

Data Storytelling Project

10-02-2023

Made By:

Monish Gosar (J025)

Atharva Rode(J056)

Kallind Soni(J065)

Prsni Kanani(J073)

Overview

The dataset we chose describes drug poisoning deaths at the U.S. and state level by selected demographic characteristics, and includes age-adjusted death rates for drug poisoning from 1999 to 2015.


Goals

1. To find the key factors affecting the death rate in the United states.
2. To Summarize the given dataset and draw conclusions.

Specifications

By Columns-

- 1) **Year** - Contains years from 1999 to 2015.
- 2) **Sex** - Consists of Male , Female and Both Sexes for the United States whereas only Both Sexes for other 51 states.
- 3) **Age** - Consists of age groups from 15 years to 75 years in a continuous interval of 10 years each for the U.S. and all ages for other 51 states.
- 4) **Race and Hispanic Origin** - All Races - All Origins (51 states) and Hispanic , Non Hispanic Black , Non-Hispanic White for the United States.
- 5) **States** - Consists of 51 states alongside the United States as a whole.
- 6) **Deaths** - Calculated deaths by states.

- 
- 7) **Regions** - Consists of West , South , Northeast , Midwest regions.
 - 8) **Population** - Calculated population of a particular state.
 - 9) **Crude-Death Rate** - Refers to rates per 100,000 population.
 - 10) **Age-Adjusted Rate**- Refers to rates that are calculated as if the deaths occurred in a population with the same age structures.

ETL PROCESS

Firstly, we collected the data from the CDC website and understood the data in excel. Then, we cleaned the existing data by eliminating the unnecessary data columns such as standard error, upper confidence limit, lower confidence limit etc. and then found the United States by Region and added a new Region column in our dataset for better findings.

We later added the data into visualization tools such as Power BI and Tableau. To find the relationship between the columns we carefully choose the variables and the data representation that would allow us to show our findings efficiently. We have chosen different data representation charts such as line graph, tree map and area chart to visually compare different variables and infer conclusions.

Understanding Visualizations

Tree Map [Slide 2]


This visualization shows the relationship between state and average crude death rate. It shows the **top 10 states with the highest average crude death rates**. A tree map helps to easily compare the visuals and obtain results. **West Virginia** is the state with the highest rate of 20.68% while Tennessee is 10th place with 14.194%. We have used the rate instead of deaths here to keep the population involved as a factor influencing it.

Area Chart [Slide 03]

The area under the top most line shows the **total deaths over the years**. The areas have been split on the basis of age by intervals of 10 years. By representing data this way we can understand that, the maximum deaths occur in the **45 to 54 age** interval followed by the 34 to 45 age interval. The least number of deaths are recorded in the age interval less than 15 years. We can also infer that over the years the death counts have gone up by only a bit for age 75+ .

Stacked Bar Chart [Slide 04]

This visual representation shows the relationship of the **number of deaths grouped by the age groups and then by their race/origin**. By seeing this stacked




bar chart we can draw the conclusion that the highest number of deaths takes place between the **ages 45-54 with non-hispanic whites** being the most affected. On the other hand the least affected age groups are less than 15 followed by age groups of 75+. This graph shows how each age group's race and origin contributes to the total number of deaths.

Line Chart [Slide 05]

This line graph throws light upon the correlation between the **number of deaths and the years**. Here we have segregated all the states of the United States into 4 categories that are Midwest, Northeast, South and West for better and clear understanding. The lines of all 4 regions show a continuous increase in the number of deaths with increasing years. The **South region** had the highest number of deaths over the years (1998-2015). We also observed that Midwest started with the least number of deaths but over time surpassed the North and the South region making it the second highest region by 2015.

Decomposition Tree [Slide 6 & 7]

The first decomposition tree reveals the **combination of age, sex, and race** with the highest crude death rate: **45-54 years old, male, and non-Hispanic black** have an average crude death rate of 25.3%. This is the highest death rate across all combinations of the given variables, indicating that this demographic is the most



vulnerable to death. The second decomposition tree illustrates the combination of age, sex, and race with the lowest crude death rate: less than 15 years of age, female, and Hispanic has an average crude death rate of 0.14%. This is the lowest death rate across all combinations of the given variables, indicating that this demographic is the least likely to experience death.

Aster plot [Slide 8]

The Aster Plot displays radial slices whose radius is based on a numeric measure and whose arc length is based on a 'weighted' measure. Here we have taken the **Average Crude death rate by region** and visually displayed the highest and the lowest regions. The highest Rate was found in the **West region** while the lowest was found in the **Mideast region**.

Bar Graph [Slide 09]

The bar graph on this slide shows the connection between the **states of the United States and the number of deaths**. Here we have shown the **top 10 states** that consist of the highest number of deaths, concluding that the state of **California** had the highest number of deaths. Death as a parameter was chosen here to show the highest number of deaths **independent of the population** of that state. The death rates are affected by the population which does not allow us to infer exact values so choosing death variable here helps us understand that.

Conclusion

We can see that California has the most number of deaths i.e 59,427 and that the overall deaths are increasing over the years while West Virginia has the highest crude death rate.

Overall the ages 45-54 years and male and Non-hispanic Black have the highest death rate whereas the Non - hispanic White have the most number of deaths (not taking into account the population).

In conclusion, there are many factors that influence the rate of death by drug poisoning across the United States, including age, sex, and race. By analyzing different visualizations, we are able to identify the demographic that is most likely to experience death and the demographic that is least likely to experience death. We can also identify the states with the highest and lowest average crude death rates, as well as the regions that have experienced the highest and lowest total number of deaths over the years. By understanding these factors, we can work to improve, see the trends and can act on them accordingly.