Manuscript for a Data Analysis Project

Katie Wells & Connor Ross

4/19/23

# 1. Summary/Abstract

This article seeks to examine the popularity of different COVID-19 vaccine manufacturers in the year 2021 in the U.S. and to develop a model in understanding what potential factors are more associated with increase vaccine administration. Understanding the popularity differences across manufactures and the profiles of the different region within the U.S. could potentially explain differences in vaccine uptake. While vaccine distribution and administration information is updated every week, other potential drivers of increase vaccine administration across the country are not updated as frequently. For this reason, we limited our scope to 2021 so as to include data relating to these variables. After cleaning and joining the necessary datasets, several machine learning models were generated to ascertain the best comparative performance to explain the differences in vaccine administration. Given the complexity of the dataset and incompatibility of others containing region characteristics we were interested in, our findings were limited. However, modeling indicates that region, particularly in the West, is a major indicator of vaccine administration. More research and consistent data is needed to target the major factors of vaccine uptake. Furthermore, looking into the characteristics and behaviors of those living in the Western region of the U.S. states may lead to interventions to promote vaccine uptake which can reduce the transmission of COVID-19 and severity of disease.

# 2. Introduction

## 2.1 General Background Information

A little over three years ago, the world was brought to its knees by a virus no one had heard of before. SARS-CoV-2, which causes the disease COVID-19, ravaged the world and caused millions of deaths (CDC, n.d.-a). Scientists and researchers worked tirelessly during 2020 to develop vaccines to stop the spread and lessen the symptoms, and by the end of the year emergency use authorizations were given to administer the first immunizations.

After the first vaccines were approved, a rollout began across the United States to distribute the new immunizations. In the beginning, priority was given to healthcare workers, long-term care facility residents, essential workers, those with high-risk conditions, and older adults. Distribution then began to other groups as supply warranted. Throughout 2021 and 2022, immunizations continued to be distributed to each state through health departments, pharmacies, doctor’s offices, and many other settings (Dooling et al., 2021). Distribution did not always equal administration, however, and many factors may have affected whether or not someone received a vaccine.

Income is an integral predictor of health outcomes; previous studies have shown that there are large health differences between high and low income groups (Zimmerman & Anderson, 2019). Studies have also shown that low income is associated with low vaccine uptake (Kawai & Kawai, 2021). Low income areas are likely to have less access to healthcare facilities, which may contribute to these numbers.

To investigate if any outside factors affected administration, we are going to compare region groups and median income data for each state to administration numbers and investigate their relationships. We expect to see states with lower median incomes (and therefore regions with lower median incomes) having lower administration numbers.

## 2.2 Description of data and data source

The main dataset contains US COVID-19 vaccine deliveries and administration data at national and jurisdiction level from all vaccine partners, including jurisdictional partner clinics, retail pharmacies, long-term care facilities, dialysis centers, Federal Emergency Management Agency and Health Resources and Services Administration partner sites, and federal entity facilities (CDC, n.d.-b). It has 109 columns and 37,500 rows of administration, distribution, and series data of COVID-19 vaccinations across the US.

To use income data in our analysis, we uses US Census Bureau ACS Income date from 2021 (Bureau, 2021). This data was taken from the American Community Survey.

## 2.3 Questions/Hypotheses to be addressed

Question: What factors (like region or median income) influenced COVID-19 vaccines across the US? Hypotheses: H0 - There is no association between region or median income per state and COVID-19 vaccine administration in the United States. HA - There is an association between region or median income and COVID-19 vaccine administration in the United States.

# 3. Methods

## 3.1 Data acquisition, import, and cleaning

The main dataset contains US COVID-19 vaccine deliveries and administration data at national and jurisdiction level from all vaccine partners, including jurisdictional partner clinics, retail pharmacies, long-term care facilities, dialysis centers, Federal Emergency Management Agency and Health Resources and Services Administration partner sites, and federal entity facilities (CDC, n.d.-b). It has 109 columns and 37,500 rows of administration, distribution, and series data of COVID-19 vaccinations across the US. To use income data in our analysis, we used US Census Bureau ACS Income date from 2021 (Bureau, 2021). This data was taken from the American Community Survey.

