey sample reflects that of the general population.

The analytic weight does not identify the survey respondent. The weight for an individual is scaled to approximate the number of people in the adult population represented by that individual based on age, gender, location, and date. Facebook passes these weights to CMU. To protect respondent privacy, CMU cannot use these weights to identify specific Facebook users, and Facebook never receives individual survey responses and cannot link them to specific users.

**Analysis.**

In this study we examined a range of different outcomes measured in the CTIS over the period April 6, 2020 to April 5, 2021. We analyzed both aggregated macrolevel data and individual-level data to highlight some of the key questions that may be examined using CTIS. Across the examples presented in this paper, we have stratified analyses by individual characteristics, including age, race/ethnicity (using categories consistent with National Center for Health Statistics), and occupation, and by geographic divisions, including Census region, Census division, state, and county.

To examine the representativeness of the study sample, we compared characteristics of the sample to data from the American Community Survey 2019 (ACS) supplemental estimates.

We evaluated reported symptoms and symptom patterns in comparison to surveillance data on confirmed COVID-19 hospitalizations from the Department of Health and Human Services ([9](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r9)) and reported COVID-19 cases and mortality aggregated by the Johns Hopkins University Center for Systems Science and Engineering ([10](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r10)). To summarize relevant symptom patterns, we defined “COVID-like illness” (CLI) as reporting a fever of at least 100 °F, along with cough, shortness of breath, or difficulty breathing, in line with a working definition of CLI used for syndromic surveillance purposes beginning in early 2020. A second indicator, which we call “CLI in Community,” was based on responses to an item on the survey that asks whether respondents know someone personally in their community who is ill with COVID-like symptoms. We also compared reported diagnoses in CTIS to cumulative diagnoses in surveillance data.

To illustrate the utility of individual-level data to provide detailed information on characteristics that may be relevant to transmission risk, we compared reported diagnoses and reports of working outside of the home while symptomatic across occupational categories. To evaluate time trends in risk exposures and mitigating behaviors, we examined reported contacts, mask use, and use of public transit over time, across a range of stratifying variables, including counties grouped by levels on the Centers for Disease Control and Prevention (CDC) Social Vulnerability Index (SVI) ([11](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r11)), which is a composite measure constructed based on 15 social variables measured at the census tract level, age groups, and Census regions.

Finally, as an illustration of the information value of the survey at various levels of granularity, we assessed reported COVID-19 vaccination intent stratified by individual characteristics and across counties.

The study was approved by the CMU Institutional Review Board, under protocol STUDY2020\_00000162. All respondents gave informed consent before participating in the survey.

**Results**

**Characteristics of the Study Sample.**

As of April 5, 2021, a total of 20.2 million responses had been collected in the US CTIS. [Table 2](https://www.pnas.org/doi/10.1073/pnas.2111454118#t02) summarizes characteristics of the survey respondents. Compared to the weighted sample, the unweighted sample had a higher proportion of women (66% vs. 52%) and a slightly higher proportion of respondents between ages 25 and 64 y (72% vs. 68%). Household size and prevalence of at least one comorbidity were similar in the unweighted and weighted samples.

**Table 2.**

Characteristics of the study sample, compared to 2019 ACS supplemental estimates

|  | **Number** | **Unweighted proportion (%)** | **Weighted proportion (%)** | **Census proportion (%)** |
| --- | --- | --- | --- | --- |
| All responses | 20,249,152 |  |  |  |
| Gender |  |  |  |  |
| Female | 11,409,227 | 66.3 | 52.3 | 50.8 |
| Male | 5,613,674 | 32.6 | 46.2 | 49.2 |
| Nonbinary/self-described | 174,124 | 1.0 | 1.5 | — |
| Age groups |  |  |  |  |
| 18–24 | 1,001,345 | 5.8 | 10.7 | 11.9 |
| 25–34 | 2,856,685 | 16.5 | 16.4 | 17.9 |
| 35–44 | 3,212,187 | 18.5 | 16.5 | 16.4 |
| 45–54 | 3,129,334 | 18.1 | 17.7 | 16.0 |
| 55–64 | 3,337,427 | 19.3 | 17.5 | 16.6 |
| 65–74 | 2,752,379 | 15.9 | 15.3 | 21.2[\*](https://www.pnas.org/doi/10.1073/pnas.2111454118#tfn2) |
| ≥75 | 1,035,551 | 6.0 | 6.0 |
| Education[†](https://www.pnas.org/doi/10.1073/pnas.2111454118#tfn3) |  |  |  |  |
| Less than high school | 226,284 | 3.1 | 4.0 | 12.0 |
| High school or equivalent | 1,152,727 | 16.0 | 17.2 | 27.1 |
| Some college, no degree | 1,744,155 | 24.1 | 24.5 | 20.4 |
| Associate’s degree | 829,618 | 11.5 | 11.2 | 8.5 |
| Bachelor’s degree | 1,765,207 | 24.4 | 23.6 | 19.7 |
| Graduate or professional degree | 1,505,785 | 20.8 | 19.5 | 12.3 |
| Region |  |  |  |  |
| Northeast | 3,426,497 | 17.4 | 17.6 | 17.4 |
| Midwest | 4,791,585 | 24.3 | 20.9 | 20.8 |
| South | 7,328,178 | 37.2 | 37.9 | 38.0 |
| West | 4,160,577 | 21.1 | 23.6 | 23.8 |
| Household size |  |  |  |  |
| 1 | 2,698,400 | 13.7 | 13.3 |  |
| 2 | 6,668,445 | 33.9 | 31.9 |  |
| 3–5 | 8,489,946 | 43.1 | 43.7 |  |
| 6–10 | 1,476,637 | 7.5 | 8.7 |  |
| >10 | 352,267 | 1.8 | 2.4 |  |
| Date of completion |  |  |  |  |
| April–June 2020 | 6,814,488 | 33.7 | 23.6 |  |
| July–September 2020 | 5,280,596 | 26.1 | 25.2 |  |
| October–December 2020 | 3,832,698 | 18.9 | 25.2 |  |
| January–March 2021 | 4,131,800 | 20.4 | 24.7 |  |
| April 2021 | 189,570 | 0.9 | 1.4 |  |
| At least 1 comorbidity | 9,949,181 | 53.7 | 52.0 |  |

EXPAND FOR MORE

\*

Value reported for 65 y and older in ACS.

†

Value reported for adults 25 and older in ACS; only collected in CTIS beginning in Wave 4.

[**OPEN IN VIEWER**](https://www.pnas.org/doi/10.1073/pnas.2111454118#t02)

Compared to 2019 ACS supplemental estimates, the weighted survey sample slightly overrepresented women, but had a broadly comparable age distribution and matched the ACS distribution across geographic regions. The weighted sample included a larger proportion of respondents with greater than a high school education, and a much smaller proportion with less than a high school education, suggesting the presence of a sampling or response bias correlated with education. This bias has remained consistent over time. As the weights provided by Facebook do not account for education, the weighting did not correct this bias.

