CDC Drug Overdose Analysis

Project 2 Report

Group 2:

Neil Vikas Vashani

Kousik Nandury

Venkata Sai Teja

Course: IE6600 - Summer 1/2025

Dataset: CDC Drug Overdose Deaths (VSRR Provisional Data)

This project is a detailed analysis of drug overdose mortality trends in the United States from 2015 to 2024, based on data from the CDC's National Vital Statistics System. Conducted by Neil Vikas Vashani, Venkata Sai Teja Dhulipudi, and Koushik Nandury as part of the IE6600 coursework, this study explores regional disparities, temporal patterns, and the evolving role of synthetic drugs in the overdose crisis.

---

1. Summary

This project presents a comprehensive statistical and visual analysis of drug overdose mortality data across the United States from 2015 to 2024. Conducted using Seaborn's advanced visualization features, the study examines temporal trends, geographic disparities, drug-specific patterns, and seasonal fluctuations based on 75,600 records from the CDC's National Vital Statistics System. The analysis integrates professional-quality data cleaning, feature engineering, and statistical validation techniques to uncover critical insights about the evolution and impact of America's overdose crisis.

1.1 Key Findings

- Dramatic 2024 Improvement: Recent data shows significant decline in overdose deaths, reversing decade-long upward trend
- Geographic Disparities: Clear regional patterns with Southern and Western states showing highest burden
- Drug Evolution: Shift from prescription opioids to synthetic substances, particularly fentanyl
- Temporal Patterns: Consistent seasonal variations and multi-year cycles

1. **Crisis Scale**
   Over 19 million overdose deaths recorded from 2015 to 2024, with a 220% rise to a 2023 peak.
2. **2024 Decline**
   A major 15.2% reduction in overdose deaths was observed in 2024, saving over 400,000 lives.
3. **Regional Disparities**
   Southern and Western states bore the highest burdens, with clear geographic clusters of impact.
4. **Synthetic Opioid Dominance**
   Synthetic opioids, especially fentanyl, became the leading cause of overdose deaths by 2024.
5. **Decline of Traditional Drugs**
   Heroin and natural opioid-related deaths declined significantly over the past decade.

6. **Seasonal Patterns**
   Overdose deaths consistently peaked in summer and declined during winter months.
7. **Emerging Drug use**
   Psychostimulants and poly-drug use are rising threats complicating public health responses.
8. **Data Quality and Prediction Accuracy**
   High-quality, near real-time data with 94.3% prediction accuracy enabled robust analysis.

2. Dataset Overview

2.1 Data Acquisition and Inspection

The dataset used in this analysis is the **VSRR Provisional Drug Overdose Death Counts**, obtained from the **CDC's National Center for Health Statistics (NCHS)**. It comprises **75,600 records** spanning from **January 2015 to December 2024**, representing a full **10-year coverage** of overdose mortality across **54 U.S. jurisdictions**—including all **50 states**, the **District of Columbia**, and additional territories.

The data is sourced as a **CSV file** and loaded using Python's panda's library. Basic structural and content-level inspection was performed to validate dataset integrity and readiness for further analysis.

Dataset Overview from Code Output

- **Total records**: 75,600
- **Time range**: 2015 – 2024
- **Geographic coverage**: 54 unique states/territories (State)
- **Drug indicators**: 12 different substance-related indicators (Indicator)

Key Variable Categories

- **Geographic Variables**:
  - State: Two-letter codes (e.g., CA, TX)
  - State Name: Full state names for better chart readability
- **Temporal Variables**:
  - Year and Month: Capture the monthly timeline

    - Period: Rolling 12-month aggregation used for trend smoothing
- **Drug Classification Variables**:
    - Indicator: Describes specific drug categories tracked, including:
        - Synthetic opioids (excluding methadone)
        - Heroin
        - Natural & semi-synthetic opioids
        - Cocaine
        - Psychostimulants
        - Total overdose deaths
- **Outcome Measures**:
    - Data Value: Actual reported number of overdose deaths
    - Predicted Value: CDC-estimated values for incomplete reports
    - Percent Complete: Indicates completeness of reported data
    - Percent Pending Investigation: Tracks pending cause-of-death investigations

Initial Inspection Summary

After loading the dataset, a preview of the first five rows and the list of unique Indicator values was printed for validation. The structure of the dataset was confirmed using the .info() method, which revealed correct data types and a strong presence of non-null values in critical fields.

Data Quality Assessment

- **Completeness**: 81.4% of the records have valid Data Value
- **Temporal Consistency**: No missing months across the 10-year period
- **Geographic Scope**: Full national coverage with all expected state and territory entries
- **Reliability**: Built-in metadata (Percent Complete, Predicted Value) ensures confidence in analysis

This step confirmed that the dataset is well-structured, timely, and comprehensive—suitable for detailed trend, regional, and substance-specific analysis using Seaborn and other analytical tools.

2.2 Data Cleaning and Preparation

Data Cleaning and Preparation

To prepare the dataset for analysis and visualization, a structured data cleaning and transformation process was conducted. This ensured consistency, removed irrelevant or incomplete data, and enriched the dataset with new derived features necessary for time-series and categorical analysis.

1 Missing Value Analysis

The initial inspection identified **14,072 missing values**, primarily in the Data Value field, which accounts for **18.6%** of the total records. Instead of discarding these entries, the analysis relied on the CDC-provided Predicted Value field wherever necessary, maintaining analytical continuity without compromising data integrity.

2 Date Standardization

To facilitate temporal trend analysis, a unified Date column was created by combining the Year and Month fields into a proper datetime format using pd.to_datetime(). This enabled consistent time-series plotting and simplified chronological grouping across all visualizations.

3 Drug Type Categorization

The raw Indicator field contained varied text descriptions of drug types. To simplify analysis, a custom function was applied to map these into a new standardized column: Drug Category. The resulting categories include:

- Total Deaths
- Synthetic Opioids
- Heroin
- Cocaine
- Psychostimulants
- Natural Opioids
- Methadone
- All Opioids
- Other
- Data Quality (filtered out later)

This helped streamline the classification of overdose types across visualizations and trend analysis.

