Muzahidul Islam, Saki Takatsu, John Harrison,
James Soltis, Isabel Pacheco Mattivi
DAT400 Spring 2025

Estimating Excess Deaths Due to Covid in New York State 2020-2022

Introduction

COVID-19 transformed New York State from one of the nation's healthiest jurisdictions in 2019 into the epicenter of excess mortality only months later. In 2020 alone, all-cause deaths jumped by roughly **30 %**—a reversal not seen since the 1918 influenza pandemic. Understanding *how many* of those deaths were truly "in excess," *which* populations were most affected, and *why* the toll evolved across 2020-2022 is critical for evaluating public-health policy and preparing for future crises.

This report quantifies excess deaths in New York State from **2020 through 2022** and asks three guiding questions:

1. **How large was the mortality shock?**
   We observed benchmark deaths against statistically expected baselines derived from 1997-2019 vital-records data.
2. **Who was most affected?**
   Exploratory data analysis (EDA) disaggregates death counts and per-capita rates by age, sex, and race/ethnicity to reveal differential trends.
3. **How well can we forecast a counterfactual "non-COVID" trajectory?**
   We compare several time-series and count-data models—simple moving average, Poisson regression, ARIMA, exponential smoothing, and the demographic Lee–Carter model—to estimate expected deaths and to isolate the pandemic's excess.

Data & Scope

We assembled annual New York vital-statistics files for 1997–2022. Each year's data include total population; male and female populations; total deaths with breakdowns by sex, eleven age bands, place of death (hospital, home/public place, nursing home, hospice, other), twelve cause-of-death categories, and race/ethnicity counts. We use 1997‑2019 to train our models and 2020‑2022 to measure excess mortality.

Data Transformations

We merged the 26 yearly files into one master dataset and fixed any minor naming inconsistencies. In our analysis code, we calculated per-capita and subgroup death rates by dividing each count by that year's total population. No other new variables or predictors were created.

Study Limitations

Our analysis is based entirely on past death counts as the sole predictor, which means we cannot leverage other important factors, such as insurance coverage levels or socioeconomic indicators, that might improve forecast accuracy. Delays in reporting and misclassification of certain deaths introduce potential bias into our excess-death estimates. Finally, these results reflect New York's specific population structure and vital statistics reporting practices; applying the same approach in areas with different demographics or data quality may yield different findings.

Literature Review

- **Excess Mortality Estimation in New York**
  Excess mortality, the number of deaths above an expected baseline, has been estimated in New York using various methods. Early approaches compared 2020–2021 deaths to averages from prior years (e.g. NYC compared spring 2020 deaths to the 2015–2019 average). More advanced methods, such as the CDC's Farrington algorithm, use historical data with seasonal adjustments to establish 95% prediction intervals, while time-series regressions (e.g., a Census Bureau study based on 2010–2019 data) provide counterfactual baselines. Although methodological choices and data lags introduce uncertainty, agencies like WHO and CDC offer standardized frameworks for measuring excess mortality.

- **Policy Responses and Their Impact on Mortality**
  New York implemented early, strict COVID-19 measures, including lockdowns, mask mandates, and aggressive vaccination, which initially couldn't prevent a severe mortality surge in spring 2020 but later contributed to improved outcomes. Analyses have shown that after May 2020, the Northeast (including NY) had significantly lower COVID mortality than regions with less stringent policies. While the early surge was catastrophic, the state's subsequent public health interventions helped lower excess deaths, as evidenced by comparative studies and regional analyses.

Muzahidul Islam, Saki Takatsu, John Harrison,
James Soltis, Isabel Pacheco Mattivi
DAT400 Spring 2025

- **COVID Data Challenges and Limitations**

  Accurate mortality data during the pandemic were hampered by under-reporting, misclassification, and delays. In New York, early controversies arose when nursing home deaths were undercounted due to reporting only on-site fatalities. Adjustments by the CDC help, but uncertainty remains regarding cause attribution and reporting lags. Despite these issues, New York's relatively complete all-cause mortality data have enabled researchers to gauge the true pandemic impact, though interpretations must account for baseline uncertainties.

- **Pre- vs. During-Pandemic Mortality Trends**

  Before COVID-19, New York experienced declining mortality and rising life expectancy, boasting one of the nation's lowest age-adjusted death rates in 2019. However, in 2020, deaths spiked sharply. Statewide deaths increased by roughly 30%, and NYC's weekly death rates more than doubled. Although 2021 saw continued excess mortality, improved treatments and vaccination helped bring rates closer to pre-pandemic levels. Life expectancy in NYC, for instance, fell dramatically in 2020, underlining the stark impact of the pandemic.

- **Demographic Patterns in Excess Mortality**

  COVID-19's toll in New York was unevenly distributed. Older adults bore the highest mortality, and racial/ethnic minorities, particularly Black and Hispanic communities, experienced substantially higher excess mortality rates due to factors like crowded living conditions and frontline work. Socioeconomic disparities and geographic variations further compounded these differences, highlighting longstanding inequities that were exacerbated during the pandemic.

- **Lockdown Measures and Mortality Trends**

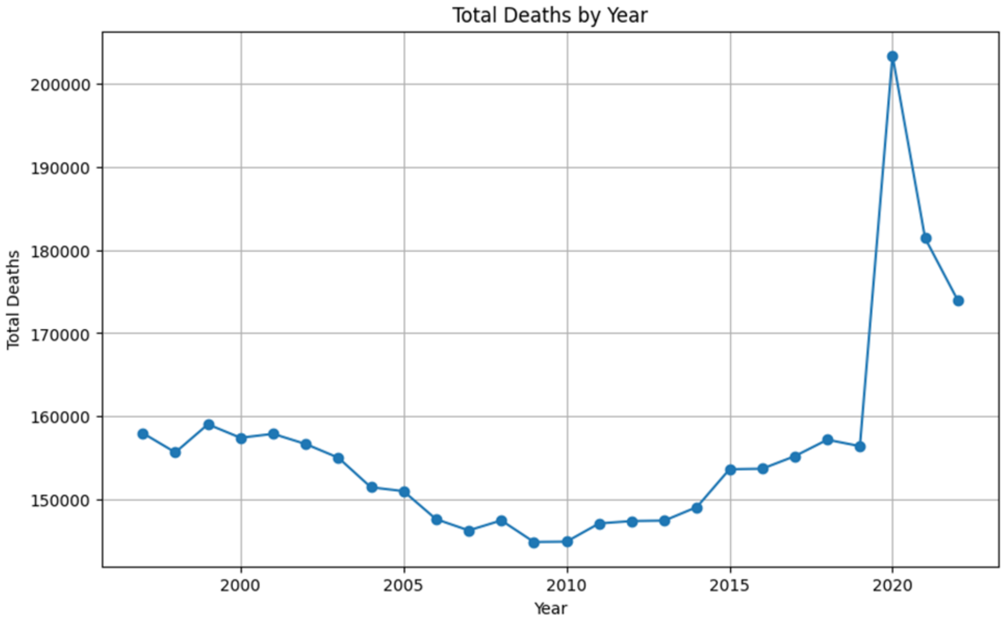
  New York's early "PAUSE" lockdown helped slow viral transmission, reducing deaths in later weeks and lowering other causes of mortality (e.g., traffic fatalities). However, the lockdown also disrupted non-COVID care, contributing indirectly to mortality from other causes and worsening mental health outcomes. Studies offer mixed results on the net effect, but overall, strict mitigation combined with subsequent public health measures helped reduce overall excess mortality.

Muzahidul Islam, Saki Takatsu, John Harrison,
James Soltis, Isabel Pacheco Mattivi
DAT400 Spring 2025

- **Nursing Home Policies and Their Effects on Mortality**
  Nursing homes in New York were particularly hard-hit during COVID-19. A controversial policy early in the pandemic led to the readmission of COVID-positive patients into nursing homes, which, combined with PPE shortages and underreporting of off-site deaths, contributed to extremely high fatality rates. Reviews have since criticized these policies and recommended reforms, emphasizing that improved infection control and preparedness are critical for protecting vulnerable populations.

EDA

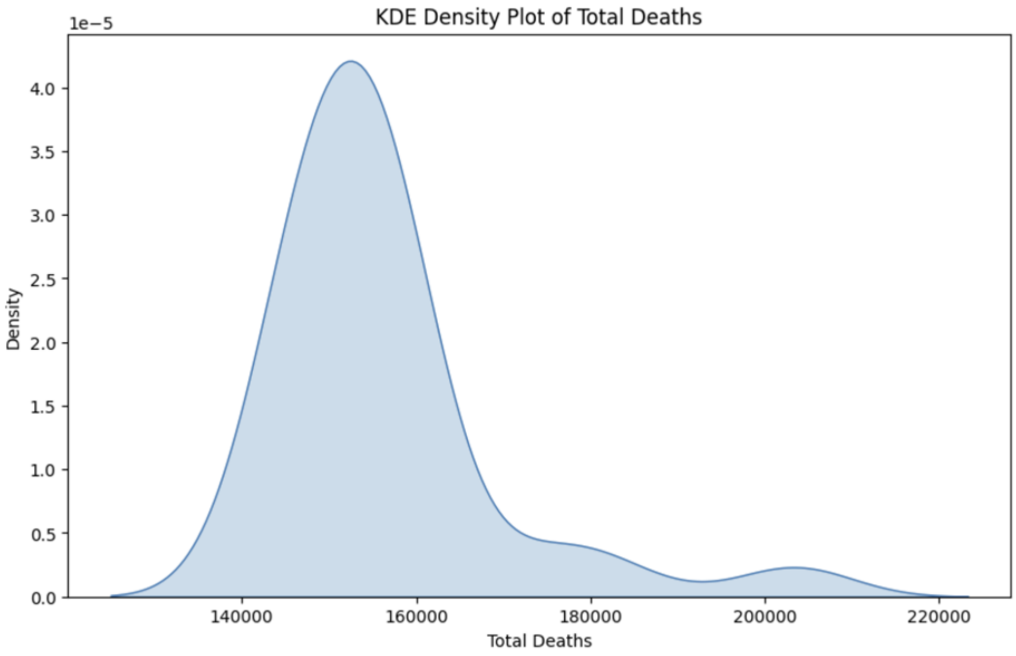
- **Overall Deaths**
  - **Chart 1: Overall Death Count**



- Observation:
  - The first chart shows that the total number of deaths in New York was around 158,000 in 1997. Death counts dropped to a low of approximately 145,000 in 2009 and then began to climb again in subsequent years.
- Interpretation:

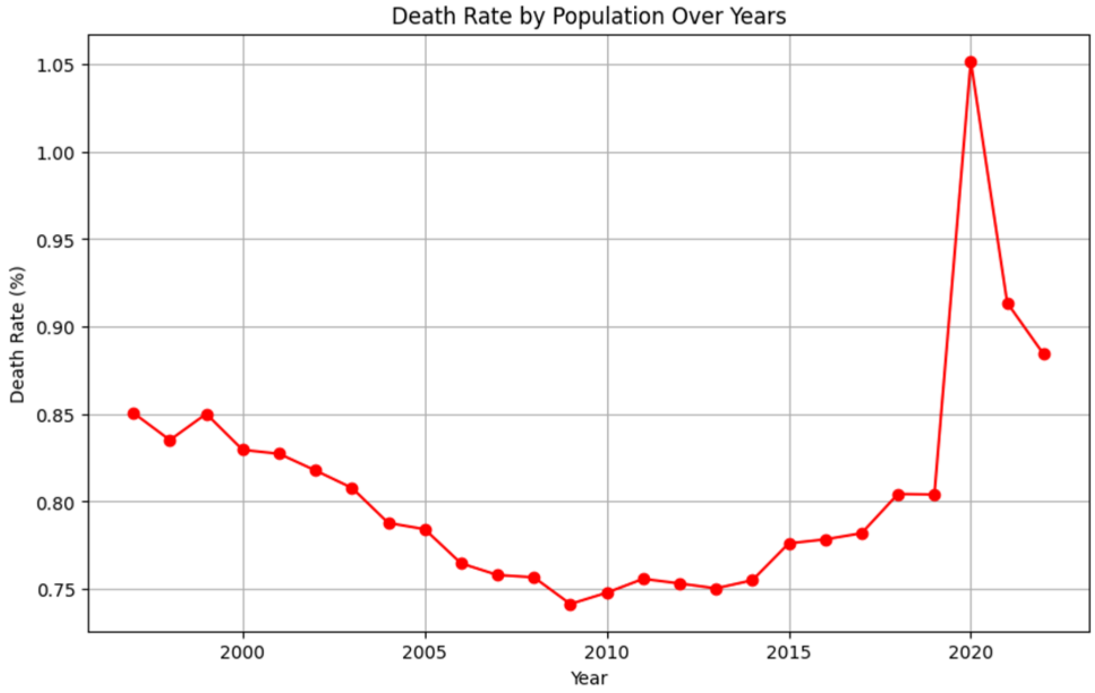
        - This trend suggests an initial period of lower mortality (possibly due to public health improvements or demographic changes) followed by an upward trend that likely reflects factors such as an aging population and shifts in health care dynamics over time.

