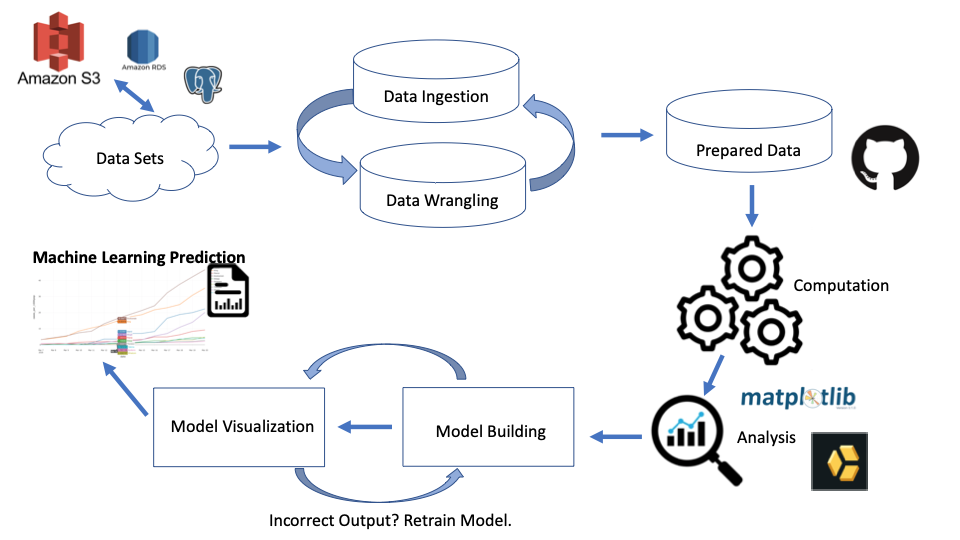
### **Team: Masks to Liberty**

### Capstone Assignment #2: Architecture & Design

**Updated Hypothesis:** This architecture design document was developed to support a capstone machine learning project that will use CDC COVID-19 vaccine administered data1 by state to estimate the number of people vaccinated on a given date in the future. This project will also determine if different vaccines (for example, Pfizer or Moderna) have any impact on the vaccination rate, if race4 has an impact on vaccination rate, and if Democratoc or Republican states, as decided by the 2020 election2, have an impact on vaccination rate. We will also look at Covid-19 symptoms to see if there was an impact on vaccine administration. This document provides both logical and physical design considerations for all related infrastructure components including servers, storage, networking, management, and virtual machines. The scope of this document is specific to the design of the machine learning model and the supporting components. We will further investigate if using this model, we can establish potential landmarks for achieving herd immunity thresh-holds?

**System Diagram:**

******

**System Description:**

* **Frameworks**: Uploading our .csv files into an Amazon S3 bucket, our group will work to mungle and wrangle the data for different avenues of exploration with SQL queries and joins. Once clean, we will build machine learning models in Python and use Github as a means of version control for different avenues of exploration. Once our model is trained, we will use future collected data to test our models predictions.
* **Sever**: Amazon Simple Storage Service (Amazon S3) + Amazon Web Services Relational Database System (AWS RDS) + PostgreSQL (database). The database is the repository where all of the data utilized for exploratory and machine learning models resides. For this capstone project, we will use Amazon RDS as the database service and will use PostgreSQL as a method to run SQL queries and create new tables, which can be easily imported into Jupyter Notebook, Terminal etc. <https://www.postgresql.org/>
* **Software:** The software architecture is designed to incorporate all CDC COVID-19 vaccine administered data1 entries including and up to March 31, 2021. This data set is taken from a GitHub source (see below for link) and will be part of a database that will be used as the foundation of the integration and wrangling work. We will also use a 2020 election data set2, a racial4 covid-19 dataset, and a Covid-19 symptom dataset5 which will be imported into the database.

