An Analysis of the Effects of COVID-19 on the NYC Area and General Populus

Jana Alghamdi, Sharazad Ali, Yurie Han, Trey McCray, Andres Villada

Introduction:

As of 2023, we live in a relatively post-pandemic era. COVID-19 affected us all in some way or another so we were intrigued to see what data trends reveal, on both a national and highly populated area, as the pandemic went on and what lessons we could learn from these trends to better prepare ourselves for future national health crises. In addition we also wanted to look at the COVID-19 history in particular in the NYC area as a small part because of how densely populated the area is. Which would allow us to add much more data in an easier way to make a comparison to the general population. We also wanted to see within the data trends, segment groupings based on age groups and/or racial demographics. We believed that certain marginalized demographics would be more negatviely effected by the pandemic due to lacking more financial resources on average to afford treatment. Additionally, we anticipated that more densely populated areas would exhibit more severe and common cases of COVID-19. Overall, the purpose of our overall study was to make an larger analysis of a known topic in COVID-19. To educate ourselves personally on its effects in the general population using many different forms of statistical analysis and modelling in general. Through the techniques we learned in class from sql databsing to making conditional plots we were able to come to our own conclusions as a group on COVID-19.

Primary Data Set:

Our primary data set is the "COVID-19 Case Surveillance Public Use Data" published by the U.S. Department of Health & Human Services. Featuring 95M+ rows, this case surveillance public use data set has 12 elements for all COVID-19 cases shared with CDC and includes demographics, any exposure history, disease severity indicators and outcomes, presence of any underlying medical conditions and risk behaviors, and no geographic data. The COVID-19 case surveillance database includes individual-level data reported to U.S. states and autonomous reporting entities, including New York City and the District of Columbia (D.C.), as well as U.S. territories and affiliates.

source: https://catalog.data.gov/dataset/covid-19-case-surveillance-public-use-data

Secondary Data Set:

1) Our first secondary data set is the "Hospital Inpatient Discharges (SPARCS De-Identified): 2019" published by the New York State Department of Health. This 2M+ row data set contains patient information for the 2019 year including discharge level detail on patient characteristics, diagnoses, treatments, services, and charges. This data file contains basic record level detail for the discharge. Although this 2019 year data set does not contain any mention of COVID-19 due to being before the pandemic, we wanted to use it to analyze just how a normal year of inpatients would look like prior to the pandemic for our future observations.

source: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/4ny4-j5zv

2) Our second secondary data set is the "Hospital Inpatient Discharges (SPARCS De-Identified): 2020" published by the New York State Department of Health. This 2M+ row data set contains patient information for the 2020 year including discharge level detail on patient characteristics, diagnoses, treatments, services, and charges. This data file contains basic record level detail for the discharge. It is also worth noting that this is the first year of the pandemic so we wanted to use this year's inpatient data to observer any strain on the hospitals in New York state.

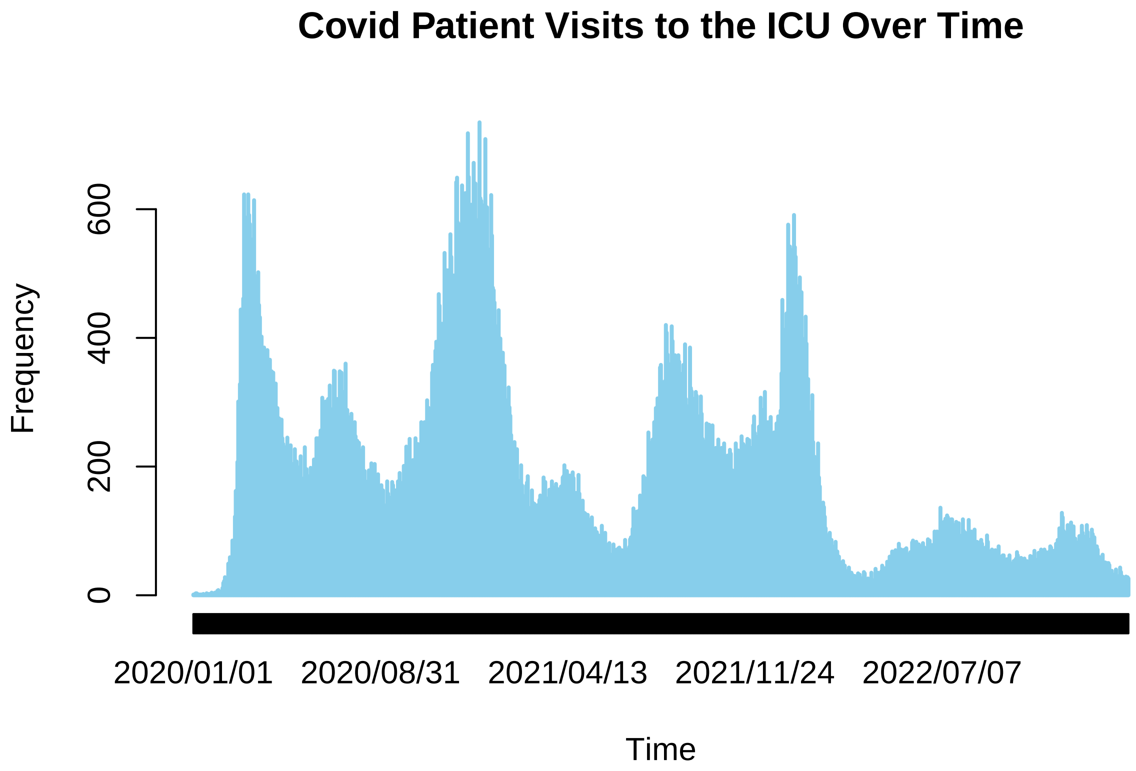
source: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/nxi5-zj9x

