**Introduction:**

In this exploratory analysis, we dig into two major CDC datasets: BRFSS\_2023\_Prevalence\_Data and COVID-19 vaccination in the United States jurisdiction 2024-10-30. Each of these datasets offers a deep and rich look at behavioral health trends and vaccination coverage across the United States; together, they reveal underlying patterns and themes critical to public health for the year. In this type of analysis, the aim is to look for central tendencies, extremes, or other patterns and associations in the information which are useful for determining health issues at present, and where possible treatment measures could be applied. It is proposed to achieve this using comprehensible quantitative narration of the data in question, interpretation of major measures through statistical summaries, and other forms of visualization in Python programming language.

**Decision of Data description:**

BRFSS\_2023\_Prevalence\_Data:

* Year, Locationabbr, Locationdesc, Geolocation: Geographical factors
* Class, Topic, Question: Concerning the health behavior/health outcome the survey sought to measure.
* Response, Break\_Out, Break\_Out\_Category: Actual individual answers to the questions to a certain demographic or behavioral segmentation.
* Sample\_Size, Data\_value, Confidence\_limit\_Low, Confidence\_limit\_High: Sample size and other measurements of quantitative data and with confidence intervals.

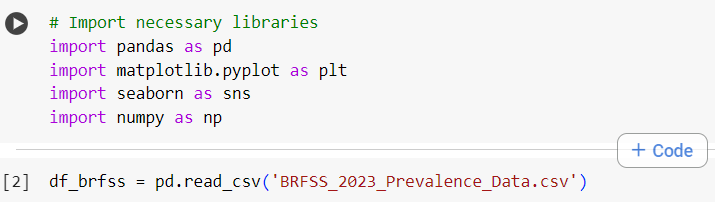
COVID-19\_Vaccinations\_in\_the\_United\_States data:

* Date, MMWR\_week, Location: name of the city and state to which the data was submitted; MMWR (Morbidity and Mortality Weekly Report) week for epidemiological context.
* Distributed & Administered: Numbers of doses shipped and delivered by vaccine types, such as Janssen, Moderna, Pfizer, and Novavax(manufacturers).
* Administered/Distributed per Age Group: Information on coverage of immunizations by population sub-groups including, 5 years and above, 12 years and above, 18 years and above, and 65 years and above.
* Series Completion & Additional Doses: Data on people who received at least one shot and people who received a booster dose or the final dose in the primary vaccination.
* Bivalent Booster Data: Details on bivalent boosters by age and manufacturers.

**Data analysis:**

Dataset(BRFSS\_2023\_Prevalence\_Data):

We have loaded the required libraries and then worked on loading the data/ reading the Excel file using read\_csv.



We tried to print the first 5 rows of our dataset.

This brief is about state demographics in the year 2023; it affords a glimpse of the demography of states such as Alabama and Alaska. It amazes me how some of them just want to understand people like their age or hearing ability through questions like “What is your age?” This goes a long way in illustrating how different folks, say males, or those within households earning more than $200,000, endure health-related complications. Moreover, with the help of geographic coordinates, we see where these people reside and can thus try to expand our search for health profiles by area. This is truly enlightening, thinking about it in terms of getting the behind-the-scenes stories to the numbers on the bivalent boosters, including age and the manufacturer.

A screenshot of a computer

Description automatically generated

The data frame has a density of 5,999 entries on 27 columns providing an insight into the Health Survey for 2023.

The information about which states or areas we are looking at is given by two columns named as Locationabbr and Locationdesc. Most of the columns are in string data type, however; we have some important floats also namely Data\_value, Confidence\_limit\_Low, and Confidence\_limit\_High.

The measured Data\_value, on average, stays at 40.36; the minimum at 0.6 and the maximum at 100. The confidence limits oscillate between points 36.61 and 44.11, thereby indicating the range within which the actual values may be found.

A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generated

The data quality is quite high as most of the columns show perfect missing values of zero, except the Topic column which only has one missing value, while Response has thirty-three. However, Data\_value, Confidence\_limit\_Low, and Confidence\_limit\_High all have 1,377 missing values and these are quite many and should not be deemed reliable. Also worth noting is that Data\_Value\_Footnote\_Symbol and Data\_Value\_Footnote combined have 4,622 missing entries meaning that there are not even supposed to be explanatory details to most of the data represented in the dataset.

