Air Quality and Community Health Dataset

About

This dataset provides various data collections of air quality from various years including the area and amount of air pollution found within New York City. Alongside, is the community health dataset which represents asthma health complications collected from various years and different areas in New York City as well.

Created By: Anisa Tse and Anna Pham
Created On: 11/22/2023
Collected During: 2009-2016

Content: There are 29 columns, and the two columns that are the focus are the “Name” column and the “Trend.Data.County.Value” column. The “Name” column gives the name of the type of air pollutant (Ozone, Sulfur Dioxide, Nitrogen Dioxide, and Fine Particulate Matter) and the “Trend.Data.County.Value” column gives the average value for that county and year.

Use Cases
Potential real-world applications of the dataset

1. Potential uses for this dataset could be in developing a geographic of age groups that are most vulnerable to health issues (ex. Adolescents; Seniors)
2. Environmentalists may use this dataset to calculate trends on air quality and set campaigns of awareness on this issue to public media.

Badges

Alert Count	4
Completeness	**1**
Racial Factor	1
Collection	**1**
Not Specific Enough	1
Description	**0**
Composition	**2**
Third Factor	1
Unpredictability	1

Alerts

1. A general county value may not be specific enough.

   Mitigation Possible: **Maybe**
   Category: **Collection**
   Potential for Harm: **Not Specific Enough**

   For the air quality, the county value may not be specific enough. A county is a large area, and the air quality for the opposite sides of the county could be different, like how the weather for a different place of a city can be different. However, we did not collect the data, so we do not know exactly how the value was calculated.

   A possible mitigation strategy could be to collect more data on specific area locations of the county, like a zip code.

2. There can be other factors that affect a person's susceptibility to asthma.

   Mitigation Possible: **Maybe**
   Category: **Composition**
   Potential for Harm: **Third Factor**

   For the asthma rates, there can be risk factors that vary from person to person. These factors can include living environment, genetics, and racial or cultural factors that causes a person to be more prone to having asthma.

   Besides collecting more information about the participant's demographics information, there is not much to be done to mitigate this issue

3. The unpredictability of future events can deviate from the predicted course from the data.

   Mitigation Possible: **Maybe**
   Category: **Composition**
   Potential for Harm: **Unpredictability**

   There are unpredictable activities and technological advancements, which can affect air quality and asthma rate in unpredictable ways in the future, making this dataset possibly unreliable for making inferences.

   A possible way to mitigate the unpredictability of activities and technological advancement, further experimental research would be beneficial in addressing the correlational relationship between air quality and asthma rates.

4. The sampling method makes it so that it cannot be applied to other places and communities of people

   Mitigation Possible: **Maybe**
   Category: **Completeness**
   Potential for Harm: **Racial Factor**

Also, this data only sampled air quality and asthma rates in New York City, thus the relationship between the two may be different outside of New York City due to geographical location differences, population risk, etc.

A possible way to mitigate the problem is to collect more data about different subpopulations of people.

Dataset Info

Description

IS THERE AN INTENDED PURPOSE FOR THE DATASET?
This dataset was created for a course assignment to allow students to get experienced with working with data and to get familiar with creating transparent documentation about the dataset. This dataset was created by merging two existing dataset; the air quality data was from Data.gov, and the asthma data was from New York State Department of Public Health. The original data from New York State Department had many more data about other health issues, like cardiovascular issues and cancer, but we chose to use data about asthma because it most directly relates to air quality. The purpose of this dataset is to assess the trend and correlation between air quality and asthma rates in New York City from 2009 to 2016.

HOW SHOULD THIS DATASET NOT BE USED?
This dataset should not be used to draw causal relationships between air quality and asthma because the data are observational data and not experimental data. In order to draw a causal relationship, experiments need to be conducted and other variables needs to be controlled. Otherwise, we would not know if the changes in asthma rates are solely due to air quality or other factors, like advancements of medical equipment.

WHAT CONCERNS MIGHT YOU HAVE ABOUT EXTRAPOLATING TRENDS OR MAKING GENERALIZED INFERENCES FROM THIS DATASET AT A POPULATION LEVEL?
There are unpredictable activities and technological advancements, which can affect air quality and asthma rate in unpredictable ways in the future, making this dataset possibly unreliable for making inferences. Also, this data only sampled air quality and asthma rates in New York City, thus the relationship between the two may be different outside of New York City due to geographical location differences, population risk, etc.

ARE THERE ANY CULTURAL OR DOMAIN KNOWLEDGE THAT WOULD MAKE THE DATASET EASIER TO UNDERSTAND?
Environmental specialists would find it easier to interpret the air quality data and people working in health would find it easier to interpret the asthma rates data. However, the data is relatively easy to interpret, so an average person should be able to interpret the data with no major issues.

Composition

HOW MANY OBSERVATIONS ARE THERE?
There are 4280 observations (rows) and 29 variables (columns)

IS THERE ANY MISSING DATA? IF SO, WHAT IS THE REASON?
There are some missing data in the Three.Year.Average.County.Value column, and they are NA values. The original dataset had the NA values, and there is no explanation to why there are NA values there. A possible reason is that the observation with the NA values and the observations with numerical values in that column are from different sources. Thus, possibly the two sources collected their data differently, and was able to provide different information.

Collection

DESCRIBE ANY PROTOCOLS FOLLOWED TO MANIPULATE OR ADJUST EXISTING DATA, AND INDICATE WHICH FIELDS WERE AFFECTED AND IN WHAT QUANTITY (%).
We created two variables from the existing data called Risk_Level and percent_diff. The Risk_Level variable only applied to the hospitalization observations, since the rates of hospitalizations had a much greater range than the observations for mortality rates. The values in the Risk_Level variable took on either "Low," "Medium," "High," or "Very High," values, which were determined by the quantile values of the hospitalization data. For observations that were lower than the 25^{th} percentile value, it was categorized as having low risk; for observations between the 25^{th} percentile and 50^{th} percentile values, they were categorized as having medium risk; for observations between the 50^{th} percentile and 75^{th} percentile values, they were categorized as having high risk; for observations above the 75^{th} percentile value, they were categorized as having very high risk. For the percent_diff variable, we took the Trend_Data