**COVID-19 Symptoms and Diagnoses.**

A large fraction of daily respondents reported new or unusual symptoms consistent with COVID-19 ([Fig. 1](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig01)). The most common single new or unusual symptom among all respondents was “tiredness or exhaustion,” with a prevalence of 3.9% over Waves 4 through 10. Patterns of symptoms were notably different among the subset of respondents who reported testing positive for COVID-19, compared to all other respondents, including a substantially higher probability of reporting loss of smell or taste (34% compared to 1.2%).

**Fig. 1.**

[A graph of a number of patients

Description automatically generated](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig01)Frequency of reported new or unusual symptoms, pooled over respondents to the CTIS, September 8, 2020 to April 5, 2021. Respondents are grouped by whether they indicated they tested positive in the past 14 d. Dots indicate the ratio of frequency among those who tested positive compared to all others.[**OPEN IN VIEWER**](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig01)

[Fig. 2](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig02) compares time series for three indicators from the CTIS (reported anosmia, CLI, and CLI in community) against the three main surveillance indicators that have been used to monitor trends in the epidemic (confirmed cases, hospitalizations, and deaths) stratified by Census region. Over the period April 6, 2020 to April 5, 2021, the three survey indicators tracked both broad temporal trends and regional patterns in the surveillance indicators, and several notable features are evident in the comparison. First, the survey-based indicators were less susceptible to the daily fluctuations and reporting anomalies that appeared in cases and deaths, including abrupt discontinuities around certain holiday periods. Second, trends and patterns in anosmia were similar to patterns in CLI, and the anosmia series provided a closer match than the other two survey indicators to the trends and patterns observed in COVID-19 hospitalizations, mirroring temporal peaks and comparative levels across regions over different waves of the epidemic. Third, CLI in community provided the most temporally stable signals while also expressing broad differences over time and space that were generally similar to those in other indicators. In a companion paper, we performed extensive correlation analyses between reported cases and various auxiliary indicators, including the survey-based CLI and CLI in community signals ([12](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r12)), showing strong correlations between cases and these two survey signals during much of the pandemic.

**Fig. 2.**

[A graph of different colored lines

Description automatically generated with medium confidence](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig02)Trends in anosmia, CLI, CLI in community, confirmed cases, hospitalizations, and deaths, by Census region, April 6, 2020 to April 5, 2021.[**OPEN IN VIEWER**](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig02)

The CTIS includes questions about testing and diagnosis, which since September 8, 2020 have been asked of all respondents. [Fig. 3](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig03) compares weekly CTIS estimates of the proportion of adults reporting that they have ever had a positive test for COVID-19 against cumulative diagnoses from surveillance reports by state in the same week. State surveillance reports were adjusted, using 2019 population estimates and CDC line-level demographic data on confirmed COVID-19 cases, to produce estimated diagnosis rates among the state’s population over age 18 y. As of April 5, 2021, reported diagnoses in the survey ranged from 3.1% in Hawaii to 19% in Idaho, and the correlation between survey reported diagnoses and surveillance reports at the state level was 0.83, indicating strong convergent validity.

**Fig. 3.**

[A chart of graphing lines

Description automatically generated with medium confidence](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig03)Comparison of proportion of respondents in CTIS reporting ever having tested positive for COVID-19 and cumulative proportion of adult population with confirmed COVID diagnosis, by state, September 8, 2020 through April 5, 2021.[**OPEN IN VIEWER**](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig03)

**Transmission Risk by Individual Characteristics.**

Since Wave 4, the CTIS has included questions about occupation, which can offer valuable insights into exposures among essential workers and also supply signals of where transmission may be concentrated. Using responses from January 2021, we examined the probability of reporting a positive COVID-19 test across different occupation categories, as well as the probability of reporting working outside the home while having symptoms consistent with COVID-19 ([Table 3](https://www.pnas.org/doi/10.1073/pnas.2111454118#t03)). Substantial heterogeneity appeared across broad groups of occupation. The large proportion of people reporting never having been tested indicates the limitations of passive surveillance. Combining questions on symptoms, testing, and working into a single indicator, we examined the fraction of people who reported both working outside the home and currently having atypical symptoms; results ranged from more than 15% for respondents in food preparation and serving related occupations, to 4% of those in arts, design, entertainment, sports, and media.

**Table 3.**

Reporting testing, symptoms, and working outside the home, by reported occupation category, in January 2021

| **Occupation group** | **% Tested positive** | **% Working with symptoms** | **% Working outside and never tested** | ***n*** |
| --- | --- | --- | --- | --- |
| Arts, design, entertainment, sports, and media | 7.2 | 4.0 | 17.1 | 20,585 |
| Building and grounds cleaning and maintenance | 11.9 | 10.8 | 45.8 | 12,528 |
| Community and social service | 13.1 | 9.0 | 23.3 | 28,223 |
| Construction and extraction | 11.8 | 11.3 | 47.7 | 10,528 |
| Education, training, and library | 10.5 | 6.8 | 24.0 | 72,098 |
| Food preparation and serving related | 13.6 | 15.5 | 39.4 | 30,817 |
| Healthcare practitioners and technicians | 15.1 | 9.6 | 25.4 | 70,793 |
| Healthcare support | 15.2 | 9.2 | 22.8 | 45,084 |
| Installation, maintenance, and repair | 10.5 | 10.9 | 49.2 | 15,511 |
| Office and administrative support | 11.0 | 6.2 | 24.4 | 84,285 |
| Other | 9.9 | 6.5 | 25.9 | 165,719 |
| Personal care and service | 12.1 | 9.2 | 32.6 | 15,115 |
| Production | 14.1 | 12.6 | 42.1 | 23,149 |
| Protective service | 14.6 | 11.8 | 33.5 | 8,314 |
| Sales and related | 11.7 | 10.8 | 36.1 | 63,066 |
| Transportation and material moving | 11.2 | 10.5 | 48.3 | 23,013 |

EXPAND FOR MORE

\*

There was no Wave 9 survey. The numbering of waves skipped from 8 to 10 to synchronize numbering conventions with the international version of the survey.

[**OPEN IN VIEWER**](https://www.pnas.org/doi/10.1073/pnas.2111454118#t03)

**Mitigating Behaviors and Policy Analysis.**

A core set of questions included since the launch of the survey has addressed contacts and preventive behaviors. The survey has been amended over time to augment these items, with the addition of questions on mask use and specific high-risk behaviors in September 2020. In the context of recurrent surges in COVID-19 around the country over the course of 2020 and 2021, these items have illuminated how contacts and mitigating behaviors can shift in response to changes in local COVID-19 risk, sometimes preceding policy changes. For example, [Fig. 4](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig04) shows selected variables relating to contacts and preventive behaviors over the period September 2020 to April 2021. Responses indicate sharp increases in risk-reducing behaviors during November and December as cases surged—including reduced contacts, increased use of masks, and reduced use of public transit—followed by relaxation of mitigating behaviors over the period January to April 2021 as cases fell.

