Task 2.4

Population Data By Geography

Data Sourcing:

This is an external data source. The source of this data is US Census Bureau. Data from the U.S. Census Bureau is highly trustworthy due to its comprehensive collection and processing standards, making it a critical resource for planning and analysis.

Data Collection:

The data is administrative data collected as part of the National Census Program. Each of the U.S. states and territories is required to record all details such as based on each year ,its population, gender and various age group within their jurisdiction. The collection method is plural, partially administrative, and overall manual and prone to human errors.

Data Contents:

The data includes detailed population data by county in the United States from 2009 to 2017.The dataset contains several key variables: County, Year, Total population, Gender, and Age group. It is also divided into different age groups and genders.

Data Relevance:

The data shows population of each US county based on its age group, gender across the United States over multiple years. As it was collected via the government vital statistics program, we can assume that it's the most trustworthy and complete version of the data available. Since data is divided

on various age groups, we can determine population that is most vulnerable to the flu, which tells us that we need to send more staff to these states.

Data Limitations:

Since this data was taken over a period of years from 2009 to 2017, there may be sometimes invalid data in it. Data is tracked at the each house hold level within a county by an governmental agency, this makes it subject to possible typos. As administrative data, it's compiled from records that may not capture real-time changes in population dynamics, such as migration or death rates, within the year they occur.

Influenza Laboratory Tests And Visits

Data Sourcing:

This is an external data source. The source of this data is CDC, which is National, Regional, and State Level Outpatient Illness and Viral Surveillance agency responsible for surveying healthcare system in US
This information is free for the public to view which makes it external. The data is trustworthy since CDC gets their information from all the healthcare providers. If the data is fraudulent, there would be huge consequences.

Data Collection:

 U.S. influenza surveillance system is a collaborative effort between CDC and its many partners in state, local, and territorial health departments, public health and clinical laboratories, vital statistics offices, health care providers, hospitals, clinics, emergency departments, and long-term care facilities.

Data Contents:

CDC lab Test dataset consists of data related to different Influenza flu type's Lab test reports based on US county and years. The Dataset explains variables related to total number of specimens collected for testing , percent of positive and negative samples of the Influenza Flu, also the variable which could not determine the Influenza type.

CDC influenza Visit dataset determines the PERCENTAGE OF VISITS FOR INFLUENZA-LIKE-ILLNESS REPORTED BY SENTINEL PROVIDERS. It has variables which explains total number of patients in different age groups wo were registered for influenza like illness.

Data Relevance:

For our study, this data set contains the first data from our Wishlist (data of all living population by age and by state). It will allow us to normalize the death rate from a state to another and compare it by age group. Since the objective is to map a staff repartition, focusing on the biggest populations may be a valuable input. This data is very relevant for both the objective and the tested hypothesis.

Data Limitations:

Since these data sets were collected by surveys, there tend to be inconsistencies. Surveys only contain a subset of the whole group, which means it doesn't include the whole population. The main limitation is from its many collection methods and the manual aspect in each method: there is a risk of human typos, and maybe of double count for one person.

Children Flu Shots data set

Data Sourcing:

The data source is external because the University of Chicago is collecting the data and sending it to the CDC. These records is accurate and trustworthy because it is verified by healthcare professionals. Given the CDC's oversight and the systematic verification of flu shot information with healthcare providers, the data is considered highly accurate and reliable for research and policymaking.

Data Collection:

Centers for Disease Control and Prevention National Center for Immunization and Respiratory Diseases and National Center for Health Statistics collects and presents it to University of Chicago. The data is collected from phone calls and emails in the form of surveys in order to monitor vaccination coverage among children 19–35 months, collected mostly manually by the phone operators.

Data Contents:

The main content of this dataset is the details of vaccination of the children for age group between 19-35 months. Dataset also explains each child's details like BREAST/FORMULA FEEDING IN DAYS,parental details,financial conditions etc

Data Relevance:

The file could help understand more the young population at risk from all ILI, but wont help us in verifying our hypothesis about elder population.

Data Limitations:

Since it is from a random sampling of parents, the sampling may not represent the whole population. It could contain some manual errors as part of a survey. The data could be biased on some demographics since people may lie (marital status, income categories, etc.).