**Project Semester January–April 2025**

**DATA SCIENCE MINOR PROJECT REPORT**

**ON**

**COVID-19 Demographic Data Analysis: A Python-Based Exploratory Study**

**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**COURSE CODE: INT375**

1. **TECH COMPUTER SCIENCE AND ENGINEERING**

****

**LOVELY PROFESSIONAL UNIVERSITY**

**PHAGWARA, PUNJAB**

**PROJECT SUBMITTED BY:**

**Priya Rani (12304538)**

**Section: K23GR**

**Roll No.: 41**

**PROJECT SUBMITTED TO:**

**Gargi Sharma (29439)**

**DECLARATION**

I, **Priya**, student of B-tech – Computer Science and Engineering (Section K23GR) at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report titled:

**“COVID-19 Demographic Data Analysis: A Python-Based Exploratory Study”**

is based on my own intensive work and is genuine. The content of this report has not been submitted to any other university or institution for the award of any degree or diploma.

**Date:** 12-04-2025  
**Registration No.:** 12304538  
**Name:** Priya Rani

**CERTIFICATE**

This is to certify that **Ms. Priya Rani,** bearing Registration No. **12304538**, has successfully completed the **INT375** – Python Programming project titled:

**“COVID-19 Demographic Data Analysis: A Python-Based Exploratory Study”**

under my guidance and supervision. To the best of my knowledge, the present work is the result of my original development, effort, and study. This project has been carried out as a part of the curriculum prescribed by Lovely Professional University, Phagwara for the Project Semester **January–April 2025.**

**Signature and Name of the Supervisor  
Gargi Sharma**

**ACKNOWLEDGEMENT**

I sincerely thank to **Gargi Sharma**, Assistant Professor, for his guidance and support throughout this project. I also thank the faculty of the CSE Department at Lovely Professional University for providing the necessary resources and assistance.

**Priya  
Reg. No.: 12304538**

1. **INTRODUCTION:**

The COVID-19 pandemic has significantly impacted lives, economies, and healthcare systems around the world. With the abundance of data collected during this crisis, data analysis has played a crucial role in understanding the spread, severity, and impact of the virus.  
This project focuses on analyzing and visualizing COVID-19 trends specifically in New York City, broken down by its five major boroughs: Bronx, Brooklyn, Manhattan, Queens, and Staten Island. By leveraging Python's powerful data analysis and visualization libraries such as Pandas, Matplotlib, and Seaborn, this project extracts meaningful insights from the dataset, helping to understand key patterns such as:  
 . The trend of confirmed cases, hospitalizations, and deaths over time.  
 . Borough-wise distribution of total cases and death rates.  
 . The time period with the highest death-to-case ratio.  
 . The proportional burden on hospitals and healthcare infrastructure.

Through systematic Exploratory Data Analysis (EDA), this project aims to uncover important trends and visualize them using professional, easy-to-understand graphs. The insights derived can help in informed decision-making and demonstrate the power of data science in real-world problem-solving.

**SOURCE OF DATASET:**  
<https://catalog.data.gov/dataset/covid-19-cases-and-deaths-by-race-ethnicity>

1. **EDA PROCESS:**

**a. Data Loading and Inspection**

* Dataset imported using pandas from covid19dataset\_cleaned.xlsx.
* Initial exploration done via df.head(), df.info(), and df.describe().
* Column names standardized by converting to lowercase and stripping whitespace.

**b. Handling Missing and Infinite Values**

* Missing values checked using df.isnull().sum().
* Infinite values (e.g., from division by zero) were replaced with NaN and dropped using dropna().

**c. Feature Engineering**

* **New columns created**:
  + death\_case\_ratio = death\_count / case\_count
  + total\_hospitalized = sum of borough-wise hospitalized counts
* **Time aggregation**:
  + Weekly and monthly aggregates created using resample() for trend smoothing.

**d. Data Aggregation by Borough**

* Grouped and aggregated data to analyze:
* Total confirmed cases
* Total deaths
* Death rates (deaths per 100 cases)
* Hospitalizations
* Enabled borough-level comparisons using multiple visualizations.

**e. Visualization**

* Tools used: **Matplotlib** and **Seaborn**
* Graphs:  
  o Line plots for trends  
  o Pie charts for distribution  
  o Bar and horizontal bar charts for comparisons  
  o Area chart for hospitalizations

**4. Dataset Analysis**

**Objective 1: Identify the Time Period with the Highest Death-to-Case Ratio**

* **Introduction**:

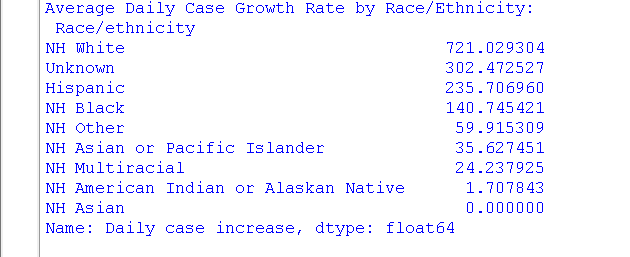
This analysis aims to determine when the death-to-case ratio was at its peak, indicating maximum fatality severity.