**Fig. 4.**

[A graph of different colored lines

Description automatically generated with medium confidence](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig04)Contacts, mask use, and use of public transport, by quartile of counties grouped by the CDC SVI (*Top*), age group (*Middle*), and Census region (*Bottom*), September 8, 2020 to April 5, 2021.[**OPEN IN VIEWER**](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig04)

Individual-level data allow for geographically detailed analysis that can also be disaggregated by demographic features. In [Fig. 4](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig04), results are stratified in three different ways to illustrate this: by quartiles of the CDC SVI, by age, and by US Census region. There were minimal differences across counties grouped by the SVI in reported mask use, moderately higher contacts among those living in more vulnerable communities, and substantially higher use of public transit in more vulnerable counties. The second row shows age differences, which indicate a pronounced gradient of higher risk mitigation among older respondents, especially with respect to reduced contacts. The third row describes regional patterns that vary across indicators, with higher contacts and lower mask use in the South and Midwest regions compared to the Northeast and West, but greater use of public transit in the Northeast and West compared to South and Midwest.

**Vaccination and Vaccine Acceptance.**

Since December 19, 2020, the CTIS has included questions on vaccination intent, and since January 6, 2021, the survey has asked about vaccination status. The combination of geographic and demographic resolution in the survey allows a uniquely detailed view on vaccination acceptance and hesitancy across different United States population groups. [Fig. 5*A*](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig05) displays results by age group, race/ethnicity, gender, and Census region, pointing to high levels of acceptance among older respondents in all categories, but lower and more variable results at younger ages. (Respondents may identify as nonbinary or self-describe their gender, but this group was typically too small to break out and report reliable hesitancy estimates by region.) [Fig. 5*B*](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig05) shows the percentage of respondents indicating that they would probably not or definitely not get vaccinated across United States counties, indicating regional patterns but also high variability across counties within a given state. As the vaccination campaign slows across the country, high-resolution information on vaccine acceptance can inform policies that aim to increase uptake toward the goal of high levels of population immunity against COVID-19.

**Fig. 5.**

[A map of the united states

Description automatically generated](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig05)Reported vaccine acceptance and hesitancy by age group, race/ethnicity, gender, and Census region (*A*) and by county (*B*) during March 2021. Note: Results are pooled over the period March 1, 2021 to April 5, 2021. The map displays results computed for specific counties having at least 50 responses recorded over that period, with all other counties in a state combined into a residual group.[**OPEN IN VIEWER**](https://www.pnas.org/doi/10.1073/pnas.2111454118#fig05)

**Discussion**

As SARS-CoV-2 spread throughout the United States during 2020 and into 2021, and policy makers faced decisions that would profoundly impact all sectors of society, the breadth and depth of information needed to support these decisions vastly exceeded the availability of data collected through existing surveillance systems designed to capture reported COVID-19 cases, hospitalizations, and deaths. A number of efforts to fill the urgent need for additional information relied on novel data collection and dissemination platforms that leveraged mobile phone technology and new media. In this study, we describe one of these efforts, the CTIS, which is the largest continuous health survey ever conducted in the United States, in operation since April 6, 2020, with more than 20 million responses collected over the first year of operation.

Comparisons to routine sources of surveillance information point to both the face validity and incremental value of the CTIS. Time trends and geographic patterns in COVID-19 outcomes measured in the CTIS—including specific symptoms strongly associated with SARS-CoV-2 infection, such as anosmia, syndromic patterns, such as COVID-like-illness, and the novel CTIS measure of CLI in community—mirror broad temporal and spatial features in standard surveillance measures on confirmed COVID-19 cases, hospitalizations, and deaths, while in many cases avoiding data artifacts and reporting anomalies that affect the official measures.

In this study we have highlighted several examples of how attributes of CTIS give it particular value and salience as an information platform for public health policy. The scope, scale, and recruitment strategy used in the survey support analysis at high geographic and temporal resolution, allowing detection of local trends on short timescales not available through other surveys, and accommodate a high level of stratification on relevant individual characteristics. Several examples illustrate the benefits of this granularity, including the ability to compare risks and preventive behaviors by occupational category, with further stratification possible by demography and geographic location; ability to describe variation in intentions; and use of key mitigating measures, including physical distancing, masking, and vaccination. Regular updating of the survey has enabled the survey content to adapt alongside the evolving policy response, for example through addition of survey items on mask use in September 2020, school mitigation strategies in November 2020, and vaccination in December 2020.

Other studies have used data from the CTIS to answer specific questions about key COVID-19 impacts and policies. A number of studies have analyzed relationships between reported risk-mitigating behaviors in the CTIS and other outcomes. For example, Rebeiro et al. ([13](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r13)) examined reported mask-wearing behavior as an outcome in relation to statewide mask-wearing requirements. Rader et al. ([14](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r14)) examined the relationship between mask wearing and physical distancing as measured in CTIS and measures of community transmission. Bilinski et al. ([15](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r15)) assessed trends across states in a number of indicators on risk perception and preventive behaviors in relation to COVID-19 case rates. Other studies have used CTIS measures to explore correlates of variation in risk. For example, Flaxman et al. ([16](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r16)) computed relative infection rates for healthcare workers vs. other respondents using information from the survey on occupation, testing, and test results. Lessler et al. ([17](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r17)) have assessed reported risks of COVID-19–related outcomes, including CLI, anosmia, or a positive COVID-19 test, in relation to whether a household includes a child who attends in-person schooling, and reported school-based mitigation measures.