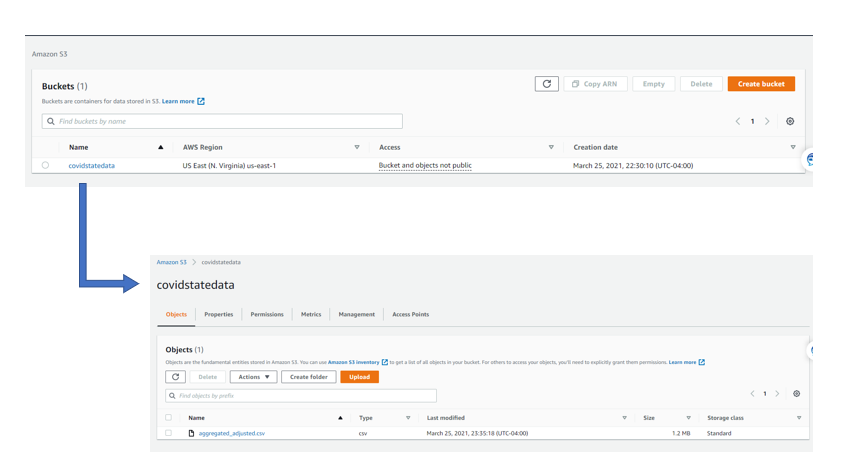
**Dataset References:**

1. CDC Covid -19 Dataset: <https://raw.githubusercontent.com/youyanggu/covid19-cdc-vaccination-data/52afaf5e5435f90dcd0da333aa17e28d14ecf45d/aggregated.csv>
2. 2020 Presidential Election Dataset: <https://cookpolitical.com/2020-national-popular-vote-tracker>
3. 2020 Election including county: <https://www.kaggle.com/unanimad/us-election-2020?select=president_county_candidate.csv>
4. Covid-19 Racial Dataset: <https://www.kff.org/other/state-indicator/covid-19-vaccinations-by-race-ethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>
5. Covid-19 Symptoms Dataset: <https://wonder.cdc.gov/vaers.html>

