A POCKET GUIDE FROM
DATA DOZEN
a visual analytics studio

WORKING WITH COVID-19 DATA:

THE DECISIONS BEHIND THE DASHBOARDS

A PRIMER

The COVID-19 pandemic has inundated the public with data and visualizations, but few folks know how these resources are actually built. Here's a peek into one analytic process that reveals just how many distinct decisions underpin this kind of data work.

In October 2020, Data Dozen launched its own series of interactive dashboards built with The New York TimesCOVID-19 data. This pocket guide is meant to accompany those dashboards and provide a window into how they were created. Each visualization is the product of numerous analytic and ethical judgment calls made during the building process. And of course, these are just the latest layer of innumerable decisions that have accumulated in the underlying data itself. This pocket guide is by no means a comprehensive collection of all the nuances or caveats that can be found in this data set or these visualizations. Instead, this primer strives to offer a helpful sampling of the types of intricacies and ethical decisions that underpin this kind of analytic work. The goal is to help demystify the building process behind such public resources and highlight the complexity that both tabular data and polished dashboards so often flatten.

A POCKET GUIDE FROM

DATA DOZEN

a visual analytics studio

DATE

def: The day new cases or deaths were announced.

INCONSISTENT SCHEDULES

Many counties report new data each day, but others maintain sparser schedules such as a Monday-Wednesday-Friday release pattern. In that case, the county has 0 cases or deaths reported for every Tuesday, Thursday, Saturday, and Sunday row in the data set.

CASES

def: The running total of reported COVID-19 cases. To isolate a single day's new case count, subtract the previous day's running total from that day's running total.

UNDERCOUNTING

These counts only include reported cases, which means that these numbers will be consistently lower than the actual (but unknowable) number of total COVID-19 cases in the United States. Folks who were sick or asymptomatic but were never tested or evaluated by a physician have no way of showing up in the data.

PROBABLE VS. CONFIRMED CASES

On April 5, 2020, the Council of State & Territorial Epidemiologists recommended that public health officials report probable cases in addition to laboratory confirmed cases. Probable cases meet specific benchmarks for clinical evidence but have not received a laboratory test. The CDC adopted this recommendation in mid-April, but many states were already reporting probable cases. To date, some states report confirmed and probable cases separately. (The New York Times combines these numbers.) Other states report their confirmed and probable cases together. And some states still only report laboratory confirmed cases.

DEATHS

def: The running total of reported COVID-19 deaths. To isolate a single day's new death count, subtract the previous day's running total from that day's running total.

OUTLIER SPIKES

Single day spikes in the number of reported cases or deaths can be attributed to many different events. They could be an influx of previously unreported cases or deaths, as was the case in New York City in April 2020. Case spikes have also been explained by comprehensive testing of an outbreak that has occurred in a concentrated community, such as a nursing home, prison, or school.

NEGATIVE COUNTS

As they receive new information, public health officials revise their counts. Some of these revisions are clear cut. For example, when it is confirmed that a COVID-19 positive patient died due to an independent cause such as an accident, suicide, or homicide, the count is revised by subtracting a death from the next day's running total. Other revisions are less clear.

STATE

def: The name of the US state, territory, or D.C. where the cases and deaths were announced.

RESIDENCE VS. TREATMENT

In the context of communicable diseases, where a patient was treated might be more relevant than their residential address. When the applicable information is available, The New York Times counts cases and deaths based on where patients are being treated. They provide an example of an early death of a Florida resident who was treated and died in Los Angeles. This individual shows up in California's total death count, not Florida's. These types of discrepancies introduce small degrees of error in our analysis when we control the number of cases or deaths by a state or county's population.

COUNTY

def: The name of the county or county equivalent (parish, borough, municipality, etc.) where the cases and deaths were announced.

PATCHWORK METHODOLOGY

There are well over 3,000 unique counties in our 50 states, 5 territories, and D.C. Each of these counties answers to different local and state governments, and each has experienced different pressures throughout this pandemic. Variance in methodology between counties is inevitable, which has often left The New York Times journalists with the responsibility of interpreting these differences as their team attempts to create a data set that is as consistent and meaningful as possible.