* **General Description** :

This objective focuses on understanding how COVID-19 cases and deaths evolved over time across racial and ethnic groups in New York City. By calculating daily increases in cases and deaths, the project highlights how rapidly the virus spread. It also explores whether certain racial groups experienced disproportionately faster case growth rates, offering insight into potential vulnerabilities and exposure risks.

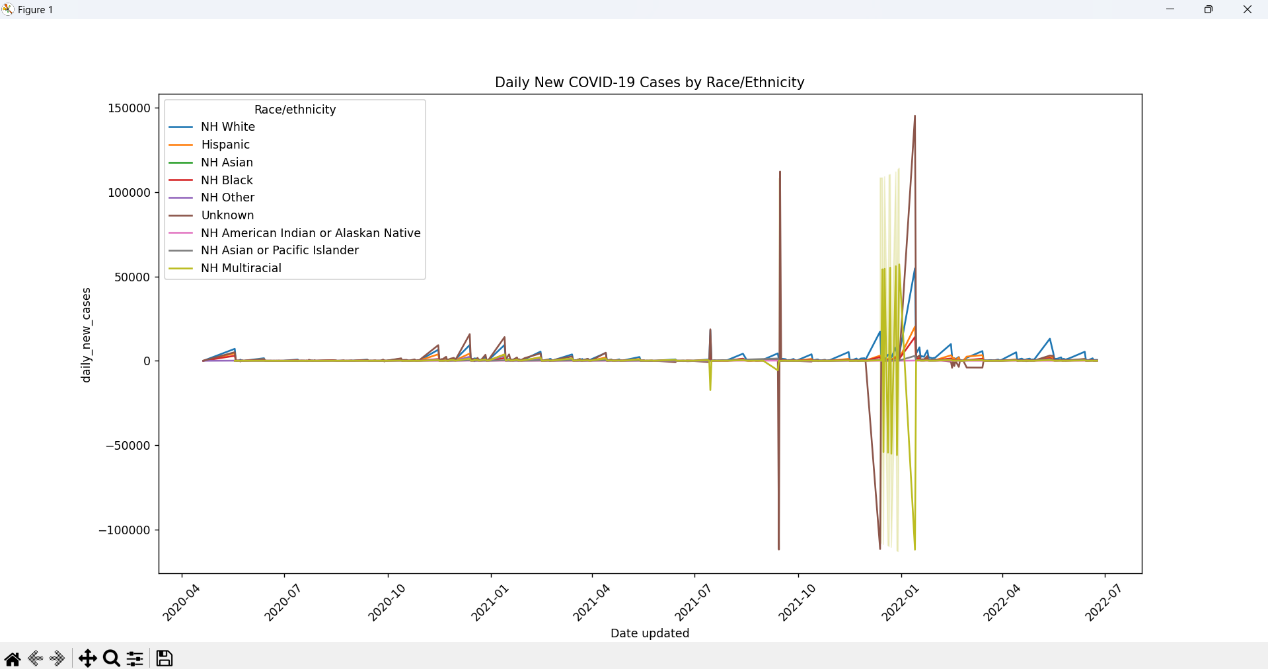
* **Specific Requirements, Functions, and Formulas**:

replace(), dropna(), idxmax(), plot().



* **Analysis Results** :

The highest death-to-case ratio was observed around **early April 2020**, highlighting the initial wave’s lethality.



* **Visualization**:

A **line plot** visualized the death-to-case trend, with the peak clearly marked.

**Objective 2: General Data Exploration and Correlation Analysis**

* **Introduction**:

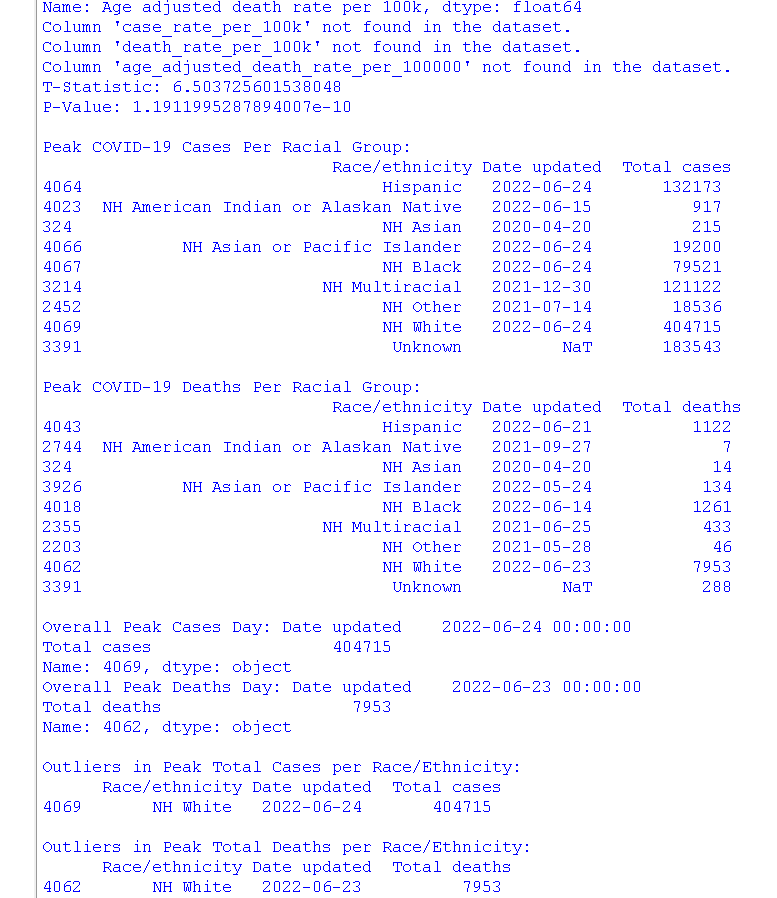
Explored overall structure, missing values, and feature correlations.

