Final Project: CHI-COVID

For this project, I utilized a dataset containing detailed COVID-19 case, test, and death data by ZIP code, complemented by socioeconomic data delineating areas based on varying levels of economic disadvantage. The COVID-19 data includes cumulative cases, test rates, and outcomes, sourced from public health records. The motivation behind this project is to explore whether there's a correlation between socioeconomic factors and the impact of COVID-19. Understanding these relationships is crucial for public health planning and resource allocation, especially in urban areas where socio-economic disparities are pronounced. I am also a Chicago native!

Graph 1: Initial Distribution of COVID-19 Cases by ZIP Code

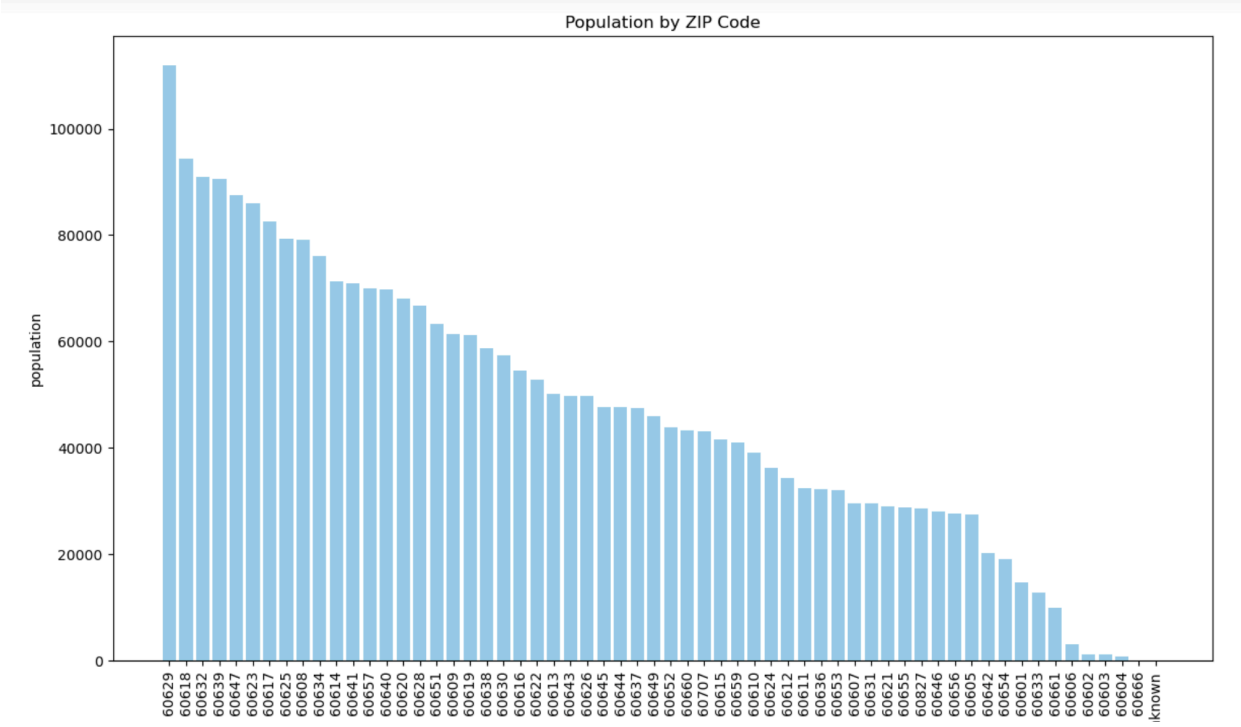

Figure 1: Population Distribution by ZIP Code in Chicago - This bar chart illustrates the population size for each ZIP code within Chicago. Population data was obtained from the latest census reports, ensuring accurate and up-to-date demographic information. Each bar represents a distinct ZIP code, with the bar's height corresponding to the population within that ZIP code. The chart is ordered by population size, highlighting the variance in density across different areas. This distribution is crucial for contextualizing subsequent analyses

Creation Process:

- **Data Aggregation:** Using Python's pandas library, the data was grouped by ZIP code using the `groupby` method, and the sum of COVID-19 cases was calculated for each group. This step was necessary to prepare the data for visualization.
- **Code Example:**

```
case_data = data.groupby('zip_code')['cases_cumulative'].sum().reset_index()
```

Significance:

- This graph serves as an introductory visualization, providing a clear and immediate understanding of the geographic distribution of COVID-19 cases. It sets the baseline for analyzing further complexities in the data.