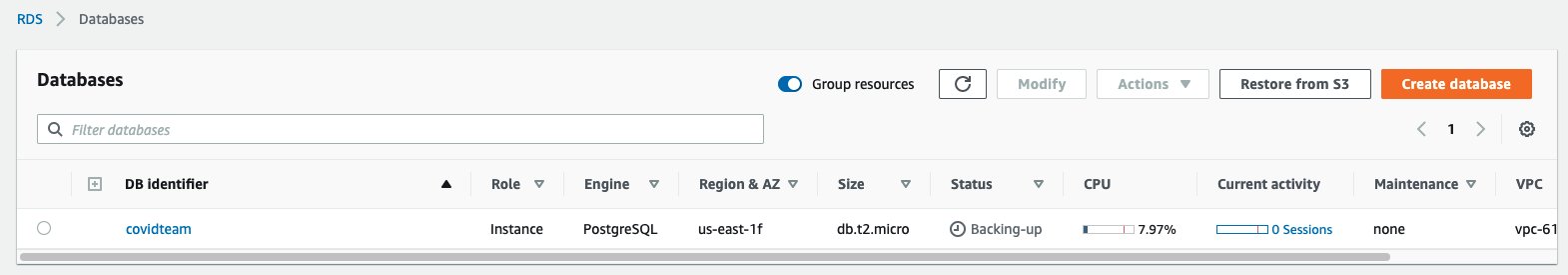
**Other Technology:**

|  |  |
| --- | --- |
| **Item** | **Value** |
| Command Line System: For the Mac, it is the Terminal. For Windows, it is PowerShell. | The command line is a text interface for your computer. It's a program that takes in commands, which it passes on to the computer's operating system to run. From the command line, you can navigate through files and folders on your computer, just as you would with Windows Explorer on Windows or Finder on Mac OS. |
| Text Editor: Sublime Text Editor <https://www.sublimetext.com/> | Text editors make it easy to compile your code and enables the user to create and edit text files. Sublime Text is the fastest text editor that you can use for writing code. |
| Language: Python <https://www.anaconda.com/>  SQL <https://www.mysql.com/> | Python is one of the more popular languages for machine learning.  SQL is a domain-specific language used in programming and designed for managing data held in a relational database management system, or for stream processing in a relational data stream management system. |
| Notebook: Jupyter Notebook <https://jupyter.org/> | Jupyter Notebooks are a powerful way to write and iterate on your Python code for data analysis. Additional features include storing your code and output and allowing you to keep markdown notes. |
| Version Control: Git <https://git-scm.com/> | Git is software for tracking changes in any set of files, usually used for coordinating work among programmers collaboratively developing source code during software development. Its goals include speed, data integrity, and support for distributed, non-linear workflows. |
| Web Interface: GitHub <https://github.com/> | GitHub, Inc. is a provider of Internet hosting for software development and version control using Git. It offers the distributed version control and source code management functionality of Git, plus its own features. |
| Data Analysis: Pandas <https://pandas.pydata.org/> | Pandas is a popular and favourite data science tool used in Python programming language for data wrangling and analysis. It is often used when it comes to cleaning, transforming, manipulating and analyzing data. |
| Graphics: Matplotlib [https://matplotlib.org/](https://matplotlib.org/%5C)  Yellowbrick:<https://www.scikit-yb.org/en/latest/>  Seaborn <https://seaborn.pydata.org/>  Sklearn: <https://scikit-learn.org/stable/supervised_learning.html#supervised-learning> | Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.  Yellowbrick is a tool built on matplotlib which may be helpful for visualizing our machine learning simply with a variety of visualization options.  **from** **yellowbrick.regressor** **import** AlphaSelection  **from** **yellowbrick.regressor** **import** PredictionError, ResidualsPlot  **from** **yellowbrick.features** **import** parallel\_coordinates  **from** **yellowbrick.classifier** **import** classification\_report  *# Displays parallel coordinates*  g = parallel\_coordinates(X, y)  *# Displays classification report*  g = classification\_report(LogisticRegression(), X, y)  Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.  Sklearn is a Python tool which can help us build and tune our machine learning.  **from** **sklearn.linear\_model** **import** LinearRegression  **from** **sklearn.model\_selection** **import** train\_test\_split  **from** **sklearn.linear\_model** **import** Ridge, Lasso |

**Amazon S3 Bucket**



**AWS RDS**



**Processes**:

Data Ingestion & Storage

* Download .csv files as identified above
* For storage, this team has uploaded the datasets to an Amazon S3 bucket to leverage Amazon’s easily usable data management system. As noted above, we have selected Amazon RDS as the database service and PostgreSQLas the database engine.
* Develop credentials and processes for accessing shared data storage
* Develop folder structure for raw/clean and work in progress (new tables)

Mungling & Wrangling

* Decide structure for organizing datasets, naming files, placement etc.
* Organize features, remove all unnecessary data
  + Generate joined tables with features of interest from raw data
* Upload experimental tables to S3 buckets with appropriate Readme
  + This way we can collaborate on experiments as work in progress with data locality
* Upload fruitful analysis to github
  + Include explanation of SQL query for replication from raw data, still may utilize cleaned tables

Computation & Analysis

* Create python code to read our table features into visualizations for exploratory data analysis
* Develop boilerplate for common aspects of analysis (such as accessing S3 buckets)
  + Rewrite code from scholars code to professional. For example: classes and functions
    - This effort will pay off as we will need to make individual models for at least 50 states, as well as potentially counties.
    - Because we will want to include other features in our model, developing well written code will allow more time for phase 2 exploratory data analysis and feature exploration with machine learning.
* TBD… Potential methods of analysis include
  + Political Split (by state, by county)
  + Demographic split (by state, by county)
  + Type of vaccine and side effects
  + Other

Modeling & Application

* From aggretate-adjusted.csv, choose a particular state to create a test model template

from Sklearn.linear\_model import LinearRegression

model = LinearRegression()

model.fit(X, y)

model.predict(X)

* X may be the date range of Dec. 15 - Mar 7th because we have further datasets to test
* y is the chosen states second dose administered feature divided by census population
* The model can than be tuned with yellowbrick resources

