RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

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

Kai-Yin Huang

12/15/2024

“Lower-income areas and higher Hispanic/Latino populations were linked to higher COVID-19 case rates, with lagged case rates identified as a key predictor of transmission dynamics.”

Table of contents

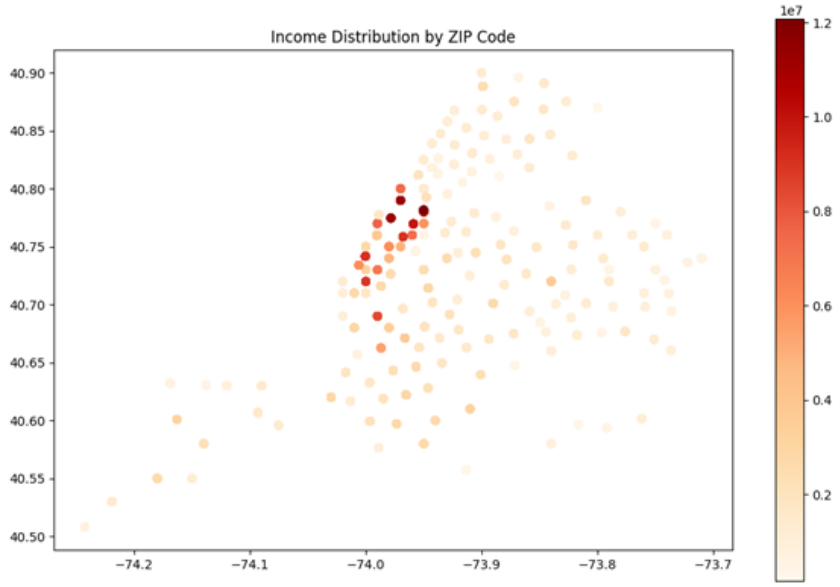
Introduction

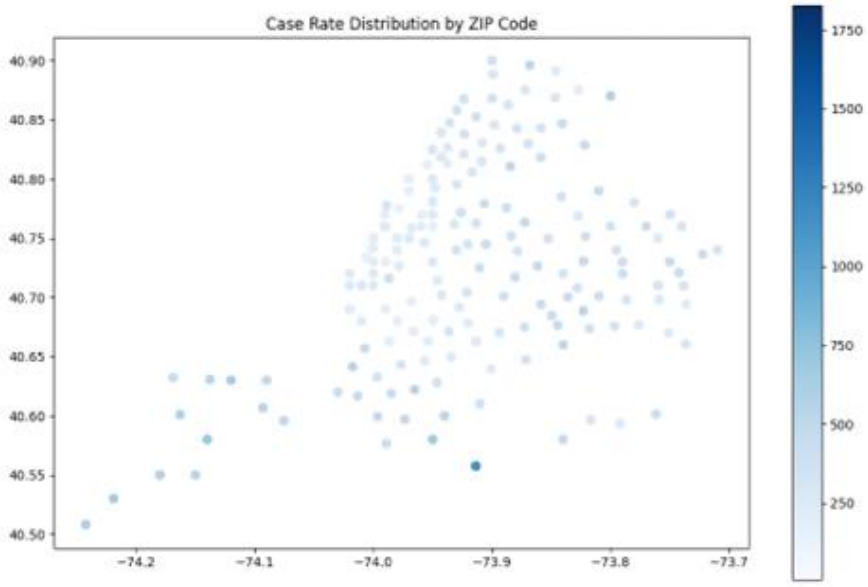
Background

The COVID-19 pandemic has exposed **significant disparities in infection rates**, particularly in densely populated urban areas like New York State. Socioeconomic and demographic factors, such as income inequality, racial composition, and population density, play a critical role in disease transmission.

Problem Statement

- Disparities in COVID-19 Infection Rates:
  - Lower income neighborhoods, higher proportions of racial minorities, and greater population density in urban areas (e.g., New York State) are disproportionately affected.