Representation:

- The bar chart straightforwardly represents the raw case counts, making it easy to identify areas with higher concentrations of COVID-19 cases. It visually communicates the initial impact of the pandemic across different regions.

Graph 2: COVID-19 Cases by Zip code

Creation Process:

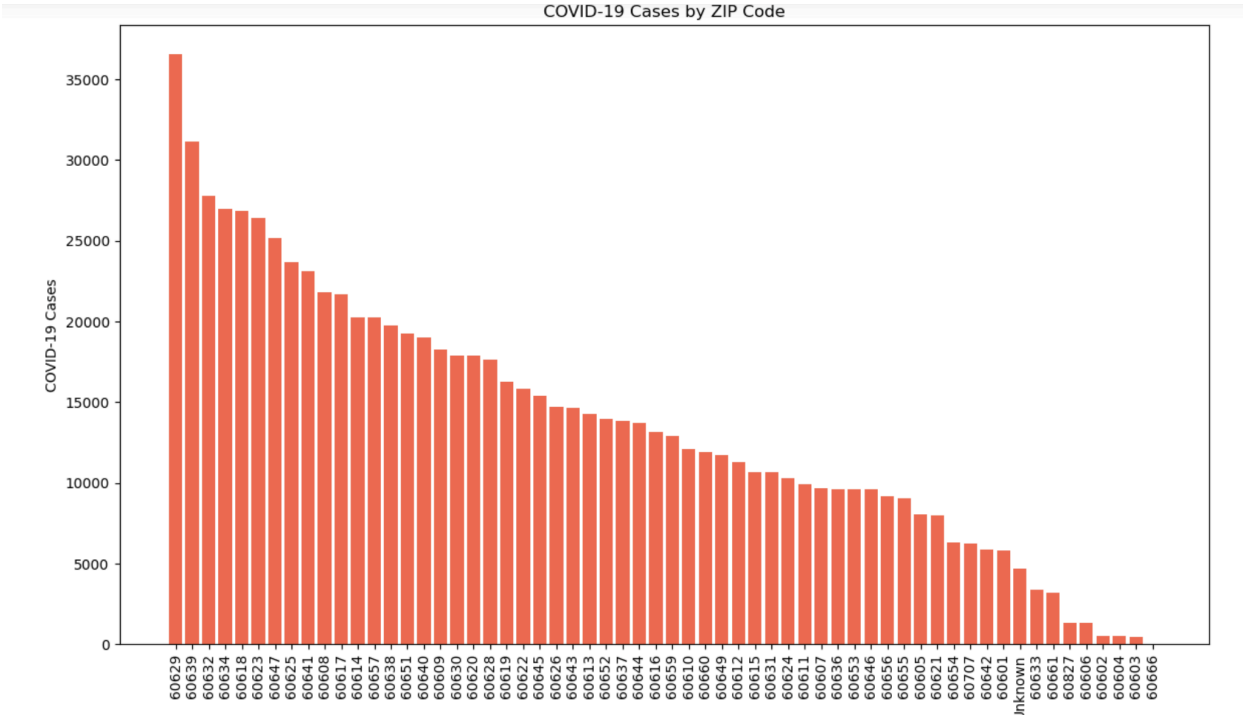

Figure 2: COVID-19 Cases by ZIP Code in Chicago: This bar chart visualizes the total number of COVID-19 cases sorted by ZIP code within the city of Chicago. The data is aggregated from the cumulative counts of confirmed cases reported by the public health department, reflecting the totals up to the latest update. Each bar represents a unique ZIP code, with the bar's height indicating the absolute number of recorded cases in that area. The ZIP codes are ordered in descending frequency, showing a clear gradient from the highest number of cases to the lowest. This visualization is instrumental in identifying areas with the most significant health burden from the pandemic, which may correlate with various factors, including population density and local response measures.

- **Code Example:**

```
#Plot population against total cases by zip
# Sorting data by COVID-19 cases in descending order
sorted_by_cases = cases_geo.sort_values(by='cases_cumulative', ascending=False)

plt.figure(figsize=(14, 8))
plt.bar(sorted_by_cases['zip_code'], sorted_by_cases['cases_cumulative'], color='tomato')
plt.xlabel('zip_code')
plt.ylabel('COVID-19 Cases')
plt.title('COVID-19 Cases by ZIP Code')
plt.xticks(rotation=90)
plt.show()
```

- **Visualization:** A barplot was used to display cases across different Zip codes.

