MAVEN Project

Jackson Isidor, Grisha Dekhtyar, Mayumi Paraiso, Khoa Dang

November 2023

Abstract

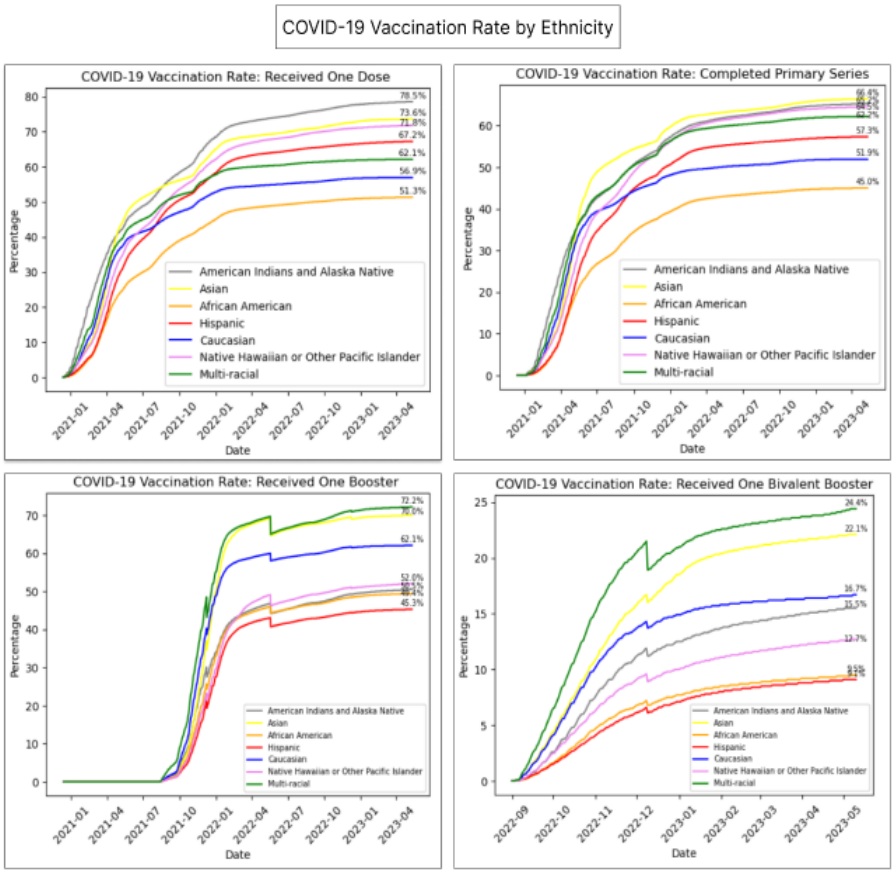
The MAVEN (Multidisciplinary Analysis of Vaccination Games for Equity) project addresses the global health threat of vaccine inequity in the fight against emerging infectious diseases. This project aims to provide a comprehensive understanding of vaccination coverage and identify key drivers of vaccine uptake. Our report investigates several of these potential factors including ethnicity, geography, and socioeconomic status, which can be used to reduce the risk of future pandemics by enabling targeted interventions where resources and information may not be as accessible. Following this analysis is the construction of a forecasting model that aims to track future trends ahead of time so that institutions, and society as a whole, can be better prepared during a pandemic.

1 The Relationship Between Ethnicity and Vaccination Uptake

The main goal of the research paper is to identify the factors that lead to the uptake of the COVID-19 vaccine. In the United States, there has been a history of racial conflicts that create barriers for historically marginalized groups, which may affect their COVID-19 vaccine uptake. To further investigate this potential issue, we analyzed ethnicity in the United States and its relationship with vaccination uptake.

1.1 Visualization of Doses by Ethnicity

Using information from the Centers for Disease Control and Prevention (CDC) [1], we plotted time-series vaccination data in the United States across various demographics with specific information on the type of dose from Dec. 2020 to May 2023.



1.1.1 Visualization Analysis

To understand the influence of ethnicity on vaccination uptake, we broke down the COVID-19 vaccination rate into four main doses: received the first dose, completed the primary series, received one booster, and received one bivalent booster. Breaking down the data into these four doses will help us understand how vaccination rates change over time, as the pandemic develops. By analyzing these graphs, some insightful patterns suggest that African Americans struggled with their vaccination uptake rates throughout the pandemic across four types of doses. Over time, it is apparent that vaccination uptake rates of Hispanic Americans drop significantly in comparison with other ethnic groups.

1.2 Vaccination Hesitancy Amongst Ethnic Groups

Carnegie Mellon University published a data set on the Data for Good website run by Meta [4], in which they recorded COVID-19 vaccination hesitancy in the United States. We sought to use this information to find if such hesitancy levels vary between ethnic groups.

1.2.1 Methodology

There are two variables we are interested in when comparing ethnicities:

1. Vaccine Acceptance: Percent who would definitely or probably choose to get vaccinated.

2. Vaccine Hesitancy: Percent who would definitely or probably NOT choose to get vaccinated.

Given that the survey data is presented as proportions across various ethnic groups, the most appropriate statistical analysis tool would be the Chi-Squared Test for Independence. We will conduct this test on two separate dates, 7 months apart, for each variable. The formula for the chi-squared statistic is as follows:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

$$O_i = \text{observed value in cell } (i),$$
$$E_i = \text{expected value in cell } (i).$$

The formula calculates whether there is a significant association between the two categorical variables, in this case, acceptance or hesitation association with ethnicity. The statistic can be used to calculate a p-value for the test, which indicates how unusual the observed proportions are if the groups are truly equal.

Next, we can calculate the squared residual to identify which groups contributed the most to the overall test statistic (i.e. the groups most different from the expected proportions).

$$\text{Squared Residual} = \frac{(O_i - E_i)^2}{E_i}$$

These hypotheses will guide our test:

- Null Hypothesis: There is no association between vaccination hesitancy/acceptance and ethnicity.
- Alternative Hypothesis: There is an association between vaccination hesitancy/acceptance and ethnicity.

