“Analysis of the Impact of COVID-19 Vaccinations on the Occurrence of Positive Tests”

Justin Alexander Bowman
STAT 4020
December 9, 2024

Introduction

The project has the goal of analyzing the impact that the COVID-19 vaccination dosage one and dosage two had on the positive test rates throughout the United States. COVID-19 (SARS-CoV-2) virus that infects its host when liquid particles enter a host's airway from an infected person's mouth ("Coronavirus disease (COVID-19)", 2024). The disease had an impact on individuals all over the world resulting in a new everyday norm. During this time Clemson responded by opening the Clemson REDDI Lab which helped develop a RT-PCR testing method for the detection of different strains of the COVID-19 virus. Many employees and I had the pleasure of helping to facilitate testing in a diagnostic lab during this time. As more people became infected, hospitals began to become overwhelmed and at max capacity. To fight against the disease, the release of a vaccination was released initially to health care workers and later to the public. The vaccination was administered in two dosages with an 8-week interval in-between as of 2024 ("Interim Clinical Considerations", 2024). Many argued that the vaccine was not necessary and had no impact on disease prevention. This project aims to see if there was an impact caused by the use of COVID-19 vaccine.

Data

Data Source

The data that was used was derived from Google Health COVID-19 Open Data Repository which contains data from over 20,000 locations around the world. The data was provided by authoritative sources (government, health, universities), general sources (news media, publications), crowdsourcing (volunteers, contributors). The data is intended to be used by organizations, public agencies, data scientists, and/or policymakers. The dataset provided real-time update up until September 15, 2022. The data source was chosen due to the reliability of the information provided from the data base.

The first data set that was used was the Epidemiology data set which focuses on the daily infection rates on in different regions around the world. The second data set that was used was Vaccinations data set which focuses on the daily administration of COVID-19 vaccine (*Covid-19 Open Data*).

Data Cleaning/Manipulation

The Epidemiology data set was initially imported to SAS as a CSV file using proc import step. To analyze the data, the proc contents step was used to determine the variables and type of variables found in the data as shown in **Table 1**. The data was then manipulated with the data step. The variables date, location_key, new_confirmed, cumulative_confirmed, and new_confirmed during the set statement. The variables Year, Month, and Day was formed from the date variable and then finally the variable date was dropped. Proc sort was then used on the data where which used the statement "where location_key="US"" to keep data for the United States and ordered by the variable Year. The same data was manipulated again with a data step. During this step, the new variable was Monthly_Sum_New was retained with the retain statement to keep a running count of each month's total of confirmed COVID-19 test. This data

was finally sorted using a sort step by the variables Year then Month. **Table 2** demonstrates the data after these steps were applied.

Table 1: Contents of Epidemiology Data

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
7	cumulative_confirmed	Num	8	BEST12.	BEST32.
8	cumulative_deceased	Num	8	BEST12.	BEST32.
9	cumulative_recovered	Char	1	$1.	$1.
10	cumulative_tested	Char	1	$1.	$1.
1	date	Num	8	YYMMDD10.	YYMMDD10.
2	location_key	Char	2	$2.	$2.
3	new_confirmed	Num	8	BEST12.	BEST32.
4	new_deceased	Num	8	BEST12.	BEST32.
5	new_recovered	Char	1	$1.	$1.
6	new_tested	Char	1	$1.	$1.

Table 2: Epidemiology Data Sorted and Edited

Obs	location_key	Year	Month	Monthly_Sum_Vaccinated_New	Monthly_Sum_Full_Vaccinated_New
1	US	2020	12	5504758	36849
2	US	2021	1	24131626	7336605
3	US	2021	2	25535781	22019808
4	US	2021	3	51323710	29845969
5	US	2021	4	41751173	45317704
6	US	2021	5	19776382	31664764
7	US	2021	6	9375389	19017261
8	US	2021	7	9411671	8562363
9	US	2021	8	13355270	9364106
10	US	2021	9	8152432	10001433

