Analyzing Health Disparities during COVID

Avinash Akella (SID: 5719153)

Table of Contents

1. Introduction

The COVID-19 pandemic has highlighted social, and racial injustice in public health. It is crucial to develop a better understanding of these disparities, in order to design health policies that promote fairness. This work attempts to shed light on these differences, by using a combination of datasets to compare total cases, deaths, previous underlying conditions, and vaccination status for people in each race.

The analysis has been carried out in Databricks, with PySpark and SparkSQL being used the most. The code shifts to Python Pandas wherever convenient, especially during visualizations based on third-party library requirements. Majority of the data is pulled from an API provided by the Centres for Disease Control and Prevention (CDC). Other sources include Kaggle, and Kaiser Family Foundation (KFF). The analysis sports visual comparisons, which are summarized here. Finally, Spark's MLflow has been used to log a clustering model.

2. Data used for analysis

2.1 Sources and brief description

Majority of the data is pulled from CDC's "COVID-19 Case Surveillance Public Use Data with Geography" dataset. It has de-identified patient case data with 69.7 million rows, and 19 columns. Each record describes a person's race, ethnicity, state, county, vaccination status, and hospitalization status among a few other details. State-wise counts, and time series data for such attributes of interest are extracted from here. Vaccination trends are obtained from another CDC dataset: "COVID-19 Vaccination Demographics in the United States,National". It has 15.2K rows, and 16 columns. Each row describes a given demographic category's vaccination status on a given date for booster, and full vaccine doses, among other fields. Time Series information on vaccination for each racial group is taken from here.

Data on population estimates, reported asthma cases, and health access are taken from Kaiser Family Foundation (KFF) estimates. Finally, data on deaths due to Influenza, and Pneumonia are taken from Kaggle. Put together, these datasets help understand how the pandemic went for different races, and allow us to quantitatively compare them on common grounds. For specific details on their schema, please refer to the Appendix.

2.2 Methodology of Collection

The CDC dataset is too big to be uploaded into the Databricks workspace. Thankfully, the CDC dataset we're using here is a Socrata open dataset that can be queried using the Socrata Open DAtaset (SODA) API. Registering for an account with CDC and creating an App Token, gives more flexibility on the number of requests that can be made, and data throttling limits. This sql-like interface allows us to fetch only the data of interest, which is much more manageable.

For the other data sources, CSV files are uploaded into an Amazon S3 bucket. Data is fetched from here, or from CDC using an API, pre-processed in a separate Databricks notebook, and clean files are re-uploaded into S3 buckets under a different name. An IAM User was created

and given the "AmazonS3FullAccess" permission for this. Refer to the data-flow diagram below for a better understanding.

