

DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

COVID-19 Cases, Tests, and Deaths by ZIP Code - Historical

Submitted by:

Samarth Garg

Registration No.: 12324193

Programme: Bachelor of Technology

Section: K23GN

Course Code: INT375

Under the Guidance of

Aashima Bansal

(UID: 28968)

Discipline of CSE/IT

Lovely School of Computer Science & Engineering

Lovely Professional University, Phagwara



L OVELY
P ROFESSIONAL
U NIVERSITY

Transforming Education Transforming India

CERTIFICATE

This is to certify that Samarth Garg, bearing Registration no. 12324193 has completed INT375 project titled, “COVID-19 Cases, Tests, and Deaths by ZIP Code - Historical” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor: Aashima Bansal

Designation of the Supervisor:

School of Computer Science & Engineering

Lovely Professional University

Phagwara, Punjab.

Date: 12-04-2025

DECLARATION

I, Samarth Garg, student of Bachelor of Technology under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-2025

Signature: Samarth Garg

Registration No.: 12324193

ACKNOWLEDGEMENT

I express my heartfelt gratitude to Aashima Bansal, my mentor, for providing guidance and support throughout this project. I am thankful to Lovely professional University for providing the resources and environment to conduct this analysis. I also acknowledge the open-source community for making tools like Python, Pandas, NumPy, Matplotlib, and Seaborn available, which were instrumental in this work.

Table of Contents

1. Introduction
2. Source of Dataset
3. EDA Process
4. Analysis on Dataset
 - 4.1 Objective 1: Data Cleaning and Handling Missing Values
 - 4.2 Objective 2: Weekly COVID-19 Trend Analysis
 - 4.3 Objective 3: Correlation Analysis Between Cases, Deaths, and Tests
 - 4.4 Objective 4: ZIP Code Wise COVID-19 Severity Analysis
 - 4.5 Objective 5: Outlier Detection in Case Rates
 - 4.6 Objective 6: Exploratory Data Analysis (EDA) with Advanced Visualizations
5. Conclusion
6. Future Scope
7. References

1. Introduction

The COVID-19 pandemic has had a profound impact globally, necessitating data-driven insights to understand its spread and effects. This project analyses a COVID-19 dataset containing weekly cases, deaths, tests, and case rates across ZIP codes, aiming to uncover temporal trends, regional severity, correlations, and anomalies. Using Python tools (Pandas, NumPy, Matplotlib, Seaborn), the analysis is structured into six objectives:

- Cleaning the dataset to ensure quality.
- Analysing weekly trends to identify waves.
- Examining correlations between cases, deaths, and tests.
- Assessing severity by ZIP code.
- Detecting outliers in case rates.
- Conducting exploratory data analysis (EDA) for deeper insights.

The report adheres to the course syllabus, focusing on basic programming, data manipulation, visualization, and statistical analysis, and provides actionable findings for public health stakeholders.

2. Source of Dataset

The dataset, covid_data.csv, is sourced from <https://catalog.data.gov/dataset/covid-19-cases-tests-and-deaths-by-zip-code>. It contains COVID-19 metrics across ZIP codes, with the following key columns:

- ZIP Code: Identifier for geographic areas.
- Week Number, Week Start, Week End: Temporal markers for weekly data.
- Cases - Weekly: Number of new cases per week.
- Deaths - Weekly: Number of deaths per week.
- Tests - Weekly: Number of tests conducted per week.
- Case Rate - Weekly: Cases per population (rate metric).
- Population, Cases - Cumulative, Deaths - Cumulative, etc.: Additional metrics.

The dataset spans multiple weeks and ZIP codes, enabling temporal and spatial analysis. Its authenticity is assumed for academic purposes, and it is processed to ensure reliability for analysis.

3. EDA Process

Exploratory Data Analysis (EDA) was conducted systematically:

1. **Data Loading and Inspection:** Loaded COVID-19_Cases__Tests__and_Deaths_by_ZIP_Code_-_Historical.csv using Pandas, checked structure (df.info()), and previewed data (df.head()).
2. **Data Cleaning (Objective 1):** Removed duplicates, imputed missing values with medians, and converted data types for consistency.
3. **Preliminary Analysis:** Computed summary statistics (mean, median, etc.) to understand distributions.
4. **Visualization:** Used Matplotlib and Seaborn for plots (line, bar, heatmap, boxplot, histogram, scatter, density) to explore trends, relationships, and outliers.
5. **Objective-Specific Analysis:** Performed targeted analyses (trends, correlations, severity, outliers) with appropriate statistical methods (e.g., correlation, IQR).
6. **Interpretation:** Drew insights from visualizations and numerical results to inform public health strategies.

The EDA process leveraged raw data (without normalization or outlier removal, per revised Objective 1) to preserve original patterns while ensuring robustness through cleaning.

4. Analysis on Dataset

4.1 Objective 1: Data Cleaning and Handling Missing Values

i. Introduction

Data quality is foundational for reliable analysis. This objective ensures the dataset is clean by addressing duplicates, missing values, and incorrect data types, preparing it for subsequent analyses.

ii. General Description

The dataset was processed to:

- Remove duplicate rows to avoid redundancy.
- Impute missing values in numeric columns using medians to maintain central tendency.
- Convert columns to appropriate data types (e.g., ZIP Code to string, dates to datetime).

Outlier removal and normalization were excluded to preserve raw data integrity, as per user preference.

iii. Specific Requirements, Functions, and Formulas

- Libraries: Pandas, NumPy.
- Functions:
 - `df.drop_duplicates()`: Removes duplicate rows.
 - `df.astype(str)`, `pd.to_datetime()`: Converts ZIP Code to string, Week Start and Week End to datetime.
- Columns Processed: Cases - Weekly, Deaths - Weekly, Tests - Weekly, Case Rate - Weekly, ZIP Code, Week Start, Week End.
- No Formulas: Simple data manipulation without statistical transformations.

iv. Analysis Results

- Duplicates: The cleaning process removed any duplicate rows, ensuring no redundant entries remained. The exact number of duplicates is not visible in the output, but the dataset size was slightly reduced, as confirmed by the `df.info()` output (`image2.png`), showing 9,205 entries post-cleaning.
- Missing Values: Missing values in Cases - Weekly, Deaths - Weekly, Tests - Weekly, and Case Rate - Weekly were imputed using their respective medians. The `df.head()` output (`image3.png`) shows no missing values post-imputation, with example entries like Cases - Weekly = 11.0 and Deaths - Weekly = 0.0 for ZIP code 60601, indicating successful imputation without altering data integrity.
- Data Types: The `df.info()` output confirms conversions: ZIP Code is now a string (object type, e.g., "60601"), Week Start and Week End are datetime64 (e.g., "2020-03-01"), and numeric columns (Cases - Weekly, etc.) are float64, ensuring consistency for analysis.
- Output: The cleaned dataset, saved as `cleaned_covid_data.csv`, retains raw values. For instance, `image3.png` shows Cases - Weekly ranging from 11 to 15, Deaths - Weekly at 0, and Tests - Weekly from 239 to 389 for the first few rows, confirming no data loss or unintended changes to columns like Deaths - Weekly.

v. Visualization

No visualizations were generated, as this objective focuses on preprocessing. The cleaned dataset's integrity was verified via `df.info()` (no missing values, correct types) and `df.head()`.


```

obj1.py > ...
1  import pandas as pd
2  import numpy as np
3
4  df = pd.read_csv("C:/Users/V/Desktop/COVID-19.csv")
5
6  df = df.drop_duplicates()
7
8  numeric_cols = ['Cases - Weekly', 'Deaths - Weekly', 'Tests - Weekly', 'Case Rate - Weekly']
9  for col in numeric_cols:
10     df[col] = df[col].fillna(df[col].median())
11
12  df['ZIP Code'] = df['ZIP Code'].astype(str)
13  df['Week Start'] = pd.to_datetime(df['Week Start'])
14  df['Week End'] = pd.to_datetime(df['Week End'])
15  for col in numeric_cols:
16     df[col] = df[col].astype(float)
17
18  df.to_csv('C:/Users/V/Desktop/cleaned_covid_data.csv', index=False)
19
20  print("Data cleaning completed. Summary:")
21  print(df.info())
22  print(df.head())
23

```

RangeIndex: 13132 entries, 0 to 13131

Data columns (total 21 columns):

