Chris Arnold

Task 1.4

1. Using the Influenza death data set as a guide, write a summary of the population data by geography US Census data. Remember, all data sets' download links are to be found in your Project Brief!

   ○ Summarize the data source. Include whether it's internal or external data, who owns the data, and how trustworthy it is.

   The source is the US Census Bureau. They are trustworthy in the reporting of the data that they get as they are a government agency. This is external from the staffing agency and the government owns the data. Like stated before the data from the government should be trustworthy from the standpoint that they accurately give out the information that they receive, however not everyone chooses to participate in the census, and sometimes the people that do, do not fill out the form correctly.

   ○ Summarize the data collection method. Is it administrative data, usage data, or survey data? Is it collected manually or automatically? Is there a time lag?

   The US Census is a survey format that is sent to every residence in the country. It is a manual form that is filled out by hand and mailed back. It measures what people choose to write on the form, so if people make mistakes or choose not to write certain things on there the data submitted at times will be inaccurate. In general I would say that most that submit the form fill it out properly but there are most likely some errors present. Also census data is only collected once every 10 years and a lot can change during that time so there is definitely a lag going on.

   ○ Write an overview of the data contents. What variables are included?

   This data set includes all of the states in the united states divided into their individual counties. The data goes from 2009-2017 giving total population within each county, then splits up into categories. First it splits it into the male/female counts. Then it divides the data into 5 year age categories going from 0-5, all the way through 85 and over.

2. Be sure to note any limitations of the data set. Could the data be biased? Is it collected infrequently? Could it contain manual errors?

   As stated before, the census is a manual form that people have to participate in by hand choose to send back to the government for processing. The reporting itself should not be biased, however we need to consider that certain groups may be opposed to giving out their information (likely small groups). Also since this is a manual collection method, it is prone to having at least some user input errors. In general, the census is likely to be close to accurate with some errors in it. Also it states in the description in the project brief that the numbers are estimates, so ther eis most likely some extrapolation going on.

3. Use the project objective and your hypothesis to determine the relevancy of the data set to your project.

   My hypothesis: If a region has a high vulnerable population, then these regions will have higher case numbers of the flu given an outbreak situation.

   This data set can most definitely help me find trending information as to where more vulnerable populations are higher in number. The age categories will help me to sort through which states and more specifically which counties have the vulnerable age brackets in higher numbers, thus would be more prone to outbreak (if the data as a whole shows this after investigation).

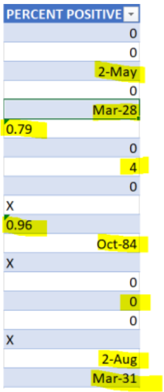
Repeat steps 1–3 for the Influenza Laboratory Tests

- Summarize the data source. Include whether it’s internal or external data, who owns the data, and how trustworthy it is.

This data source is from the CDC and is used for public safety reasons. This is also an external source however the data does get collected at individual institutions that have to submit their numbers for this set to be complete. The CDC would be a trustworthy source because their interest is public health and safety without pushing any specific agenda or bias.

- Summarize the data collection method. Is it administrative data, usage data, or survey data? Is it collected manually or automatically? Is there a time lag?

This would be an administrative data set that comes from each state in the United States. Without knowing specifically, I can say almost without a shadow of a doubt because of the below small screenshot that this data is manually entered.



PERCENT POSITIVE
0
0
2-May
0
Mar-28
0.79
0
4
0
X
0.96
Oct-84
X
0
0
0
X
2-Aug
Mar-31

There are multiple formats of entry and some (the date format) that may not even be able to used. The data comes in weekly (for the most part) for every state so the lag should be minimal to none depending on submit deadlines for data collection.

- Write an overview of the data contents. What variables are included?

This data set gives the specimen numbers given per state on a certain week of the year. Beyond that we have columns designating subtyping of the strain of the flu found if it was performed. If it was not performed or unable to be performed that was counted as well. Also included is percentage positive rates on the total tests that were submitted to the CDC for that week.

4. Be sure to note any limitations of the data set. Could the data be biased? Is it collected infrequently? Could it contain manual errors?

The data from the CDC is not biased, however there could be a small amount of hospitals or reporting agencies that report more positives than were truly accurate if hospital/facility funding is given out based on this number. The data is collected quite frequently, in most cases once per week when taking a first glance at the data. As stated above, this looks to be a manual entry data set so there are going to be errors present that have to be sorted through.

5. Use the project objective and your hypothesis to determine the relevancy of the data set to your project.

My hypothesis: If a region has a high vulnerable population, then these regions will have higher case numbers of the flu given an outbreak situation.