Given sufficiently large group sample sizes and expected counts, while assuming independent observations, all of the conditions for a chi-squared test for independence have been met, and we can continue with the evaluation.

1.2.2 Results

The following results are the output for each chi-squared test conducted. There are four total tests - two investigating vaccine acceptance and two investigating vaccine hesitance. Both sets have one test on January 31, 2021 and one test on July 31, 2021.

- Acceptance Result: 01/31/2021
  - Chi-squared Statistics: 13627.041
  - P-value: 0.0
- Acceptance Result: 07/31/2021
  - Chi-squared Statistics: 1295
  - P-value: 6.9 * e^{-278}
- Hesitation Result: 01/31/2021
  - Chi-squared Statistics: 13618.276
  - P-value: 0.0
- Hesitation Result: 07/31/2021
  - Chi-squared Statistics: 1295
  - P-value: 6.9 * e^{-278}

1.2.3 Conclusion of Results

Since all of the Chi-squared Tests for Independence returned a p-value that is less than 0.05, we reject the null hypothesis. There is evidence of an association between ethnicity and both vaccine acceptance and hesitancy for both periods of January 2021 and July 2021.

1.2.4 Analysis

By examining the outcomes, we observe that the patterns in vaccination rate by ethnicity from section 1.1 are consistent with the findings from the chi-squared tests. Further analysis of differences between groups through squared residuals reveals a notable disparity between African Americans and other groups. African Americans exhibit a significantly higher observed percentage of vaccine rejection. This is evident in the more than 8,000 counts of "No" responses, surpassing the expected frequency. Being that this is the highest value of the squared residuals, we can conclude that African Americans had the greatest effect on the chi-squared statistic, meaning they were most different from other groups. Additionally, Asian Americans have a strong vaccination rate during the pandemic showcased by their high acceptance and low hesitation rate, which resulted in another large squared residual. Furthermore, Hispanic Americans have a profound number of vaccination acceptance, demonstrated by their adequate vaccination uptake rate for the first dose and completion of the primary series. This represents a difference in the other direction from African Americans, as Asian Americans are much more accepting of vaccines than other groups. These results closely follow the trends of the figures in section 1.1.

1.2.5 Limitations

While test conditions were met, a precaution worth noting is that we only have data available to conduct the chi-squared test from January 2021 to July 2021, so our results only apply to this specific time frame. The COVID-19 pandemic spans much longer, so our results can only be applied to the time frame given and should not be expanded further.

2 Vaccination Trends Among Californian Zip Codes

Using a large-scale dataset from the California Department of Public Health (CDPH), the data was processed, cleaned, and sub-divided COVID-19 for visualization and analysis. This data featured the vaccination statistics for every California zip code from January 2021 to August 2023.

2.1 Geographical Analysis

California is a state renowned for its diversity in many aspects. The state's geopolitical divide a key contributor to the diversity of the state. From the agricultural Central Valley, to the technological Silicon Valley, and the media juggernaut of Los Angeles, the various industries provide a diverse set of socioeconomic conditions statewide. Examining past studies [5], one of the most prevalent socioeconomic factors that contributed to vaccination uptake was latitude and longitude. Two of the most culturally geographically different regions of California is the Central Valley, spanning from Bakersfield, up north to Redding. This region is home 7.2 million people, and is an agricultural powerhouse

for the state and entire country. By contrast, the Central Coast is home to 2.3 million people, and is known for its viticulture and tourism industries. Taking some of the most populated zip codes from the main population centers from both of regions and plotting the total percent of their population that is vaccinated visualizes the individual vaccination progression for these communities.

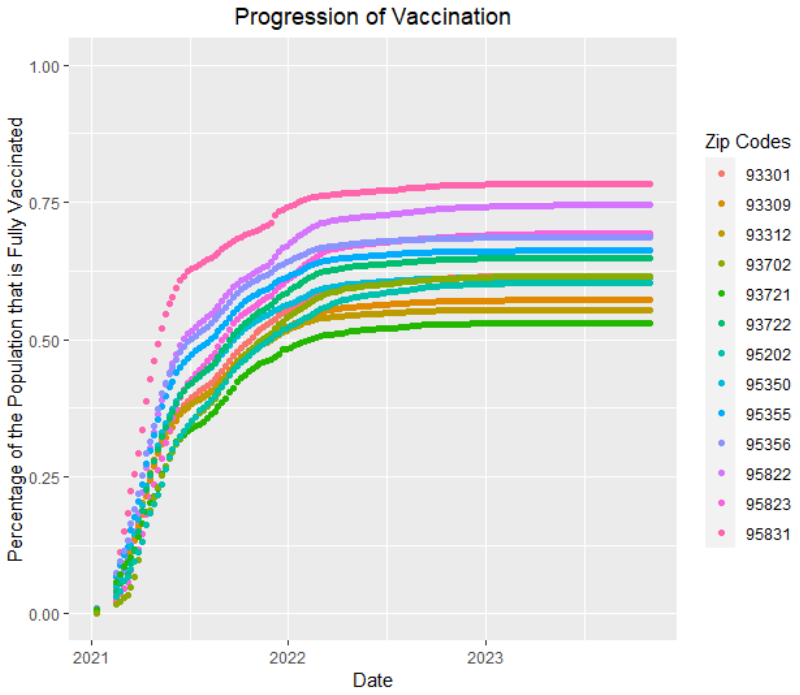


Figure 1: Time series visualization of COVID-19 Vaccination progress for several zip codes from the major population centers in the Central Valley, including Fresno (93721, 93722, 93702), Modesto (95355, 95350, 95356), Sacramento (95823, 95822, 95831), and Bakersfield (93301, 933090, 93312)

