Project 1 Write Up

Effects of Income Level on Disease Rates

•The Big Question: What impact does income level have on infection rates of preventable diseases?

When making our initial inquiries into what we would analyze for our project we decided to compare Income Level with various diseases pulled from Census data. This data covered the years 2006-2010, including 8 diseases in total. After combing through the data we decided we would track only Hepatitis A and Pertussis.

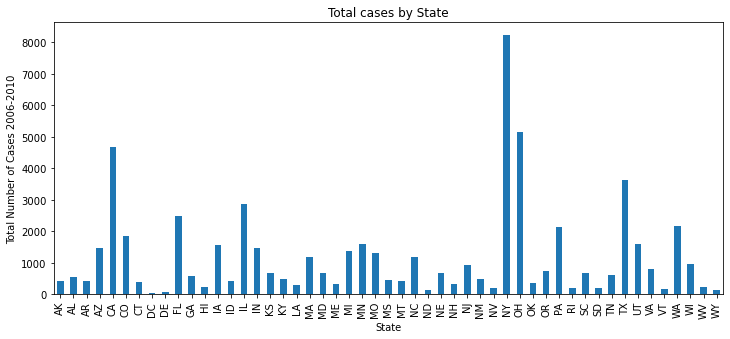
•High-level:

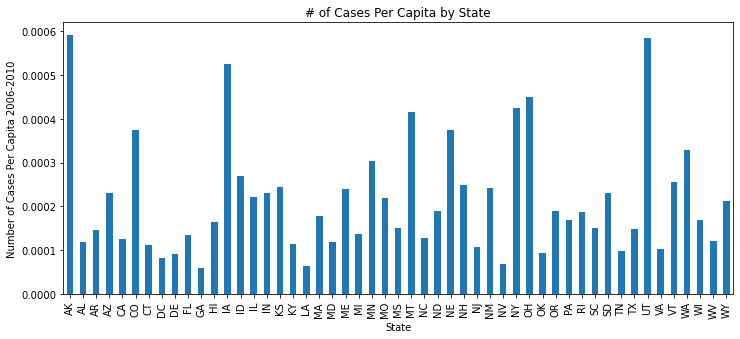
​ •What is the average income for each state?

We knew we would focus on average income but as that can be identified as slightly too general we decided to hone our focus and make it based on average household income specifically. That specific approach made it easier to encompass a larger slice of the population as well as measure data on a per-capita scale more accurately.

​ •How many reported cases of various diseases are there in each state?

As for this information, we started with the total number of 8 diseases included within the data and slowly eliminated them based on curability and relation to income (based on our initial hypothesis). Once we were able to pin it down to Hepatitis A and Pertussis we were able to proceed to finding out how many cases there were per-state

and how it related to average household income. 



•Is there a correlation between a state’s average income level and the prevalence of certain diseases?

Our initial hypothesis was that there would be a positive correlation between disease and income as income directly relates to quality of healthcare in the United States and these are curable diseases. So in turn that ended up being the final hypothesis we would test.

Clean Up Process

•Epiweek – used by CDC, but not a usable date format

•Split into Year and Week and placed in new columns

•Cases listed as objects in DataFrame

•Removed blanks and converted to integers

•Census data lists state names in full, disease data listed abbreviations

•Changed state names in census data to abbreviations to allow for DataFrames to be merged

•Needed to integrate a loop to pull multiple years from Census API

Conclusions

•Correlation scatterplot shows that instances of disease are higher in lower income states

•Average income at which cases start to drop off was higher than expected

•Lower income is associated with many contributing factors to higher rates of disease, including:

•Lack of access to care and affordable health insurance

•Lower quality education, both from underfunded schools and an inability to afford higher education

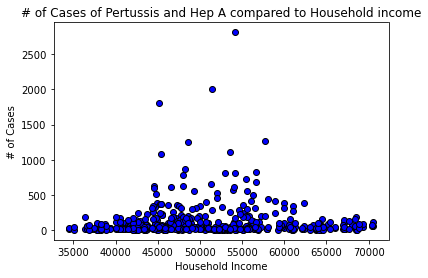
Insights

•When did the other diseases get eliminated or drop so low that they stopped being reported?

•Is there a correlation between the states with the highest/lowest infection rates and the highest/lowest average income?

•Track changes over time – chart each year to see how it changes

•What other contributing factors could be added to the analysis? Religion? Climate? Population density?

****

* Questions that continue to focus tactical opportunities from a state and family size understanding. This data will continue to narrow data and increase visibility to further discovery of health correlation and other associations among data currently available.

