Rutgers, The State University of New Jersey

Master of Information Technology and Analytics

# Socioeconomic and Demographic Drivers of COVID-19 Case Rates: Insights from Regression Analysis

Kai-Yin Huang

Supervisor: Professor Michalis Xyntarakis

December 9, 2024

# Content

# 1. Introduction

## 1.1 Background

The **COVID-19 pandemic** has caused unprecedented public health and economic challenges worldwide. As the virus spreads unevenly across communities, significant **disparities in infection rates** have been observed, highlighting the role of **socioeconomic and demographic factors** in influencing disease transmission. In densely populated urban areas like New York State, these disparities are particularly pronounced, with some neighborhoods experiencing disproportionately high case rates.

Existing studies suggest that factors such as income inequality, racial demographics, and population density are key determinants of health outcomes. Lower-income communities often face limited access to healthcare, higher population densities, and increased exposure to essential service work, all of which exacerbate vulnerability to infectious diseases. Similarly, racial and ethnic minorities have been disproportionately affected due to systemic inequities in social determinants of health.

While prior research has explored the relationship between socioeconomic factors and COVID-19 outcomes, few studies have incorporated high-resolution data at the ZIP-code level or considered temporal factors such as lagged case rates. This study aims to fill this gap by investigating the interplay between socioeconomic characteristics, racial composition, and COVID-19 case rates using **ZIP-code-level** data from New York State during the critical early phases of the pandemic.

By understanding these relationships, this analysis seeks to provide actionable insights for **policymakers** to design targeted interventions and allocate resources effectively to mitigate the impact of future pandemics on vulnerable communities.

## 1.2 Problem Statement

The COVID-19 pandemic has exposed significant disparities in infection rates across different socioeconomic and demographic groups. In urban areas such as New York State, neighborhoods with lower income levels, higher proportions of racial minorities, and greater population density have been disproportionately affected. However, the mechanisms driving these disparities remain poorly understood, particularly in terms of how lagged case rates, interaction effects between income and racial composition, and other socioeconomic factors contribute to the observed trends.

Existing research often lacks the granularity to capture these dynamics at the neighborhood level, and the temporal aspects of disease transmission are frequently overlooked. This creates a critical gap in understanding the multifactorial drivers of COVID-19 case rates, which is essential for designing equitable and effective public health interventions.

### 1.3 Aims and Objectives

Aims:

To provide actionable insights for targeted policy interventions through the investigation of socioeconomic and demographic determinants of COVID-19 case rates in New York State at the ZIP-code level during the pandemic's early stages.

Objectives:

1. To analyze the relationship between socioeconomic factors (e.g. income levels, population density) and the case rates of COVID-19.

2. Assess the impact of demographic racial characteristics, including the proportions of Hispanic/Latino and other ethnic groups, on case rates.

3. Examine the role of lagged case rates in predicting current case rates by capturing temporal transmission dynamics.

4. Explore interaction effects, such as between income levels and racial demographics, and their influence on case rates.