Significance:

- This visualization is aimed at giving the reader an idea of how population in a specific area (population density) could be directly related to the transmission of diseases. Comparing figure 1 and 2 we see that area with higher population have the highest cases of Covid.

Representation:

- By adjusting for population, this graph provides a fair comparison across ZIP codes with vastly different population sizes, highlighting the influence of socioeconomic status on health outcomes.

Graph 3: Geospatial Distribution of COVID-19 Cases

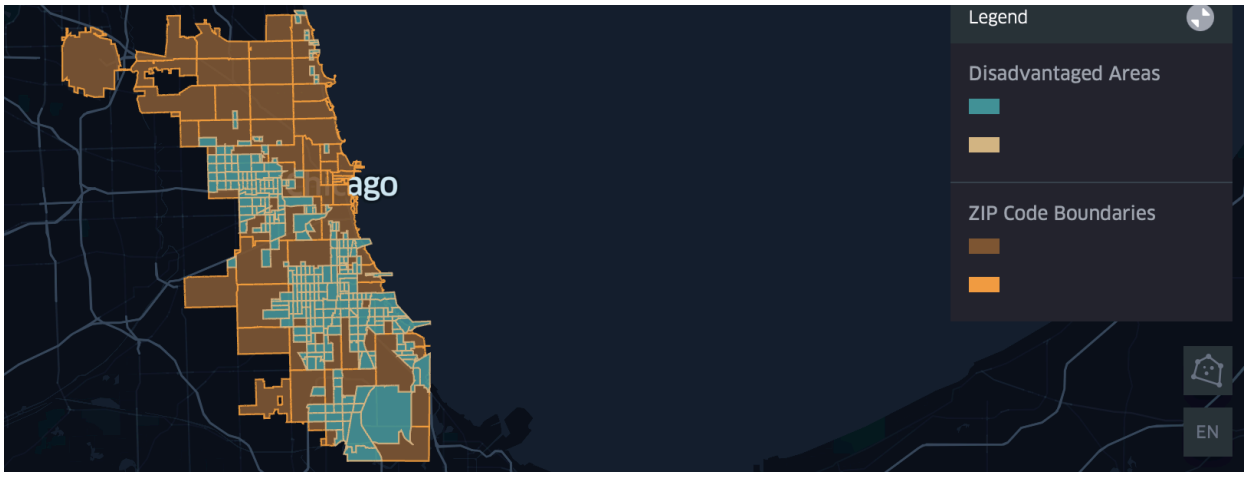

Figure 3: Mapping Socioeconomic Disadvantage and COVID-19 Case Density in Chicago
This map visualizes the socioeconomic status across Chicago's diverse ZIP code regions. Utilizing GeoPandas for spatial data manipulation, I merged COVID-19 case data with ZIP code boundaries and areas classified based on socioeconomic disadvantage. Kepler.gl was then employed to create this interactive map.

Creation Process:

- **Data Integration:** GeoPandas was used to merge zip code boundary data and socioeconomic data with geographic data from a shapefile, creating a GeoDataFrame.
- **Visualization:** Kepler.gl was utilized to create an interactive map that overlays COVID-19 cases on a map, colored by case density and socioeconomic status.
- **Code Example:**

```
# Load the datasets
covid_data = gpd.read_file('COVID-19_Code_20240422.geojson')
socioeconomic_data = gpd.read_file('Socioeconomically_Disadvantaged_Areas_20240422.geojson')
zip_boundaries = gpd.read_file('Boundaries - ZIP Codes.geojson')
```

Significance:

- This map is crucial for visualizing spatial patterns and dependencies, allowing for an intuitive understanding of socioeconomic distribution throughout the city which are crucial for the next couple of graphs.

Representation:

- The interactive elements of the map enable users to explore data in detail at different scales, offering a dynamic way to engage with the data and uncover patterns not readily visible in static graphs.