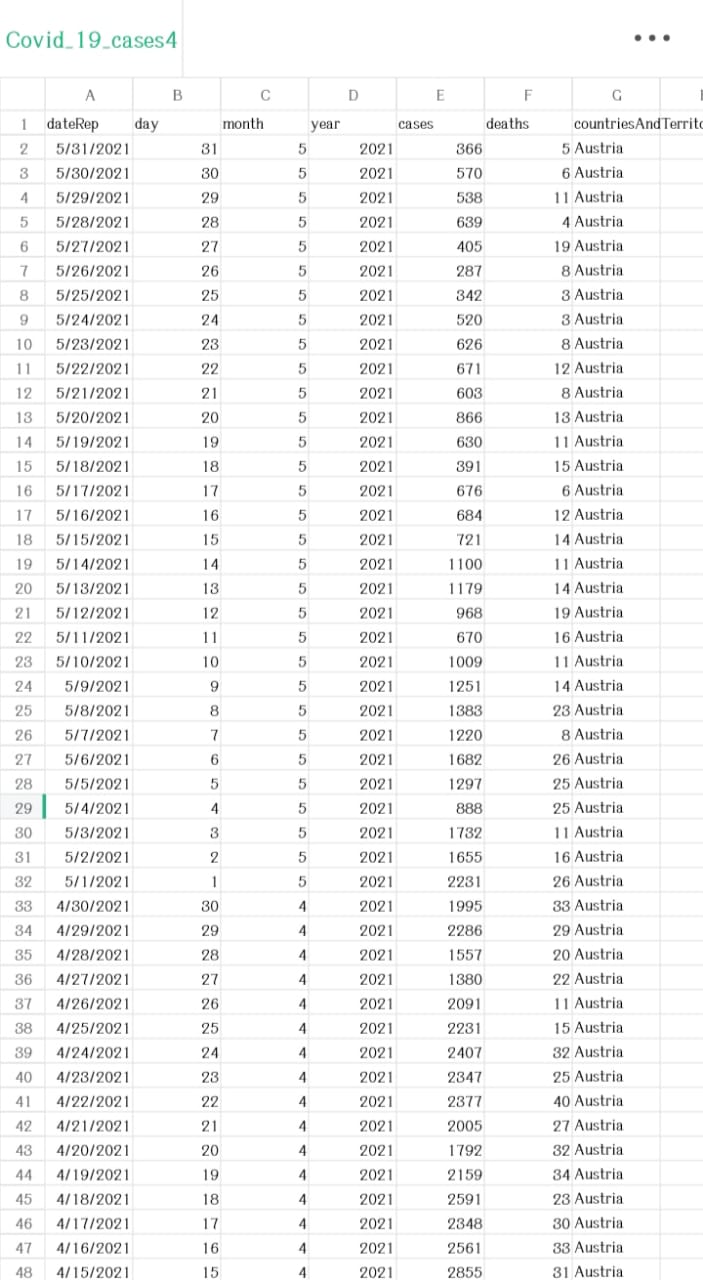
Symptom measures from the survey have also been used to aid in forecasting of COVID cases and deaths. Through the COVID-19 Forecast Hub, the CDC collects standardized forecasts from dozens of teams. Rodríguez et al. ([18](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r18)) incorporated symptom surveillance data from CTIS into a deep-learning framework for real-time forecasting. In a companion paper in this issue ([19](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r19)), we demonstrate that symptom surveillance data and other auxiliary data streams (such as medical insurance claims) can improve forecasting and hotspot prediction accuracy over short (1 to 3 wk) time intervals.

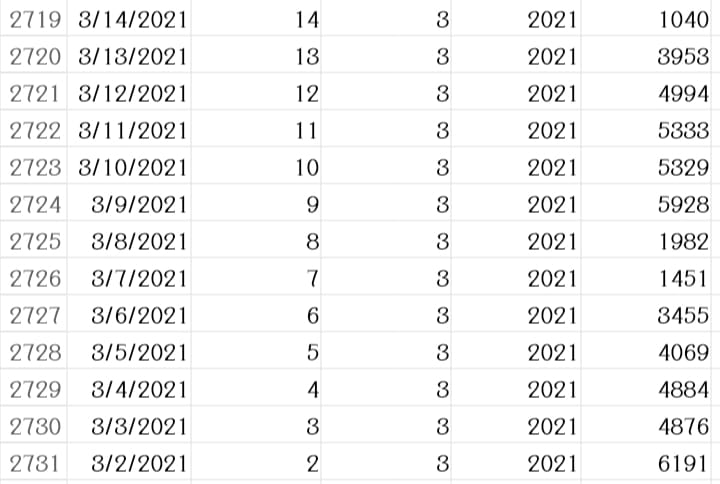
Several limitations are important to note. First, because the survey uses Facebook active users as its sampling frame and because participation in the survey is strictly voluntary, respondents may not be fully representative of the United States population, despite incorporation of survey weights, which adjust for nonresponse and coverage biases based on a limited number of covariates. Comparison to the ACS indicates that our sample overrepresents respondents who are college-educated. Research users of the survey microdata can use additional demographic or other survey variables to construct improved poststratification adjustments to correct this for their purposes. However, any nonresponse biases not accounted for by Facebook’s nonresponse weights would be much more difficult to correct.

Additionally, many of the outcome measures related to COVID-19 are based on self-reports, which may diverge from more objective measures due to recall bias, social desirability bias, and other sources of survey bias and measurement error. On the other hand, broad comparisons of indicators, such as cumulative COVID-19 diagnoses, suggest that measurement of key COVID-19 outcomes are relatively robust to response biases that may be present in the sample. Ultimately, the value of such a large-scale survey is not in accuracy afforded by its sample size, since survey biases persist no matter the size of the survey; smaller surveys, more carefully constructed to reduce sampling biases, would likely yield more accurate estimates ([20](https://www.pnas.org/doi/10.1073/pnas.2111454118#core-r20)). Instead, since these survey biases are unlikely to change rapidly in time or in space, CTIS can accurately track trends in key signals, even if the daily point estimates are systematically biased. This is demonstrated by the strong correlations between survey estimates of CLI in community and reported COVID case rates, for example; while CLI in community is not an unbiased population estimate of COVID case rates, it nonetheless provides useful information about trends in cases. The principal value of CTIS is hence in the detailed spatial and demographic comparisons it makes possible, and in its ability to track changes continuously over time and correlate them with key outcome measures.

Although CTIS was initially designed with a relatively limited scope, including a particular focus on syndromic surveillance, its value has ultimately derived in large part from its flexibility as a surveillance platform that can be rapidly adapted to changing information needs. Running a survey of this size has involved many challenges, particularly as it expanded to include key measures of public knowledge, attitudes, and behaviors, and as public health needs evolved continuously during the pandemic. Despite these challenges, however, CTIS has provided both a valuable public information resource during a global health emergency, as well as a potential model for ongoing health surveillance needs. Similar online surveys are likely to play important roles in future epidemics and pandemics by supplementing public reporting systems with information that is difficult to gather any other way.

**GIVEN DATASET:**





2731 rows x 5 columns

**NECESSARY STEPS TO FOLLOW:**

**1)Import Libraries:**

Begin by importing the necessary libraries. In most cases,need Pandas, a powerful data manipulation library in Python.

**CODE:**

import pandas as pd

**2) Choose the Data Source:**

Determine the source of your dataset. It can be a CSV file, Excel file, SQL database, or any other structured data format.

**3)Load the Dataset:**

Use the appropriate Pandas function to load the dataset into a Pandas DataFrame. For example, if your data is in a CSV file:

# Assuming your file is named "data.csv"

df = pd.read\_csv("data.csv")

For Excel files:

# Assuming your file is named "data.xlsx" and the sheet name is "Sheet1"

df = pd.read\_excel("data.xlsx", sheet\_name="Sheet1")

For SQL databases:

# Assuming you have a SQL connection and your query is stored in the variable "sql\_query"

df = pd.read\_sql\_query(sql\_query, sql\_connection)

**4)Explore the Loaded Data:**

After loading, it's crucial to explore the data briefly to understand its structure and ensure it was loaded correctly. Use functions like head(), info(), and describe()

**CODE:**

print(df.head()) # Displays the first few rows of the DataFrame

print(df.info()) # Provides information about the DataFrame, including data types and null values

print(df.describe()) # Generates summary statistics of the numerical columns.

**5)Handle Missing Data:**

Check for missing data and handle it appropriately. You can remove rows with missing values or fill them with suitable values using dropna() or fillna() functions.