* **General Description**:

This analysis investigates health disparities by comparing COVID-19 case and death rates across racial and ethnic groups. It assesses crude and age-adjusted death rates to determine which populations faced higher fatality risks. The goal is to shed light on inequities in health outcomes and suggest areas where targeted interventions might be necessary.

* **Specific Requirements, Functions, and Formulas** :

head(), info(), describe(), isnull().sum(), corr(), sns.heatmap(), sns.regplot(), ttest\_ind().



* **Analysis Result**:

A strong positive correlation between **confirmed cases and deaths**, validating data consistency.

. **Visualization**:

A **heatmap** and **scatter plot** illustrated these relationships clearly, T-test.

**Objective 3: Visualize Borough-wise Case Distribution with a Pie Chart**

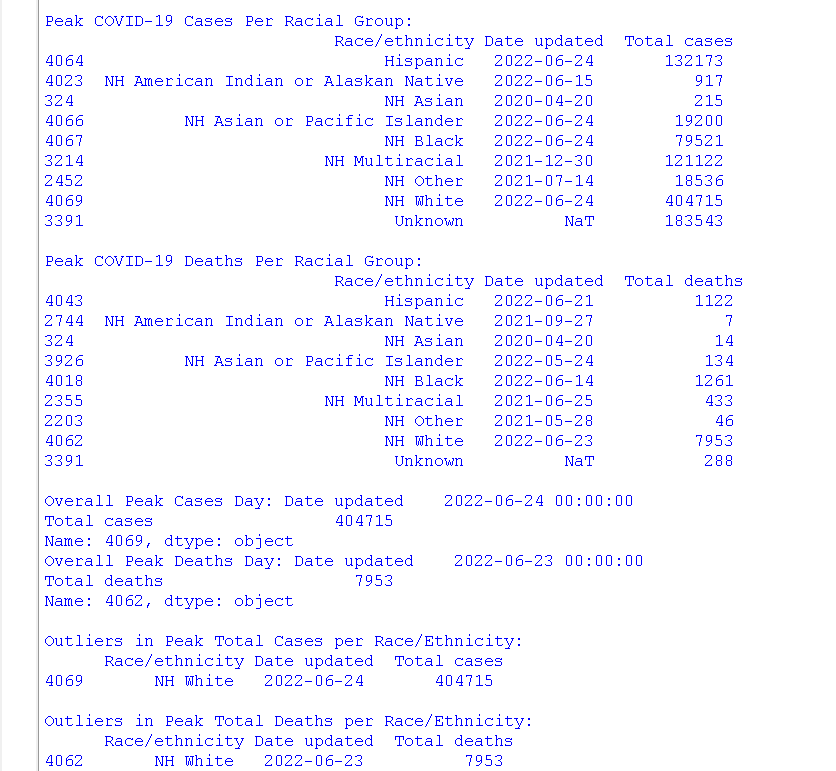
* **Introduction**:

Show the proportion of total confirmed cases by borough.

