Deliverable 2 Report

1.1 Analysis Questions

- Questions regarding different factors that may affect vaccination rates (specifically MMR vaccination rates)
    - Education attainment
        i. For this, we differentiate education levels by looking at high school (and higher) completion percentages and undergraduate (and higher) completion percentages
    - Total spending per pupil
    - Total spending per capita on health
    - Population density
    - Exemption regulations
        i. Although this is categorical data, R regression packages actually allows it as an input and handles it automatically in linear models
    - Median household income
- Are specific states/groups of states trending in similar/different directions?
- Are any of the above factors also relevant for states/groups of states trending in a direction?

1.2 Findings & Results

Most of our initial efforts were spent cleaning and merging the datasets. This data wrangling and cleaning was done in R (as was the analysis). 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 section 1.1).

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)**.** *Running the regression revealed that two of the factors were significant* (correlation with MMR vaccination rates): spending per capita (health expenditures) and bachelor's and higher education attainment. 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, required us to test for multicollinearity. None or average multicollinearity 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 the figure below 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.

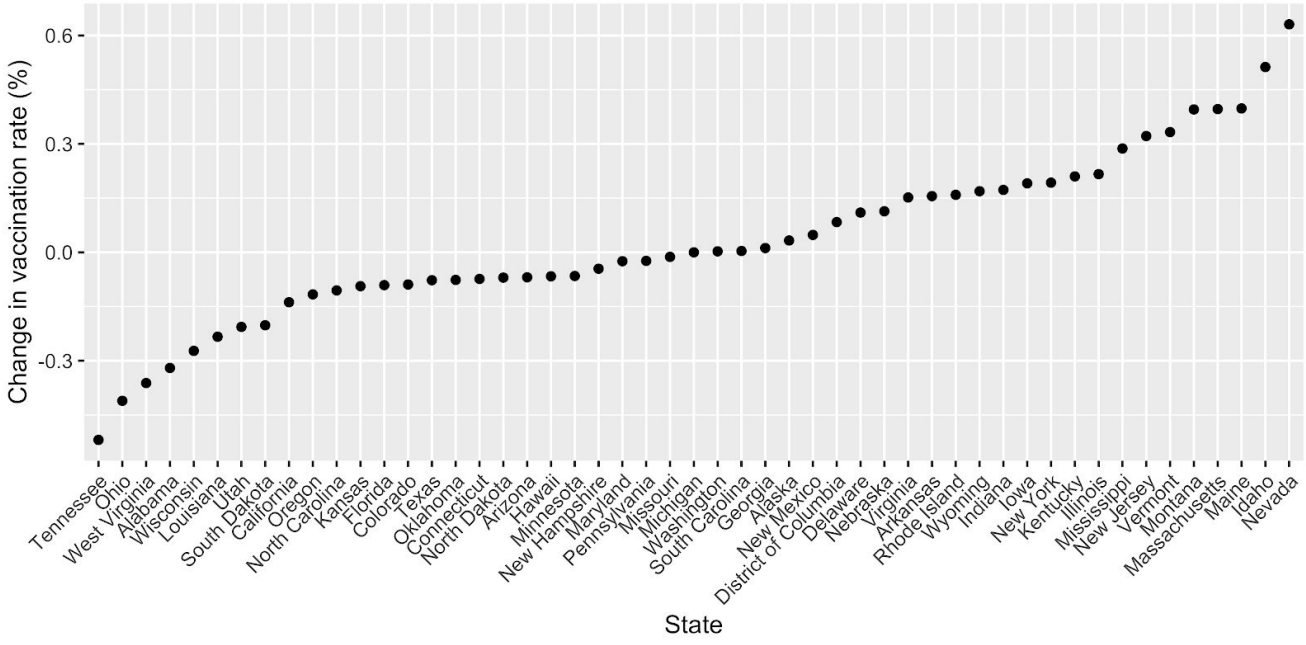

Figure 1. Percentage Change in Vaccination Rates by State (2007 - 2017).