Nirav Bariya

02/15/2024

1. Summary of the population data by geography US Census data

Data Source: This data source is external. It was collected by US Census Bureau and not by the staffing agency. This is the most accurate data that we can find anywhere as it is published by US Census Bureau. Therefore, this dataset can be trusted. However, the figures in the dataset under different fields is an estimate of the exact value, hence the total of all fields may not be same as the figure in total column.

Collection Method: This dataset is an administrative dataset as this can be utilized for developing policies and for governance purposes. This type of data collection is done through 5 possible modes-a paper census form, an online form, a telephone interview, a high-quality administrative record, or an in-person interview.[1] These methods seem to be manual in nature. There is a time period of 1 year between data collection and in some case, it is more than 1 year. This may affect the uniformity of the analysis and impede aggregation of data as for some regions the data may not be available for all the periods.

Data Content: The data contains county-wise total population and population for different age groups for every year from 2009-2017. As mentioned earlier, the data is not available for all the counties for all years.

Data Relevance: This data is one of the most important datasets as it is directly linked to one of the business requirements of creating a staffing plan based on vulnerable population. We can identify vulnerable population from this dataset and this will drive our staffing plan.

Limitations of the Dataset

Due to the various methods employed to collect the data, it is likely that the dataset is representative of the population. However, there are some years for which the population data is missing in some county. This will limit the aggregation process and comparing yearly population state-wise. Further, the data was collected every year, so there was a considerable gap between when the survey was conducted and when the results were published creating a scope of error. Some of the survey methods include feeding manual data in the system which may involve error. I believe because this data is published by a government agency it is likely to have minimum errors and it is the best option available for this analysis.

Relevancy of the Dataset:

The dataset contains county-wise population and helps to identify total vulnerable population in each county and in each state. One of the business requirements asks to identify states with large vulnerable population and classify each state as either low risk, medium risk, or high risk. Thus, this will be one of the most important datasets that we will be working with.

2. Summary of Patient Visits data set

Data Source: Patient Visits data set is an external dataset as it is not provided by the staffing agency but by hospitals in all states and is aggregated by region. This dataset is published by Centres for Disease Control and Prevention (CDC) making it a trustworthy dataset available.

Collection Method: This dataset is of survey type as it may be used for understanding the spread of the disease by the government. It is provided by outpatient medical professional. Therefore, this is a manually created dataset. The data is for time period 2010-2019 and therefore any analysis done at present may not be relevant for 2024.

Data Content: The dataset contains county-wise weekly patient visits for Influenza like symptoms and weekly total positive patients between 2010 and 2019. It also informs us about number of providers in each county every week. It also has different age group columns, %weighted ill, and %unweighted ill.

Data Relevance: Since we are required to provide a staffing plan to tackle the problem of Influenza season, we will require the details about weekly patient visits, weekly positive cases, and number of providers each week. This will help us device a staffing plan according to the need of each county. Hence, this is a relevant dataset to meet one of the business requirements.

Limitations of the Dataset: The dataset doesn't precisely tell, among the positive cases, how many belonged to each of the age category. Therefore, it would be difficult to identify which age group was affected the most. The dataset is of survey type and hence is not the complete account of influenza cases. Also, the non-vulnerable population may not have appeared for the test due to less severity of symptoms. This might make the data biased and may not represent the actual number that was affected due to Influenza. Further, since the data was manually collected there is a possibility of human error. The data is published weekly and I assume it is collected on daily basis.

3. Summary of Lab Tests Dataset

Data Source: Lab Tests dataset is an external dataset compiled and published by CDC. The data comes from 100 Public Health providers and 300 clinical laboratories located across US. As it is published by a government agency, it is a trustworthy dataset.

Collection Method: The data set is of survey type as it helps in understanding the spread of Influenza. It is collected manually as it provided by health workers. The data is only available between 2010-2015. Hence, there is a considerable time lag between when it was published when we are conducting the analysis.

Content: The dataset contains weekly positive influenza tests by states between the year 2010 and early 2015. The dataset also elaborates on the type of virus detected during the test.

Limitation: As noted earlier, the dataset is quite old and the results of the analysis may not be relevant for current year. Further, percent positive column has mixed data and it is difficult to interpret. Thus, this column can not help us in the analysis process. Since this is a survey

data it is not a complete count of total positive tests in the United States. Further, it could include human error as it is a manually entered data. There may be implicit bias in the data in that non-vulnerable population may not have appeared for lab tests because of non-severity of symptoms.

Relevance: As we are required to present a staffing plan which needs a spatial distribution of medical personnel, we will require the data about the positive cases for each of the region. Hence, this dataset is relevant for the analysis.

4. Children Flu Shots Data set

Data Source: Children Flu Shots data set is external data set because it is provided by The National Immunization Surveys (NIS) through telephone interviews with parents. It is conducted by the University of Chicago. As it is published by NIS, it is a trustworthy data set.

Collection Method: This is a survey data conducted by the University of Chicago through interviews of parents in all states. Since the process is of interview type the data is manually collected. The data is for year 2017 only. Since it is only for one year some of the responses of parents might change and hence there is a time lag.

Content: The data set has range of information such as- child identifier, unique household identifier, year of interview, age category of child, duration of breastfeeding in days (recode), number of people in household (recode), relationship of respondent to child (recode), education of mother categories imputed (recode), first born status of child (imputed), income to poverty ratio (recode), income to poverty ratio: imputed (recode) and so on. There are a total of 73 variables which are results of questions asked in an interview.

Limitation: The dataset is of the year 2017. This is a limitation as some of the answers of parents might have changed over the years such as education qualification. Further, answer to many questions is NA. This may impede our ability to draw conclusions and relationships among different variable in the data set. There are questions related to education and income of the household that can easily be misrepresented by the interviewee. Therefore, there is a scope of creating bias in the data set. Again, it is a one-time data and there is no data about follow up interviews. There may be a chance of human error in the data set.

Relevance: As we will not be looking at relationship between different variable of the dataset to identify any relationship between the children who received flu shots and the variables this is an irrelevant dataset for us.