**Difficulties and limitations.**

Ususally I would not start a project recap with the section ‘difficulties’, but in my case I thought it is necessary to explain my problems first, since it impacted the my whole vision of the project. During the project, I faced some tough hurdles and limitations, mainly related to getting Python set up for running FastAPI and/or Flask. These issues boiled down to two specific error messages that proved rather frustrating:

* "The term 'conda' is not recognized as the name of a cmdlet function script or operable program" error.
* "'Running cells with python.exe' requires the ipykernel package"

Despite my best efforts, including reinstalling Python, switching from conda to Python as the interpreter, and tweaking my PATH variable, which I kind of fixed, but then I started getting the seconf error message... I spent two evenings on the internet for potential fixes, but unfortunately, none of them seemed to work. After exhausting my options, I decided to at least write a hypothetical implementation of how I would implement an API is a seperate file, which you can find called ‘api’. It's important to note that I couldn't verify the correctness of the code since I couldn't run the app due to my errors, but I at least when I checked my code using good old Chat GPT 4, it did not see any major problems. Ofcourse, GPT cannot be trusted. I have attached my error screenshot in the projec files as well. Please note that all ‘import’ statements are underscored due to my unresolved environemtn issues. The libraries were installed beforehand but because of the PATH problems, could not be located (?)

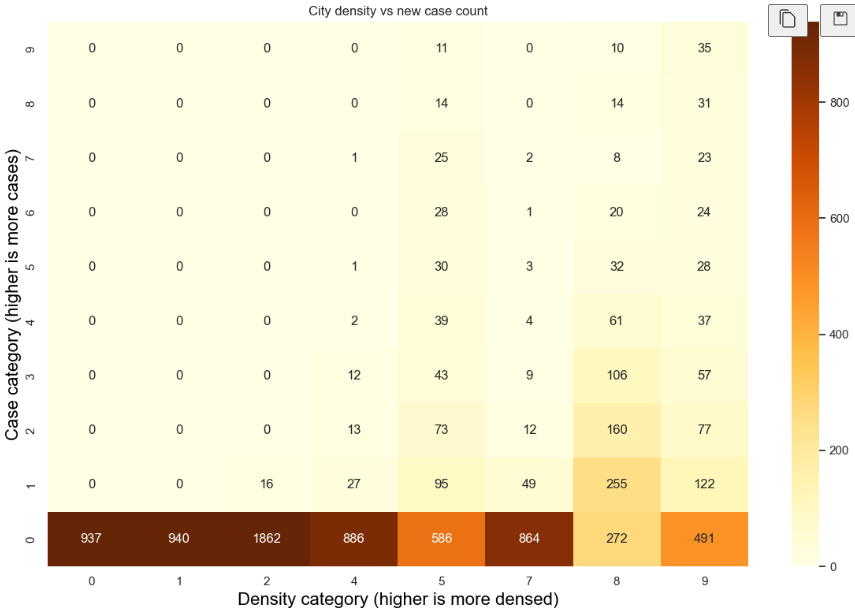
These roadblocks significantly altered the project's course, causing it to deviate from my original vision and the project requirements. To work around these issues, so that I can at least do the data analysis part, I created a function named 'get\_snow\_data.' This function connects directly to Snowflake using a user-defined SQL query as input. While this approach could have been adapted also for NoSQL databases, I opted not to incorporate semi-structured data into the project, so I did not define a similar function for that. These challenges served as valuable learning experiences, emphasizing the importance of a robust environment setup and effective troubleshooting skills in software development projects.

Moreover, these issues prevented me from visualizing the Dash apps in HTML, prompting me to plot graphs directly in Python. To facilitate this approach, I opted to use Jupyter Notebook, which allowed me to segment and organize different sections of my code more conveniently. These challenges served as valuable learning experiences, underscoring the significance of a robust environment setup and effective troubleshooting skills in the realm of software development.

I've included several datasets in my COVID-19 data project, and you can find the links to these datasets in the comments section of the .ipynb file. These datasets were stored in an AWS S3 bucket as CSV files, and I created a function called 'read\_from\_s3' to make it easy to retrieve them. This function connects to my AWS account, specifically targeting the 'myeduards' bucket, and users can input the name of the CSV file they want to access directly from the S3 bucket using this function. While it's possible to make this function more customizable by adding an extra input variable called 'bucket', which would allow users to specify a different bucket if needed, I personally didn't find it necessary for my project.