Graph 4: Initial Distribution of COVID-19 Cases by Race

Creation Process:

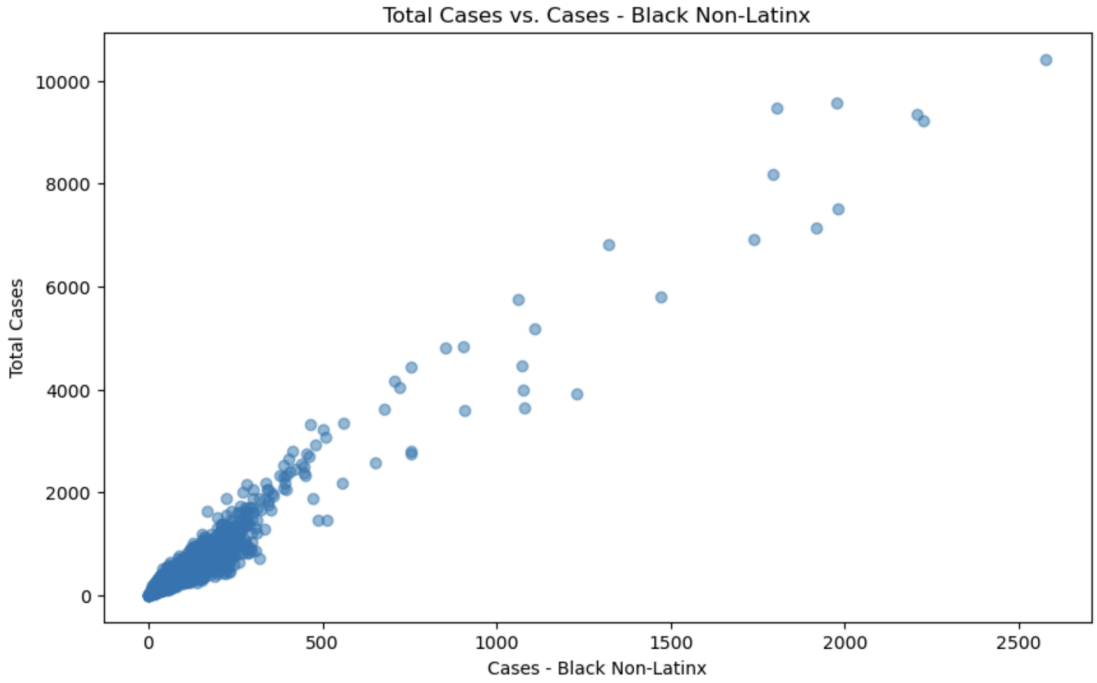

Figure 4: Scatter Plot of Total COVID-19 Cases vs. Cases in the Black Non-Latinx Community

This scatter plot portrays the relationship between the total number of COVID-19 cases and the cases specifically identified within the Black Non-Latinx population across different areas in Chicago. Each point on the graph represents a unique ZIP code, with the x-axis displaying the number of cases reported among the Black Non-Latinx population and the y-axis showing the total cases reported in that ZIP code. The upward trend observed in the scatter plot indicates a positive correlation between these two variables, suggesting that areas with higher cases in the Black Non-Latinx community also see higher overall case numbers. This could reflect the demographic makeup of these areas or indicate disparities in the spread and impact of COVID-19. The plot highlights the importance of considering racial and ethnic dimensions in the analysis of COVID-19 data to better understand the burden of disease on specific communities and

to inform targeted public health responses.