H

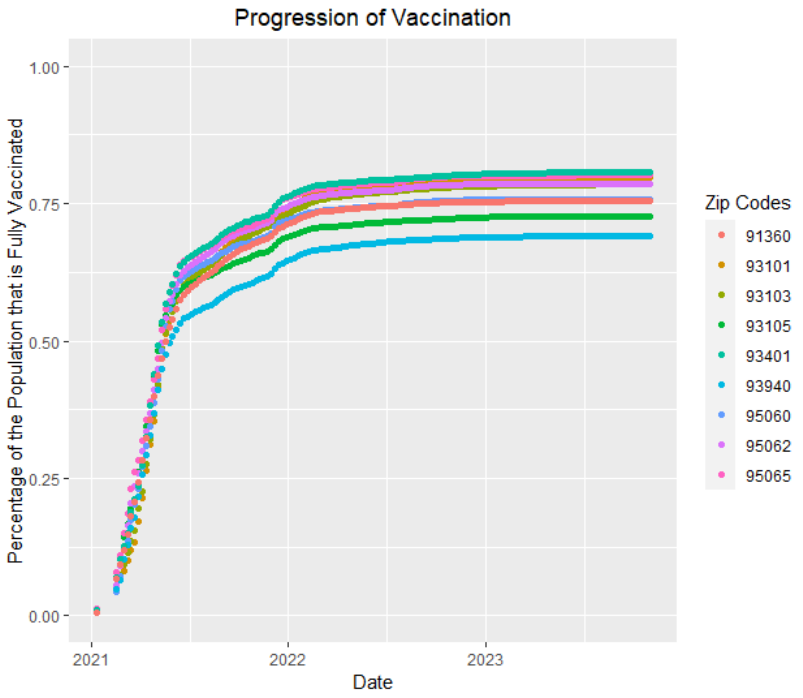


Figure 2: Time series visualization of COVID-19 Vaccination progress for several zip codes from the major population centers on the Central Coast, including Monterey(93940), San Luis Obispo (93401), Santa Barbara (93101, 93103, 93105), Santa Cruz (95060, 95062, 95065), and Thousand Oaks(91360)

This provides us with a good basis of certain trends that can be seen throughout the different regions. Visually, the Central Coast has less variability in its vaccination progression compared to the Central Valley. There are many socioeconomic differences between these two regions, and examining certain socioeconomic factors throughout the state can give insight on how they impact COVID-19 Vaccination.

2.2 Median House Hold Income

One of the many socioeconomic factors that is observed to cause an impact on COVID-19 vaccination progression is median household income. Median household income is a common predictor of relative wealth that a community holds, and that wealth disparity can also be a factor in regards to vaccination distributions. Examining the progression of which populations receive the vaccination in zip codes with high median household income compared to low median household income can reveal the magnitude of effect this disparity has on vaccination rate.

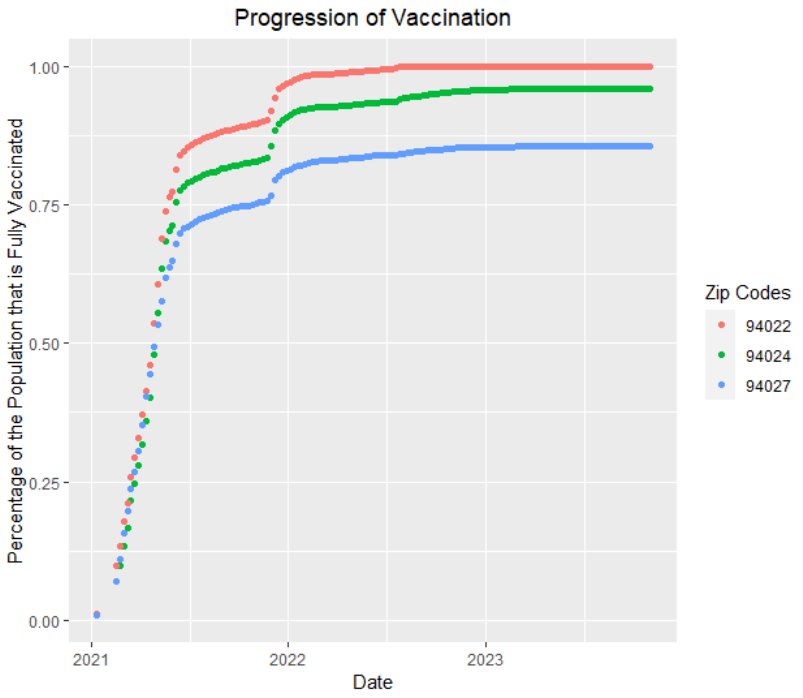


Figure 3: Time-series visualization of Covid-19 vaccination progress for the three zip codes with the highest household median income in California, including the cities of Los Altos: 94024 and 94022($220,970 and $208,984) and Atherton: 94027($250,001) [2]

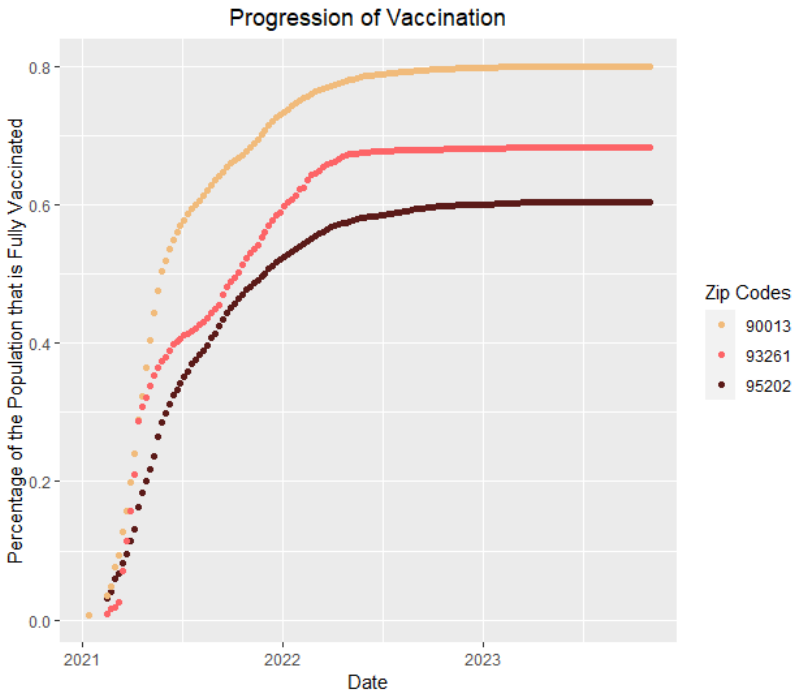


Figure 4: Time series visualization of COVID-19 Vaccination progress for the three zip codes with the lowest household median income in California, including the cities of Stockton: 95202($14,549), Los Angeles: 90013(2,808), and Richgrove: 93261 ($25,169) [2]