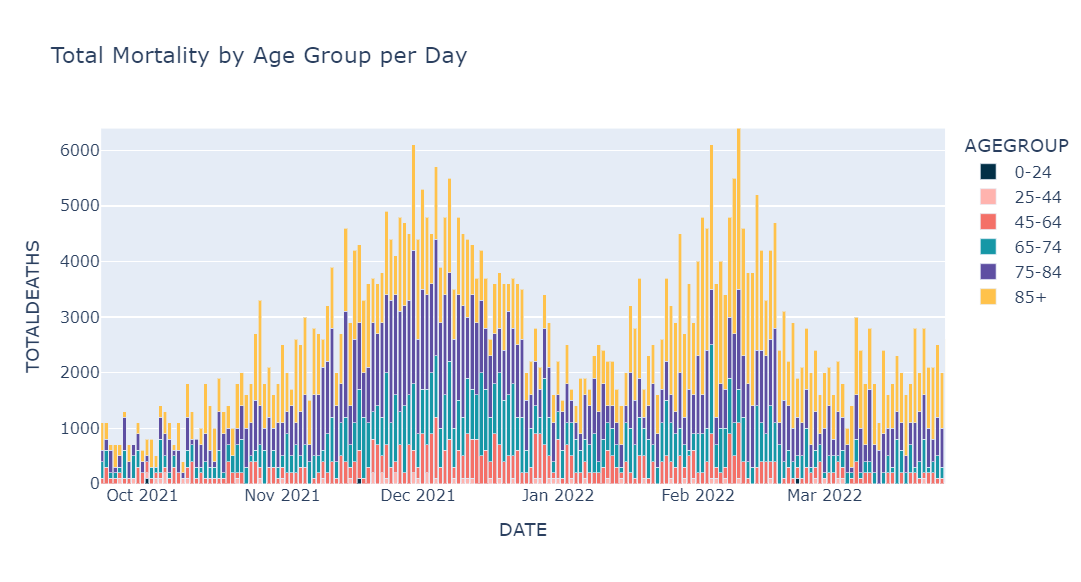
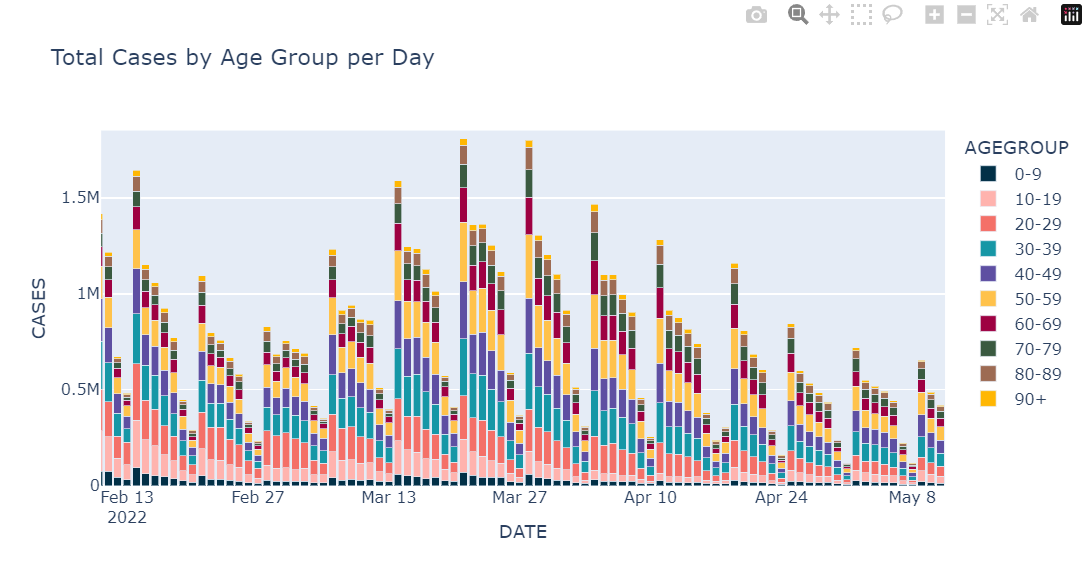
**Results**