GEOGRAPHIC ANOMALIES

All of the counts are reported at the county level with three notable exceptions. New York City reports the data for their five counties as a single unit. Both Kansas City, Missouri and Joplin, Missouri report their data as a city municipality instead of by their local counties. This means that the separate counts for the overlapping counties only include counts for individuals living outside the city's municipal limits.

UNKNOWN & MISSING COUNTIES

When a patient's county is unknown or pending, their case or death is attributed to a particular state under the category of "Unknown County." This often revised later when more is known. If a U.S. county has never reported a case or a death, the

FIPS NUMBER

def: The 5-digit Federal Information Processing Standard number that corresponds to the county or county equivalent's name. This is blank for unknown counties and the three geographic anomalies that are not reporting at a county level.

> The New York Times' thorough documentation is a testament to Lisa Gitelman's keen observation that "Raw data is an oxymoron."

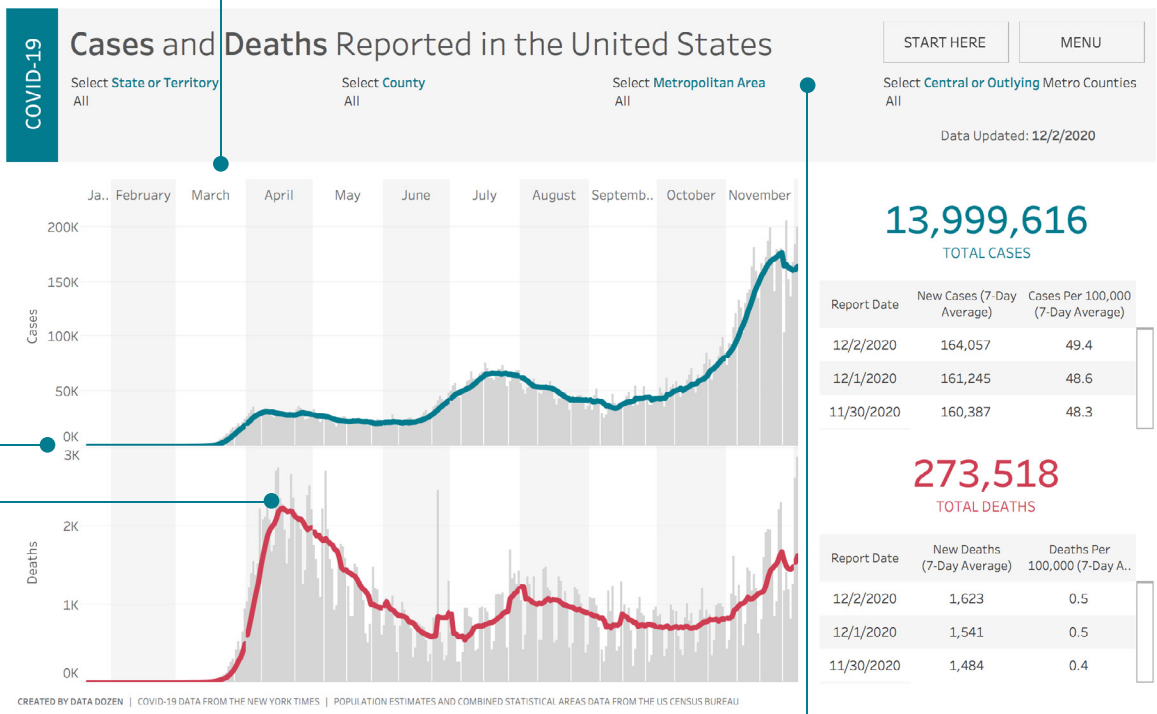
Every data set is the accumulation of innumerable judgment calls made by the humans who created it. When we build our work on top of any data set, we inherit all of these decisions, good or bad, and then proceed to add our own. The New York Times county-level COVID-19 data set has just six variables, but behind this tidy collection of columns and rows lies an abundance of nuance. Here are just a few of the caveats baked into each field.