2.2.1 Analysis

Although this only a comparison of the two extremities of median household income in California, there is a clear indication that the zip codes located in areas were the median household income is higher corresponds to a higher percentage of the population being vaccinated. The zip codes with the highest median household incomes are all located in the Silicon Valley, while the zip codes with the lowest median household incomes are located throughout the state. Median household income is not a perfect predictor of vaccination progression amongst populations, as seen with zip code 90013, located in one of the poorest parts of Los Angeles(average household income of $22,808) [2], including parts of Skid Row, which contains one of the highest populations of homeless people in the United States. Despite this, the percentage of the population that is fully vaccinated evens out at about 80 percent, rivaling the zip codes located in the much wealthier Silicone Valley. This could be largely due to the fact that, being located in Los Angeles, the most populated city in the State, COVID-19 vaccination distributions were much more prevalent, especially compared to a similarly poor zip code such as 95202 which has a median household income of

$14,549 [2], located in Stockton, a significantly less wealthy major city.

2.3 Unemployment Rate

Another common predictor of relative wealth in a community is its unemployment rate. While median household income is commonly associated with geographical location, unemployment rate is a standard metric that correlates with the local economy. Higher unemployment rates can indicate stagnation in local economies, as well as lower disposable incomes among individuals, which can decrease the standards of living, making overall accessibility for health services potentially more difficult and unattainable, even with COVID-19 vaccinations being free.

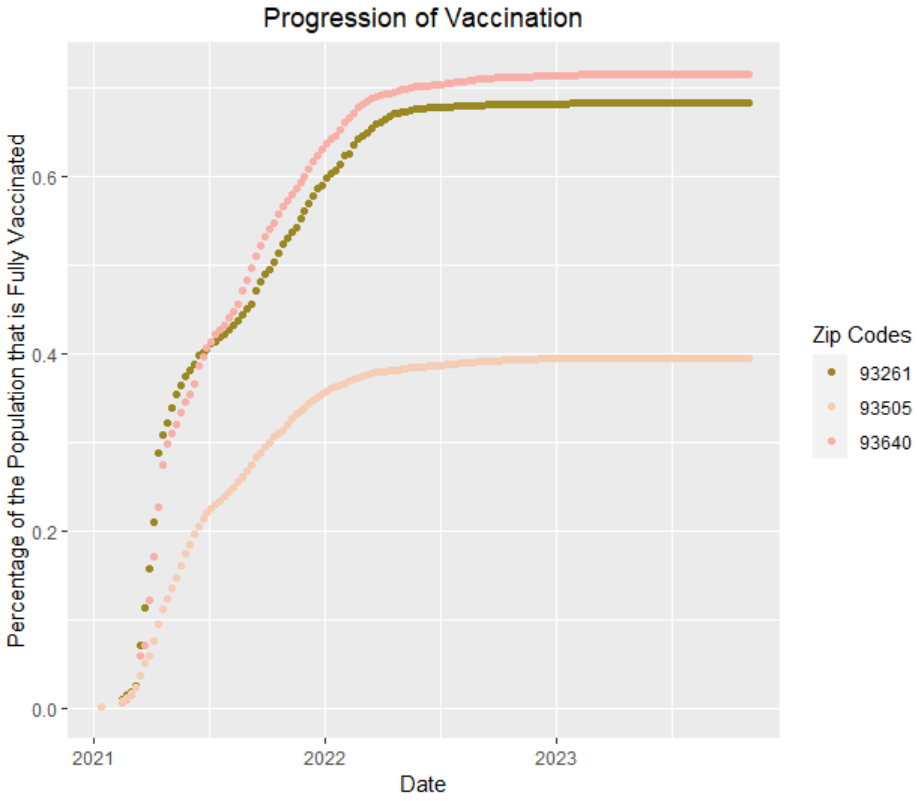


Figure 5: Time series visualization of Covid 19 Vaccination Progress for the three zip codes with the highest unemployment rates in California, including the cities of Richgrove :93261 (29.8%) , California City: 93505 (22.9%)and Mendota: 93640(22.9%) [3]