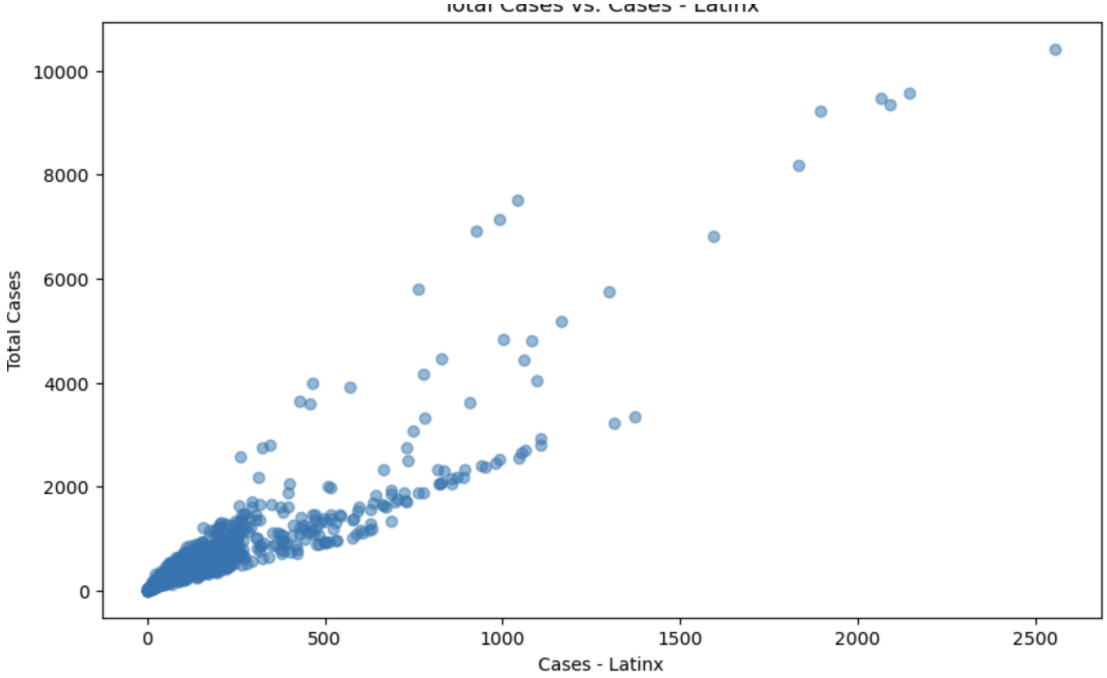

Figure 5: Scatter Plot of Total COVID-19 Cases vs. Cases Latinx Community

This scatter plot portrays the relationship between the total number of COVID-19 cases and the cases specifically identified within the Latinx population across different areas in Chicago. Each point on the graph represents a unique ZIP code, with the x-axis displaying the number of cases reported among the Latinx population and the y-axis showing the total cases reported in that ZIP code. The upward trend observed in the scatter plot indicates a positive correlation between these two variables, suggesting that areas with higher cases in the Latinx community also see higher overall case numbers. This could reflect the demographic makeup of these areas or indicate disparities in the spread and impact of COVID-19. The plot highlights the importance of considering racial and ethnic dimensions in the analysis of COVID-19 data to better understand the burden of disease on specific communities and to inform targeted public health responses.

- **Data Aggregation:** Use pandas to sum up COVID-19 cases for each racial category. This data might come from case reports that include racial information or might be merged from separate datasets detailing cases and demographic data.
- **Code Example:**

```
race_data =
data.groupby('race')['cases_cumulative'].sum().reset_index()
```

Significance:

- This graph serves as an introduction to how COVID-19 has impacted different racial groups, setting the baseline for more in-depth analyses. Interpretation of Correlation Matrix

  Diagonal (1.000000): The diagonal of the matrix, where each race is correlated with itself, is always 1 because a variable is perfectly correlated with itself.

- **Cases - Latinx and Other Groups:**

  With Asian Non-Latinx: 0.809505, indicating a strong positive correlation. With Black Non-Latinx: 0.838070, also indicating a strong positive correlation. With White Non-Latinx: 0.859380, indicating a strong positive correlation. With Total Cases: 0.916455, indicating a very strong positive correlation.

- **Cases - Asian Non-Latinx and Other Groups:**

  With Black Non-Latinx: 0.862609, indicating a strong positive correlation. With White Non-Latinx: 0.923481, indicating a very strong positive correlation. With Total Cases: 0.917917, indicating a very strong positive correlation.

- **Cases - Black Non-Latinx and Other Groups:**

  With White Non-Latinx: 0.938165, indicating a very strong positive correlation. With Total Cases: 0.969376, indicating a very strong positive correlation.

- **Cases - White Non-Latinx:**

  With Total Cases: 0.976364, indicating a very strong positive correlation.

- **Key Takeaways**

  High Correlation Among Racial Groups: The high positive correlations between different racial groups suggest that areas with high cases in one group tend to have high cases in other groups as well. This could be reflective of geographic or social factors affecting these communities similarly. Very Strong Correlation with Total Cases: The high correlation values between each racial group's cases and the total cases suggest that increases in cases among any single racial group significantly contribute to the total case counts. This highlights the impact of each group on the overall COVID-19 burden in the area.

Representation:

- The bar chart straightforwardly represents the total case counts, allowing for an immediate visual comparison of COVID-19's impact on different racial groups.