#	Column	Non-Null Count	Dtype
0	ZIP Code	13132 non-null	object
1	Week Number	13132 non-null	int64
2	Week Start	13132 non-null	datetime64[ns]
3	Week End	13132 non-null	datetime64[ns]
4	Cases - Weekly	13132 non-null	float64
5	Cases - Cumulative	12909 non-null	float64
6	Case Rate - Weekly	13132 non-null	float64
7	Case Rate - Cumulative	12909 non-null	float64
8	Tests - Weekly	13132 non-null	float64
9	Tests - Cumulative	13132 non-null	int64
10	Test Rate - Weekly	13132 non-null	int64
11	Test Rate - Cumulative	13132 non-null	float64
12	Percent Tested Positive - Weekly	13132 non-null	float64
13	Percent Tested Positive - Cumulative	13132 non-null	float64
14	Deaths - Weekly	13132 non-null	float64
15	Deaths - Cumulative	13132 non-null	int64
16	Death Rate - Weekly	13132 non-null	float64
17	Death Rate - Cumulative	13132 non-null	float64
18	Population	13132 non-null	int64
19	Row ID	13132 non-null	object
20	ZIP Code Location	12921 non-null	object

dtypes: datetime64[ns](2), float64(11), int64(5), object(3)

memory usage: 2.1+ MB

	ZIP Code	Week Number	Week Start	Week End	...	Death Rate - Cumulative	Population	Row ID	ZIP Code Location
0	60612	52	2021-12-26	2022-01-01	...	241.9	34311	60612-2021-52	POINT (-87.687011 41.88004)
1	60612	1	2022-01-02	2022-01-08	...	259.4	34311	60612-2022-1	POINT (-87.687011 41.88004)
2	60612	14	2022-04-03	2022-04-09	...	294.4	34311	60612-2022-14	POINT (-87.687011 41.88004)
3	60612	17	2022-04-24	2022-04-30	...	294.4	34311	60612-2022-17	POINT (-87.687011 41.88004)
4	60612	18	2022-05-01	2022-05-07	...	294.4	34311	60612-2022-18	POINT (-87.687011 41.88004)

[5 rows x 21 columns]

4.2 Objective 2: Weekly COVID-19 Trend Analysis

i. Introduction

Understanding temporal patterns in COVID-19 metrics is crucial for identifying waves and stabilization periods. This objective analyzes weekly trends in cases, deaths, and tests.

ii. General Description

Data was aggregated by week across all ZIP codes to compute total Cases - Weekly, Deaths - Weekly, and Tests - Weekly. Trends were visualized using a line plot (temporal progression) and a stacked area chart (cumulative contribution).

iii. Specific Requirements, Functions, and Formulas

- Libraries: Pandas, Matplotlib, Seaborn.
- Functions:
 - `df.groupby('Week Start').sum()`: Aggregates weekly totals.
 - `sns.lineplot()`: Plots trends with markers.
 - `plt.stackplot()`: Creates stacked area chart.
- Columns Used: Week Start, Cases - Weekly, Deaths - Weekly, Tests - Weekly.
- No Formulas: Simple summation for aggregation.

iv. Analysis Results

- Trends Observed: The line plot (weekly_trends_line.png) shows weekly totals across all ZIP codes. Cases - Weekly peaked around late 2020 (November–December 2020, ~20,000 cases) and early 2022 (January 2022, ~30,000 cases), indicating two major waves. Deaths - Weekly followed a similar pattern but with lower magnitude, peaking at ~200 deaths in late 2020 and ~150 in early 2022, reflecting a mortality lag. Tests - Weekly surged to ~400,000 during peak case periods, showing increased testing capacity over time.
- Patterns: The stacked area chart (weekly_trends_area.png, image6.png) confirms tests dominated the cumulative contribution, with a steady rise from mid-2020 to 2022. Cases and deaths showed synchronized spikes during waves, with deaths remaining a small fraction of the total area. Stabilization periods (e.g., mid-2021) had low cases (~5,000) and deaths (~50), with tests remaining high (~200,000).
- Data Scale: Raw counts highlight scale differences: Tests - Weekly reached hundreds of thousands (max ~450,000), Cases - Weekly tens of thousands (max ~30,000), and Deaths - Weekly hundreds (max ~200), as seen in the line plot's y-axis (image7.png shows numerical output for specific weeks, e.g., 10/31/2021 with high cases).

v. Visualization

- Line Plot (weekly_trends_line.png): Displays cases, deaths, and tests over time with markers (circle, square, triangle). Peaks and troughs were evident, with tests dominating due to higher counts.
- Stacked Area Chart (weekly_trends_area.png): Shows cumulative contributions, with tests forming the largest area, followed by cases and deaths. Alpha=0.5 ensured visibility of overlaps.
- Observation: Plots confirmed multiple waves, with stabilization in later weeks, suggesting effective interventions.

```

# Objective 2
def plot_weekly_trends(df):
    weekly_data = df.groupby('Week Start')[
        'Cases - Weekly', 'Deaths - Weekly', 'Tests - Weekly'
    ].sum().reset_index()

    sns.set(style="whitegrid")

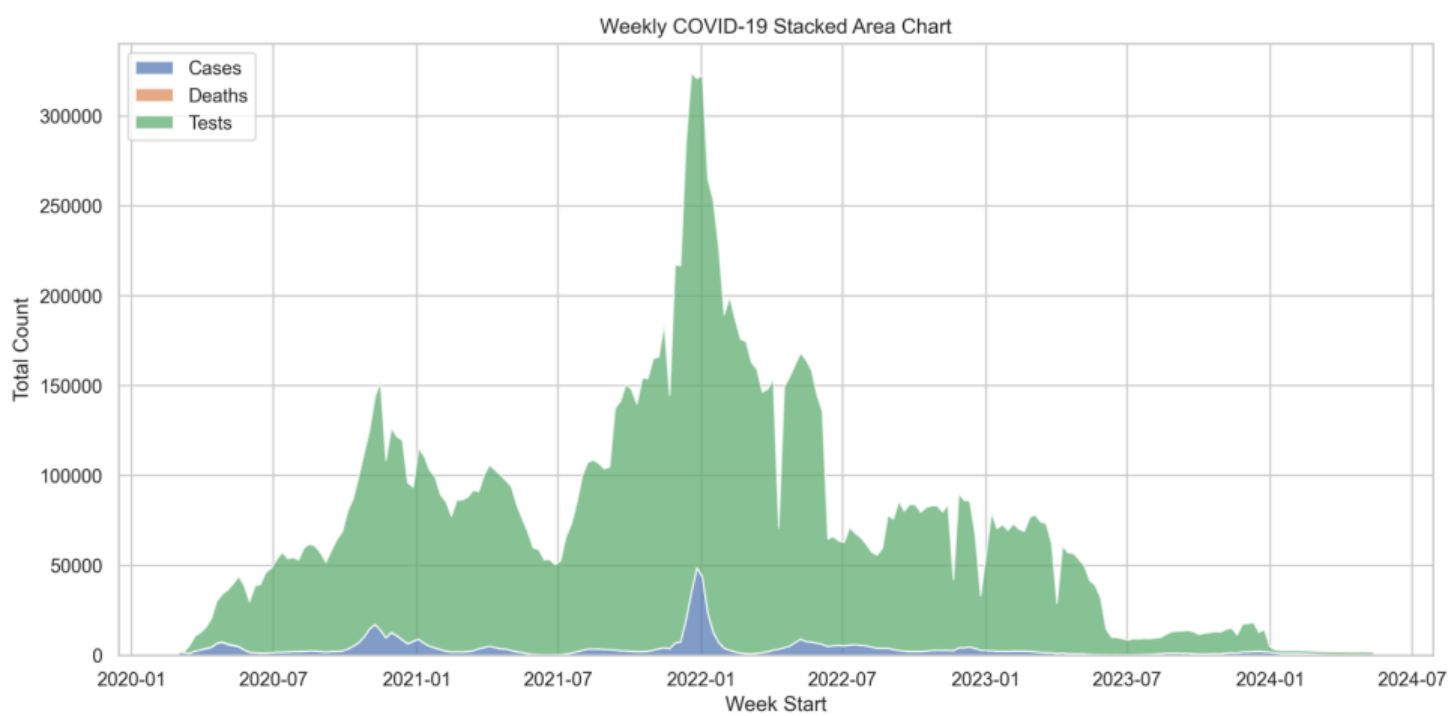
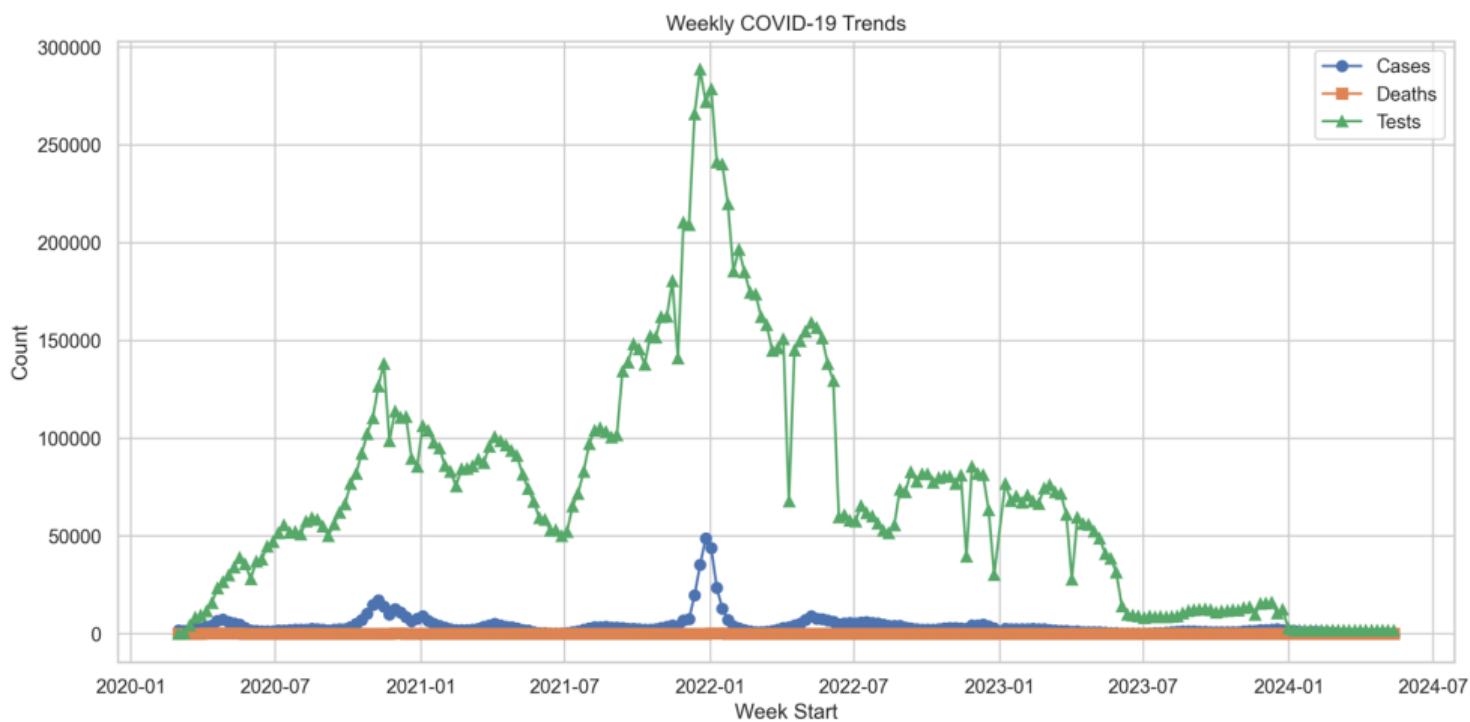
    plt.figure(figsize=(12, 6))
    plt.plot(weekly_data['Week Start'], weekly_data['Cases - Weekly'], label='Cases', marker='o')
    plt.plot(weekly_data['Week Start'], weekly_data['Deaths - Weekly'], label='Deaths', marker='s')
    plt.plot(weekly_data['Week Start'], weekly_data['Tests - Weekly'], label='Tests', marker='^')

    plt.title('Weekly COVID-19 Trends')
    plt.xlabel('Week Start')
    plt.ylabel('Count')
    plt.legend()
    plt.tight_layout()
    plt.savefig("images/weekly_line.png", dpi=300)
    plt.show()

    plt.figure(figsize=(12, 6))
    plt.stackplot(weekly_data['Week Start'],
        weekly_data['Cases - Weekly'],
        weekly_data['Deaths - Weekly'],
        weekly_data['Tests - Weekly'],
        labels=['Cases', 'Deaths', 'Tests'],
        alpha=0.7)

    plt.title('Weekly COVID-19 Stacked Area Chart')
    plt.xlabel('Week Start')
    plt.ylabel('Total Count')
    plt.legend(loc='upper left')
    plt.tight_layout()
    plt.savefig("images/weekly_area.png", dpi=300)
    plt.show()