RUTGER_

Aims

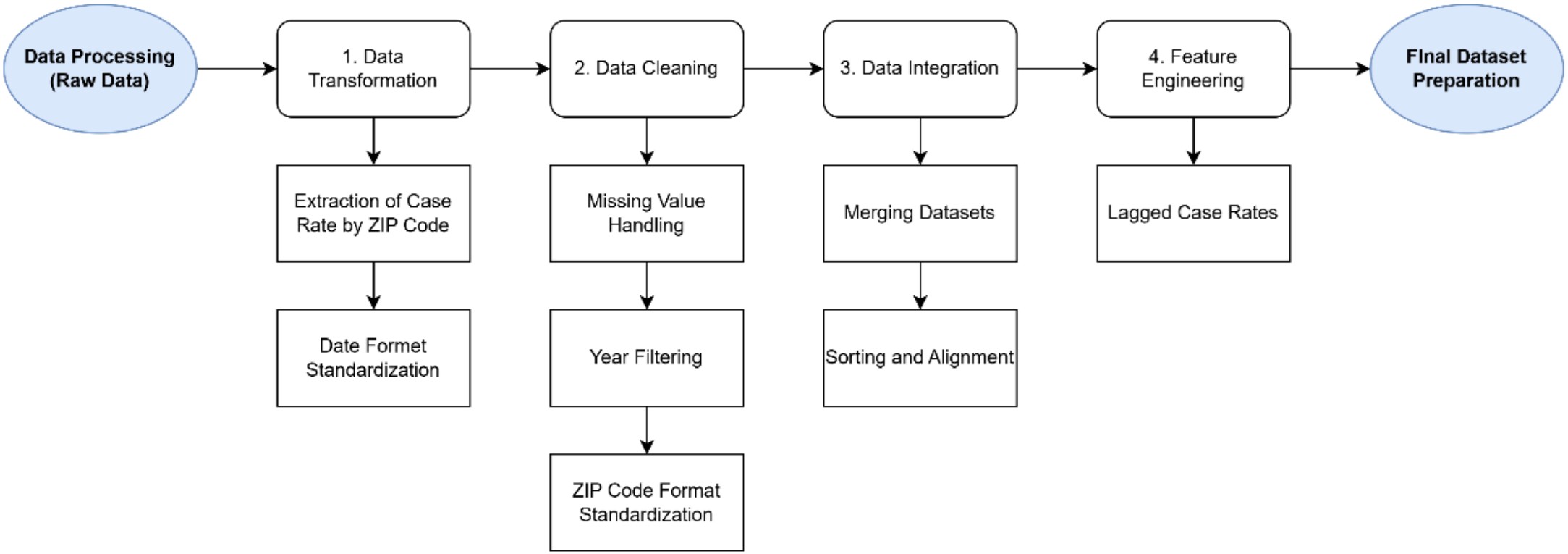
To provide **actionable insights for targeted policy interventions** through the investigation of socioeconomic and demographic determinants of COVID-19 case rates in New York State at the **ZIP-code level** during the pandemic's early stages.

Data and Methodology

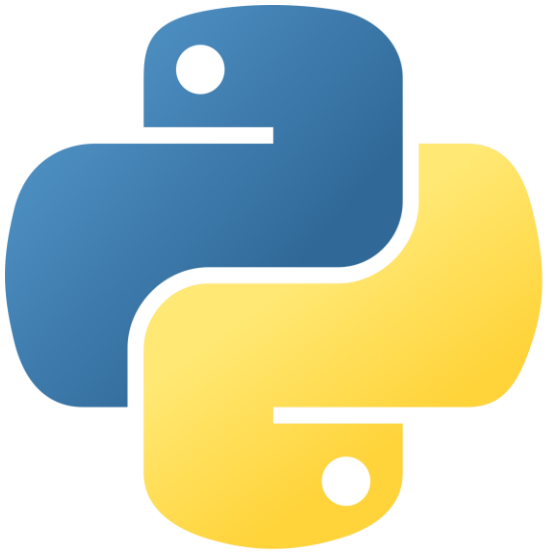
Data Fields

Field Name	Description
week_ending	The ending date of the week (used for temporal analysis.
case_rate	COVID-19 case rate per 100,000 people.
zip	ZIP code in a simplified format.
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.
dist_highway, dist2_large_airport, dist2_medium_airport, dist_to_shore	Distances to key infrastructure: highways, large/medium air-ports, and shorelines.
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_0_4 to age_75up	Population counts for different age groups (e.g., 0-4 years, 5-12 years, 13-17 years, etc.).
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).