**from** **yellowbrick.regressor** **import** PredictionError, ResidualsPlot

**from** **yellowbrick.regressor.alphas** **import** AlphaSelection

* We may choose to instead train and test our model all on an aggregate adjusted time frame since we will attempt to include features from other datasets later.
  + from sklearn.model\_selection import train\_test\_split
  + from sklearn.metrics import r2\_score

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, train\_size=.8, shuffle=False, random\_state=1)

Reporting & Visualization

* Write a paper that describes our process and analysis
* Create a powerpoint presentation that explains our paper and focuses on machine learning
  + Include in paper and powerpoint visualizations of our data and machine learning
    - 70% Machine Learning focus in capstone presentation

**Appendix**

Using a github repository of CDC data, as well as daily updates of the most current CDC data, we train our machine learning to estimate the amount of vaccine administered to the states population, divided by the population of the state, to get a percent vaccinated.

Github history: ~58 data-points over 3 months (from beginning of year 2021 to march 7th)

Daily updates: 1 Datapoint per day

By utilizing a database system such as SQLite, we can create tables out of old and current CDC data, and join the tables by index, state, time(date), doses administered, % of pop vaccinated, and other metrics provided in the tables to create projections for the future, which may then be validated or invalidated by more current data, so as to train our machine. Using tools such as matplotlib, pyplot, pandas, (and others), we can distill our tables into specific metrics and periods of time, as well as create visualizations for the data.

From CDC COVID-19 data:

Instance: ‘**LongName’**, ‘**Date**’

Features: ‘**Doses\_Administered’, ‘Census2019’**

Engineered Features: ‘**Doses\_Administered’ / ‘Census2019’**

Labels:

percent-vaccinated = ‘**Doses\_Administered’ / ‘Census2019’**

Daily Updates from CDC: <https://covid.cdc.gov/covid-data-tracker/#vaccinations>

Instance: ‘**LongName’**, ‘**Date**’

Features: ‘**Doses\_Administered’, ‘Census2019’,**

**‘Percent of Total Pop Fully Vaccinated by State of Residence’**

Policy Hypothesis: We can use our dataset to validate or invalidate criticisms of policy decisions made about certain states.

In a WP article, a controversy was cited regarding Maryland’s policy of guaranteeing the second dose for those who have the first. One state politician referred to this as vaccine hoarding while another claimed it was an ethical imperative.

Source:<https://www.washingtonpost.com/local/second-doses-vaccine-maryland-virginia/2021/02/15/252d4eea-6c89-11eb-9ead-673168d5b874_story.html>

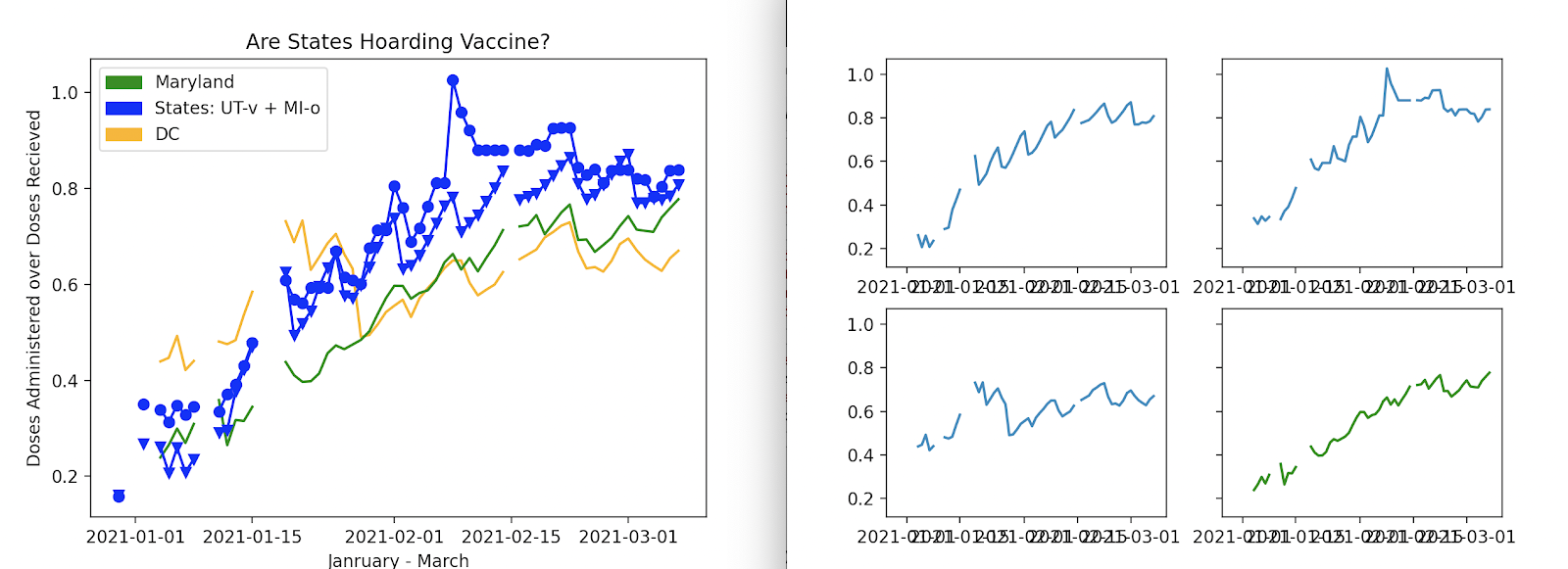
Instance: ‘**LongName’**, ‘**Date**’