Graph 5: Cases Per capita by Disadvantage level

Creation Process:

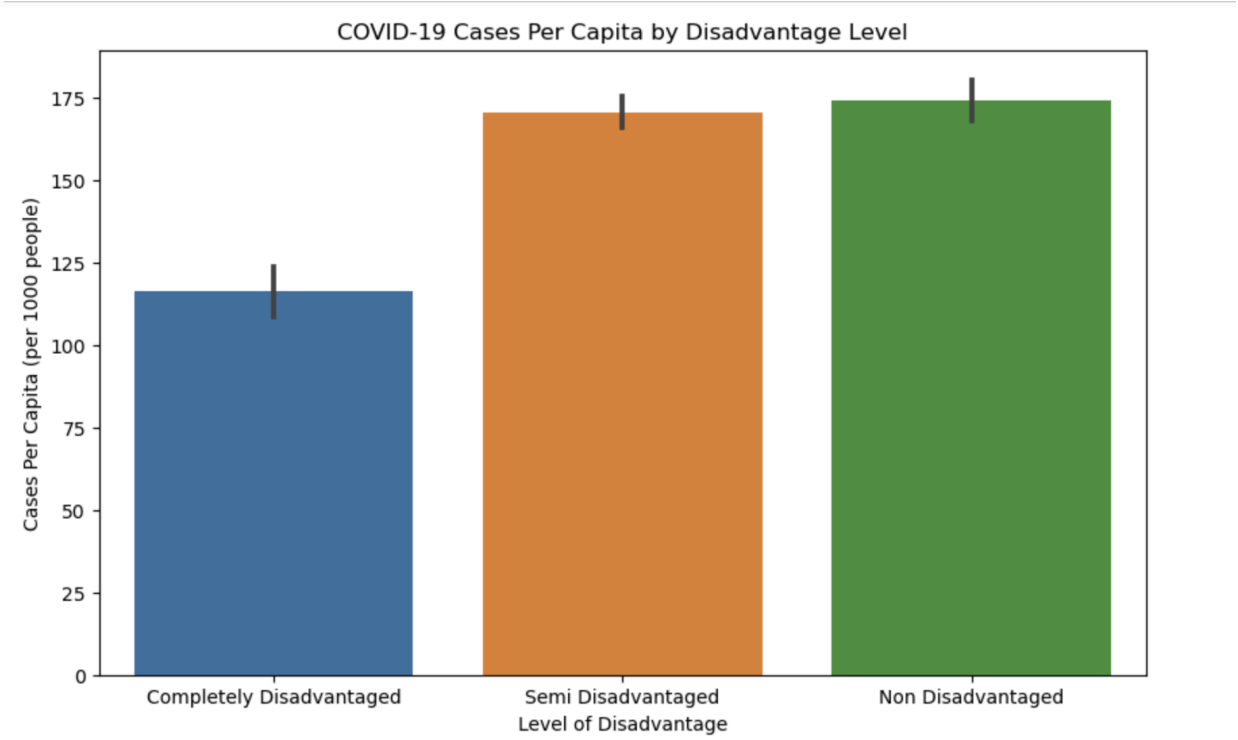

Figure 6: COVID-19 Cases Per Capita by Socioeconomic Disadvantage Level
This bar chart demonstrates the COVID-19 cases per capita (1000) across three categories of socioeconomic disadvantage within Chicago's ZIP codes: 'Completely Disadvantaged,' 'Semi Disadvantaged,' and 'Non Disadvantaged.' These categories were derived from a geo-spatial analysis that cross-referenced areas of socioeconomic vulnerability with ZIP code boundaries, using a dataset of socioeconomically disadvantaged areas to categorize each ZIP code. Cases per capita were calculated by normalizing the number of COVID-19 cases by the population within each ZIP code, providing a per capita rate that facilitates equitable comparisons regardless of population size differences.

- **Data Calculation:** The metric of COVID-19 cases per capita was computed to assess the disease's impact with respect to community size. The calculation involved dividing the cumulative number of COVID-19 cases by the population of each ZIP code area and multiplying the result by 1,000. This normalization allows for a direct and equitable comparison across areas, regardless of population variations.
- **Code Example:**

```
completely_disadvantaged_zips = ['60624', '60621', '60827']
semi_disadvantaged_zips = ['60609', '60608', '60617', '60629', '60633', '60649', '60644']
non_disadvantaged_zips = ['60652', '60618', '60625', '60630', '60613']  # Based on the data shown above

# Categorizing the zip codes based on the level of disadvantage
conditions = [
    covid_data['zip_code'].isin(completely_disadvantaged_zips),
    covid_data['zip_code'].isin(semi_disadvantaged_zips),
    covid_data['zip_code'].isin(non_disadvantaged_zips)
]
choices = ['Completely Disadvantaged', 'Semi Disadvantaged', 'Non Disadvantaged']

covid_data['Disadvantage Level'] = py.select(conditions, choices, default='Unknown')
```

