Data Source

I chose this data set because I wanted to use public health data that provided a geographically detailed description of a variety of health measures. This data is organized by state and city, with 36 different measures for 2020 and 2021. Reports are also available for past years.

PLACES data is external data owned by the Centers for Disease Control and Prevention (CDC) and Robert Wood Johnson Foundation. It combines survey data and model based estimates from the United States CDC and Division of Population Health, both trustworthy government agencies. It is a mix of manual data collection from surveys and automatic estimates generated from Behavioral Risk Factor Surveillance System (BRFSS), the Census Bureau, and American Community Survey. This data set includes all 50 states plus the District of Columbia and reports estimates for 36 measures including health outcomes, preventative services use, chronic disease related heath risk behaviors, disabilities and health status. 29 measures use 2021 data and 7 use 2020. Time lag is caused by some surveys being conducted every other year and the time it takes to compile and analyze the data. Survey responses may be intentionally or unintentionally reported and/or recorded inaccurately, some measures are only collected every other year, and the small area model used for estimates cannot detect effects due to local interventions. These limitations constrict the applicability of results to program and policy evaluations. Additionally, bias in data collection could be compounded using a model, which itself could be biased.

Data Profile

To start this data set had 228,770 rows and 21 columns. I dropped the following columns because they were unnecessary or redundant*: 'StateDesc', 'DataSource', 'Data_Value_Unit' ,'Data_Value_Type', 'Data_Value_Footnote' ,'Low_Confidence_Limit', 'High_Confidence_Limit' ,'LocationID', 'CategoryID', 'MeasureId' , 'DataValueTypeID', 'Short_Question_Text', 'Geolocation', 'Data_Value_Footnote_Symbol'.* 74 '*LocationName*' values were NULL. This was found to be because they were aggregate values for the U.S instead of individual county and state estimates. I created and exported a new file '**US.pkl'** and removed the rows from the working data frame. 2913 rows were found to be duplicates, these were removed and exported into '**dups.pkl'**. For clarification I changed the '*Data_Value'* column to '*Value_Percent'*, and then created a new column '*Value'* reflecting the actual incidence by multiplying the percentage and the total population. Because the 2020 data only covered 7 of the 36 measures I removed it from the dataframe to not add confusion. The final exported data frame has 176,370 rows and 8 columns. Below is a summary of the final data frame variables and descriptive statistics for the qualitative values.

Year	NA	Structured	Quantitative	Discrete
StateAbbr	Time-invariant	Structured	Qualitative	Nominal
LocationName	Time-invariant	Structured	Qualitative	Nominal
Category	Time-invariant	Structured	Qualitative	Nominal

Measure	Time-invariant	Structured	Qualitative	Nominal
Value_Precent	Time-variant	Structured	Quantitative	Continuous
TotalPopulation	Time-variant	Structured	Quantitative	Discrete
Value	Time-variant	Structured	Quantitative	Discrete

	Mean	Min	Max
Year	2020	2020	2021
Value_Percent	28.5	1.6	91.3
TotalPopulation	100,288	57	9,829,544
Value	28,055	1	8,473,067

Bias in healthcare data is a large issue being investigated and uncovered by many parties at all levels of data collection and analysis. The time lag and known limitations of survey data, as well as the potential compounding of bias by the small area model used to create reported estimates, must be considered in the application of results to policy development and assessment. Although this data set is aggregate estimates, it is important to remember that these numbers reflect real people experiencing these health outcomes. The privacy and security of survey results is of upmost importance, especially since anonymized data can still be potentially identifying when outcomes are rare or the population small, and healthcare data can be used to target and discriminate against certain groups.

Questions

Compare frequency of different measures across states.

Do measures trend differently within the same state?

How do different categories relate to one another (ie Prevention and Health Outcomes, Prevention and Health Status, etc.)?

- Identify leading factors for good/bad health outcomes