STARTING AT ZERO

The y-axis is fixed to start at 0. This decision creates a crisp horizontal line across the bottom of both the cases and deaths bar chart, a pattern maintained when the audience filters to a subset of the data, like a single state or county. This makes reading the charts very simple, but glosses over the complexity introduced by negative counts discussed on page 2. When sizable corrections are issued and a day's case or death count is negative, the fixed y-axis prevents any negative grey bar from appearing on the chart (although the 7-day average line does adjust accordingly). Ultimately this decision prioritizes legibility and consistency at the risk of presenting the data as more fixed than it is.

LONGITUDINAL PERSPECTIVE

The x-axis starts on January 21, 2020 when the United States reported its first COVID-19 case. Showing the pandemic's US timeline in its entirety was an intentional decision. Instead of focusing on just the last few weeks, the goal here was to acknowledge the full duration of the pandemic. By lining up the months with their corresponding grey bands along the x-axis, we're inviting the audience to reflect on how they might have felt at different points in time. Rarely is our emotional response directly proportional to the objective levels of risk we're navigating. Seeing the data framed by month encourages such an internal dialogue.



7 DAY AVERAGE

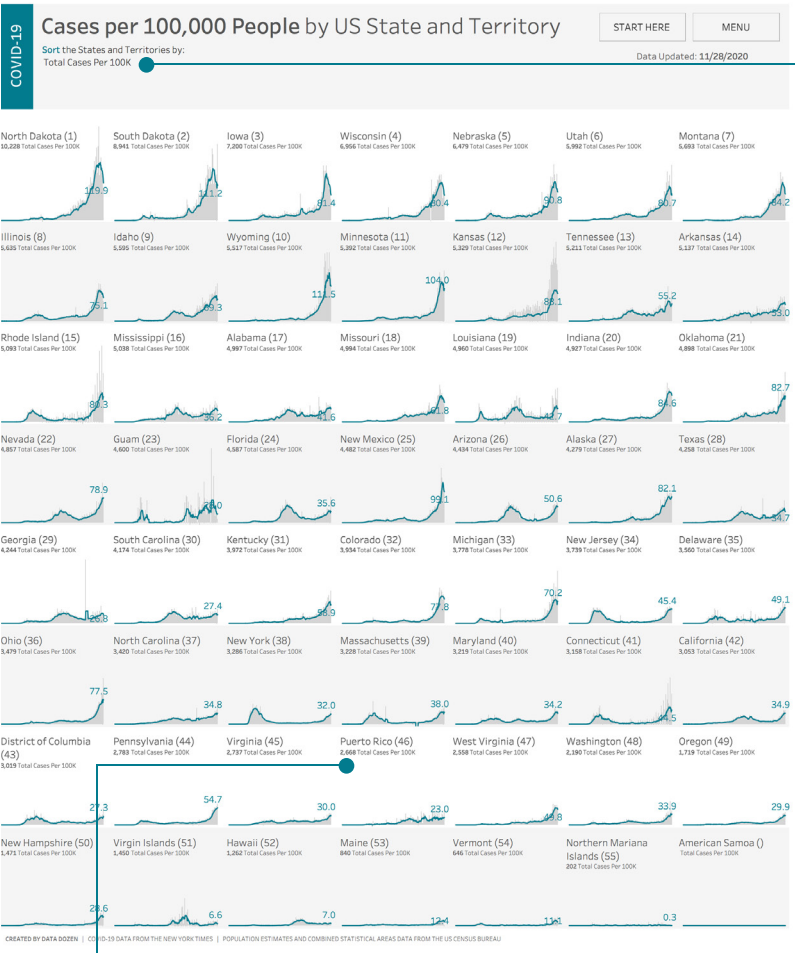
As we can see in the grey bars, there's a predictable cadence of peaks and valleys each week caused by reporting processes. In order to provide an accurate representation of the counts despite this procedural choppiness, we added a 7-day average represented by the teal and red lines. These lines add up the reported cases or deaths from the previous 7 days, and divides them by 7 to create a steady indicator of whether the numbers are actually going up or down. A 7-day average allows us to compensate for a full week's predictable dips (e.g., we always have a Sunday in our average) without diluting the immediacy like a 14-day average could.