5. Provide data-driven recommendations for public health strategies aimed at reducing disparities in infection rates.

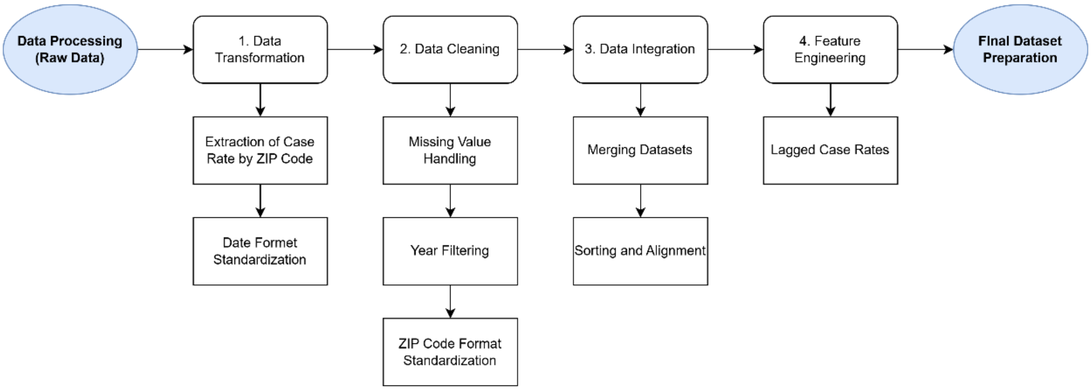
# 2. Data and Methodology

### 2.1 Data Field

| **Field Name** | **Description** |
|---|---|
| week_ending | The ending date of the week (used for temporal analysis. |
| zip_code_column | The ZIP code identifier in the dataset. |
| case_rate | COVID-19 case rate per 100,000 people. |
| zip | ZIP code in a simplified format. |
| lat, lng | Latitude and longitude of the ZIP code area. |
| city, state_id, state_name | City name, state abbreviation, and full state name. |
| population | Population of the ZIP code area. |
| density | Population density per square mile. |
| county_name | The name of the county associated with the ZIP code. |
| po_box | Indicator for whether the ZIP code is a P.O. Box. |
| dist_highway, dist2_large_airport, dist2_medium_airport, dist_to_shore | Distances to key infrastructure: highways, large/medium air-<br>ports, and shorelines. |
| number_of_business | Total number of businesses in the ZIP code area. |

| | |
|---|---|
| adjusted_gross_income | Adjusted gross income (AGI) for the ZIP code area. |
| total_income_amount | Total income reported for the ZIP code area. |
| number_of_returns | Number of tax returns filed for the ZIP code area. |
| age_all_ages | Population count for all age groups combined. |
| age_0_4 to age_75up | Population counts for different age groups (e.g., 0-4 years, 5-12 years, 13-17 years, etc.). |
| all_race_ethnicity | Total population across all racial and ethnic groups. |
| Asian_Pacific_Islander | Percentage of the population identifying as Asian or Pacific Islander. |
| Black_African_American | Percentage of the population identifying as Black or African American. |
| Hispanic_Latino | Percentage of the population identifying as Hispanic or Latino. |
| White | Percentage of the population identifying as White. |
| lagged_case_rate | Case rate from the previous week (used for temporal lag analysis). |

## 2.2 Data Processing



The dataset used in this analysis contains multiple features, including socioeconomic indicators, population demographics, and COVID-19 case rates at the ZIP-code level. To ensure the data's usability and reliability for statistical modeling, the following preprocessing steps were undertaken:

1. Data Transformation

i. Extraction of Case Rates by ZIP Code

- Columns containing COVID-19 case rates for 5-digit ZIP codes were identified using regular expressions (CASERATE_\d{5}).
- The dataset was reshaped from wide format to long format using the melt function, creating a column for zip_code and case_rate.

- ZIP codes were extracted and standardized as strings for consistency.

ii. Date Format Standardization:

- The week_ending column in all datasets (caserate_data, age_data, and race_data) was converted to a consistent datetime format to facilitate merging and filtering.

2. Data Cleaning

i. Missing Value Handling:

- Missing values were identified in age, race, and demographic data. Rows with missing values in critical fields (e.g., case_rate, demographic variables) were dropped to ensure data quality.

ii. Year Filtering:

- Data was filtered to include only records from the year 2020, focusing on the early pandemic period.

iii. ZIP Code Format Standardization:

- ZIP codes across all datasets were converted to strings to ensure consistent formatting for merging.

3. Data Integration

i. Merging Datasets:

- The case rate data was merged with demographic, age, and race datasets using zip_code and week_ending as the primary keys.
- An inner join ensured that only records with complete information across all datasets were retained.

ii. Sorting and Alignment:

- The merged dataset was sorted by zip_code and week_ending to maintain temporal and geographic consistency.

4. Feature Engineering

i. Lagged Case Rates:

- A new variable, lagged_case_rate, was created to capture the case rate from the previous week for each ZIP code, enabling temporal analysis.
- Rows where lagged_case_rate could not be calculated (e.g., the first week of data for each ZIP code) were removed.

ii. Final Dataset Preparation:

- The final dataset was inspected to ensure completeness and exported as a CSV file for further analysis.