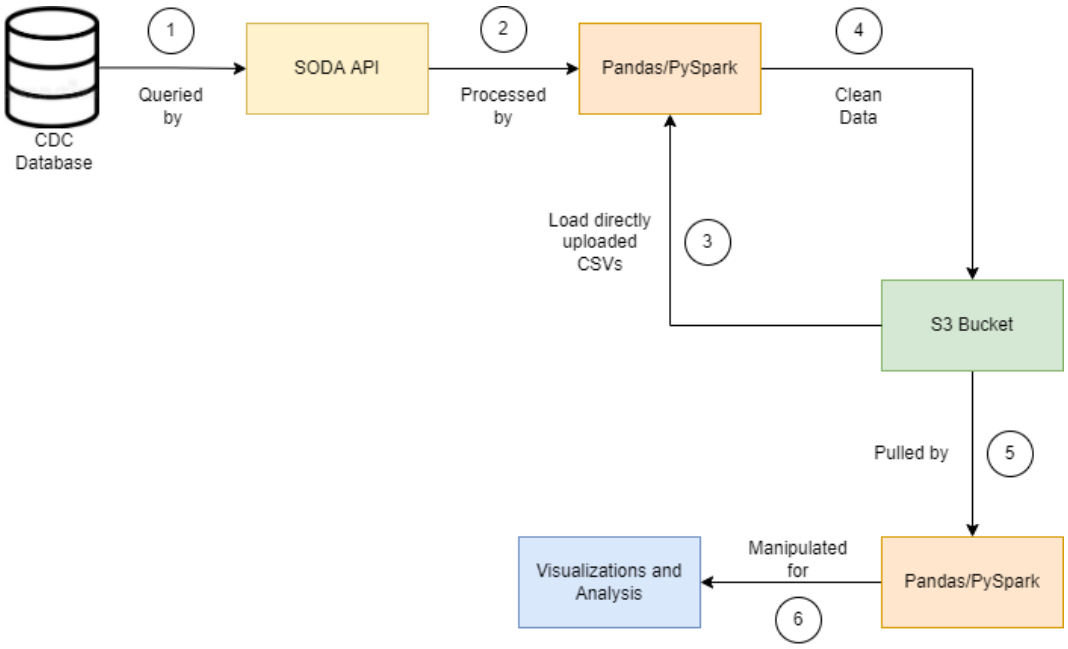

Figure 1: Data flow architecture diagram

3. Data manipulations / Pre-processing

The data was gathered from disparate sources, and consolidation was not easy as there were inconsistencies with date ranges, and column names. After lots of preprocessing, multiple intermediate representations were created to allow for effective visualization and analysis. The aim was to use the most appropriate, and efficient representation that is suitable for a query.

3.1 Description

Data collected from the CDC website had valid date ranges. As some of these datasets have information collected from patient forms, there are many missing or un-reported values that are left blank. These have been filled with 0 as a placeholder. For this dataset, we don't have to worry about fields marked as 'Missing' or 'NA' etc. This is because we plot the data with the Plotly library which only uses rows with valid states and FIPS codes. We even filter out information pertaining to our races of interest, so we don't need to (or can) make sense of the 'Missing'/'Unknown' race values.

Although, the KFF and Kaggle datasets needed special attention to fill missing values because aggregations on null columns gives a null overall result. The KFF population dataset had state names as a column, which had to be converted to state codes to match with that of CDC. A lot of inconsistencies exist in the CDC dataset, mostly because it is curated from patient forms. As the data is mostly missing at random, the best strategy would be to use only meaningful records and drop the rest. The dataset being large, we hope to have enough information to draw meaningful insights.

Basic data visualizations of different factors grouped by the percentage of people affected in each race, gives a sense of what the racial disparity looks like. Native Hawaiian and American Indian populations saw the highest percentage of cases, while American Indians and Whites saw the most deaths. The African American population had a higher percentage of underlying conditions, including Asthma, Pneumonia, and Influenza, while they were the last in vaccination percentages. The white population was second to last in vaccination percentages.

3.2 Summary Visualization

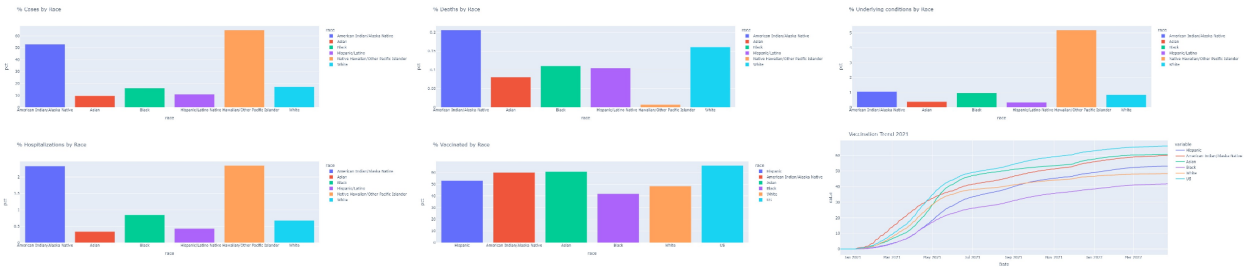

Figure 2: Some analyses from data exploration, from left to right, top to down: % of cases, % of deaths, % of underlying conditions, % of hospitalizations, % of vaccinations, and vaccination trend, all calculated by race.

4. Analysis

4.1 Questions

How were the cases in each state, by race?
The case percentage is noticeably different for each state and race. The White population seems to have been affected more than the rest. Arkansas, Louisiana, New York see higher % of cases for the African American population, compared to the rest. A majority of the American Indian/Alaska Native population in Arizona have been affected by COVID.

What was the case trend like for each race?
The American Indian/Alaska Native communities have been severely affected, almost always topping the charts. The more recent wave (January 2022), was the worst ever for Native Hawaiian/Other Pacific Islanders. The % of cases for the African American, Asian, and White populations have been varying a lot, beating each other in different times.