**CODE:**

df.dropna(inplace=True) # Drops rows with missing values

# OR

df.fillna(value, inplace=True) # Fills missing values with a specific value

**6)Data Cleaning and Transformation:**

Depending on your analysis goals, you might need to clean the data by removing outliers, standardizing formats, or transforming variables.

**CODE:**

# Example: Converting a column to datetime format

df['date\_column'] = pd.to\_

datetime(df['date\_column'])

**7)Save Processed Data:**

If you make significant changes to the dataset, consider saving the processed data for future use.

**CODE:**

df.to\_csv("processed\_data.csv", index=False) # Saves the DataFrame to a CSV file without index column

**PREPROCESSING THE DATASET:**

Preprocessing a dataset is a crucial step in data analysis and machine learning. It involves cleaning, transforming, and organizing the data to make it suitable for analysis or modeling. Here are the essential steps for preprocessing a dataset.

**1)Handling Missing Data:**

Identify and handle missing data points. Options include removing rows with missing values or filling missing values with the mean, median, or a specific value.

**CODE:**

df.dropna(inplace=True) # Drops rows with missing values

# OR

df.fillna(value, inplace=True) # Fills missing values with a specific value

**2)Handling Categorical Data:**

Convert categorical variables into numerical representations. This can be done using techniques like one-hot encoding, where each category becomes a binary column.

**CODE:**

df = pd.get\_dummies(df, columns=['categorical\_column']) # One-hot encoding

**3)Data Standardization:**

Standardize or normalize features if they are on different scales. Standardization ensures that the features have the same scale, which is important for many machine learning algorithms.

**CODE:**

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

df[['feature1', 'feature2']] = scaler.fit\_transform(df[['feature1', 'feature2']])

**4)Data Transformation:**

Transform variables if necessary. Common transformations include log transformations for skewed data to make it more normally distributed.

**CODE:**

import numpy as np

df['transformed\_column'] = np.log(df['original\_column'])

**5)Feature Engineering:**

Create new features from existing ones if they can provide valuable information to the analysis or model.

**CODE:**

df['new\_feature'] = df['feature1'] \* df['feature2']

**6)Outlier Detection and Removal:**

Identify and handle outliers if they exist in the dataset. Outliers can significantly impact the results of data analysis and machine learning models.

**CODE:**

from scipy import stats

z\_scores = np.abs(stats.zscore(df['numeric\_column']))

df\_no\_outliers = df[(z\_scores < 3)] # Keeps only the data points within 3 standard deviations

**7)Data Splitting (for Machine Learning):**

If preparing the dataset for machine learning, split it into features (X) and target variable (y). This is necessary for supervised learning tasks.

**CODE:**

X = df.drop('target\_column', axis=1) # Features

y = df['target\_column'] # Target variable

**8) Feature Scaling (for Machine Learning):**

Scale the features to ensure that all features contribute equally to the analysis or model. Common techniques include Min-Max scaling or Standardization.

**CODE:**

from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

X\_scaled = scaler.fit\_transform(X)

**IMPORTANCE OF LOADING AND PREPROCESSING THE DATASET:**

**1. Data Quality Assurance:**

Loading a dataset allows you to assess its quality. By inspecting the loaded data, you can identify missing values, inconsistencies, and errors. Preprocessing steps such as handling missing data and outliers enhance the quality and reliability of the dataset.

**2. Data Understanding:**

Loading the dataset provides an initial understanding of its structure, variables, and format. This understanding is essential for making informed decisions during preprocessing. It helps you identify the types of transformations and cleaning operations required for meaningful analysis.

**3. Ensuring Compatibility:**

Preprocessing ensures that the data is in a format compatible with the analysis or machine learning algorithms you plan to use. Different algorithms have different requirements; preprocessing helps in transforming data to meet these requirements, improving the performance and accuracy of the models.

**4. Enhancing Model Performance:**

For machine learning tasks, preprocessing steps like feature scaling and handling categorical data are vital. Scaling ensures that features are on a similar scale, preventing certain features from dominating the model. Handling categorical data appropriately ensures that machine learning algorithms can interpret these variables correctly, leading to more accurate predictions.

**5. Handling Missing Data:**

Real-world datasets often have missing values. Preprocessing techniques such as imputation (filling missing values) or removing incomplete records ensure that analyses and models are based on complete information, preventing biased or skewed results.

**6. Improving Interpretability:**

Well-preprocessed data is easier to interpret. Clear, consistent, and transformed data allows for better visualization and understanding of patterns and trends, leading to more meaningful insights.

**7. Reducing Computational Costs:**

Preprocessing, including techniques like dimensionality reduction, can significantly reduce the computational resources required for analysis and modeling. This is particularly important when dealing with large datasets, making the process more efficient.

**8. Facilitating Collaboration:**

When datasets are properly loaded and preprocessed, they become easier to share and collaborate on. Clean, well-documented data ensures that other team members or researchers can understand and work with the data effectively.

**CHALLENGES INVOLVED IN LOADING AND PREPROCESSING DATASET:**

Loading and preprocessing datasets, especially large and real-world datasets, come with several challenges. Here are some common challenges faced in these processes:

**1)Data Inconsistency:**

Datasets often come from multiple sources, leading to inconsistencies in data formats, units, and naming conventions. Handling these inconsistencies is challenging, as standardizing the data is crucial for meaningful analysis.

**2)Missing Data:**

Real-world data frequently contains missing values, which need to be appropriately handled. Deciding whether to remove incomplete records or fill in missing values without introducing bias is a challenge in preprocessing.

**3)Noisy Data:**

Noisy data includes irrelevant or erroneous information. Identifying and filtering out noise is essential, as it can significantly impact the quality of analysis and modeling.

**4)Large Volume of Data:**

Big data challenges include processing and managing vast volumes of data efficiently. Loading and preprocessing large datasets require powerful computing resources and specialized algorithms to handle the volume.

**5)Data Security and Privacy:**

Ensuring data privacy and security is crucial, especially when working with sensitive data. Compliance with regulations such as GDPR adds complexity to handling and preprocessing datasets, requiring careful handling of personally identifiable information (PII).

**6)Handling Categorical Data:**

Categorical variables need to be transformed into numerical representations for analysis. Deciding on the appropriate encoding technique, especially for variables with many categories, can be challenging.

**7)Feature Engineering:**

Creating new meaningful features from existing ones requires domain knowledge. Deciding which features to engineer and how to combine them effectively can be a challenge.

**8)Scalability:**

Scalability challenges arise when processing datasets that grow over time. Preprocessing methods need to be scalable to handle the increasing volume of data efficiently.

**9)Time Complexity:**

Real-time data processing requires algorithms and techniques that can handle data streams efficiently. Traditional preprocessing methods might not be suitable for real-time applications.

**10)Data Imbalance:**

In classification tasks, datasets may be imbalanced, where some classes have significantly fewer instances than others. Handling this imbalance during preprocessing is crucial to prevent biases in the analysis or modeling results.

**11)Versioning and Reproducibility:**

Ensuring version control of the preprocessing steps is vital for reproducibility. Changes in preprocessing techniques or parameters can significantly impact the results, making it essential to document and version preprocessing procedures.