3) Our third secondary data set is the "Hospital Inpatient Discharges (SPARCS De-Identified): 2021" published by the New York State Department of Health. This 2M+ row data set contains patient information for the 2021 year including discharge level detail on patient characteristics, diagnoses, treatments, services, and charges. This data file contains basic record level detail for the discharge. This year's inpatient data has the context of imposed lockdowns and the beginning of vaccinations to account for in observations.

source: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Iden tified/tg3i-cinn

Nationwide related COVID-19 Data and Observations

- COVID-19 Patient Visits to the ICU Over Time:



Question of Interest:

What can we infer and notice about the change in COVID-19 patient visits in the ICU over time?

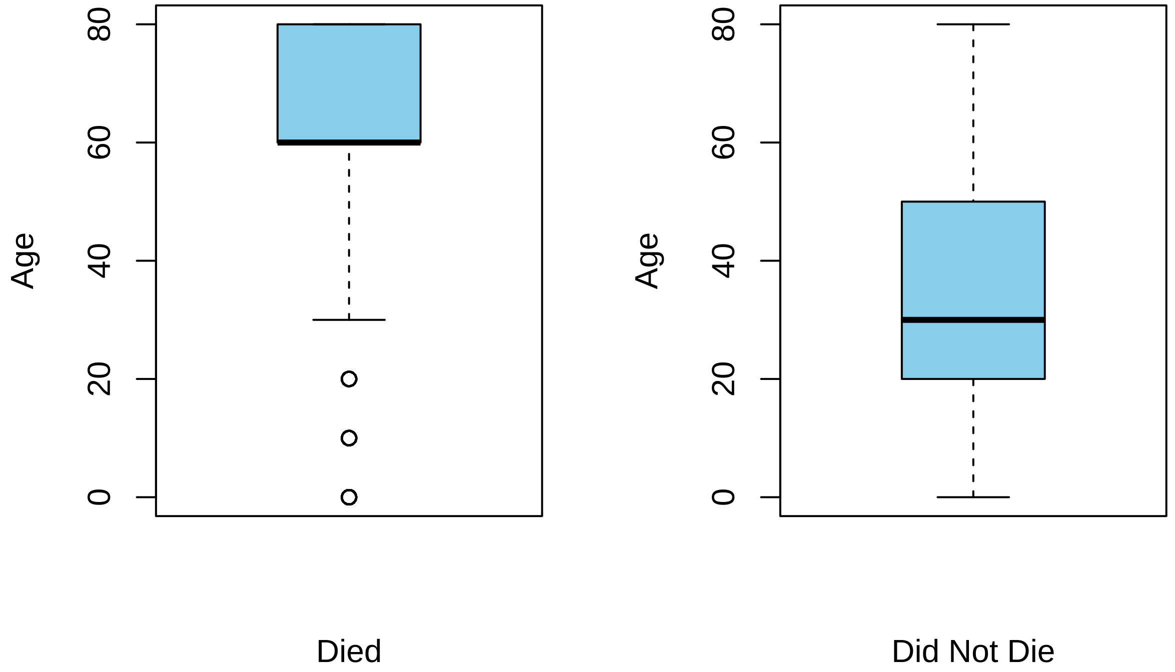
Introduction:

The plot is a running Time vs Frequency chart of COVID-19 Patients in the ICU over time. The x-axis is the timing of ICU occurrences from Jaunary of 2020 to December of 2021. The y-axis is the daily number of ICU patients who have COVID-19 over time. The chart shows the increases and decreases of the ICU frequency filled into the chart of the given time-span.

Interpretation:

Based off of the plot of "Covid Patients Visits to the ICU overtime," the appearance of COVID-19 first began with the initial spike between January of 2020 and June of 2020 with the frequency of patient reaching a consistent peak of roughly 550 patients entering the ICU every day. With the decrease of the ICU patients after the initial peak, one can assume the timeline coincides with the enforcement of lockdowns, which kept people inside and prevented a significant number of ICU visits from June 2020 to November 2020. It is also important to note that during this time period is when booster vaccinations were rolling out as well, likely driving the decline of COVID-19 inpatients. Another spike occurred in early 2021 with a relaxation of lockdown procedures, leading to an increase in COVID-19 cases. When this plot is combined with the overall daily number of COVID-19 cases, a similar alignment in plotting exists. It is inferred that this is in regards to the changes between enforced lockdowns and relaxed lockdowns.