METROPOLITAN AREAS

Many cities in the United States span multiple counties and some, like Data Dozen's hometown of Kansas City, span multiple states. However, few folks can quickly name all of the counties that would be relevant to the local prevalence of COVID-19. With this in mind, we added to the data the 172 Combined Statistical Areas (CSAs) delineated by the United State Office of Management and Budget. These CSAs group individual counties that are economically or socially connected (employment interchange of at least 15%). The goal was to enable any metropolitan resident to quickly filter the dashboard down to just their city's data. However, such a strategy still doesn't provide a comprehensive picture of cities that span multiple countries, such as the San Diego-Tijuana metropolitan area. This also privileges urban audiences by providing filters that directly address their data needs, while similar considerations were not taken for rural audiences.

> Transforming rows and columns into charts and graphs is an act of curation. As we build each dashboard, we decide what to emphasize, what to exclude, and how it will feel for our audience to walk through this data.

Despite all its caveats, this data set is an immensely useful resource. (Imagine the alternative: what would our personal, local, or national conversations look like if we had no comprehensive data set capturing our nation's cases and deaths?) With the creation and maintenance of such a powerful resource, the next question becomes how to disseminate its knowledge. Data dashboards are more intuitive and approachable communication tools than data tables, but as we create them, we inevitably frame which pieces of the data our audience sees. Here are just a few examples of the judgment calls behind the first dashboard in this series.



CONTROLLING FOR POPULATION

Decisions pertaining to the data's structure early on in a project often determine who shows up (and who doesn't) in the final visualizations. To create meaningful comparisons across states here, we needed to standardize the data by population. The inclusion of each county's unique FIPS number by The New York Times enables analysts to extend the functionality of the data by lining up each county's COVID-19 counts with its population, demographics, or other county-level information. However, if a county has never reported a COVID-19 case, that county does not appear in The New York Times data set. This means that when we attempt to tack on the county's population, it doesn't find a matching row in the COVID-19 data, and those residents don't get added to the state's total population. To compensate for this, we added a complete list of counties to the end of The New York Times data set to ensure that when we built visualizations like this one, we had a complete population for each state.

QUANTIFYING SUFFERING

With a comparable chart for each state, the question becomes how to order them. By including a dropdown menu that changes how the states are sorted, we're enabling our audience to see how the story changes depending on the ranking method they choose. For example, as of December 1, 2020, when ranked by the number of cases per 100,000 residents, North Dakota is first; by deaths per 100,000 residents, New Jersey is first; by total cases, Texas is first; and by total deaths, New York is first. Fundamentally, this is a question about how to quantify and compare suffering. How do we identify disproportional outbreaks while also equally respecting each individual life affected? As we prioritize one goal, we often infringe on the other. When we rank the states by cases per 100,000 residents, which is our default here, we can quickly spot the most disproportionate outbreaks. But given their small populations, we end up potentially over emphasizing the suffering in places like North and South Dakota despite the fact that the number of people affected there is but a fraction of those affected in large states like Texas, California, Florida, and New York.

TERRITORIES

This dashboard uses a visualization technique called small multiples, which creates an identical chart for each category being compared. To build this chart type, we start with a seemingly simple question: what are all the categories? Your immediate response probably depends on where you live. If you live in the contiguous United States, 50 categories probably come to mind. If you live in Washington D.C., you're probably envisioning 51. And if you live in American Samoa, Guam, the Northern Mariana Islands, the U.S. Virgin Islands, or Puerto Rico (where the population exceeds that of 21 other American states), you're probably envisioning 55 or 56 categories, depending on whether the capital crossed your mind. This dashboard made the intentional decision to give equal weight to all of the states, territories, and D.C. by giving each its own chart with the same physical real estate on the canvas. During this global pandemic, if the analysts in the United States don't include the cases and deaths in its territories, no country will.

> Comparison is one of the most natural ways to explore a data set. The onus is on the analyst to ensure our visualizations compare apples to apples, and in turn, to question what gets lost in our methods of standardization.