A screenshot of a computer

Description automatically generated

We tried Checking for non-empty 'Data\_value' column entries and Converted 'Data\_value' to numeric if it's not already

A screenshot of a computer code

Description automatically generated

value counts for your categorical columns:

The data collected reflect that Alabama has the most entries (2,863), while relatively little response originated from states such as California or Connecticut. Under the class category analysis, Demographics comes out at 1,826 while Tobacco use has the least recorded response at 11. The Topic of Gender is also discussed most often, thereby revealing the importance of Disability status at 469. Of all the breakout categories, Household Income and Age Group have responses with a recurrence of 1,257 and 1,078 respectively. Concisely, the primary findings revealed more emphasis on demographic and socioeconomic health determinants.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The sample size is another variable to be discussed an average value of the Sample size is approximately 518, but the variability of the participant numbers is quite high, with the Sample size varying from 11 to 11 162. The Data\_value indicates the variation of health responses and equals to 40.36 with a minimum of 0.6 and maximum of 100 and an SD of 32.30. Sample\_Size: minimum =68 % and maximum = 1770 % with median value 189%, Data\_value: minimum value = 11.5 % and maximum = 48.9% with a median value of 29.7 %. This also depicts a high variability between the two elements: the sample size and the health indicators in the dataset. Altogether, the fluctuations can be discussed as an indication of the fact that general tendencies in the sphere of health are rather diverse, as far as the results of the survey are concerned.

A screenshot of a computer

Description automatically generated

The count plot illustrates response distribution with reference to geographical reference as depicted on the x-axis. Alabama is most dominant in the responses, other states are less, it may mean that either Alabama's sample proportion is higher in the health survey or Alabama's population is denser. Alaska and Colorado come next but with significantly fewer answers which may be indicative of a regional variation in participation or ad-wide awareness. The rest of the places like California Connecticut and others have very low counts hence they seem to have a small presence in the data set. This visualization shows that geographical disparity affects the general evaluation and understanding of health-related trends and results and thus should not be ignored.

A close-up of a computer code

Description automatically generated

A graph with blue squares

Description automatically generated

The box plot shows the Data Values for each of the Breakout Categories and gives an overview of range and median of the health survey responses. Each box in the figure is actually the interquartile range and the line in each box represents the median of those groups.

In the case of the Household Income and Age Group, scale stretching shows more data values indicating more dispersion of the responses. Outliers are displayed as points outside the whiskers; these circles show high data points which may influence the overall results for categories.

The difference in medians is observed, which seems to indicate the existence of health inequalities; for example, the Education Attained and Race/Ethnicity have lower medians. This representation appropriately draws attention to the patterns and fluctuations of the evaluated parameters and could serve to investigate the existent conditions and changes in their dynamics.

A close-up of a white background

Description automatically generated

A diagram of a box plot

Description automatically generated

The plots are based on Sample Size and Data Value: how the number of respondents indeed influences the health indicators reported.

The major trend is a positive direction which suggests that as the sample size increases most often data values also increase. However, the spread is rather apparent, and some small samples possessed high data values while some large samples reported lower values.

The situation where points congregate at the lower data values may mean that respective categories have small responses regardless of the size of the sample used. The argumentative potential of alpha transparency is that it enables clearer identification of the areas of the densest points by making the intersections more contrasting.

This plot raises concern about the sample size while undertaking health check surveys suggesting that broader cross-sectional health check survey samples may affect the reliability of the reported health indicator.

A close-up of a computer code

Description automatically generated

A screen shot of a diagram

Description automatically generated

The trends observed in the Data Value histogram are declining to the left, which demonstrates that the greater part of health-related responses must be less than 50. The KDE overlay has some of the most prominent peaks in the lower numeric values, which indicates that most of the indicators report low rates. There are fewer high values, illustrating that the surveyed population seldom achieves the best health results. This pattern points out possible areas of concern to public health and potential areas of concern that may need attention. Altogether, the distribution indicates the desirability of elaborating on the causes which have led to obtaining the indicated lower health rates.

A graph with blue lines and a line going up

Description automatically generated

Same as in the case of previous variables, a positive coefficient concerns Sample Size and Data Value. So, it can be stated that as the sample size increases, the data values also increase. The correlation coefficient getting a figure close to 1 indicates a strong relation and thereby increasing the sample size might improve the accuracy of the reported data. For this reason, the heatmap comes with a gradient color so that the strength of this correlation can be easily identified by the warmer color ramp.

A computer code with text

Description automatically generated

A screenshot of a data analysis

Description automatically generated

Dataset 2(COVID-19\_Vaccinations\_in\_the\_United\_States data):

We tried to load the packages and read the CSV file (read\_csv).