The Vaccinations data set was initially imported to SAS as a CSV file using proc import step. To analyze the data, the proc contents step was used to determine the variables and type of variables found in the data as shown in **Table 3** and **Table 4**. The data was then manipulated with the data step. The variables date, location_key, new_persons_fully_vaccinated, and new_persons_vaccinated during the set statement. The variables Year, Month, and Day was formed from the date variable and then finally the variable date was dropped. Proc sort was then used on the data where which used the statement “where location_key=”US”” to keep data for the United States and ordered by the variable Year. The same data was manipulated again with a data step. During this step, the new variables Monthly_Sum_Vaccinated_New and Monthly_Sum_Full_Vaccinated_New were retained with the retain statement to keep a running count of each month’s total COVID-19 first and second dosage for the month. This data was

finally sorted using a sort step by the variables Year then Month. **Table 5** demonstrates the data after these steps were applied.

Table 3: Contents of Vaccinations Data

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
12	VAR12	Char	1	$1.	$1.
14	VAR14	Char	1	$1.	$1.
18	VAR18	Char	1	$1.	$1.
20	VAR20	Char	1	$1.	$1.
24	VAR24	Char	1	$1.	$1.
26	VAR26	Char	1	$1.	$1.
6	cumulative_persons_fully_vaccina	Num	8	BEST12.	BEST32.
4	cumulative_persons_vaccinated	Num	8	BEST12.	BEST32.
22	cumulative_persons_vaccinated_ja	Char	1	$1.	$1.
16	cumulative_persons_vaccinated_mo	Char	1	$1.	$1.
10	cumulative_persons_vaccinated_pf	Char	1	$1.	$1.
8	cumulative_vaccine_doses_adminis	Num	8	BEST12.	BEST32.
1	date	Num	8	YYMMDD10.	YYMMDD10.
2	location_key	Char	2	$2.	$2.
5	new_persons_fully_vaccinated	Num	8	BEST12.	BEST32.
23	new_persons_fully_vaccinated_jan	Char	1	$1.	$1.
17	new_persons_fully_vaccinated_mod	Char	1	$1.	$1.
11	new_persons_fully_vaccinated_pfi	Char	1	$1.	$1.
29	new_persons_fully_vaccinated_sin	Char	1	$1.	$1.
3	new_persons_vaccinated	Num	8	BEST12.	BEST32.
21	new_persons_vaccinated_janssen	Char	1	$1.	$1.
15	new_persons_vaccinated_moderna	Char	1	$1.	$1.
9	new_persons_vaccinated_pfizer	Char	1	$1.	$1.
27	new_persons_vaccinated_sinovac	Char	1	$1.	$1.
7	new_vaccine_doses_administered	Num	8	BEST12.	BEST32.

Table 4: Contents of Vaccinations Data Continued

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
25	new_vaccine_doses_administered_j	Char	1	$1.	$1.
19	new_vaccine_doses_administered_m	Char	1	$1.	$1.
13	new_vaccine_doses_administered_p	Char	1	$1.	$1.
31	new_vaccine_doses_administered_s	Char	1	$1.	$1.
30	total_persons_fully_vaccinated_s	Char	1	$1.	$1.
28	total_persons_vaccinated_sinovac	Char	1	$1.	$1.
32	total_vaccine_doses_administered	Char	1	$1.	$1.

Table 5: Vaccinations Data Sorted and Edited

Obs	location_key	Year	Month	Monthly_Sum_Vaccinated_New	Monthly_Sum_Full_Vaccinated_New
1	US	2020	12	5504758	36849
2	US	2021	1	24131626	7336605
3	US	2021	2	25535781	22019808
4	US	2021	3	51323710	29845969
5	US	2021	4	41751173	45317704
6	US	2021	5	19776382	31664764
7	US	2021	6	9375389	19017261
8	US	2021	7	9411671	8562363
9	US	2021	8	13355270	9364106
10	US	2021	9	8152432	10001433

The two edited data sets were then merged using a proc merge statement by the variables Year and then Month. The data was then edited using a data step. The format statement was used on the variables cumulative_confirmed, Monthly_Sum_Vaccinated_New, Monthly_Sum_Full_Vaccinated_New, and Monthly_Sum_New to have the numeric values include commas to increase data readability. An if statement was also used to ensure the data that is not required was set to be a missing value. Proc sort was then used to resort the data to ensure it was sorted by the variable Year. **Table 6** demonstrate the data after these steps were applied. This data is what will be used for the series plot.