- Mortality Distribution over Age Groups:



Question of Interest:

Is there a difference in mortality risk for COVID-19 between the age groups?

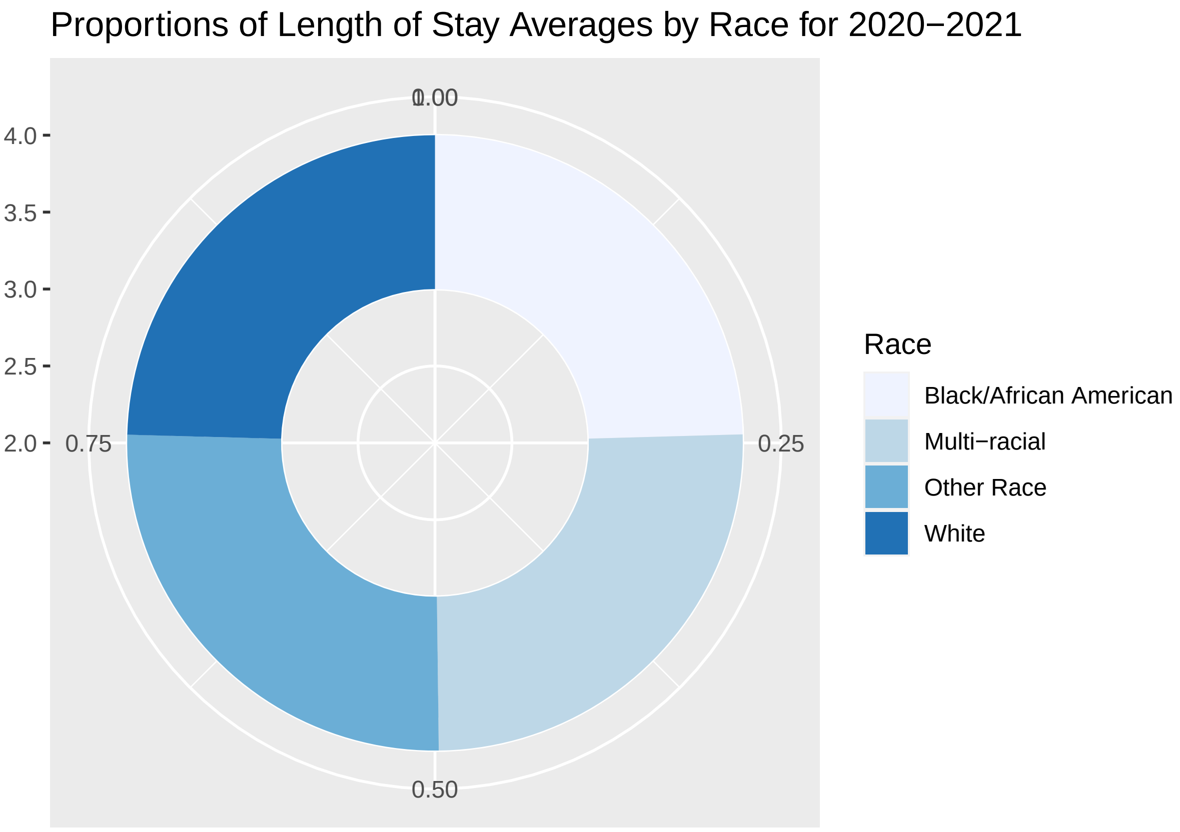
Introduction:

A box and whisker plot that represents the age group distribution of those diagnosed with COVID-19 from 2020 to 2021 and died is compared to another box and whisper plot representing the age distribution of those diagnosed with COVID-19 from 2020 to 2021 but did not die. The age distributions are separated by decades in age (10-19 years, 20-29 years, . . . , and 80+ years). The box plot for those who died from COVID-19 has higher values in terms of its 1st quartile, median, and 3rd quartile values when compared to the box plot for those who did not die.

Interpretation:

The data for those who died is mainly concentrated towards the older ages (50% of the data lies from 60-80 years old). Those who did not die were younger compared to those who did die. Additionally, the data for those who did not die is more spread out. However, because the dataset does not specify the specific age of each person diagnosed with COVID-19 and separates them by age distribution, the exact age for each quartile can not be determined.

- Proportions of Length of Stay Averages by Race by Year:



Question of Interest:

Is there a difference in length of stay for COVID-19 illness across races?