## 2.3 Analytical Methods

2.3.1 Regression Analysis

To investigate the relationship between socioeconomic, demographic factors, and COVID-19 case rates, Ordinary Least Squares (OLS) regression was used as the primary analytical approach. The analysis was performed in several stages:

a. Univariate Regression:
- Explored the individual relationship between adjusted gross income and case rates using data for the week ending April 4, 2020.
- Assessed the strength and direction of the relationship, as indicated by the regression coefficient and $R$^2.

b. Multivariate Regression:
- Included additional predictors, such as population density, racial composition (e.g., Hispanic/Latino population), and age distribution (e.g., age 18-24, age 65-74).
- Evaluated the relative importance of each variable and their combined explanatory power ($R$^2).

c. Lagged Variable Analysis:
- Incorporated lagged case rates as a predictor to capture temporal dynamics in disease spread.
- Investigated the role of previous week's case rates on current case rates while controlling for socioeconomic and demographic factors.

d. Interaction Effects:
- Modeled the interaction between income and Hispanic/Latino population to assess how the relationship between income and case rates varies across racial groups.
- Interaction terms were created as products of log-transformed variables for enhanced interpretability and model stability.

e. Log-Transformed Regression:
- Logarithmic transformations were applied to variables with high skewness (e.g., adjusted gross income, population density) to address heteroscedasticity and improve model fit.

f. Robust Standard Errors:
- To account for potential heteroscedasticity, robust standard errors (HC3) were used in multivariate models.

2.3.2 Software and Tools

- Python Libraries:
  - Pandas: Data preprocessing, merging, and feature engineering.
  - Stats models: Building OLS regression models and assessing results.
  - NumPy: Logarithmic transformations and numerical computations.
  - Plotly: Interactive visualizations like scatter plots, bar charts, and maps for exploring case rates, demographics, and socioeconomic factors.
  - Matplotlib: Static visualizations, including line and scatter plots, for clear and publication-ready trend analysis.
  - Seaborn: Statistical visualizations like density and box plots for analyzing case rate distributions across demographics and socioeconomic variables.

- Regression Diagnostics:
  - Multicollinearity was monitored using condition numbers and variance inflation factors (VIFs).
  - Residual analysis and $R$^2 were used to evaluate model fit.

# 3. Regression Analysis and Results

### 3.1 Descriptive Statistics

The dataset contained 6,191 observations with variables representing socioeconomic (e.g., adjusted gross income, population density), demographic (e.g., racial proportions, age distribution), and COVID-19-related metrics (e.g., case rates, lagged case rates). Key descriptive statistics include:

- Adjusted Gross Income:
  - Mean: $2,208,431
  - Median: $1,400,926
  - Range: $167,611 to $12,075,230
- Case Rates:
  - Significant variability across ZIP codes and weeks was observed.

### 3.2 Univariate Regression Analysis

Using data from April 4, 2020, a simple OLS regression was conducted between adjusted gross income and case rates:

- Key Results:
  - Adjusted Gross Income (β=−4.042×10^−5, p<0.001): A significant negative relationship, indicating that higher income ZIP codes had lower case rates.
  - $R$^2 = 0.269: Income explained approximately 26.9% of the variance in case rates.
- Interpretation:
  - Wealthier neighborhoods likely had greater access to healthcare and fewer exposure risks during the early pandemic stages.

### 3.3 Multivariate Regression Analysis

A multivariate regression model was developed using socioeconomic, demographic, and lagged case rate variables:

- Key Predictors:
  - Adjusted Gross Income (β=−2.957×10^-5, p<0.001): Income remained a significant negative predictor of case rates.
  - Population Density (β=−0.0044, p<0.001): Denser areas showed slightly reduced case rates, potentially due to stricter lockdown measures.
  - Age Groups:

    - 18-24 (β=0.1414, p<0.001): Younger adults contributed to increased transmission.
    - 65-74 (β=0.4906, p<0.001): Older age groups showed higher vulnerability to COVID-19.

### 3.4 Interaction Effects

Interaction terms revealed significant insights:

- Income-Hispanic Interaction (β=1885.52, p<0.001):
    - In neighborhoods with higher Hispanic/Latino populations, the protective effect of income diminished, highlighting systemic inequities.

### 3.5 Lagged Variables

Incorporating lagged case rates improved model performance:

- Lagged Case Rate (β=0.9602, p<0.001):
    - A strong positive relationship indicated that current case rates were heavily influenced by prior week's case rates.
- Model Fit:
    - $R$^2 = 0.823: The inclusion of lagged variables and interactions explained over 82% of the variance.