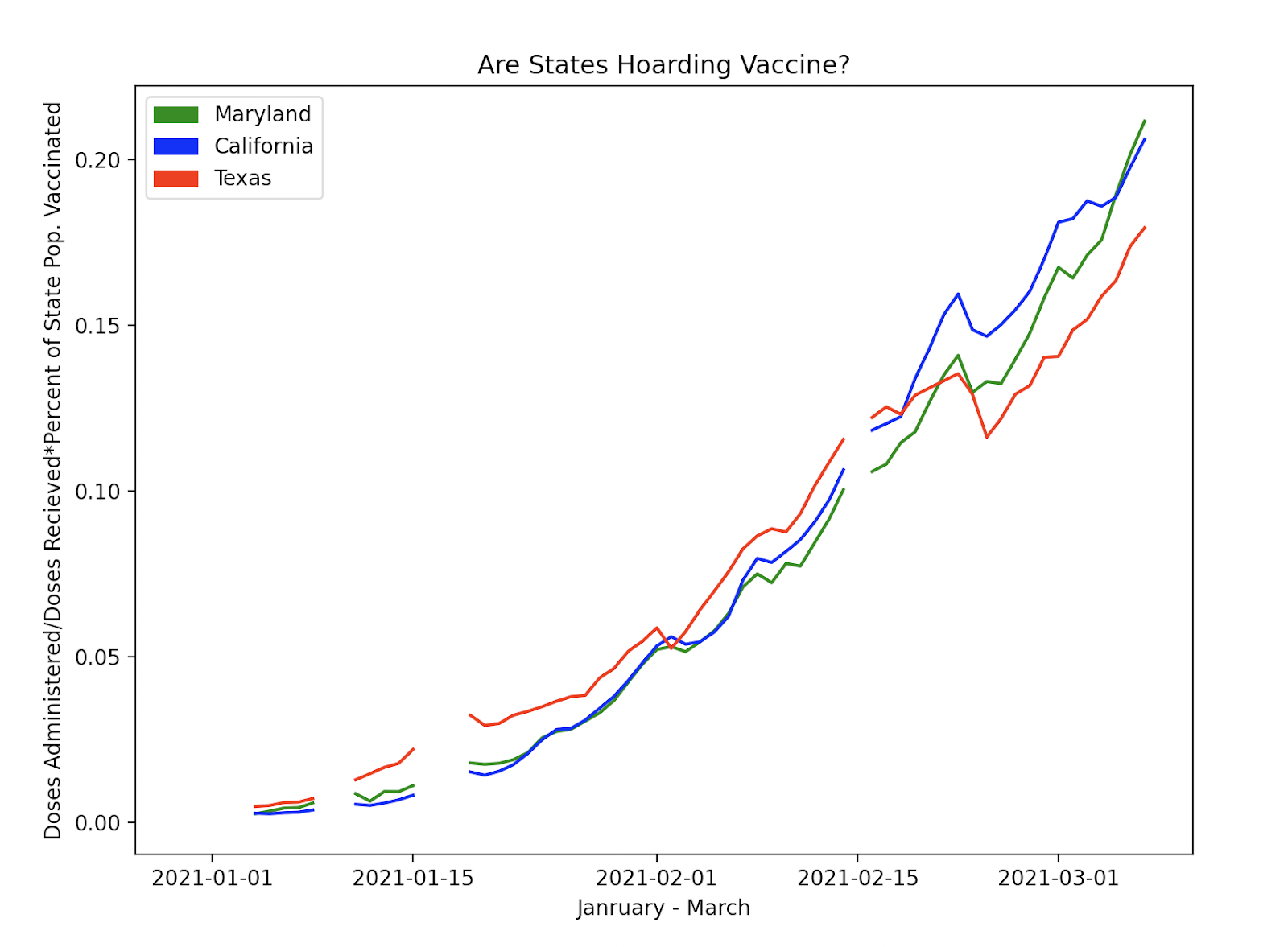
Features: ‘**Doses\_Administered’, ‘Doses\_Distributed’, ‘Census2019’**