Data Preprocessing



Methodology



$$Y = a + bX + \in$$

Key Results of Regression Analysis

Summary of Regression Analysis Results

Model Name	Dependent Variable	Independent Variables	R-squared	Key Findings
Univariate Regression Analysis	Case Rate	Adjusted Gross Income	0.269	Higher income levels are negatively associated with case rates.
Multivariate Regression Analysis		Population Density, Age (18-24, 65-74), Race (Hispanic, White)	0.344	Age and race composition significantly influence case rates.
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.
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.
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.

Summary of Regression Analysis Results

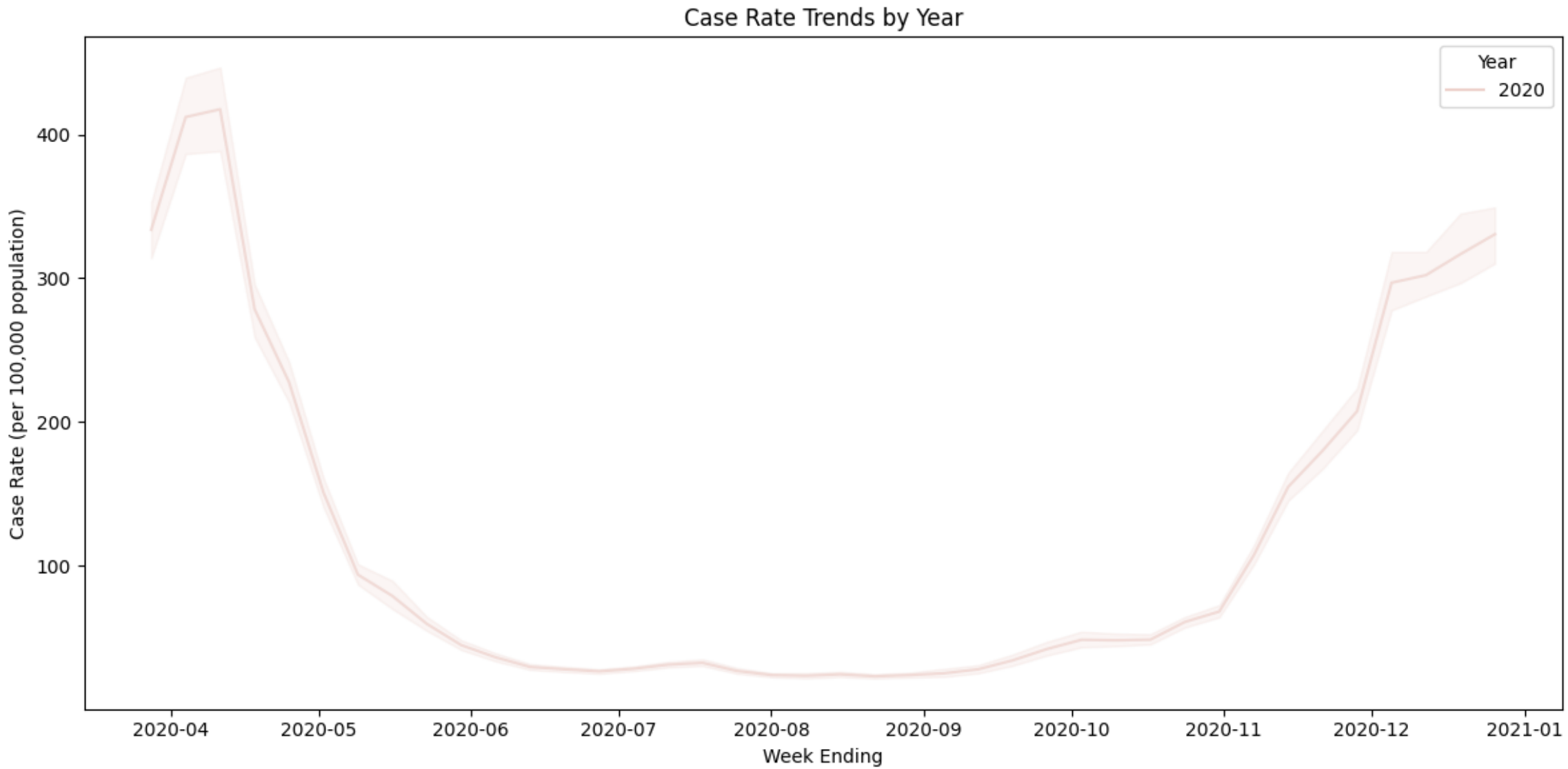
Model Name	Dependent Variable	Independent Variables	R-squared	Key Findings
Univariate Regression Analysis	Case Rate	Adjusted Gross Income	0.269	Higher income levels are negatively associated with case rates.
Multivariate Regression Analysis		Population Density, Age (18-24, 65-74), Race (Hispanic, White)	0.344	Age and race composition significantly influence case rates.
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.
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.
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