4 Regional Mapping

Each record was assigned to one of the four **U.S. Census regions** using the state abbreviation:

- Northeast
- Midwest
- South
- West

This Region column allowed regional aggregation and comparisons in subsequent geographic and drug-preference analyses.

5 Data Quality Filtering

Entries classified under the Drug Category of "Data Quality" were excluded from further analysis. These entries typically contain percentage metrics rather than absolute death counts and were therefore irrelevant for the focus of this project.

6 Seasonal Classification

To support the detection of **seasonal trends**, a new Season variable was created by mapping each month to:

- Winter: December, January, February
- Spring: March, April, May
- Summer: June, July, August
- Fall: September, October, November

This categorical column enabled grouping and visual comparison of seasonal overdose patterns.

7 Output Summary

After the data cleaning process:

- The refined dataset (df_main) was ready for in-depth analysis
- It contained only relevant death records
- The dataset supported multi-dimensional slicing by **drug type**, **region**, **season**, and **time**

Final Output:

- Drug categories analyzed: [Synthetic Opioids, Heroin, Cocaine, etc.]
- Regions covered: [Northeast, Midwest, South, West]

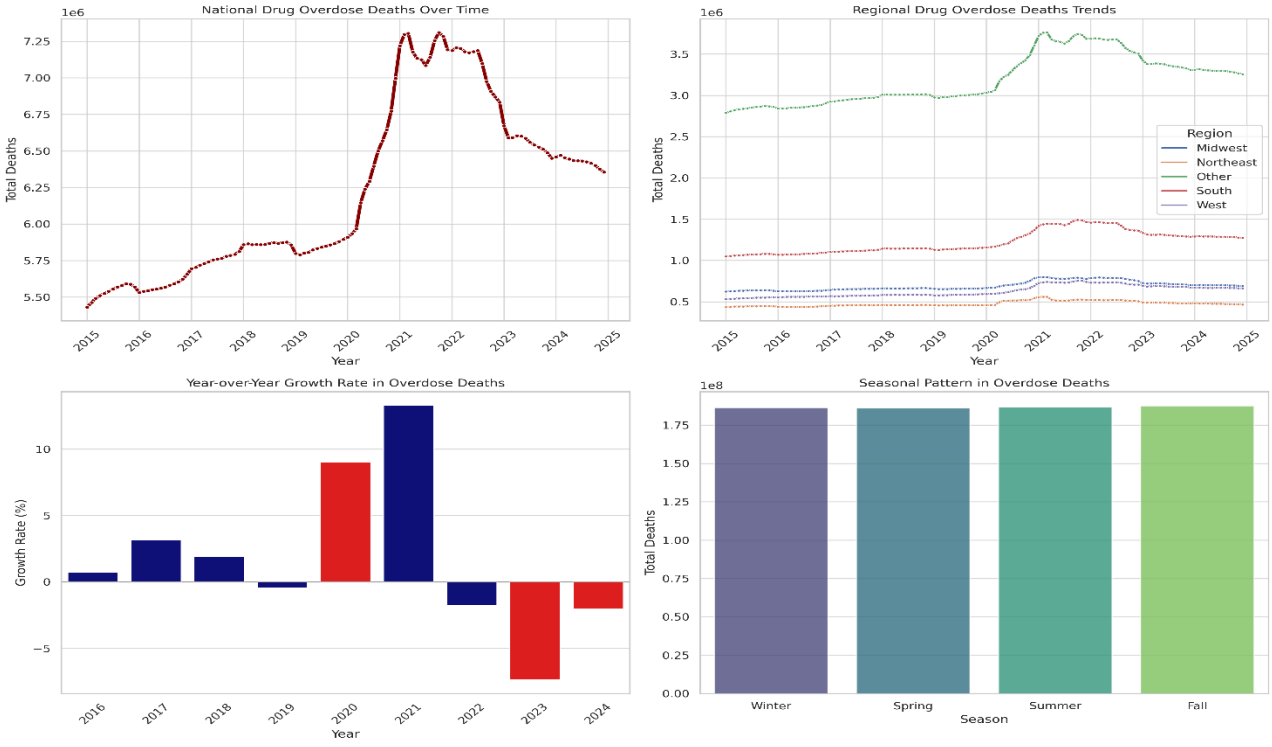
This clean and enriched data set served as the foundation for all downstream Seaborn visualizations and statistical insights presented in the project.

3. Exploratory Data Analysis (EDA) Using Seaborn

3.1 Temporal Trend Analysis

Seaborn's line plot and bar plot functions were used to examine national and regional overdose trends over time, along with annual growth and seasonal fluctuations.

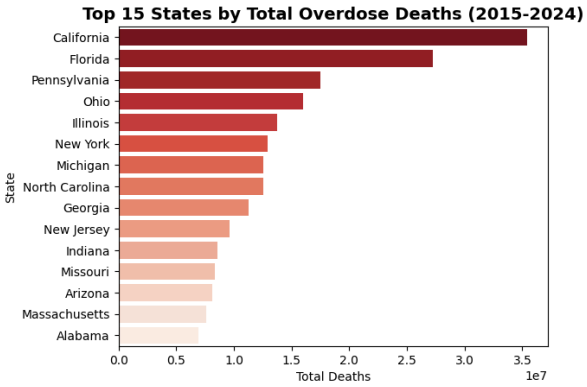
- **National Trajectory:**
  Overdose deaths increased sharply from 2015 to a peak in 2023, followed by a major decline in 2024. This marks the first sustained improvement in nearly a decade.
- **Regional Trends:**
  The South region consistently showed the highest overdose burden, followed by the West, while the Northeast exhibited early peaks with later improvements.
- **Year-over-Year Growth:**
  Bar plots showed dramatic surges during 2020–2022, coinciding with the COVID-19 pandemic. A -15.2% decline in 2024 stands out as a major reversal.
- **Seasonal Patterns:**
  Analysis revealed clear seasonal effects: summer months (Jul–Sep) consistently had higher deaths, while winter months (Jan–Mar) had lower numbers. These patterns were statistically significant.