What were the deaths in each state, by race?
Most of the deaths seemed to have occured in California, Nevada, Florida and New York. The African American community was the worst hit in New York, which is an important find because New York was among the worst hit states during COVID. Their death rates are nearly 2.5 times that of Whites, and Asians. Florida, Nevada, and California were the worst for Whites, with their death rates being 2 to 3 times that of Asians, African Americans, and Hispanic/Latinos.

What percentage of cases led to deaths for each race? What was that trend like?

The conversion is high for Whites in general, but during peak covid periods, it's slightly worse for Hispanic/Latinos, and sometimes Asians. Maybe they're not given priority during these periods? The disparity is quite large in a few months: July 2020, December 2020, and November 2021.

*What was the percentage of deaths **per population**, for each race?*
In more recent times, this value has reduced for the Asian and the Hispanic community upto some extent, even at times when it increased for the Whites. But it seems to have stayed the same/worsened for the African American community. Overall, Whites have a higher death % than the other races.

This is interesting, because the % of deaths for the other races seems smaller, but the proportion of cases that are leading to deaths are higher for them. Could it be a difference in access to health care? different underlying conditions? or vaccination status? These are some of the questions we'll try to address next.

How were the races affected by underlying health conditions?
Some states have a comparatively higher proportion of populations with some underlying health conditions, and these seem to be shared across races. Interestingly, Louisiana, Arkansas, and New York see a high proportion of African American population with underlying health conditions, which is higher than that for any other race in these states. This could explain why these states saw the highest proportion of cases for this race.

Also, New York, one of the majorly hit states, sees a higher proportion of Hispanic, African American and Asian population with underlying health conditions compared to Whites. African Americans are 4 times, and Asians and Hispanic/Latinos are 2 times as much affected.

How were the races affected by covid/related diseases? Such as Asthma, Pneumonia, Influenza?
The relative vulnerability of people belonging to different races seems consistent across different diseases. African Americans are slightly more vulnerable compared to Whites, while the American Indian/Alaska Natives are especially vulnerable. This seems consistent with the overall death rates due to COVID alone, and hints at a disparity that is more general across multiple diseases.

The African American population seems to have a high proportion of Asthma rates in a few states. American Indian/Alaska Natives, and Hispanic/Latinos also have disproportionately high rates in few states like Virginia, Michigan, and Nebraska, where it's nearly twice that of Whites.

But do people of different races have equal access to hospitalization?

How was the health access and hospitalization rate for each race?
This shows a stark disparity between different races. A comparatively high % of the Hispanic/Latino community report not visiting their doctors in the past 12 months because of costs. The disparity is quite high in the South and Midwest. Whites seem to be doing much better compared to the rest, indicating they have better access / lesser cost hindrance to health care. An agglomerative clustering method allows us to compare health access across different

states and understand what general category a state falls into. The clustering result is attached in Section 4.3.

But what did the hospitalization rates really look like during covid?
Hospitalization was very high for American Indian/Alaskan Natives in Arizona. It was particularly high for African Americans in New York, which makes sense as they saw a lot of cases in that state. Hospitalization was quite low for Hispanic/Latinos, which could be because they didn't have proper access to health care.

What did the trend for percentages of hospitalizations look like?
As suspected, hospitalization was quite low for Hispanic/Latinos. African Americans were hospitalized at a higher rate during many periods, despite many of them complaining of improper access to health care. One possible explanation for this is: They were severely affected, and had no choice but to get admitted to the hospital.

Let's attempt to validate this claim, at least to some extent. One way to do this is to see:

What the ICU admittance rates were for people belonging to different races?
The numbers are very low to make a solid comparison. One seemingly interesting point is that the % of ICU admits for the African American population in New York is higher compared to that of other races. This suggests it was indeed serious for them. Looking at the % of ICU admit per cases trend, the ordering from worst to best was roughly: Native Hawaiian/Other Pacific Islander, Asian, Hispanic/Latino or Black, and White.
So it seems like it was getting serious for them, but we don't see a corresponding increase in their hospitalization rates.

May CDC articles point out that a lot of cases and deaths could have been avoided with vaccinations, But

How is the vaccination response for each race?
Asians are front-runners in vaccinations. African Americans seem to lag behind everyone else, followed by Whites. The percentage gap between them is quite significant, indicating hindrances or preferences that are beyond the scope of this data.

