Prevalence of COVID-19 in the United States

DATA ANALYSIS

Martyna Ryba

EXECUTIVE SUMMARY

COVID-19 is a disease caused by a virus and was discovered in December 2019 in Wuhan, China. The pandemic has affected people across the world in many ways for the past 2.5 years. In this analysis, we will explore the spread of COVID-19 across the United States as well as the effect of vaccines.

DATA SOURCE

All data was sourced from the CDC.

- United States COVID-19 Community Levels by County (link, alternate link)
- Possible additional data:
    - COVID-19 Vaccinations in the United States (link)
    - United States COVID-19 Cases, Deaths, and Laboratory Testing (NAATs) by State, Territory, and Jurisdiction (link)

LIMITATIONS AND ETHICS

The data is regularly updated, most recent update was August 11, 2022. There is no PII in the data, therefore can be used for analysis without concern for exploiting patient privacy.

DATA CLEANING AND CONSISTENCY CHECKS

1. United States COVID-19 Community Levels by County Dataset:
    1. Deleted columns county_fips for redundant information, we do not need a county cde for our analysis, the county name is more clear.
    2. Deleted all columns regarding the health_service_area. This was again redundant information as we already have two different geographically columns to categorize the data.
    3. Changed covid-19_community_level from mixed data type to string.
    4. No duplicate values found.
    5. Total of 188 missing values found, cannot change to NA as this will change the data type to mixed and affect analysis.
        i. Missing 5 values in county population column. However there is still data in the covid_cases_per_100k column. Although there are missing values in the population, we still have information on COVID-19 rates to use for the analysis.
        ii. There are 149 missing values from the covid_inpatient_bed_utilization column. This is small portion of the dataset, and other columns will be useful for analysis.
        iii. There are 34 missing values in covid_hospital_admissions_per_100k. Once again this is a small portion of the data set, other columns are useful for analysis.
2. COVID-19 Vaccinations in the United States Dataset:
    1. Renamed State/Territory/Federal Entity column to state for merge.
    2. Made subset with relevant columns.
    3. Renamed columns for uniformity.
    4. No mixed type columns.
    5. No duplicates.
    6. Total 9 missing values from 3 rows, other information in rows useful for analysis, will not remove. Will not replace to avoid creating mixed data type.
3. United States COVID-19 Cases, Deaths, and Laboratory Testing (NAATs) by State, Territory, and Jurisdiction
    1. Made subset with relevant columns.
    2. Renamed columns for uniformity. Renamed State/Territory column to state for merge
    3. Changed ‘total_percent_positive’ from mixed type to string.
    4. No duplicates.
    5. Deleted Federated States of Micronesia row due to lack of values.
4. Data sets merged and exported.

DATA PROFILE

The data is 79,002 rows by 18 columns.

Column	Column Description	Data Type	Time Variant
county	County name	Qualitative, nominal	no
state	State name	Qualitative, nominal	no
county_population	County population (2019 census estimate)	Quantitative, discrete	yes
covid_inpatient_bed_utilization	Percent of staffed inpatient beds occupied by COVID-19 patients (7-day average)	Quantitative, continuous	yes
covid_hospital_admissions_per_100k	New COVID-19 admissions per 100,00 population (7-day total)	Quantitative, discrete	yes
covid_cases_per_100k	New COVID-19 cases per 100,000 population (7-day total)	Quantitative, discrete	yes
covid-19_community_level	COVID-19 community level [Low, Medium, High]	Qualitative, ordinal	yes
date_updated	Date of data release	Qualitative, ordinal	No
total_doses_given	Total doses of COVID-19 vaccine given	Quantitative, discrete	yes
doses_per_100k	Doses given per 100k people	Quantitative, discrete	yes
percent_with_at_least_one_dose	Percent of population with at least one dose of vaccine	Quantitative, continuous	yes
people_fully_vaccinated	Total number of people fully vaccinated	Quantitative, discrete	yes
percent_fully_vaccinated	Percent of people fully vaccinated	Quantitative, continuous	yes
total_cases	Total cases of COVID-19	Quantitative, discrete	yes
case_rate_per_100k	Rate of cases per 100k people	Quantitative, discrete	yes
total_deaths	Total deaths from COVID-19	Quantitative, discrete	yes
death_rate_per_100k	Rate of deaths from COVID-19	Quantitative, discrete	yes
total_percent_positive	Total percent of positive COVID-19 tests	Quantitative, continuous	yes

QUESTIONS TO ASK

- How are COVID-19 cases and deaths distributed over the United States?
- How do vaccines impact COVID-19 cases?
- How do vaccines impact deaths from COVID-19?
- How do vaccines effect hospitalizations?