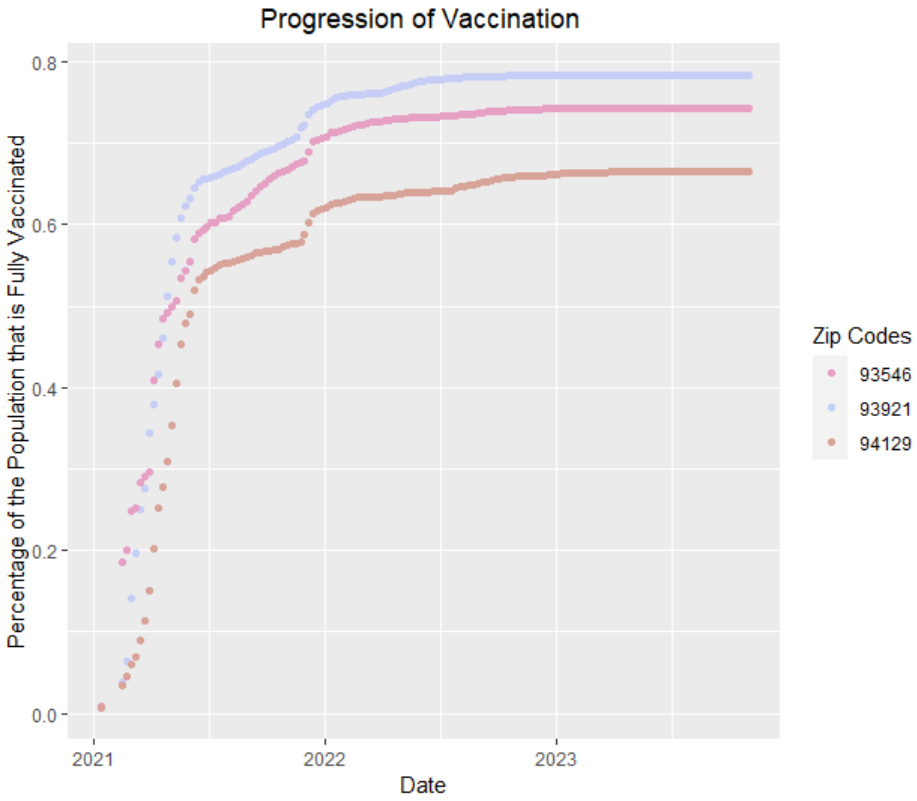


Figure 6: Time series visualization of Covid 19 vaccination progress for the threee zip codes with the lowest unemployment in California, including the cities of Carmel by the Sea: 93921 (0.1%), San Francsisco: 94129 (1.1%), and Mammoth Lakes: 93546(0.8%) [3]

2.3.1 Analysis

With unemployment rate being more of an independent metric from geographical location, it becomes more apparent what kind of relationship a zip codes economical situation has on its vaccination progression. An interesting pattern that can be seen with the zip codes with the lowest unemployment rate is that they are located in places that are of touristic relevance. The beach town of Carmel by the Sea, the popular ski destination of Mammoth Lakes, and the cultural landmark that is San Francisco creates jobs, which in turn helps further develop the surrounding area, making healthcare, and in turn, vaccine distribution to be more available. This is contrasted by the zip codes with the highest unemployment rate, being small, rural towns located in the valley or desert. Zip code 93505, encompassing most of California City is especially noteworthy, considering the failed history of development in the area especially. Something

interesting to note amongst the zip codes that tend to have lower household median income and high unemployment rate is that they tend to be located in or near the Central Valley. Going back to what was established previously, there are many geopolitical factors that differently effect different parts of California, and classifying these populations based on these factors can give us further insight on what causes vaccine uptake in populations. Using advanced metrics such as the Healthy Places Index score can help quantify all these socioeconomic differences and compare different populations.

2.4 Healthy Places Index Score

The Healthy Places Index (HPI) score is a comprehensive representation of health equity in various areas within the state of California. The score is a percentile relative to all other tracts in California, as a measure of how much healthier a particular tract's conditions are compared to all tracts in the state. This measure is a combined average of 23 factors, categorized into 8 domains: Economic, Education, Transportation, Social, Neighborhood, Housing, Clean Environment, and Healthcare Access. Factors include, but are not limited to: percentage of bachelor degree holders (or higher), percentage of registered voters who voted in the 2020 general election, and the percentage of households that have access to an automobile.

2.4.1 Relationship Between HPI and Vaccinations

HPI scores for zip codes (chosen tract) throughout California were utilized in conjunction with 1) percentage of fully vaccinated individuals, and 2) percentage of partially vaccinated individuals within the respective zip code, recorded on November 21, 2023.

Regression and best-subset analyses were performed to determine whether a significant correlation existed between HPI score and total vaccination rate, as well as HPI score and partial vaccination rate.

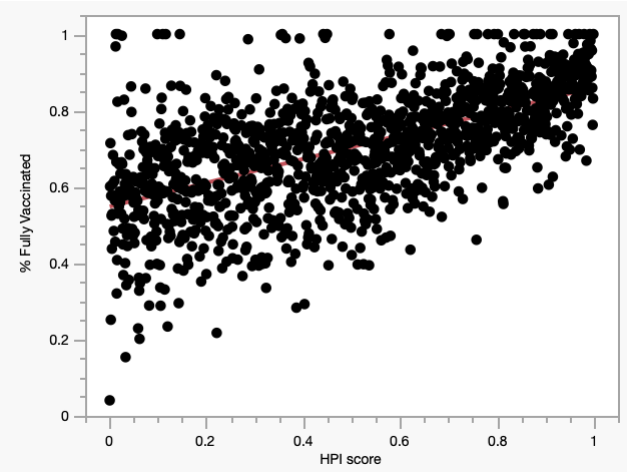


Figure 7: Regression plot of percentage of fully vaccinated people within a zip code by the corresponding HPI score for 1298 zip codes in California.

2.4.2 Analysis

The regression plot of percentage of fully vaccinated individuals vs. HPI score shows a moderate positive trend in the data points of 1298 zip codes in California.

The least squares regression model of HPI score to percentage of fully vaccinated individuals gives the predictor, HPI score, a t-ratio of 26.97 and a p-value less than 0.0001, indicating that HPI score is significantly positively associated with percent of fully vaccinated individuals. Furthermore, it indicates that HPI is a significant predictor of percent of fully vaccinated individuals.

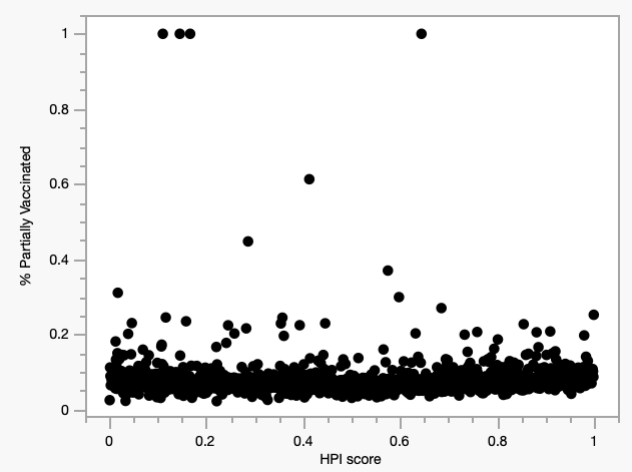


Figure 8: Regression plot of percentage of partially vaccinated people within a zip code by the corresponding HPI score for 1298 zip codes in California.

2.4.3 Analysis

The regression plot of percentage of partially vaccinated individuals vs. HPI score shows no trend.