Table 6: Merged Data Sorted and Edited

Obs	new_confirmed	cumulative_confirmed	Year	Month	Monthly_Sum_New	Monthly_Sum_Vaccinated_New	Monthly_Sum_Full_Vaccinated_New
1	1	9	2020	1	10	.	.
2	9	82	2020	2	72	.	.
3	20131	158,732	2020	3	158,632	.	.
4	28654	916,239	2020	4	735,275	.	.
5	20610	1,626,491	2020	5	680,589	.	.
6	46496	2,476,880	2020	6	833,500	.	.
7	68412	4,406,097	2020	7	1,870,266	.	.
8	39130	5,874,328	2020	8	1,412,019	.	.
9	45516	7,084,590	2020	9	1,169,543	.	.
10	92551	8,998,888	2020	10	1,867,535	.	.

The data was furthered transposed using the transpose step where the Variable Month values were transformed into columns. The columns cumulative_confirmed, Monthly_Sum_Vaccinated_New, Monthly_Sum_Full_Vaccinated_New, and Monthly_Sum_New were transformed to be rows under the column name Vaccine Status by the variable Year. The final data step was used to add the final edits on the data. The columns that were originally the Month values were renamed to be the associated three letter month code. The values in Vaccine Status were edited to make the values more appropriate for table values. **Table 7** and **Table 8** demonstrate the data after these steps were applied. This data is what will be used for a table to better illustrate the data.

Table 7: Transposed Data Sorted and Edited

Obs	Year	Vaccine Status	JAN	FEB	MAR	APR	MAY
1	2020	Cumulative Positive Test	9	82	158,732	916,239	1,626,491
2	2020	Monthly Sum of 1st Covid Shot	.	.	.	.	.
3	2020	Monthly Sum of 2nd Covid Shot	.	.	.	.	.
4	2020	Monthly Sum of Positive Tests	10	72	158,632	735,275	680,589
5	2021	Cumulative Positive Test	25,605,016	27,833,375	29,508,492	31,266,636	32,145,209
6	2021	Monthly Sum of 1st Covid Shot	24,131,626	25,535,781	51,323,710	41,751,173	19,776,382
7	2021	Monthly Sum of 2nd Covid Shot	7,336,605	22,019,808	29,845,969	45,317,704	31,664,764
8	2021	Monthly Sum of Positive Tests	5,687,860	2,102,990	1,625,607	1,697,024	831,267
9	2022	Cumulative Positive Test	72,993,107	76,721,814	77,720,577	78,902,217	81,715,030
10	2022	Monthly Sum of 1st Covid Shot	7,352,967	2,643,305	1,523,949	1,529,910	1,311,401

Table 8: Transposed Data Sorted and Edited Continued

Obs	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1	2,476,880	4,406,097	5,874,328	7,084,590	8,998,888	13,444,501	19,694,998
2	.	.	.	.	.	.	5,504,758
3	.	.	.	.	.	.	36,849
4	833,500	1,870,266	1,412,019	1,169,543	1,867,535	4,304,708	6,059,498
5	32,532,211	34,050,793	38,266,551	42,340,798	44,791,056	47,356,230	53,349,300
6	9,375,389	9,411,671	13,355,270	8,152,432	6,372,295	11,563,469	9,982,649
7	19,017,261	8,562,363	9,364,106	10,001,433	7,601,084	4,472,061	8,448,278
8	377,572	1,341,593	4,138,884	3,862,665	2,286,750	2,417,097	5,857,798
9	84,840,692	88,530,240	91,551,803	302,448	.	.	.
10	1,329,268	1,559,736	1,173,306	606,501	.	.	.

Analyzing the Data

A line graph was used to shows a continuous visualization of the data over a period of time using the proc sgplot step using series statements. Three graphs were produced with each graph representing the variable Year. The x variable was set to represent the twelve months in a year. Three different y variables (Monthly_Sum_New, Monthly_Sum_Vaccinated_New, and Monthly_Sum_Full_Vaccinated_New) were used for visual comparison to show how each value changed during the same period of time represented in **Figure 1**, **Figure 2**, and **Figure 3**.

