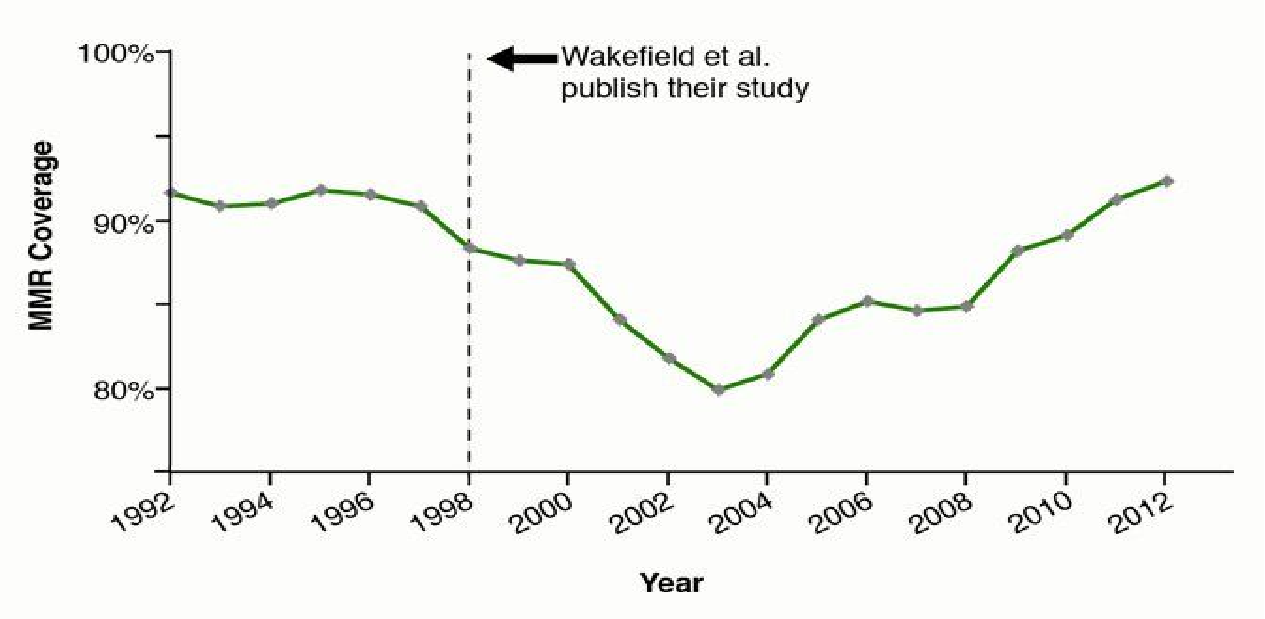
Deliverable 3

Albert Fung, Jeffrey Ng

Introduction

"Thick darkness has gathered over our squares, our streets, and our cities … [w]e find ourselves afraid and lost" (Pope Francis, 2020).[1] 2020 marks an abysmal year within the United States. With the coronavirus (covid-19) running uncontrolled for much of the year, the nation's health systems and institutions have come under intense scrutiny. Yet, the topic of health in the country has been hotly debated long before the coronavirus pandemic arose. Heated discourses in the country have long centered on the areas of public health, including the healthcare system, health information, and in more recent years, vaccinations. The first vaccination in the United States occurred in 1800[2] and since then, there has been staunch opposition. For example in 1879, a wealthy businessman founded the Anti-Vaccination Society of America in New York.[3] And more lately, anti-vaccination groups (known as anti-vaxxers) have surged in popularity amongst many areas in the country due to a faulty study that linked vaccines with autism (see Figure 1).



[1] *Quotes of fear, defiance and hope as the coronavirus pandemic spans the globe*. (2020, June 28). https://www.reuters.com/article/health-coronavirus-quotes-factbox/quotes-of-fear-defiance-and-hope-as-the-coronavirus-pandemic-spans-the-globe-idINKBN23Z058.

[2] *All Timelines Overview*. www.historyofvaccines.org/timeline.

[3] Novak, S. (2020, April 28). *The Long History of America's Anti-Vaccination Movement*. Discover Magazine. https://www.discovermagazine.com/health/the-long-history-of-americas-anti-vaccination-movement.

Figure 1. MMR coverage rates in the UK before and after faulty study.[4] The coverage rates are for the UK. This paper was eventually fully retracted in 2010. Even before the full retraction, 10 of the 13 authors cosigned a partial retraction of the paper's main interpretation in 2004. Many anti-vaxxer groups and individuals, however, still use this faulty study today as justification for their movements.