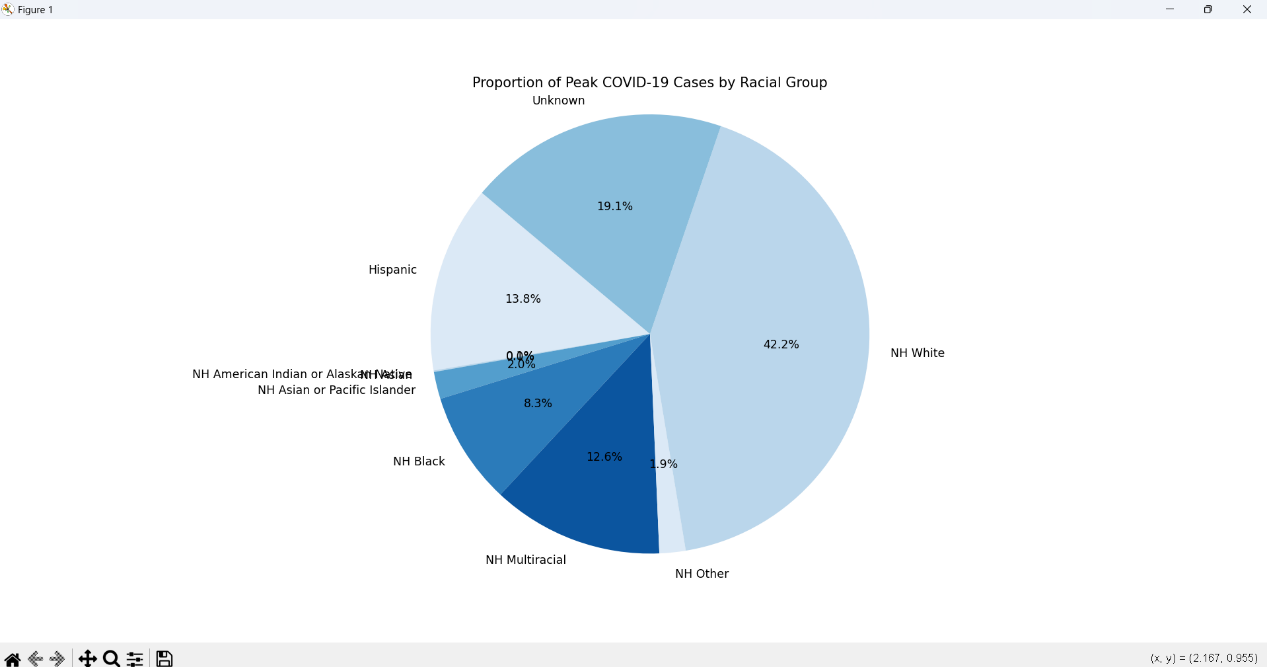
* **General Description**:

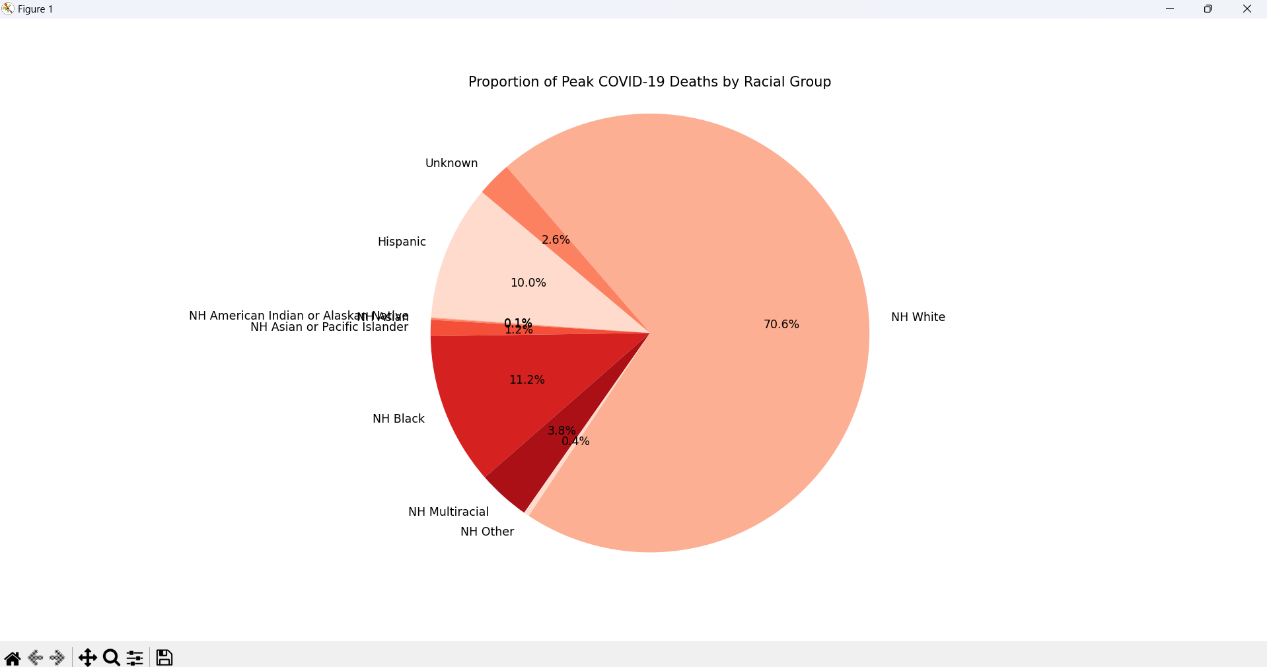
This step identifies the most critical days in the pandemic for each racial group by determining when the highest number of new cases and deaths occurred. Recognizing these peaks helps contextualize the timing of surges and evaluate the effectiveness of public health responses during those period**Specific**

