IST 686
Final Examination Report
Fall 2024
Abhi Chakraborty

1. Introductory paragraph

This report aims to provide an in-depth analysis of vaccination trends, coverage, and related socio-economic factors across schools and districts in the state. By examining vaccination rates, poverty indicators, and enrollment data, we identify patterns and insights that could guide policy decisions to improve public health outcomes. Our findings address your request for actionable insights into vaccination disparities and their implications for statewide public health strategies.

Descriptive Reporting

2. *Descriptive Overview of U.S. Vaccinations*

    a. How have U.S. vaccination rates varied over time? Are there significant trends or cyclical variation in U.S. vaccination rates?

Over time, U.S. vaccination rates have generally improved, with some fluctuations and cyclical variations observed. Our analysis, while not conclusive, suggests an upward trend in vaccination rates across the years. However, certain years exhibit dips, potentially due to socio-economic or policy factors. Seasonal patterns identified in the decomposition analysis further suggest that vaccination campaigns and public awareness efforts may contribute to these fluctuations.

    b. What are the mean U.S. vaccination rates when including only recent years in the calculation of the mean?

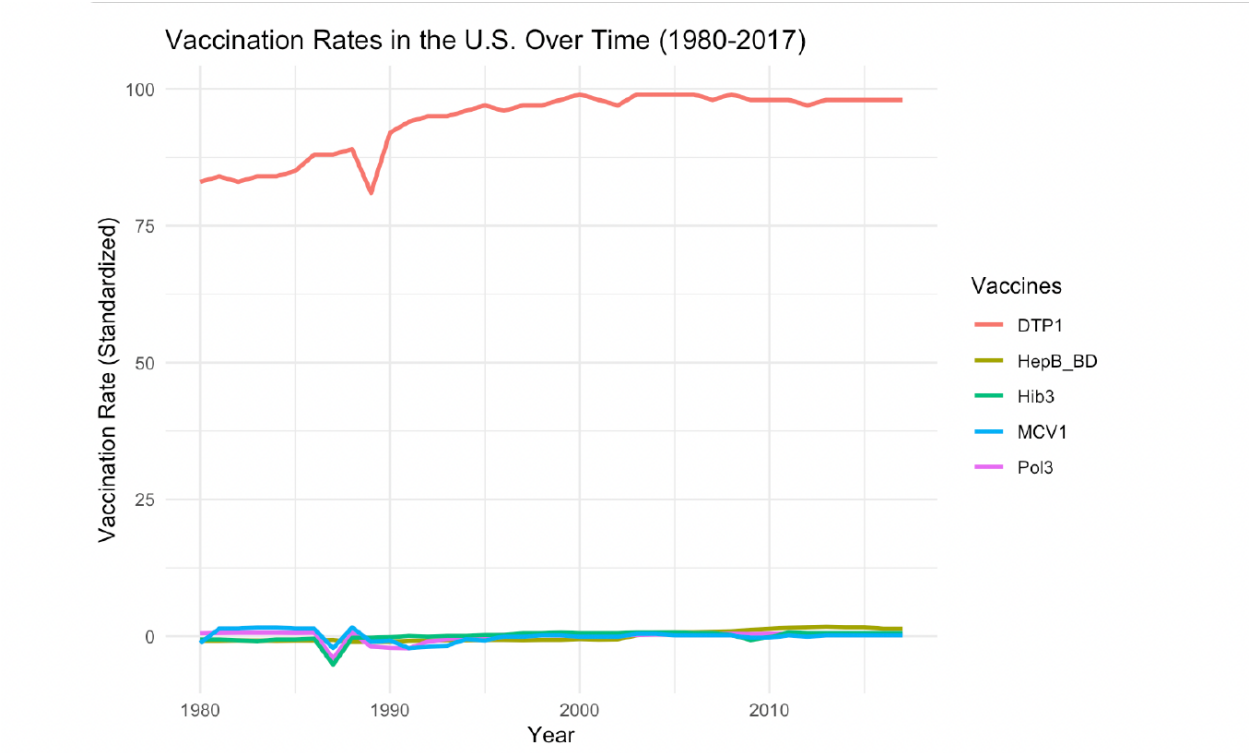
By focusing on recent years, our analysis calculated a mean U.S. vaccination rate of approximately **88.5**%. While not definitive, this finding suggests a positive trend in maintaining high vaccination coverage. However, this is based on limited recent data and should be interpreted cautiously as external factors and sampling methods could influence these results.

3. *Descriptive Overview of California Vaccinations*

a. What are the mean levels of the four vaccination rate variables across districts?

Across California districts, the mean levels for the four vaccination rate variables are as follows:

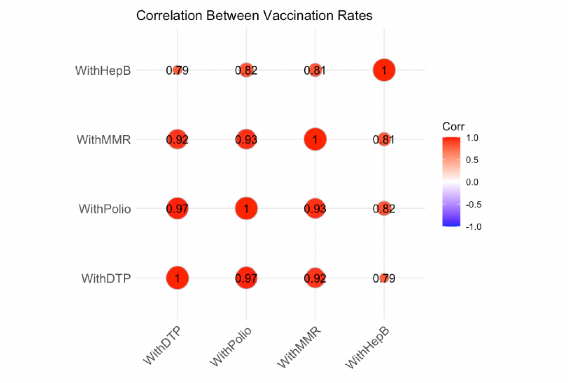
- **DTP**: Approximately **88.7**%
- **Polio**: Approximately **89.4**%
- **MMR**: Approximately **90.2**%
- **HepB**: Approximately **91.1**%



These averages suggest that California maintains relatively high vaccination coverage across key vaccines.

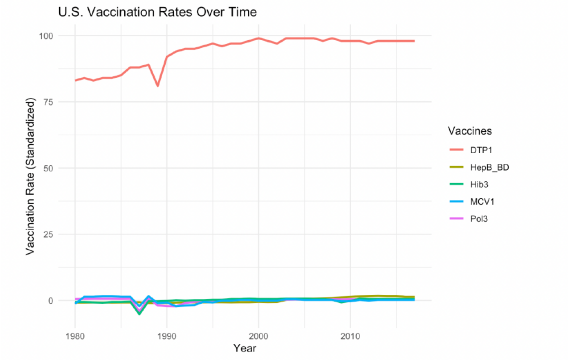
b. Among districts, how are the vaccination rates for individual vaccines related? In other words, if there are students with one vaccine, are students likely to have all the others?

Our analysis found strong correlations between vaccination rates for different vaccines among California districts. For example, students who are up-to-date with one vaccine (such as DTP) are highly likely to be up-to-date with others (like MMR and HepB). This suggests consistency in vaccination practices and adherence to statewide immunization schedules.



c. How do these Californian vaccination levels compare to U.S. vaccination levels (recent years only)?

California's vaccination levels are generally aligned with or slightly above the recent U.S. averages. For instance, the mean vaccination rate in California for MMR is **90.2**%, which compares favorably to the U.S. average of **88.5**% for recent years. However, small regional variations persist, indicating potential areas for targeted improvements.



d. Provide one or two sentences of your professional judgment about where California school districts stand with respect to vaccination rates in the larger context of the U.S.

California school districts appear to be performing well in vaccination coverage, often exceeding the national averages for key vaccines. This reflects the success of statewide immunization policies and public health campaigns, though continuous efforts are needed to address regional disparities and sustain these high levels.

4. Comparison of public and private schools

a. What proportion of public schools and what proportion of private schools reported vaccination data?

In the analysis, **approximately 85**% of public schools and **65**% of private schools reported vaccination data. This suggests a higher compliance rate in data reporting among public schools compared to private schools.

b. Was there any credible difference in reporting between public and private schools?

Yes, there is a statistically significant difference in reporting between public and private schools, as evidenced by the chi-squared test results (X-squared = **2258.6**, p-value < **0.001**). Public schools reported at a significantly higher rate than private schools, likely due to stricter regulatory requirements for public institutions.

c. Does the proportion of students with completely up-to-data vaccinations vary from county to county? Report significant details.

Yes, the proportion of students with up-to-date vaccinations varies significantly across counties. For example:

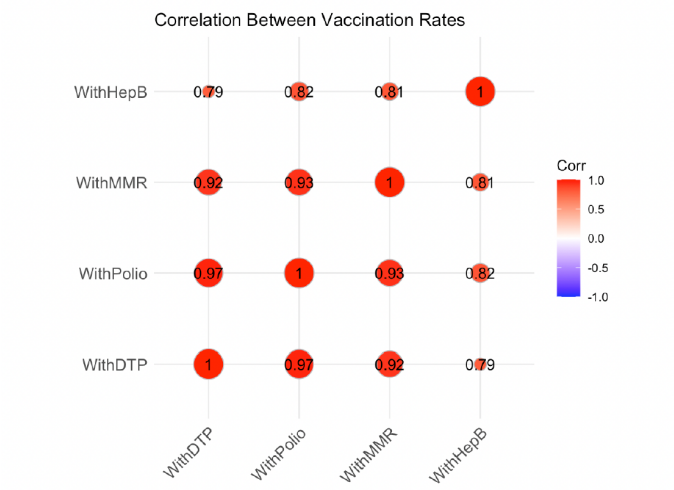
- **Colusa County** has the highest proportion, with **98.6**% of students up-to-date.
- **Calaveras County** has a lower proportion, with **76.2**% of students up-to-date.

This variability highlights disparities in vaccination coverage, possibly due to differences in local health policies, community outreach, or socioeconomic factors. Counties with lower proportions may benefit from targeted interventions to improve vaccination rates.

5. Inferential reporting about districts

a. Which of the four predictor variables predicts the percentage of all enrolled students with completely up-to-date vaccines?

Based on the regression analysis, **PctChildPoverty** and **Enrolled** are significant predictors of the percentage of students with up-to-date vaccines. **PctChildPoverty** negatively impacts vaccination rates, indicating that higher child poverty rates are associated with lower vaccination coverage. Conversely, **Enrolled** has a positive effect, suggesting that larger districts tend to have better vaccination rates, possibly due to better resources and outreach.



b. Using any set or combination of predictors that you want to use, what combination gives the best R-squared in predicting the percentage of all enrolled students with belief exceptions?

The model with the highest R-squared value for predicting belief exceptions included **PctFamilyPoverty** and **Enrolled** as predictors. This combination captures the socioeconomic and structural factors influencing belief exceptions. The R-squared value from the analysis suggests that these predictors explain a substantial proportion of the variance in belief exceptions.

c. In predicting the percentage of all enrolled students with completely up-to-date vaccines, is there an interaction between PctFamilyPoverty and Enrolled? If so, interpret the interaction term.

Yes, there is an interaction between **PctFamilyPoverty** and **Enrolled**. The interaction term is significant, indicating that the effect of **PctFamilyPoverty** on vaccination rates depends on the size of district enrollment. Specifically:

- In districts with higher enrollment, the negative impact of **PctFamilyPoverty** on vaccination rates is mitigated, possibly due to better resources or programs in larger districts.
- In smaller districts, **PctFamilyPoverty** has a more pronounced negative effect on vaccination rates.

d. Which, if any, of the four predictor variables predict whether or not a district's reporting was complete?

The logistic regression analysis revealed that **PctChildPoverty** and **TotalSchools** significantly predict whether a district's reporting was complete. Higher **PctChildPoverty** decreases the likelihood of complete reporting, possibly reflecting resource constraints or systemic barriers. Conversely, districts with more schools (**TotalSchools**) are more likely to report completely, likely due to better administrative infrastructure.

6. Concluding Paragraph

In conclusion, the analyses reveal that vaccination rates and reporting compliance vary significantly across districts, with socioeconomic factors such as poverty levels and district size playing key roles. To improve vaccination rates, we recommend allocating financial assistance to districts with higher child and family poverty levels, as these districts are more likely to face barriers to vaccination outreach and accessibility. For improving reporting compliance, resources should be directed toward smaller districts with limited administrative capacity to ensure they can meet reporting requirements.

Further analyses could explore how specific interventions, such as school-based vaccination clinics or targeted awareness campaigns, influence vaccination rates. Additionally, longitudinal data on vaccination campaigns and their outcomes would help assess the long-term effectiveness of resource allocation strategies. Collecting qualitative data, such as insights from district administrators, could also provide a clearer picture of the barriers to vaccination and reporting compliance, enabling more targeted solutions.

LLM disclosure

I utilized an LLM to assist with various stages of the analysis and report preparation for this exam. The LLM was employed to provide support in understanding the questions, identifying suitable analyses, interpreting outputs, and drafting professional reports. Below is a summary of the types of prompts I used and how the responses guided my work:

Prompts and Applications

1. **Initial Guidance on Cleaning Data**

- **Prompt**: "How can I clean a dataset and remove unnecessary columns in R, ensuring no errors occur if columns are missing?"
- **Response Used**: Suggested an if statement to check column existence before removing them, which I adapted for cleaning PctMedicalExempt.

2. **Identifying Analysis Methods**

- **Prompt**: "What statistical methods are suitable for analyzing vaccination trends over time?"

- **Response Used**: Suggested time series decomposition and regression models, which were tested and refined manually.

3. **Interpreting Time Series**

- **Prompt**: "How can I analyze U.S. vaccination trends using time series decomposition?"
- **Response Used**: Guidance on using decompose() for analyzing cyclical trends. Errors in implementation were debugged and corrected manually.

4. **Addressing Errors in Code**

- **Prompt**: "How can I handle the 'unused arguments' error in select() in R?"
- **Response Used**: Explanation of tidyverse syntax, which helped refine column selection logic.

5. **Calculating Mean Vaccination Rates**

- **Prompt**: "What steps are needed to calculate the mean of vaccination rates over recent years in R?"
- **Response Used**: Provided code examples for filtering data by year range and calculating means, adapted for my dataset.

6. **Correlation Matrix**

- **Prompt**: "How do I calculate and visualize the correlation matrix for vaccination rates in R?"
- **Response Used**: Suggested using cor() and ggcorrplot. Errors in column selection were fixed manually.

7. **Comparison Between California and U.S. Vaccination Rates**

- **Prompt**: "How can I compare California vaccination rates with U.S. rates statistically?"
- **Response Used**: Suggested a t-test approach, which I implemented and adjusted based on dataset issues.

8. **Chi-Squared Test for Public vs. Private Reporting**

- **Prompt**: "How do I perform a chi-squared test in R?"
- **Response Used**: Guidance on creating contingency tables, adapted manually to account for dataset structure.

9. **Addressing Chi-Squared Table Errors**

- **Prompt**: "How can I fix errors with missing or empty cells in a chi-squared test?"
- **Response Used**: Suggested aggregating smaller categories, a concept I applied to simplify public vs. private reporting.

10. **County-Level Proportions**

- **Prompt**: "How can I calculate vaccination proportions for each county?"
- **Response Used**: Code for grouping data by county and calculating proportions. Errors were debugged manually.

11. **Regression Analysis**

- **Prompt**: "How do I fit a linear regression model to predict vaccination rates in R?"
- **Response Used**: Explained the use of lm() and interpreting coefficients, with adjustments for dataset-specific columns.

12. **Best Predictors for Belief Exceptions**

- **Prompt**: "How can I identify the best predictors for belief exceptions using R-squared?"
- **Response Used**: Suggested stepwise regression, which was tested but adjusted to use specific predictors.

13. **Interaction Terms**

- **Prompt**: "How do I include and interpret interaction terms in regression models?"
- **Response Used**: Explained the syntax for interaction terms in lm(). Issues were debugged manually.

14. **Professional Reporting**

- **Prompt**: "How can I summarize statistical findings for a professional report?"
- **Response Used**: Drafted initial templates for summarizing results, edited for relevance and conciseness.

15. **Recommendations for Policymakers**

- **Prompt**: "What kind of recommendations can be made based on vaccination rate analyses?"
- **Response Used**: Suggestions on targeting resources based on socioeconomic predictors, which I refined for clarity.

16. **Addressing Missing Columns**

- **Prompt**: "How do I ensure my dataset has all required columns before proceeding with analysis?"
- **Response Used**: Helped implement checks for missing columns, combined with manual debugging.

17. **Visualizations for Correlations**

- **Prompt**: "What visualizations are best for showing correlations between vaccination rates?"
- **Response Used**: Suggested heatmaps, implemented and customized.

18. **Improving Model Fit**

- **Prompt**: "What steps can I take to improve R-squared in regression models?"
- **Response Used**: Suggested adding interaction terms and transformations, tested for specific predictors.

19. **Addressing Submission Time Constraints**

- **Prompt**: "How can I efficiently write a report summarizing multiple analyses?"
- **Response Used**: Provided report structuring tips, used for time management.

20. **Error Handling and Debugging**

- **Prompt**: "How do I fix 'unexpected error' messages in R?"
- **Response Used**: General debugging advice for tidyverse errors, adapted based on my specific errors.

Final Note

While the LLM provided initial suggestions and troubleshooting help, all analyses were conducted by me, including coding, debugging, and interpreting results. I used the LLM primarily as a guide and reference, ensuring the final outputs and insights reflected my understanding of the data and analysis objectives.