6.1: Sourcing Open Data

Source

In early 2020, a worldwide pandemic was triggered by a novel coronavirus referred to as COVID-19. The United States stood out as the nation with the highest number of documented COVID-19 cases. Two datasets, namely covid_us_county.csv and us_county.csv, were made available by Johns Hopkins University and can be accessed for download on Kaggle. These datasets encompass data regarding the cumulative counts of COVID-19 cases and fatalities categorized by county and gender.

Collection

Information was gathered and consolidated from the public health departments of numerous counties and states across the United States. This comprehensive dataset represents a wide geographical range and encompasses a substantial volume of data collected from various regions and administrative divisions within the country.

Limitations

The datasets in question are individual datasets that must be integrated into a unified dataset. However, it's important to note that these datasets lack information concerning COVID vaccinations and the age of patients. Additionally, they do not encompass data related to hospitalizations associated with COVID-19 cases. It's crucial to recognize that this data is constrained to individuals who have access to healthcare facilities, are willing to seek medical attention at a hospital, and have the means to do so. Consequently, the reported cases and deaths may potentially underestimate the true figures due to the exclusion of individuals who may not have these privileges or resources.

Ethics

The dataset doesn't include any HIPAA-related data. Nevertheless, it's worth noting that certain counties within the dataset may have relatively small populations. In such cases, even non-HIPAA data could potentially be utilized to indirectly identify patients. Therefore, it's crucial to perform thorough data analysis and implement robust privacy safeguards to ensure that patient identities remain protected and confidential. This proactive approach to data handling and analysis is essential for maintaining data privacy and security, especially when working with datasets that may inadvertently reveal sensitive information about individuals.

Data Profile

Dataset: *covid_us_county.csv*

Variable	Description	Structured/Unstructured	Qualitative/Quantitative	Discrete/Continuous	Nominal/Ordinary
fips	County code in numerical format	Structured	Qualitative		Nominal
county	United States´s County	Structured	Qualitative		Nominal
state	United States´s State	Structured	Qualitative		Nominal
state code	Two letter abbreviation of State	Structured	Qualitative		Nominal
lat	Latitude of the County	Structured	Quantitative	Continuous	
long	Longitude of the County	Structured	Quantitative	Continuous	
cases	Total COVID-19 cases	Structured	Quantitative	Discrete	
deaths	Total COVID-19 deaths	Structured	Quantitative	Discrete	
date	Date	Structured	Qualitative		Ordinal

Dataset : *us_county.csv*

Variable	Description	Structured/Unstructured	Qualitative/Quantitative	Discrete/Continuos	Nominal/Ordinary
fips	County code in numerical format	Structured	Qualitative		Nominal
county	United States´s County	Structured	Qualitative		Nominal
state	United States´s State	Structured	Qualitative		Nominal
state code	Two letter abbreviation of State	Structured	Qualitative		Nominal
female	Female Population	Structured	Quantitative	Discrete	
male	Male Population	Structured	Quantitative	Discrete	
population	Total Population	Structured	Quantitative	Discrete	
female_percentage	Percentage of Female population	Structured	Quantitative	Continuous	
median_age	County´s median Age	Structured	Quantitative	Continuous	
lat	Latitude of the County	Structured	Quantitative	Continuous	
long	Longitude of the County	Structured	Quantitative	Continuous	

Data Cleaning

Dataset: *covid_us_county.csv*

Column	Type of Inconsistency	Actions
fips	Rename column	Change name to better fit data. New name is "county_code"
county	Mixed data types	county set to string
county	2% of records were unassigned according to Kaggle	Couldn't find the records, so I didn't delete them. The records could have been deleted from deleting the 10,810 records in the country_code column.
county_code	Missing values	Dropped missing values, because only makes up p <1% of data. (Total 10810 records)
state code	Mixed data types	state_code set to string
Multiple dated column names	Data inconsistency	Dropped columns because they were not informative

Dataset : *us_county.csv*

Column	Type of Inconsistency	Actions
fips	Rename column	Change name to better fit data. New name is "county_code"
state code	Mixed data types	state_code set to string

The two datasets were combined to create a new dataset which we will work with deaths_cases_gender.csv

Variable	Description
fips	County code in numerical format
county	United States´s County
state	United States´s State
state code	Two letter abbreviation of State
female	Female Population
male	Male Population
population	Total Population
female_percentage	Percentage of Female population
median_age	County´s median Age
lat	Latitude of the County
long	Longitude of the County
cases	Total COVID-19 cases
deaths	Total COVID-19 deaths
date	Date
region	the region column groups states into U.S. regions using the information from Wikipedia

Key Questions

- Which counties had the most COVID-19 impact?
- Which states saw the most COVID-19 impact?
- When did COVID-19 cases reach their highest point?
- When did COVID-19 deaths reach their highest point?
- Does the size of the population affect the percentage of cases and deaths?
- Does the median age of people affect the percentage of cases and deaths?

Gerami Jacob Dugarte Teran