The current pandemic has brought vaccination concerns back into the spotlight. As companies and governments scramble to create and distribute vaccines for the coronavirus, many in the United States worry about the number of people who will actually take it. The nation's history of anti-vaccination movements and the potential surge of mistrust and misinformation in the past few months severely diminishes many people's willingness to take a vaccine that they view as rushed and possibly even dangerous - even if there are enough valid studies and trials that may say otherwise. Because vaccination has become such a critical matter, our goal was to examine the myriad of different factors that may impact vaccination rates in the United States. Using the vaccination rates by each state (and the District of Columbia), we investigated whether common factors such as educational attainment, income, and total state spending on health were correlated with vaccination rates in each state. Along those lines, we also scrutinized trends by states and groups of states over the time period of 2007-2017. That is, are any of the states trending together or vice versa? Taking this a step further, we investigated why certain states or certain groups of states are trending in a direction by utilizing the factors that were examined previously. Specifically, for example, if some state was trending downward in terms of their vaccination rates between 2007-2017, our goal was, therefore, also to investigate whether factors like income, total state spending etc., were correlated with such a trend in order to explain the reasoning for any increases or decreases in vaccination rates over time. This project will be making heavy use of regression analysis (see the Methods section below). Through our research and results, we hope to make clear some of the factors that influence vaccination rates within the U.S., thus helping leaders and state institutions understand the components to focus on in order to effectively promote their people to respect vaccination guidelines - thereby reducing deaths and increasing long-term wellness.

[4] Helft, L. (2014, September 5). *The Autism-Vaccine Myth*. PBS. https://www.pbs.org/wgbh/nova/article/autism-vaccine-myth/.

Data Sources and Datasets

Due to our data focusing on state by state information, much of our data was collected from government agencies such as the Centers of Disease Control (CDC) and the US Census Bureau. To supplement this, data was also gathered from Non-government organizations (NGOs). Some NGOs gather information directly from the government while others do their own polling and data collection.

The sources for our datasets and their respective names can be found in our “data” folder within the repository as listed in the bibliography section of our report (see below). Some of the data sources came as downloadable csv or Excel files, while others had to be copied manually. As a result, those that had to be copied manually may exclude information that was deemed irrelevant for our purposes (for similar reasons, datasets in our data folder may have information that was dropped).

Data Cleaning

Our data cleaning and wrangling can be summarized in three steps. First, we manually deleted irrelevant data and changed all of the names for Washington, D.C. to “District of Columbia” in order to maintain a standardized naming convention.[5] Secondly, we preprocessed many of the numerical columns, removing dollar signs and commas. This makes further analysis much easier by formatting the data into simple categorical or numerical information. Finally, we merged all of the relevant datasets by state. By having all of our data in one dataset, we can reduce the number of files needed to be read to just one. The merged and final dataset is named “full_dava.csv” and can be found “data/full_data.csv”.

Features in Our Data

States:

- This label simply represents each state within the US. We are primarily using state data. There are 50 states plus the District of Columbia (DC), which totals to 51 observations. Since we're looking at state data, we found data and created this final dataset by merging the

[5] For Washington, D.C., many sites vary in their naming conventions (for example D.C. versus District of Columbia or even DC (no periods)). For the sake of standardizing the name, making formatting simpler, and making the label obvious, Washington, D.C. references were renamed to “District of Columbia”.

observations with the state name.

Labels 2007-2017:

These labels represent the percentage of each state's population that has received the measles, mumps, and rubella (MMR) vaccine in that year. The MMR vaccine is one of the most commonly recommended vaccines for individuals and is typically taken at a young age in two doses. This vaccine is very commonly required by law for children to attend school. 2017 is the primary vaccination rate that was used. We also used, however, others as well to gain additional insight into trends over the years.

Exemption:

- This variable represents the exemptions that each state allows for all mandatory vaccines. There are three types of exemptions: religious exemption, medical exemption only, and personal belief exemptions. Religious and medical exemptions are self-explanatory. Personal belief exemptions are very broad and may overlap with other exemptions such as religious exemptions. From strictest vaccination legislation to the most lax: medical exemptions only, religious exemptions, personal relief exemptions.

Total Health Spending:

- This feature represents the total health spending that each state spends per year (represented in millions of dollars). This includes “spending for all privately and publicly funded personal health care services and products (hospital care, physician services, nursing home care, prescription drugs, etc.) by state of residence. It should be noted, however, that we decided it would be prudent to find the per capita since states have widely varying populations. Therefore,
just examining the total health spending per state would fail to take into account that the population may play a role (states with lower populations likely spend much less and vice versa).

Population

- Represents the population of each state. This feature can be used for a myriad of things. For example, it was used in conjunction with the "Total Health Spending" feature to find the total health spending per capita in each state.

spending_per_capita

- This feature represents the total health spending per capita for each state. This was a variable that we created by dividing the "Total Health Spending" observations by the "Population"

observations. This feature will give us more reliable readings on how much each state spends on
health by removing any skewing of the data that might affect "Total Health Spending" due to population size differences.

population_density
- Represents the number of people per square mile for each state. This feature may have multiple uses. It provides insight into the geography and demographics of different states.