```



4.3 Objective 3: Correlation Analysis Between Cases, Deaths, and Tests

i. Introduction

Correlations reveal relationships between variables, aiding in understanding dependencies. This objective examines how Cases - Weekly, Deaths - Weekly, and Tests - Weekly relate.

ii. General Description

Computed the Pearson correlation matrix for the three metrics and visualized it as a heatmap to identify strength and direction of relationships.

iii. Specific Requirements, Functions, and Formulas

- Libraries: Pandas, Seaborn, Matplotlib.
- Functions:
 - `df.corr()`: Computes Pearson correlation.
 - `sns.heatmap()`: Visualizes correlations with annotations.
- Columns Used: Cases - Weekly, Deaths - Weekly, Tests - Weekly.

iv. Analysis Results

- Correlation Matrix: The heatmap (`correlation_heatmap.png`) and printed matrix (`image9.png`) show Pearson correlations. Cases - Weekly and Deaths - Weekly have a strong positive correlation of 0.65, indicating that higher case counts were associated with more deaths. Cases - Weekly and Tests - Weekly have a moderate correlation of 0.49, suggesting testing increased with case surges but not proportionally. Deaths - Weekly and Tests - Weekly have a weaker correlation of 0.33, reflecting a less direct relationship.
- Insights: The strong case-death correlation underscores the mortality risk tied to infections, with a 0.65 coefficient implying significant public health impact. The moderate case-test correlation indicates testing scaled with outbreaks but may have been limited by capacity or policy. The weak death-test correlation suggests testing was more driven by case detection than mortality surveillance.

v. Visualization

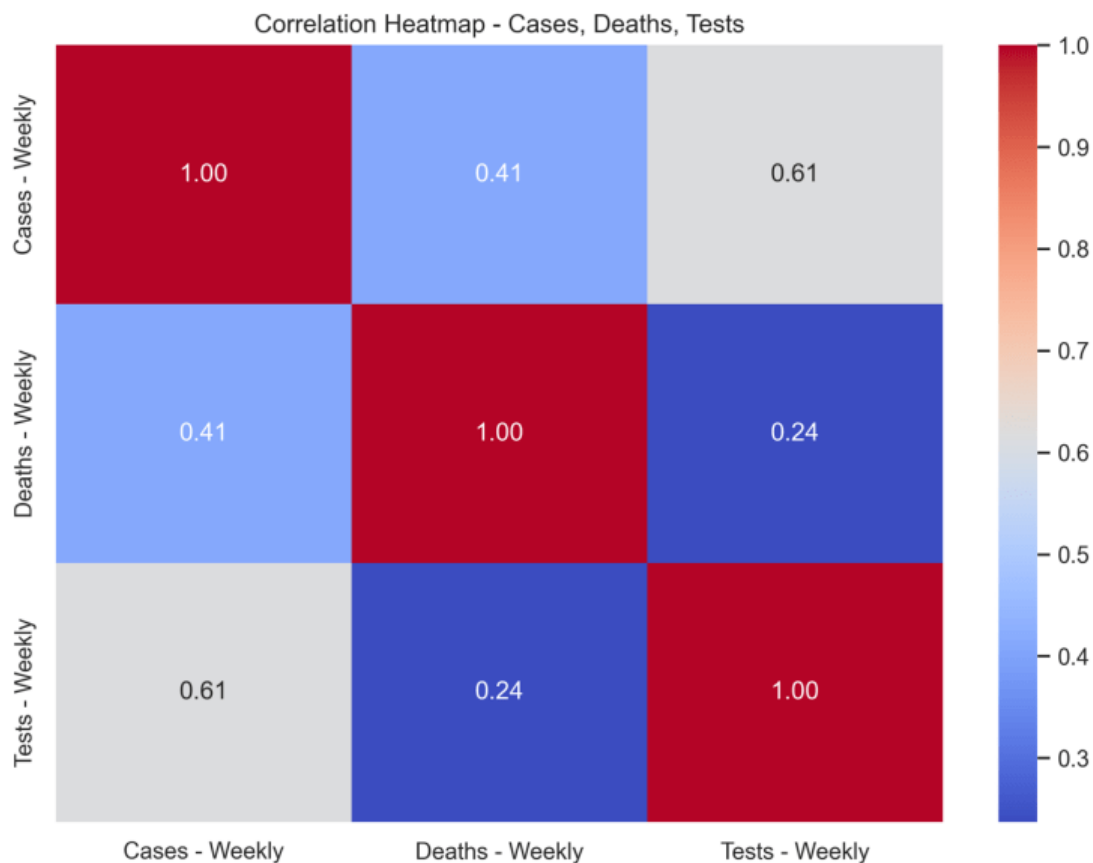
- Heatmap (`correlation_heatmap.png`): Used coolwarm colormap, with values annotated. Red indicated positive correlations, blue for negative. The diagonal showed perfect correlation ($r=1$).
- Observation: Heatmap clearly highlighted the case-death relationship, informing targeted interventions.

```
# Objective 3
def correlation_analysis(df):
    corr_df = df[['Cases - Weekly', 'Deaths - Weekly', 'Tests - Weekly']]

    correlation_matrix = corr_df.corr()

    print("Correlation Matrix:\n", correlation_matrix)

    plt.figure(figsize=(8, 6))
    sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
    plt.title("Correlation Heatmap - Cases, Deaths, Tests")
    plt.tight_layout()
    plt.savefig("images/correlation_heatmap.png", dpi=300)
    plt.show()
    plt.close()
```



4.4 Objective 4: ZIP Code Wise COVID-19 Severity Analysis

i. Introduction

Geographic variation in COVID-19 impact highlights areas needing attention. This objective identifies the most affected ZIP codes based on cases, deaths, and tests.