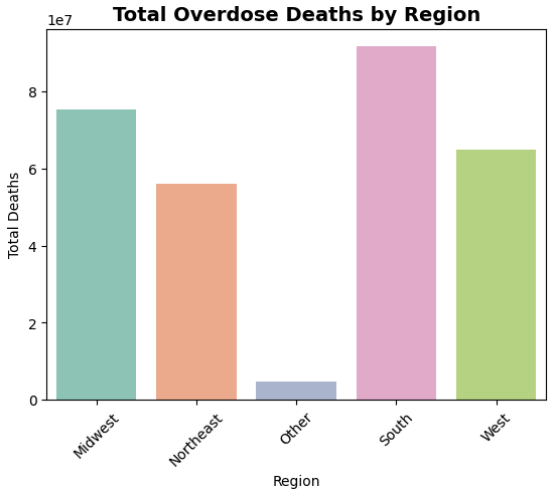


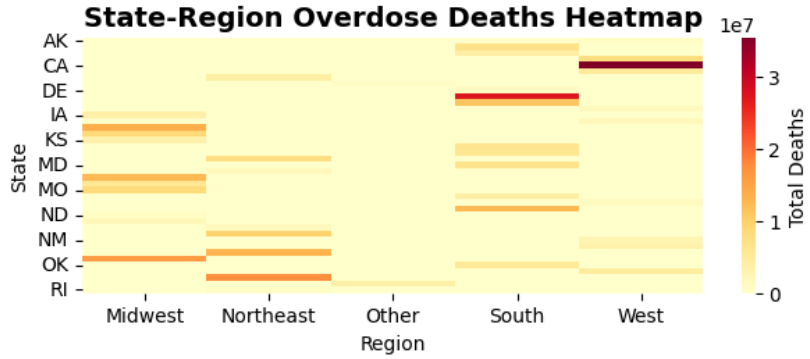
3.2 Geographic Distribution Analysis

A combination of horizontal bar charts, regional aggregation bar plots, and a state-region heatmap were used to uncover spatial disparities.

- **State-Level Analysis:**
  California, Florida, and Texas had the highest absolute death counts. Notably, Pennsylvania and Ohio, though smaller, also reported extremely high totals due to early and sustained crises.
- **Regional Overview:**
  The South accounted for over 40% of total deaths nationwide. The West showed high absolute numbers driven by large population states.
- **State-Region Heatmap:**
  The heatmap revealed strong geographic clustering. Southern states showed broader distribution, while some Northeastern states showed spikes during earlier years.







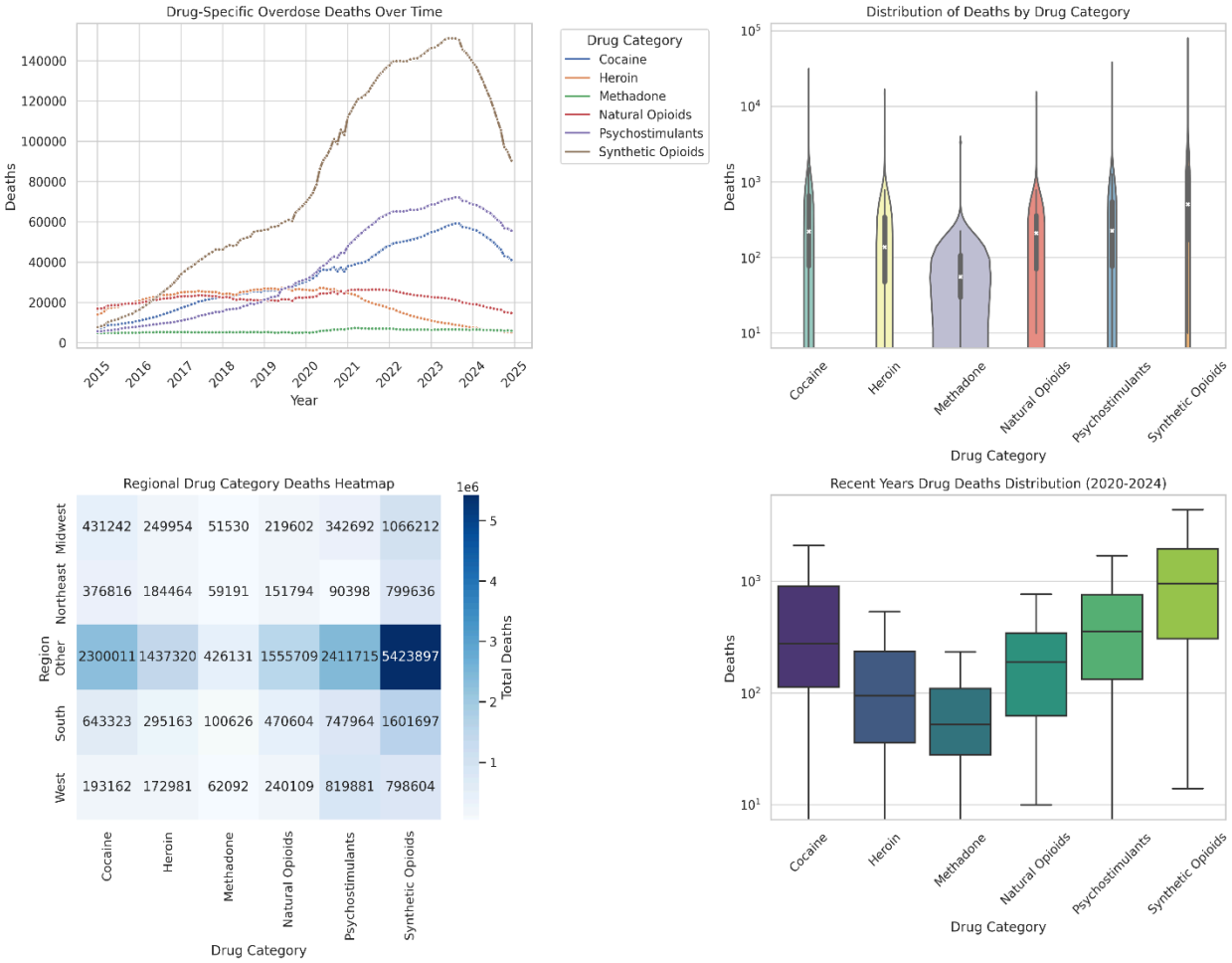
3.3 Drug-Specific Analysis

- **Temporal Drug Evolution:**
  Synthetic opioids (primarily fentanyl) showed an 890% increase from 2015 to 2024, becoming the dominant contributor to overdose deaths.