The data was directly imported from their respective websites into the repository, read into the ‘wrangling.qmd’ file, and saved as an RDS file. The CDC US vaccination data cleaning included reordering the ‘date’ column, filtering by only the year 2021, pivoting long to reformat into a more analysis-friendly setup, adding a ‘region’ column. The cleaned data was saved into a new RDS file ‘SubData’. The ACS income data cleaning included removing unnecessary columns, converting state names to abbreviations to match the CDC dataset, and adding a region column. These two datasets were merged using ‘inner\_join()’ by the ‘state’ column into the ‘Clean\_Complete’ dataset. ‘Clean\_Complete’ was then filtered by the last MMWR week as that had the cumulative number of vaccines distributed/administered; we also removed column that were not pertinent to our analyses, created binary predictors based on the median and mean household income for the US (provided by the ACS dataset). The final version of ‘Clean\_Complete’ was saved in a new RDS file for use in our analyses. Step-by-step instructions can be found in the ‘wrangling.qmd’ file.

To explore the data, we created two main figures using the ‘ggplot2’ package; one viewing distribution counts over 2021 and one viewing administration counts over 2021. Both are also categorized by vaccine manufacturers. These can be viewed in the results section of this document; more exploratory graphs and the code for all figures can be found in the ‘exploration.qmd’ file.

## 3.2 Statistical analyses

For our initial analyses, we fit simple linear models using overall administered vaccines as the outcome and region as the predictor. More fits can be found in the ‘fitting.qmd’ file. To determine what the best predictor of vaccine administration in our dataset was, we used a tree-based machine learning model. Our code was modeled after the ‘tidymodels’ package framework; it used cross-validation and split the cleaned data into training and testing sets for comparison. We also created a null model to compare our final model to to evaluate its performance. The exact steps can be found in the ‘analysis.qmd’ file.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

Here are the main results from our exploratory analysis where we plotted the distributed and administered vaccines over the observed period. Individual bar graphs for each region can be found in the ‘exploration.qmd’ file in the ‘processing\_code’ file.

|  |
| --- |
|  |

|  |
| --- |
|  |

## 4.2 Basic statistical analysis

Here are the main results from our initial fitting analysis. Note: we may not keep all of these fits in our final project but we calculated them all to view the results. Further checking of the models can be found in the ‘fitting.qmd’ in the ‘analysis\_code’ folder.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -4.080986e+04 | 8.911512e+03 | -4.579454 | 4.7e-06 |
| RegionNorthEast | 1.372110e+05 | 1.231630e+04 | 11.140597 | 0.0e+00 |
| RegionSouthEast | -3.766822e+05 | 1.205792e+04 | -31.239408 | 0.0e+00 |
| RegionSouthWest | -4.360264e+05 | 1.718853e+04 | -25.367286 | 0.0e+00 |
| RegionWest | 8.410008e+04 | 1.231966e+04 | 6.826495 | 0.0e+00 |
| Distributed | 8.316494e-01 | 4.335000e-04 | 1918.240535 | 0.0e+00 |

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 5019195.3 | 127037.3 | 39.509619 | 0e+00 |
| RegionNorthEast | 974680.7 | 183695.6 | 5.305955 | 1e-07 |
| RegionSouthEast | 951843.1 | 179657.9 | 5.298088 | 1e-07 |
| RegionSouthWest | 4110723.8 | 254074.6 | 16.179200 | 0e+00 |
| RegionWest | 1086846.3 | 183695.6 | 5.916561 | 0e+00 |

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -1.031285e+05 | 5664.8526336 | -18.20498 | 0 |
| Distributed | 8.297177e-01 | 0.0004641 | 1787.81982 | 0 |