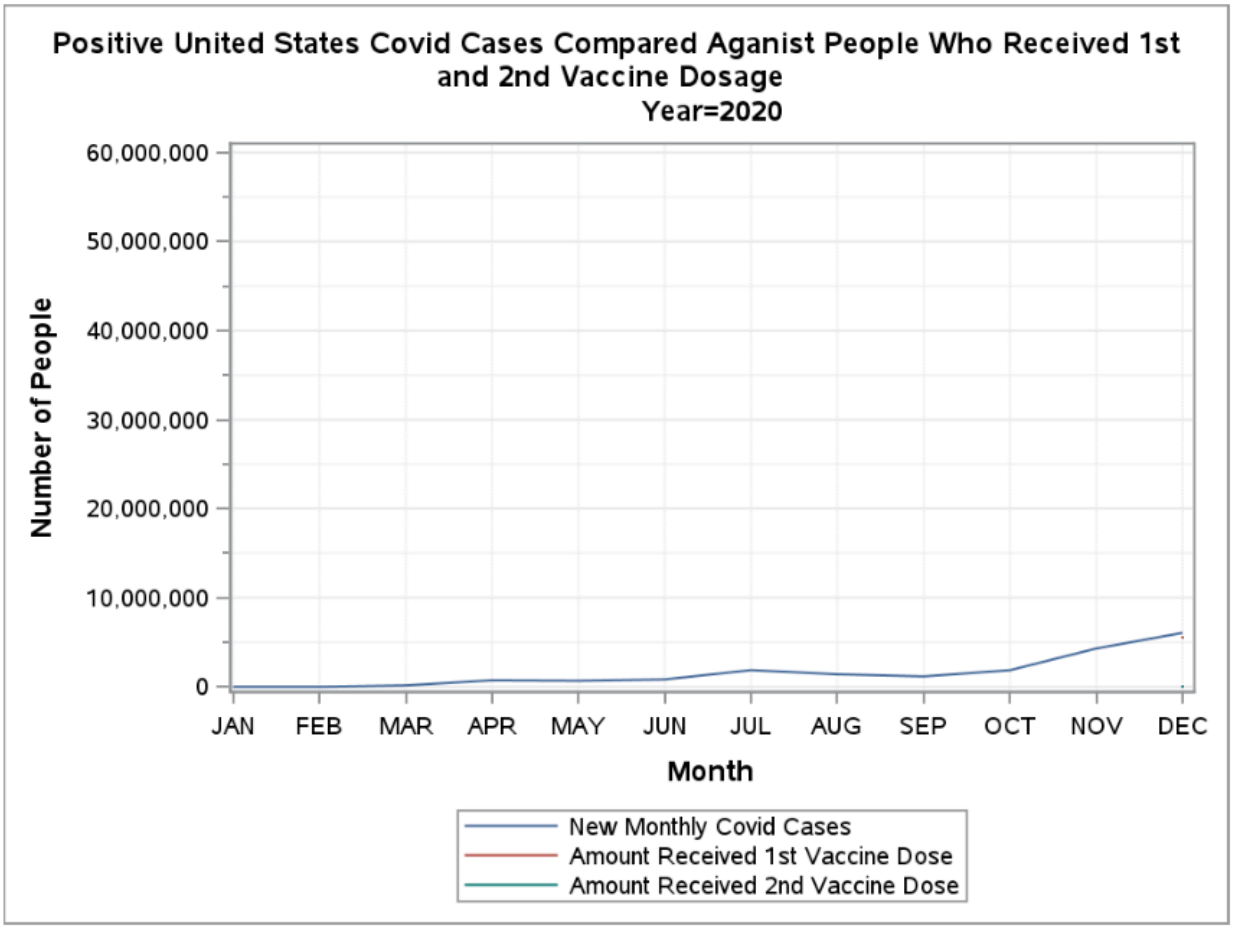

Figure 1: 2020 Line Graph of COVID-19 Cases and Vaccine Doses

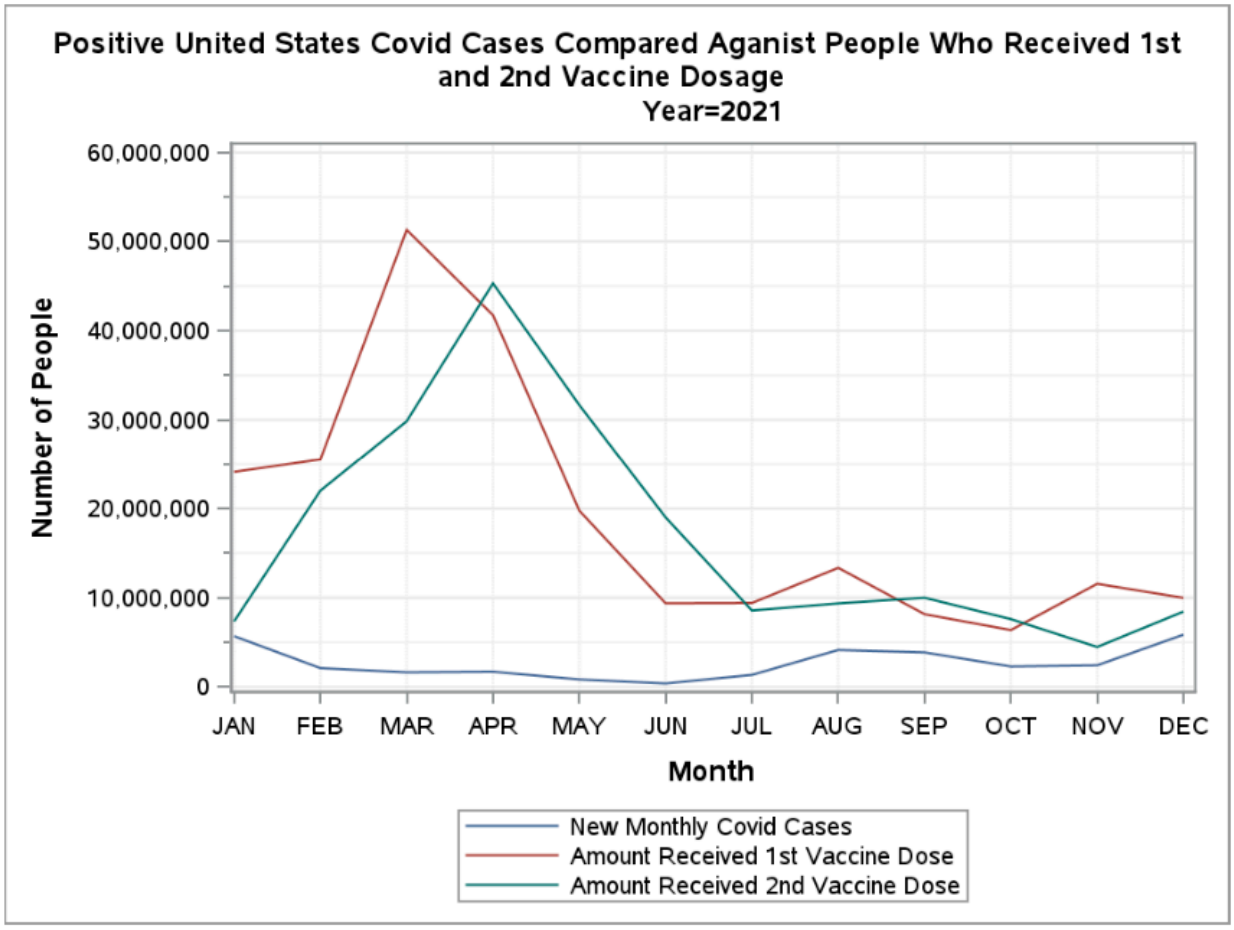

Figure 2: 2021 Line Graph of COVID-19 Cases and Vaccine Doses