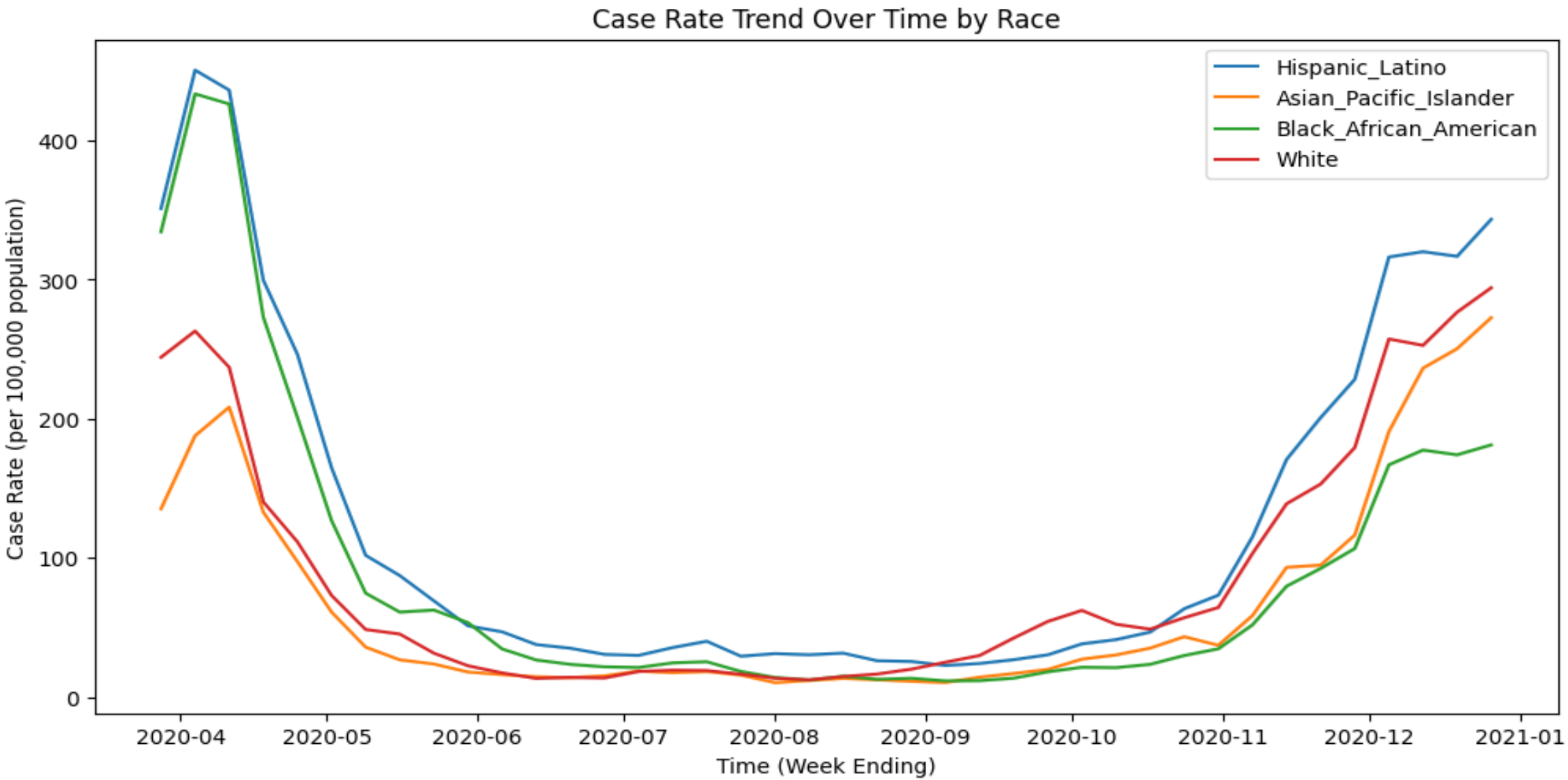
1. Interaction Effects Model (R-squared = 0.823):
    i. Strengths: High explanatory power, captures temporal dynamics, includes interaction between income and racial composition.
    ii. Use Case: Predicts future case rates for timely public health interventions.

