**The data set:**

The dataset chosen is the Centres for Disease Control & Prevention’s dataset on Deaths due to Drug Poisoning in the United States of America. We cover the dataset’s various attributes by taking a top – down approach: *Country -> State -> County* and analysing the trends within each.

The main questions answered are:

1. What has the trend of deaths due to drug poisoning been like over the course of a decade?
2. Where have we seen maximum number of such cases?
3. What age group/sex is most affected by such deaths?
4. What classification of county’s are most likely to have maximum number of cases?

**Data Cleaning and ETL**

1. **The process of understanding and loading:**
   1. The process of understanding this varied dataset began with first carrying out basic scanning through the various rows and columns. This allows us to get more comfortable with the parameters.
   2. Looking up the terms like crude death rates (raw value) made it easier to relate to the kind of data that was dealt with.
   3. While digging deeper into the concept of death rates due to drug poisoning, I came across multiple articles and papers (all sourced in the power point). This gave me a much deeper understanding of the plight faced by people battling the opioid crisis.
2. **The data cleaning process:**
   1. The cleaning of this data set took some understanding of what I needed cleaned up. Setting up questions before I started my analysis allowed me to leverage the data into a system I needed.
   2. It started off with checking the basic descriptive statistics of the deaths.

Table

Description automatically generated

* 1. Since I needed to analyse the states and distribution of deaths, I checked for any discrepancies with the state names and county names in the sheet.

|  |  |
| --- | --- |
| No table of figures entries found. | **Sum of Crude Death Rate** |
| Alabama | 205.792 |
| Alaska | 278.4286 |
| Arizona | 311.3212 |
| Arkansas | 206.4536 |
| California | 199.6508 |
| Colorado | 276.9633 |
| Connecticut | 286.5854 |
| Delaware | 320.1935 |
| District of Columbia | 346.2309 |
| Florida | 292.3046 |
| Georgia | 192.411 |
| Hawaii | 197.7532 |
| Idaho | 202.3027 |
| Illinois | 229.014 |
| Indiana | 264.7443 |
| Iowa | 131.4518 |
| Kansas | 178.3475 |
| Kentucky | 385.7651 |
| Louisiana | 284.876 |
| Maine | 276.5839 |
| Maryland | 340.7148 |
| Massachusetts | 320.9456 |
| Michigan | 272.3743 |
| Minnesota | 148.6295 |
| Mississippi | 185.8186 |
| Missouri | 276.8893 |
| Montana | 217.3668 |
| Nebraska | 106.9373 |
| Nevada | 377.5313 |
| New Hampshire | 319.0144 |
| New Jersey | 251.3328 |
| New Mexico | 423.5064 |
| New York | 191.8714 |
| North Carolina | 249.3562 |
| North Dakota | 88.7183 |
| Ohio | 343.5263 |
| Oklahoma | 310.316 |
| Oregon | 225.9317 |
| Pennsylvania | 357.4493 |
| Rhode Island | 351.787 |
| South Carolina | 240.8178 |
| South Dakota | 108.2312 |
| Tennessee | 319.1003 |
| Texas | 173.7015 |
| United States | 23768.0509 |
| Utah | 338.9503 |
| Vermont | 240.1048 |
| Virginia | 193.7716 |
| Washington | 267.0563 |
| West Virginia | 501.2135 |
| Wisconsin | 225.4354 |
| Wyoming | 225.5357 |
| **Grand Total** | **37029.16** |

* 1. Benford’s Law:

An expected result was not obtained from this test. Zeroes were removed using multiplier method.

Chart, bar chart

Description automatically generated

* 1. Z Test (outliers)

Using Z Test I analysed the Z score values less than -3.0 and greater than 3.0. It was noted that 31 fell out of the range greater than 3.0. Of these only 3 were beyond 4.0 with one heavy outlier having a score of 5.47. These observations were considered while creating the data visualizations. However, since it lies in the United States category it was not analysed in great detail.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sex | Age | Race and Hispanic Origin | State | Deaths | Population | Crude Death Rate |
| Male | 25-34 years | Non-Hispanic White | United States | 6874 | 12848868.00 | 53.50 |

**The visualizations:**

The visualizations contain:

1. Line Chart
2. Bar Graph
3. Stacked Bar Graph
4. Filled Map
5. Pie Chart
6. Decomposition Tree
7. Scatter Plots & (Cluster Analysis)

Each of these visualizations serve a specific purpose in highlighting, analysing, and profiling of the dataset parameters. Making use of the different kind of analytics, the power point touches on all 3 out of 4 of the analytics, descriptive, diagnostic, and prescriptive.

Descriptive:

This analysis answered most of the questions pertaining What happened? The first 6 visualizations mentioned above covered this delivering better hindsight by analysing the existing information.

These visualizations gave a deeper understanding about how the data has developed over a period, the number of such casualties, the gender and age group of the people affected and an in depth understanding of the regions affected by drug toxicity deaths.

Diagnostic:

This analysis answered the deeper question of why these deaths occurred. This has been covered in detail in the ppt’s West Virginia analysis as well as the conclusion.

Prescriptive:

This analysis was heavily dependent on the current scenario of opioids and the drug crisis. How we can reduce the effect of these drugs and prevent them from taking over people’s lives is something that is touched upon towards the very end.

**Summary:**

To summarize the entire project, I learned about the flow of understanding how to deal with data, cleaning it and summarizing it to come up with a proper visualization supporting my observations. The art of finding a story with data is something that this project pushed me to do. This project not only led me to understand just how serious the problem of drug abuse is but also the extent to how much it affects people’s lives personally. In depth study and understanding of them problem allows you to tell a better story.