My goal is to analyze a set of data that can give clear insights about a larger population by targeting a smaller one. Following the saying, as above so below.

We know that is mathematically and conceptually proven that the characteristics of macrocosm can be found in the characteristics of a microcosm. In our case here, NYC is the microcosm, and the global earthling population is the macrocosm. By studying and understanding what affects the health of the NYC population, we can gain insight into what may affect the health of the population on a global scale.

My reason for choosing this dataset (1) is the understanding that the NYC population being so dense and multicultural that it is the perfect representation of global health data taken from a microcosm. NYC is a melting pot and is representative of the bell curve of population by showcasing characteristics that can be categorized as Universal.

The source of our dataset is a NYC Community Health Survey. The DOHMH, Division of Epidemiology, Bureau of Epidemiology Services conducts an annual telephone survey known as the New York City Community Health Survey (CHS). The CHS provides comprehensive data on the health of the residents of New York City, including estimates at the neighborhood, borough, and citywide levels, covering a broad range of chronic diseases and behavioral risk factors.

The data is carefully analyzed and shared to inform health program decisions and increase the understanding of the relationship between health behavior and health status. The subsequent dataset generated from survey is a primary source of information as it came straight from the population themselves, thus being free from some of the common bias related to statistical research; selection bias, survivorship bias, omitted variable bias (2).

Using the knowledge and skills learned from the Python for Data Analytics course, I devised a process to generate the codes for this project by researching which Python functions, libraries, methods relevant to each of my objectives and consequently creating code with them as it relates to my dataset.

After creating the basic code to import the libraries, my dataset, and cleaning, filtering, displaying the dataset; I created codes to compute histograms and boxplots for the following variables: 'No Health Insurance', 'Smoking Status (current smokers)', 'Binge Drinking', 'Obesity', 'Flu shot in last 12 months, adults ages 65+ (not age-adjusted)', and 'Self-reported Health Status (excellent/very good/good)'.

The histograms allow visualization of the frequency distribution of the variables, helping to understand the shape of the data distribution. The boxplots provide a summary of the statistical distribution of these variables. They give insights into the quartiles, median, and possible outliers in the data.

By analyzing these plots, one can understand how these health indicators are distributed across the New York City population. For example, we can observe how many people are without health insurance or how prevalent obesity and binge drinking are; and how those health behaviors are related to health outcomes.

I have computed code that calculated Statistical Measures from our dataset (1). This section calculates the mean, median, mode, and standard deviation for all numeric columns in the dataset.

These measures provide a summary of central tendency and variability in the data:

The mean provides the average value, the median gives the middle value, and the mode represents the most frequent value. The standard deviation indicates the extent of variability or dispersion in the data. These measures can provide insights into the average state of health and health behaviors among New Yorkers and the variability in these measures.

Subsequently, I tasked myself with creating Correlation Graphs, to investigate the correlation between the various datapoints in our dataset (1).

Mathematically, the script computes a correlation matrix and visualizes it using a heatmap. It also identifies the top five pairs of variables with the highest correlation.

The heatmap provides a visual representation of the correlation between all numerical variables in the dataset. The color scale in the heatmap corresponds to the strength and direction of the correlation.

The top five correlations can help identify health behaviors or factors that tend to occur together. For example, if 'Obesity' and 'Binge Drinking' are among the top correlated pairs, this might suggest a relationship between these two health behaviors.

By interpreting these correlations, along with the distribution and summary statistics, I can provide a comprehensive analysis of the health status and behaviors of New Yorkers. It's essential to bear in mind that correlation does not imply causation, and further analysis would be necessary to investigate any causal relationships.

Based on the results of my analysis, I have concluded that the columns of my datasets I have chosen to analyze do justice in the matter of drawing a parallel between various health behaviors and health outcomes. Though, not explicitly implying causation, this analysis points the directions of risk factors that can contribute to the relevance and prevalence of the health outcomes presented in our dataset.

For example, it is reasonable to infer that an individual that has no health insurance and no personal doctor, that participate in binge drinking, who is an active smoker, drinks more than 1 sugar-sweetened beverages per day is likely to be at risk of obesity, colon cancer, and not to report good or excellent self-reported health status.

**References**

1. Publisher data.cityofnewyork.us. (2023, May 20). New York City Community health survey. Catalog. https://catalog.data.gov/dataset/dohmh-community-health-survey-2010-2016

2. Technologies, D. (2022, September 20). Statistical bias in machine learning: Types &amp; examples. Datatron. https://datatron.com/what-is-statistical-bias-and-why-is-it-so-important-in-data-science/#:~:text=The%20most%20common%20sources%20of,Omitted%20variable%20bias