4.2 Summary

The disparities in COVID-19 cases exist across multiple factors. These disparities are quite nuanced, and they surface when analyzed at specific conditions. Racial/ethnic minorities have lower access to health care, and it seems to affect their hospitalization rates, even when they seem to really require it. Vaccinations trend tries to explain why the COVID case percentage might be high for the White, and African American populations. Lastly, a disparity exists for other underlying / related conditions as well, and its nature closely aligns with that of COVID.

4.3 Figures

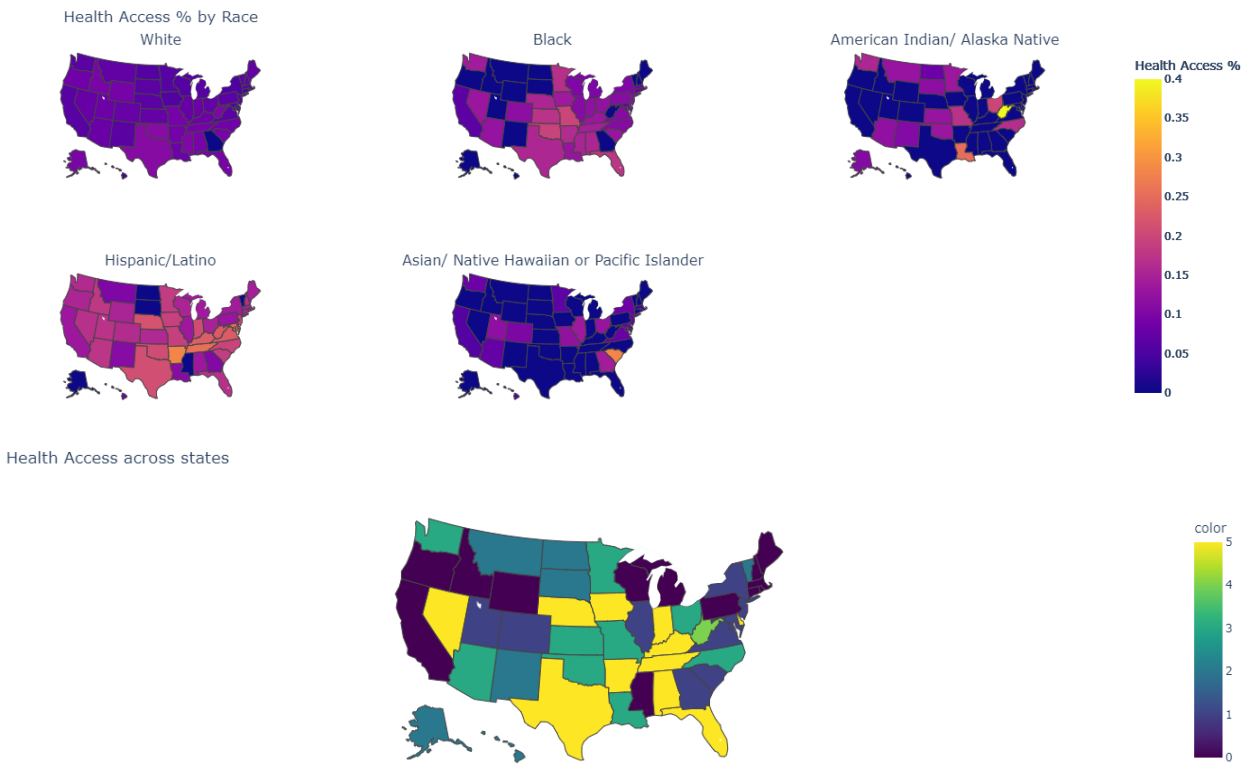

Figure 3: Top: Health access % per race. Bottom: Clustering of states based on health access rates.

5. Lessons learned

The entirety of the project was done on Databricks, so familiarity with the environment was a natural consequence. Checking job logs after executions, and experimenting with cluster settings helped write more efficient code and use compute resources more effectively.

Another good implementation was to mount the S3 bucket to the DataBricks FileSystem (DBFS), so it can virtually be used as a local filestore, without syncing data locally. This saves a lot of space in the DBFS for intermediate outputs to be stored. An additional nuance here is to store the access keys (for connecting to S3) in Spark secret keys. Although this is not available for the Databricks community edition, it's an essential feature for Standard.

The main challenge in the project was to analyze the large dataset and convert it into a manageable form. Its sheer size is a major hindrance to even basic analytical questions. Fortunately, Spark SQL provides an effective way to query these datasets to extract basic statistics. Creating views provides an abstraction, allowing one to explore the dataset and store query results. Although this was initially done with the entire dataset hosted on S3, we migrated to SODA API as it seemed like a more viable option. It essentially has the same effect by querying over the dataset and storing its results after pre-processing.