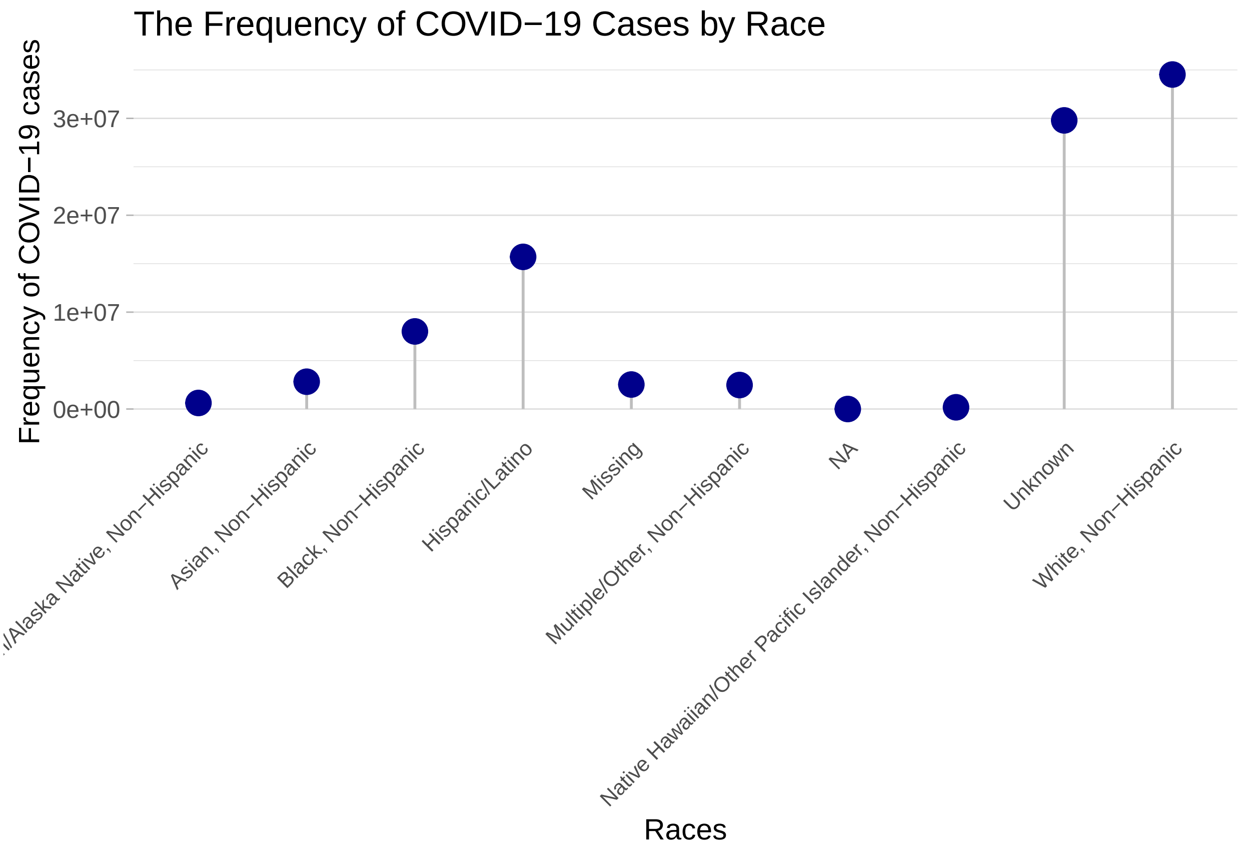
Introduction:

The donut plot above consists of the proportions of length of stay averages by race by year for 2020 and 2021. A legend for each race is placed on the right side and the donut chart and legend is color coded for readability. Additionally there are labels around the donut for reading and estimating the amount of proportion each race category takes up for each section of the donut.

Interpretation:

Given the plot that produced and the averages by race generated by the dplyr library, it can be concluded that there is not a significant difference in length of stay for COVID-19 illness across races. Visually speaking, the plot has an even distribution of proportions between the races: Black/African American, Multi-racial, Other Race, and White (the fact that the Other Race category is about the same proportion as the other races is reasonable evidence to conclude that other races have approximately the same length of stays on average as well). By investigating the data, it is shown that the maximum average length of stay by race was the Other Race category with 8.893880 days. The minimum average length of stay was Black/African American with 8.577218. Therefore, given the very small difference between the maximum and minimum average days of length of stay, it is determined that there was no significant difference of Length of Stay by race for COVID-19 patients.

- Frequency of COVID-19 Patients Over Races:



race_ethnicity_combined	freq
White, Non-Hispanic	34524397
Unknown	29790224
Hispanic/Latino	15703401
Black, Non-Hispanic	8007592
Asian, Non-Hispanic	2817740
Missing	2528751
Multiple/Other, Non-Hispanic	2475372
American Indian/Alaska Native, Non-Hispanic	623991
Native Hawaiian/Other Pacific Islander, Non-Hispanic	178012
NA	7

Question of Interest:

How accurate is our plot of the frequency of COVID-19 cases by race compared to the overall population in the US?