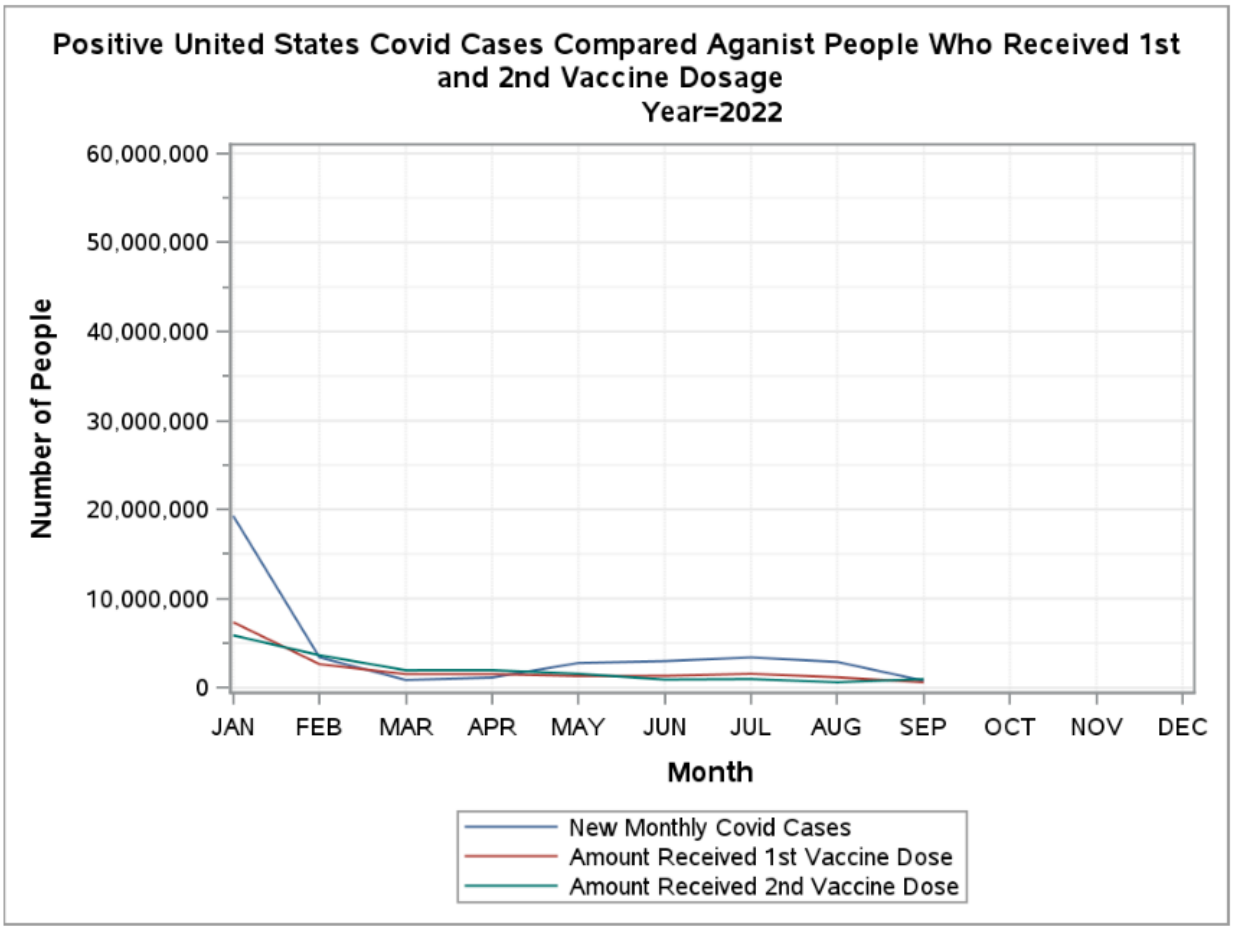

Figure 3: 2022 Line Graph of COVID-19 Cases and Vaccine Doses

A chart of the transposed data was also created to give an easier visualization of the data shown in **Figure 4**. The data was chosen to be transposed to allow for the data to be visualized wider format instead of having a narrow point of view. The data was then printed out using a proc print statement.