Lazy evaluations save time and allow Spark to optimize execution. Errors, if any, aren't raised until an action such as display or show are performed. Chaining of queries is done wherever possible to reduce serialization and deserialization. A broadcast hash join was compared with the normal join, and its advantages in terms of speed are immediately apparent. The population dataset here is pretty small compared to the rest, so a broadcast join is appropriate.

Lastly, it was interesting to explore the benefits of MLflow, and how it can be useful for production. Models fine-tuned with Hyperopt and SparkTrials, can be logged for future use, or hosted as a service (only in Standard edition).

6. Future Work

The analysis performed in this work has mostly been exploratory. A lot more data and a lot of statistical testing would be required to validate the claims made here. The CDC data has a lot of inconsistencies, in that the information on deaths, hospitalizations, underlying conditions, etc are not fully recorded. For instance, the CDC reports race/ethnicity of people vaccinated at the national level, but the race/ethnicity information is missing for nearly 40% of the vaccinated people. Moreover, this is not reported at the state level, limiting understanding of how disparities may vary across the US. If the data is not missing at random then, ironically, this would be a disparity or bias in itself.

Another data related issue is that the population information is an estimate from the 2020 Census, so it's not accurate. In fact, this leads to the percentage of cases in Arizona for American Indian/Alaska Natives to go beyond 100%. Assuming we have all the requisite data, a more thorough analysis of the time series data would involve comparing differences, ratios, or cumulative values of the time series. To make it more complicated a Granger test could be performed to see if one series can be used to forecast another. Lastly, an ARMA model can be used and an F-test can be constructed to test the hypothesis of a common set of parameters.

7. Notebook Instructions

Notebook to run: "Covid Health Equity"
Contains data analysis, visualizations and summary.
Environment details: 15.25 GB, DBR 10.4 LTS ML Cluster with Spark 3.2.1 and Scala 2.12.
Default interpreter: Python

Reference notebook: "PreProcessing"
Contains data pre-processing, and schema summary.

8. References

1: Advancing Health Equity Requires More and Better Data. Link
2: Health Equity Considerations and Racial and Ethnic Minority Groups. Link
3: Disparities in COVID-19 Illness. Link
4: COVID-19 exacerbating inequalities in the US. Link
5: COVID-19 preventable mortality. Link
6: Unvaccinated COVID-19 hospitalizations cost billions of dollars. Link
7: Comparison of two time series. Link
8: Connecting to S3. Link1, Link2, Link3
9: Creating Secrets. Link1, Link2

9. Appendix A: Schemas

The schemas of some of the tables are listed below. For the tables not listed, their schemas are simple and straightforward. Refer to the 'PreProcessing' code 'notebook for full details of schemas and sources.

Column Name	Description	Type
case_month	The earlier of month the Clinical Date (date related to the illness or specimen collection) or the Date Received by CDC	Plain Text
res_state	State of residence	Plain Text
state_fips_code	State FIPS code	Plain Text
res_county	County of residence	Plain Text
county_fips_code	County FIPS code	Plain Text
age_group	Age group (0 17 years; 18 49 years; 50 64 years; 65 years; Unknown: Missing: NA if value suppressed for privacy protection.)	Plain Text
sex	Sex (Female: Male: Other Unknown: Missing: NA. if value suppressed for privacy protection.)	Plain Text
race	Race (American Indian/Alaska Native: Asian; Black; Multiple/Other Native Hawaiian/Other Pacific Islander: White: Unknown: Missing NA if value suppressed for privacy protection.)	Plain Text
ethnicity	Ethnicity (Hispanic Non-Hispanic: Unknown: Missing: NA if value suppressed for privacy protection.)	Plain Text
case_positive_specimen_interval	Weeks between earliest date and date of first positive specimen collection	Number
case_onset_interval	Weeks between earliest date and date of symptom onset	Number
process	Under what process was the case first identified? (Clinical evaluation: Routine surveillance; Contact tracing of case patient; Multiple: Other Unknown: Missing)	Plain Text
exposure_yn	In the 14 days prior to illness onset, did the patient have any of the following known exposures: domestic travel, international travel cruise ship or vessel travel as passenger or crew member workplace airport/airplane, adult congregate living facility (nursing assisted living. or long-term care facility). school/university/childcare center, correctional facility, community event/mass gathering animal with confirmed or suspected COVID-19 other exposure contact with a known COVID-19 case? (Yes Unknown, Missing)	Plain Text
current_status	What is the current status of this person? (Laboratory-confirmec case, Probable case)	Plain Text
symptom_status	What is the symptom status of this person? (Asymptomatic, Symptomatic Unknown, Missing)	Plain Text
hosp_yn	Was the patient hospitalized? (Yes, No. Unknown, Missing)	Plain Text
icu_yn	Was the patient admitted to an intensive care unit (ICU)? (Yes, No. Unknown, Missing)	Plain Text
cleath_yn	Did the patient die as result of this illness? (Yes: No: Unknown: Missing: NA if value suppressed for privacy protection.)	Plain Text
underlying_conditions_yn	Did the patient have one or more of the underlying medical conditions and risk behaviors diabetes mellitus, immunosuppressive condition, autoimmune condition, current smoker, former smoker substance abuse or misuse, disability psychological/psychiatric pregnancy. other. (Yes, No. blank)	Plain Text