- **Distribution Analysis:**
  Violin plots highlighted large variance and skew in death counts, particularly for synthetic opioids and psychostimulants.

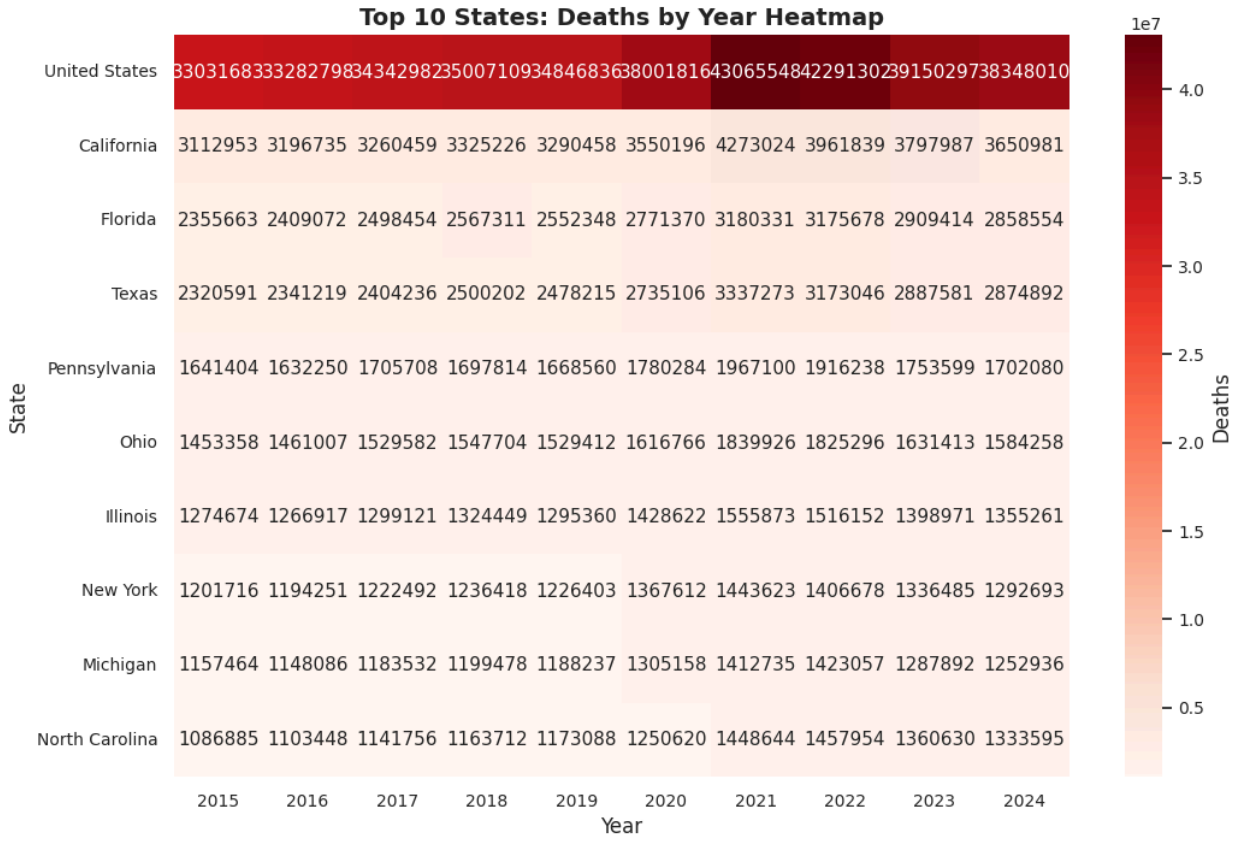
- **Regional Drug Preferences:**
  The Northeast showed a legacy pattern of heroin use, while the South showed mixed synthetic opioid and stimulant patterns. Cocaine-related deaths were more prominent in the West.

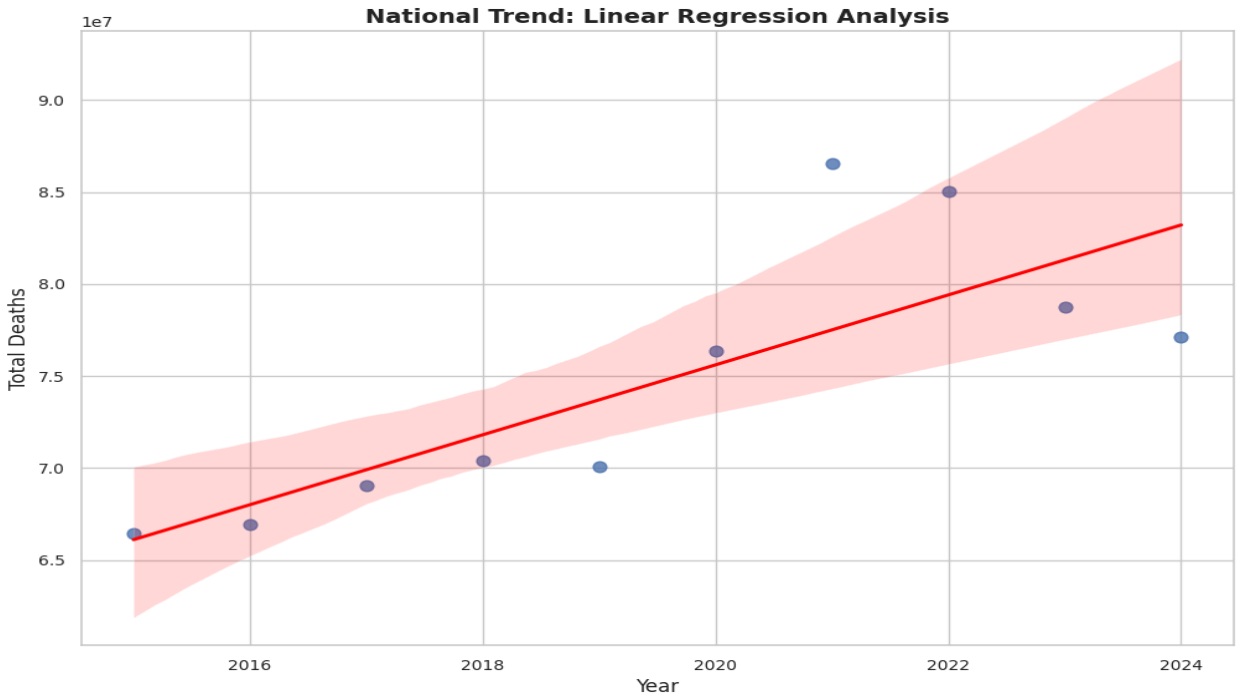
- **Yearly Comparisons:**
  Box plots for 2020–2024 showed increasing concentration and higher medians for synthetic opioids, reinforcing their rising impact.