Year=2020

Vaccine Status	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
Cumulative Positive Test	9	82	158,732	916,239	1,626,491	2,476,880	4,406,097	5,874,328	7,084,590	8,998,888	13,444,501	19,694,998
Monthly Sum of 1st Covid Shot	.	.	.	.	.	.	.	.	.	.	.	5,504,758
Monthly Sum of 2nd Covid Shot	.	.	.	.	.	.	.	.	.	.	.	36,849
Monthly Sum of Positive Tests	10	72	158,632	735,275	680,589	833,500	1,870,266	1,412,019	1,169,543	1,867,535	4,304,708	6,059,498

Year=2021

Vaccine Status	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
Cumulative Positive Test	25,605,016	27,833,375	29,508,492	31,266,636	32,145,209	32,532,211	34,050,793	38,266,551	42,340,798	44,791,056	47,356,230	53,349,300
Monthly Sum of 1st Covid Shot	24,131,626	25,535,781	51,323,710	41,751,173	19,776,382	9,375,389	9,411,671	13,355,270	8,152,432	6,372,295	11,563,469	9,982,649
Monthly Sum of 2nd Covid Shot	7,336,605	22,019,808	29,845,969	45,317,704	31,664,764	19,017,261	8,562,363	9,364,106	10,001,433	7,601,084	4,472,061	8,448,278
Monthly Sum of Positive Tests	5,687,860	2,102,990	1,625,607	1,697,024	831,267	377,572	1,341,593	4,138,884	3,862,665	2,286,750	2,417,097	5,857,798

Year=2022

Vaccine Status	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
Cumulative Positive Test	72,993,107	76,721,814	77,720,577	78,902,217	81,715,030	84,840,692	88,530,240	91,551,803	302,448	.	.	.
Monthly Sum of 1st Covid Shot	7,352,967	2,643,305	1,523,949	1,529,910	1,311,401	1,329,268	1,559,736	1,173,306	606,501	.	.	.
Monthly Sum of 2nd Covid Shot	5,883,737	3,648,173	1,961,658	1,972,932	1,562,629	920,854	974,165	612,090	971,482	.	.	.
Monthly Sum of Positive Tests	19,293,809	3,402,565	871,622	1,142,441	2,776,951	2,985,527	3,397,529	2,892,177	785,170	.	.	.

Figure 4: Transposed COVID-19 Cases and Vaccine Doses

Figure 1 visualization is not as useful for analyze as the red line, representing Amount Received 1st Vaccine Dose, and the green line, representing Amount Received 2nd Vaccine Dose, do not appear until DEC. This would be due to the fact that the vaccine was not available until later in the pandemic. **Figure 2** shows the vaccine could possibly be having. As the red and green line increase, indicating more individuals getting the vaccine, the blue line, representing the monthly new COVID-19 cases, starts to decrease. As the months progress after APR, there are less individuals receiving the vaccine. There is an increase in the blue line representing an increase in confirmed COVID-19 cases. This pattern continues from **Figure 2** to **Figure 3** highlighting a pattern demonstrating the potential affects the vaccine has on reducing the spread of the disease. **Figure 4** shows a more numeric breakdown as showing as there was more vaccinations, the prevalence of the disease reduced. This helps to further my assumption that vaccines had a positive impact on reducing the spread of COVID-19.

After completing this project, I was able to gain a much better understanding of using SAS. I also learned how to use real life database and clean the data into a much more manageable form for analysis. The skills that I developed from STAT 4020 were able to be applied to real world data that can further be applied to workforce skills.

Exporting

Exporting the tables and graphs were able to be accomplished by starting the SAS file with ODS pdf and setting my output file path with the file statement. I personally chose to use the peralj style for export as out of all the styles available provided, in my opinion, the most professional result. To close the file, the ODS pdf CLOSE statement was used to stop the ODS statement and have the file produced.

References

Google. (n.d.). *Covid-19 open data - google health*. Google. https://health.google.com/covid-19/open-data/raw-data

N/A. (2024a). *Coronavirus*. World Health Organization. https://www.who.int/health-topics/coronavirus#tab=tab_1

N/A. (2024b, October 31). *Clinical guidance for covid-19 vaccination*. Centers for Disease Control and Prevention. https://www.cdc.gov/vaccines/covid-19/clinical-considerations/interim-considerations-us.html#:~:text=An%208%2Dweek%20interval%20between,pericarditis%20associated%20with%20these%20vaccines.

Author Information

My name is Justin Bowman, and I am currently a senior Computer Science major with a minor in math at Clemson. I worked for the Clemson's REDDI Lab which had a big impact on my choice for the topic. I have shifted from working for the REDDI Lab to working at Clemson's Office of Research and Organizational Development assisting with data collection. My focus is on furthering my knowledge of data collection to better prepare me for a job in the workforce.