Figure 4: COVID-19 Case Surveillance Public Use Data with Geography. Source

Column Name	Description	Type
Date	Date data are reported on CDC COVID Data Tracker	Date & Time
Demographic_category	Age, sex or race/ethnicity of person receiving vaccination	Plain Text
Administered_Dose1	Total count of people with at least one dose in demographic category	Number
Administered_Dose1_pct_known	Percent among persons with at least one dose who are Hispanic/Latino	Number
Administered_Dose1_pct_US	Percent among persons with at least one dose, who have demographic information available on age, race/ethnicity or sex	Number
Series_Complete_Yes	Total count of fully vaccinated people in demographic category	Number
Administered_Dose1_pct_agegroup	Percent among persons with at least one dose in demographic category	Number
Series_Complete_Pop_pct	Percent among fully vaccinated persons in demographic category.	Number
Series_Complete_Pop_Pct_known	Percent among fully vaccinated persons who are Hispanic/Latino	Number
Series_Complete_Pop_Pct_US	Percent among fully vaccinated persons, who have demographic information available on age, race/ethnicity or sex	Number
Booster_Doses_Vax_pct_agegroup	Percent of people 12+ in a demographic category with a booster dose	Number
Booster_Doses_Pop_Pct_known	Percent of people 12+ with a booster dose where selected demographic category is known	Number
Booster_Doses_Vax_Pct_US	Percent of people 12+ with a booster dose who have known demographic information	Number
Booster_Doses_Pop_Pct_known_Last14Days	Percent of people 12+ with a booster dose in the last 14 days where selected demographic is known	Number
Booster_Doses_Yes	People 12+ with a booster dose	Number
Booster_Doses_Yes_Last14Days	People 12+ with a booster dose in the last 14 days	Number

Figure 5: COVID-19 Vaccination Demographics in the United States,National. Source

Column Name	Description	Type
Location	State name	Text
All Adults	% who report having asthma among adults of all races	Number
{All other columns}	% who report having asthma among people of that race	Number
Footnotes	Column for collection notes (dropped)	Text

Figure 6: Adults Who Report Currently Having Asthma by Race/Ethnicity. Source

Column Name	Description	Type
Data as of	Last update date	Date
Start Date	Start date of data collection	Date
End Date	End date of data collection	Date
State	State name	Text
Age group	Age group under consideration	Text
Race and Hispanic Origin Group	Race/ethnicity of group under consideration	Text
{All other columns}	Deaths related to that particular disease	Number
Footnotes	Column for collection notes (dropped)	Text

