BST260 Project: Excess Mortality

Rex Manglicmot

Introduction

Puerto Rico is a lush island in the beautiful tropical Caribbean that encapsulates a rich cultural history and draws in millions of visitors annually. However, a critical public health concern lies beneath its beauty: examining and understanding excess mortality (the number of raw observed deaths minus modeled expected deaths).

Excess mortality is an important measure to study because it conveys key insights into the impacts of unexpected natural disasters like typhoons and socioeconomic crises, such as consumer goods limitations, on the population health of Puerto Rican citizens. Further, excess mortality is a beneficial public health measure because it requires minimal data, relying only on historical death records. Furthermore, excess mortality avoids reliance on the accuracy of cause-of-death assignments on death certificates, making it more robust in contexts with limited diagnostic precision. Lastly, excess mortality allows for comparisons and contrasts of trends across regions and time, showcasing a sweeping view of the overall impact of events like pandemics and natural disasters on populations.

By enumerating deaths beyond the expected from historical population data, researchers and government policymakers can assess societal disruptions and devise action plans to alleviate future harm. One of the most significant events in Puerto Rico's history was Hurricane María, a disastrous storm that ravaged the island in September 2017. This Category 4 hurricane caused remarkable destruction, disrupting electricity and health infrastructure, and leaving many Puerto Rican citizens without access to basic services for months on end. Beyond its immediate physical toll, Hurricane María catalyzed a public health crisis, with its impact felt most acutely among vulnerable populations such as the elderly and individuals with pre-existing conditions. Because hospitals were overwhelmed, these populations needed the most care. Thus, within the DS Labs package, using the puerto_rico dataset provides not only a lucid picture of the hurricane's immediate effects before and after but also sheds light on long-term vulnerabilities within Puerto Rico exacerbated by health systems (i.e., hospitals), delayed recovery efforts, and systemic inequalities affecting individuals.

Therefore, this BST260 Intro to Data Science project aims to contribute to the public health body of literature through an exhaustive study of excess mortality in Puerto Rico with respect

to Hurricane María. The project's approach underscores the critical need for adaptive, resilient public health infrastructure that can effectively respond to unprecedented environmental challenges. Specifically, the four main goals of this analysis are:

1) Examine population sizes by age group and sex to identify demographic patterns and their implications for public health.
2) Calculate expected mortality and standard deviations for each week before 2017, broken down by age group and sex, to establish baseline trends.
3) Explore historical data for anomalies in mortality rates before 2017, removing periods of abnormal mortality to refine expected death rate calculations.
4) Estimate excess deaths for each week of 2017–2018, particularly in the aftermath of Hurricane María, to assess its impact on different age groups and sexes.

Through these four goals, this project seeks to unravel critical insights about the interplay between environmental disasters and public health in Puerto Rico. Further, the findings hope to shed light and contribute to discussions on preparation and recovery strategies, thereby providing a blueprint for improving Puerto Rico's health systems' responses to future crises in the event another natural disaster occurs in the not-so-distant future.

Methods

Data Source and Collection

The dataset used in this study, puerto_rico_counts, was obtained from the dslabs package and contains demographic and mortality data for Puerto Rico. puerto_rico_counts includes five key variables: age group (categorical representation of age groups), date (daily observation dates), sex (male or female), population (numeric values representing population estimates), and outcome (numeric values for mortality counts). puerto_rico_counts spans 1985 to 2022, providing a medium to investigate time series for mortality trends across the variables above. The date variable specifically represents daily observations, making it suitable for time series analysis and detecting trends, seasonal variations, or sudden spikes in mortality over nearly four decades.

Data Cleaning and Preprocessing

The data cleaning process followed a systematic approach to ensure the puerto_rico_counts dataset answered the 4 four main objections and consisted of exploratory data analysis. First, an initial exploration of the dataset was performed by inspecting the structure, column names, and dimensions using functions such as head(), colnames(), and dim(). This step provided a preliminary understanding of the variables and confirmed the presence of five columns and 499,644 observations.

Duplicate rows were checked and removed using the distinct() function. Since the dimensions of the dataset remained the same, it was confirmed that no duplicate rows existed. To validate key combinations (age group, date, and sex), a count was performed, and no duplicate key combinations were found, indicating that these variables uniquely identify the data.

Next, missing values were assessed using a column-wise check with summarise_all(~sum(is.na(.))). The results showed no missing data in any of the variables. puerto_rico_counts data types were then verified using the str() function, ensuring proper formatting for each column: age group was a factor, date was in date format, and population and outcome were numeric. This validation step confirmed that the data types aligned with the expected structure of the dataset.

The values in population and outcome were checked for logical consistency. Summary statistics for these columns were generated using the summary() function, and filtering methods were applied to detect negative or unrealistic values.

Analytical Techniques

After cleaning and preprocessing, the dataset was used for time series analysis and demographic comparisons. The date variable was critical in tracking mortality trends over time and was transformed into years and weeks respective to the 4 main objectives.

Mortality counts (outcome) were evaluated based on population estimates to calculate mortality rates, enabling the identification of excess mortality across specific age groups and sexes.

Age groups based on similar mortality patterns were combined to better uncover the data for analyzing mortality trends over broader age ranges. For example, groups such as "0-4," "5-9," and "10-14" were merged into a single "0-14" category, while older groups like "60-64," "65-69," and "70-74" were aggregated into "60-74."

Visualizations were integral to analyzing the data's findings. For example, time series and line plots were visualized to track mortality trends and seasonal variations among the age groups aforementioned. Additionally, faceted plots displayed subgroup-specific trends, allowing for side-by-side comparisons across age, sex, and time.

Results

Q1

Figure 1 graph shows population trends over time by age group and sex. For both males and females, younger age groups (e.g., 15–19, 20–24) exhibit a gradual decline in population size over time. Middle-aged groups (e.g., 35–39, 40–44) show relatively stable trends, while older age groups (e.g., 65–69, 70–74) have increasing population sizes. Females consistently have higher population counts in older age groups compared to males.

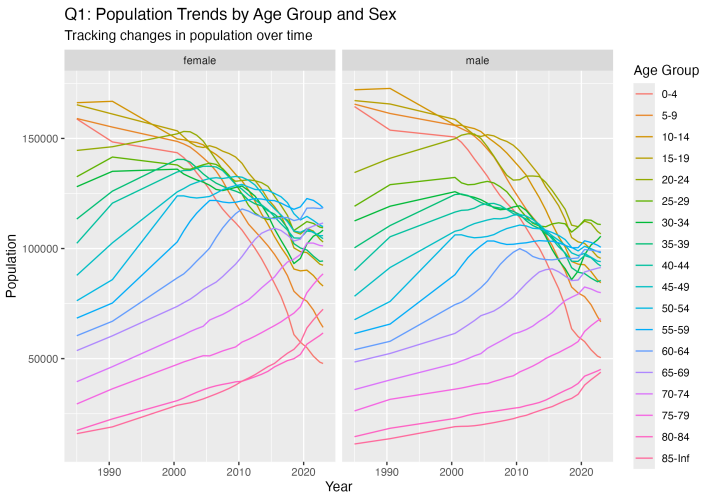

Figure 1: Population Trends by Age Group and Sex

Q2

(Please see supplementary.qmd for reasons why groups were combined. That figure was moved because we are only allowed to have up to 4 figures in the report.)

(That figure in the supplementary shows mortality trends by age and sex. Younger age groups (0-14, 15-19, 20-24) have low, stable mortality with minimal sex differences. In adults aged 25-59, mortality increases with age, and males consistently exhibit higher rates than females. For older adults (60-74), mortality rises further, with a widening male-female gap. Among seniors (75+), mortality is highest, though the gap narrows in the oldest group (85+) as female rates begin to align with male rates.

Based on these trends, similar age groups can be combined for analysis. The 0-14 group consolidates ages 0-4, 5-9, and 10-14 due to their low and stable mortality. The 15-29 category combines 15-19, 20-24, and 25-29, as they share comparable mortality levels and minimal variability. Adults aged 30-59 (30-34 to 55-59) can be grouped together, reflecting a gradual but consistent increase in mortality. Similarly, ages 60-74 (60-64, 65-69, and 70-74) and Seniors aged 75+ (75-79, 80-84, and 85+) can be combined.)

(This question asks for expected mortality and sds, so I treated them as separate plots).

Figure 2 shows **expected mortality** trends across combined age groups. Older groups (60–74, 75–79, 80–84, 85+) consistently have higher mortality rates, with males exhibiting notably higher values and more variability than females. In contrast, younger groups (0–14, 15–29) display significantly lower and more stable mortality, with minimal sex-based differences. Adults (30–59) show moderate mortality, where males maintain a steady upward trend compared to females.

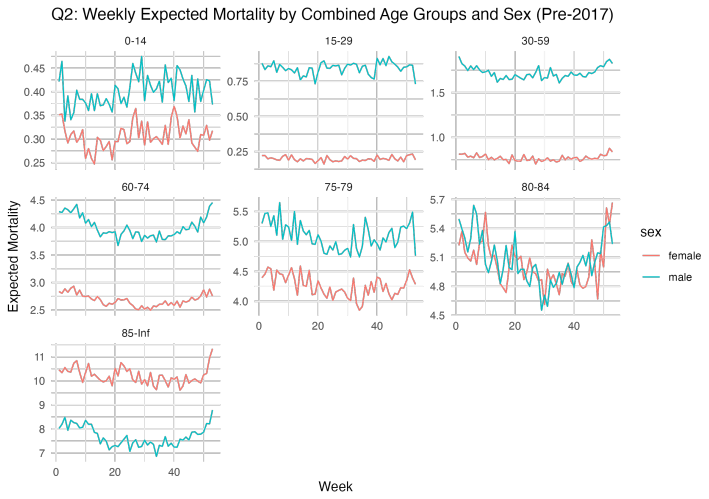

Figure 2: Weekly Expected Mortality by Combined Age Groups & Sex (Pre-2017)

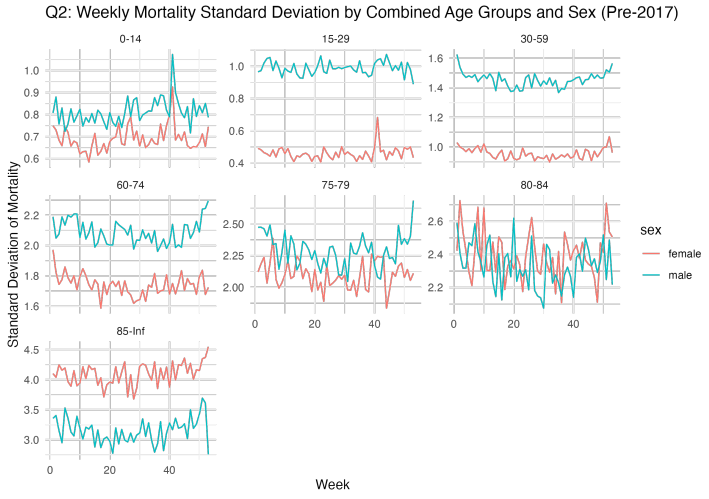

Figure 3: Weekly Mortality SD by Combined Age Groups & Sex (Pre-2017)

Figure 3 highlights the variability (**standard deviation**) in mortality rates. Older age groups (75–79, 80–84, 85+) have the highest standard deviations, particularly among males, indicating significant fluctuations. The 60–74 group also shows notable variability, though less pronounced. Younger groups (0–14, 15–29) exhibit lower standard deviations, reflecting stability and minimal sex differences. In the 30–59 group, males consistently show greater variability than females.

Q3

(Please see supplementary for mortality trends with red points marking excess mortality values exceeding values > mean + 2SD. Essentially, it shows that excess mortality is more common in older age groups (60-74, 75-79, 80-84, 85+), where males consistently show higher mortality rates than females. Younger groups (0-14, 15-29) exhibit lower mortality with fewer anomalies.)

Figure 4: Adjusted Weekly Expected Mortality by Age Groups & Sex

Figure 4 shows adjusted mortality trends after removing weeks with excess mortality (values > mean + 2 SD). The trends are smoother, with males maintaining higher mortality rates, especially in older age groups (30-59, 60-74, 75+). Younger groups (0-14, 15-29) display stable, low mortality levels. The sharp decline at the end reflects incomplete data.

Q4

(Please see supplementary as it highlights excess mortality during 2017–2018, with cyan points marking weeks where mortality exceeded the baseline (mean + 2 SD). Essentially, it shows that the most affected age groups were older adults (60–74, 75–79, 80–84) and seniors (85+),

where excess deaths were most pronounced after Week 38, corresponding to Hurricane María's landfall.)

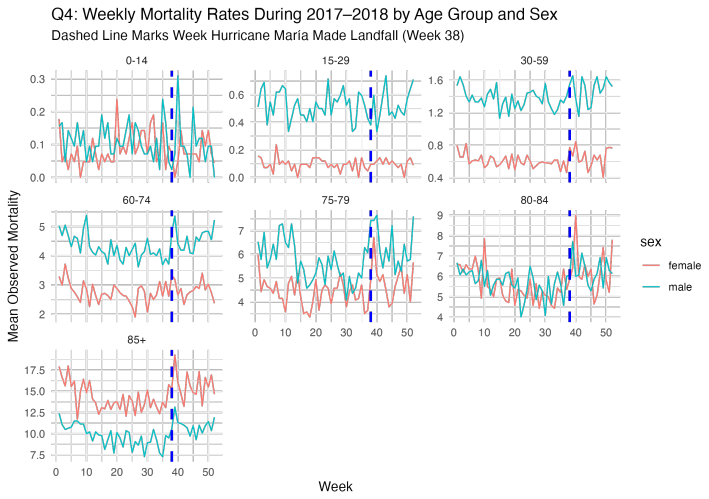

Figure 5: Weekly Mortality Rates During 2017-18 by Age Groups & Sex

Figure 5 shows clear differences between men and women. In adults (30–59) and seniors (65+), males exhibited consistently higher mortality rates compared to females, especially during periods of excess mortality. However, in the 85+ age group, females experienced a sharper increase in mortality following Hurricane María, surpassing male rates. Children (0–14) and young adults (15–29) were the least affected, with minor deviations observed for both sexes. Older age groups, particularly males in mid-age and seniors, were disproportionately impacted by excess mortality, while the oldest females (85+) showed a unique spike post-María.

Discussion

Question Analyses

The results from Q1 revealed clear patterns in older age groups, particularly 60-74, 75-79, 80-84, and 85+. It showed the highest mortality rates with a difference between sexes. For instance, males exhibited higher mortality than females across all groups, a gap that widened with age. However, younger groups like 0-14 and 15-29 displayed lower, more stable mortality rates. These two findings show natural expectations that mortality increases with age while emphasizing the greater vulnerability of older males.

The results Q2 expanded on mortality variability by examining standard deviations. Older groups (without the age group 60-74), particularly 75-79, 80-84, and 85+, displayed the highest fluctuations, especially among males. This result suggests that older populations' mortality is

higher and unpredictable, possibly due to external factors like seasonal illnesses or environmental stressors as their immune system wanes. The younger age groups, however, in particular, 0-14 and 15-29, displayed lower SDs and minimal differences between the sexes (male and female). This discovery could be due to their mature and developed immune system and resilience to common diseases. Altogether, these findings highlight that both age and sex are factors influencing both mortality levels and their fluctuations.

The identification of excess mortality in Q3, (in the supplmentary directory; there is a plot with red dots) defined as mortality above mean + 2 SD revealed key periods of elevated mortality, primarily in older age groups. Excess mortality was more frequent in groups 60-74, 75-79, 80-84, and 85+, where both sexes experienced spikes, though males were consistently more affected. Younger age groups showed fewer anomalies, reflecting lower mortality overall. After removing these periods, adjusted trends (Figure 4) were smoother, with males continuing to show higher mortality rates, particularly in older groups.

Q4 analyzed mortality during 2017-2018, specifically around the week Hurricane María made landfall (Week 38, 2017). A sharp increase in mortality was observed in particular in older groups (60-74, 75-79, and 85+) in which males showed higher levels compared to females. Adults aged 30-59 also showed notable increases, though less pronounced. For younger groups, such as 0-14 and 15-29, mortality remained relatively stable with minimal deviations. The results again convey the uneven effect of Hurricane María on older populations and males. Thus, these findings suggest greater vulnerability to such populations in post-disaster conditions like natural disasters.

Implications

The findings collectively reveal the influence of age, sex, and external events on mortality. However, upon closer review, older age groups are consistently at higher risk, with males facing very high mortality across all analyses. This may be attributed to biological differences, lifestyle factors, or greater exposure to chronic conditions in their golden ages. Therefore, more societal policies and gerontology public health research are needed to monitor such vulnerable populations during periods of elevated risk, such as flu seasons or periods of natural disasters.

Excess mortality analysis provides critical insights for identifying anomalies and understanding baseline trends. Removing outliers in Q3 unearthed the underlying yet prevalent mortality patterns in Puerto Rico's population. Again, mortality is higher for older people.

Limitations and Future Research

While the analysis reveals significant trends, there are limitations. The dataset focuses on mortality but does not capture other contributing factors, such as socioeconomic conditions, healthcare access, or pre-existing comorbidities. The dataset contained only five variables, which is not a good way to elucidate mortality excess.

Another assumption in this project is that the Puerto Rico population and mortality counts provided in the data are accurate and complete. However, it is reasonable to think that not all data collection was complete. In times of natural disaster, data collection comes in second while care for the population becomes primary.

While the dataset captures daily mortality trends over an extensive period, it does not account for external factors, such as underlying public health concerns (e.g., access to quality water), which may also have influenced mortality rates. Additionally, aggregating age groups might smooth out significant variations within narrower age ranges and may not truly reflect this project's findings on the Puerto Rican population.

Conclusion

This BST260 project demonstrates the importance of data cleaning, visualizations, and interpretation analyses in understanding mortality trends in Puerto Rico. It helps identify periods and populations of excess mortality to create and modify public health strategies. Again, older age groups and males remain the most affected populations, particularly during events like Hurricane María. Active pursuance of solutions to these vulnerabilities in the elderly population through government conversations, public education institutions, and news media will be essential in mitigating future mortality spikes and protecting at-risk populations. As such, if such solutions were to be implemented we would expect to see a normalization of mortality rates across all ages races as future data colleciton continues for the beautiful state of Puerto Rico.