Median_income
- Represents the median household income in per state. The values are indicated in dollars.

HighSchool_Plus
- Represents the percentage of residents in each state that have completed at least
high school. This percentage also includes those who have completed studies after high school such as an undergraduate degree (bachelor's).

Bachelor_Plus
- Represents the percentage of residents in each state that have completed at least a bachelor's (undergraduate degree). This also includes those who have pursued further education (postgraduate education, etc).

spending_per_pupil
- Represents the amount of dollars spent per elementary and secondary (K-8) student by State.

Methods

After cleaning the data and doing research into various methodologies that could be used, we decided that regression analysis would be best suited for our purposes. Performing regressions would allow us to find whether any of the factors are correlated with vaccination rates. To do this, we set our single dependent variable as the vaccination rates and our independent variables as the various different factors (as outlined in the features section).

Because the vast majority of our factors are continuous, linear regression was the regression model chosen. For the categorical features in our data, R’s tidyverse package includes a lm function (linear model), which handles categorical inputs automatically. With so many factors, we did our analysis using a multiple linear regression. That is, we had one dependent variable and multiple independent variables in our regression formula. Initial regressions were used using vaccination rates **from 2017**, however other years were also examined (to examine whether other trends/interesting data existed)**.**

In addition to linear regression, we also wanted to find the trend in vaccination rates throughout the 2007-2017 time period. To do this, we again use linear regression to determine the best fitted line with respect to the vaccination rate of a state. We then took the slope of this line and used it as the rate of change in vaccination for that specific state. Using these results, we regressed the trend in vaccination rates against the features outlined in the above section.

Results

Running the regression revealed that two of the factors were significant (indicating a correlation with MMR vaccination rates): spending per capita (health expenditures) and percentage of residents with at least a bachelor's degree. To evaluate significance, we examined p-values of 5% or lower (95% confidence). For 2017 vaccination rates, p-values indicated that spending per capita and bachelor's and higher education completion were significant. In order to confirm that these factors were in fact relevant, however, we were required to test for multicollinearity. None or average multicollinearity amongst the independent variables is a necessary assumption for linear regressions and to test for it within our independent variables, we examined the tolerance and the vif (variance inflation factor). Both are available through the olsrr (Tools for Building OLS Regression Models) package in R. Through tolerance and vif, we discovered that the second factor (percentage of residents completing at least an undergraduate (bachelor's) degree) had extremely high multicollinearity. To remedy this, we removed it from our regression. On the other hand, spending per capita had no major issues with multicollinearity.

For further insight, we also performed this analysis on vaccination rates of other years (primarily vaccination rate data was in 2017) and the outcomes were widely varying in terms of which factors were significant. Spending per capita, however, was consistently significant regardless of the year of the vaccination rate data. Therefore, out of all the factors examined, state healthcare spending per capita is the only consistent feature that has correlation with vaccination rates.

Examining the trends of states and groups of states, we found no real reason or factors for states that have the similar trends in vaccination rates from 2007 to 2017. Interestingly, we did find that the 5 states with the worst trends (states with the highest trend downward from the period of 2007 - 2017) were all Republican states with the exception of Wisconsin (which in that time period has fluctuated between Republican and Democratic). **See Figure 2 for all the state's trends over the 2007 - 2017 period** (ordered from negative trends to positive trends). We did not find any reason for the bottom 5 being Republican (again, with the exception of Wisconsin), nor did we find any similar trends for the states with the most positive trends (best increases in vaccination rates). The states that are ranked highest for their change in vaccination rates were a mixture of both Republican and Democratic.

Furthermore, when considering the same factors (health spending per capita, population density, household income, education attainment, etc.) used in the previous regression analysis, *we did not find any correlation between them and the state's trend over the period.* This result was obtained by regression the factors on the slope of the line from our previous linear regression (since the goal was to find correlation between the factors and the vaccination rate changes over the 2007-2017 period). None of the subsequent coefficients were statistically significant as indicated by their p-values.