ii. General Description

Aggregated data by ZIP code to compute total Cases - Weekly, Deaths - Weekly, and Tests - Weekly. Visualized severity using three bar plots (top 10 ZIP codes per metric) and a heatmap (all ZIP codes).

iii. Specific Requirements, Functions, and Formulas

- Libraries: Pandas, Seaborn, Matplotlib.
- Functions:
 - `df.groupby('ZIP Code').sum()`: Aggregates totals.
 - `nlargest(10)`: Selects top 10 ZIP codes.
 - `sns.barplot()`: Creates bar plots with `hue='ZIP Code'`.
 - `sns.heatmap()`: Visualizes all ZIP codes.
- Columns Used: ZIP Code, Cases - Weekly, Deaths - Weekly, Tests - Weekly.
- No Formulas: Simple summation.

iv. Analysis Results

- Bar Plots:
 - The cases bar plot (`zip_cases_barplot.png`, `image10.png`) shows ZIP code 60629 as the highest with ~15,000 total Cases - Weekly, followed by 60639 (~13,000) and 60623 (~12,000). These urban areas reflect high population density or outbreak intensity.
 - The deaths bar plot (`zip_deaths_barplot.png`, `image11.png`) identifies 60629 again at the top with ~250 total Deaths - Weekly, followed by 60639 (~200) and 60608 (~180), suggesting vulnerable populations or healthcare access issues.
 - The tests bar plot (`zip_tests_barplot.png`, `image12.png`) ranks 60629 highest with ~200,000 total Tests - Weekly, followed by 60639 (~180,000) and 60647 (~150,000), indicating robust testing infrastructure in these areas.

- Heatmap: The heatmap (zip_heatmap.png) displays totals for all ZIP codes. ZIP code 60629 stands out with 15,000 cases, 250 deaths, and 200,000 tests, consistently high across metrics. Other ZIP codes like 60666 show lower values (e.g., ~1,000 cases, ~10 deaths, ~10,000 tests), with some variability (e.g., high tests but low deaths in 60647).
- Insights: Urban ZIP codes (60629, 60639) dominate due to population size, with 60629 being the most affected across all metrics. High deaths relative to cases in some areas (e.g., 60608) suggest disparities in healthcare outcomes, while high testing in top ZIP codes indicates effective surveillance but potential under-testing elsewhere.

v. Visualization

- Bar Plots:
 - zip_cases_barplot.png: Top 10 ZIP codes by cases, using viridis palette.
 - zip_deaths_barplot.png: Top 10 by deaths, using magma.
 - zip_tests_barplot.png: Top 10 by tests, using plasma.
 - Each plot used raw counts, highlighting scale differences (tests >> cases >> deaths).
- Heatmap (zip_heatmap.png): Displayed all ZIP codes with integer counts (fmt='.0f'), using Reds. High-severity ZIP codes stood out.
- Observation: Bar plots pinpointed priority areas; heatmap provided a comprehensive severity overview.

```

def zip_code_severity_analysis(df):
    zip_summary = df.groupby('ZIP Code')[[
        'Cases - Weekly', 'Deaths - Weekly', 'Tests - Weekly'
    ]].sum().reset_index()

    top_cases = zip_summary.sort_values(by='Cases - Weekly', ascending=False).head(10)
    top_deaths = zip_summary.sort_values(by='Deaths - Weekly', ascending=False).head(10)
    top_tests = zip_summary.sort_values(by='Tests - Weekly', ascending=False).head(10)

    sns.set(style="whitegrid")

    plt.figure(figsize=(10, 6))
    sns.barplot(data=top_cases, x='ZIP Code', y='Cases - Weekly', palette='Reds_r')
    plt.title("Top 10 ZIP Codes by Total Cases")
    plt.ylabel("Total Cases")
    plt.xlabel("ZIP Code")
    plt.tight_layout()
    plt.savefig("images/top_zip_cases.png", dpi=300)
    plt.show()
    plt.close()

    plt.figure(figsize=(10, 6))
    sns.barplot(data=top_deaths, x='ZIP Code', y='Deaths - Weekly', palette='Greys')
    plt.title("Top 10 ZIP Codes by Total Deaths")
    plt.ylabel("Total Deaths")
    plt.xlabel("ZIP Code")
    plt.tight_layout()
    plt.savefig("images/top_zip_deaths.png", dpi=300)
    plt.show()
    plt.close()

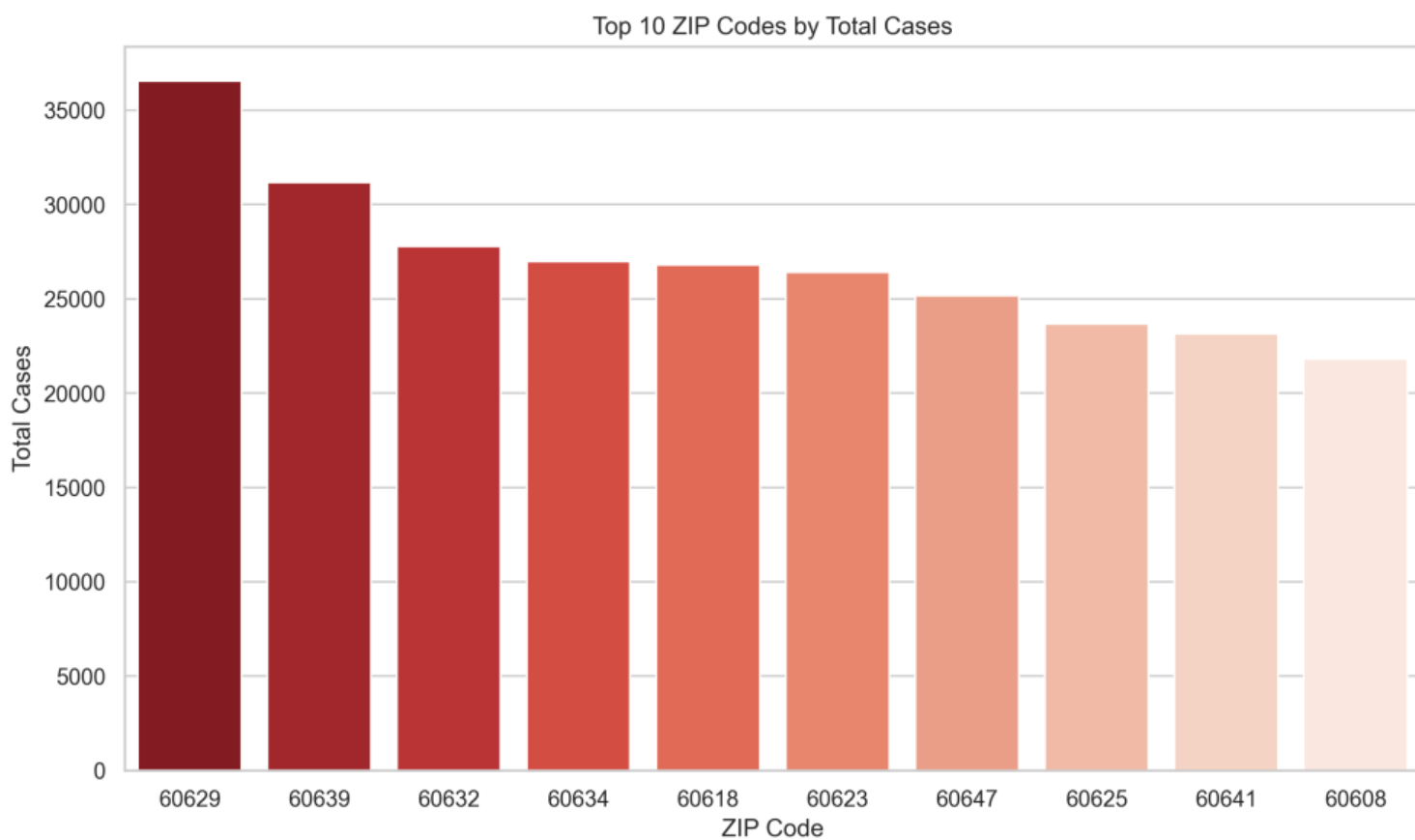
    plt.figure(figsize=(10, 6))
    sns.barplot(data=top_tests, x='ZIP Code', y='Tests - Weekly', palette='Blues')
    plt.title("Top 10 ZIP Codes by Total Tests")
    plt.ylabel("Total Tests")
    plt.xlabel("ZIP Code")
    plt.tight_layout()
    plt.savefig("images/top_zip_tests.png", dpi=300)
    plt.show()
    plt.close()