Figure 7: Conditions contributing to deaths COVID-19. Source

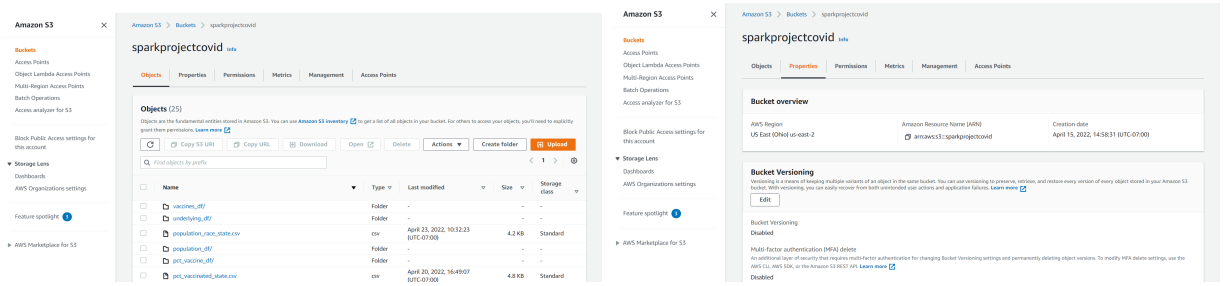

Figure 8: Amazon S3 Bucket

10. Appendix B: Visualizations

Some visualizations from the analysis are shown below for reference and understanding. Refer to the 'Covid Health Equity' notebook for an interactive visualization and full details on how they're created.