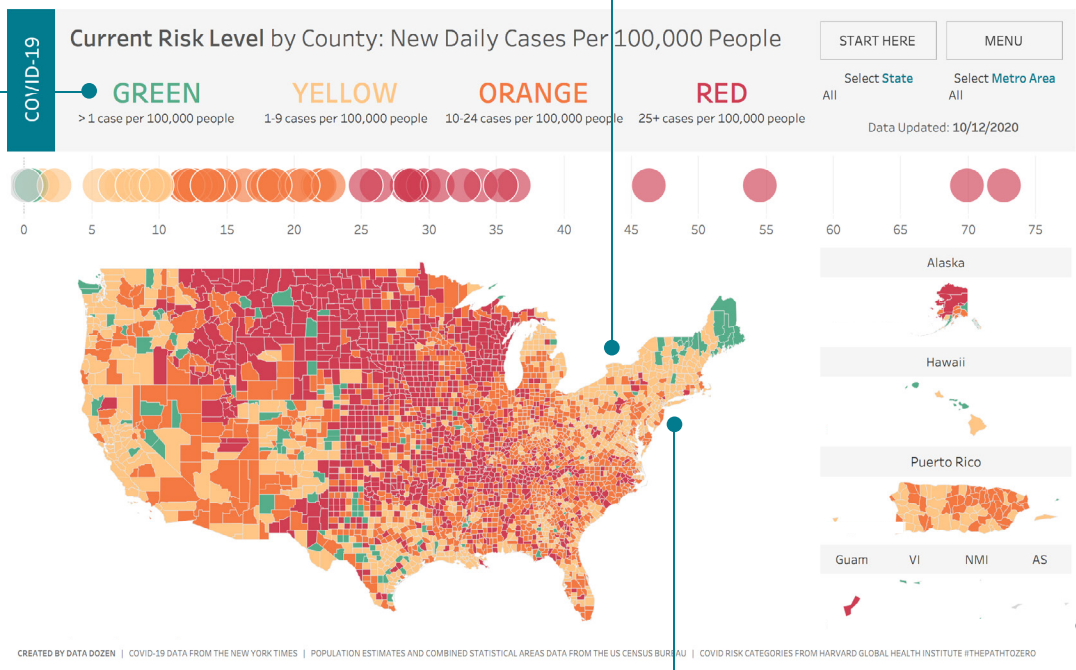
State breakdowns, like the dashboard above, have become commonplace in COVID-19 analysis. In the United States, the delineation by state is so ingrained in our reality that such comparisons feel approachable and informative. However, state populations vary enormously. To compare the number of cases by state is mostly just to compare each state's population. It tells us very little about the severity or velocity of each state's outbreak. In order to get a meaningful comparison, we need a standard number we can calculate for each state. Here we examined the number of cases per 100,000 residents. Then the question becomes what pieces of the story have we obscured by prioritizing relativity.

RISK LEVEL CATEGORIES

These risk level categories were designed by a collaborative team of public health and policy researchers led by the Harvard Global Health Institute. Their risk level categories offer a framework built on epidemiological expertise that's also easy for non-experts to understand. They accompany each color coded risk level with a short description (on track for containment, community spread, accelerated spread, tipping point) and policy advice (ranging from monitor with testing and contact tracing to stay-at-home orders necessary). On the dashboard, we've included these descriptions in a popup box that appears when the audience hovers over the different categories in the color legend. As analysts, we are constantly leveraging content expertise to understand and contextualize the data we visualize. This is undeniably necessary. After all, we can't be content experts in everything we ever visualize. But we must think carefully about which content experts we've decided to trust, and consequently, whose biases, blindspots, or errors we might unintentionally amplify as we incorporate their methodology into our own visualizations.

COUNTY-LEVEL GRANULARITY

With each visualization we create, we decide how to aggregate the data. In the first dashboard (page 3), we show the data for the entire nation. In the second dashboard (page 4), we aggregate the data by state. In this third dashboard, we've picked the finest level of detail this data set has to offer, breaking out the data by each county. While the national and state level data can be very helpful in a variety of situations, the county-level granularity most closely contextualizes what individual citizens are experiencing. For example, as a resident of Buffalo, New York, the rise and fall of cases in Erie County is much more applicable to your daily decisions than the numbers for the state of New York or the US as a whole.