Engineered Features: **Doses\_Distributed’ / ‘Doses\_Administered’ \* ‘percent\_vaccinated’**

Labels:

percent-vaccinated = ‘**Doses\_Administered’ / ‘Census2019’**

score= ‘**Doses\_Distributed’ / ‘Doses\_Administered’ \* ‘percent\_vaccinated’**

Our CDC data visualizations seem to suggest that Maryland is not vaccine hoarding, at very least in relation to Texas, California, Utah, Michigan, or the District of Columbia. If we added features of current and previous infections, we could potentially further validate or invalidate this policy decision. 

Hypothesis 2

Combining CDC Data from the Adjusted Aggregate source with data on past infections by state, can we create a hypothetical projection of herd immunity by drawing a meaningful relationship between vaccinations and the number of daily infections?

Source:<https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data> ^^I had to edit a Texas input that was negative, data needs review

Instance: ‘**submission\_date’, ‘state’**

Features: ‘**tot\_cases’, new\_case**

Engineered Features: n/a

Labels: n/a (This analysis combines Adjusted Aggregate’s Engineered Features\*\*)

Below are some graphs which juxtaposed in:

Red, the number of daily infections (Jan 2020 - March 2021)

Orange, the total number of infections (Jan 2020 - March 2021)

Blue, the states percent of the population vaccinated against covid 19 (Dec 2020 - Mar)