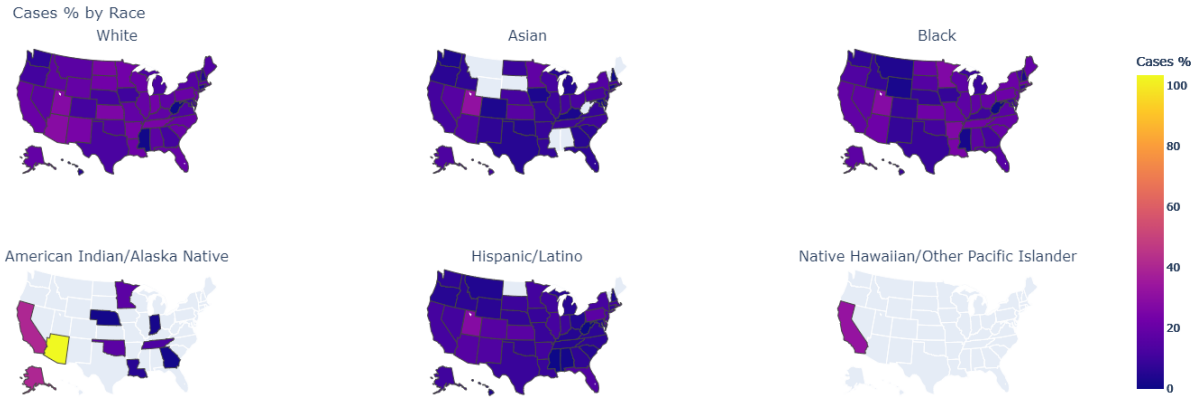

Figure 9: Case % per race

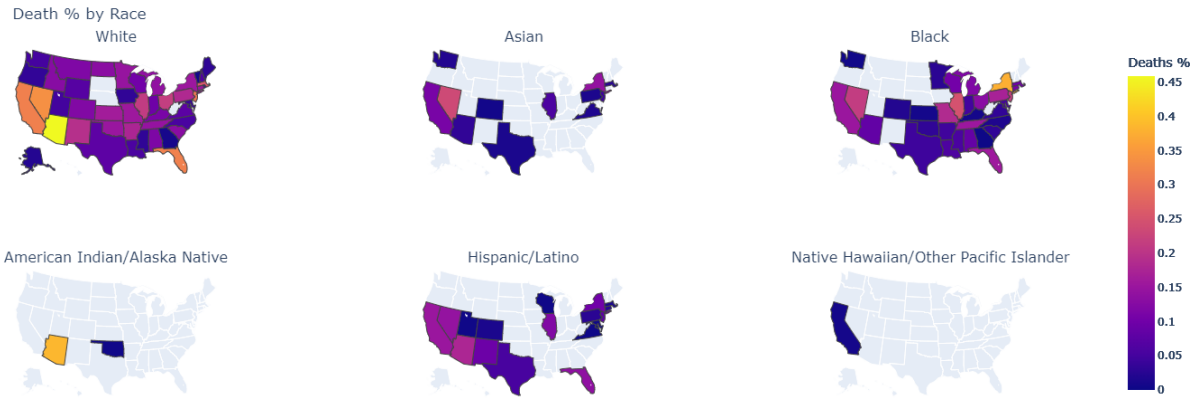

Figure 10: Deaths % per race

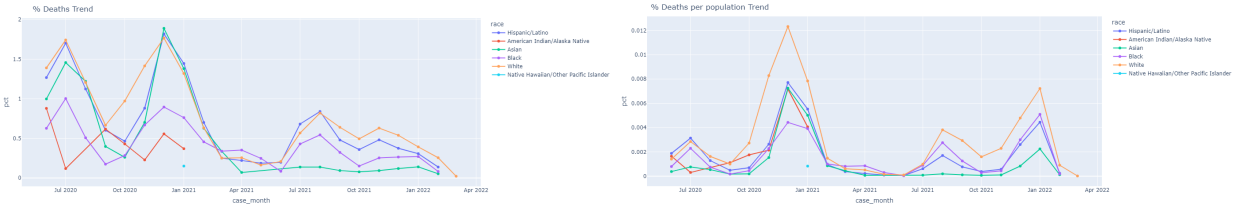

Figure 11: Left: % of cases that led to deaths. Right: % of population that had COVID deaths.

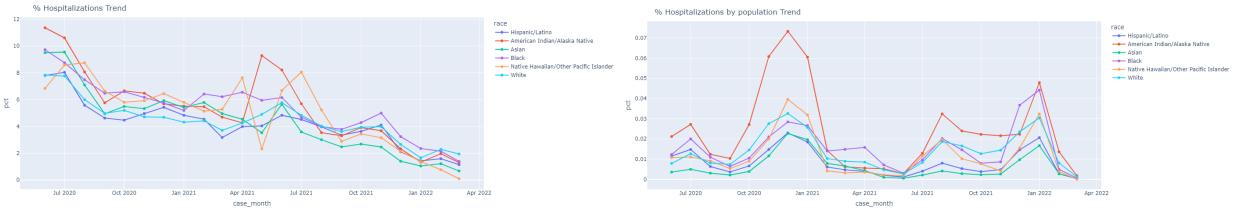

Figure 12: Left: % of cases that led to hospitalizations. Right: % of population that was hospitalized

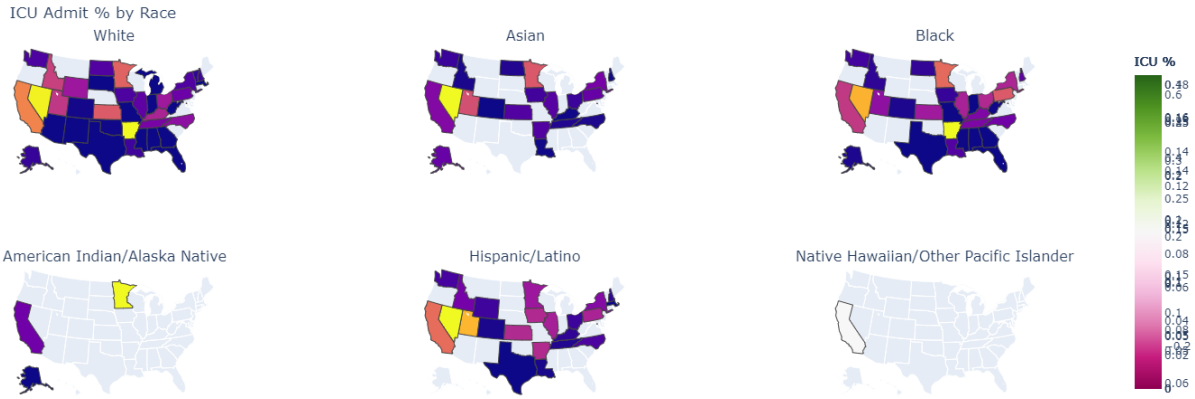

Figure 13: ICU Admit % per population.

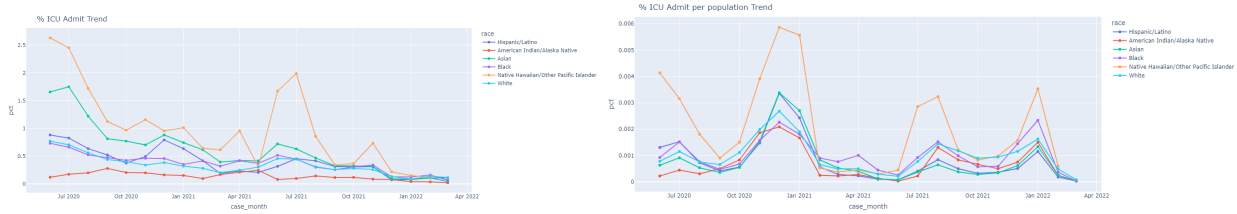

Figure 14: Left: % of cases that required ICU admittance. Right: % of population that required ICU admittance.