- **Chart 2: Distribution of Death Counts – KDE Density Plot**



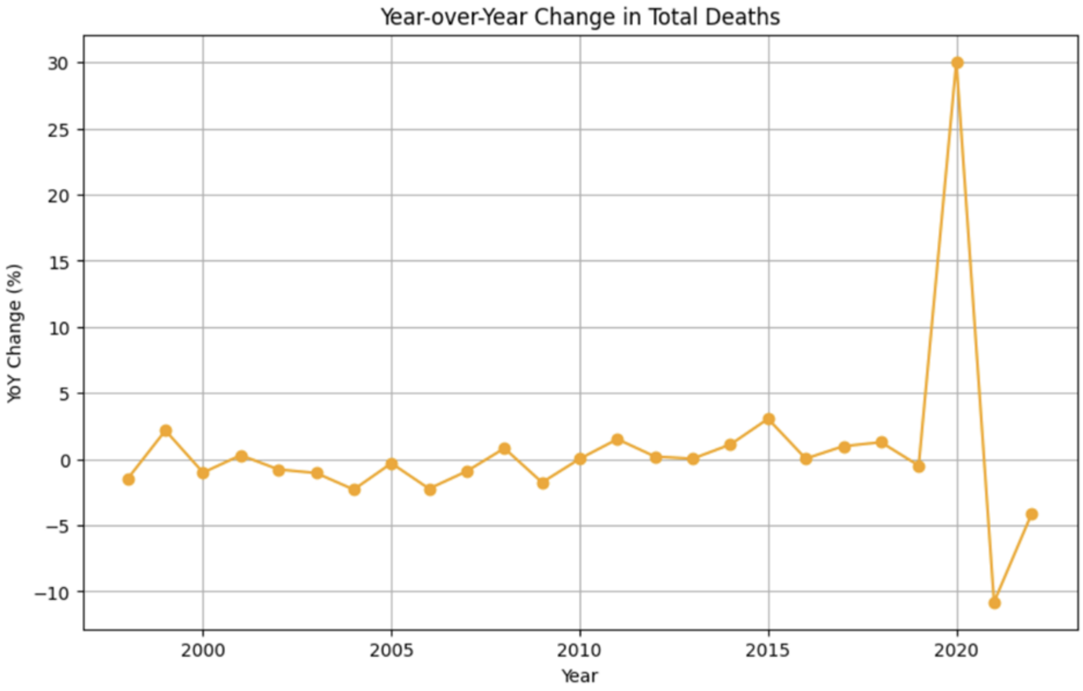
    - Observation:
        - The KDE density plot indicates that most annual death counts are concentrated around the 150,000 mark, suggesting a stable baseline during non-pandemic years.
    - Interpretation:
        - This stable central tendency provides a benchmark to compare against the dramatic deviations observed during the COVID-19 period. It confirms that, outside of the pandemic impact, the overall death counts in New York tend to cluster near 150,000.

- **Chart 3: Overall Mortality Rate**



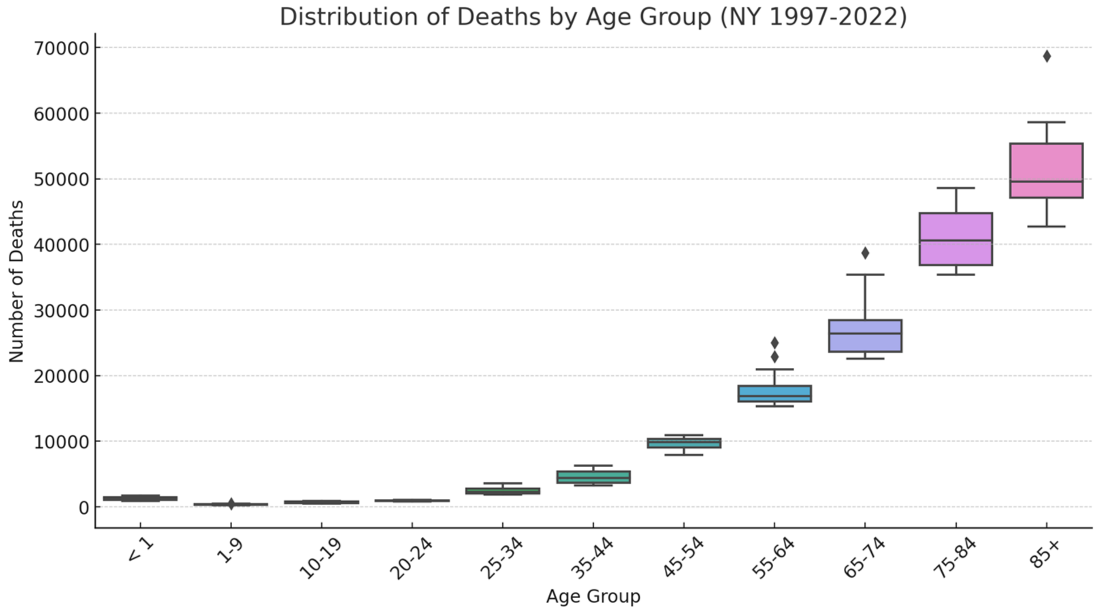
    - Observation:
        - Although the absolute death counts have hovered near similar levels (around 150,000) since the early 2000s, the mortality rate has been decreasing. This is evident in the death rate chart, which shows a downward trend over time.
    - Interpretation:
        - The decrease in the mortality rate, despite stable death counts, is likely due to population growth. With a larger population base, the same or similar number of deaths translates into a lower per-capita (or per 100,000) mortality rate. Contributing factors may include improvements in overall healthcare access and outcomes for certain demographics, even as challenges persist for others.

- **Chart 4: Year-over-Year Percentage Change in Total Deaths**



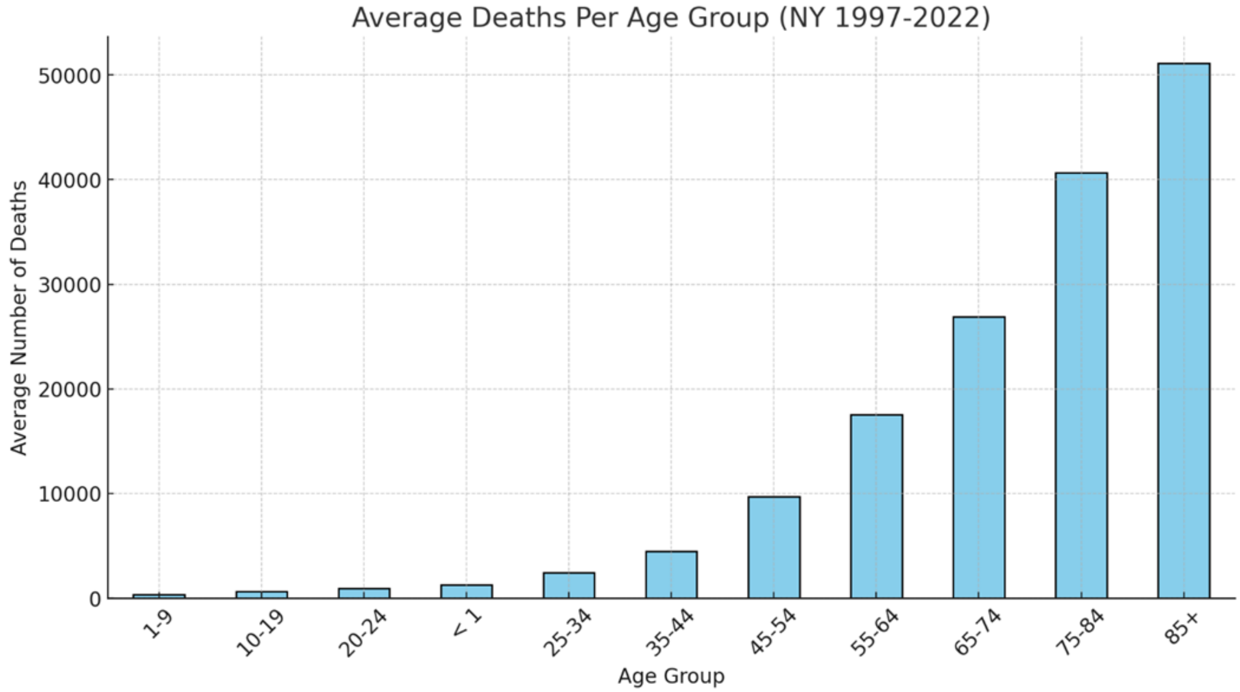
- Observation:
    - In most years, the year-over-year percentage change in total deaths remains close to a baseline, with moderate increases and decreases from year to year. However, there is a stark deviation during the COVID-19 period, where a nearly 30% increase is observed relative to the previous year.
- Interpretation:
    - This sharp, abnormal increase during 2020 highlights that the number of deaths in that period was in excess of what historical trends would predict, an excess mortality signal attributable to the pandemic's impact.

- **Age**



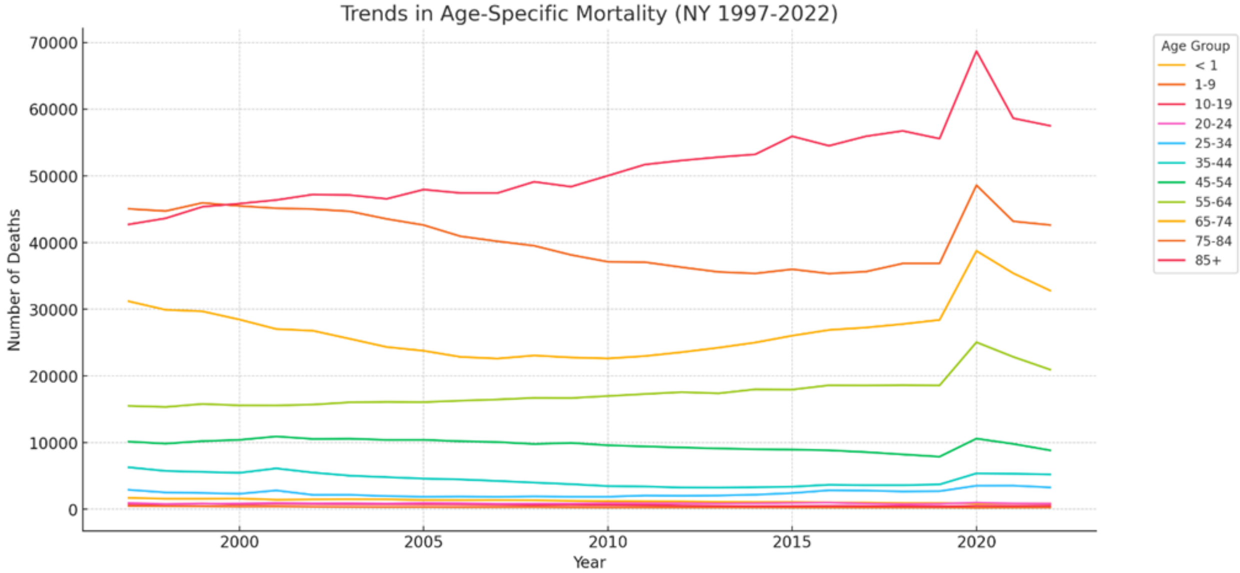
- Box & Whisker Plot of Distribution of Deaths by Age Group (NY 1997–2022):
  - Observation: New York's death distribution by age group from 1997 to 2022 is shown in the boxplot, which clearly and predictably shows an age-related trend in mortality. Throughout the 25-year period, the mortality rates for children and adolescents were reasonably constant and low, with younger age groups including <1, 1–9, and 10–19 continuously displaying low numbers of deaths with limited volatility. The number of fatalities steadily rises starting with the 25–34 age group, but this rising tendency becomes noticeably more prominent in senior age groups, especially those aged 55–64 and beyond. The 65–74, 75–84, and particularly the 85+ group had the highest number and distribution of mortality, with a median of about 50,000 deaths and outliers as high as 70,000.
  - Interpretation: Given that older persons have exponentially larger mortality risks, this distribution emphasizes the fact that age is one of the best indicators of mortality. The large interquartile ranges and greater volatility in older groups imply that senior mortality is more susceptible to outside health emergencies and demographic changes, such as the aging population. The close clusters in younger age groups, on the other hand, show more stability over years and lower death rates. All things considered, this graph emphasizes the

urgent need for focused public health initiatives meant to safeguard senior citizens, particularly in times of crisis like the COVID-19 outbreak, when they are most at risk.



- Bar Chart of Average Deaths Per Age Group (NY 1997–2022):
  - Observation: The average annual death toll for each age group in New York from 1997 to 2022 is shown in the bar chart. There is a definite exponential trend that shows that average deaths rise progressively with age. The lowest average yearly fatalities are seen in the youngest age groups, which are 1–9, 10–19, and 20–24. All of these groups dip considerably below 2,000. Since newborns are more vulnerable than other child and adolescent groups, the <1 group's statistics are somewhat higher. We see a slow rise that starts with the 25–34 age group and gets noticeably more noticeable in the 55–64 and older age groups. The elderly have a high mortality burden, as seen by the sharpest increases in the 75–84 and 85+ age categories, the latter of which average over 50,000 deaths annually.
  - Interpretation: The substantial relationship between age and mortality risk is shown by this distribution. The graph highlights the vulnerability of older people, especially when it comes to healthcare planning,

resource allocation, and pandemic preparedness, by confirming that they account for the vast majority of fatalities. Even while younger people make only a small portion of the total number of deaths, the dramatic rise observed in senior age groups emphasizes the necessity of ongoing public health assistance and focused initiatives for the elderly, particularly during emergencies like the COVID-19 pandemic. The number of people in high-risk age groups is gradually rising due to larger demographic trends such an aging population and increasing life spans.

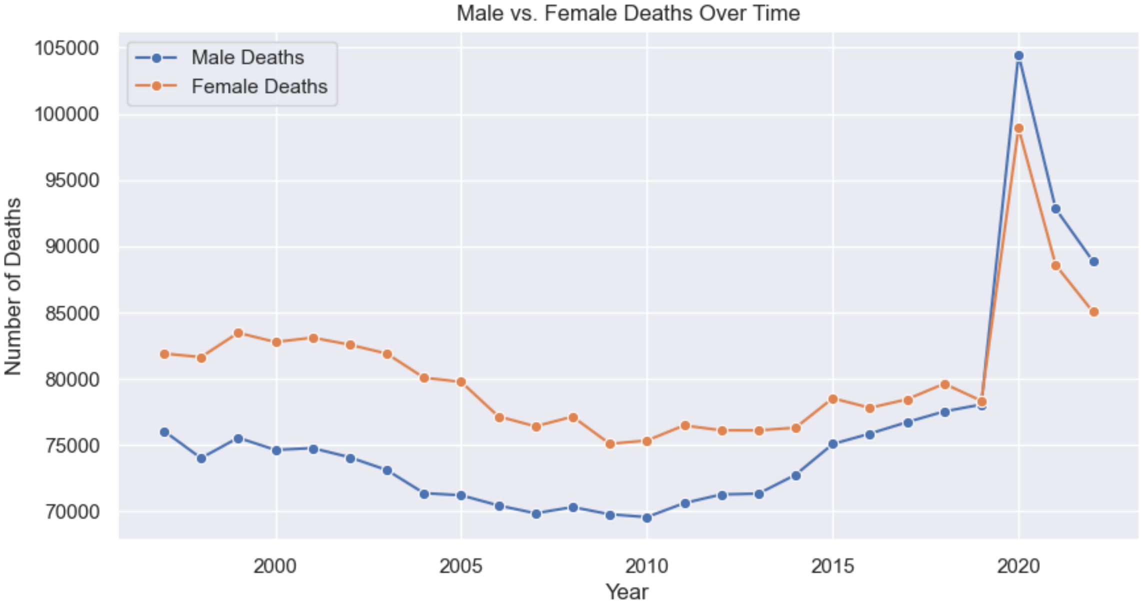


- Line Chart for Trends in Age-Specific Mortality (NY 1997–2022):
    - Observation: The yearly trends in the number of fatalities in New York across different age groups from 1997 to 2022 are depicted in this line graph. The image reveals several important trends. The most obvious pattern is the steadily high mortality rate among the 85+ age group, which not only holds the top spot over the time frame but also sees a dramatic increase around 2020, reaching a peak of about 70,000 fatalities at the height of the COVID-19 epidemic. Likewise, the 75–84 and 65–74 age groups show comparatively high and steadily rising mortality rates over time, with notable spikes starting in 2020. These increases, which show how the virus disproportionately affects older persons, are obviously consistent with the start of the pandemic.Younger age groups, on the other hand, such those aged 1–

9, 10–19, and 20–24, continue to have low death rates and little variation over time. The 55–64 and 45–54 age groups showed a minor downward trend from the late 1990s to the early 2010s, but mortality rates started to increase once more in the middle of the decade, especially for the 55–64 age group, until reaching a high point in 2020. This is probably a result of both pandemic-related mortality and population aging.