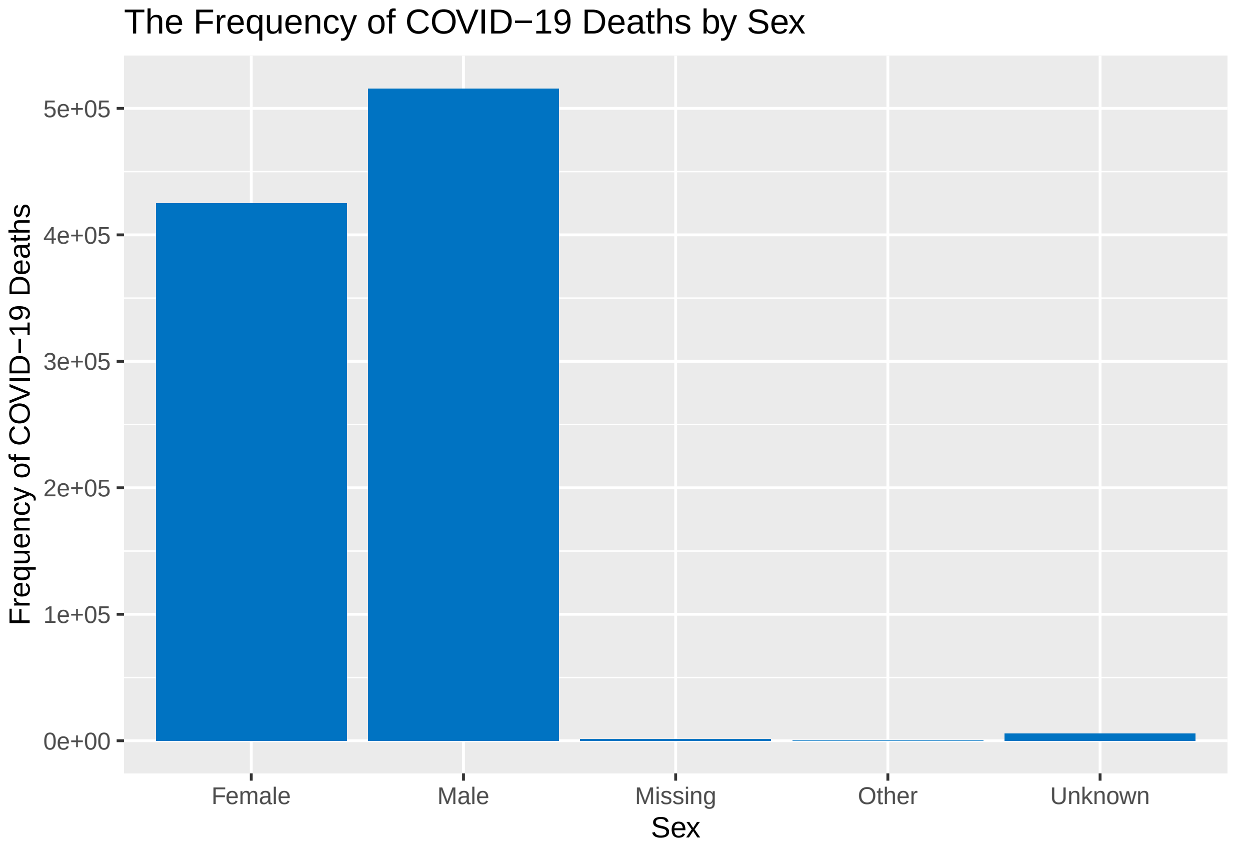
Introduction:

The lollipop chart represents the number of COVID-19 cases (y axis) that occurred in each racial group. In total, there are nine distinct racial groups (the x axis) reflected in the graph, each lollipop segment representing a different race. Three of the segments exhibit individuals that are not of one distinct racial group, these groups are: missing, unknown, and other. The 'missing' category reflects individuals who left out their race from the data, the 'unknown' category is participants who do not know their racial group, while the 'other' category belongs to individuals who could be mixed, or do not conform to any of the distinct racial groups. The overall purpose of this graph is to showcase the relationship between race and COVID-19 cases.

Interpretation:

As shown in the chart, the maximum number of COVID-19 cases occurred among the White racial group, while Hispanic people had the second highest number of cases. Individuals of Native Hawaiian/Other Pacific Islander descent exhibited the lowest number of COVID-19 cases (excluding 'missing', 'other', and 'unknown'). From these statistics, one can conclude that white people in America overall had higher rates of COVID-19. This data makes sense because the percentage of white people, which is around 60.1%, makes up most of the US. We have that 16,248,211 of white people got COVID-19 out of the 231.9 million. Thus, the weight that the percentage holds corresponds to why we see in the data above they have the highest cases of COVID-19. Similarly with Hispanic people in which they make up almost 18.5% of the US. Out of the 62.57 million Hispanic people in the US,6,715,048 of them got COVID-19. Lastly, Native Hawaiian/Other Pacific Islander make up 0.2%, meaning that out of the 1.4 million, 78,123 of them got COVID-19. Thus, it is very important when looking at the frequency of COVID-19 cases of the ethnicity/race of a certain population to consider what percentage each ethnicity/race makes up of the whole population at hand.