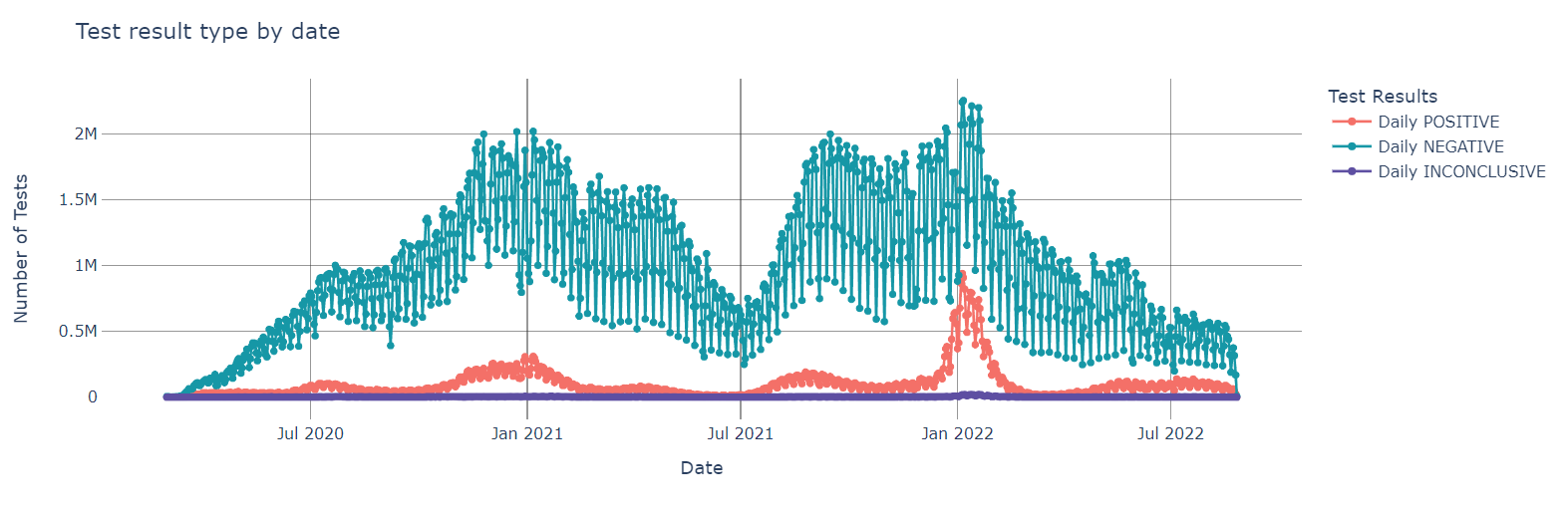
In order to visualize my analysis, I made use of both Dash and Seaborn. One particularly intriguing aspect of my analysis was exploring the correlation between city density (measured in people per square mile) and infection rates. To conduct this correlation analysis, I devised a method where I divided the daily number of cases into ten distinct categories. Python calculated these categories based on the minimum and maximum daily case counts, effectively creating intervals. For instance, Category 1 for daily cases would encompass values ranging from 1 to 50 people. I applied a similar categorization process to city density. The outcome of this analysis was a heatmap that visually represented the relationship between city density and daily case numbers. It's important to note that I filtered out two factors: Firstly, I excluded negative case counts, as the CDC occasionally adjusted the daily case figures retrospectively, causing inconsistencies in the data. For instance, if on January 1, there were initially reported 100 cases, but the CDC later corrected it to 70 cases, and on January 2, there were 80 cases, the data for January 2 would appear as -10, which could significantly impact our analysis. Secondly, I omitted the initial stages of the COVID-19 outbreak when most counties in the US had not yet reported any COVID-19 cases. This led to a daily case category of zero, even though it was irrelevant since the virus had not yet reached those areas.