HPI score has a t-score of -1.03, which indicates a weak negative association, but the p-value of 0.3012 represents insignificance of HPI score as a predictor of percentage of partially vaccinated individuals in a given zip code in California.

2.4.4 Relationship Between HPI Indicators and Vaccinations

In order to further explore HPI score and potential associations with COVID-19 measures like vaccination rate, the researchers broke down HPI score into its 8 indicators using data from the California Healthy Places Index HPI 3.0 Map. The scores recorded were percentile scores of a California zip code relative to all California zip codes for the following indicators: Economic, Education, Transportation, Social, Neighborhood, Housing, Clean Environment, and Healthcare Access.

A best-subset regression procedure was used to identify the best subset of indicators for the same COVID-19 measures: 1) percentage of fully vaccinated individuals, and 2) percentage of partially vaccinated individuals.

Best subsets were created by minimizing the Bayesian information criterion (BIC), which is a measure that optimizes parsimony and penalizes complexity, in forward stepwise selection. The subset with the smallest BIC was chosen as the best subset.

Summary of Fit

RSquare	0.496109
RSquare Adj	0.49455
Root Mean Square Error	0.106886
Mean of Response	0.704552
Observations (or Sum Wgts)	1298

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	14.543943	3.63599	318.2573
Error	1293	14.772102	0.01142	Prob > F
C. Total	1297	29.316045		<.0001*

Parameter Estimates

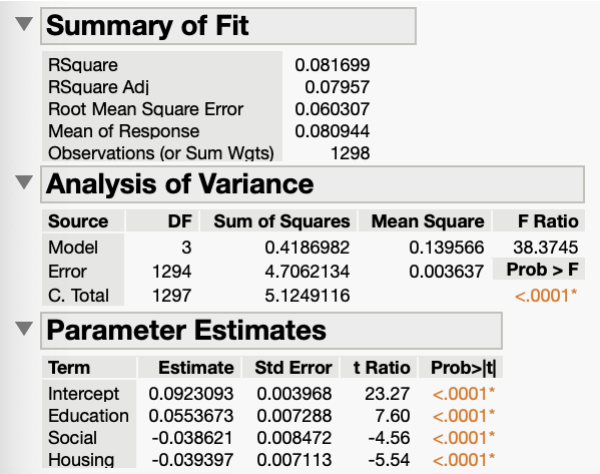
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.5756069	0.007032	81.85	<.0001*
Economic	0.3216087	0.020675	15.56	<.0001*
Education	0.0532518	0.016855	3.16	0.0016*
Social	0.0738216	0.017144	4.31	<.0001*
Housing	-0.19261	0.012702	-15.16	<.0001*

Figure 9: Regression output of best subset of indicators for percentage of fully vaccinated people within a zip code.

2.4.5 Analysis

Best-subset regression for the percentage of fully vaccinated individuals produced the subset with 4 indicators: Economic, Education, Social, and Housing. This subset scored a BIC of -2083.1.

Using this subset to find the least squares regression equation resulted in all predictors (Economic, Education, Social, and Housing) with p-values less than 0.002. Economic, Education, and Social had a positive t-value, and as such, a positive association with percentage of fully vaccinated individuals. Meanwhile, Housing had a negative t-value, indicating a negative association with the percentage of fully vaccinated individuals.



Summary of Fit

RSquare	0.081699
RSquare Adj	0.07957
Root Mean Square Error	0.060307
Mean of Response	0.080944
Observations (or Sum Wgts)	1298

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	0.4186982	0.139566	38.3745
Error	1294	4.7062134	0.003637	Prob > F
C. Total	1297	5.1249116		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0923093	0.003968	23.27	<.0001*
Education	0.0553673	0.007288	7.60	<.0001*
Social	-0.038621	0.008472	-4.56	<.0001*
Housing	-0.039397	0.007113	-5.54	<.0001*

Figure 10: Regression output of best subset of indicators for percentage of partially vaccinated people within a zip code.

2.4.6 Analysis

Best-subset regression for the percentage of partially vaccinated individuals produced a subset with 3 indicators: Economic, Social, and Housing. This subset scored a BIC of -3575.0.

Using this subset to find the least squares regression equation resulted in all predictors (Economic, Social, and Housing) with p-values less than 0.0001 across the board. Education had a positive t-value, indicating a positive association with the percentage of partially vaccinated individuals. Meanwhile, Social and Housing had negative t-values, indicating a negative association with the percentage of fully vaccinated individuals.

2.4.7 Relationship Between HPI Indicator Components and Vaccinations

In order to derive more insight into the factors associated with both full and partial vaccination percentage, the researchers used the individual factors that make up each domain as potential predictors for the response variables. As with the HPI score and domain, factors were scores as percentiles of all zip codes in California.

The same procedures were used as before. The indicators Economic, Education, Social, and Housing were split into the relevant factors listed below:

1. Economic (Employment Rate, Income per Capita, Above Poverty)
2. Education (Pre-School Enrollment, High School Enrollment, Bachelor's Education or Higher)
3. Social (2020 Census Response Rate, Voting in 2020)

4. Housing (Low-Income Renter Severe Housing Cost Burden, Low-Income Homeowner Severe Housing Cost Burden, Housing Habitability, Uncrowded Housing, Homeownership)

The 13 factors listed above were used in best-subset regression and least-squares regression in order to identify associations between the factors and percentage of fully vaccinated individuals and percentage of partially vaccinated individuals in a given zip code.