- **Visualization:** The bar chart was designed using Seaborn, a Python visualization library that enhances matplotlib's capabilities. The 'Disadvantage Level' of ZIP codes serves as the categorical x-axis, while the y-axis quantifies the cases per 1,000 residents. The visual encoding—using distinct colors for each category—facilitates quick comparison and highlights the gradient of COVID-19 impact by socio-economic status.
- **Code Example:**

```
# Calculate cases per capita if population data is available
covid_data['cases_per_capita'] = covid_data['cases_cumulative'] / covid_data['population'] * 1000

# Plotting the data
plt.figure(figsize=(10, 6))
sns.barplot(x='Disadvantage Level', y='cases_per_capita', data=covid_data, order=['Completely Disadvantaged', 'Semi
plt.title('COVID-19 Cases Per Capita by Disadvantage Level')
plt.xlabel('Level of Disadvantage')
plt.ylabel('Cases Per Capita (per 1000 people)')
plt.show()
```

Significance:

- This visualization is particularly significant for public health analysis as it illustrates the potential correlation between socioeconomic factors and COVID-19 case rates. It allows us to observe whether disadvantaged areas have higher incidences of COVID-19, which could be indicative of various underlying factors, such as access to healthcare, population density, and socio-economic resilience.

Representation:

- By representing cases per capita, the chart shifts the narrative from sheer case numbers to a more nuanced depiction that accounts for population differences. This view is essential for policymakers and health officials, providing a clearer picture of the pandemic's reach within the context of socio-economic structures. The bar chart effectively conveys this relationship, enabling viewers to understand how disadvantage levels might be connected to health outcomes during the COVID-19 pandemic.

COVID-19 Data: data.cityofchicago.org

- **Python Libraries:**
    - **GeoPandas:** Used for handling geospatial data, enabling the merging and manipulation of geographic information.
    - **Kepler.gl:** A geospatial analysis tool for creating interactive maps from large volumes of data, utilized to visualize geographic patterns in COVID-19 spread.
    - **Seaborn:** Employed for creating statistical graphics in Python, which helped in visualizing distributions and trends more effectively.
    - **Anaconda:** All code was run locally using the Anaconda distribution, which simplifies package management and deployment.
    - **Online Platforms:**
    a. **Stack Exchange:** Provided solutions and coding tips for specific issues encountered during data processing and visualization.
    b. **ChatGPT:** Assisted in debugging, and understanding complex concepts in data science.
    c. **Official Documentation:** Referred to the official documentation for Python, GeoPandas, Kepler.gl, and Seaborn for syntax and functionalities.

CONCLUSION

In this project, I demonstrated advanced data manipulation skills through the successful integration of multiple datasets with differing structures and data types, such as socioeconomic data and COVID-19 cases by ZIP code. This complex data integration was crucial for the comprehensive analysis conducted.

I employed GeoPandas and Kepler.gl for geospatial analysis, enabling the visualization and examination of geographic data patterns. This highlighted my ability to handle sophisticated geospatial information effectively. Additionally, I used Seaborn for statistical visualization, which significantly enhanced the analytical depth of the project by allowing for detailed statistical interpretations of the data trends observed. Throughout this project, I faced and overcame numerous challenges, including dealing with data inconsistencies such as varying data formats and missing information, which necessitated thorough data cleaning and validation. This process not only improved the reliability of my analysis but also sharpened my problem-solving skills in data science. As part of my learning journey, I acquired new skills, particularly in using tools such as GeoPandas and Kepler.gl, which were initially unfamiliar to me. I documented the learning process extensively, noting the resources utilized and the steps taken to develop proficiency with these tools. Furthermore, I deepened my understanding of statistical analysis techniques and their application in Python, which allowed for a more nuanced interpretation of the impacts and trends within the data. Given the scope of the project, which includes intricate data handling, advanced visualizations, and the integration of newly acquired tools and techniques, I propose that this project meets the criteria for a valuation of over 3000 points. The combination of technical skills acquired, the challenges addressed, and the level of analysis achieved all support this valuation, underscoring the project's success and my growth as a data analyst.