- Frequency of COVID-19 Deaths by Sex:



Sex	Count
Female	425226
Male	515816
Missing	1360
Other	7
Unknown	5752

Question of Interest:

Is there statistical evidence for sex differences in death caused by COVID-19?

Introduction:

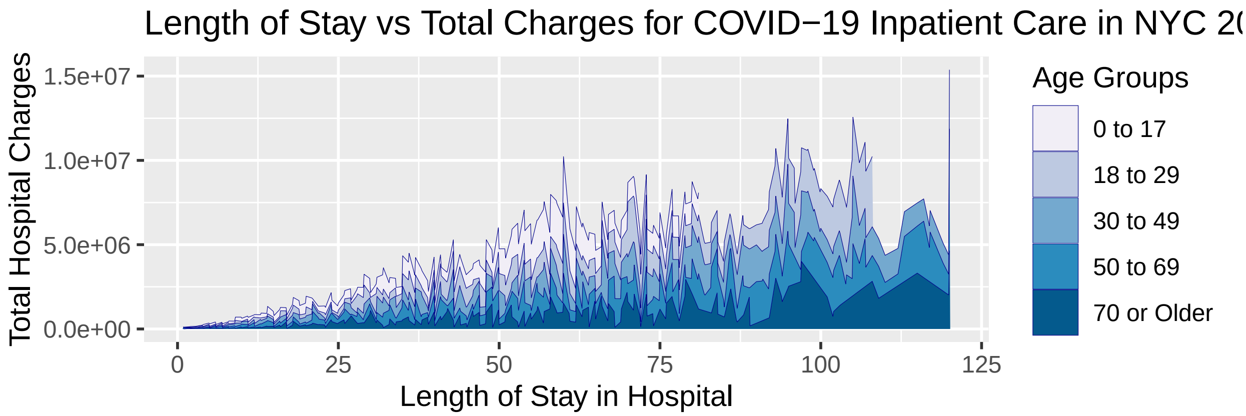
The graph above plots the frequency of COVID-19 deaths by sex from 2020 - 2023. The x-axis represents each patient's sex, which was identified as being either "Female," "Male," "Missing", "Other", or "Unkown." The y-axis represents the amount that each of the labels on the x-axis died having COVID-19.

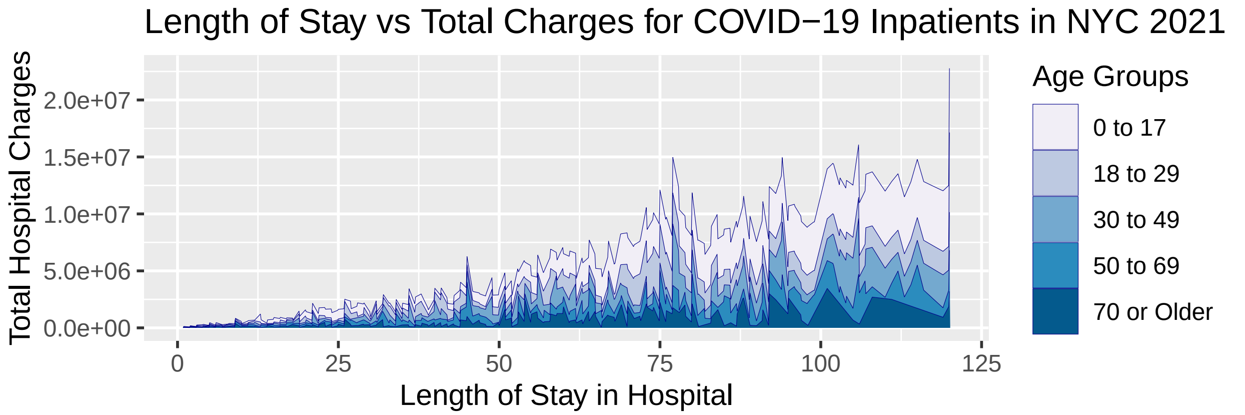
Interpretation:

According to the graph, males ranked the highest when it came to the reported COVID-19 deaths, followed by females. The other groups of 'Missing,' 'Other,' and 'Unknown' had a very minimal frequency when compared to the females' and males' frequency. However, the difference between females and males is minimal, meaning that is barely a difference between the sexes and the COVID-19 deaths throughout the years of 2020 to 2023. The slight difference between them might be due to the fact that more males reported COVID-19 deaths than females. Additionally, males having a higher frequency of COVID-19 may reflect the fact that men tend to work outside of the house at a higher rate than women, thus exposing them to the disease more. Another factor is the fact that men tend to have a weaker immune system than women.

New York State wide COVID-19 Data and Observations

- Length of Stay over Total Charges for COVID-19 Patients for 2020-2021:





Age.Group	correlation
0 to 17	0.8652675
18 to 29	0.8626409
30 to 49	0.8193177
50 to 69	0.8165139
70 or Older	0.7485752

Question of Interest:

What is the correlation between length of stay in the hospital and total hospital charges within each age group?