Summary of Fit

RSquare	0.294758
RSquare Adj	0.29472
Root Mean Square Error	0.165011
Mean of Response	0.629938
Observations (or Sum Wgts)	185014

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	10	2105.3930	210.539	7732.257
Error	185003	5037.3908	0.027	Prob > F
C. Total	185013	7142.7839		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.4643302	0.001381	336.28	<.0001*
hous_homeownership	0.045058	0.002243	20.09	<.0001*
hous_uncrowded	-0.148304	0.002222	-66.75	<.0001*
hous_own_severe	-0.054771	0.001503	-36.43	<.0001*
soc_census_response	0.0686568	0.001941	35.37	<.0001*
soc_voting	-0.037008	0.002545	-14.54	<.0001*
educ_preschool	0.0412467	0.001642	25.12	<.0001*
educ_highschool	0.0000433	8.875e-6	4.88	<.0001*
educ_bachelors	0.2755544	0.003033	90.84	<.0001*
econ_employed	0.1015549	0.002168	46.85	<.0001*
econ_above_poverty	0.0340895	0.003634	9.38	<.0001*

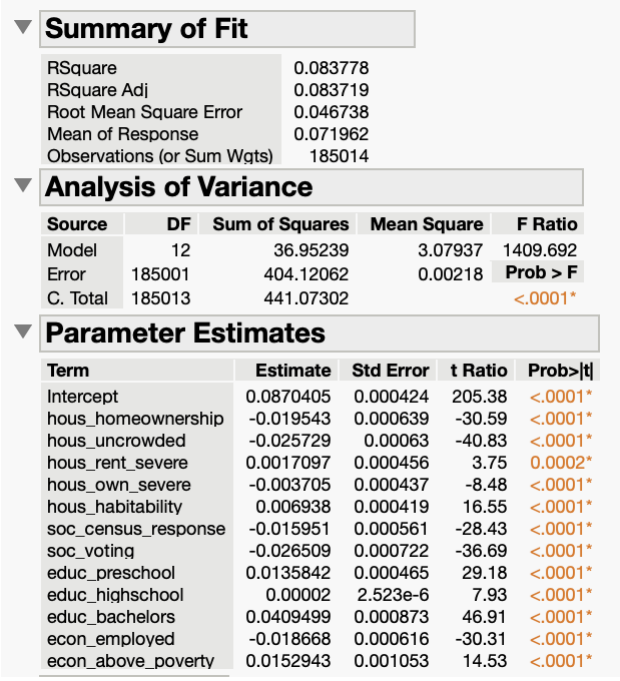
Figure 11: Regression output of best subset of factors per indicator (Economic, Education, Social, and Housing) for percentage of fully vaccinated people within a zip code.

2.4.8 Analysis

For percentage of fully vaccinated individuals, the best subset consists of 10 out of the 13 factors, namely: Employment Rate, Above Poverty, Pre-School Enrollment, High School Enrollment, Bachelor's Education or Higher, Census Response Rate, Voting, Low-Income Homeowner Severe Housing Cost Burden, Uncrowded Housing, and Homeownership. This subset obtained a BIC of -141513.

The corresponding least-squares regression output presented p-values of less than 0.0001 for all factors included. Factors with positive t-values are: Above Poverty, Employed, Preschool Enrollment, High School Enrollment, Bachelor's Education or Higher, 2020 Census Response, and Homeownership. Factors with

negative t-values are: Voting in 2020, Low-Income Homeowner Severe Housing Cost Burden, and Uncrowded Housing.



Summary of Fit

RSquare	0.083778
RSquare Adj	0.083719
Root Mean Square Error	0.046738
Mean of Response	0.071962
Observations (or Sum Wgts)	185014

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	36.95239	3.07937	1409.692
Error	185001	404.12062	0.00218	Prob > F
C. Total	185013	441.07302		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0870405	0.000424	205.38	<.0001*
hous_homeownership	-0.019543	0.000639	-30.59	<.0001*
hous_uncrowded	-0.025729	0.00063	-40.83	<.0001*
hous_rent_severe	0.0017097	0.000456	3.75	0.0002*
hous_own_severe	-0.003705	0.000437	-8.48	<.0001*
hous_habitability	0.006938	0.000419	16.55	<.0001*
soc_census_response	-0.015951	0.000561	-28.43	<.0001*
soc_voting	-0.026509	0.000722	-36.69	<.0001*
educ_preschool	0.0135842	0.000465	29.18	<.0001*
educ_highschool	0.00002	2.523e-6	7.93	<.0001*
educ_bachelors	0.0409499	0.000873	46.91	<.0001*
econ_employed	-0.018668	0.000616	-30.31	<.0001*
econ_above_poverty	0.0152943	0.001053	14.53	<.0001*

Figure 12: Regression output of best subset of factors per indicator (Economic, Education, Social, and Housing) for percentage of partially vaccinated people within a zip code.

2.4.9 Analysis

For percentage of partially vaccinated individuals, the best subset consists of 12 out of the 13 factors, all factors but Income per Capita. The subset obtained a BIC of -608267.

The corresponding least-squares regression output presented p-values of less than 0.0003 for all factors included. Factors with positive t-values are: Above Poverty, Bachelor's Education or Higher, High School Enrollment, Preschool Enrollment, Housing Habitability, and Low-Income Renter Severe Housing Cost Burden. Factors with negative t-values are: Employment Rate, Voting in 2020, 2020 Census Response, Low-Income Homeowner Severe Housing Cost Burden, Uncrowded Housing, and Homeownership.

2.4.10 Discussion

Interestingly, regression analysis shows that while HPI scores are significantly useful for predicting the percentage of fully vaccinated individuals, it is not

significantly useful for predicting the percentage of partially vaccinated individuals. Figure 2 seemed to indicate that the percentage of partially vaccinated individuals fall under a certain range of mostly below 20 percent regardless of how much healthier a community is compared to others, as indicated by HPI score.

Meanwhile, conducting regression analysis on the components of HPI score, from the domains to each of their factors, seems to provide more insight on relevant measures as they tie to COVID-19 metrics.

Scores used for every factor, domain, and HPI score are percentiles, and it is cautioned that negative and positive associations should not be taken to mean causation in that lower values of indicators lead to lower percentages in COVID vaccination rates and vice versa. Rather, zip codes with combinations of corresponding predictors, in which they perform worse or better relative to all other zip codes, tend to have higher percentages of fully vaccinated individuals/percentages of partially vaccinated individuals.

In terms of significance, these procedures may shed light on what factors should be taken into account when it comes to policy-making on a community-based level. Economic, social, education, and housing can be key observable factors when evaluating health equity with a COVID-19 vaccination lens.

