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.

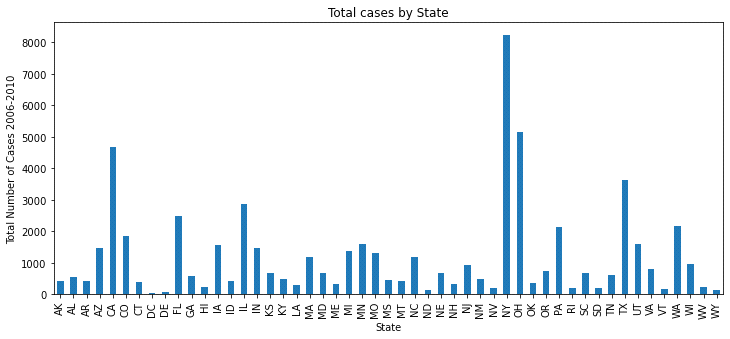
•High-level:

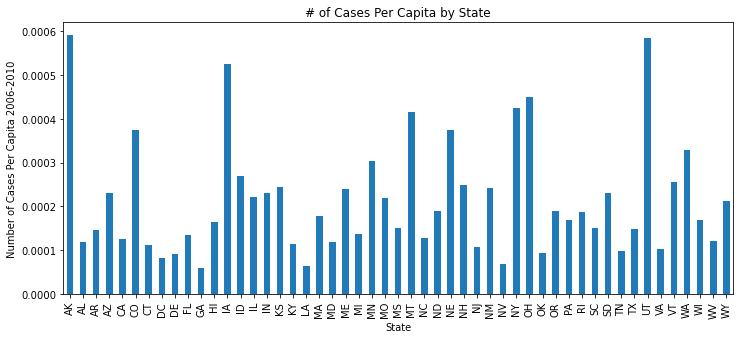
​ •What is the average income for each state?

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

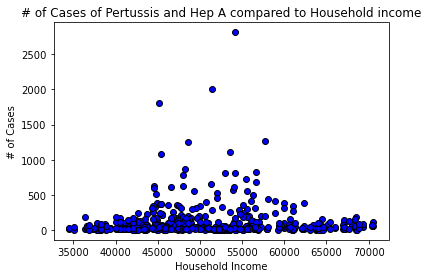
* Issue: Epiweek – used by CDC, but not a usable date format
* Solution: Split into Year and Week and placed in new columns, by converting epiweek column into a string and then splitting the string and placing the new values into columns.
* Issue: Cases listed as objects in DataFrame
* Solution: replaced placeholder values ‘\\N’ from cases column with a ‘nan’ value, then converted cases column to integer as we anticipated the need to use formulas when doing the analysis.
* Issue: Census data lists state names in full, disease data listed abbreviations
* Solution: Changed state names in census data to abbreviations using .replace() to allow for DataFrames to be merged
* Issue: Needed to gather multiple years from Census API ACS1
* Solution: utilize a loop to perform multiple api calls to the census api and gather a Json object that has all 5 years of data.

Conclusions

Correlation scatterplot shows that there is no strong correlation between income and disease rate. This was quite surprising as it seemed fairly intuitive that income level would have an influence on disease prevalence, especially ones that are preventable through vaccines and hygiene. Lower income is associated with many contributing factors to higher rates of disease, including lack of access to care and affordable health insurance, as well as lower quality education, both from underfunded schools and an inability to afford higher education

Insights

The lack of correlation observed raised further questions, as did the limited disease data for the years we selected. For example, 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? What other contributing factors could be added to the analysis? Religion? Climate? Population density? And further, what would be needed to track changes over time?



* 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.