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The COVID-19 Vaccination dataset has records of 2368 and features of 109, with different aspects of the vaccination ranging from manufacturers. The most relevant are the number of administered first doses by Vaccine Manufacturers (Janssen, Moderna, Pfizer); and the frequencies of administered first doses per 100000 population by age group. In the case of the number of vaccines distributed, the mean values differ significantly; on the one hand, Moderna and Pfizer; on the other hand, Novavax and unknown manufacturers – which results vary significantly from those mentioned before. The spread from minimum to maximum value also shows that vaccination was significantly low in some weeks while it was higher in others. In general, this dataset is useful for understanding the process of vaccination and can be utilized for further research of tendencies that occurred due to the presence of antibodies.

A screenshot of a computer program

Description automatically generated

We tested this on two rows dated August 31, 2022, and the results were quite good with Iowa and South Carolina. At the same time, South Carolina got the largest total amount of 11,845,275 doses of the vaccine and Iowa got 7,504,855. We also obtained that list to verify the number of doses disseminated by Janssen, Moderna, and Pfizer. However, we saw that some of the columns, especially those for bivalent boosters, had a significant number of missing values, in fact 384 in total. This highlighted some missing elements in the data, which are important that we didn’t get a clear picture of. We looked into the various types of data in our dataset, with total doses, meaning whole numbers, and percentages, meaning decimals.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

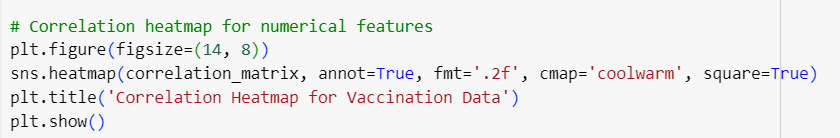
Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated



The shapes of the histograms of the distributed and administered vaccines present considerable information about the magnitude of vaccination.

Distribution of Distributed Vaccines:

The first one charts the number of vaccines that were administered and on top of this there is a density plot. This means it might be right skewed in case you plot it because few states received a lot of vaccine doses than most other states.

Distribution of Administered Vaccines:

The second one depicts the number of the vaccinated population. You might observe from the same context as in the first plot that this distribution is positively skewed. That is, while one or several of the districts reported a comparatively lower number of vaccinations, one or several of the other districts reported many shots. This indicates that in some of these, compliance has been good because health facilities have improved or because compliance was willingly adopted by some members of the community.

A white rectangular object with black text

Description automatically generated

A graph showing a number of vaccine

Description automatically generated

Boxplot of Distributed Vaccines by Location:

This plot depicts the distribution of vaccines across the various regions and the observation made is indicated below.

Here, each box indicates IQR and shows the middle half of the data.

The line inside the box represents the amount that was given out on average.

Places that have larger boxes have greater fluctuations, and any points that are out of range mean that those places received either much more or much less vaccines than the rest of them because of rationalization of the logistics.

Boxplot of Administered Vaccines by Location:

The following plot has been designed to show the number of vaccines delivered at each location.

Like the first boxplot, it also contains the median and the IQR so vaccine uptake can be compared.

Larger squares once more suggest a higher variance of vaccination rates, while individual points could be interpreted as high or low administration spots.

Such information can help in highlighting areas of inadequate or unequal distribution of vaccines and or vaccine uptake so that resources can be dispatched there to support effective vaccine implementation.