GEOGRAPHIC ANOMALIES

As mentioned on page 2, there are three geographic anomalies that do not report their data at the county level: New York City, Kansas City, Missouri, and Joplin, Missouri. In order to display this data on a county-level map like this, several judgment calls were made. First, the five counties that make up New York City were grouped together into a single polygon on the map, mirroring the joint way in which the data is reported. However, the solution for Kansas City and Joplin was not as simple. These municipalities span multiple counties, which are also reporting their own counts that exclude the city's counts. To visualize this data on a county-level map, we grouped all of the cases for the city into the county where the majority of the city belongs (Jackson County for Kansas City and Jasper County for Joplin). We noted this decision in a popup box that appears when our audience hovers over these counties on the dashboard, striving to embed our documentation into the visualization itself.

> Rarely is a data visualization specialist also an expert on the content they are visualizing. Like journalists, they rely on domain experts to help unpack and contextualize the information they're sharing with a broader audience.

A data analyst has the skills to structure a data set for analysis and identify internal patterns, outliers, and relationships within the data. But that doesn't guarantee they know what those data points actually mean in the real world. For example, this county-level dashboard intends to tease apart geographical variation that gets lost when we aggregate the data by state. This allows us to see where cases are clustering and compare whether individual counties have more or less cases than their neighbors. But it doesn't tell us whether the number of reported cases is inherently good or bad news. To know that, we need context from epidemiologists. An amount of cases that might seem minuscule to an average citizen might signify a tipping point to a communicable disease expert. With domain insights, a data analyst can then leverage their visual communication expertise to effectively disseminate what the content specialist knows.

SNAPSHOT APPROACH

This dashboard leverages the same small multiples chart type as the state dashboard on page 4, but instead of making one chart per state, we have one chart per date. There are many strategies for visualizing time. Here we've opted for a snapshot approach. Instead enabling the audience to scrub through all available dates themselves, we've built a map for the 1st and 15th of every month. This not only encourages a comprehensive examination of the pandemic, but it offers a dashboard format that is easy to print or share online, without requiring the audience to visit the interactive version themselves. Such format decisions might limit the interactive flexibility of a dashboard while simultaneously expanding its potential reach.

COLOR LIMITATIONS

As mentioned on page 5, the COVID-19 risk level categories used here were put forth by an interdisciplinary team of experts led by the Harvard Global Health Institute. We opted to keep the green, yellow, orange, and red categories associated with their original framework. However, this color palette has several limitations that can be seen in this dashboard. The first is a perennial problem with red-green palettes. While these colors have strong connotations associated with bad and good, approximately 7% of the male population in the United States has some form of red-green color blindness. (Future revisions of these dashboards intend to include a color blind toggle.) Secondly, this color palette maxes out at 25+ daily cases per 100,000 residents. While the categorical nature of this color palette allows us to quickly identify which counties and regions fall above or below these four benchmarks, it cannot infinitely distinguish hotspots. For example, by November 2020, the vast majority of the counties are red, and it is impossible to tell where the most severe outbreaks are occurring.

CONTIGUOUS UNITED STATES

When asked to imagine a map of the United States, many folks envision the contiguous US, or the 48 states that share borders to form a single land mass. This dashboard leans into that association, replicating the familiar polygon for a tidy visual that lends itself to intuitive comparisons. These 48 states account for over 98% of America's population. But if you're one of the 5,716,122 individuals who calls the United States home and yet doesn't see your county represented here, this dashboard erases you and your community from the map of America. As analysts, we're constantly looking for ways to streamline our visualizations to make them more intuitive and approachable. But as we do, we must also constantly ask ourselves what complexities we're flattening, what stereotypes or biases we're propagating, and when we're valuing aesthetics over accuracy.

> Time is one of the most common variables we visualize. As we choose a strategy to display the passage of time, we are choosing the perspective from which our audience will understand and remember the data.

The COVID-19 data set is first and foremost a time series data set. More than anything else, the questions we hope to answer with this information revolve around time, particularly the most recent days or weeks that can help us evaluate the current risks and needs our communities are facing right now. The previous dashboard worked to address the immediacy of such questions by fixing the visualizations to the latest date available. However, this dashboard serves a different purpose. By offering just two maps per month and encapsulating the data's full timespan, we're encouraging the audience to step back and consider a more comprehensive perspective of the pandemic. As we decide how to anchor the audience in time, we guide not only their insights but their internal reflections as well.