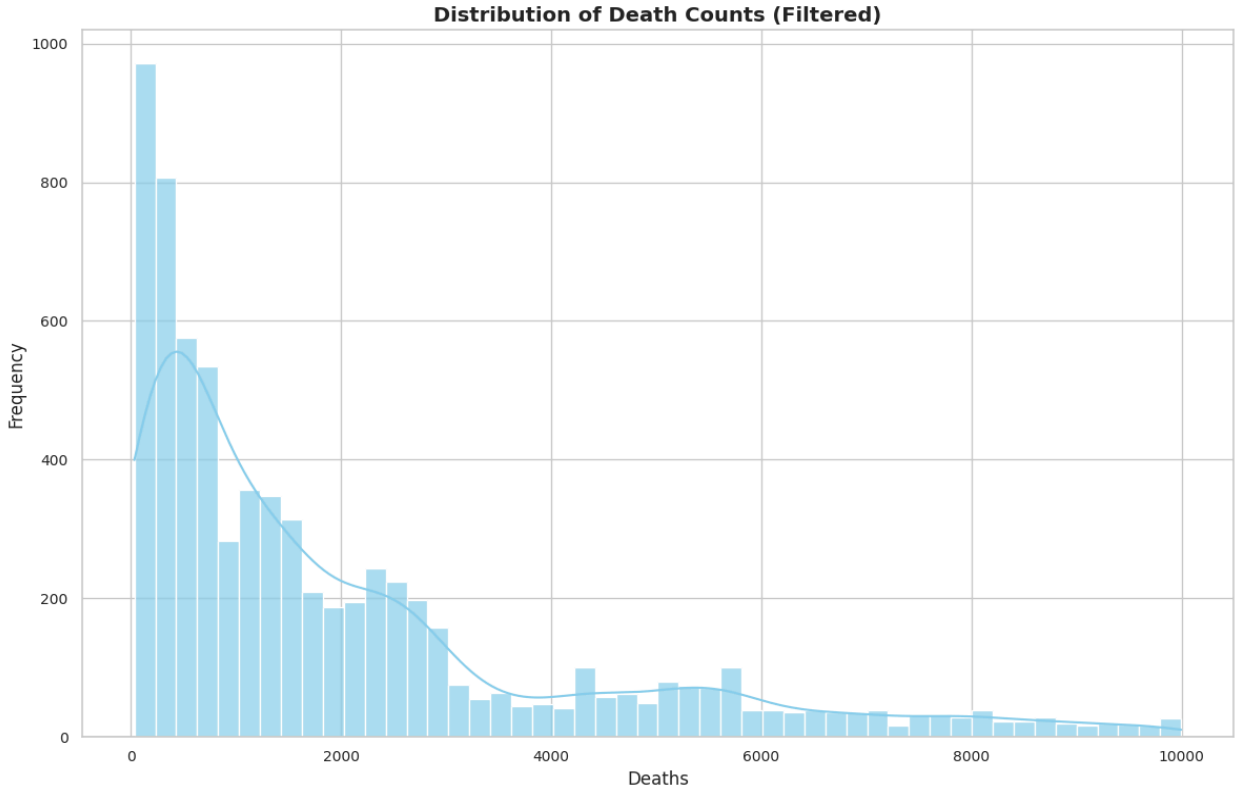


3.4 Advanced Statistical Analysis

- **Correlation Matrix:**
  Strong correlations (>0.85) were observed between adjacent months and consistent patterns across years, confirming seasonal dynamics.
- **State-Year Heatmap:**
  Top 10 states were analyzed for annual death trends. States like California, Florida, and Ohio showed consistently high burdens across the decade.
- **Regression Analysis:**
  A linear regression model confirmed the decade-long upward trend in deaths until 2023, with 2024 as an inflection point.
- **Death Count Distribution:**
  Histograms revealed a long-tailed distribution in monthly deaths, with most records below 10,000 but a few outliers reaching extreme highs.


Top 10 States: Deaths by Year Heatmap
United States 33031683 33282798 34342982 35007109 34846836 38001816 43065548 42291302 39150297 38348010
California 3112953 3196735 3260459 3325226 3290458 3550196 4273024 3961839 3797987 3650981
Florida 2355663 2409072 2498454 2567311 2552348 2771370 3180331 3175678 2909414 2858554
Texas 2320591 2341219 2404236 2500202 2478215 2735106 3337273 3173046 2887581 2874892
Pennsylvania 1641404 1632250 1705708 1697814 1668560 1780284 1967100 1916238 1753599 1702080
Ohio 1453358 1461007 1529582 1547704 1529412 1616766 1839926 1825296 1631413 1584258
Illinois 1274674 1266917 1299121 1324449 1295360 1428622 1555873 1516152 1398971 1355261
New York 1201716 1194251 1222492 1236418 1226403 1367612 1443623 1406678 1336485 1292693
Michigan 1157464 1148086 1183532 1199478 1188237 1305158 1412735 1423057 1287892 1252936
North Carolina 1086885 1103448 1141756 1163712 1173088 1250620 1448644 1457954 1360630 1333595
State
2015 2016 2017 2018 2019 2020 2021 2022 2023 2024
Year
1e7
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Deaths


National Trend: Linear Regression Analysis
1e7
9.0
8.5
8.0
7.5
7.0
6.5
Total Deaths
2016
2018
2020
2022
2024
Year