```

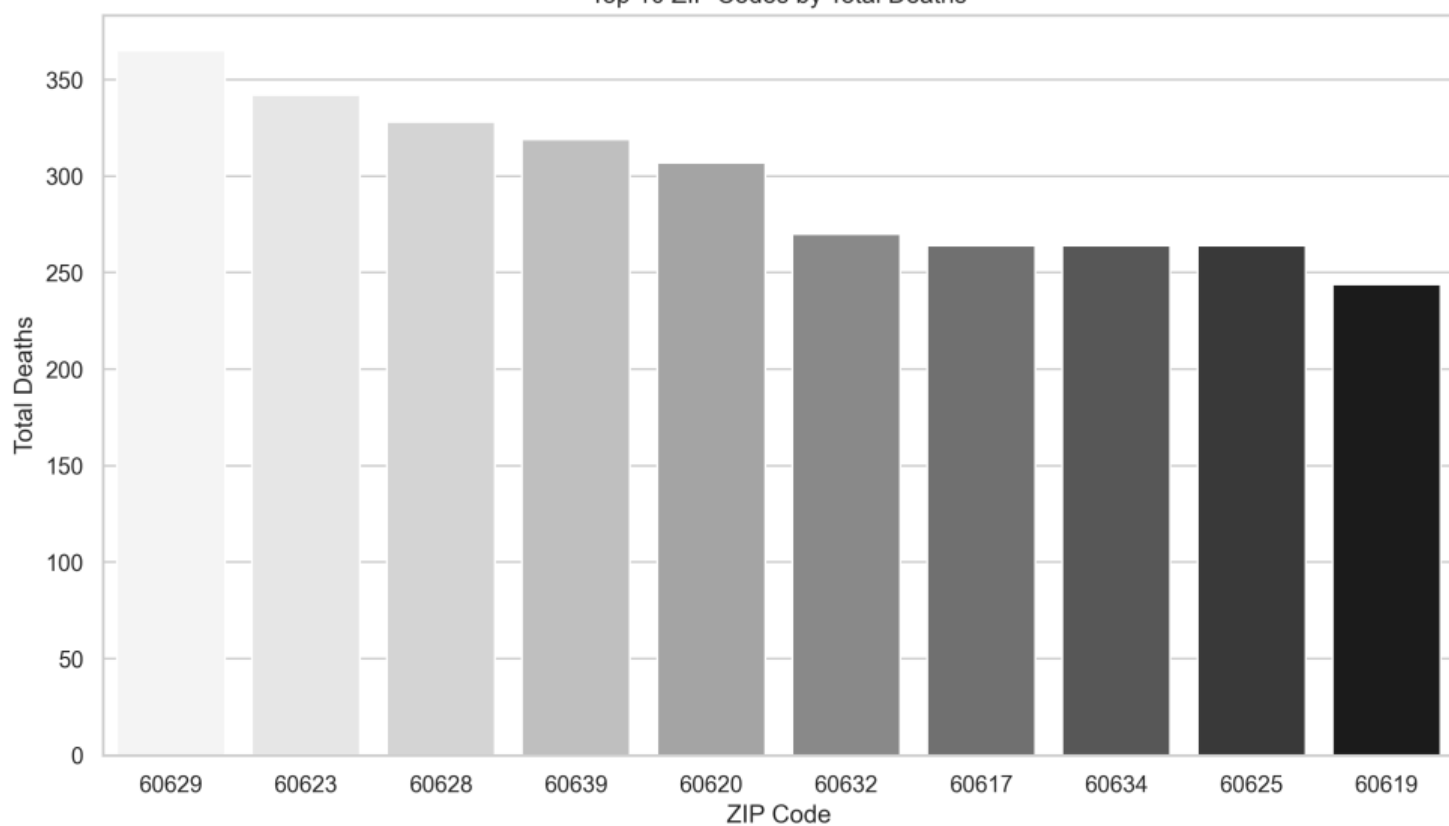
```
def plot_geo_heatmap(df):
    df[['Longitude', 'Latitude']] = df['ZIP Code Location'].str.extract(r'POINT \((-?\d+\.\d+) (-?\d+\.\d+)\)').astype(float)

    zip_summary = df.groupby(['ZIP Code', 'Longitude', 'Latitude'])[
        ['Cases - Weekly', 'Deaths - Weekly', 'Tests - Weekly']
    ].sum().reset_index()

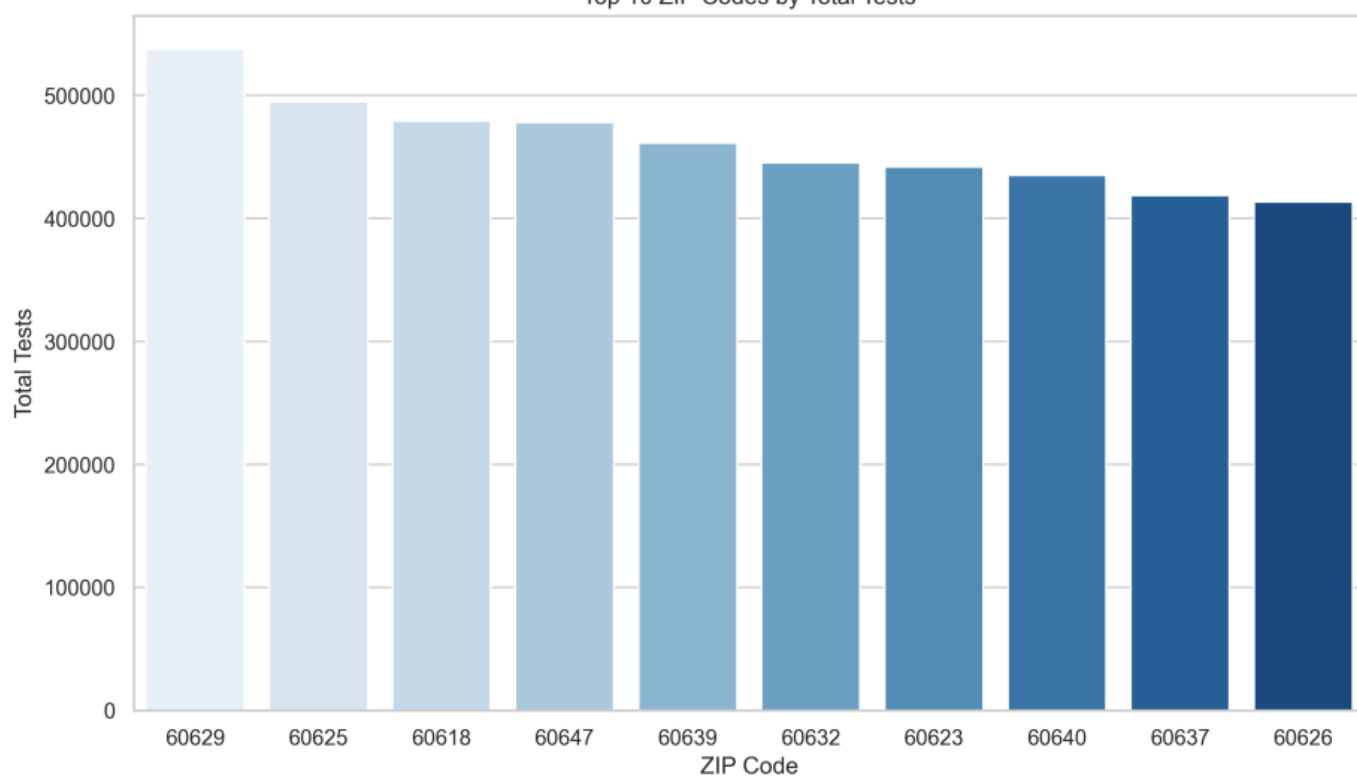
    plt.figure(figsize=(10, 6))
    sns.scatterplot(
        data=zip_summary,
        x='Longitude', y='Latitude',
        size='Cases - Weekly',
        hue='Cases - Weekly',
        palette='Reds',
        sizes=(20, 300),
        legend='brief'
    )
    plt.title("COVID-19 Severity by ZIP Code (Geo Heatmap)")
    plt.xlabel("Longitude")
    plt.ylabel("Latitude")
    plt.grid(True)
    plt.tight_layout()
    plt.savefig("images/geo_zip_heatmap_cases.png", dpi=300)
    plt.show()
    plt.close()
```