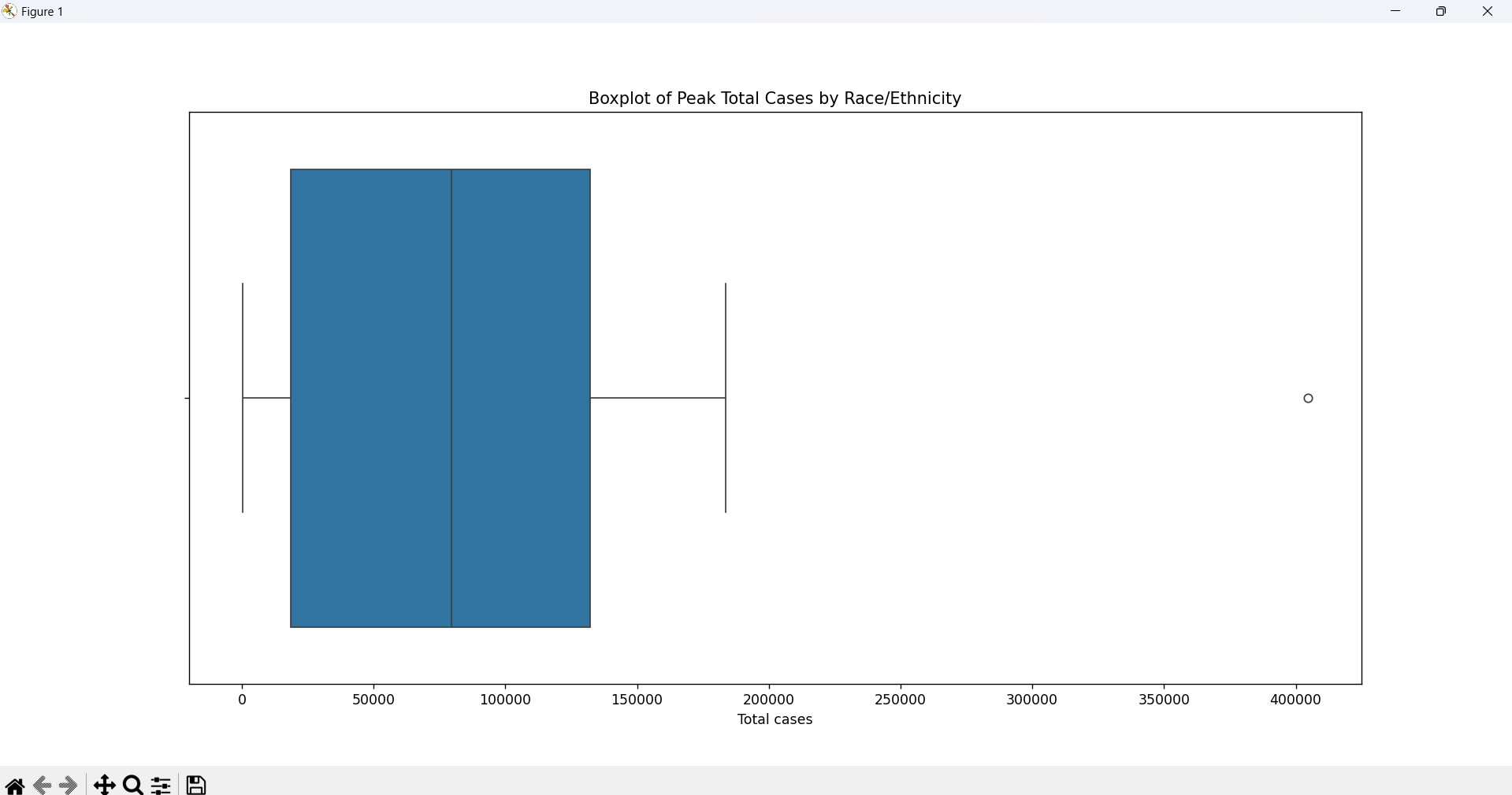
**Requirements, Functions, and Formulas**: plt.pie().