COVID-19 DATA

Data sets tracking the number of COVID-19 cases and deaths in the United States.

THE NEW YORK TIMES
https://github.com/nytimes/covid-19-data

TABLEAU COVID-19 DATA HUB
https://data.world/covid-19-data-resource-hub/covid-19-case-counts

JOHNS HOPKINS UNIVERSITY
https://github.com/CSSEGISandData/COVID-19

THE COVID TRACKING PROJECT
https://covidtracking.com/

POPULATION DATA

Data sets with 2019 population estimates for American states and territories.

US COUNTIES
https://www.census.gov/content/census/en/data/datasets/time-series/demo/popest/2010s-counties-total.html

PUERTO RICAN MUNICIPALITIES
https://www.census.gov/content/census/en/data/datasets/time-series/demo/popest/2010s-total-puerto-rico-municipios.html

AMERICAN SAMOA
https://www.cia.gov/library/publications/the-world-factbook/geos/aq.html

GUAM
https://www.cia.gov/library/publications/the-world-factbook/geos/gq.html

NORTHERN MARIANA ISLANDS
https://www.cia.gov/library/publications/the-world-factbook/geos/cq.html

U.S. VIRGIN ISLANDS
https://www.cia.gov/library/publications/the-world-factbook/geos/vq.html

COUNTY FIPS CODES
https://en.wikipedia.org/wiki/List_of_United_States_FIPS_codes_by_county

METROPOLITAN AREAS
https://www.census.gov/geographies/reference-files/time-series/demo/metro-micro/delineation-files.html

JOURNALISM

Examples of COVID-19 visualizations in the media.

THE NEW YORK TIMES
https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

NPR
https://www.npr.org/sections/health-shots/2020/09/01/816707182/map-tracking-the-spread-of-the-coronavirus-in-the-u-s

CNN
https://www.cnn.com/interactive/2020/health/coronavirus-us-maps-and-cases/

THE WASHINGTON POST
https://www.washingtonpost.com/graphics/2020/national/coronavirus-us-cases-deaths/

COVID-19 DASHBOARDS

COVID-19 data dashboards made available to the public.

COVID-19 RISK LEVELS
https://globalepidemics.org/key-metrics-for-covid-suppression/

CENTERS FOR DISEASE CONTROL
https://covid.cdc.gov/covid-data-tracker/#cases_casesper100klast7days

WORLD HEALTH ORGANIZATION
https://covid19.who.int/

JOHNS HOPKINS UNIVERSITY
https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

UNIVERSITY OF CHICAGO
https://geodacenter.github.io/covid/index.html

DARTMOUTH COLLEGE
https://www.dartmouthatlas.org/covid-19/hrr-mapping/

TABLEAU PUBLIC
https://www.tableau.com/covid-19-coronavirus-data-resources

Here are all the sources used to create this dashboard series along with additional resources related to COVID-19 data.

A special thank you to the team at The New York Times for creating and maintaining such an important resource for our country during this pandemic. Thank you also to the collaborative team led by the Harvard Global Health Institute for creating the COVID-19 risk categories and disseminating their knowledge in a way the public can understand. This project would not have been possible without their work.

THANK YOU FOR READING

VISIT THE INTERACTIVE DASHBOARDS HERE

WWW.DATADOZEN.COM/TABLEAU-COVID-19-DASHBOARD-SERIES/

Thank you so much for reading this pocket guide about how Data Dozen built its COVID-19 dashboards. You can explore the full interactive dashboard series by visiting our website. Just click here.

These dashboards were created by Data Dozen, LLC, a visual analytics studio based in Kansas City. We hope you find these dashboards helpful during these unprecedented times. If you have any questions, comments, or ideas for improvement, we'd love to hear from you at contact @ datadozen.com. Please note the following disclaimer: We strive to provide information that is complete, accurate, and useful, but our content may not be suitable for every situation, organization, or individual. The content made available by Data Dozen, LLC is provided on an as-is basis, and you agree that you use this information entirely at your own risk. Under no circumstances will Data Dozen, LLC be held responsible or liable for any errors in the content or any direct or indirect damages resulting from the use of this information. Please exercise your best judgment when using the information and data provided here. Thank you.