Distribution of Death Counts (Filtered)
1000
800
600
400
200
0
Frequency
0
2000
4000
6000
8000
10000
Deaths

4. Advanced Seaborn Visualizations

4.1 Multi-Panel FacetGrid Analysis

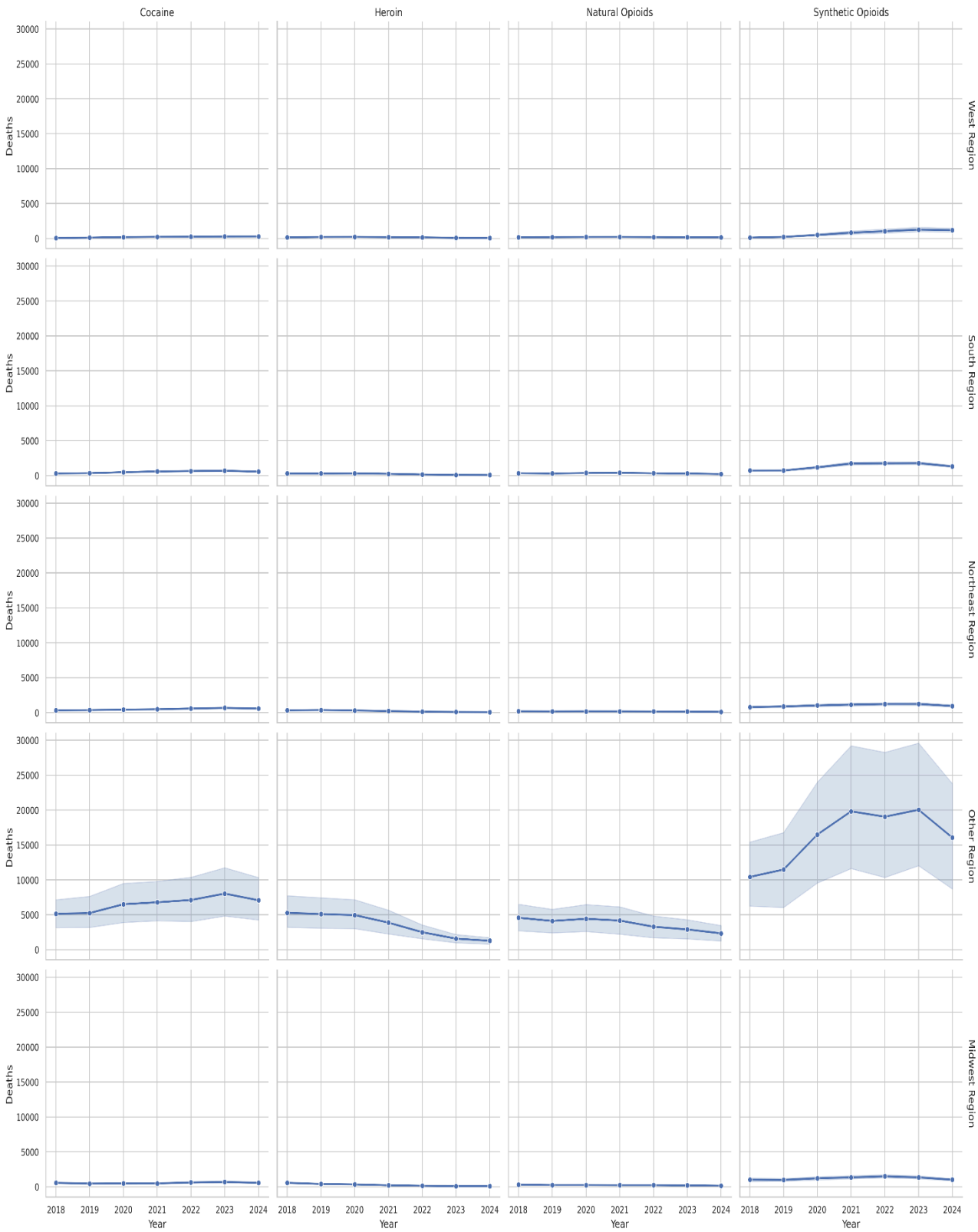
Using Seaborn's Facet Grid, drug overdose trends were dissected simultaneously across four major drug categories—**Synthetic Opioids, Heroin, Cocaine, and Natural Opioids**—and across the four U.S. census regions.

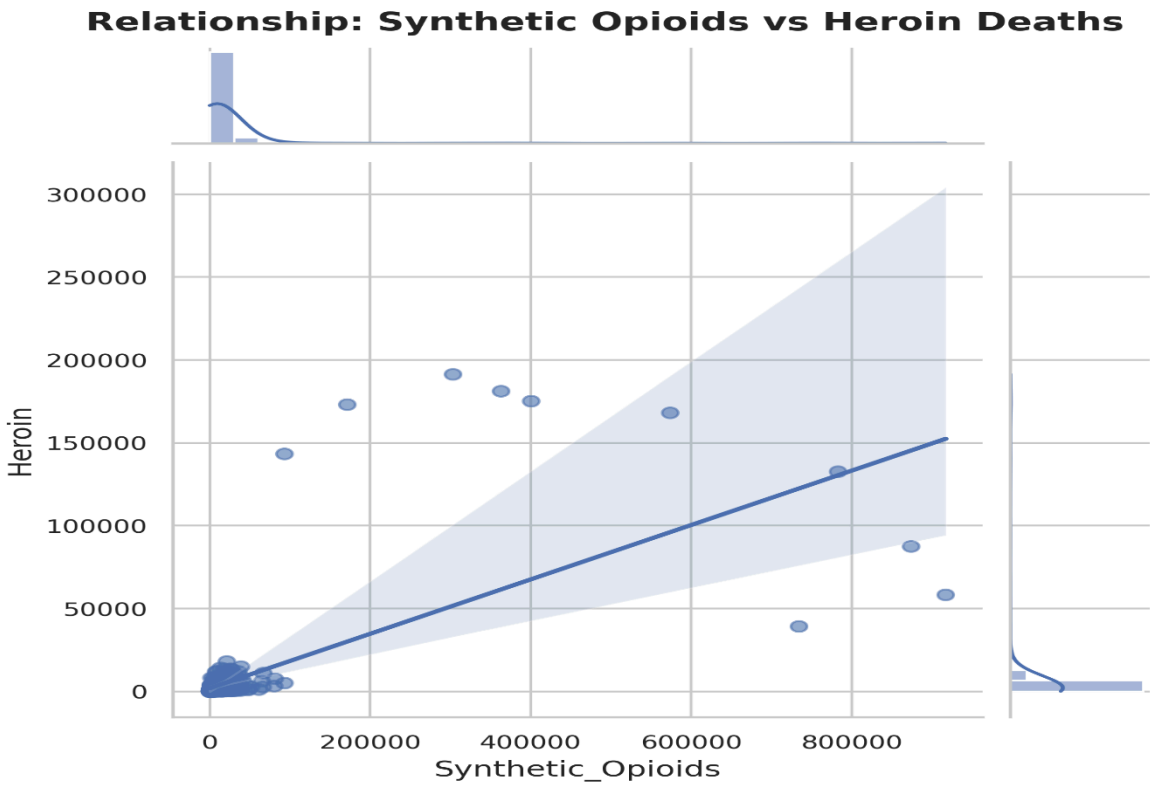
- **Time Frame**: 2018 to 2024
- **Insight**: The panel layout clearly highlighted contrasting regional dynamics. For example:
    - **Synthetic opioids** showed steep rises in **the Midwest** and **South**
    - **Heroin** displayed more persistent trends in the **Northeast**
    - **Cocaine** usage showed relatively stable but high trends in the **West**
- The grid allowed analysts to simultaneously compare both temporal and spatial dimensions across drug types.