### 3.6 Expanded Model with November-December Data

A final model using November-December 2020 data included additional variables:

- Population Density (β=−22.13, p<0.001): Surprisingly negative, potentially reflecting behavioral adaptations (e.g., social distancing).
- Age 65 and Up (β=−0.5232, p<0.001): Higher age groups contributed less to case rates in later months, possibly due to early vaccine rollouts.
- Hispanic/Latino Proportion (β=2.0937, p<0.001): Continued to significantly increase case rates.
- Lagged Case Rate (β=0.7738, p<0.001): A persistent driver of current case rates.

### 3.7 Model Diagnostics

- Adjusted $R$^2: Models consistently demonstrated high explanatory power ($R$^2 = 0.744 to 0.8400).
- Residual Analysis: Residuals were approximately normal, though slight heteroscedasticity was observed and corrected using robust standard errors.
- Multicollinearity: The condition numbers flagged moderate multicollinearity, but key variables remained interpretable.

### 3.8 Summary of Regression Analysis Results

| Model Name | Dependent Variable | Independent Variables | R-squared | Key Findings |
|---|---|---|---|---|
| 3.2 Univariate Regression Analysis | Case Rate | Adjusted Gross Income | 0.269 | Higher income levels are negatively associated with case rates. |
| 3.3 Multivariate Regression Analysis | | Population Density, Age (18-24, 65-74), Race (Hispanic, White) | 0.344 | Age and race composition significantly influence case rates. |
| 3.4 Interaction Effects | | Log Income, Hispanic Proportion, Lagged Case Rate | 0.823 | Case rates are strongly influenced by lagged case rates and interaction effects with income. |
| 3.5 Lagged Variables | | Log Income, Log Density, Age 65+, Interaction Terms | 0.840 | Population density, age, and income interactions play key roles in predicting case rates. |
| 3.6 Expanded Model with November-December | | Income, Hispanic Proportion, Lagged Case Rate | 0.744 | Temporal factors, including prior case rates, significantly explain case rate variations in late 2020. |

Recommended Regression Models:

1. <u>Interaction Effects Model (R-squared = 0.823):</u>

✓ Strengths: High explanatory power, captures temporal dynamics, includes interaction between income and racial composition.

✓ Use Case: Predicts future case rates for timely public health interventions.

2. <u>Lagged Variables Model (R-squared = 0.840):</u>

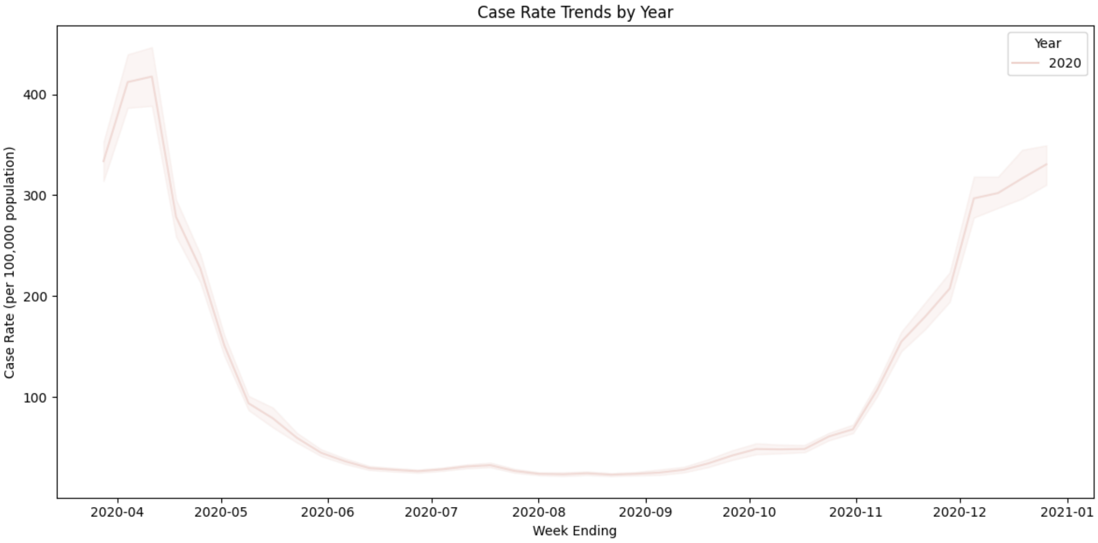
✓ Strengths: Highest explanatory power, integrates income, age, density, and temporal factors for comprehensive insights.

✓ Use Case: Guides interventions in high-risk areas like densely populated or aging communities.

# 4. Visualization

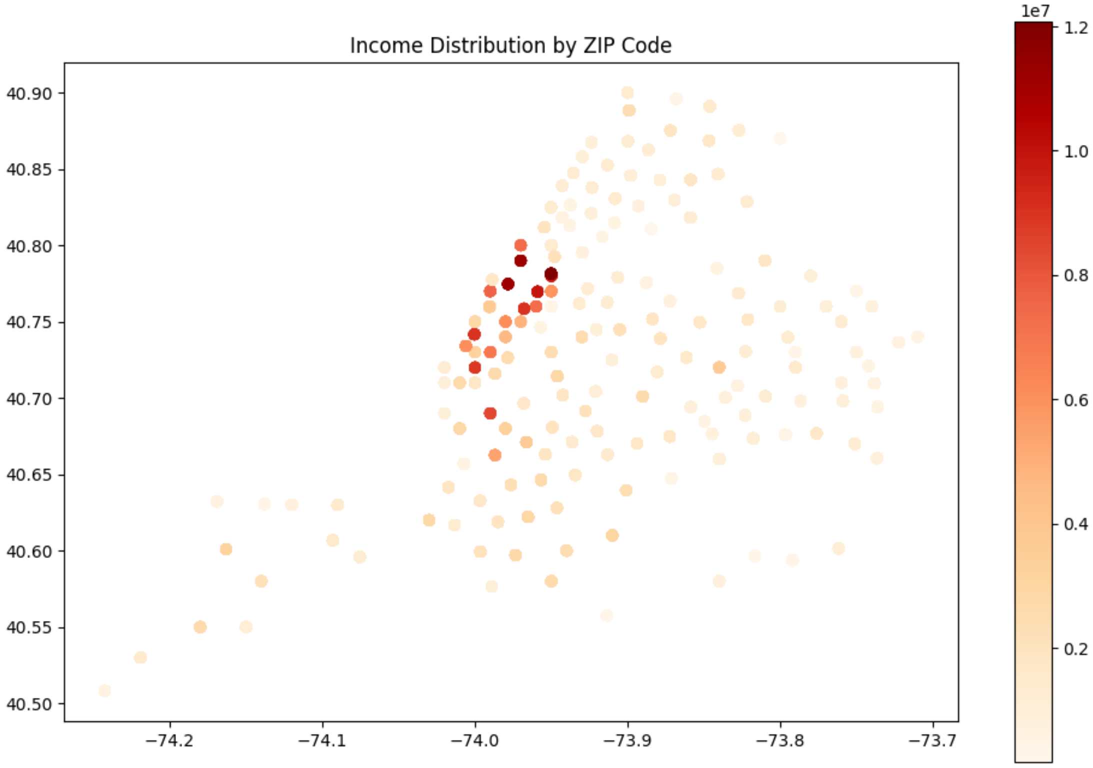
## 4.1 Case Rate Trends by Year

This figure shows the annual trend of COVID-19 case rates in 2020. Case rates reached a low point mid-year and surged significantly toward the end of the year.



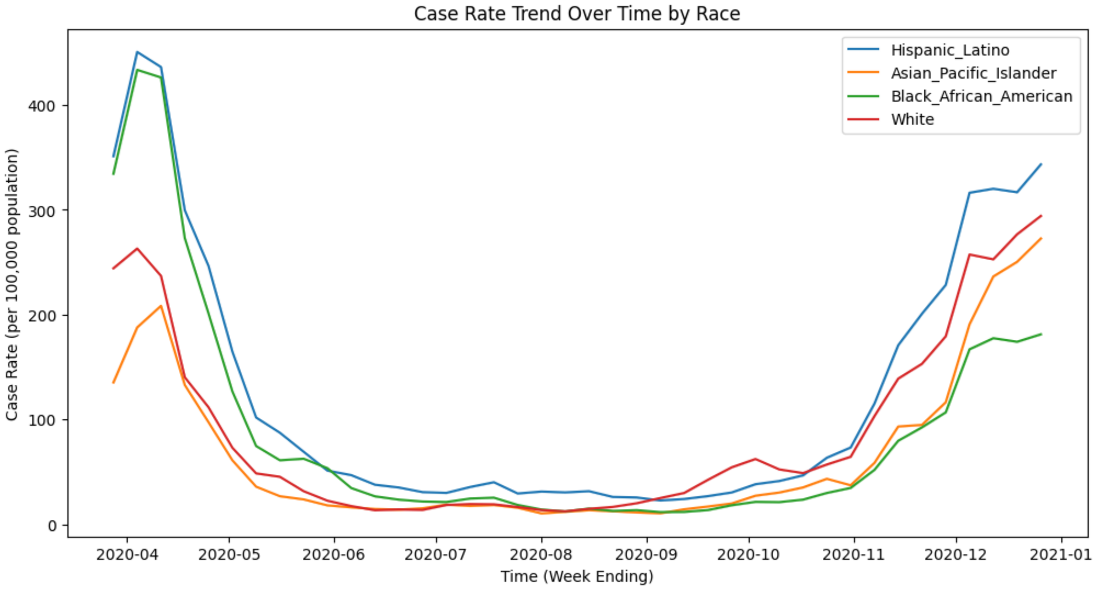
## 4.2 Income Distribution by ZIP Code

This figure displays the income distribution for each ZIP Code. The darker the color, the higher the income.



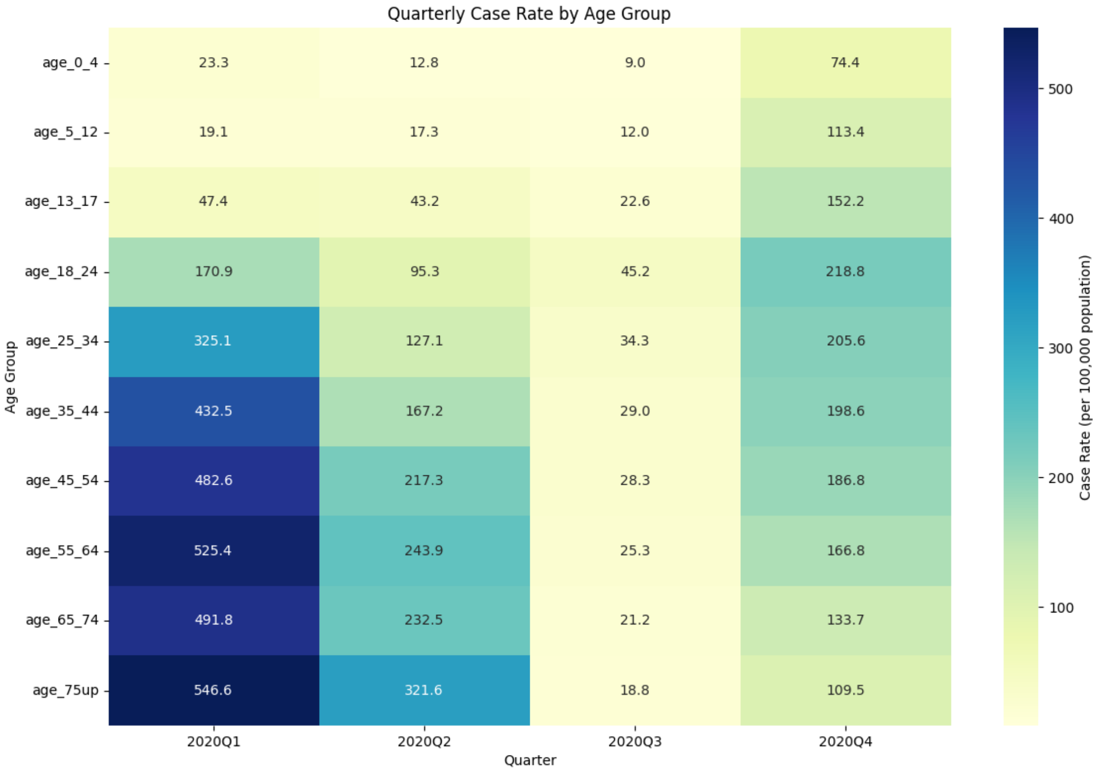
### 4.3 Case Rate Trend Over Time by Race

This figure depicts the changes in COVID-19 case rates over time for different racial groups, highlighting disparities in the pandemic's impact across racial demographics.



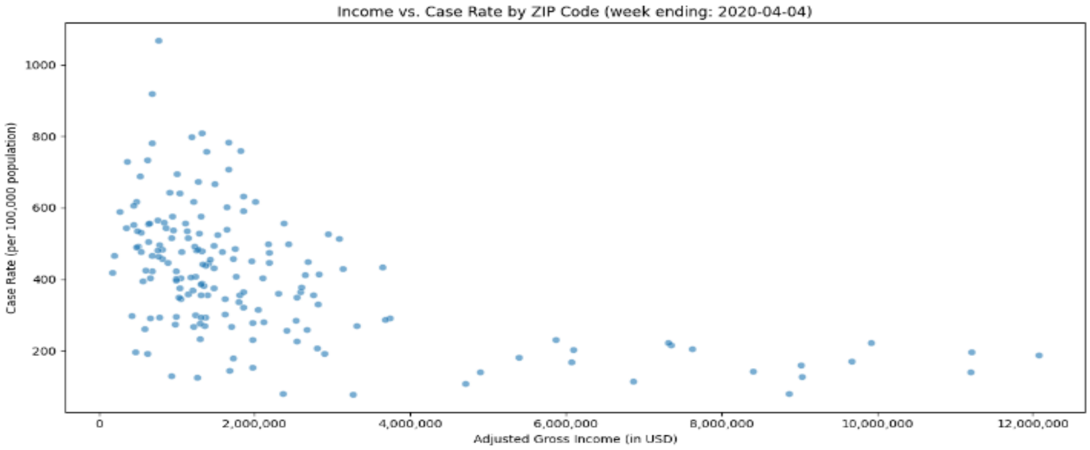
### 4.4 Case Rate Trend by Age Group

This heatmap shows the case rates of different age groups across the four quarters of 2020.

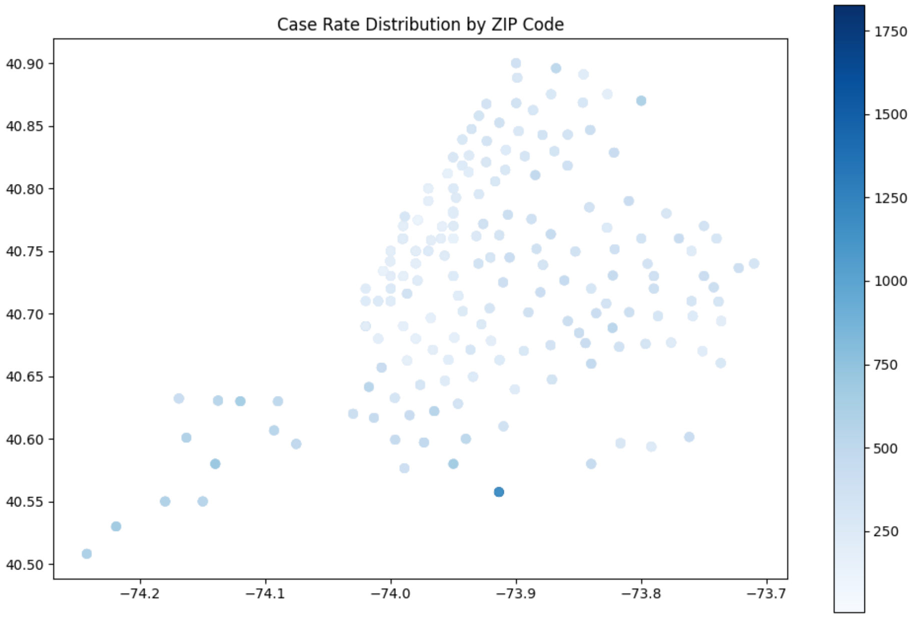


## 4.5 Income and Case Rate Relationship

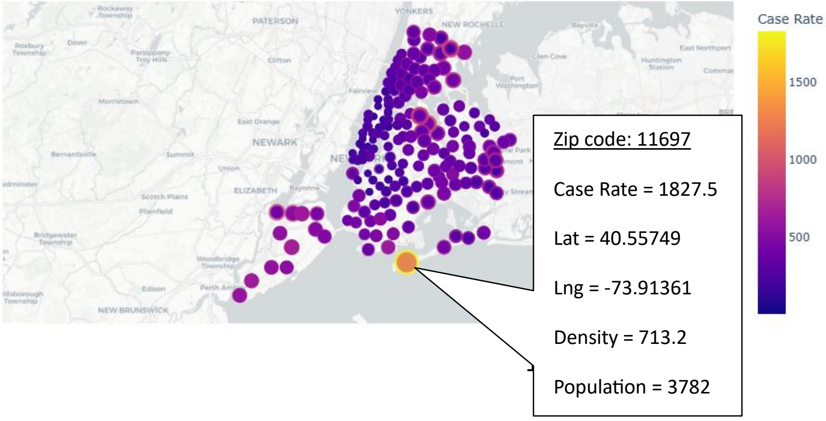
This scatter plot examines the relationship between income and case rates, revealing a trend where lower-income areas tend to have higher case rates.



This figure illustrates the distribution of COVID-19 case rates across ZIP Codes, highlighting areas with higher case rates and their respective spatial patterns.



Using Plotly's interactive visualization, the highest COVID-19 case rate was identified in ZIP code 11697.



Zip code: 11697

Case Rate = 1827.5

Lat = 40.55749

Lng = -73.91361

Density = 713.2

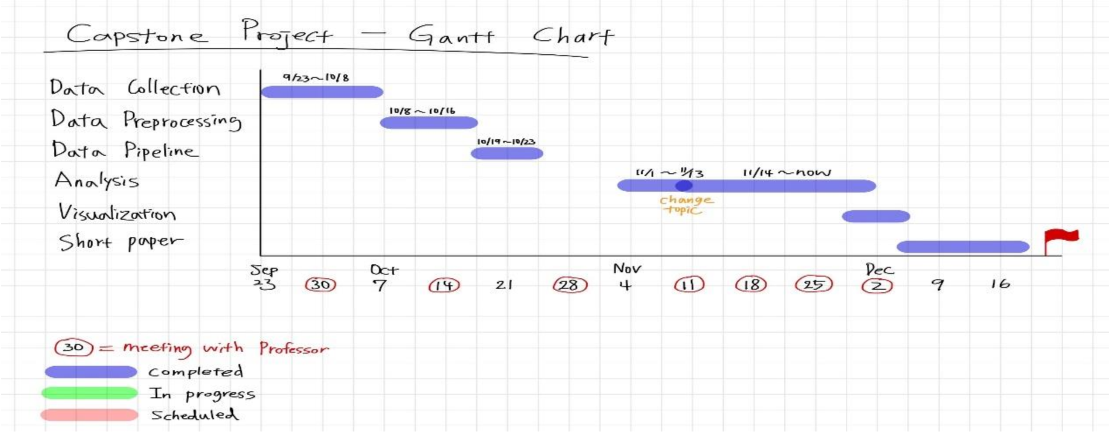
Population = 3782

# 5. Conclusion

- Lower-income areas experienced higher infection rates, indicating a negative correlation between income levels and case rates.
- The proportion of Hispanic/Latino populations exhibited a positive association with case rates, highlighting demographic influences on infection rates.
- Regression analysis identified lagged case rates as a significant predictor, emphasizing the temporal dynamics of COVID-19 transmission.

Effective data analysis and statistical methods are fundamental for understanding complex public health challenges and making informed decisions. This study demonstrates how rigorous analysis of demographic, socioeconomic, and temporal factors can uncover key patterns, such as the link between income levels and infection rates or the role of lagged case rates in predicting virus spread. By leveraging these insights, data-driven strategies can guide targeted interventions, resource allocation, and policymaking. Accurate and thoughtful use of data ensures that public health responses are not only evidence-based but also equitable and impactful, highlighting the critical role of analytics in addressing health disparities and preparing for future crises.

# 6. Appendix



# 7. References and Sources

- References

https://www.census.gov/library/stories/state-by-state/new-york-population-change-between-census-decade.html

https://www.pnas.org/doi/pdf/10.1073/pnas.2006853117

https://scholar.google.com/scholar?hl=en&as_sdt=0%2C31&q=covid+19+air+pollution&btnG=

https://www.nature.com/articles/s41598-021-90483-1

- Sources

https://github.com/nychealth/coronavirus-data/blob/master/Geography-resources/README.md

https://github.com/nychealth/coronavirus-data/blob/master/trends/caserate-by-modzcta.csv

https://www.kaggle.com/datasets/erdi28/zip-codes-demographics

https://www.kaggle.com/datasets/tsnowak/us-geographic-codes

https://www.kaggle.com/datasets/new-york-city/ny-demographic-statistics-by-zip-code?select=demographic-statistics-by-zip-code.csv

https://plotly.com/python/choropleth-maps/

https://plotly.com/python/line-charts/

https://plotly.com/python/line-and-scatter/

https://www.statology.org/pandas-lag/