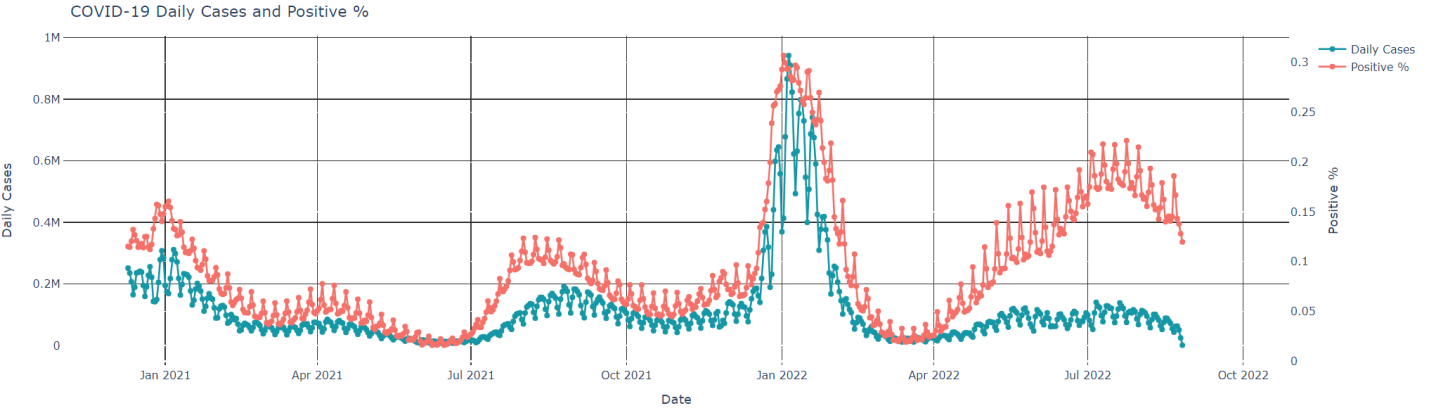


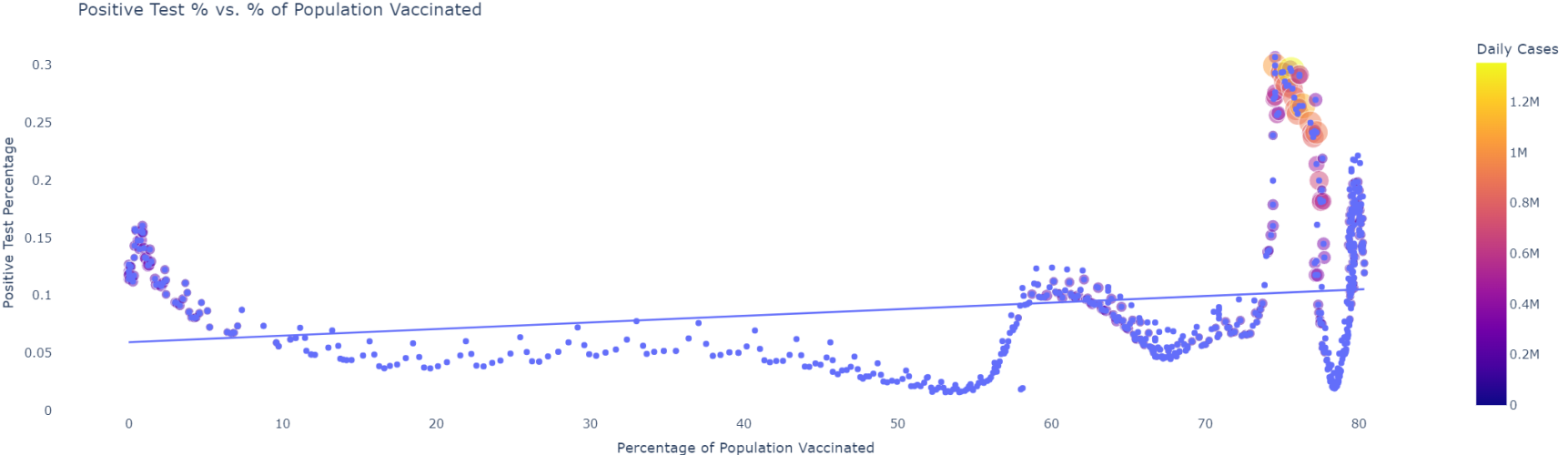
The results of the analysis both intrigued and perplexed me. Firstly, it was evident that counties with lower population density tended to have lower COVID-19 cases, which aligns with logical expectations. However, when it came to counties with higher population densities, the correlation between the number of daily cases and density was less apparent. In fact, it was more common to find counties with high density but relatively low daily case counts. This finding raises two possible explanations: firstly, that the effectiveness of lockdown measures was so pronounced that even densely populated urban areas experienced limited COVID-19 cases (nice joke), or secondly, that there might be some issues with the data itself.

For instance, in the case of San Francisco, there were numerous data entries where the daily case count remained unchanged, while the remaining data indicated very low daily case numbers. This doesn't align with reality, especially given that San Francisco, categorized as having 'extreme' density, later reported thousands of daily cases in 2022. It leads me to believe that there were inconsistencies or errors in data reporting during the earlier years of 2020 and 2021. Disregarding San Francisco's data, the patterns in the dataset become more evenly distributed across the various categorizations.



Another observation emerges when examining the charts depicting 'Total cases by age group per day' and 'Total mortality by age group per day.' It’s apparent that approximately one-third of the daily cases originates from age groups spanning from 0 to 39 years, while on the mortality chart, there were relatively few recorded cases of individuals younger than 44 succumbing to the virus. This stark contrast underscores the virus's disproportionate impact on the older generation compared to the younger one. Although it would have been valuable to create a heatmap similar to the ‘Daily cases versus city density’ analysis to visually represent this age-mortality correlation, due to time constraints in the project, such an investigation left on hold. But even an intern level Data analyst could guess that this correlation is strong.