**LOADING THE DATASET:**

**import pandas as pd**

**# Replace 'your\_dataset.csv' with the actual file path of your CSV dataset**

**file\_path = 'https://www.kaggle.com/datasets/chakradharmattapalli/covid-19-cases'**

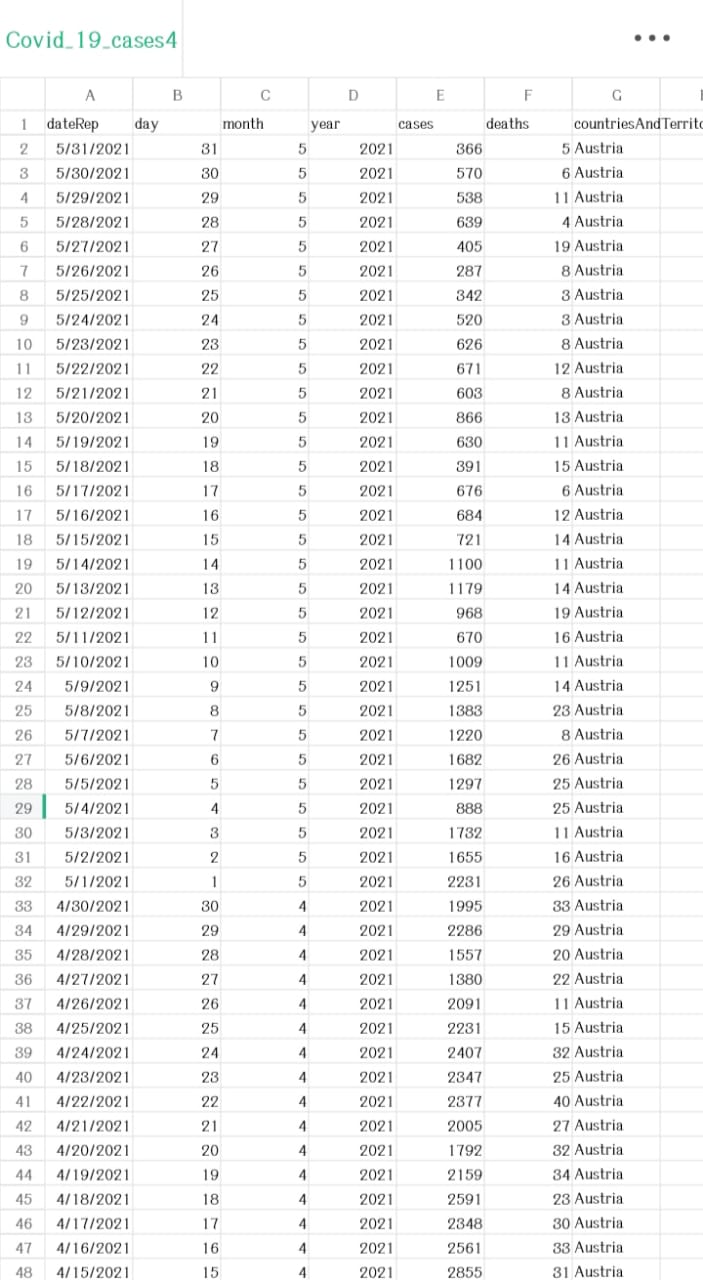
**# Load the dataset into a Pandas DataFrame**

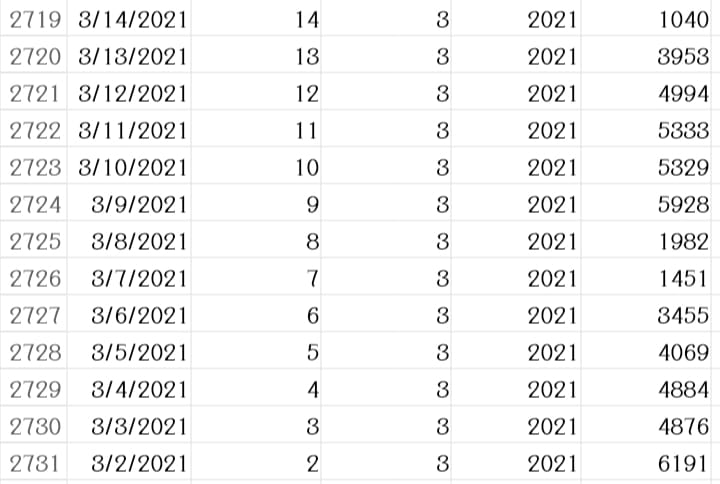
**df = pd.read\_csv(https://www.kaggle.com/datasets/chakradharmattapalli/covid-19-cases)**

**# Now 'df' contains your dataset and you can perform various operations on it**

**print(df.head()) # Display the first few rows of the loaded dataset**

**OUTPUT:**

****

****

**PREPROCESSING THE DATASET:**

* Data preprocessing is the process of cleaning, transforming, and integrating data in order to make it ready for analysis.
* This may involve removing error and inconsistencies, handling missing values, transforming the data into a consistent format, and scaling the data to a suitable rows and columns.
* #Import necessary libraries

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler, LabelEncoder

# Load your COVID-19 dataset into a Pandas DataFrame

data = pd.read\_csv(‘https://www.kaggle.com/datasets/chakradharmattapalli/covid-19-cases')

# Handle missing values (if any)

data = data.dropna()

# Encode categorical variables (if any)

label\_encoders = {}

categorical\_columns = ['column1', 'column2'] # Specify the categorical columns in your dataset

for column in categorical\_columns:

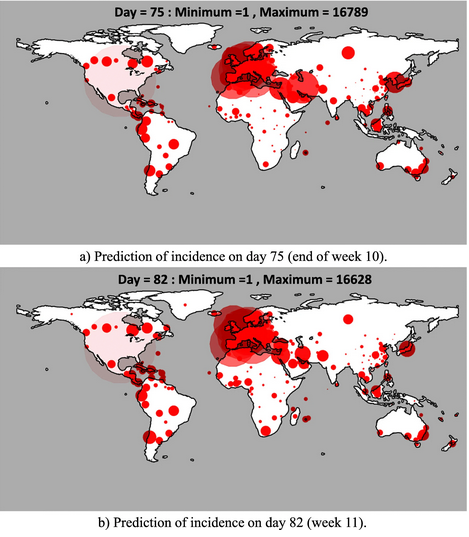
label\_encoders[column] = LabelEncoder()

data[column] = label\_encoders[column].fit\_transform(data[column])

# Split the data into features (X) and target variable (y)

X = data.drop(columns=['target\_column']) # Replace 'target\_column' with the name of your target variable

y = data['target\_column']