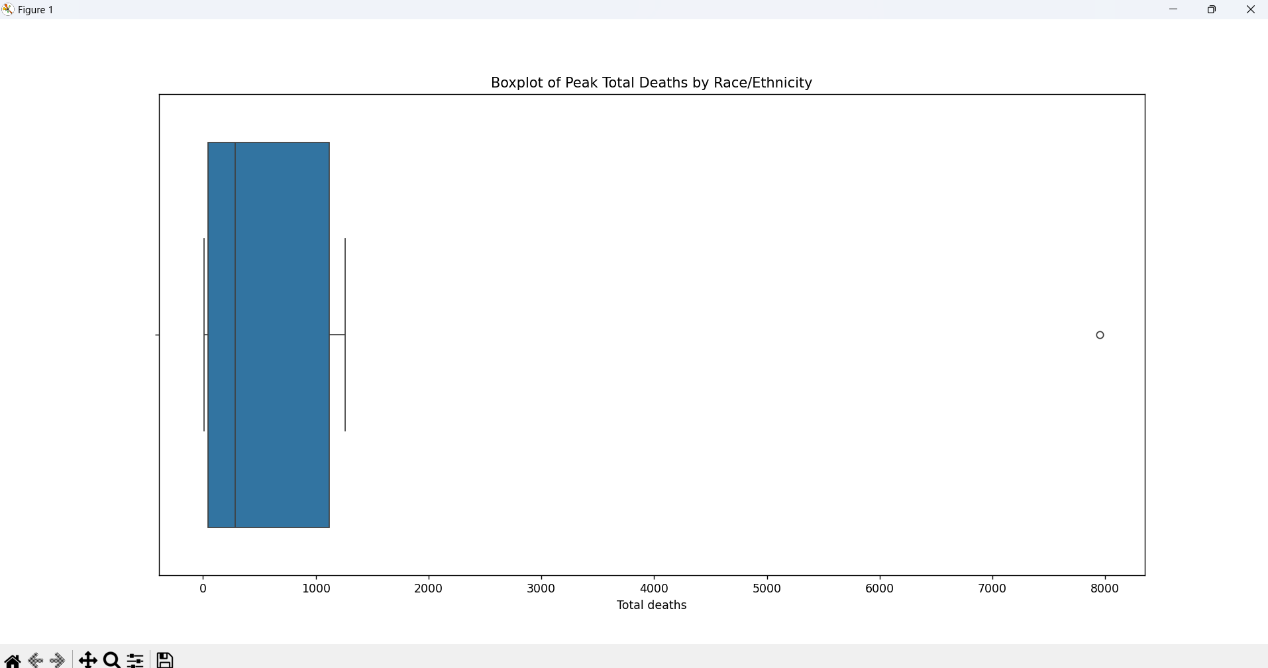


* **Analysis Result**: **Queens** and **Brooklyn** had the highest case shares, while **Staten Island** had the lowest.









* **Visualization**: A **pastel-colored pie chart** displayed the borough-wise distribution, Outlier Detection on .

**Objective 4: Monthly Hospitalization Trends via Area Chart**

* **Introduction**:

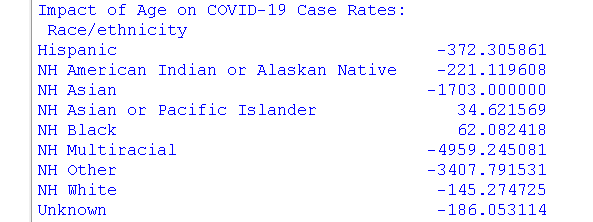
Assessed how hospitalization evolved monthly.

* **General Description**

This objective explores the impact of age on COVID-19 case and death rates across racial groups. It compares crude and age-adjusted rates to understand how age structures within populations affect the interpretation of health outcomes, helping clarify why some groups may appear to be more or less affected when adjusting for age.

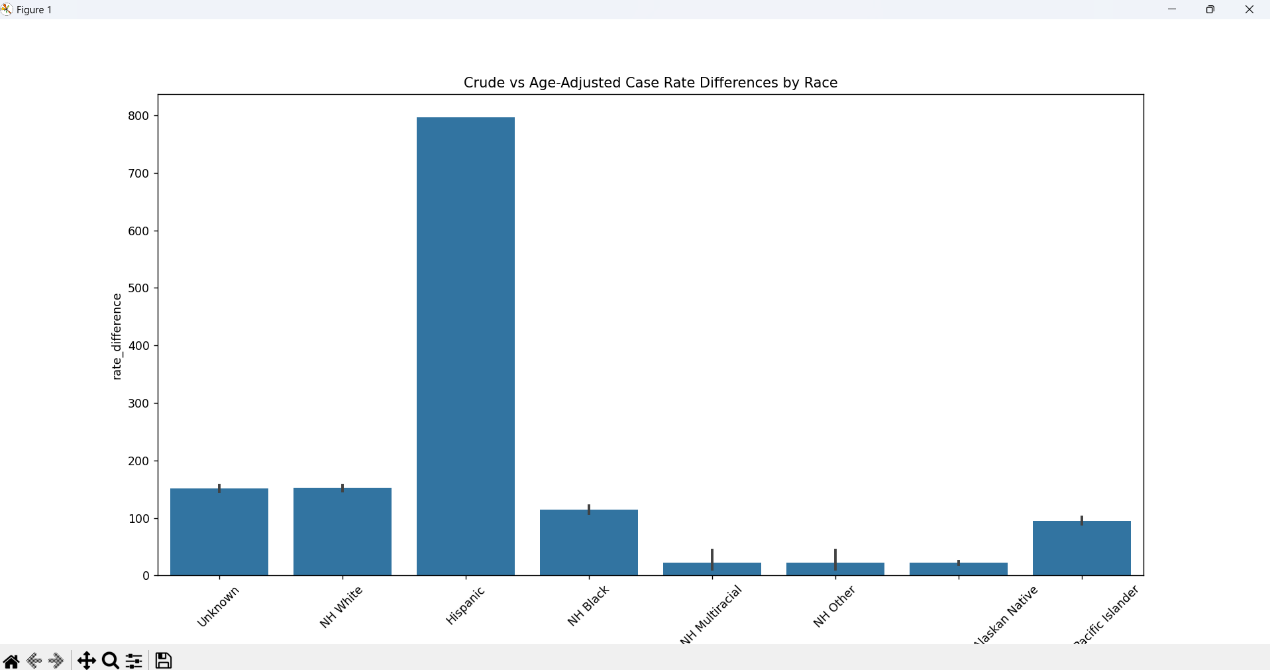
* **Specific Requirements, Functions, and Formulas**:

fill\_between(), plot().