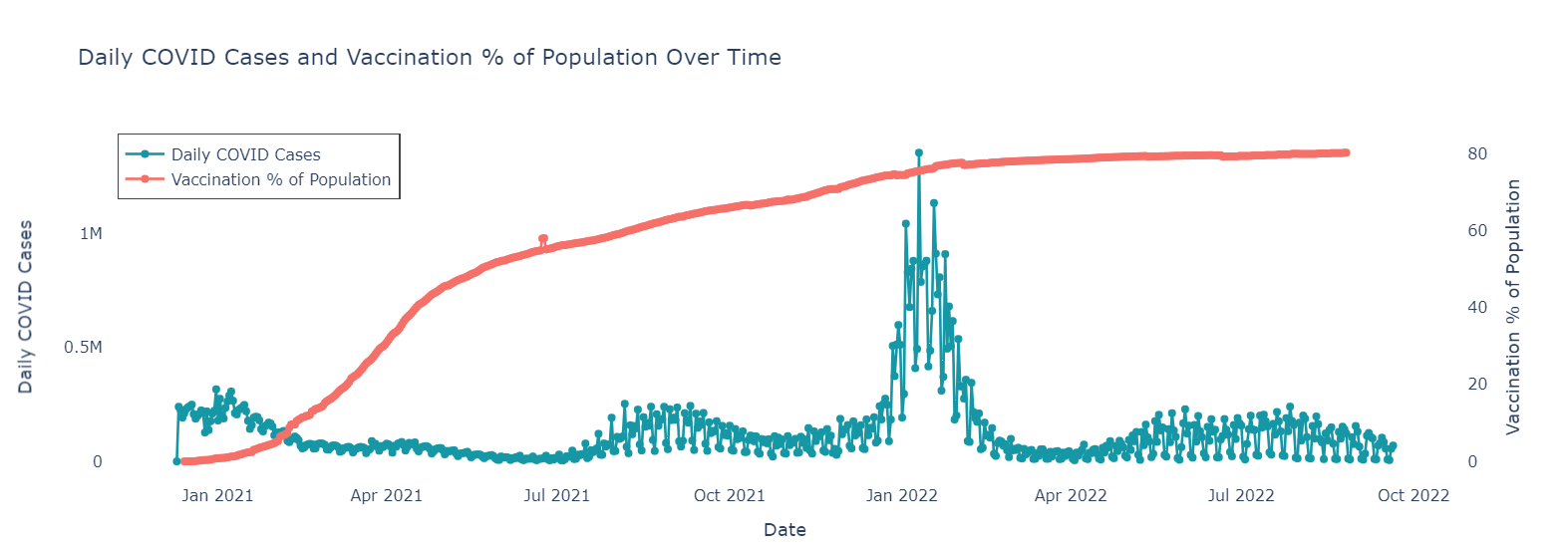


The three charts above reveal a curious pattern in our data. The initial chart underscores a straightforward correlation: as the number of individuals undergoing COVID-19 testing increased, the corresponding percentage of positive/negative/inconclusive cases naturally grew, which aligns with our expectations. However, when we factor in the second graph titled 'COVID-19 Daily Cases and Positive Percentage,' a more nuanced and compelling insight emerges. One of two things emerges from the data:

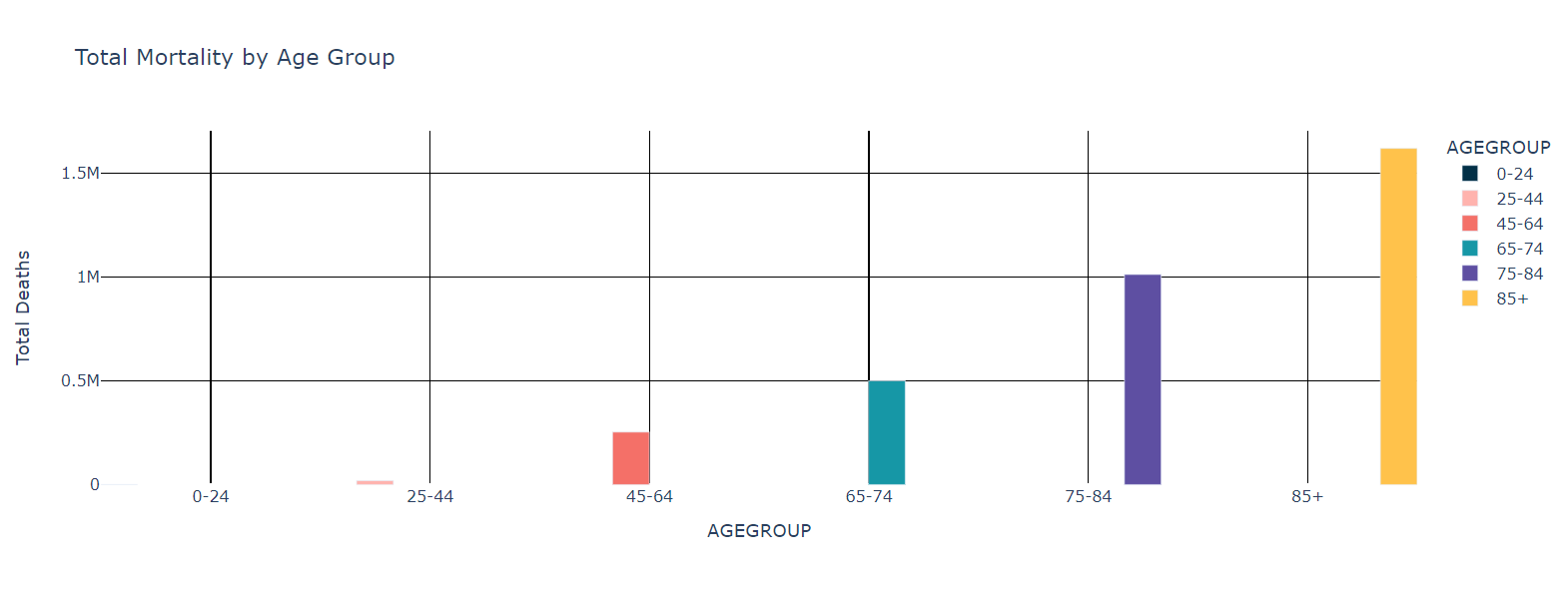
1. As the COVID-19 pandemic unfolded, our testing capabilities improved over time.

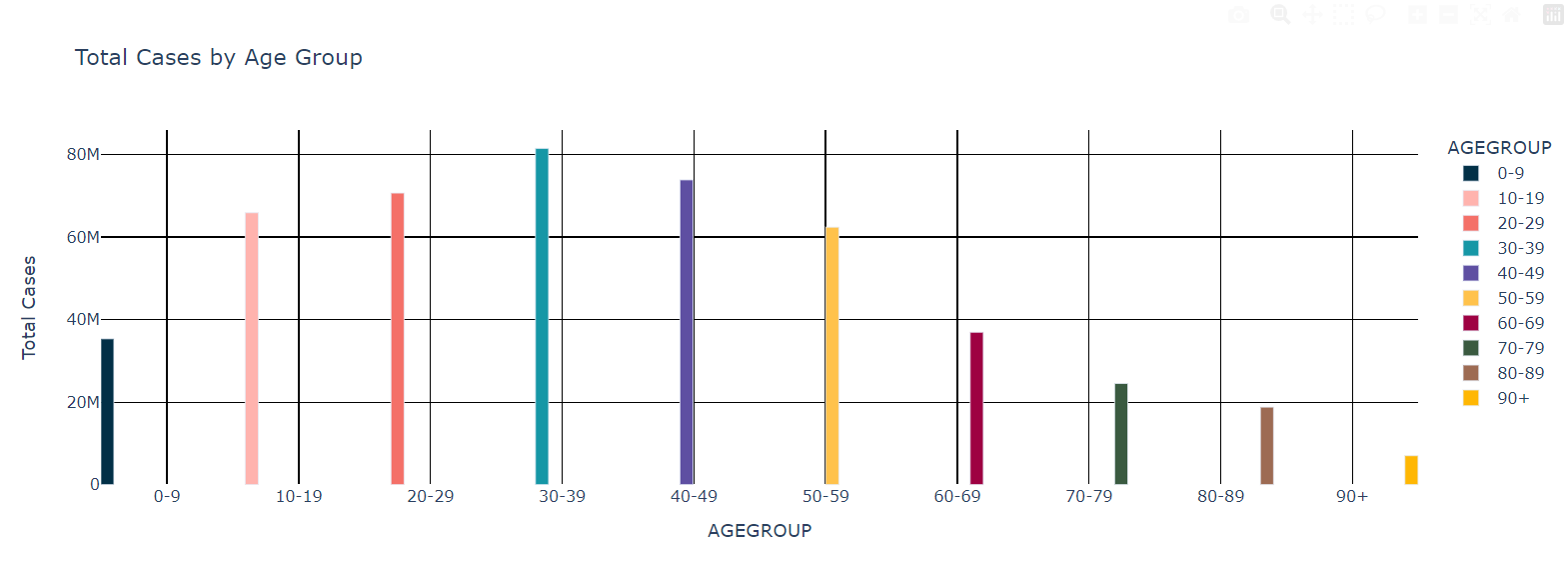
2. Our vaccines got less effective in battling COVID-19 as the new and more contageous variants emerged.