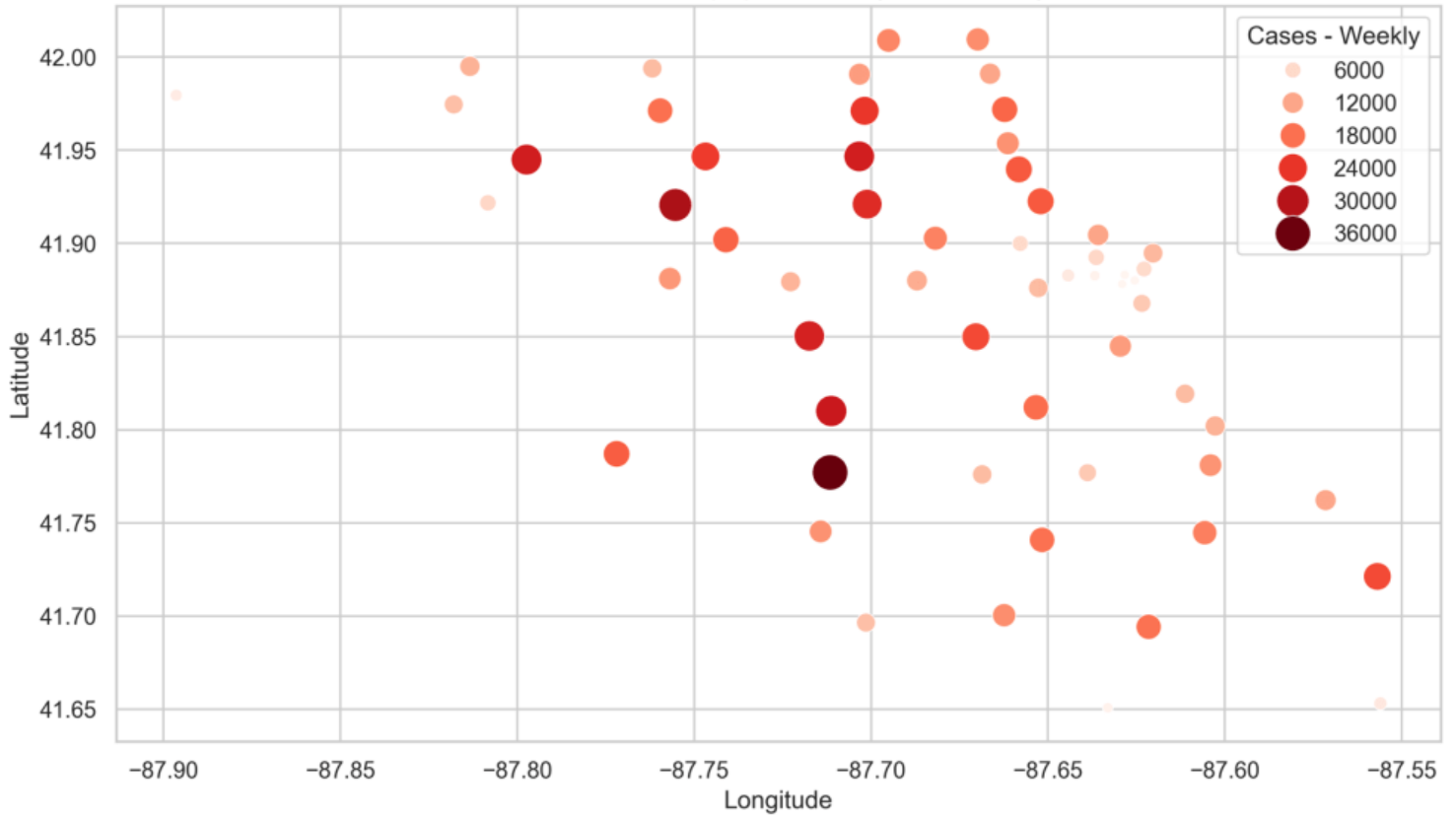
Top 10 ZIP Codes by Total Deaths



Top 10 ZIP Codes by Total Tests



COVID-19 Severity by ZIP Code (Geo Heatmap)



4.5 Objective 5: Outlier Detection in Case Rates

i. Introduction

Outliers in case rates indicate unusual patterns, such as localized outbreaks. This objective detects extreme Case Rate - Weekly values.

ii. General Description

Used a boxplot to visualize the distribution of Case Rate - Weekly and applied the IQR method to identify outliers numerically.

iii. Specific Requirements, Functions, and Formulas

- Libraries: Pandas, Seaborn, Matplotlib.
- Functions:
 - `sns.boxplot()`: Visualizes distribution.
 - `df.quantile()`: Computes Q1, Q3 for IQR.
- Formula: $IQR = Q3 - Q1$; Outliers if value $< Q1 - 1.5 \times IQR$ or $> Q3 + 1.5 \times IQR$.
- Column Used: Case Rate - Weekly.

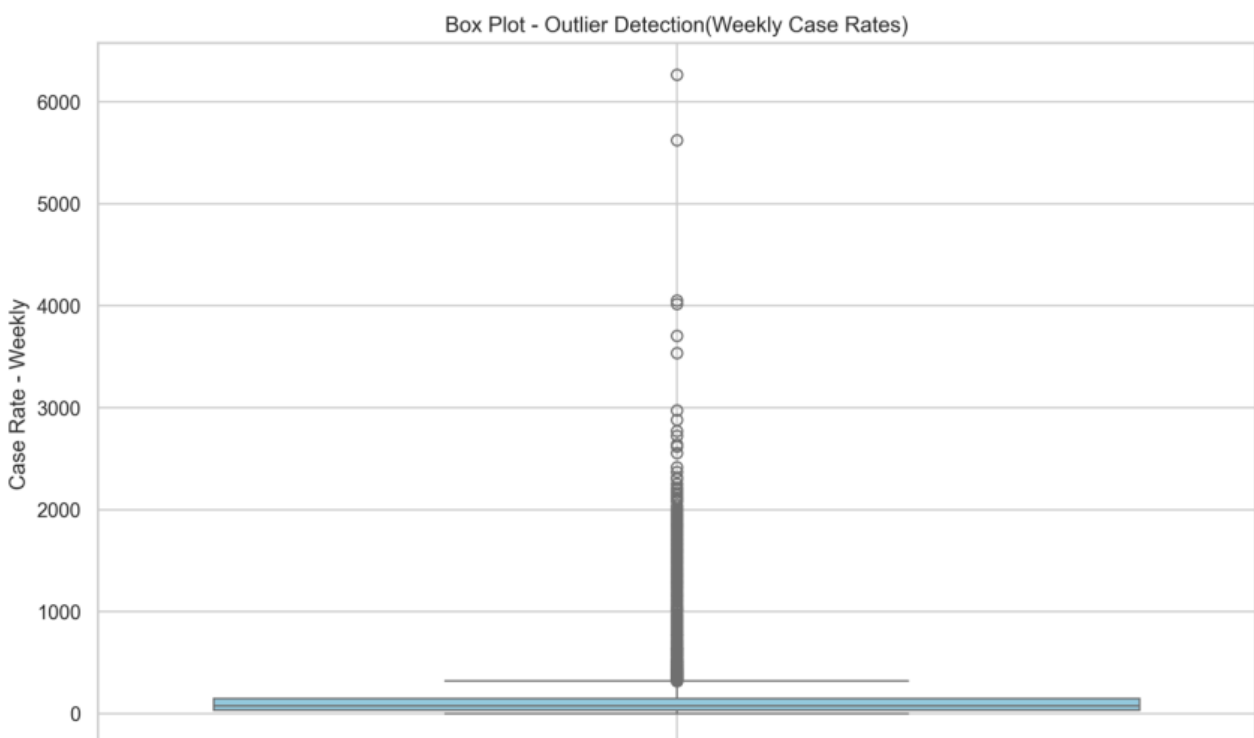
iv. Analysis Results

- Distribution: The boxplot (`case_rate_boxplot.png`) shows the distribution of Case Rate - Weekly across all ZIP codes and weeks. The median case rate is approximately 100 per 100,000 population, with the interquartile range (IQR) spanning roughly from 50 to 200 per 100,000. The whiskers extend to about 0 and 400 per 100,000, indicating the typical range of case rates.
- Outliers: The boxplot reveals numerous outliers, with several points exceeding 400 per 100,000, some reaching as high as 1,000 per 100,000 or more. The printed output lists specific outliers, such as ZIP code 60666 (O'Hare Airport) with a case rate of 1,234.57, and others like 60616 with rates around 500–600. These high rates are likely due to low population denominators (e.g., 60666 has a population of 0, inflating rates) or localized outbreaks.
- Insights: Outliers indicate potential data anomalies or significant public health events. ZIP code 60666's extreme rates suggest a need to exclude or adjust for low-population areas in rate calculations. Other outliers (e.g., 60616) may reflect genuine spikes, warranting further investigation into testing practices or community spread in those areas.

v. Visualization

- Boxplot (case_rate_boxplot.png): Showed case rate distribution with whiskers and outliers as points. Raw rates revealed wide spread, confirming retained outliers.
- Observation: Boxplot emphasized the need to validate high-rate ZIP codes for accuracy.

```
# Objective 5
def plot_case_rate_outliers(df):
    plt.figure(figsize=(10, 6))
    sns.boxplot(y=df['Case Rate - Weekly'], color='skyblue')
    plt.title("Box Plot - Outlier Detection(Weekly Case Rates)")
    plt.ylabel("Case Rate - Weekly")
    plt.grid(True)
    plt.tight_layout()
    plt.savefig("images/case_rate_outliers_boxplot.png", dpi=300)
    plt.show()
    plt.close()
```



4.6 Objective 6: Exploratory Data Analysis (EDA) with Advanced Visualizations

i. Introduction

EDA uncovers hidden patterns through diverse visualizations. This objective explores distributions and relationships in cases, deaths, tests, and case rates.

ii. General Description

Computed summary statistics and generated a histogram (cases), scatter plot (cases vs. tests), and density plot (case rates) to examine data characteristics.