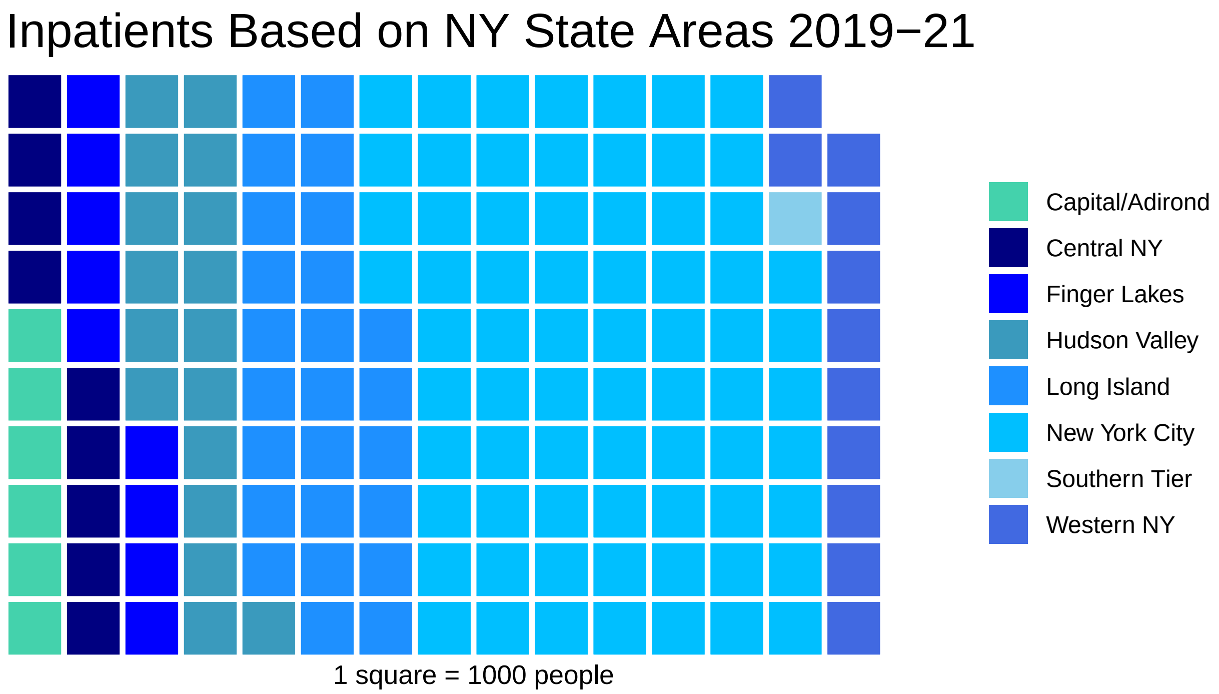
Introduction:

The plots above graph the relationship between total hospital charges for COVID-19 inpatient care in a NYC hospital and the number of days the patient stayed in the hospital. Data was collected and graphed in the plots from 2020 and 2021, respectively. The x-axis represents the length of stay each patient stayed in the hospital for, and the y-axis represents the total hospital charges for that patient. Through a stacked area chart, each age group (separated by 0-17 years old, 18-29, 30-49, 50-69, and 70 or older), is categorized and stacked on top of each other to analyze how the age group's length of stay changes over the progression of total hospital charges.

Interpretation:

The 2020 plot is fairly similar to the 2021 plot, which shows a somewhat positive correlation between the length of stay in the hospital and the total hospital charges, with a longer length of stay correlating to higher hospital charges. Additionally, the plots both show that the younger the age group, the higher the charges are for the same length of stay when compared to the older age groups. However, the highest charges occur for 18 to 29 year olds at an extended length of stay. For 2020, the highest charge was for 18 to 29 year olds at the hospital for approximately 105 days. For 2021, the highest charge was for 18 to 29 year olds at the hospital for approximately 80 days. As each age group gets older, each group reaches their highest charge at a higher length of stay. This is likely due to younger people having less access to health insurance, from either being unemployed or having less life savings. Additionally, it is less likely for younger people to be staying in the hospital for an extended period of time due to less severity in their illness when compared to the older age groups. In order to further analyze the plot, the correlation between the length of stay and total charges was analyzed for each age group. Across all age groups, the correlation proved to be very strong, as all of the age groups had a correlation in the ~.80 range, while the 70 or older age group had a correlation of .75. This exhibits strong evidence that there is a direct relationship between length of stay and total charges: as the length of stay increases, the total charge for each patient also increases across all age groups.

- Inpatients Based on NY State Areas (2019-2021)



Hospital Service Area	Count
Capital/Adirond	6640
Central NY	9369
Finger Lakes	9359
Hudson Valley	17056
Long Island	25671
New York City	71377
Southern Tier	1804
Western NY	11177

Question of Interest:

Is COVID-19 more likely to transmit in some areas than others?