Starting from April 2022, there was a remarkable surge in the positive percentage relative to the total daily cases, showing either one of the conclusions above. To substantiate this trend, a third chart named 'Positive Test Percentage versus Percentage of Population Vaccinated' demonstrates a noteworthy correlation between the percentage of the population vaccinated and the positive test rate. This makes me believe that our vaccines indeed did not help in the battle of making the virus spread less fast.



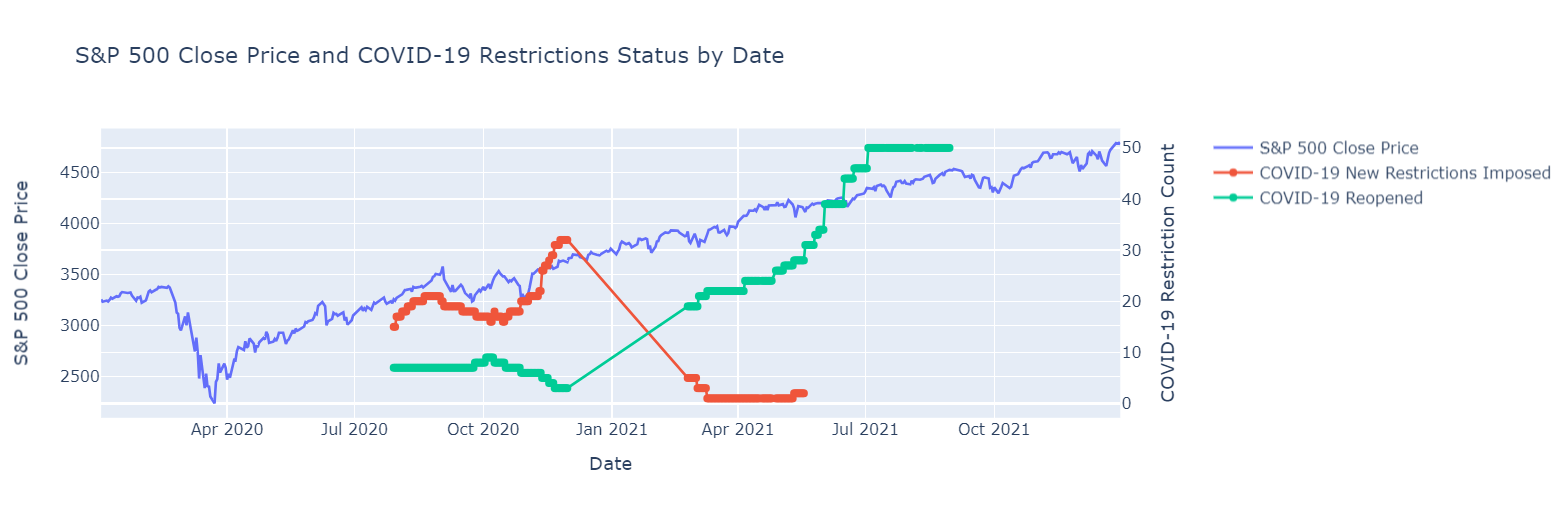
As vaccination rates rise, daily COVID cases initially drop, indicating a positive impact. However, fluctuations persist due to factors like new variants and waning immunity. Higher vaccination rates (above 60% where mass imunity starts to kick in) correlate with lower daily cases. Towards the end, vaccination rates plateau, posing concerns if many remain unvaccinated. Recent trends show low daily cases despite plateau, hinting at population immunity or effective treatments.