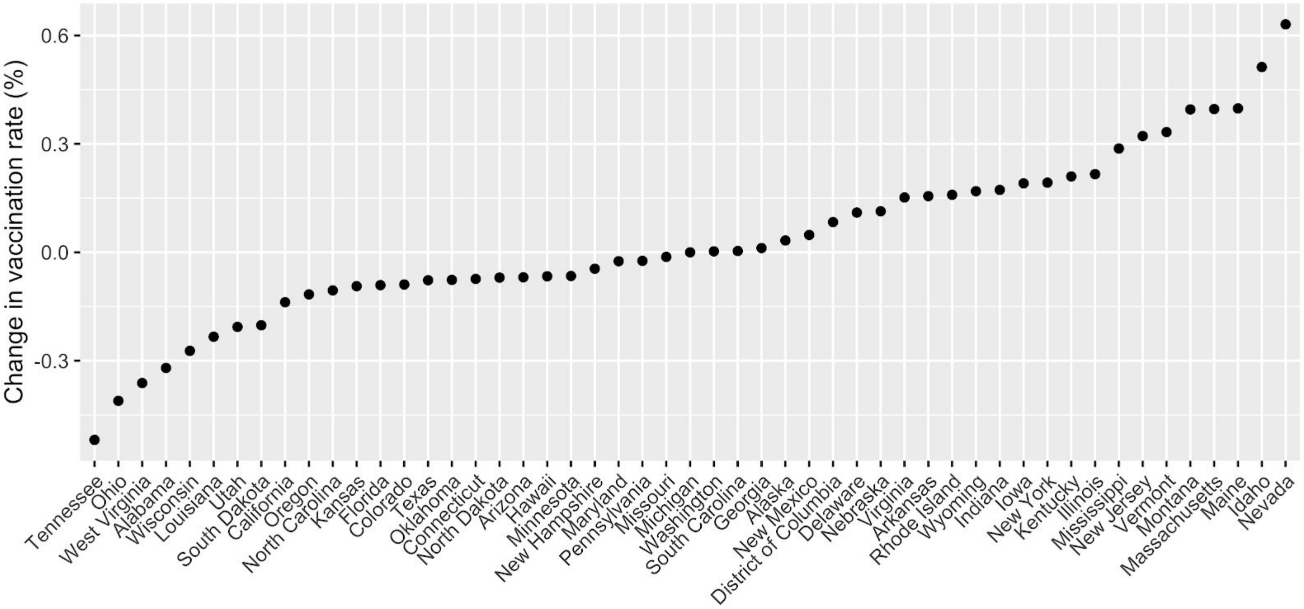


Figure 2. Percentage Change in Vaccination Rates by State (2007 - 2017). From decreasing to increasing (left to right), the figure shows the trend in vaccination rates over the period of 2007-2017. States with negative change indicates that their vaccination rates have fallen over the time period and vice versa.

Challenges/Further Research

- CDC used small sample sizes to determine vaccination rates (n=100-700), which led to high variance
- Lack of state-level polling data for public perception of vaccines and medical professionals in general
- Difficult to find data for the same time frames
- Data was often in strange formats, or in nice infographics without the actual dataset
-

Bibliograph/References/Works Cited (APA)

All Timelines Overview. www.historyofvaccines.org/timeline.

Helft, L. (2014, September 5). *The Autism-Vaccine Myth*. PBS. https://www.pbs.org/wgbh/nova/article/autism-vaccine-myth/.

Novak, S. (2020, April 28). *The Long History of America's Anti-Vaccination Movement*. Discover Magazine. https://www.discovermagazine.com/health/the-long-history-of-americas-anti-vaccination-movement.

Quotes of fear, defiance and hope as the coronavirus pandemic spans the globe. (2020, June 28). https://www.reuters.com/article/health-coronavirus-quotes-factbox/quotes-of-fear-defiance-and-hope-as-the-coronavirus-pandemic-spans-the-globe-idINKBN23Z058.

Data Sources:

- MMR Vaccination Rate by State (vaccination_19_35months_exemption.csv, vaccination_not_poverty.csv, vaccination_poverty.csv) https://www.cdc.gov/vaccines/imz-managers/coverage/childvaxview/data-reports/mmr/trend/index.html

- State Vaccination Exemptions (combined with vaccination dataset in vaccination_19_35months_exemption.csv)
  https://www.ncsl.org/research/health/school-immunization-exemption-state-laws.aspx

- Educational Attainment by State (educational_attainment.csv)
  https://www.census.gov/newsroom/releases/xls/cb12-33table1states.xls

- State Health Budgets (healthExpendituresState.csv)
  https://www.kff.org/other/state-indicator/health-care-expenditures-by-state-of-residence-in-millions/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

- Population by State (population_by_state.csv)
  https://www.infoplease.com/us/states/state-population-by-rank

- Median Household Income by State (medianHouseholdIncome.csv)
  https://www2.census.gov/programs-surveys/cps/tables/time-series/historical-income-households/h08.xls

- State Education Spending per Pupil (EducationSpending.csv)
  https://www.governing.com/gov-data/education-data/state-education-spending-per-pupil-data.html

- Population Density by State (populationDensity.csv)
  https://worldpopulationreview.com/state-rankings/state-densities