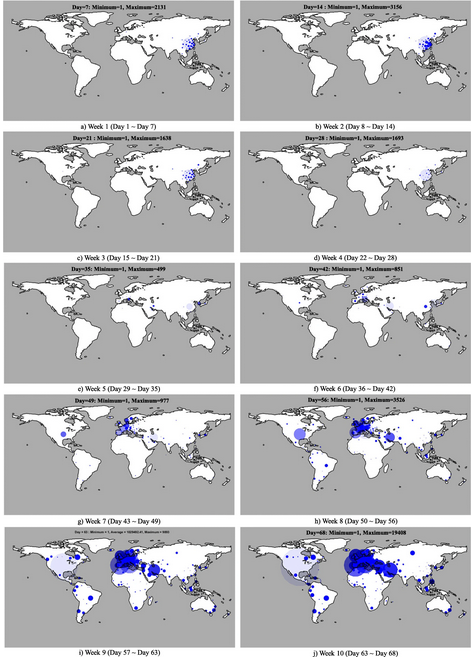


# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

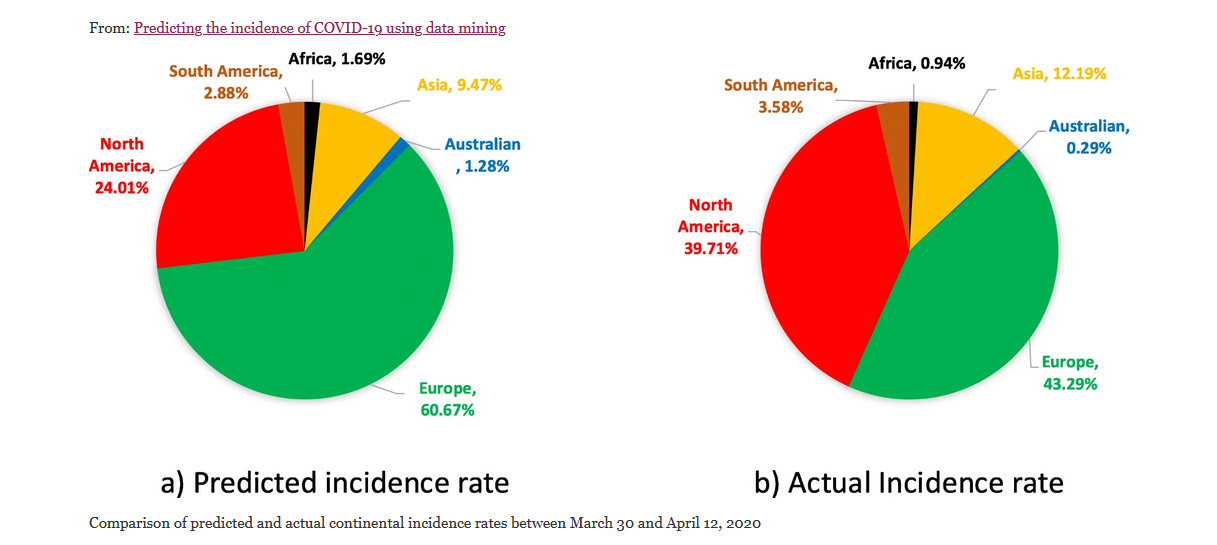
# Standardize features by removing the mean and scaling to unit variance

scaler = StandardScaler()



X\_train\_scaled = scaler.fit\_transform(X\_train)

X\_test\_scaled = scaler.transform(X\_test)



import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Load your preprocessed COVID-19 dataset into a Pandas DataFrame

# (Assuming you already have a preprocessed DataFrame named 'data')

# data = pd.read\_csv(‘https://www.kaggle.com/datasets/chakradharmattapalli/covid-19-cases’)

sns.lineplot(x='Date', y='Cases', data=data)

plt.xlabel('Date')

plt.ylabel('Number of Cases')

plt.title('Daily COVID-19 Cases Over Time')

plt.xtic')

Bar plot - COVID-19 cases by country

plt.figure(figsize=(12, 6))

sns.barplot(x='Country', y='Cases', data=data)

plt.xticks(rotation=90)

plt.xlabel('Country')

plt.ylabel('Number of Cases')

plt.title('COVID-19 Cases by Country')

plt.show()

Line plot - Daily cases over time

plt.figure(figsize=(12, 6))

sns.lineplot(x='Date', y='Cases', data=data)

plt.xlabel('Date')

plt.ylabel('Number of Cases')

plt.title('Daily COVID-19 Cases Over Time')

plt.xticks(rotation=45)

plt.show()

Box plot - Distribution of cases by continent

plt.figure(figsize=(10, 6))

sns.boxplot(x='Continent', y='Cases', data=data)

plt.xlabel('Continent')

plt.ylabel('Number of Cases')

plt.title('Distribution of COVID-19 Cases by Continent')

plt.show()

Heatmap - Correlation between numerical variables

correlation\_matrix = data.corr()

plt.figure(figsize=(10, 8))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm')

plt.title('Correlation Heatmap')

plt.show()

Pairplot - Relationship between multiple variables

sns.pairplot(data[['Cases', 'Deaths', 'Recovered', 'Population']])

plt.suptitle('Pairplot of COVID-19 Data', y=1.02)

plt.show()

**VISUALIZATIONS:**

PLOTTING SIMPLE PLOT:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

data = pd.read\_csv('case\_time\_series.csv')

Y = data.iloc[61:,1].values

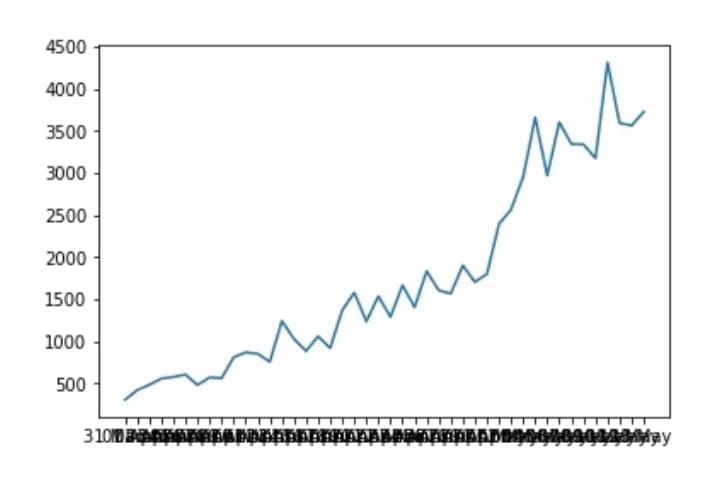
R = data.iloc[61:,3].values

D = data.iloc[61:,5].values

X = data.iloc[61:,0]

plt.plot(X,Y)