A graph showing a number of vaccine

Description automatically generated

A graph showing a number of vaccinations

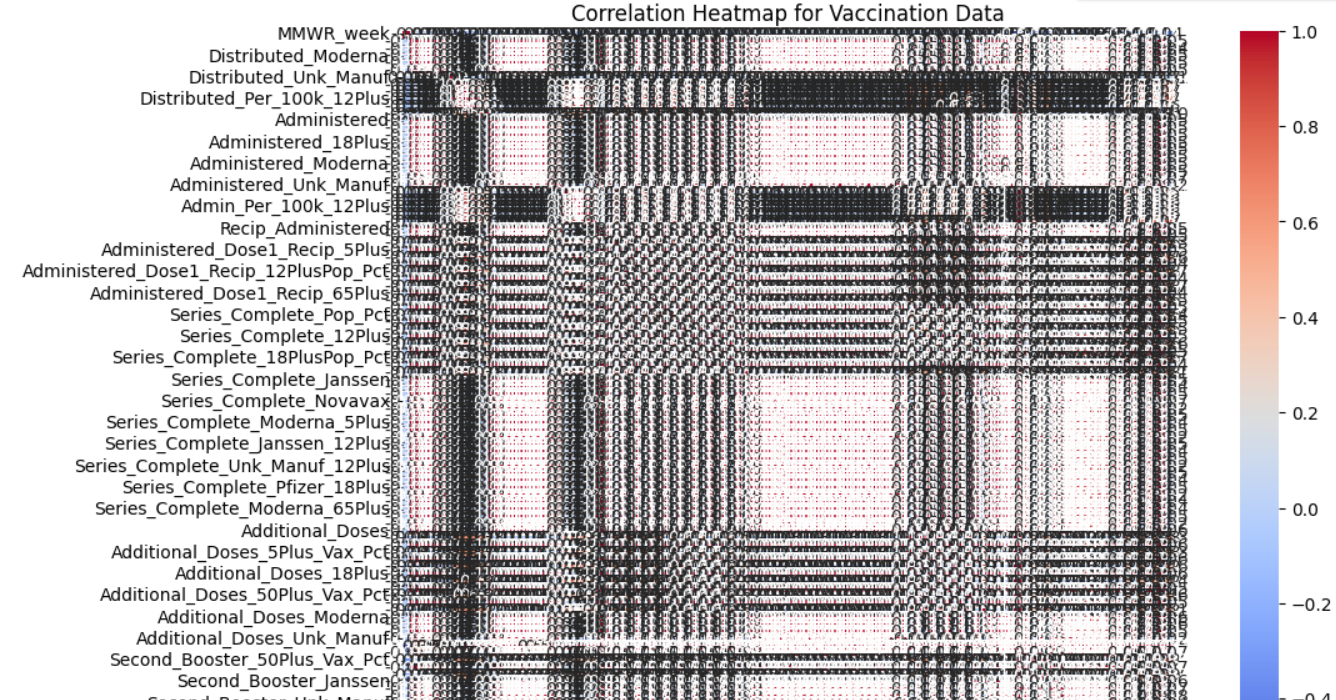
Description automatically generated

The scatter plot used also depicts a positive correlation between distributed and administered vaccines implying that the places that distributed more also administered more. All the colors distinguish different places of distribution and administration, as well as groups of similar concentrations. A few densities show high administration rates for low distribution meaning that they have good strategies or purposes. On the other hand, some distribution is high, and administration is low suggesting miscellaneous issues such as hesitancy or access. In sum, the plot provides tendencies that may help the interventions of public health to enhance vaccination.

A graph showing a number of vaccine

Description automatically generated with medium confidence

The correlation heatmap gives an initial insight into the nature of relations between the numeric variables in the vaccination dataset. Strong positive correlations, especially with the contrast, “Distributed” on one axis and “Administered” on the other show that as the value increases in distributed vaccine, the value of administered increases as well. Thus, lack of negative correlations does not imply benign associations as it was previously seen with values approaching zero proving that certain variables have no correlation and thus are independent.



A grid with black and red lines

Description automatically generated with medium confidence

Conclusion:

The findings from the current study were informative concerning COVID-19 vaccine distribution and administration. Although a considerable number of vaccines were provided, it is not clear that all recipients got the doses that should have been given; hence, possible gaps in vaccine utilization may exist. The graph comparison also pointed to a positive correlation between distributed and administered vaccines; the scope of distributed vaccines was higher than the administered vaccines. In the correlation heatmap, high correlation equations were seen among the numerical attributes especially between distributed and administered vaccines.

The evaluation of the BRFSS 2023 Prevalence Data helped us understand different aspects of health risk factors and disease prevalence by demographic characteristics. Specific recommendations showed signs of disparities across various geographical areas as well as populations. The correlation analysis showed a significant correlation between some of the risk factors and health status that can be targeted for improvement.

**References:**

* **Behavioral Risk Factor Surveillance System (BRFSS)**

Centers for Disease Control and Prevention (CDC). (2021). Behavioral Risk Factor Surveillance System (BRFSS). Retrieved from [Behavioral Risk Factor Surveillance System (BRFSS) Prevalence Data (2011 to present) | Data | Centers for Disease Control and Prevention (cdc.gov)](https://data.cdc.gov/Behavioral-Risk-Factors/Behavioral-Risk-Factor-Surveillance-System-BRFSS-P/dttw-5yxu/data)

* + **COVID-19 Vaccination Dataset**

Centers for Disease Control and Prevention (CDC). (2023). COVID-19 Vaccination Data in the United States. Retrieved from <https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc/data_preview>

**Appendices:**

Chat GPT:

1. What insights can I gain from the BRFSS\_2023 Prevalence?
2. What visualizations can I create to better understand the distribution of health behaviors in the BRFSS data?