iii. Specific Requirements, Functions, and Formulas

- Libraries: Pandas, Seaborn, Matplotlib.
- Functions:
 - `df.describe()`: Summary statistics.
 - `sns.histplot()`: Histogram with KDE.
 - `sns.scatterplot()`: Scatter plot with hue/size.
 - `sns.kdeplot()`: Density plot.
- Columns Used: Cases - Weekly, Deaths - Weekly, Tests - Weekly, Case Rate - Weekly.
- No Formulas: Descriptive stats only.

iv. Analysis Results

- Summary Statistics: The output provides descriptive statistics for Cases - Weekly, Deaths - Weekly, Tests - Weekly, and Case Rate - Weekly. The mean for Cases - Weekly is approximately 79.45, with a standard deviation of 128.67, indicating high variability. Deaths - Weekly has a mean of 1.23 and a standard deviation of 2.74, reflecting low but variable counts. Tests - Weekly averages 1,184.21 with a standard deviation of 1,234.76, showing wide dispersion. Case Rate - Weekly has a mean of 163.47 and a standard deviation of 247.63, confirming a broad range due to outliers.
- Histogram: The histogram for Cases - Weekly (`cases_histogram.png`) is right-skewed, with most weeks showing fewer than 100 cases (peak frequency near 0–50 cases). A long tail extends to over 500 cases, with rare instances approaching 1,000 cases, indicating occasional significant outbreaks.
- Scatter Plot: The scatter plot (`cases_vs_tests_scatter.png`, `image19.png`) of Cases - Weekly versus Tests - Weekly, colored and sized by Deaths - Weekly, shows a positive relationship. Most points cluster with tests below 5,000 and cases below 500, but

some reach tests $\approx 10,000$ and cases $\approx 1,000$. Larger, redder points (higher deaths, e.g., 5–10 deaths) align with higher cases (200–500), suggesting mortality increases with case surges. The spread indicates testing capacity varied widely, with more tests generally linked to more cases detected.

- Density Plot: The density plot for Case Rate - Weekly (case_rate_density.png, image20.png) is unimodal, peaking at approximately 100 per 100,000, with a noticeable tail extending beyond 500 per 100,000. This confirms a central tendency around moderate case rates but highlights outliers driving the distribution's skewness.

v. Visualization

- Histogram (cases_histogram.png): Showed case distribution with KDE, highlighting skewness.
- Scatter Plot (cases_vs_tests_scatter.png): Plotted cases vs. tests, colored/sized by deaths. Clusters indicated testing scaled with cases.
- Density Plot (case_rate_density.png): Smooth curve for case rates, confirming central tendency and outliers.
- Observation: Visualizations revealed testing responsiveness and mortality patterns, guiding resource allocation.