This set can be used to find the correlation coefficient between the vulnerable population numbers and the positivity rates seen in each state. I could possibly complete the project with this data set and "Population data by geography", and "Influenza deaths by geography, time, age, and gender" used in conjunction with each other.

Repeat steps 1–3 for the Patient Visits data sets.

- Summarize the data source. Include whether it's internal or external data, who owns the data, and how trustworthy it is.

This data source is from the CDC and is used for public safety reasons. This is also an external source however the data does get collected at individual institutions that have to submit their numbers for this set to be complete. The CDC would be a trustworthy source because their interest is public health and safety without pushing any specific agenda or bias.

- Summarize the data collection method. Is it administrative data, usage data, or survey data? Is it collected manually or automatically? Is there a time lag?

This would be administrative data from a number of providers in within each state in the United States. There would be pretty minimal lag because the information has week of the year markers meaning that the data is collected weekly. This also appears to be a manual entry method because there are different input types in the % unweighted column.

- Write an overview of the data contents. What variables are included?

This provides a weekly report on the positivity rate of patients within all states. There is a weighted percentage of positive tests and unweighted percentage of positive tests, a column for ages 0-4, 24-49, 25-64, 5-24, 50-64, and 65, there is an ILI total, number of providers, and total number of patients seen.

6. Be sure to note any limitations of the data set. Could the data be biased? Is it collected infrequently? Could it contain manual errors?

Like stated in previous data sets, this is the CDC which is a government agency so there should not be any bias. The data is collected weekly so there is pretty frequent comparison values to different parts of the year/months within the year. It looks like there is manual entry error in the %unweighted ILI column as there are two different entry types. One that shows percentages and another that shows larger numbers. I would have to dig in a lot to figure out what the numbers are that are not percentages. Also since it is manually entered, there could be errors in some of the weeks for some of the states.

7. Use the project objective and your hypothesis to determine the relevancy of the data set to your project.

My hypothesis: If a region has a high vulnerable population, then these regions will have higher case numbers of the flu given an outbreak situation.

It disappointed me that there are columns of data for different age brackets, however all of them are filled in with an x, and x appears to mean that there is no data for that cell. With the % unweighted column being pretty scattered in data entry type, it would take a lot of cleaning to get accurate values out of the data set. In general, this gives me less than the previous data set and looks like it would take more work to make it useable so I would most likely not use this unless I absolutely had to, in order to replace some missing data from another data set or enhance another data set.

Repeat steps 1–3 for the Children Flu Shots data set.

- Summarize the data source. Include whether it's internal or external data, who owns the data, and how trustworthy it is.

This data source is from the CDC which is external to the staffing agency. The data is owned by the government, so again it should be a trustworthy representation of the data that was collected. The institutions could be manipulating data inputs if they have funding tied to low income. If you were trying to tie low income to higher risk because of environment, there is a possibility of slightly manipulated data. I would say that for the most part, the data should be accurate with this being a governmental organization.

- Summarize the data collection method. Is it administrative data, usage data, or survey data? Is it collected manually or automatically? Is there a time lag?

  This was survey data. If you click on the link to the source it takes you to a page that says "National Immunization Surveys." They are phone surveys that manually collect the data through vocal response. Since this is all done by phone, there can be a lot of time in between the first and last collected data creating a possible decent lag in data. It could be that after the survey, someone changes their mind and does to get vaccinated. If you are trying to get an honest representation, even the fact that some of the questions are being asked can make people second guess their own opinion or stance. Also the way that questions are asked can sometimes shut people out and give inaccurate responses. Just browsing through the data for one minute reveals that there are a decent amount of people unwilling to give out their financial data. While they have that right, it can put a major hinderance on people looking to make connections through the data on a large scale.

- Write an overview of the data contents. What variables are included?

  This set contains demographic information on families and their vaccination status for the flu. It contains things like financial bracket data, number of children, racial data, sex, state location, housing data, insurance status, and many more. It is a very good all-encompassing large set of data.

8. Be sure to note any limitations of the data set. Could the data be biased? Is it collected infrequently? Could it contain manual errors?

The big limiting factor in this data set would be people's willingness to give answers, and people's willingness to give out accurate data. It would be less likely that there are manual errors because the data being entered is not entered by the people taking the surveys. There is a small chance that some people with accents could cause the person administering the data could cause an input error. In general, the responses that are documented should be what the people said over the phone.

9. Use the project objective and your hypothesis to determine the relevancy of the data set to your project.

My hypothesis: If a region has a high vulnerable population, then these regions will have higher case numbers of the flu given an outbreak situation.

This data set could be used if it is shown in the data set that lower income families show a lower case of vaccination. If this can be tied to certain regional locations that could indicate a higher vulnerability to outbreak.