Introduction:

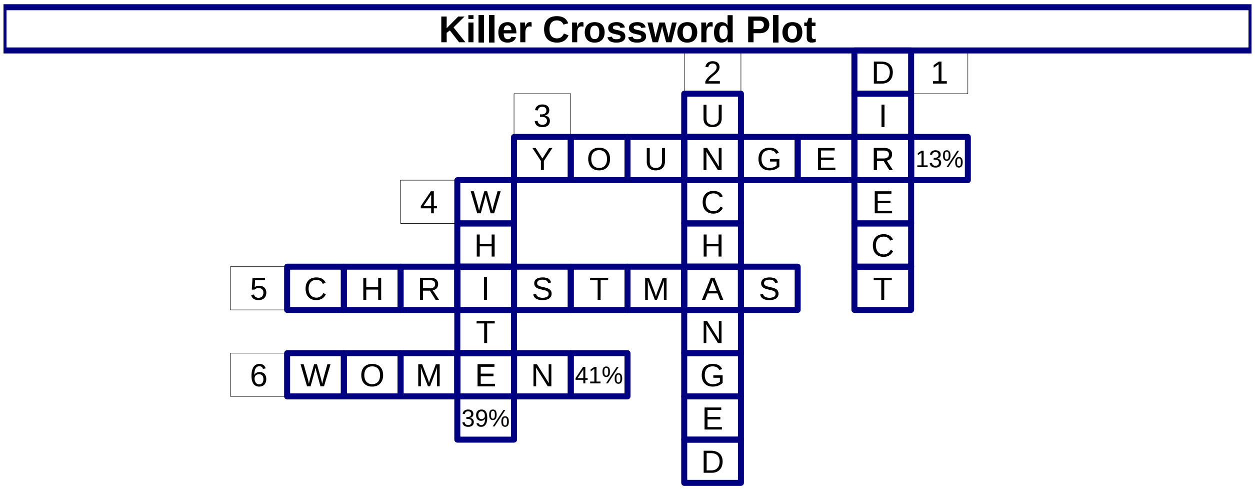
The plots above is a waffle plot detailing the number of inpatients and what NY state area they belonged to for the years 2019, 2020, and 2021. Thus this waffle plot is a combination of the waffle plots for 2019, 2020, and 2021. The waffle plot is color-coded for convenience and each square represents 1000 people. The square without a title in the legend is for inpatients who did not identify with an area and did not self-report.

Interpretation:

In the plot above it's evident that the New York City area had the highest number of inpatients. This is in constrast to less populated areas which demonstrated less inpatient admissions. This demonstrates that there is a positive correlation between densely populated areas with the number of inpatients admitted. In addition when all waffle plots of each year are plotted individually, it can be seen that the proportions of each New York area and their number of inpatients do not change, demonstrating a lack of outward or inward migration likely due to the pandemic and lockdowns.

General COVID-19 Observations

- Killer Plot



Across

3. Age Group with Highest Hospital Charges.

5. What holiday did patients visit the ICU the most

6. Which Sex had the highest frequency of covid?

Down

1. Relationship between length of stay and charges incurred

2. The trend in migration in NY remained −−−−−−−−?

4. Which race had the highest number of covid cases?

Question of Interest:

What are the key data trends from our previous plots?

Introduction and Interpretation:

The killer plot we decided to make was a crossword puzzle representing the outcomes we found from the overall data visualizations. As you can see, at the bottom puzzle are the hints of the crossword indicating the questions for each of the outcomes. For example, in the ‘across’ section of the killer plot, the third statement is “Age Group with Highest Hospital Charges” which matches to number three on the plot with the answer being, “Younger Age group at 13%”

Conclusion

After analyzing the data, there are various findings that we arrived at. The general demographic breakdown of our results revealed that White individuals had the highest frequency of cases but a lower rate than some of the other racial groups, and men exhibited the highest rate of cases compared to other sexes. Additionally, after looking at the hospital trends overall, we noticed that the trends in ICU patients over time are indicative of the chronological order of the pandemic; for example, during lockdown rulings there seemed to be fewer cases, while during peaks of the pandemic, ICU patient numbers rose (especially in more populated areas). An interesting finding to note is that younger age groups incurred higher charges for the same length of hospital stay, which we attributed to their lack of health insurance. Across all races, there seemed to be an even length of hospital stay, which substantiates proof that race does not necessarily impact the severity of COVID-19 symptoms.

Future Works

After analyzing the data, there are various findings that we arrived at. The general demographic breakdown of our results revealed that White individuals had the highest frequency of cases but a lower rate than some of the other racial groups, and men exhibited the highest rate of cases compared to other sexes. Additionally, after looking at the hospital trends overall, we noticed that the trends in ICU patients over time are indicative of the chronological order of the pandemic; for example, during lockdown rulings there seemed to be fewer cases, while during peaks of the pandemic, ICU patient numbers rose (especially in more populated areas). An interesting finding to note is that younger age groups incurred higher charges for the same length of hospital stay, which we attributed to their lack of health insurance. Across all races, there seemed to be an even length of hospital stay, which substantiates proof that race does not necessarily impact the severity of COVID-19 symptoms.