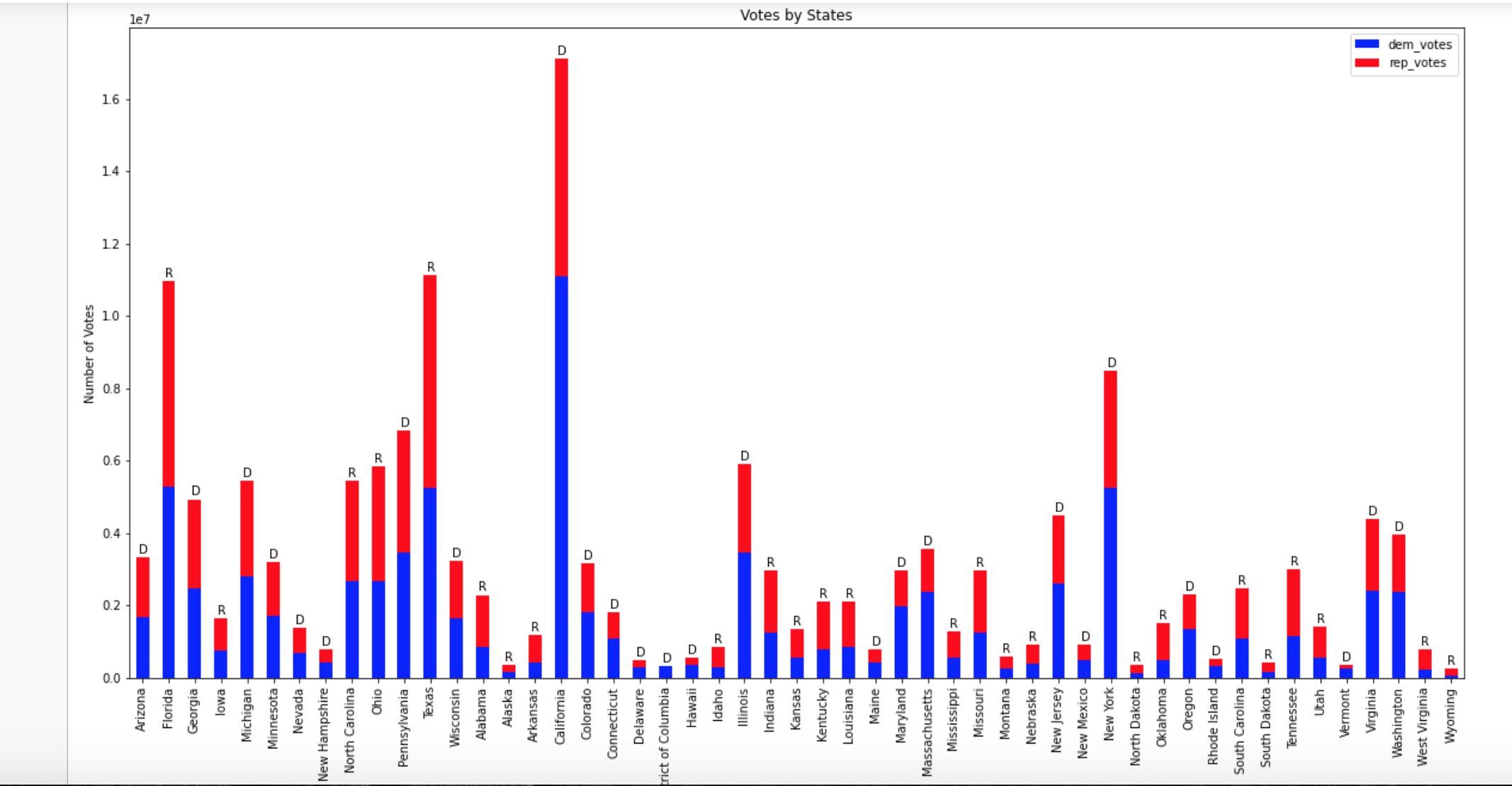
Green, the total number of vaccinations [e6] (December 2020 - March 7th 2021)

|  |  |  |
| --- | --- | --- |
| Maryland | Texas | California |
| / |  |  |
| DC | Utah | Michigan |
|  |  |  |

Political Hypothesis: If we classify our states into red and blue, relating to 2020 election results, can we find data points or trends that show correlation between political stance and data relating to Covid-19 vaccination. In other words, can we predict if classifying data into Democrat or Republican will have an impact on the vaccine administered data? What is the margin of difference that would fit the model?

The bar graph below represents the first step in visually displaying the data into Democrat or Republican states.

Source: <https://cookpolitical.com/2020-national-popular-vote-tracker>



We can now join this dataset with the COVID data set to plot vaccines administered by Democratic or Republican states.

Instance: ‘**LongName’**, ‘**Date**’ red state average, blue state average

Potential Features: **‘Doses\_Distributed’,** ‘**Doses\_Administered’, [Others]**

Potential Engineered Features: red state average, blue state average,

Potential Labels: ^^ Depends on source and methods^^  
We can show statistical correlation or no correlation between if a state went red or blue, and x features of our CDC covid data.

Another political dataset to consider: Includes county data:

<https://www.kaggle.com/unanimad/us-election-2020?select=president_county_candidate.csv>

We will also determine if vaccine side effects have an impact on vaccine administration? Which vaccine has the most side effects between Pfizer, Moderna, and Johnson and Johnson? How many vaccines have been administered and how many side effects.