## 4.3 Full analysis

Comparing the best tree model to the null model fitted with the training dataset.

| cost\_complexity | tree\_depth | .metric | .estimator | mean | n | std\_err | .config |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | rmse | standard | 10850738 | 5 | 1323450 | Preprocessor1\_Model01 |

| .metric | .estimator | mean | n | std\_err | .config |
| --- | --- | --- | --- | --- | --- |
| rmse | standard | 10034816 | 5 | 1058155 | Preprocessor1\_Model1 |
| rsq | standard | NaN | 0 | NA | Preprocessor1\_Model1 |

Comparing the last tree fit with the null model fitted with the testing data.

| .metric | .estimator | .estimate | .config |
| --- | --- | --- | --- |
| rmse | standard | 1.589127e+07 | Preprocessor1\_Model1 |
| rsq | standard | 5.776640e-02 | Preprocessor1\_Model1 |

| .metric | .estimator | mean | n | std\_err | .config |
| --- | --- | --- | --- | --- | --- |
| rmse | standard | 11986004 | 5 | 5556718 | Preprocessor1\_Model1 |
| rsq | standard | NaN | 0 | NA | Preprocessor1\_Model1 |

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

We believe given the consistency and availability of COVID-19 vaccination across the U.S. there is great potential to drill down on the drivers of increase vaccination uptake and identify regions having great success and regions of concern. This study showed that the Western region of the U.S. is associated with increase COVID-19 vaccine administered in the year 2021. Further, from our machine learning modeling region appears to be a major indicator of vaccine administration despite including income by state. Our findings should spur more research looking at the characteristics of these regions to ascertain these drivers from a national viewpoint. Our research was limited by the context and formatting of datasets with other variable that characteristics these areas. Datasets were incompatible due to less frequent or sparse updates as well as insufficient information to reliably join or add to our existing dataset containing COVID-19 vaccine numbers.

## 5.3 Conclusions

Region, particularly living in the Western region of the U.S., is a significant predictor of overall COVID-19 vaccine administration for the year 2021. While variables that could drive these numbers or speak to the behaviors and resources of these regions are not available in the vaccine administration dataset that is published weekly, there are other datasets available that can be cleaned, aggregated, and joined to study potential drivers. However, datasets with these variables are not updated as regularly and many are formatted in such a way that makes it difficult to join with our primary vaccine dataset for study. More research and resources should be devoted to cultivating data that can be analyzed to determine which regions in the U.S. was better vaccine uptake and what are the drivers of vaccine uptake or hesitancy from a national perspective.

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

# 6. References

Bureau, U. C. (2021). *American Community Survey Income Data over the Past 12 Months*. Retrieved from <https://data.census.gov/table?q=United+States&t=Income+and+Earnings&g=010XX00US,$0400000&y=2021>

CDC. (n.d.-a). *CDC COVID-19 Data Tracker*. Retrieved from <https://covid.cdc.gov/covid-data-tracker/#datatracker-home>

CDC. (n.d.-b). *COVID-19 Vaccinations in the United States,Jurisdiction*. Retrieved from <https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc>

Dooling, K., Marin, M., Wallace, M., McClung, N., Chamberland, M., Lee, G. M., … Oliver, S. E. (2021). *The Advisory Committee on Immunization Practices’ Updated Interim Recommendation for Allocation of COVID-19 Vaccine — United States, December 2020*. *69*(51).

Kawai, K., & Kawai, A. T. (2021). Racial/Ethnic and Socioeconomic Disparities in Adult Vaccination Coverage. *American Journal of Preventive Medicine*, *61*(4), 465–473. <https://doi.org/10.1016/j.amepre.2021.03.023>

Zimmerman, F. J., & Anderson, N. W. (2019). Trends in Health Equity in the United States by Race/Ethnicity, Sex, and Income, 1993-2017. *JAMA Network Open*, *2*(6), e196386. <https://doi.org/10.1001/jamanetworkopen.2019.6386>