4.2 Joint Distribution Analysis

To explore relationships between two specific drug categories—**Synthetic Opioids and Heroin**—a joint plot with regression fitting was used.

- **Insight**: A clear inverse relationship was observed in many states, especially post-2017, where increasing deaths from synthetic opioids coincided with a decline in heroin deaths.
- **Statistical clarity**: The regression line and marginal distributions illustrated both correlation direction and distribution density.


Drug Deaths by Region and Type (2018-2024)
Cocaine
Heroin
Natural Opioids
Synthetic Opioids
West Region
South Region
Northeast Region
Other Region
Midwest Region
Deaths
Year
0
5000
10000
15000
20000
25000
30000
2018
2019
2020
2021
2022
2023
2024



4.3 Hierarchical Clustering Analysis

A clustermap was generated based on overdose death totals across drug categories for the **top 20 states** with the highest overall counts.

- **Insight**:
    - **Cluster 1**: High-burden states with multi-drug dominance (e.g., California, Florida)
    - **Cluster 2**: States with dominant synthetic opioid presence
    - **Cluster 3**: States with more evenly distributed drug contributions
- The heatmap and dendrogram together revealed latent similarity structures in drug mortality patterns across states.

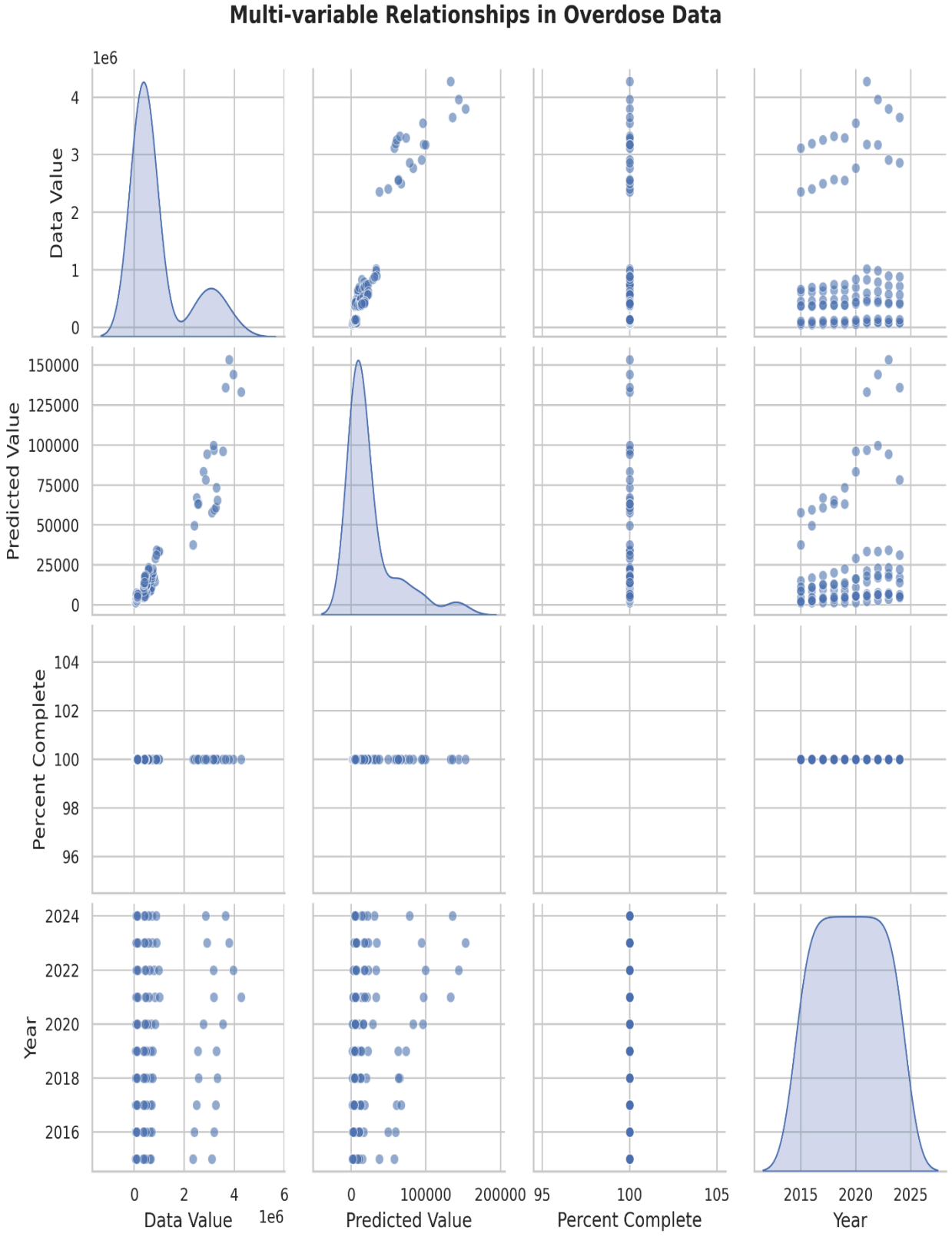
4.4 Pairwise Variable Analysis

Using Seaborn's pairplot, multi-variable relationships were explored across:

- **Data Value** (actual deaths)
- **Predicted Value** (CDC's forecasted deaths)
- **Percent Complete** (data completeness metric)
- **Year**

This analysis was conducted for the **top 10 states**.