Features: Manufacturer and Side Effects

Slice number administered and number problems

Did a certain race or age group experience more side effects than another?

Input Sources:

* Number of Vaccines per Brand:
* Number of Side Effects per Brand
* Ages of people that have received the vaccine and have had side effects
* Races that have received the vaccine and have had side effects

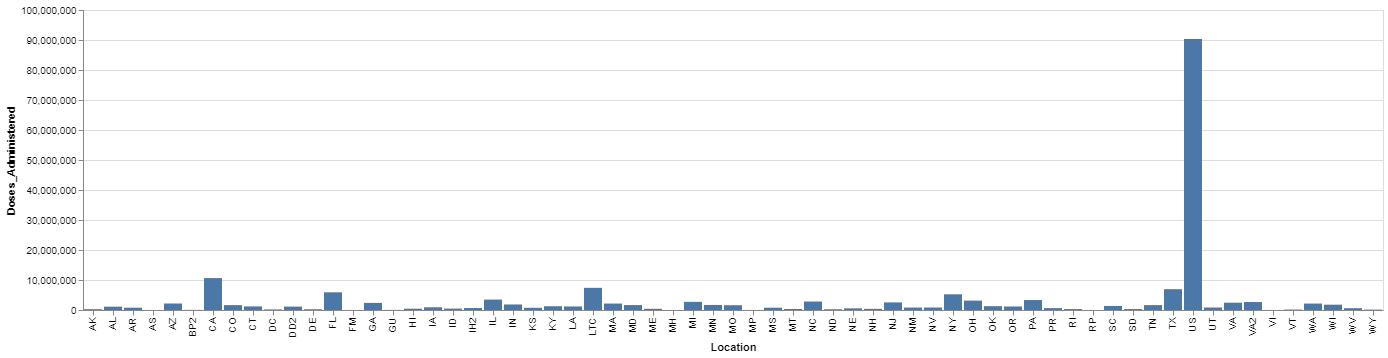
1. <https://wonder.cdc.gov/vaers.html>
2. [CDC: Serious reactions to first 14 million COVID-19 vaccines extremely rare (news5cleveland.com)](https://www.news5cleveland.com/news/continuing-coverage/coronavirus/cdc-serious-reactions-to-first-14-million-covid-19-vaccines-extremely-rare)
3. [First Month of COVID-19 Vaccine Safety Monitoring — United States, December 14, 2020–January 13, 2021 | MMWR (cdc.gov)](https://www.cdc.gov/mmwr/volumes/70/wr/mm7008e3.htm?s_cid=mm7008e3_e&ACSTrackingID=USCDC_921-DM50013&ACSTrackingLabel=MMWR%20Early%20Release%20-%20Vol.%2070%2C%20February%2019%2C%202021&deliveryName=USCDC_921-DM50013)
4. [Coronavirus (COVID-19) Vaccinations - Statistics and Research - Our World in Data](https://ourworldindata.org/covid-vaccinations)
5. [Government Database Shows 7,844 Injuries and 181 Deaths Following COVID Vaccine as of Jan. 15 says Children's Health Defense (prnewswire.com)](https://www.prnewswire.com/news-releases/government-database-shows-7-844-injuries-and-181-deaths-following-covid-vaccine-as-of-jan-15-says-childrens-health-defense-301217037.html)
6. [COVID-19 Vaccine Safety Surveillance | FDA](https://www.fda.gov/vaccines-blood-biologics/safety-availability-biologics/covid-19-vaccine-safety-surveillance)
7. <https://vaers.hhs.gov/data/datasets.html>

* Which state has been and do we expect to be the most efficient at administering its received vaccines (by insert date)?
* Which state has been the most effective distributing its vaccines to the state population

Exploratory questions:

* Does the vaccine type (e.g. Moderna v. Pfizer) have an impact on vaccine administration?

Descriptives from the available data1:



* Does race have an impact on vaccine administration?

Descriptives from the available data4:

