Manuscript for a Data Analysis Project

Katie Wells & Connor Ross

4/6/23

# 1. Introduction

## 1.1 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.

## 1.2 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.

## 1.3 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.

# 2. Summary/Abstract

*Write a summary of your project.*

# 3. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.*

## 3.1 Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

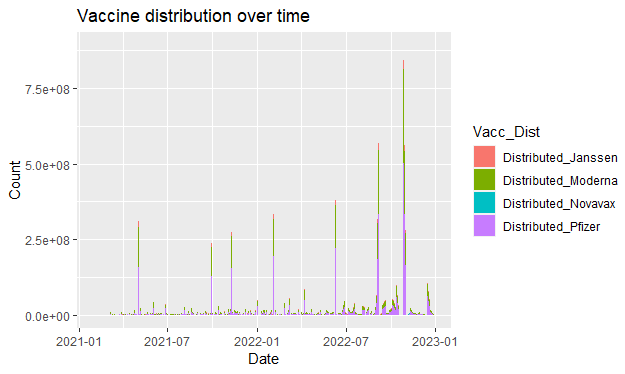
## 3.2 Statistical analysis

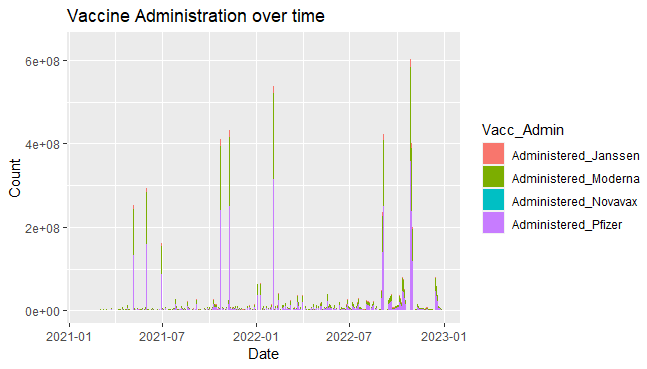
After placing each state in a region of the US, we will conduct a time series analysis to determine the change in popularity over time. Further, we intended to plot overall administration of manufacturer-specific vaccines in each region to determine popularity after adjusting for distribution and time.

# 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.

resulttable1 <- readRDS(here("code", "analysis\_code", "lm7.rds"))  
kable(resulttable1)

| 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 |

resulttable2 <- readRDS(here("code", "analysis\_code", "lm8.rds"))  
kable(resulttable2)

| 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 |

resulttable3 <- readRDS(here("code", "analysis\_code", "lm9.rds"))  
kable(resulttable3)

| 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.

resulttable4 <- readRDS(here("code", "analysis\_code", "best\_tree\_df.rds"))  
kable(resulttable4)

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

resulttable5 <- readRDS(here("code", "analysis\_code", "null\_train\_df.rds"))  
kable(resulttable5)

| .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.

resulttable6 <- readRDS(here("code", "analysis\_code", "tree\_last\_df.rds"))  
kable(resulttable6)

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

resulttable7 <- readRDS(here("code", "analysis\_code", "null\_test\_df.rds"))  
kable(resulttable7)

| .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

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*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>