**OUTPUT:**

****

**import numpy as np**

**import pandas as pd**

**import matplotlib.pyplot as plt**

**data = pd.read\_csv('case\_time\_series.csv')**

**Y = data.iloc[61:,1].values**

**R = data.iloc[61:,3].values**

**D = data.iloc[61:,5].values**

**X = data.iloc[61:,0]**

**plt.figure(figsize=(25,8))**

**ax = plt.axes()**

**ax.grid(linewidth=0.4, color='#8f8f8f')**

**ax.set\_facecolor("black")**

**ax.set\_xlabel('\nDate',size=25,color='#4bb4f2')**

**ax.set\_ylabel('Number of Confirmed Cases\n',**

**size=25,color='#4bb4f2')**

**ax.plot(X,Y,**

**color='#1F77B4',**

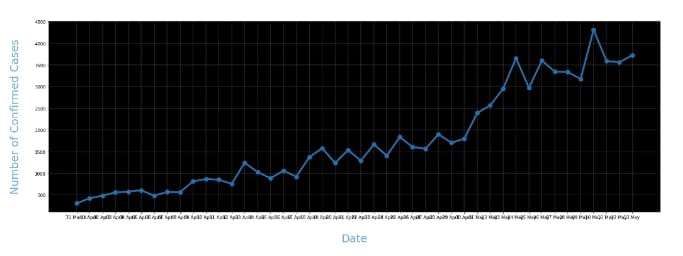
**marker='o',**

**linewidth=4,**

**markersize=15,**

**markeredgecolor='#035E9B**

**OUTPUT:**

****

**import numpy as np**

**import pandas as pd**

**import matplotlib.pyplot as plt**

**data = pd.read\_csv('case\_time\_series.csv')**

**Y = data.iloc[61:,1].values**

**R = data.iloc[61:,3].values**

**D = data.iloc[61:,5].values**

**X = data.iloc[61:,0]**

**plt.figure(figsize=(25,8))**

**ax = plt.axes()**

**ax.grid(linewidth=0.4, color='#8f8f8f')**

**ax.set\_facecolor("black")**

**ax.set\_xlabel('\nDate',size=25,color='#4bb4f2')**

**ax.set\_ylabel('Number of Confirmed Cases\n',**

**size=25,color='#4bb4f2')**

**plt.xticks(rotation='vertical',size='20',color='white')**

**plt.yticks(size=20,color='white')**

**plt.tick\_params(size=20,color='white')**

**for i,j in zip(X,Y):**

**ax.annotate(str(j),xy=(i,j+100),color='white',size='13')**

**ax.annotate('Second Lockdown 15th April',**

**xy=(15.2, 860),**

**xytext=(19.9,500),**

**color='white',**

**size='25',**

**arrowprops=dict(color='white',**

**linewidth=0.025))**

**plt.title("COVID-19 IN : Daily Confirmed\n",**

**size=50,color='#28a9ff')**

**ax.plot(X,Y,**

**color='#1F77B4',**

**marker='o',**

**linewidth=4,**

**markersize=15,**

**markeredgecolor='#035E9B**

**OUTPUT:**

****

**data = pd.read\_csv('district.csv')**

**data.head()**

**re=data.iloc[:30,5].values**

**de=data.iloc[:30,4].values**

**co=data.iloc[:30,3].values**

**x=list(data.iloc[:30,0])**

**plt.figure(figsize=(25,10))**

**ax=plt.axes()**

**ax.set\_facecolor('black')**

**ax.grid(linewidth=0.4, color='#8f8f8f')**

**plt.xticks(rotation='vertical',**

**size='20',**

**color='white')#ticks of X**

**plt.yticks(size='20',color='white')**

**ax.set\_xlabel('\nDistrict',size=25,**

**color='#4bb4f2')**

**ax.set\_ylabel('No. of cases\n',size=25,**

**color='#4bb4f2')**

**plt.tick\_params(size=20,color='white')**

**ax.set\_title('Maharashtra District wise breakdown\n',**

**size=50,color='#28a9ff')**

**plt.bar(x,co,label='re')**

**plt.bar(x,re,label='re',color='green')**

**plt.bar(x,de,label='re',color='red')**

**for i,j in zip(x,co):**

**ax.annotate(str(int(j)),**

**xy=(i,j+3),**

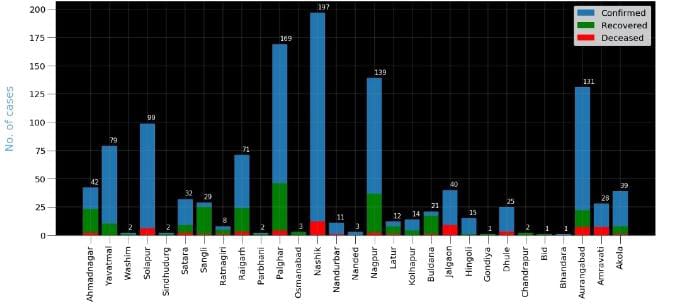
**color='white',**

**size='15')**

**plt.legend(['Confirmed','Recovered','Deceased'],**

**fontsize=20)**

**OUTPUT:**

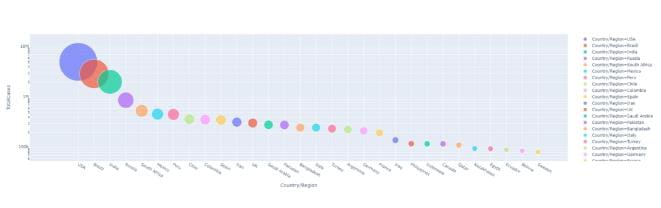
****

**px.scatter(dataset1.head(30), x='Country/Region', y='TotalCases',**

**hover\_data=['Country/Region', 'Continent'],**

**color='Country/Region**

**OUTPUT:**

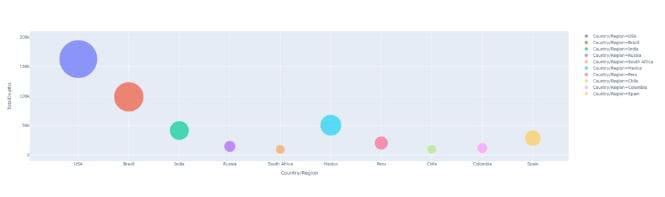
****

**px.scatter(dataset1.head(10), x='Country/Region', y= 'TotalDeaths',**

**hover\_data=['Country/Region', 'Continent'],**

**color='Country/Region', size= 'TotalDeaths', size\_max=80)**

**OUTPUT:**

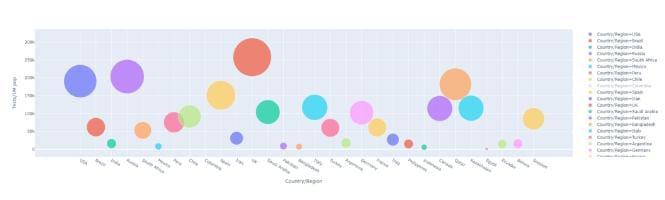
****

**px.scatter(dataset1.head(30), x='Country/Region', y= 'Tests/1M pop',**

**hover\_data=['Country/Region', 'Continent'],**

**color='Country/Region', size= 'Tests/1M pop', size\_max=80)**

**OUTPUT:**

****

**CONCLUSION OF PHASE 4 PROJECT:**

In the phase 4 conclusion of development part 2, we performed the covid 19 cases analysis and create visualizations, death cases , recovered, decreased, maximum causes. then we created some visualizations using data visualization libraries(e.g., matplotlib, seaborn) for **covid 19 cases analysis** by using the given dataset.