- Interpretation: In summary, this graph emphasizes two key conclusions: first, that mortality risk is primarily influenced by age, particularly during health emergencies; and second, that the COVID-19 pandemic significantly altered mortality trends in almost every age group, but particularly in the elderly. Age-stratified models for predicting and intervention methods are supported by these findings, which also highlight the significance of age-specific planning and resource allocation in public health response efforts.

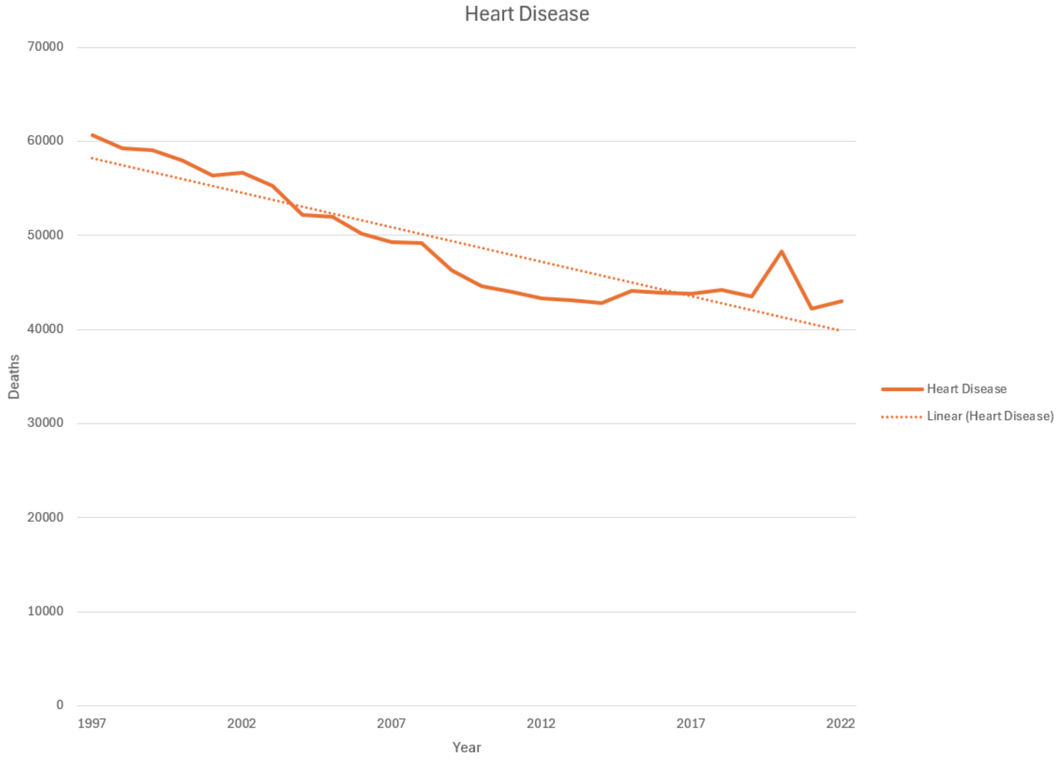
- **Gender**
  - **Male vs. Female Deaths Over Time**



- Observation:

            - From 1998 to 2010, there was a steady decline in the number of deaths of both genders, with female deaths consistently exceeding those of males. After 2010, deaths for both genders began to slowly, but steadily increase. Around 2020, deaths surged drastically from around 78,000 to over 100,000.
        - Interpretation:
            - The consistent higher number of female deaths compared to males is likely due to a higher population of females. The sharp increase corresponds closely with the COVID-19 pandemic, indicating a significant impact on both genders, with males experiencing a notably sharper increase, which suggests a potential higher vulnerability to severe outcomes from COVID-19.

- **Cause of Death**
    - **Chart 1: Heart Disease Death Counts Over Time**

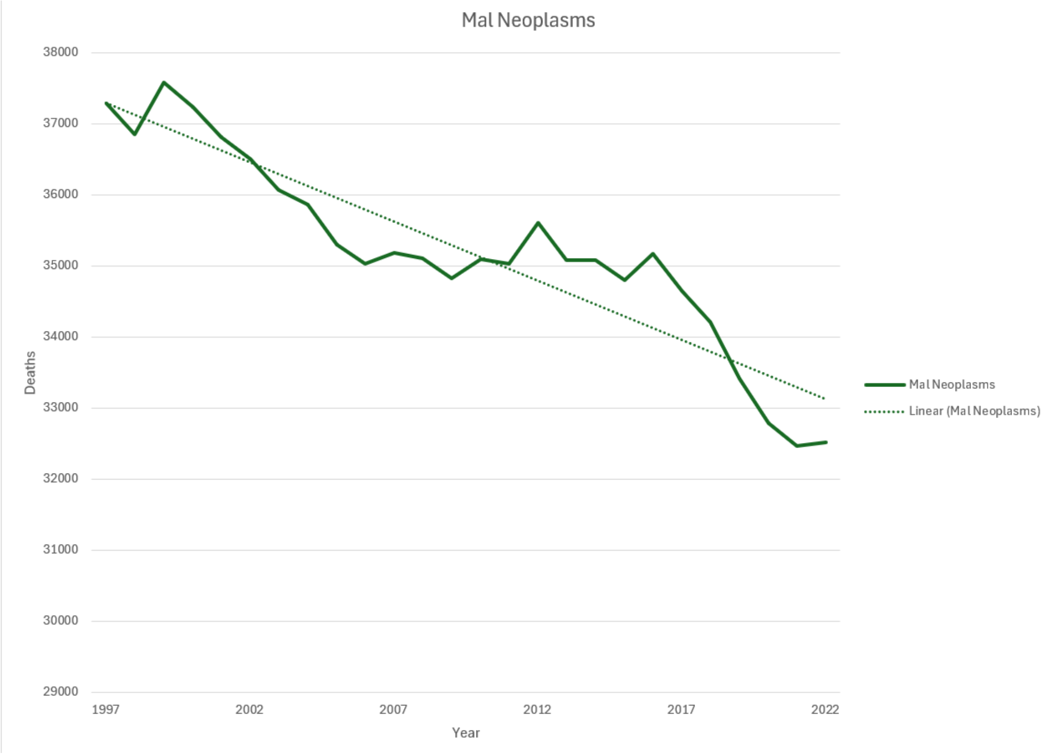


        - Observation:
            - We notice that heart disease has a steady decline in deaths during the 15 years of our data from about 60000 in 2000 to about 43500 In 2015. However, in recent years, a slight uptick

in deaths has appeared after 2015 compared to the line of best fit.

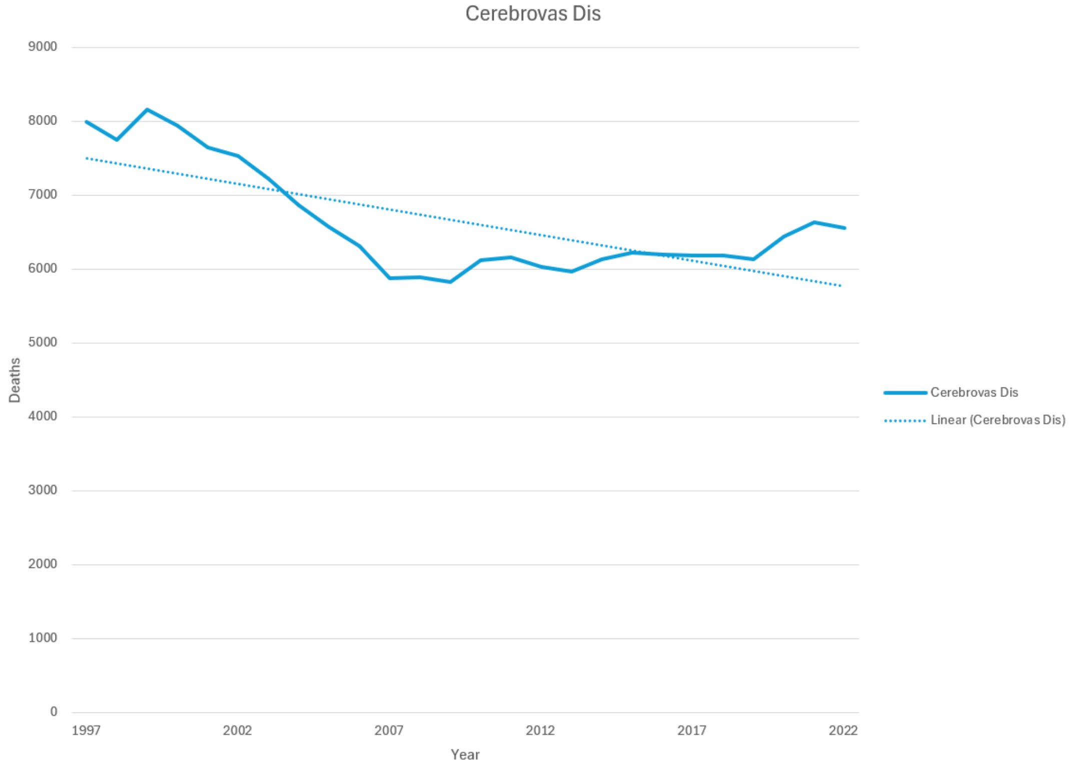
- Interpretation:
    - Heart disease is one of the biggest killers in the United States, so it makes sense that as time goes on, research on how to treat the disease has greatly decreased the amount of people dying from the disease. As treatment becomes more mainstream, it becomes more accessible to those who may not have top-tier health coverage. Although we do see the curve flattened out in later years, this could be due simply to the increase in population between the time periods.

- **Chart 2: Malignant Neoplasm Death Counts Over Time**



- Observation:
    - Malignant Neoplasms is another term used for a malignant, cancerous tumor. Over time, there is a sharp decline in the number of deaths cancer takes, although between 2007 and 2017 we observe a relatively consistent number of deaths per year.

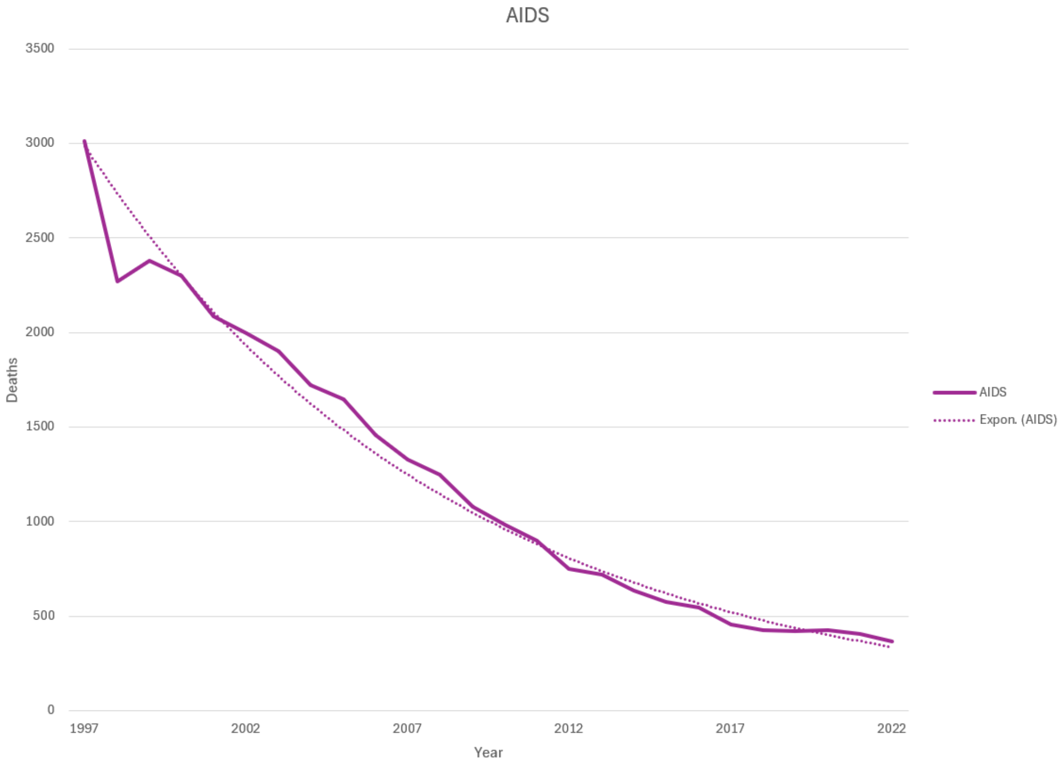
    - Interpretation:
        - Cancer is one of America’s most feared diseases as hospitals have erected dedicated wings just to research and study it. Because of population fear, this disease is one of the most studied diseases in the nation and thus treatment for it would become greatly better as time goes on, leading to less deaths by cancer.
- **Chart 3: Malignant Neoplasm Death Counts Over Time**



    - Observation:
        - Between the years of 2000 to 2008 there is a great decrease in the number of deaths caused by cerebralvascular diseases. Yet, after this time we can observe a slow, yet consistent increase from 2009 onward.
    - Interpretation:
        - Cereberalvascular disease refers to diseases related to blood and the brain which include, but are not limited to, strokes, brain bleeds and brain aneurysms. Although there are some

medications on the market to help prevent these events from happening, these deaths are a lot more spur of the moment when compared with other diseases, so we expect that as more of the population grows older, more of these deaths are likely to occur even if our trendline suggests otherwise.

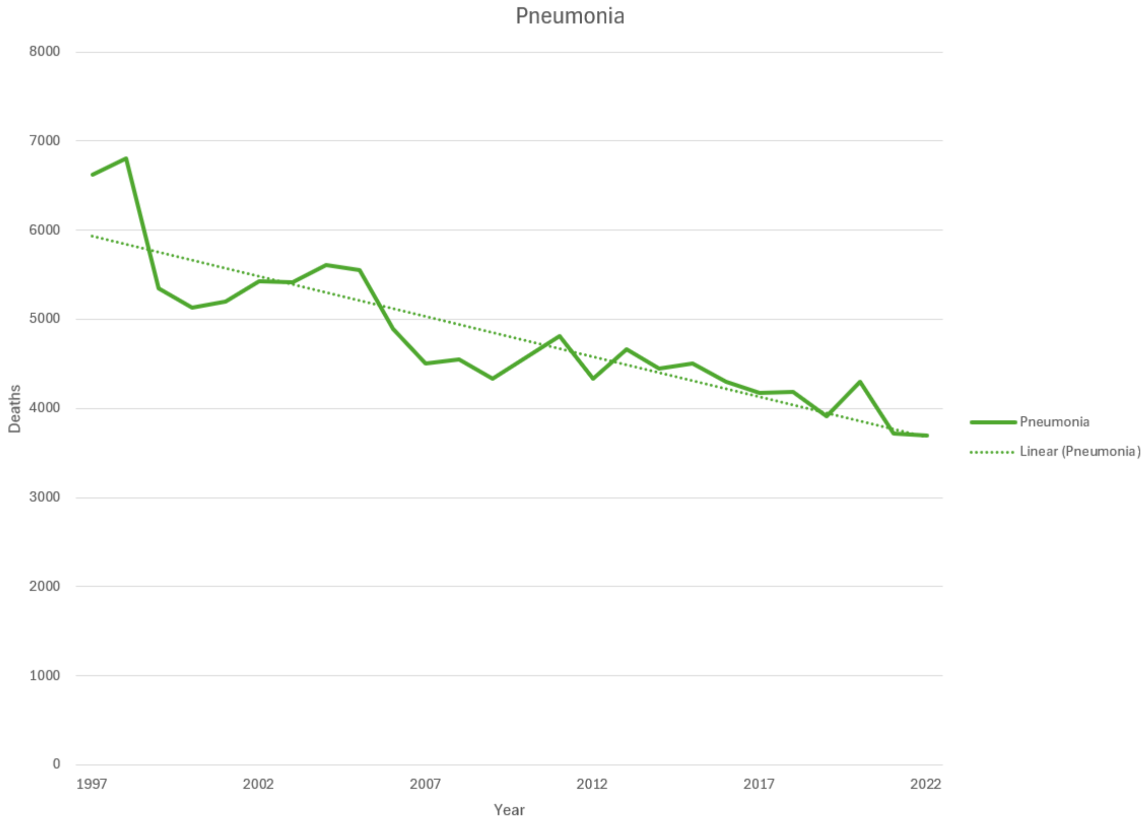
- **Chart 4: AIDS Death Counts Over Time**



  - Observation:
    - Over time we see a relatively steep negative trend in terms of deaths related to AIDS that fits incredibly well with an exponential trendline from the years 2002 up to 2022. The number of deaths has even fallen below 1000 per year every year since 2010.
  - Interpretation:
    - Although AIDS has not been cured, developments in HIV have allowed many people with HIV achieve remission of the virus. Because of these developments in HIV, less people are getting AIDS and in turn less people are dying from AIDS. Even with

AIDS, we now have a greater understanding of how to effectively live with the disease compared to twenty years ago.

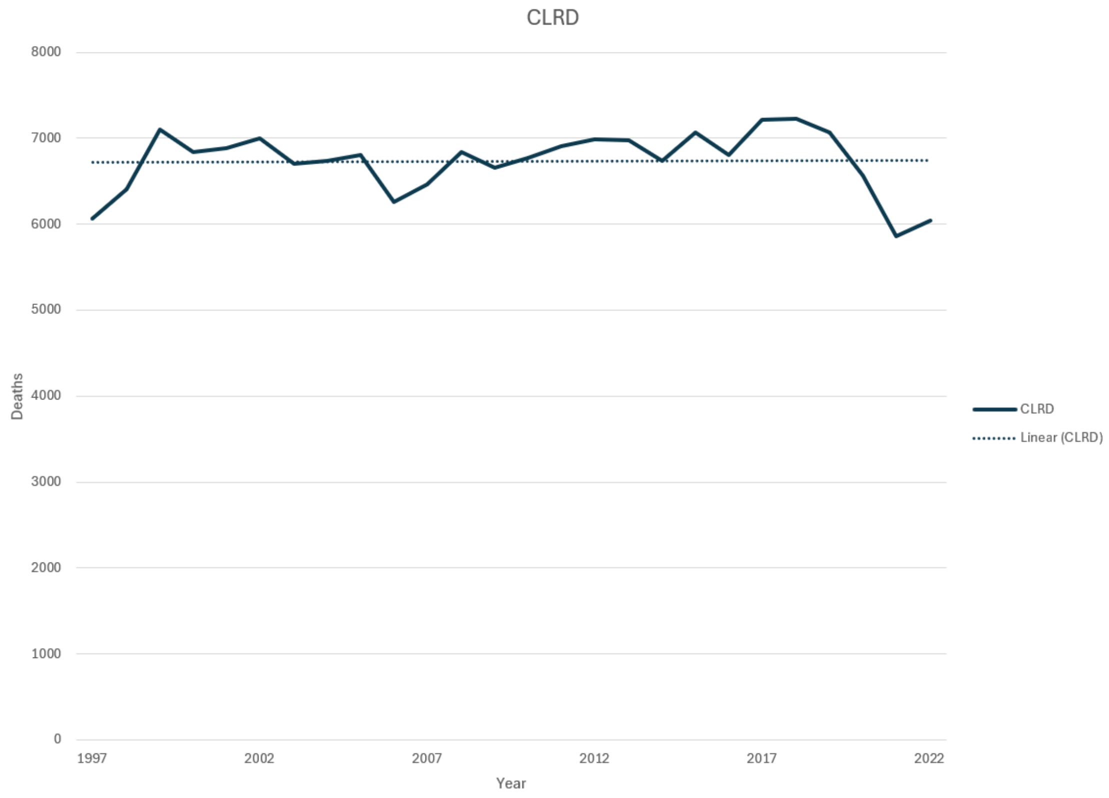
- **Chart 5: Pneumonia Death Counts Over Time**



  - Observation:
    - The number of deaths due to pneumonia does not appear to follow a distinct pattern at certain time frames like other causes of death. However, there is a noticeable downward trend starting from 1997 toward the present which is represented by the trendline.
  - Interpretation:
    - As time moves on we do expect advances in some of America's most common diseases to become more sophisticated and effective, which is what we see in the data for pneumonia. Although pneumonia is more likely to take the life of someone who is older, we could see those who are older

in age dying of other diseases or getting more effective treatment when they come down with pneumonia.

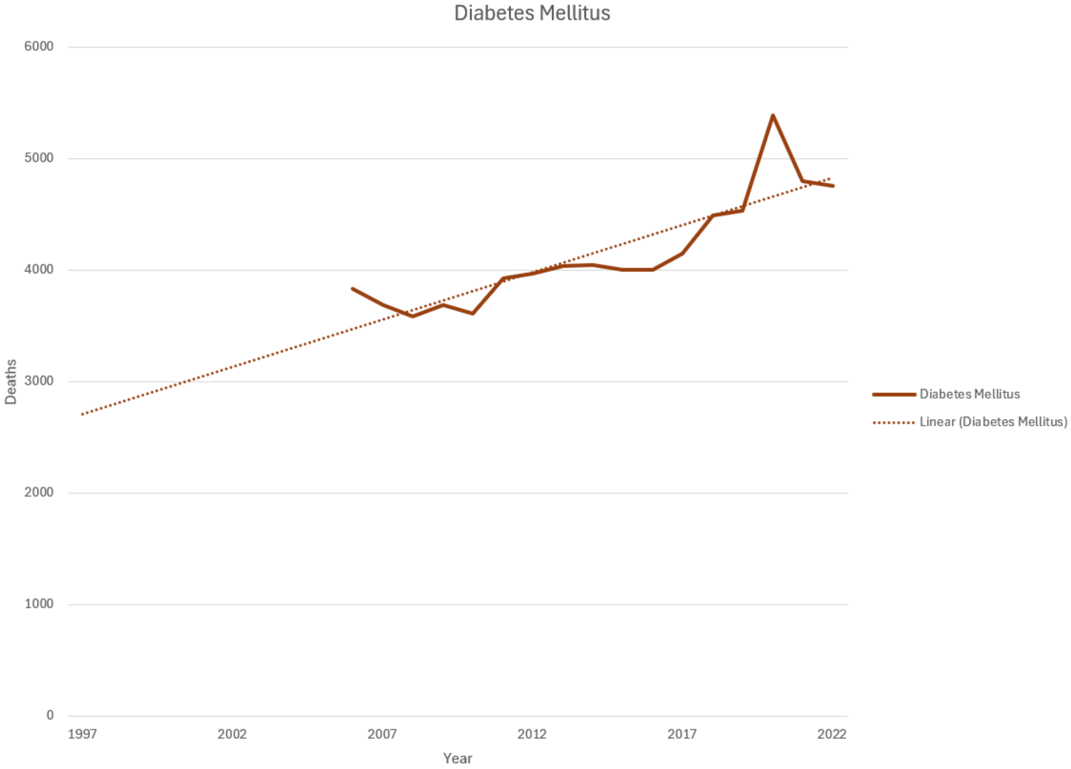
- **Chart 7: CLRD Death Counts Over Time**



- Observation:
  - The number of people dying from chronic lower respiratory diseases over time is eerily consistent as shown by our trendline. The only big dips we see are from 1997 to 199 and after 2020, otherwise the total deaths we expect is about 6800 a year.
- Interpretation:
  - Chronic lower respiratory diseases is an all-encompassing term that includes diseases such as asthma and COPD. Some of these diseases are genetic and some can be developed later in life through unhealthy habits such as smoking, which is probably why we see the diseases being so consistent in

deaths. The dip near the end may be caused by those with weak lungs dying from COVID-19 instead of their CLRDs.

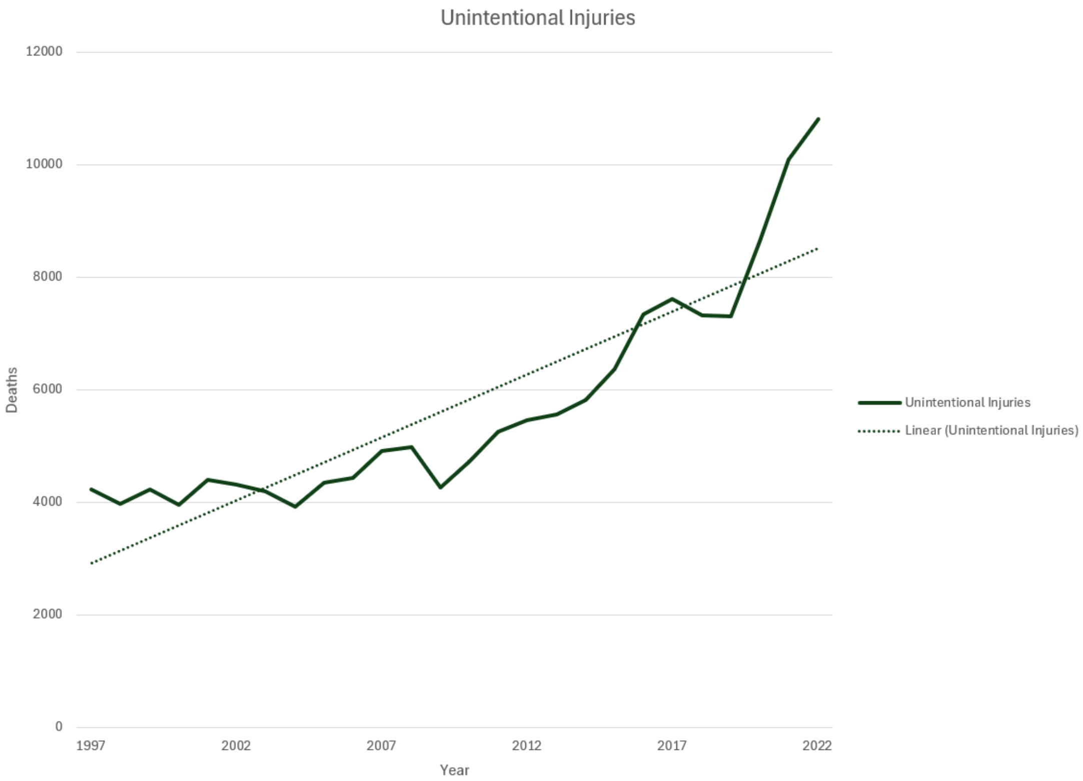
- **Chart 8: Diabetes Mellitus Death Counts Over Time**



  - Observation:
    - Although data for diabetes does not reach back until 1997, we do see a positive upward trend that was not seen in many of the other causes of death in diabetes, especially with a spike relative to the number of deaths in 2020.
  - Interpretation:
    - Diabetes is sometimes genetic, so when the population increases that means more people are prone to contracting diabetes and thus dying from it. However, unhealthy lifestyle choices can also lead to diabetes as well as the obesity rate has increased steadily over time (NIDDK).

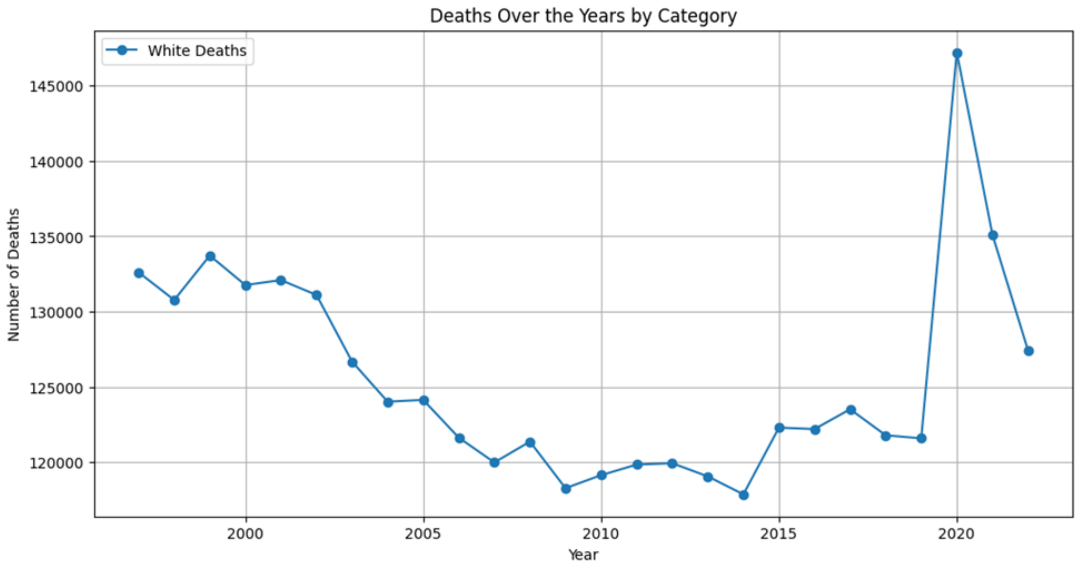
- **Chart 9: Accidental Death Counts Over Time**

Muzahidul Islam, Saki Takatsu, John Harrison,
James Soltis, Isabel Pacheco Mattivi
DAT400 Spring 2025



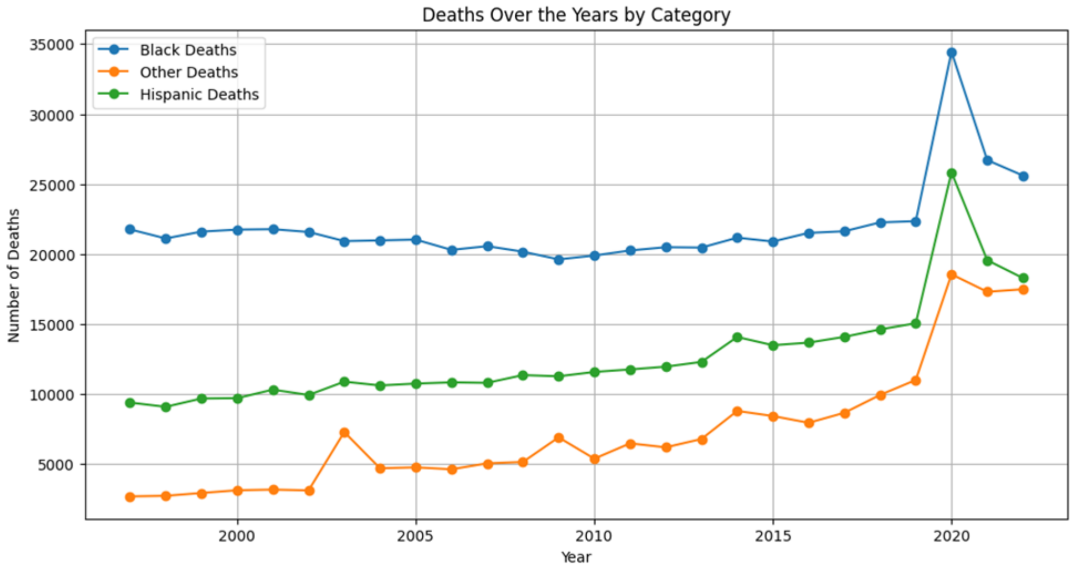
- Observation:
    - From the years 1997 to 2010 there was a steady number of accidental deaths, but after that we see a steep increase in the number of people that lose their lives in accidents. This is especially apparent during the 2019-2022 period.
- Interpretation:
    - The number of accidental deaths is intuitively related to the population increase as more people alive means more accidents can happen. However, this recent steep uptick in the 2019-2022 period could be related to the COVID-19 quarantine where people try new things at home that are dangerous such as home repairs that lead to their deaths.

- **Race & Ethnicity**
    - **Chart 1: White Death Counts Over Time**



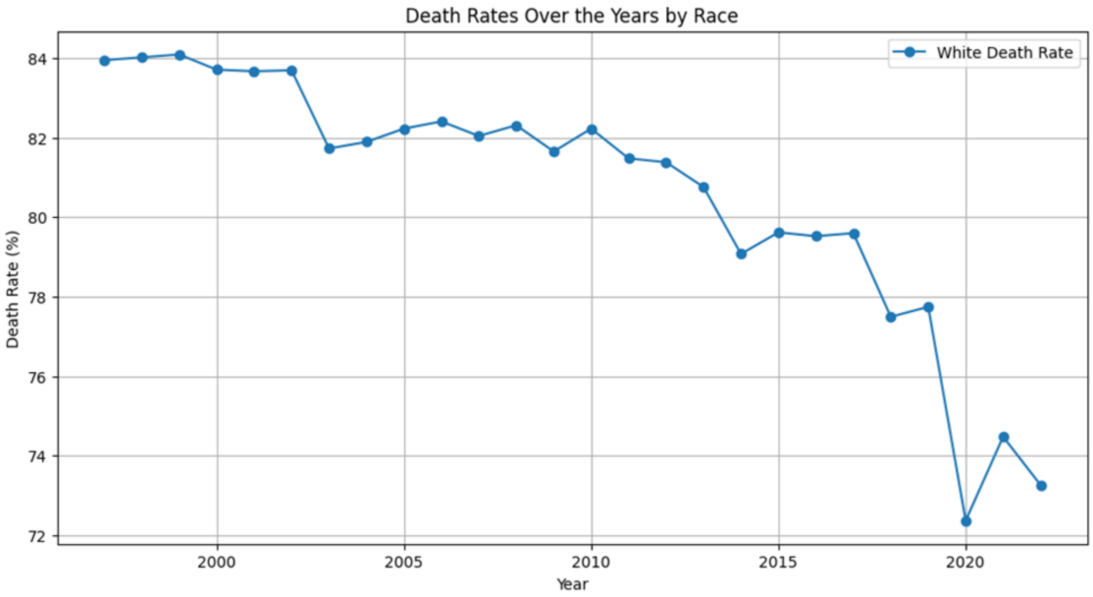
- Observation:
    - The chart shows a decrease in total death counts for Whites, from about 132,000 in 2000 down to around 118,000 by 2010, followed by an uptick starting near 2015.
- Interpretation:
    - This downward trend until 2010 may indicate improvements in healthcare and living conditions among the white population or demographic shifts. The post-2010 increase could reflect other factors such as changing age structures or emerging health challenges.

- **Chart 2: Black, Hispanic, and Other Death Counts Over Time**



- Observation:
  - In contrast to Whites, the death counts for Black, Hispanic, and 'Other' race groups have generally increased over time. There is a pronounced spike during the COVID-19 outbreak; after 2020, the counts remain high, especially for the less-represented racial groups.
- Interpretation:
  - This suggests that these groups have experienced a cumulative adverse impact from the pandemic, perhaps due to underlying health disparities, socioeconomic factors, or unequal access to healthcare.

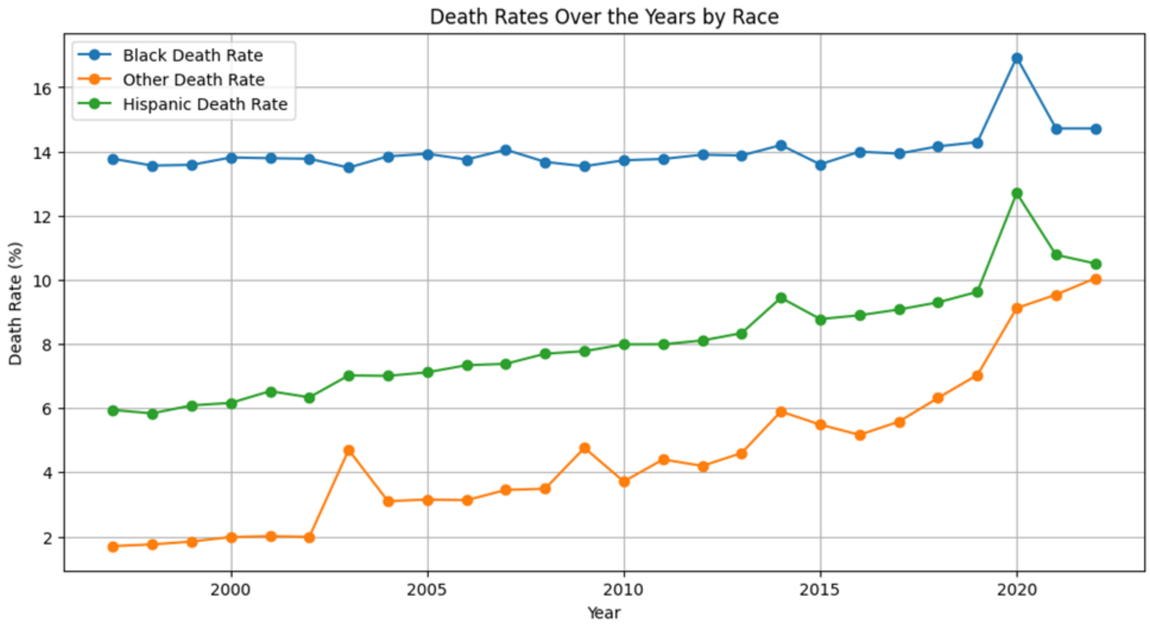
- **Chart 3: White Mortality Rate Over Time**



- Observation:
    - White mortality rates, calculated as the number of deaths relative to population size, have steadily decreased over the same period.

- Interpretation:
    - The reduction in the white death rate might be attributed to improved healthcare outcomes for Whites, coupled with demographic changes where the proportion of Whites in the population may be declining relative to other groups.

- **Chart 4: Mortality Rates for Black, Hispanic, and Other Groups Over Time**



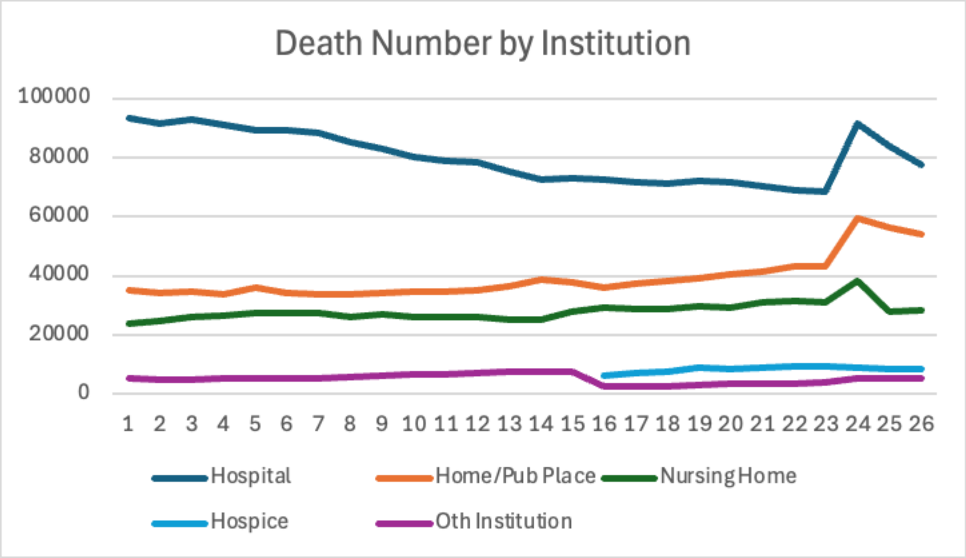
- Observation:
    - The death rate for Black individuals appears relatively stable over time. However, for Hispanics and the ‘Other’ category, the mortality rates have shown an upward trajectory, especially with a noticeable climb post-2020 COVID.

- Interpretation:
    - The sustained increase in mortality rates for Hispanics and Others, particularly after the initial COVID spike, indicates that these groups continue to suffer disproportionately, highlighting persistent inequities in healthcare access, socioeconomic status, or underlying health risks.

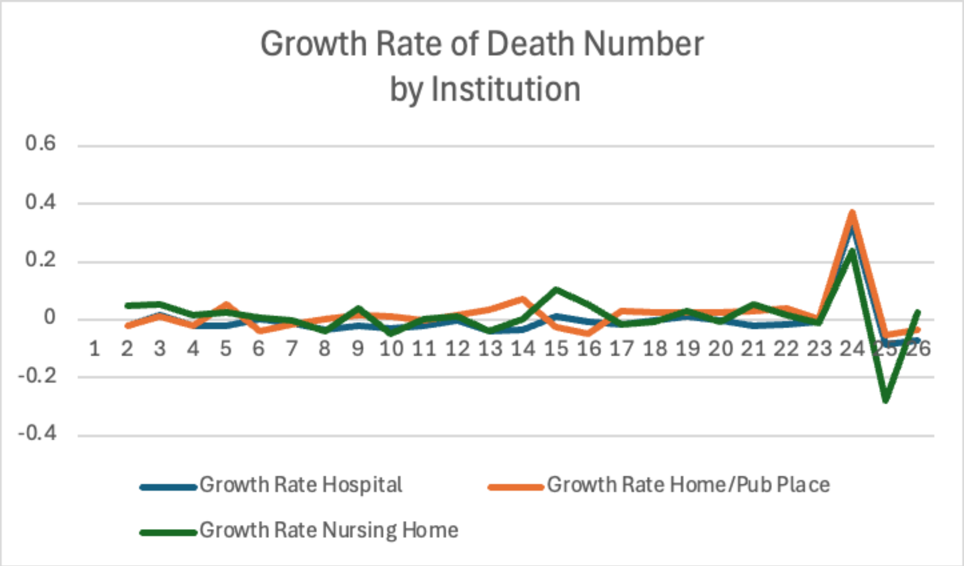
- **Overall Insights**
    - Differential Trends:
        - While Whites show an initial decline in both counts and death rates, followed by a reversal, the minority groups (Black, Hispanic, and Others) not only have increasing absolute death counts but also show adverse trends in mortality rates post-COVID.
    - Implications:

- These differences underscore the importance of accounting for demographic shifts and inequities when estimating excess mortality. The elevated death counts and sustained high mortality rates among minority groups suggest that the pandemic's impact was not uniform but exacerbated existing disparities.

- **Institutions**
  - **Chart1: Death Number by Institution**

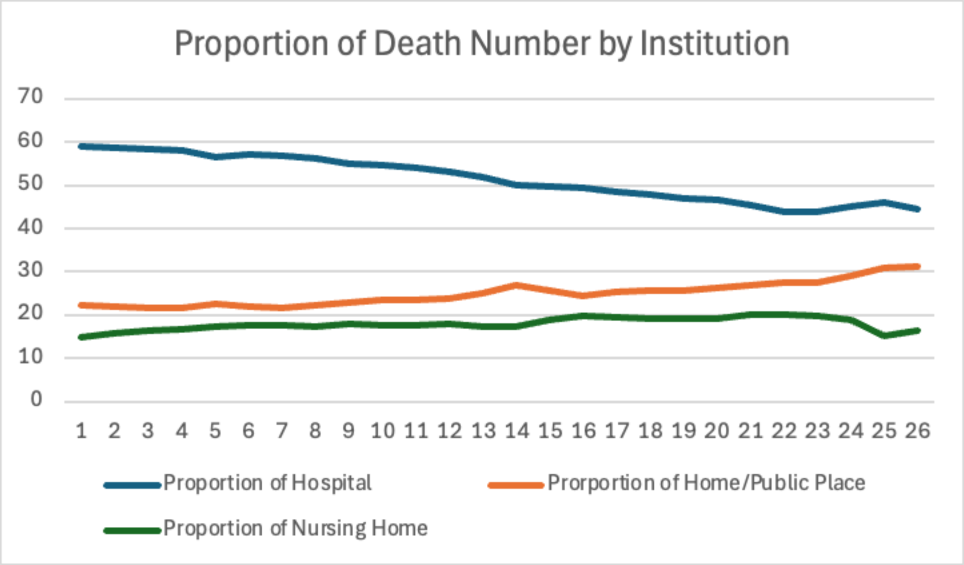


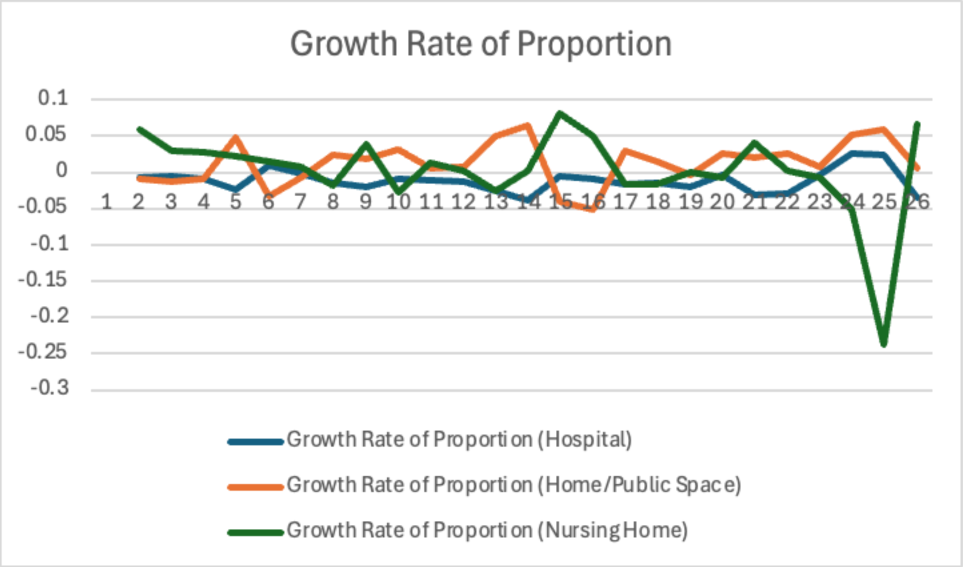
  - **Observation**
    - There was a biggest growth in the number of death in Home/Public Place (+37.15%) and Hospital(+33.43%), the institution that had the least growth in number was nursing home(+23.68%).
    - There was less than 5% change for Hospice and Other institution.
  - **Implication**
    - The sharpest increase in deaths occurred in homes/public places and hospitals, suggesting that the healthcare system may have been overwhelmed during peak periods, leading to more individuals dying before reaching or after being discharged from medical care. In contrast, the smallest increase was observed in nursing homes, which may reflect earlier containment efforts or reduced admissions during the pandemic.
  - **Chart2: Growth of Death Number by Institution**



- **Observation**
  - After the jump in number of death in 2020, there was a drop among all the three institutions.
  - Hospital (-8.54%), Home/Public Place (-5.18%) and Nursing Home (-2.79%)
- **Implication**
  - This drop suggests a partial stabilization of the healthcare system and public health response after the initial COVID-19 surge. The relatively smaller decrease in nursing homes may indicate ongoing vulnerability in long-term care facilities, while the sharper declines in hospitals and homes/public places reflect improved access to care or the effects of vaccination and public health interventions.

- **Chart3, 4: Proportion of Death Number by Institution and Growth of Proportion**



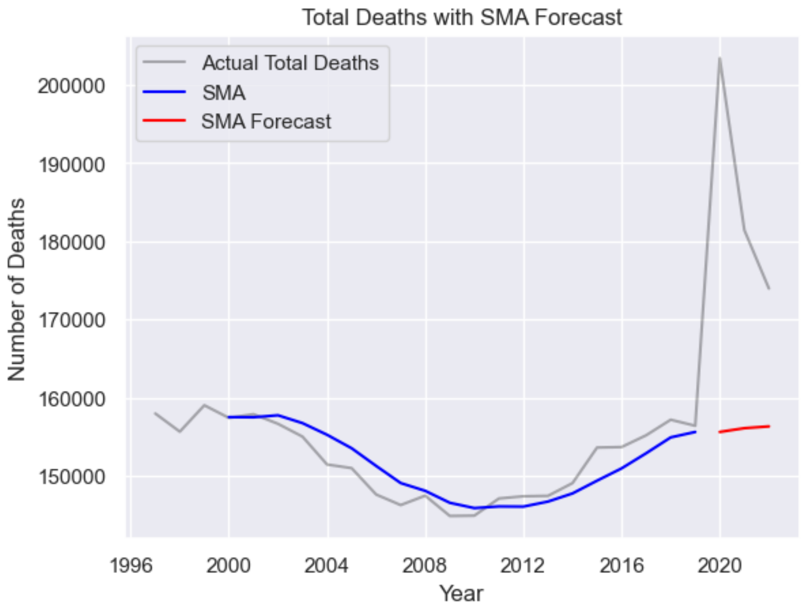


- **Observation**
    - There was growth in the proportion of the number of death in Hospital(+2.54%) and Home/Public Place (+5.18%). There was a drop in the proportion for Nursing Home(-5.15%).
- **Implication**
    - This suggests a shift in where people were dying during the pandemic. This trend may reflect changing care preferences, reduced nursing home admissions, or overwhelmed long-term care systems. The increase in home/public place deaths, in

particular, could indicate barriers to accessing institutional care or a desire to remain at home during a health crisis.

Model Building

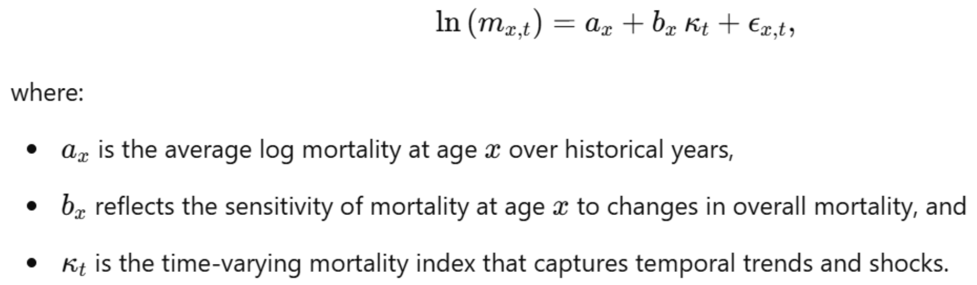
- **Moving Averages**
  - Simple moving averages (SMA) calculate the average of a target variable during a determined period of time. This model follows the following formula: SMA = $\frac{A_1+A_2+\cdots+A_n}{n}$, where A_n is the value at period n and n is the total number of periods. SMA smooths out noise in the data in order to see a clearer and more interpretable view of trends.



  - The graph above presents the actual total deaths from 1998 to 2022 alongside an SMA model calculated using a window (n) size of 4 years. Different windows of 3 to 7 years were tested initially, but a period of 4 years was chosen as the most optimal based on its strong fit to historical data. The historical data spanned from 1998 to 2019 and is used to predict total deaths for the years 2020, 2021, and 2022. The SMA model (in blue) effectively captures the historical trends in our mortality data, smoothing fluctuations. However, starting in 2020, the actual death totals diverged sharply upward from SMA predictions (in red), highlighting the significant impact of the

COVID-19 pandemic. The divergence clearly illustrates the unexpected nature of the pandemic, as actual deaths far exceeded the model's forecasts. Statistically, the model produced an R^2 of approximately 0.7688, indicating a strong fit to historical data. The RMSE of 2,145.747 deaths suggests the model typically predicts total deaths with moderate accuracy in years unaffected by COVID-19.

- **Lee-Carter Model**
  - The Lee–Carter model is a widely used demographic framework for modeling and forecasting mortality. It expresses the log of the age-specific mortality rate as:

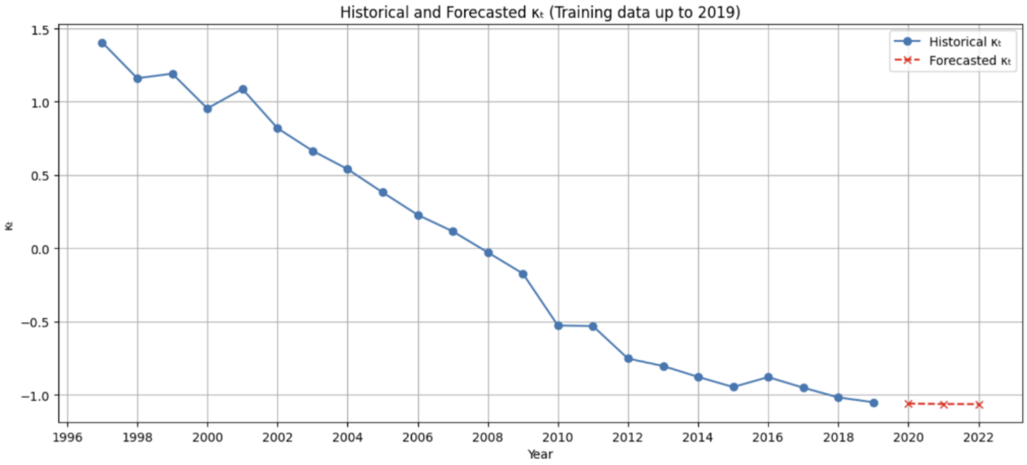


$$\ln\left(m_{x,t}\right) = a_x + b_x\,\kappa_t + \epsilon_{x,t},$$

where:

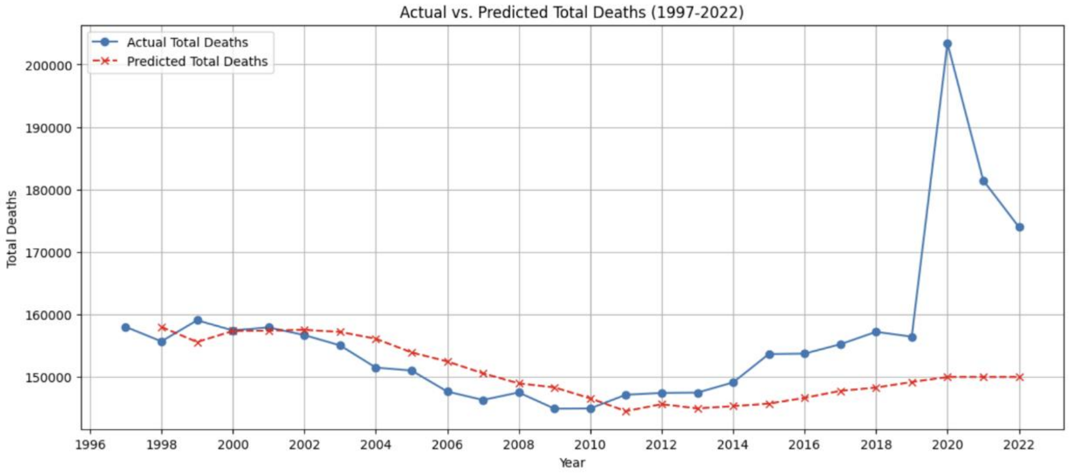
- a_x is the average log mortality at age x over historical years,
- b_x reflects the sensitivity of mortality at age x to changes in overall mortality, and
- κ_t is the time-varying mortality index that captures temporal trends and shocks.

    - In our analysis of New York's mortality data (with age-specific death counts from 1997 to 2022), a_x and b_x are estimated from historical data using singular value decomposition (SVD). The key parameter k_t summarizes the overall mortality level in each year. For instance, k_t starts around +1.4 in the late 1990s and declines steadily to approximately -1.0 by 2020. This downward trend indicates that, after accounting for age patterns, the baseline mortality improved over time until the disruptive effects of the COVID-19 pandemic.
  - **Chart 1: κ_t Trend and Forecast**



- This chart displays the historical values of k_t alongside their forecasted values. The plot shows that k_t started at around +1.4 in 1997 and decreased to near -1.0 by 2020, indicating an overall improvement (i.e., reduction) in baseline mortality over time. The forecasted k_t for subsequent years is derived solely from the trend observed up to 2019. The behavior of k_t is central to the Lee–Carter model because it captures the aggregate mortality trend that, when combined with age-specific patterns, produces the final mortality forecasts.

○ **Chart 2: Actual vs. Predicted Total Deaths**



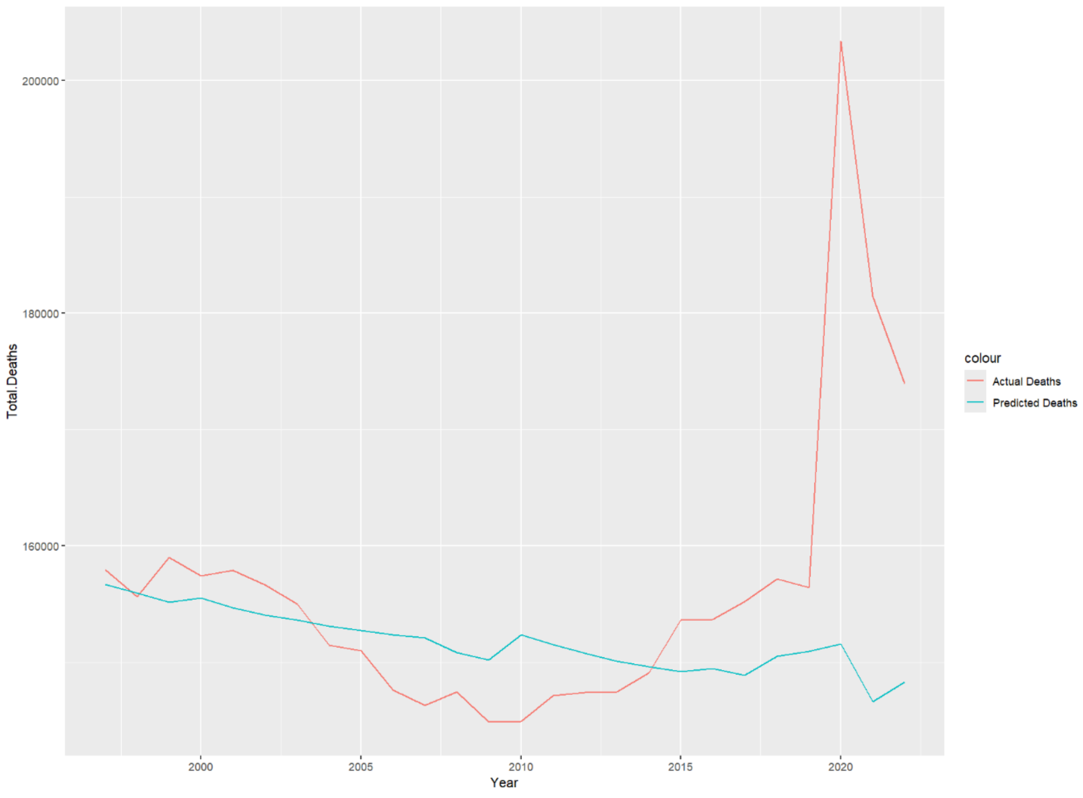
- This chart shows the actual total death counts for New York (from 1997 to 2022) alongside the predictions generated by the Lee–Carter model. The model tends to overpredict deaths in the earlier years and then shifts to underpredict later on. This pattern is reflected in our performance metric, with an RMSE of about 4,488 deaths for the backtest period (1998–2019). The RMSE provides a quantitative measure of the model's deviation from actual observed values and helps assess its predictive accuracy.

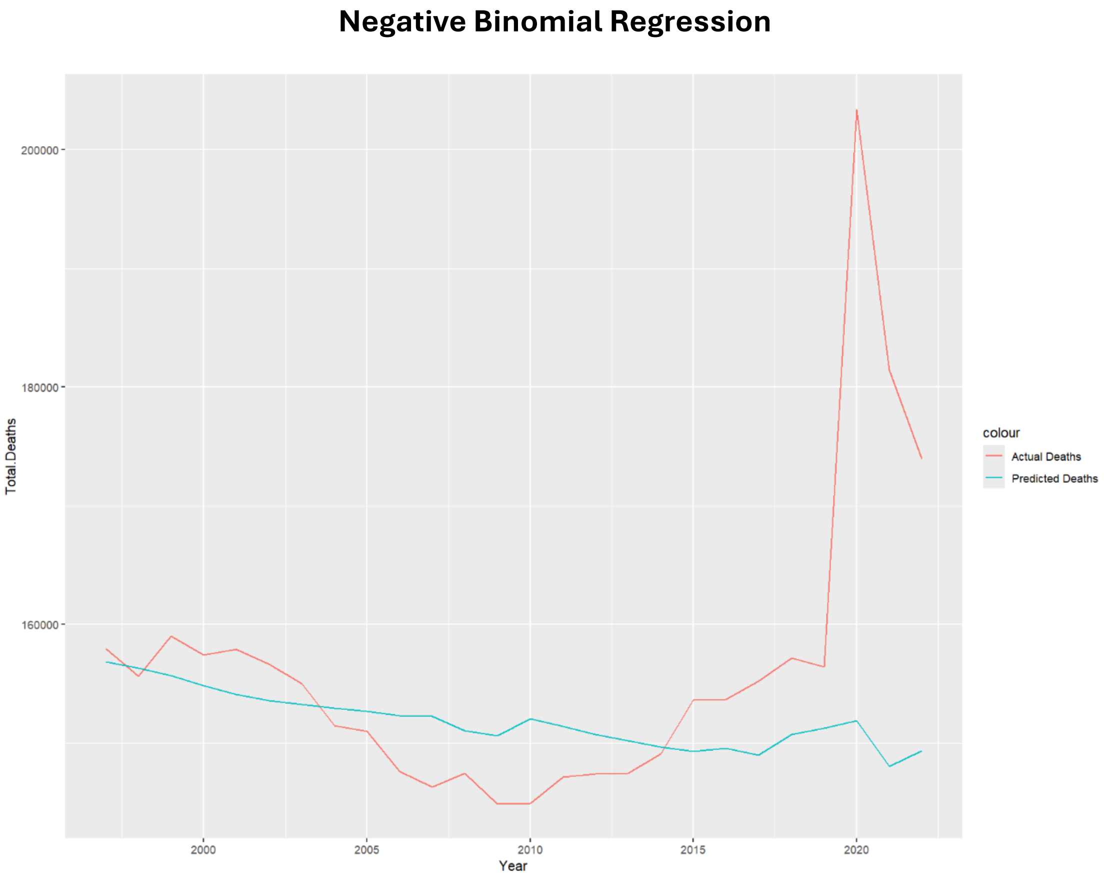
- **Poisson, Lasso, Negative Binomial**
    - Poisson and Negative Binomial regression are both similar in their analysis except for one key difference in variance. Both models are useful in modeling count data, which is what we were presented with when modeling the total deaths. The Poisson model exponentiates all the coefficients, meaning that the regression equation is:

      $$\boldsymbol{ln(y) = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n}$$

      The Poisson regression assumes the variance is equal to the mean, which the negative binomial does not assume and uses a dispersion parameter. Below are each of these models:

Poisson Regression

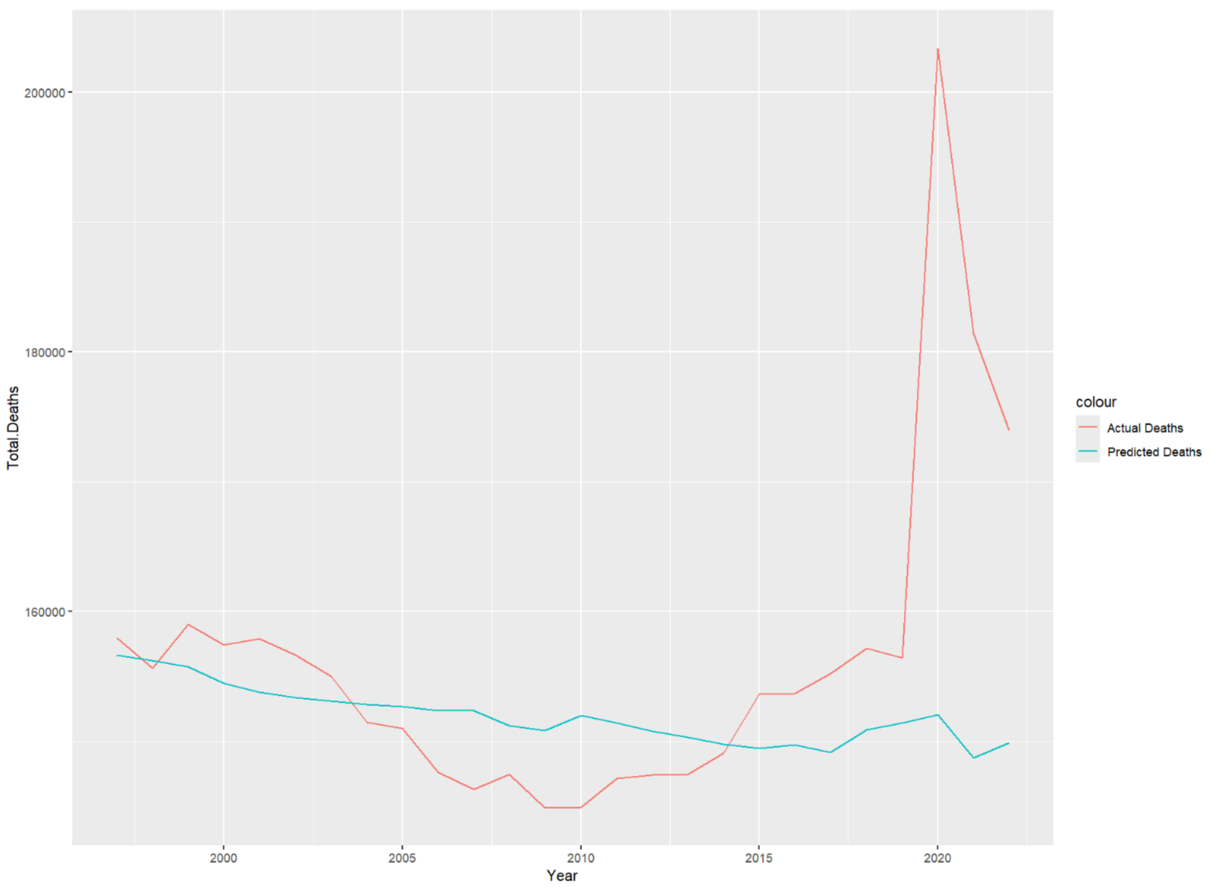




    - These two charts are very similar because they have the same limitation as the data. Unlike the other mentioned methods of modeling predicted deaths, both the Poisson and Negative Binomial regression relied on predictors to aid in their regression. However, the predictors in the data that were ready to use without transformation was total population, as every other variable were direct subdivisions of the total death variable. You can see the similarities in the RMSE (from 1997-2019) of the two models, as the Poisson RMSE was 4096.572 while the Negative Binomial RMSE was 4097.209 which is a difference of less than 1. The models themselves give an alright prediction for the average deaths among the years when they were trained, but they have trouble identifying the more minute details because they need more predictors to fine tune.

  - The LASSO model was also used to attempt to predict the total number of deaths. LASSO specifically penalizes overfitting with too many variables, which was hypothesized to be better for our data due to the number of predictors available for regression being mostly population based. In our case, the least squares method of LASSO regression was employed to produce the following prediction:

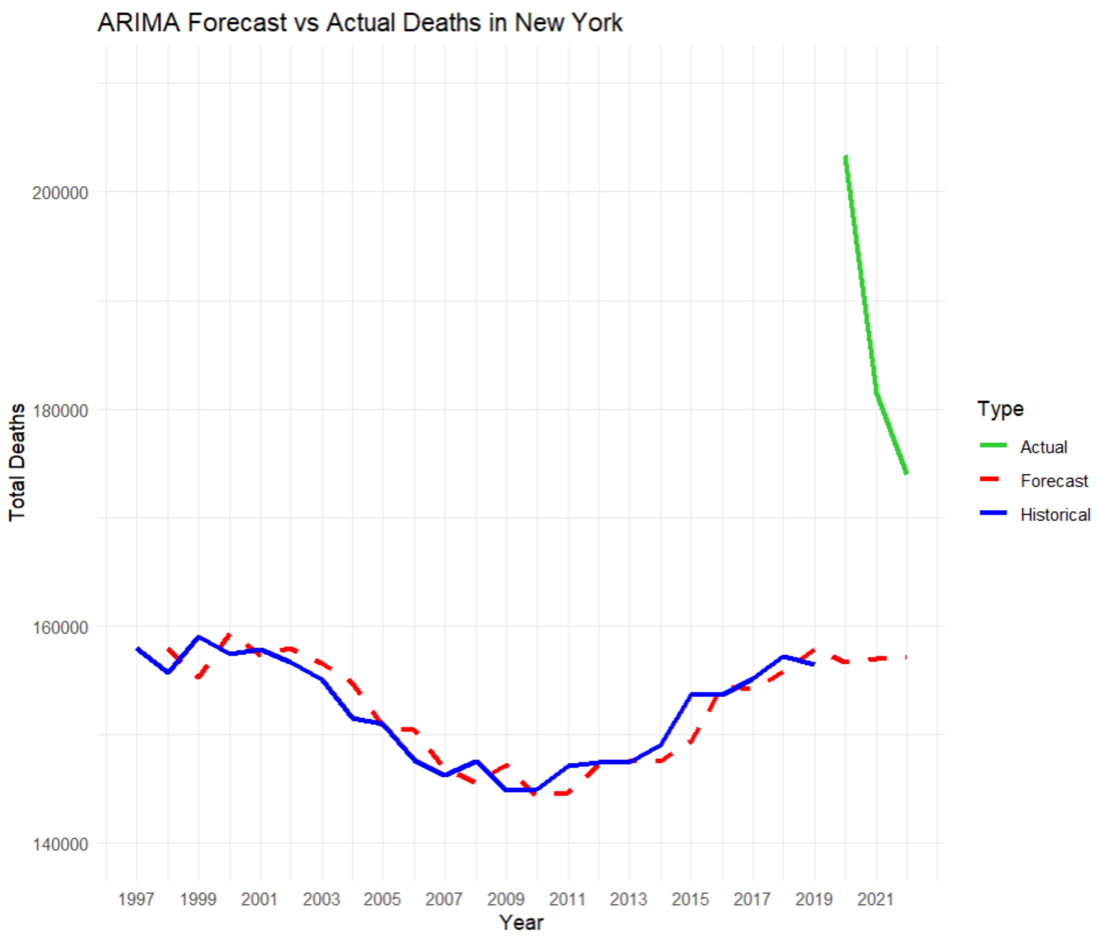
LASSO



- Similarly to the exponential and negative binomial, the LASSO regression model looks incredibly similar in terms of where the model overestimates and underestimates the data. However, the penalties that the LASSO regression applies did not seem to create a more accurate model concerning RMSE, as the LASSO (1997 to 2019) performed worse with 4116.8. In terms of total RMSE evaluation, the Poisson model would be the best out of the three but with correlated predictors not being used the models themselves were not as fruitful as hoped for.

- **ARIMA**

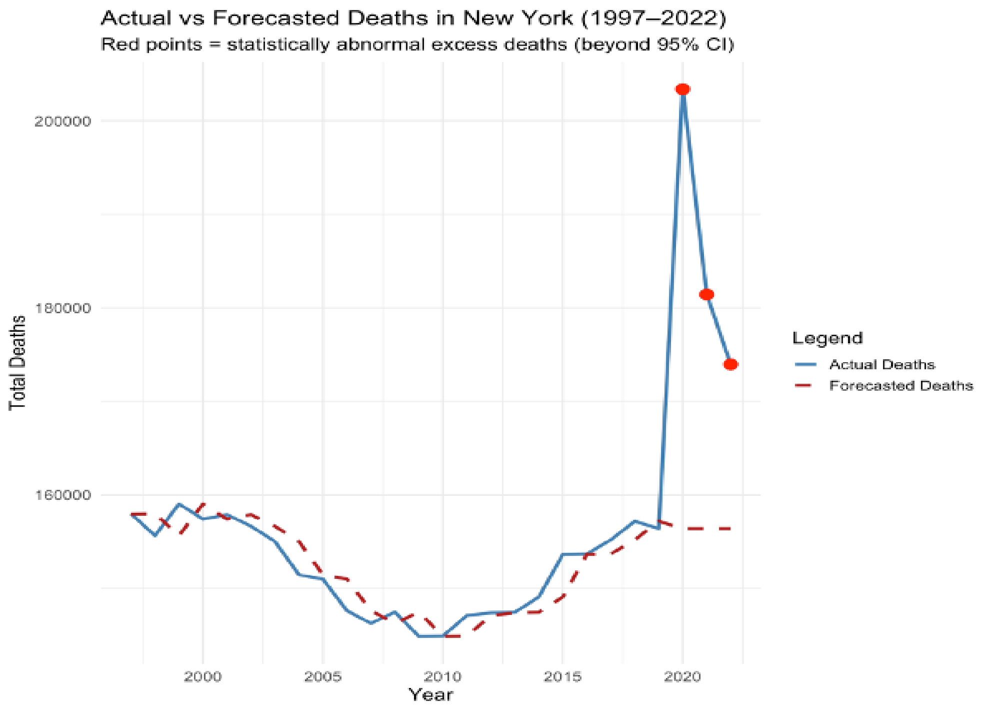
- Our ARIMA model parameters were selected to minimize AIC and BIC values after differencing the series to achieve stationarity, permitting the model to capture long-term mortality patterns. The ARIMA (AutoRegressive Integrated Moving Average) model was used to forecast total deaths in New York based on historical data from 1997 to 2019. ARIMA models are especially effective for time series with underlying trends and random fluctuations without the need for external predictors.
- The following figure shows the ARIMA forecast (red dashed line), the historical data used to train the model (blue line), and the actual total number of deaths (green line). The ARIMA model successfully captures the decreasing and stabilizing pre-pandemic death trends by closely adhering to the historical trend from 1997 to 2019. But beginning in 2020, real mortality far outpaced the model's prediction, underscoring the COVID-19 pandemic's unanticipated and disruptive character. The model's failure to predict previously unheard-of exogenous shocks is reflected in the divergence.



- A statistical match to pre-COVID historical data was demonstrated by the ARIMA model, which obtained an R2 of around 0.8131. During the training period, the model generally predicted total fatalities with good accuracy, as indicated by the RMSE of 1,995.36 deaths. In light of the unusual spike brought on by COVID-19, RMSE is a particularly useful measure of the magnitude of the deviations when reality deviated significantly from historical trends.

- **Exponential Smoothing**
  - The Exponential Smoothing (ETS) model forecasts time series data by applying greater weight to more recent observations, making it responsive to change. ETS models consist of error (E), trend (T), and seasonality (S) components, which can be additive, multiplicative, or absent.
  - In this analysis, an ETS(M,N,N) model—multiplicative error, no trend, no seasonality—was fitted to New York's total death data from 1997 to 2019.

This structure assumes a stable level over time, with error size proportional to the level of deaths. While effective for stable historical patterns, the model is less suited to forecasting sudden disruptions like the COVID-19 pandemic.



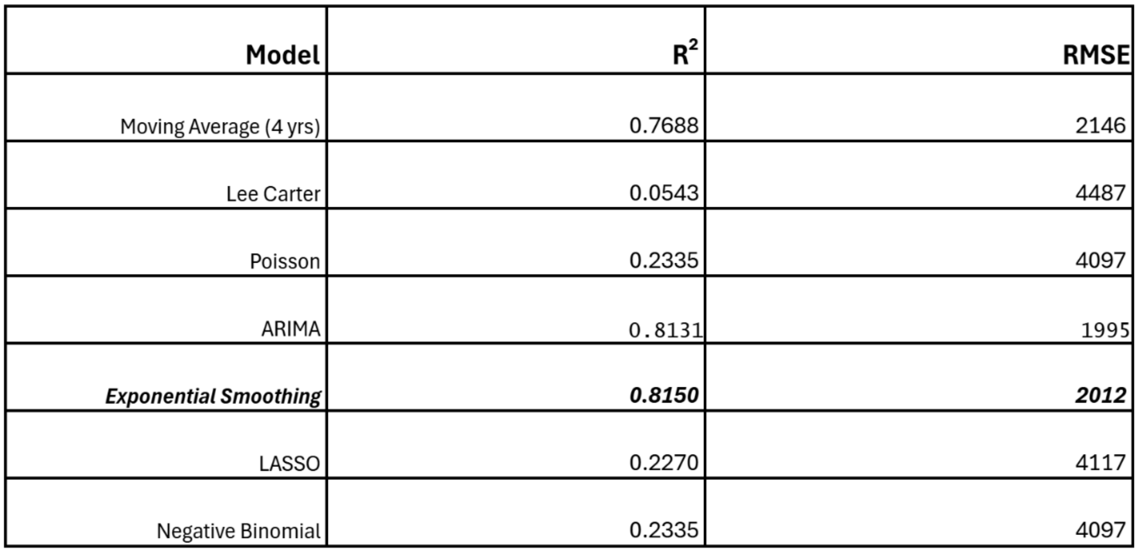
- This plot shows a comparison between the actual deaths (blue line) and forecasted deaths (red dashed line) in New York from 1997 to 2022, with red points indicating statistically abnormal excess deaths (beyond a 95% confidence interval).

- From the plot, the most notable abnormal spike occurs in the year 2020, which likely reflects the impact of the COVID-19 pandemic. The sharp rise in deaths in 2020 (indicated by the red points) indicates a significant deviation from what was forecasted. This suggests that external events, such as the pandemic, caused a surge in deaths that was not captured by the standard ETS model.
- The ETS (Exponential Smoothing) model, with a high alpha value of 0.9999, gives significant weight to recent observations, making it responsive to changes in the data. While the model's performance is generally strong, as indicated by relatively low MAPE (1.04%) and reasonable AIC and BIC values, it shows a large discrepancy in 2020 due to an external shock, likely the COVID-19 pandemic. This is reflected in

the higher RMSE (2012.26) and MAE (1578.48), which suggest difficulty in fitting the data accurately during abnormal events.

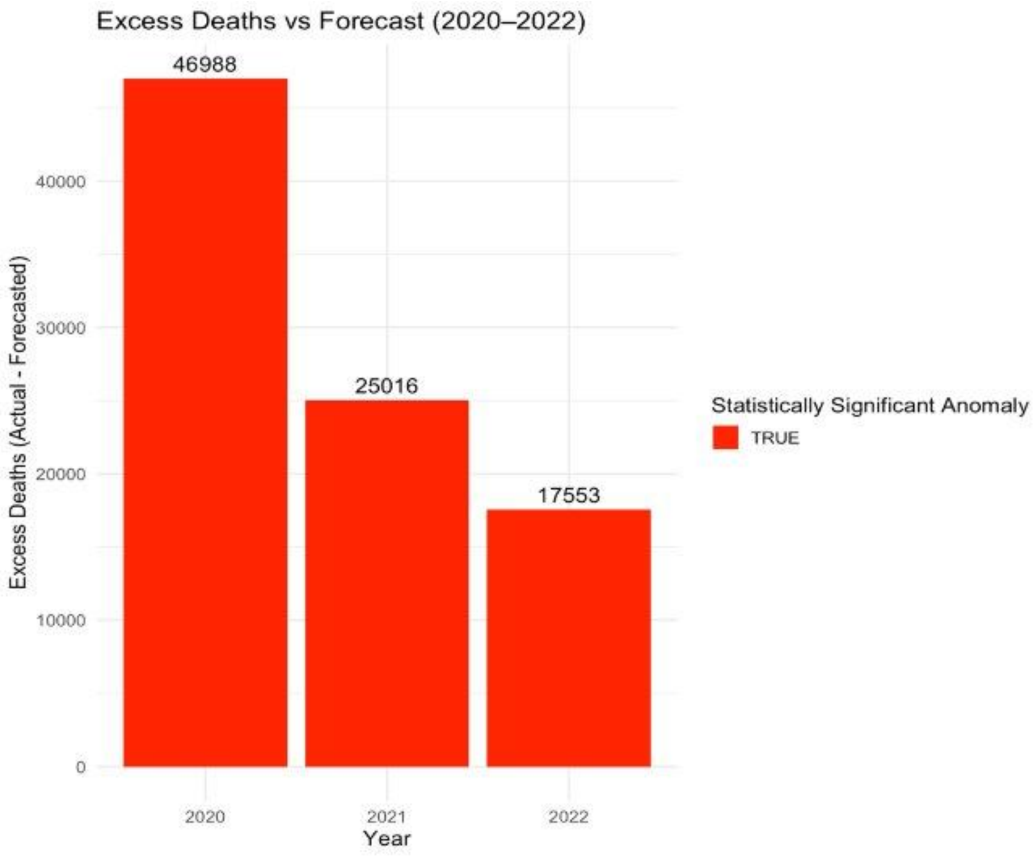
Results & Analysis

- **Model Comparison**



Model	R^2	RMSE
Moving Average (4 yrs)	0.7688	2146
Lee Carter	0.0543	4487
Poisson	0.2335	4097
ARIMA	0.8131	1995
Exponential Smoothing	***0.8150***	***2012***
LASSO	0.2270	4117
Negative Binomial	0.2335	4097

- **Excess Deaths**



- The exponential smoothing model was chosen due to its easy interpretation and minimal error.
- The excess deaths formula is represented by

$$\text{Excess Deaths} = \sum_{t=2020}^{2022} (D_t - \hat{D}_t)$$

  where D_t is the observed (actual) deaths in year and $\widehat{D}_t$ is the forecast from the exponential-smoothing model.
- For the State of New York the prediction is, due to Covid:
  $\mathit{Excess\ Deaths} = 89557$

Conclusion

Looking at three decades of pre-COVID data, we estimate that roughly 89,557 extra deaths occurred in New York State between 2020 and 2022. To reach that number, we first built several forecasting models, exponential smoothing (our main benchmark), ARIMA, Lee–

Carter, simple moving averages, and count - data regressions, to establish what mortality would have looked like absent the pandemic. The chosen exponential-smoothing model fit the pre-pandemic pattern well (RMSE ≈ 2012; R^2 ≈ 0.815) and agreed, within a few thousand deaths, with the other approaches.

The gap between predicted and observed deaths shows where COVID-19 hit hardest: Hispanic, Black, and other minority groups saw the sharpest percentage increases, and the largest absolute jump was among people aged 65 and older. We also observed that deaths from heart disease, diabetes, and chronic lung conditions rose above their long-term trends, suggesting the virus disrupted care for ongoing illnesses.

Our study is subject to several limitations. First, it relies solely on historical death counts and does not incorporate additional predictors, such as socioeconomic status or health-care access, that might explain demographic disparities. Second, reporting lags and misclassification of causes may bias the magnitude and timing of excess-death estimates. Finally, because New York's population structure and vital-statistics practices differ from other regions, these findings may not generalize directly to places with different demographics or reporting quality.

Despite these caveats, the uneven burden revealed by our analysis points to clear policy priorities for future public-health emergencies. In similar situations, directing resources—testing sites, vaccination campaigns, outreach, and treatment capacity—toward underrepresented racial and ethnic communities and to older adults could substantially reduce excess mortality. Strengthening support in long-term care facilities, improving data reporting infrastructure, and embedding equity-focused metrics into emergency response plans will help ensure that the next crisis does not amplify existing health gaps.

References

- Thompson, C. N., Baumgartner, J., Pichardo, C., et al. (2020). *COVID-19 Outbreak — New York City, February 29–June 1, 2020*. MMWR Morbidity and Mortality Weekly Report, 69(46), 1725–1729. https://www.cdc.gov/mmwr/volumes/69/wr/mm6946a2.htm
- Stoto, M. A., Schlageter, S., & Kraemer, J. D. (2022). *COVID-19 mortality in the United States: It's been two Americas from the start*. PLOS ONE, 17(4), e0265053. https://pmc.ncbi.nlm.nih.gov/articles/PMC9049562/
- Pugh, T., Harris, J., Jarnagin, K., Thiese, M. S., & Hegmann, K. T. (2022). *Impacts of the statewide COVID-19 lockdown interventions on excess mortality, unemployment, and employment growth*. Journal of Occupational and

Environmental Medicine, 64(9), 726–730. https://pmc.ncbi.nlm.nih.gov/articles/PMC9426308/

- Foster, T. B., Fernandez, L., Porter, S. R., & Pharris-Ciurej, N. (2024). *Racial and ethnic disparities in excess all-cause mortality in the first year of the COVID-19 pandemic*. Demography, 61(1), 59–85. https://pmc.ncbi.nlm.nih.gov/articles/PMC7996479/
- World Health Organization. (2023). *Methods for estimating the excess mortality associated with the COVID-19 pandemic (2020–2021)*. Geneva: WHO. https://www.who.int/publications/m/item/methods-for-estimating-the-excess-mortality-associatedwith-the-covid-19-pandemic
- Office of the New York State Comptroller. (2022, March 15). *State's pandemic response to nursing homes hindered by ill-prepared state agency* [Press Release]. https://www.osc.ny.gov/press/releases/2022/03/dinapoli-states-pandemic-response-nursing-homes-hindered-ill-prepared-state-agency
- NIDDK. "Overweight & Obesity Statistics." *National Institute of Diabetes and Digestive and Kidney Diseases*, National Institute of Health, Sept. 2021, https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity#:~:text=the%20above%20table-,Nearly%201%20in%203%20adults%20(30.7%25)%20are%20overweight.,9.2%25)%20have%20severe%20obesity.