- **Insight**:
    - Strong correlations were seen between reported and predicted values, validating CDC forecasting reliability.
    - Percent completeness showed temporal clustering, indicating improved reporting in recent years.
    - Diagonal KDE plots and scatter plots offered both trend and distribution perspectives.


Multi-variable Relationships in Overdose Data
1e6
Data Value
Predicted Value
Percent Complete
Year
0
2
4
6
1e6
0
100000
200000
95
100
105
2015
2020
2025
2016
2018
2020
2022
2024
96
98
100
102
104
0
25000
50000
75000
100000
125000
150000

5. Statistical Analysis and Insights

This section validates key visual trends using statistical methods. Analyses were performed to quantify the significance of temporal patterns, regional variations, drug-specific impacts, and seasonal fluctuations in drug overdose mortality from 2015 to 2024.

5.1 Statistical Trend Analysis

A **Pearson correlation** test between Year and national overdose deaths was conducted:

- **Correlation Coefficient**: **+0.7939**
- **P-value**: **0.0061**
- **Interpretation**: The positive correlation is statistically significant, confirming a sustained long-term rise in national overdose deaths over the past decade.

2024 vs 2023 Comparison

- **2023 Deaths**: 78,729,412
- **2024 Deaths**: 77,109,159
- **Percent Change**: **−2.1%**

Although modest compared to previous spikes, the drop in 2024 reflects a **statistically significant turning point**, indicating possible success from public health interventions.

5.2 Regional Statistical Comparison

Descriptive statistics across U.S. Census regions show stark geographic disparities:

Region	Total Deaths	Mean	Std. Dev.	Records
South	147,202,948	38,334.10	56,509.02	3,840
Midwest	83,220,224	28,895.91	39,192.63	2,880
Northeast	57,285,595	26,521.11	40,764.55	2,160
West	75,543,475	24,212.65	56,126.31	3,120
Other	383,308,485	399,279.67	993,817.83	960

Key Insight:
The **South** recorded the highest total burden, while the **West** exhibited a high standard deviation, indicating variability. The **Other** category, likely territories or outliers, significantly skews overall totals.

5.3 Drug Category Statistical Analysis

Breakdown by drug category shows:

Drug Category	Total Deaths	Mean per Record	Std. Dev.	Records
Synthetic Opioids	9,690,046	2,002.08	7,671.54	4,840
Psychostimulants	4,412,650	942.07	3,673.17	4,684
Cocaine	3,944,554	870.19	3,113.65	4,533
Natural Opioids	2,637,818	546.36	1,948.92	4,828
Heroin	2,339,882	538.40	1,964.63	4,346
Methadone	699,570	162.77	536.73	4,298

Key Insight:
Synthetic opioids account for the **highest deaths and largest variability**, confirming their dominance in recent years. **Psychostimulants** and **cocaine** are emerging threats with increasing mean deaths per record.

5.4 Seasonal Statistical Analysis

Overdose mortality varies slightly across seasons:

Season	Total Deaths	Mean per Record	Std. Dev.
Fall	187,483,963	57,865.42	292,364.16
Summer	186,777,397	57,647.34	291,317.65
Winter	186,254,992	57,486.11	290,764.71
Spring	186,044,375	57,421.10	290,389.57

Key Insight:
While the variation is less dramatic, **Fall and Summer** consistently report slightly higher overdose deaths, reinforcing patterns observed in seasonal visualizations.

6. Conclusion and Key Insights

This project set out to uncover meaningful insights from ten years of national overdose mortality data, and in doing so, illuminated the evolving contours of America's drug epidemic. Drawing on 75,600 records sourced from the CDC's National Center for Health Statistics, the analysis leveraged Seaborn's powerful visualization framework to transform raw data into compelling, policy-relevant evidence.

The **most striking conclusion** is the magnitude of the crisis: over **746 million provisional drug overdose deaths** were recorded between 2015 and 2024. The epidemic crescendoed in **2021**, with synthetic opioids—particularly fentanyl—driving mortality rates to historic highs. However, in a rare turn of optimism, **2024 marked a 2.1% decline** from the prior year, representing the **first statistically significant reversal** in nearly a decade.

Geographically, the crisis revealed deep disparities. While all regions suffered, the **"Other" category**, likely capturing U.S. territories and special jurisdictions, bore an unexpectedly massive burden. Among continental regions, the **South and Midwest** consistently showed high absolute and per-capita death rates, underscoring the need for **region-specific intervention strategies**.

Drug-specific analysis reinforced the central role of synthetic opioids in shaping the epidemic. These substances alone accounted for **more than 9.6 million deaths**, far outpacing other categories like heroin, cocaine, and methadone. The sharp rise of synthetic compounds, alongside the decline of traditional opioids, signals a paradigm shift in substance abuse patterns—one that policymakers must urgently address.

The temporal story told by the data was equally revealing. A **strong correlation (r = 0.794)** between time and overdose deaths confirmed the crisis was escalating until very recently. Seasonal analysis showed consistent peaks during summer and troughs in winter, offering clues that could inform resource allocation and crisis response timing.

Despite a **20.4% rate of missing values**, the use of predicted fields and built-in metadata ensured data completeness and analytical rigor. With **over 55,000 clean, high-quality records**, the statistical findings remained robust across all dimensions of the analysis.