2. Lagged Variables Model (R-squared = 0.840):
    i. Strengths: Highest explanatory power, integrates income, age, density, and temporal factors for comprehensive insights.
    ii. Use Case: Guides interventions in high-risk areas like densely populated or aging communities.

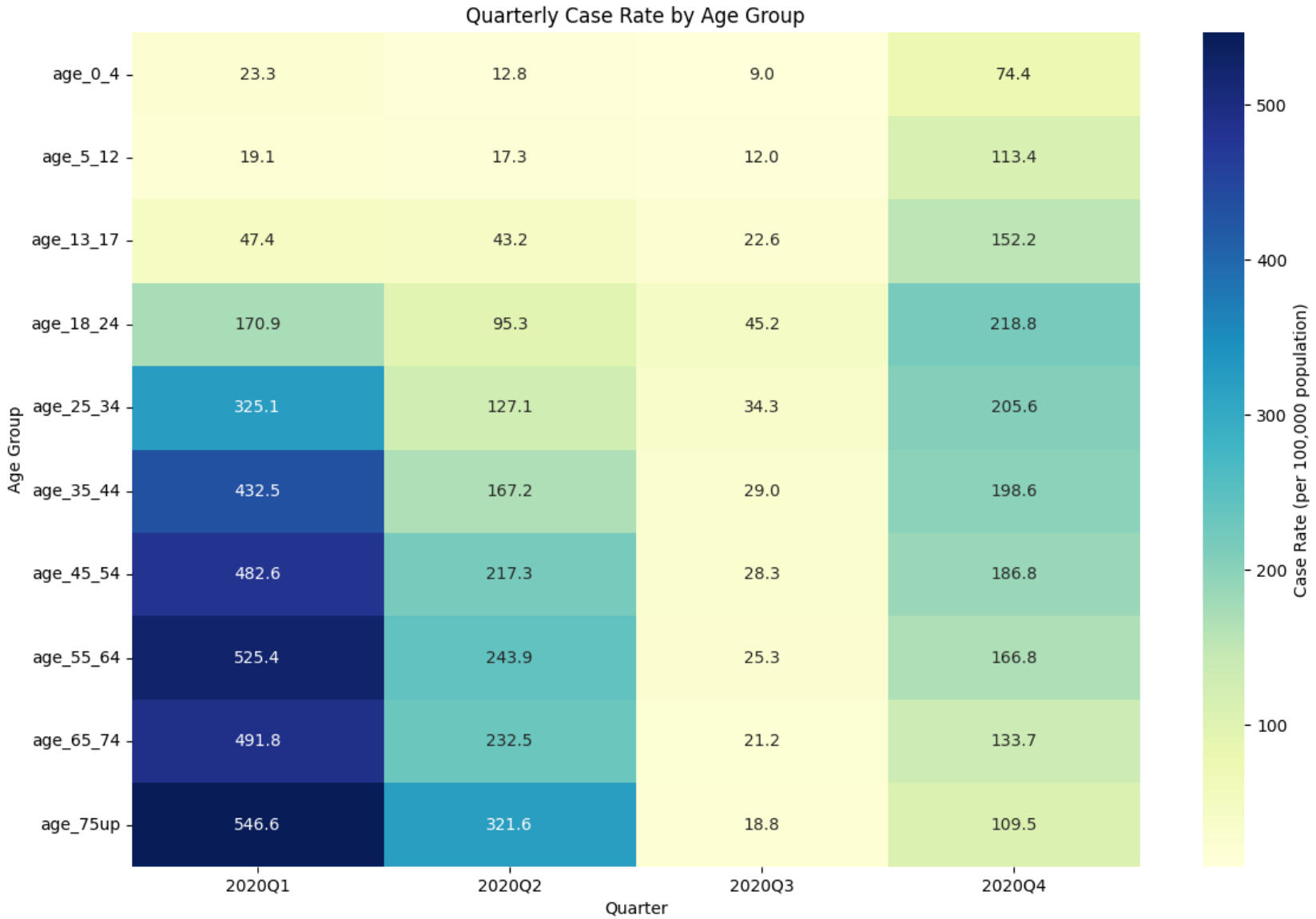
Visualization


Case Rate Trends by Year
Year
2020
Case Rate (per 100,000 population)
400
300
200
100
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11
2020-12
2021-01
Week Ending