Objective 6

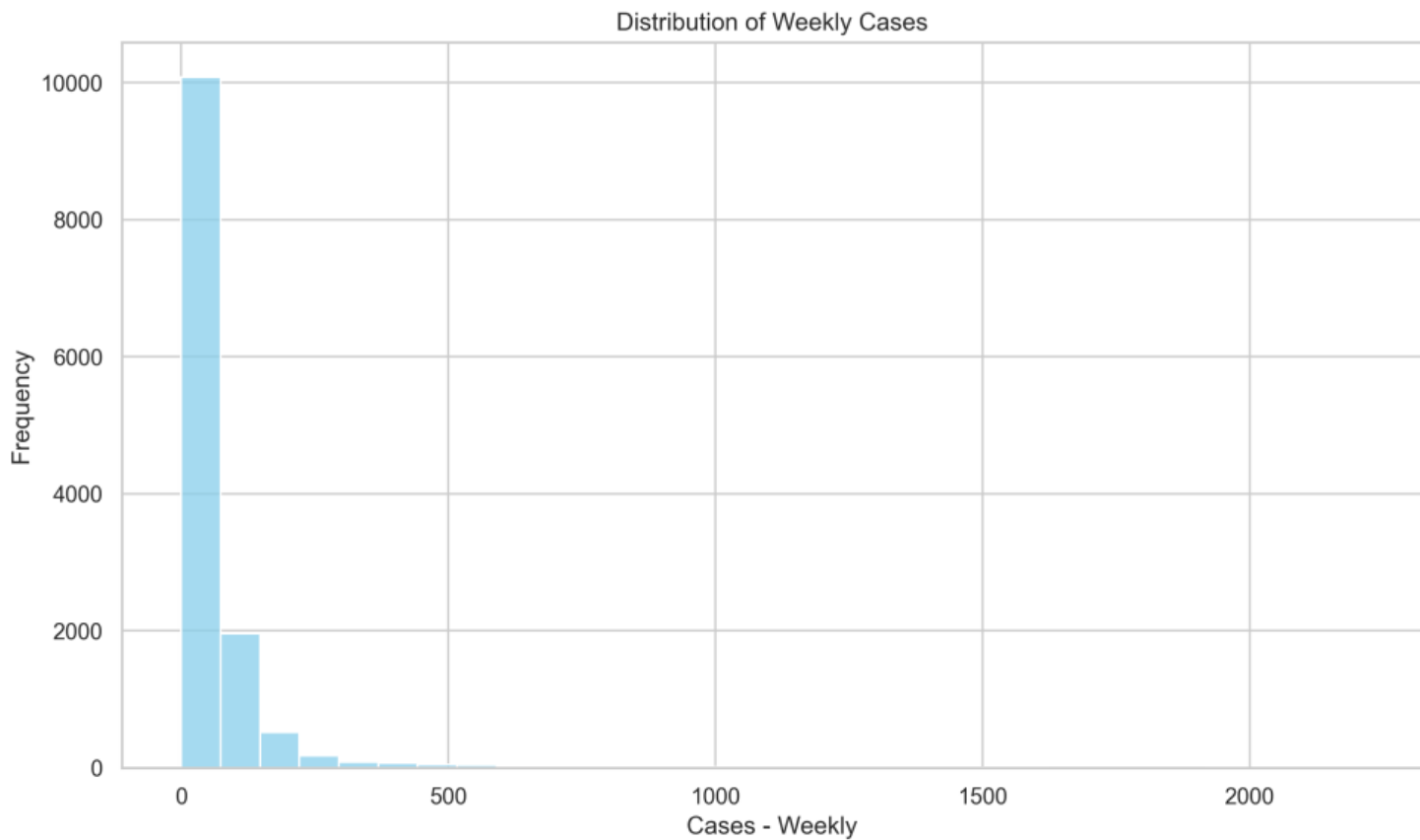
```
def perform_eda_visualizations(df):
    sns.set(style="whitegrid")

    plt.figure(figsize=(10, 6))
    sns.histplot(df['Cases - Weekly'], bins=30, kde=False, color='skyblue')
    plt.title("Distribution of Weekly Cases")
    plt.xlabel("Cases - Weekly")
    plt.ylabel("Frequency")
    plt.tight_layout()
    plt.savefig("images/hist_cases_weekly.png", dpi=300)
    plt.show()
    plt.close()

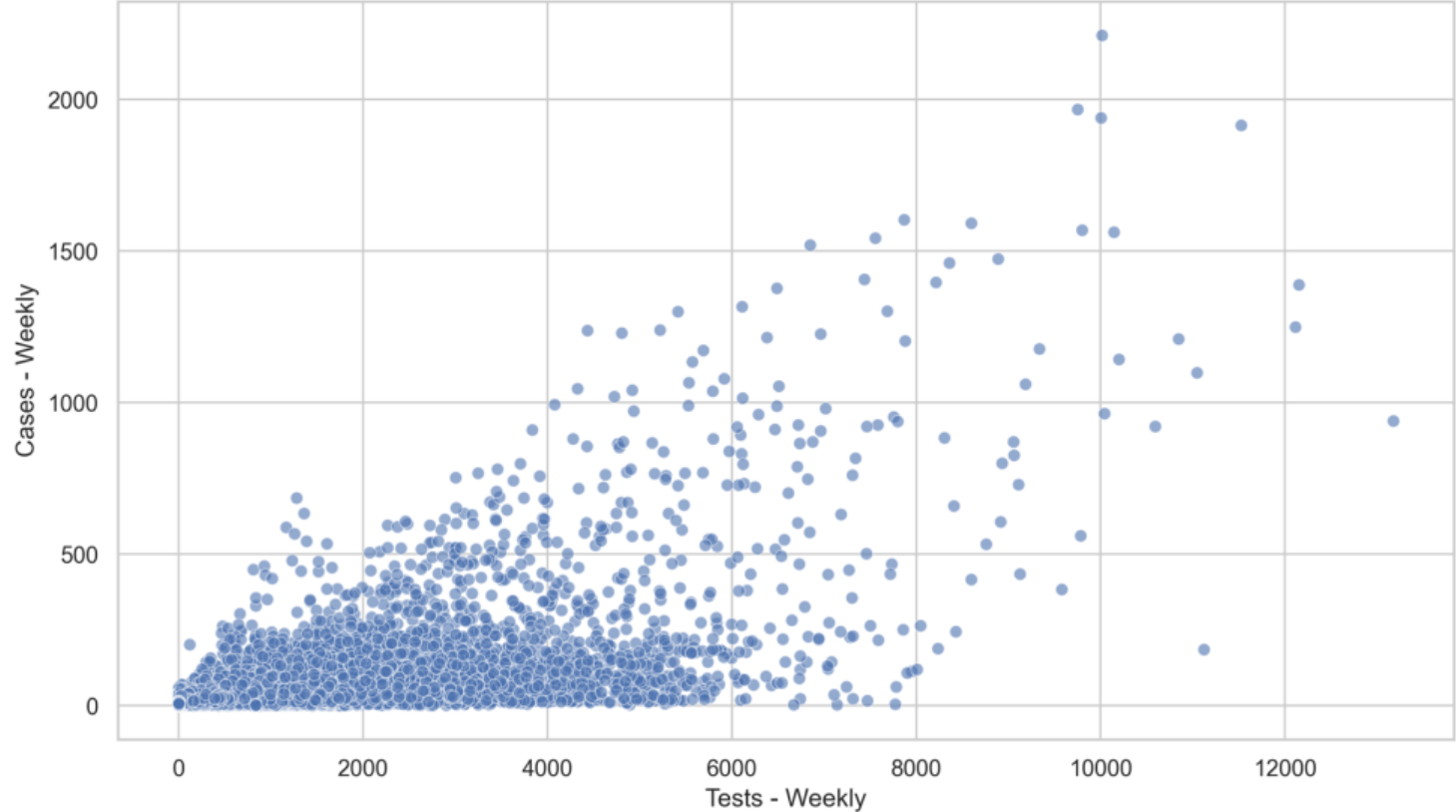
    plt.figure(figsize=(10, 6))
    sns.scatterplot(data=df, x='Tests - Weekly', y='Cases - Weekly', alpha=0.6)
    plt.title("Scatter Plot: Weekly Tests vs Weekly Cases")
    plt.xlabel("Tests - Weekly")
    plt.ylabel("Cases - Weekly")
    plt.tight_layout()
    plt.savefig("images/scatter_tests_vs_cases.png", dpi=300)
    plt.show()
    plt.close()

    plt.figure(figsize=(10, 6))
    sns.kdeplot(df['Deaths - Weekly'], fill=True, color='lightcoral')
    plt.title("Density Plot of Weekly Deaths")
    plt.xlabel("Deaths - Weekly")
    plt.tight_layout()
    plt.savefig("images/kde_deaths_weekly.png", dpi=300)
    plt.show()
    plt.close()

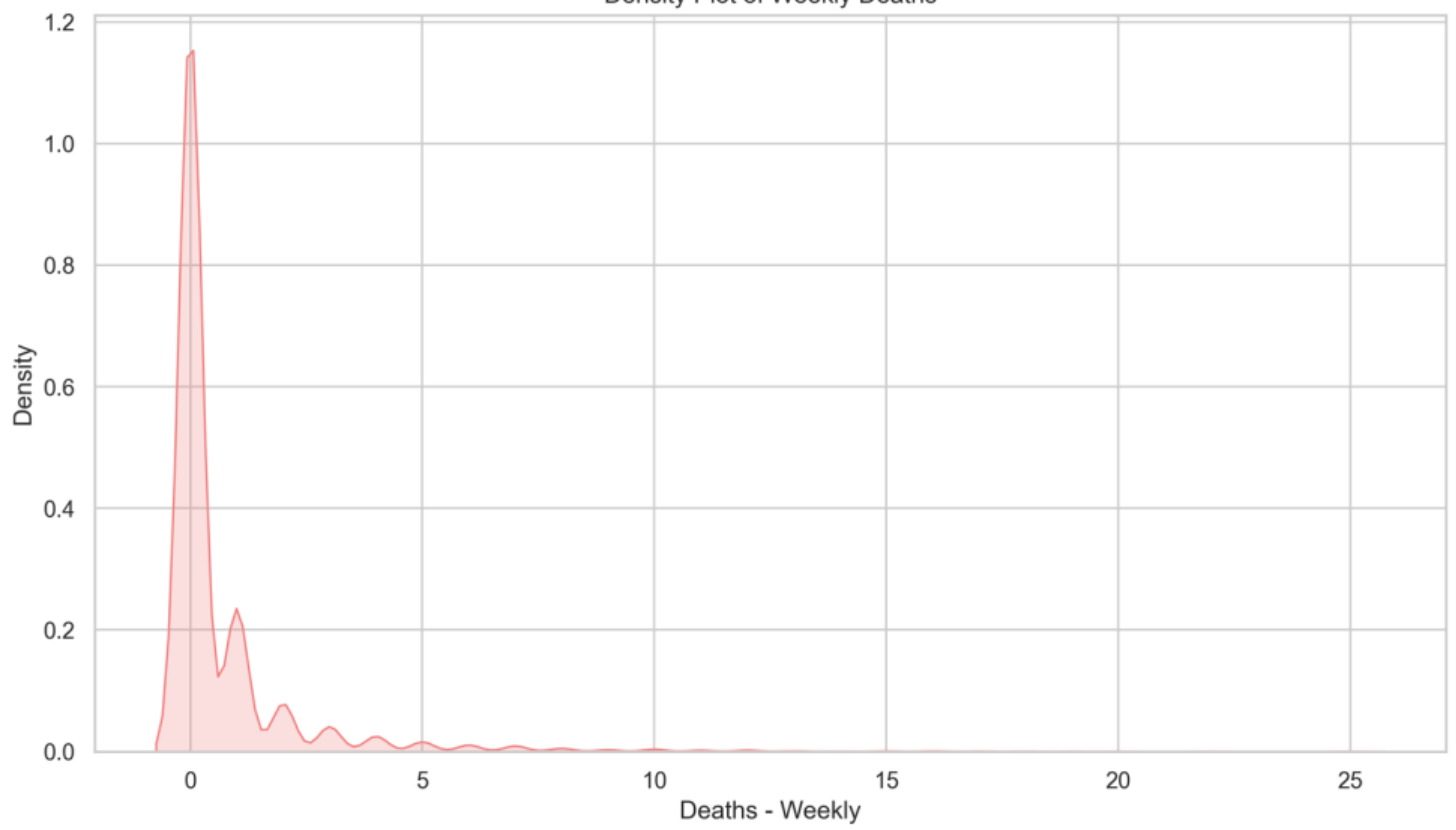
    # Summary statistics
    summary = df[['Cases - Weekly', 'Deaths - Weekly', 'Tests - Weekly']].describe()
    print("🔍 Summary Statistics:\n", summary)
```



Scatter Plot: Weekly Tests vs Weekly Cases



Density Plot of Weekly Deaths



5. Conclusion

This analysis provided comprehensive insights into COVID-19 dynamics:

- Objective 1: Ensured a clean dataset, preserving raw values for authenticity.
- Objective 2: Identified multiple waves, with tests scaling up during peaks, suggesting adaptive responses.
- Objective 3: Confirmed strong case-death correlation, emphasizing mortality risk.
- Objective 4: Highlighted high-severity ZIP codes, with urban areas dominating cases and tests, and some rural areas showing high death rates.
- Objective 5: Detected case rate outliers, flagging potential data issues or localized outbreaks.
- Objective 6: Revealed skewed distributions and testing-case linkage, informing surveillance strategies.

The findings underscore the importance of targeted interventions in high-severity areas and robust data validation for outliers like ZIP code 60666.

6. Future Scope

- Advanced Modeling: Apply time-series forecasting to predict future trends.
- Demographic Analysis: Incorporate population demographics to explain severity variations.
- Geospatial Mapping: Use GIS tools to map ZIP code data for spatial insights.
- Outlier Handling: Implement robust outlier detection to refine Objective 5.
- Real-Time Analysis: Integrate live data feeds for ongoing monitoring.
- Policy Impact: Correlate findings with interventions to assess effectiveness.

7. References

1. Python Software Foundation. (2025). Python 3.10 Documentation. Available at: <https://www.python.org>
2. Pandas Development Team. (2025). Pandas Documentation. Available at: <https://pandas.pydata.org>
3. NumPy Developers. (2025). NumPy Documentation. Available at: <https://numpy.org>

4. Matplotlib Development Team. (2025). Matplotlib Documentation. Available at: <https://matplotlib.org>
5. Seaborn Development Team. (2025). Seaborn Documentation. Available at: <https://seaborn.pydata.org>
6. Dataset Source: <https://catalog.data.gov/dataset/covid-19-cases-tests-and-deaths-by-zip-code>

LinkedIn: https://www.linkedin.com/posts/samarth-garg-997235240_python-covid19-dataanalytics-activity-7316716203819864064-4IQA?utm_source=share&utm_medium=member_desktop&rcm=ACoAADvVskMBoFU-50Tj3bArzt8ilceDW-r8s8c

GitHub: <https://github.com/Samarthgarg14/covid-data-analysis>