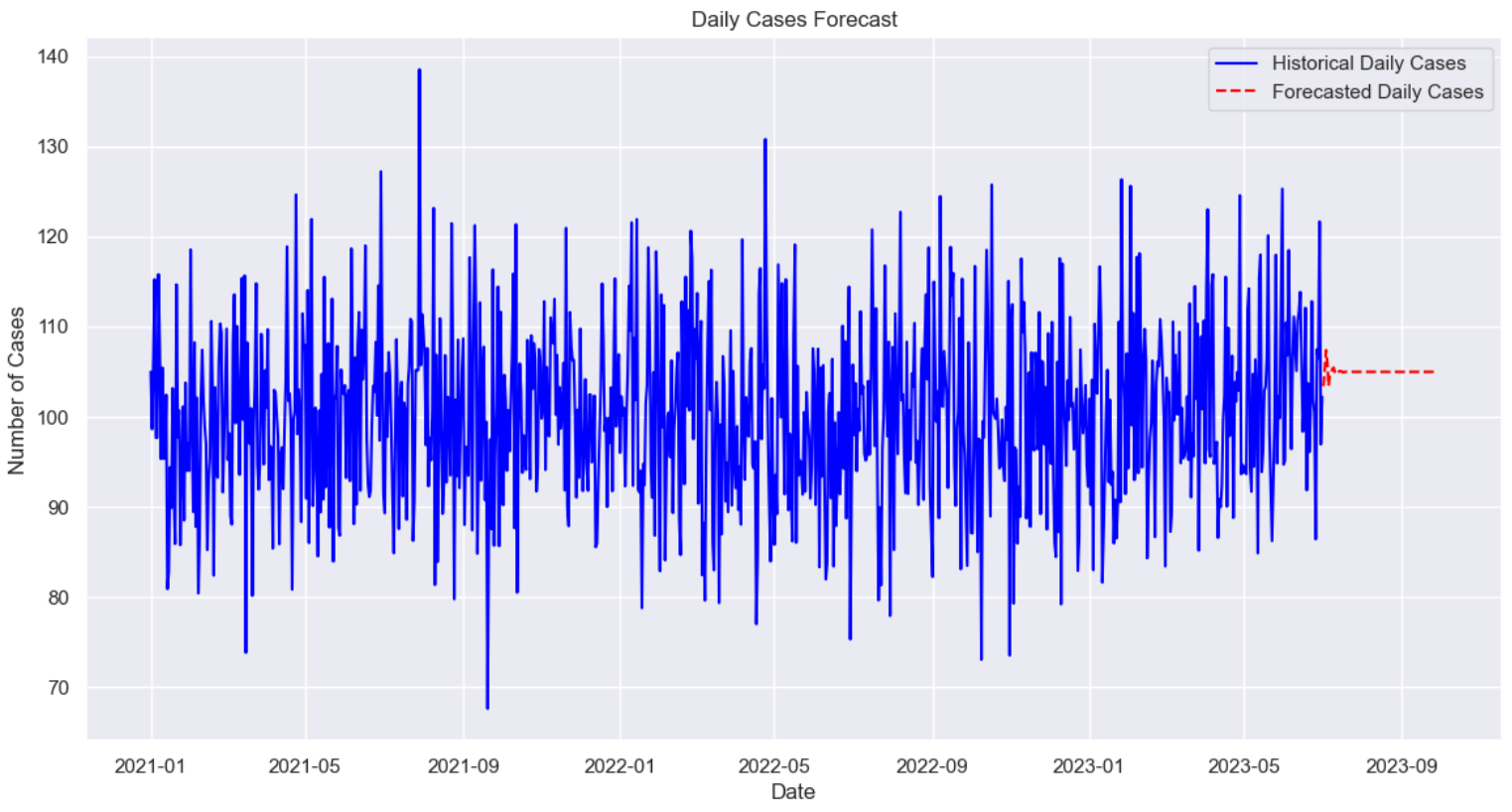
The above two charts further indicate the following factors:

* The younger you are, the more likely you will get COVID-19
* The older you are the higher the likelyhood of dying from COVID-19

These two points indicate that younger people are less avoidant to avoid socialization, which makes it spread faster in similar age groups as people typically communicate more with their age peers. But if indeed a person is on the older side, than it far more likely that a recovery will be tougher and in many more cases – fatal.

**BONUS!**

I also tried to inspect how COVID-19 restrictions impacted the stock market, but the data from the chart above is too inconclusive... There was not enough information about new imposed estrictions and reopenings, so there is a massive gap between December 2020 and April 2021. It is safe to assume that despite of the first wave of lockdows, which dropped the market in March – April 2020, new rules did not significantly impact the stock market.



I also tried to forecase the daily cases using a simple ML model, which did not work as expected as the input data is very sporadic, the model has a hard time to predict the daily forecast even in the near future as 30 days.