RUTGERS

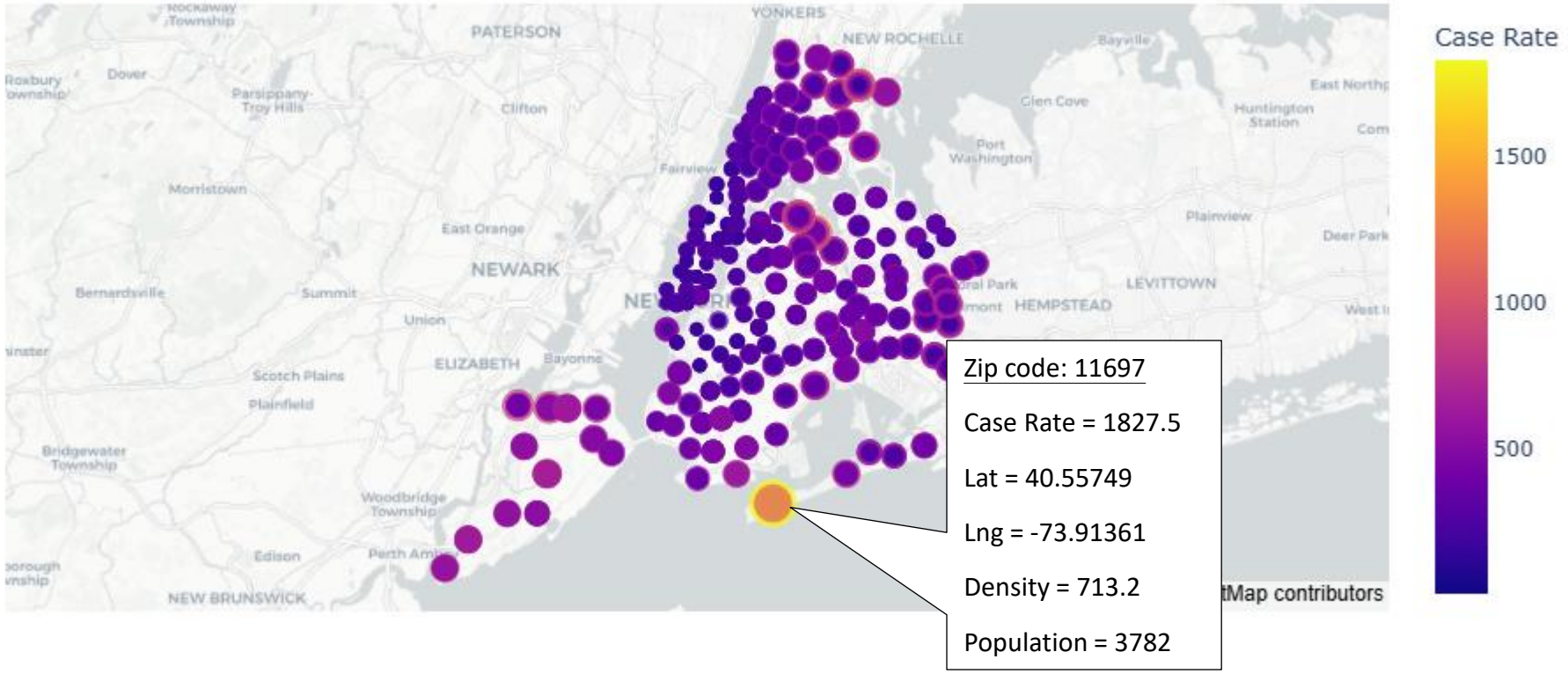

Case Rate Trend Over Time by Race
Hispanic_Latino
Asian_Pacific_Islander
Black_African_American
White
Case Rate (per 100,000 population)
0
100
200
300
400
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11
2020-12
2021-01
Time (Week Ending)

RUTGERS


Quarterly Case Rate by Age Group
age_0_4
age_5_12
age_13_17
age_18_24
age_25_34
age_35_44
age_45_54
age_55_64
age_65_74
age_75up
Age Group
23.3
12.8
9.0
74.4
19.1
17.3
12.0
113.4
47.4
43.2
22.6
152.2
170.9
95.3
45.2
218.8
325.1
127.1
34.3
205.6
432.5
167.2
29.0
198.6
482.6
217.3
28.3
186.8
525.4
243.9
25.3
166.8
491.8
232.5
21.2
133.7
546.6
321.6
18.8
109.5
2020Q1
2020Q2
2020Q3
2020Q4
Quarter
500
400
300
200
100
Case Rate (per 100,000 population)