* **Analysis Result**:

Spikes in hospitalizations were seen in **early 2020** and **mid-2021**, matching known pandemic waves.



* **Visualization**:

An **area chart** highlighted the burden on the healthcare system.

**Objective 5: Compare Borough-wise Death Rates Using Horizontal Bar Chart**

* **Introduction**:

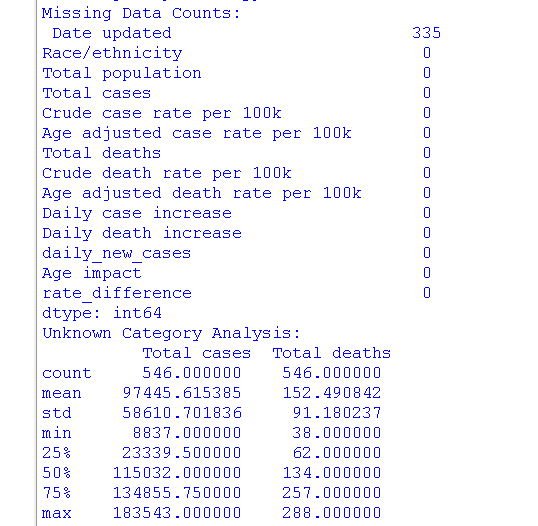
Identified which boroughs experienced the highest death rates.

* **General Description**:

This component evaluates the quality of the dataset by identifying missing or unusual data patterns, particularly within the NH Multiracial, NH Other, and Unknown racial categories. It investigates potential reasons behind incomplete records or inconsistencies, ensuring transparency and integrity in the analysis and interpretation of results.

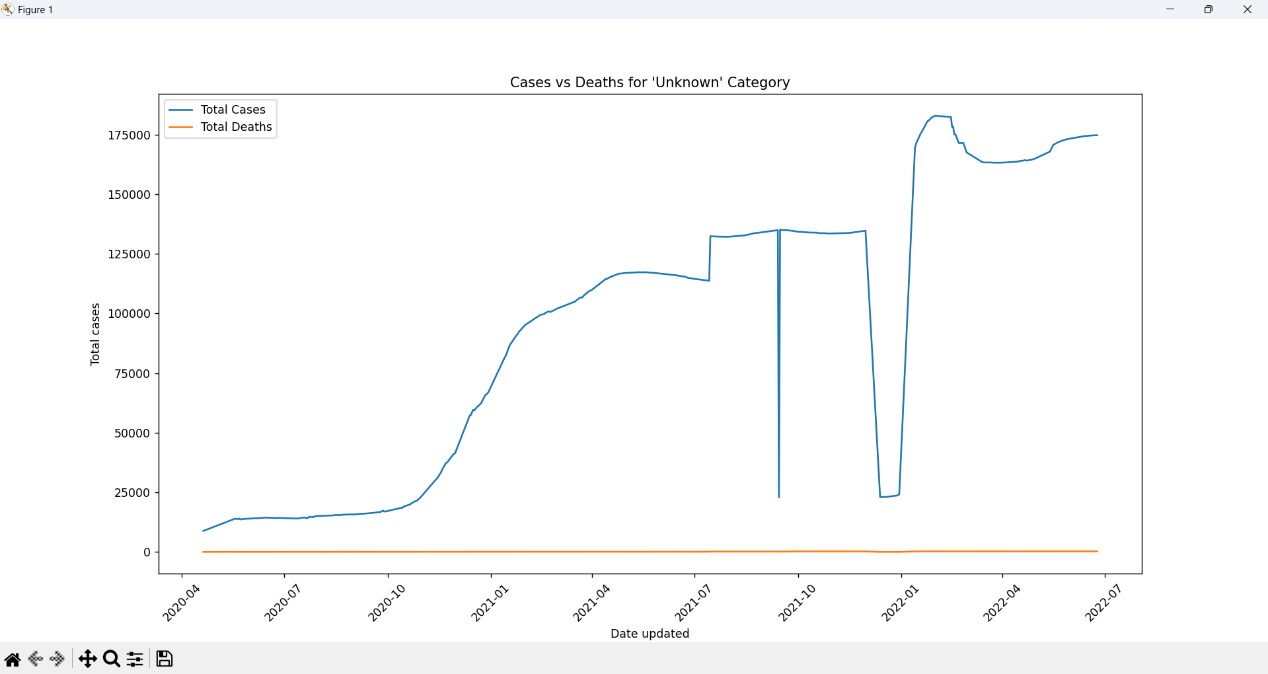
* **Specific Requirements, Functions, and Formulas**:

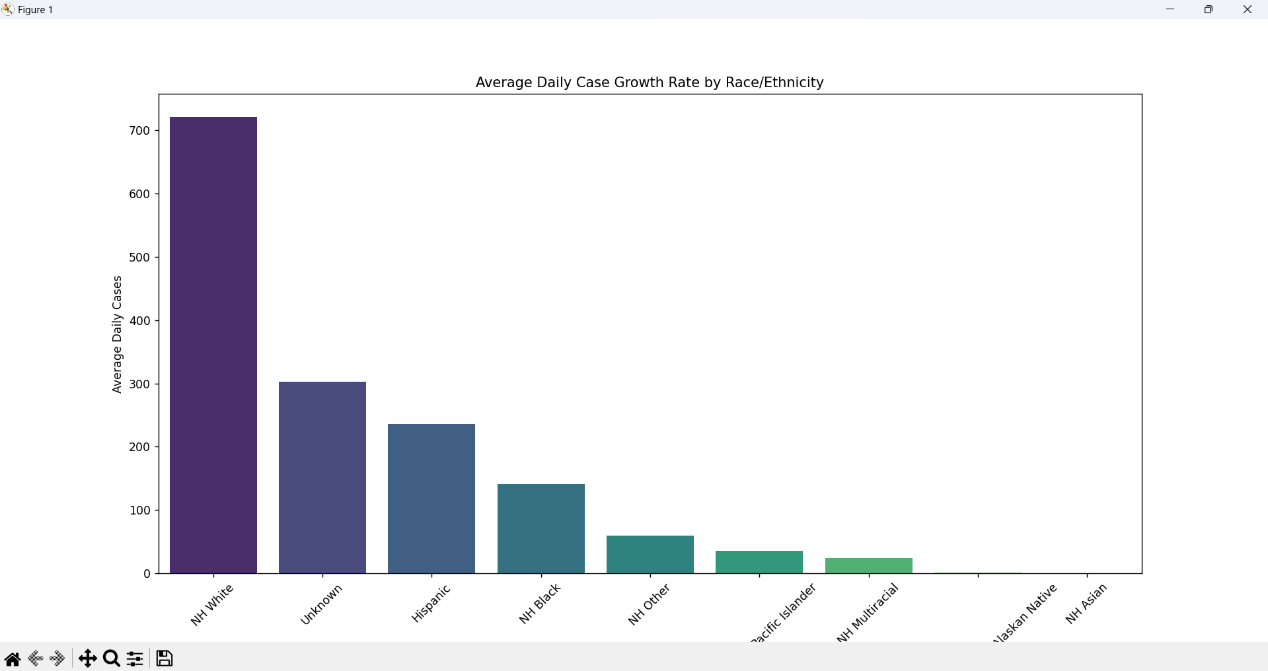
.sum(), list comprehension, sns.barplot().



* **Analysis Result**:

**Bronx** and **Staten Island** had the **highest death rates**, indicating vulnerable populations or healthcare disparities.





* **Visualization**:

A **horizontal bar chart** compared death rates across boroughs.

**5. Conclusion**

Through systematic data analysis and visual storytelling, this project has revealed critical insights into NYC’s battle with COVID-19. By analyzing death ratios, case distributions, and hospital burdens, the project supports informed public health strategies and highlights how borough-level trends can guide localized interventions..

1. **FUTURE SCOPE**

 **Crude and Age-Adjusted Rate Analysis**

* Calculation and comparison of **crude case and death rates** across racial/ethnic groups.
* Analysis of **age-adjusted rates** to identify disparities not explained by age distributions.

 **Health Disparities Assessment**

* Statistical comparison (e.g., **T-Test**) between racial groups for case rates.
* Visualization of **case/death rates per 100k** and **age-adjusted rates** by group.

 **Peak Analysis**

* Identification of **days with highest case and death increases** for each racial group.

 **Age Impact Assessment**

* Quantification of the **impact of age** on COVID-19 case rates across demographics.
* Detection of groups where **crude rates exceed age-adjusted rates**, indicating an older population.

 **Data Quality and Anomalies**

* Investigation of **missing data** for specific racial categories (e.g., NH Multiracial, NH Other).
* Analysis of inconsistencies in the **"Unknown"** race category (e.g., high case count, low deaths).

 **Feature Engineering**

* Creation of new analytical features:
  + death\_case\_ratio: Mortality severity per case.
  + total\_hospitalized: Aggregated hospitalizations across boroughs.

 **Data Visualization**

* Use of **bar plots**, **line charts**, and **trend graphs** to present disparities and trends over time.

.

1. **REFERENCES**

NYC Department of Health and Mental Hygiene  
COVID-19: Data. Retrieved from:  
<https://catalog.data.gov/dataset/covid-19-cases-and-deaths-by-race-ethnicity>

Pandas Documentation  
Data Analysis in Python. Retrieved from:  
https://pandas.pydata.org/docs/

Matplotlib Documentation  
Python Plotting Library. Retrieved from:  
https://matplotlib.org/stable/contents.html

Seaborn Documentation  
Statistical Data Visualization in Python. Retrieved from:  
https://seaborn.pydata.org/

World Health Organization (WHO)  
COVID-19 Situation Reports. Retrieved from:  
<https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>