**Data Science Program Final Project**

**Covid Vaccine Data: for the State of California**

**Executive Summary**

On March 11, 2022 represents the three-year anniversary since the World Health Organization declared Covid-19 a global pandemic, and shortly after, former President Donald Trump had declared a national emergency, with states beginning to issue stay-at-home orders, mask mandates, and capacity limits at businesses like restaurants and bars.[[1]](#footnote-1)

If 2020 was dominated by the news of how Covid-19 spread across the globe, putting normal life on hold for most people and overwhelming health care providers, then 2021 was by far focused on ending the pandemic through vaccine distribution through 2022.[[2]](#footnote-2)

With research into hundreds of vaccine candidates, there are now 3 vaccines approved by the FDA for use in the United States. As of May 13, [the CDC was reporting](https://covid.cdc.gov/covid-data-tracker/#vaccinations) that more than 580.9 million vaccine doses had been administered.[[3]](#footnote-3) Moreover, California counties like Imperial, Merin, Santa Clara, San Mateo, and San Francisco have administered the most doses per 100,000 of the total population. And with close to 75% of the California population fully vaccinated, updated guidance has changed the State.[[4]](#footnote-4)

This data project will explain the data information of vaccine rollout in the State of California as it progressed from 2021 into 2022 for the counties and demographics of community members. This data project will provide a point of view and ask questions to understand the impact of the Covid-19 vaccination administrated data sets.

**Business Objectives**

To showcase the skills that Jasmin, Lacey, and Arkajia have acquired through the Data Science program. They will be using R, Python, SQL, and other programs to process data wrangling, analyze, and visualization of the “Covid Vaccine Data: for the State of California,” dataset made available by the California Open Data and CA Health and Human Services on Google Data Sets.

This data is from the same source as the Vaccine Progress Dashboard at <https://covid19.ca.gov/vaccination-progress-data/> which summarizes vaccination data at the county level by county of residence. Where county of residence was not reported in a vaccination record, the county of provider that vaccinated the resident is included. This applies to less than 1% of vaccination records. The sum of county-level vaccinations does not equal statewide total vaccinations due to out-of-state residents vaccinated in California.

At the end of the data project, Jasmin, Lacey, and Arkajia should be able to explain their work in layman’s term, and present their findings to the students, faculty, staff, and potential employers, along with other interested parties via Zoom.

**Background**

Jasmin, Lacey, and Arkajia have selected the “Covid Vaccine Data: for the State of California” dataset because the topic carries interest in healthcare and affected the lives of millions of people in the United States, in general. They plan to collect understanding from this document for vaccine rollout in the State of California as it progressed from 2021 into 2022 for the counties and demographics of community members.

On December 11, 2020, The Food and Drug Administration issues an Emergency Use Authorization (EUA) for the first COVID-19 vaccine – the Pfizer-BioNTech COVID-19 vaccine.[[5]](#footnote-5)

Thereafter, December 18, 2020, The U.S. Food and Drug Administration issues an Emergency Use Authorization for the second COVID-19 vaccine – the Moderna COVID-19 vaccine. [[6]](#footnote-6)

Lastly, on February 27, 2021, FDA approves emergency use authorization for Johnson and Johnson one shot COVID-19 vaccine.[[7]](#footnote-7)

**Scope**

Jane and Jessica will be using the software taught in the program to complete the project. They will be intentional on using tools of their interest or tools that may aid finding a job. They may choose to use additional software/tools, but that is not required.

**Functional requirements**

Data Wrangling: The downloaded dataset should be successfully cleaned up for analyzing. Columns and unusable columns should be removed. As the dataset is fairly large, Jane and Jessica should consider sub-setting the dataset in a proper manner, meaning the subset should be a random selection of the data. The datatypes for each column should also be converted to a usable format for the needed analysis.

Data Analysis: Jane and Jessica will familiarize themselves with the dataset. They should have a good understanding of what each column means, and how the values are measured. They will brainstorm on questions to ask, and what they might gather from the dataset. Then, they will identify the proper functions to create models, predictions, etc.

Data Visualization: Once Jane and Jessica have a comprehensive understanding of and insight gathered from the dataset, they will work on visualizing the findings. They may decide to use Tableau or other graphing programs, and compile the visuals and texts in a Power Point slideshow.

Presentation: Working with school leaders, Jane and Jessica will schedule a time to present their findings via Zoom. They should be able to communicate in a clear and easy-to-understand manner. The presentation should be kept around 20 minutes. They should be dressed professionally for this occasion.

**Personnel requirements**

Jane and Jessica are the two developers. They will need to work closely for this project to succeed. They will touch base once a day via Zoom or Slack to problem-solve or to check in on work progresses. Once a week, they will review the past week workload and plan out the next week. They will take turns being the scrum master, and report their progress to their instructor (Product Owner.)

Once a week, they will meet with their instructor. They should be prepared to ask questions and seek guidance for the next steps.

They may also consult with their coding mentor.

**Delivery schedule**

Week 1: Import dataset into preferred software to begin data wrangling. Any unnecessary columns should be removed. Educate ourselves on breast cancer. Set up Github.

Week 2: Study the dataset and ask questions. What are some possible correlations? Is the data normally distributed? What are some predictive models we can make from it? Visualize the data to see if there is any interesting findings.

Week 3: Modeling/Optimization (Combined Stepwise - Forward and Backward Selection) and Machine Learning (Random Forest.)

Week 4: Review and validate findings from the previous week, and draw insights/conclusions.

Week 5: Compile findings into a Power Point slideshow. Go over it with their instructor and friend/family member to ensure that the presentation is clear and logical. Work on the style and layout of the presentation so it is delightful on the eyes.

Week 6: Make final touches to the Power Point presentation. Jane and Jessica should not attempt to come up with a brand-new analysis. There will not be enough time to verify their findings. They should practice presenting at least a couple times with the two of them, and at least once with their instructor.

**Other requirements**

All programs used should be free of charge. Though Jane and Jessica may decide to use a paid service, such as a more advanced version of Tableau.

**Assumptions**

The software programs and platforms Jane and Jessica use should be available, up-to-date, and not broken.

**Limitations**

If something should come up for Jane and Jessica during this six-week period, the project may be delayed. If the instructor or mentor have scheduled or unscheduled time-off, the project may be delayed as well. Jane and Jessica may experience a roadblock in their work, which may push back the completion date.

**Risks**

The risks that may arise are such like natural disasters, power outages, family emergencies or broken software/hardware. Jane and Jessica are eager to complete the program so there should be no motivation issues. The instructor and mentor are phenomenal so there is no concern of no help from them. The risk of this project being incomplete is minimal. They will be successful in completing this project!

1. [↑](#footnote-ref-1)
2. https://www.ajmc.com/view/a-timeline-of-covid-19-vaccine-developments-in-2021 [↑](#footnote-ref-2)
3. https://covid.cdc.gov/covid-data-tracker/#vaccinations\_vacc-total-admin-rate-total [↑](#footnote-ref-3)
4. https://covid19.ca.gov/vaccination-progress-data/ [↑](#footnote-ref-4)
5. [↑](#footnote-ref-5)
6. [↑](#footnote-ref-6)
7. https://www.cdc.gov/museum/timeline/covid19.html#Late-2020 [↑](#footnote-ref-7)