RUTGERS

Case Rate by ZIP Code



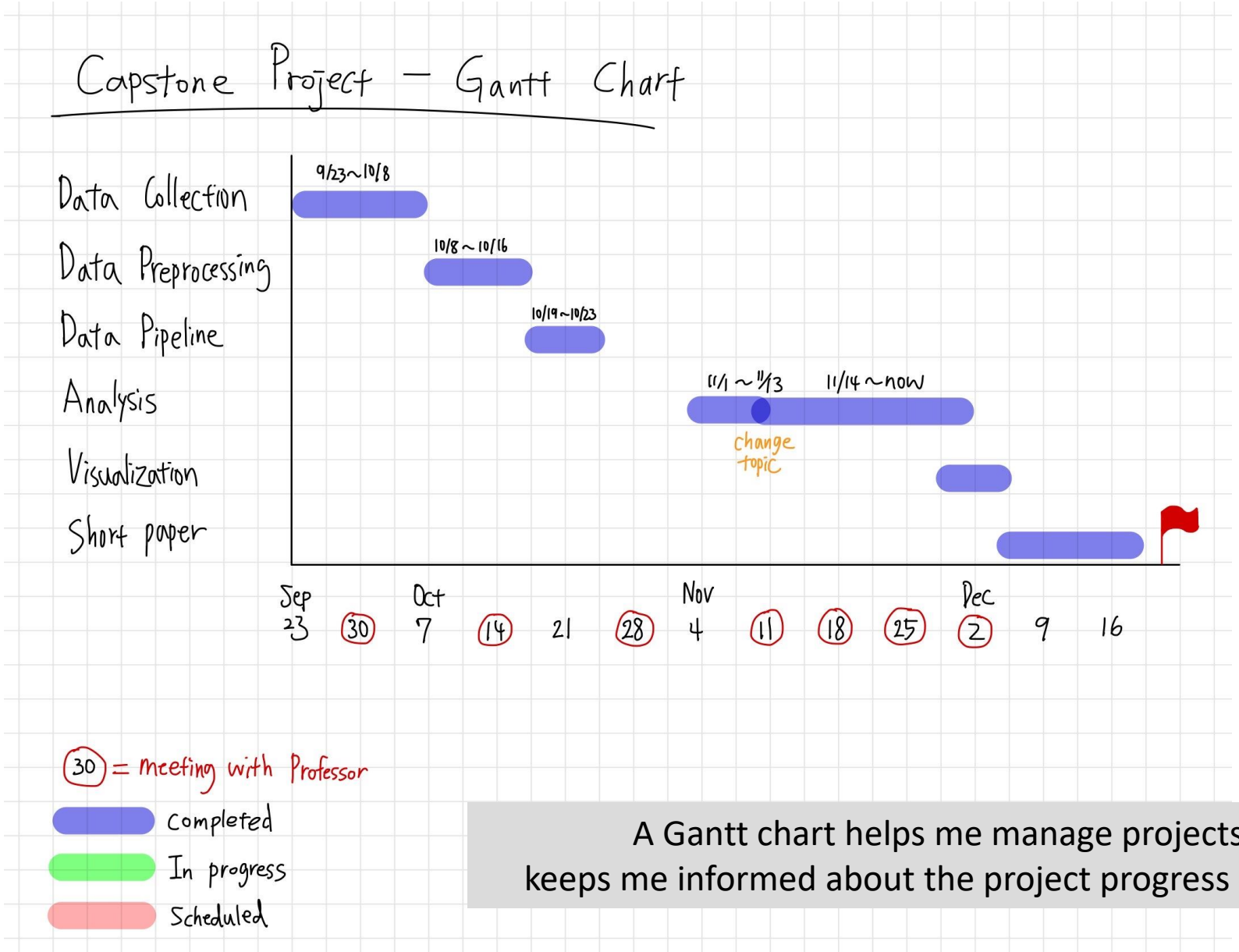
RUTGERS

Conclusion

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.

Appendix



A Gantt chart helps me manage projects and keeps me informed about the project progress at any time.

RUTGERS

Thank You