It is recommended that further exploration of systemic factors relating to the observed factors be conducted. For example, what are the policies and culture surrounding zip codes that have higher burdens of housing on both renters and owners relative to other zip codes? Diving into the roots of multiple factors may provide even greater insight on the interplay of social, individual, and structural factors that affect vaccination behavior within communities.

3 Vaccination Forecasting Using ARIMAX

Utilizing a Google Health GitHub repository [6], we prepared and analyzed COVID-19 data for a time-series forecast of vaccination rates. The data used contains information on COVID-19, vaccinations, cases, government restrictions, and mobility (visits to places in community) in the United States.

3.1 Model Building Process

1. Selecting the appropriate model is the most important part of the process. We elected to use the ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables) model that uses auto-regression and external variable predictors to forecast trends. This is suitable for forecasting vaccination rates over time, while taking account for other variable trends at the same time.

2. The first step of constructing the model was to investigate correlations between factors and the target variable we want to predict (new daily vaccinations). In doing so, we found that confirmed cases, deaths, and the percentage change in visits to transit stations and residential areas

had the strongest associations with vaccinations. This gives us an initial indication that these variables will be valuable predictors in the model.

3. The next few steps involve checking model conditions, optimizing parameters, and splitting the data into training and testing sets.

4. Once all of these steps are performed thoroughly, the final step is to fit the model to the training data and make predictions with the testing data.

3.2 Results and Scores

The goal of the model building process and fitting is to provide a prediction that accurately follows the actual trend of vaccinations. There are many metrics used to evaluate such accuracy, including the R-squared value of 0.44, suggesting approximately 44 percent of the variance in vaccination trends is explained by the ARIMAX model and the chosen exogenous variables. This shows a moderate level of explanatory power in the model, while other metrics like log likelihood, AIC and BIC showed a stronger model fit.

3.3 How to Apply

Once a strong model is obtained, the goal is to apply it. After running the code, the model summary provides coefficients that can be used to construct an equation for the model. However, the equation can become complex and impractical to interpret or apply as the number of variables increase. For that reason, I stored the model in a Python variable named "results," which can be referenced later in the same notebook or saved and exported to any other system. This can be used to evaluate the model further or make predictions outside the training set.

3.4 Forecast Visualization

It is important to note that scores are not always a one-size-fits-all metric and require further interpretation through examining the Actual VS. Predicted plot. Using the stored model mentioned in the previous section, I made vaccination predictions between 2022-03-15 and 2022-07-01, then compared the predictions to the actual vaccination rates between those dates. By interpreting the figure below, we can properly evaluate the model's effectiveness and provide recommendations for usage.

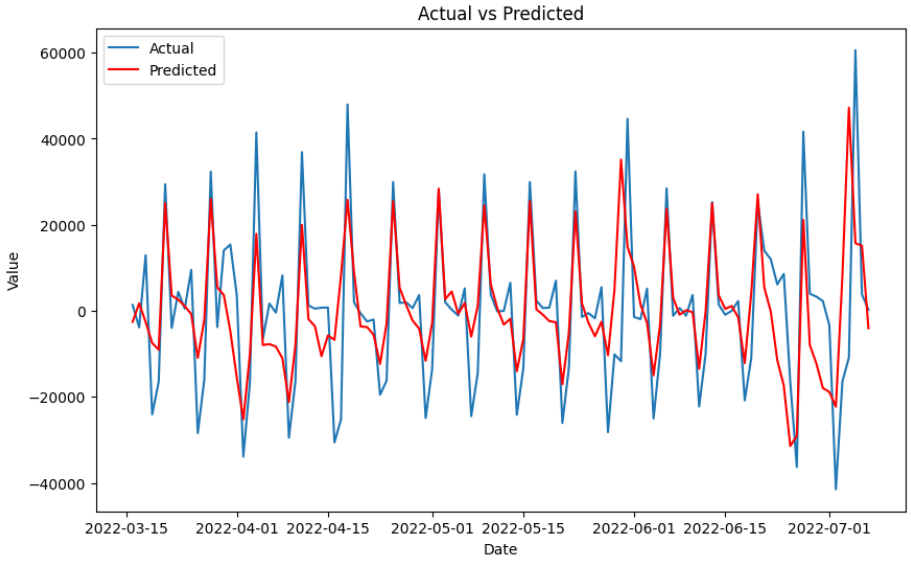


Figure 13: The blue line indicates new vaccinations between 2022-03-15 and 2022-07-01. The red line indicates the ARIMAX model predictions during the same time frame.

3.5 Performance

The model demonstrates a reasonable level of accuracy in forecasting vaccination numbers. Its predictions closely follow the actual trend, capturing most peaks and troughs in the vaccination rates. However, it's essential to note that while the model generally captures the overall trend, it generally underestimates the magnitude of the peak.

3.6 Utility and Recommendation

The predictive model can serve as a valuable tool for providing insights into the expected trends in vaccination rates. Its ability to factor in COVID-19 cases, fatalities, and mobility data makes it a conservative but reliable predictor.

Practical Use Cases:

- Indicative Information: It can provide guidance on when vaccination rates might rise or fall based on the observed variables.

- Planning and Resource Allocation: Offers insight into potential high-demand periods for vaccines, aiding in logistical planning and resource allocation.

- Policy Decisions: Helps policymakers anticipate and prepare for fluctuations in vaccination rates in response to changing pandemic conditions.

References

[1] Centers for Disease Control and Prevention. CDC Covid Data Tracker, 2023. Accessed 5 Dec. 2023.

[2] United States Zip Codes.org. Highest household income zips in california, 2023.

[3] United States Zip Codes.org. Lowest unemployment zips in california, 2023.

[4] Data For Good at Meta. Meta Covid-19 Trends and Impact Survey, n.d. Accessed 5 Dec. 2023.

[5] Cheong Q et al. Predictive modeling of vaccination uptake in us counties: A machine learning-based approach. *J Med Internet Res*, 2021.

[6] O. Wahltinez et al. Covid-19 open-data: curating a fine-grained, global-scale data repository for SARS-CoV-2. 2020.