Beza Arega

New York City Leading Cause of Death Report

I chose the dataset New York City Leading Cause of Death for my project because it provides vital data on the various health problems that the residents in the city face. In this report I will show insight to the demographics of the data and the trends over the years. This dataset was from the year 2007-2021 and has 6 variables which are Leading Cause, Gender, Race Ethnicity, Deaths, Death Rate, and Age Adjusted Death Rate.These variables help us gain a better understanding of the data.

In this dataset, there were a total of 2102 death counts throughout the years. I used a range of graphs to gain a better insight about the data. The 3 different graphs used for the visualizations are Histogram, Count plot and Catplot. I visualized the different leading causes for these deaths using a Histogram. I was able to conclude the year with the highest and lowest death rate with a count plot. The catplot was used to show the death rate throughout the years based on race & ethnicity and gender.

I applied a histogram to analyze the leading cause of death in New York. The highest leading causes of death were disease of heart, malignant neoplasm (cancer), and influenza and pneumonia.The lowest leading causes were tuberculosis, meningitis, pregnancy childbirth and the puerperium and human immunodeficiency virus disease. The histogram provided a clear visualization of the leading causes, by listing all the causes and summarizing the large dataset.

The second graph displays the death rate throughout the years 2007-2021. There was a steady rate throughout the years with slight increase and decrease then there was a significant increase in 2019. The year 2019 had the highest death rate. The lowest death rate was in the years 2013 and 2020. Count plot was perfect to show the timeline of the rates throughout the years. It marked all the years and made it easy to compare the rate throughout the years.

To better understand why 2019 had the highest death rate, I used a histogram. The graph shows the highest leading causes of death in 2019 were disease of heart, malignant neoplasm (cancer), diabetes mellitus, and influenza and pneumonia.

I chose a cat plot to visualize the correlation between the death rate throughout the years by race & ethnicity and by gender. The race & ethnicity of this dataset seems to be equal throughout the years, there was an even amount of white non hispanic, hispanic, asian and pacific islander, and black non hispanic. The only outliers are the races & ethnicities that were not mentioned. New York is a very diverse state which is probably why there wasn't any difference in death rates based on race and ethnicity. Although there wasn’t any difference by race & ethnicity, there was a difference in rates when it came to gender. Females seem to have more deaths than males. There isn’t a significant difference between the death counts between the 2 genders, but we can see that the female death count is slightly higher than the males throughout the years. There are years in which the 2 genders are equal like in 2010 and 2019. The cat plot was great to compare the distribution of the yearly death rates by race & ethnicity and by gender because it provided clear visuals to see trends or any patterns throughout the years.

In this analysis, I used a range of graphs such as histogram, count plot, and cat plot to visualize the dataset. The histogram showed the highest and lowest leading causes of death in New York and the count plot visualized the year with the highest and lowest death rate. The cat plot was a great way to show the pattern between the death rate throughout the years based on race & ethnicity and gender. There were equal amounts when looking at the rates throughout the years with each race and ethnicity. There was a difference in the rates when it was based on gender, since females had a slightly higher death rate than male throughout the years. The different visualizations helped to better understand the dataset. For future analysis, I would say having an additional variable of Age in the dataset would enhance a better insight. Overall, the dataset provided enough information to visualize and gain better understanding.