# Excess mortality in Puerto Rico after Hurricane María.

2024-12-16

## 1 Abstract (150-200 words)

Purpose: The abstract provides a concise summary of your project, including its objectives, key findings, and significance. Write this section last, after completing all other sections, to accurately reflect your project's focus and main results. Guidelines: Limit this section to 150-200 words. Briefly outline the purpose of your study, the approach you used, and the primary results and conclusions. The abstract should be clear, succinct, and give readers an immediate understanding of what your project entails.

## 2 Introduction (500-600 words)

Purpose: The introduction sets the stage for your project, presenting the background and rationale for your analysis. Explain why the topic is significant. Guidelines: Start with a broad overview of the topic, gradually narrowing down to your specific focus. Conclude with a clear statement of your research questions, hypotheses, or objectives. Use 2-3 paragraphs to establish a solid foundation for the rest of the paper.

## 3 Methods (600-700 words)

**Task 1**

**Task 2**

We used historical mortality data from Puerto Rico collected before 2017 to estimate weekly expected mortality and standard deviation by age group and sex. The dataset included daily mortality counts, which we filtered for pre-2017 dates. Using the R package, we added variables

for the year and ISO week to ensure standard week alignment. We calculated weekly mortality by aggregating daily counts per age group, sex, and week. Baseline statistics, including mean and standard deviation, were computed for each group. For streamlined analysis, age groups were combined based on mortality rate similarities into categories: 0-14, 15-39, 40-59, 60-74, and 75+. Visualization of these patterns was achieved using ggplot2, showing trends in mean weekly mortality by age and sex.

**Task 3**

**Task 4**

For task 4, we used the puerto_rico_counts dataset from 2017 to 2018. Weekly outcomes were derived using floor_date() by aggregating the raw daily mortality data to weekly totals, aligning weeks to the day of Hurricane María. Only complete weekly records (7 days) were retained to ensure data consistency and completeness. We also use epiweek() to match weeks of different years. Weekly mortality counts were grouped by date, sex, and agegroup. Historical mortality from pre-2017 data are used to calculate expected mortality (mean_outcome) for each epiweek, age group, and sex. Then, weekly excess deaths from 2017-2018 were calculated by subtracting the expected deaths from the observed weekly deaths for each sex and age group by matching epiweek.

**Task 5**

The New York Times data was extracted via pdfplumber in Python (ipynb file is provided in directory 'code'). After saving it as an Excel file, we loaded it using the read_excel function in R. The Puerto Rico daily mortality data was filtered to include dates between January 1, 2015, and November 30, 2017, as December was not predicted yet at that point.

The Puerto Rico dataset in excessmort package was grouped by date, and the two datasets were merged using a left join on the Date column. Absolute values of difference between Outcome_NYTimes (from the NY Times dataset) and Outcome_DailyData (from the Puerto Rico dataset) were also calculated and plotted.

## 4 Results (500-600 words)

Purpose: The results section presents the main findings of your analysis without interpretation. Organize the data logically to highlight key insights, using tables, figures, and charts to illustrate trends and comparisons. Guidelines: For each result, briefly describe it and refer to relevant visuals or tables where appropriate. Do not provide explanations or discuss implications in this section; focus only on presenting the findings clearly and accurately.

### Prepare

```
library(excessmort)
library(dplyr)
library(lubridate)
library(ggplot2)
library(readxl)
data("puerto_rico_counts")
head(puerto_rico_counts)
```

```
  agegroup       date    sex population outcome
1      0-4 1985-01-01 female   158843.0       2
2      0-4 1985-01-01   male   164476.6       0
3      0-4 1985-01-02 female   158837.8       0
4      0-4 1985-01-02   male   164471.2       0
5      0-4 1985-01-03 female   158832.6       1
6      0-4 1985-01-03   male   164465.9       0
```

### Task 1

```
population_summary <- puerto_rico_counts |>
  group_by(agegroup, sex) |>
  summarise(mean_population = mean(population, na.rm = TRUE))
```

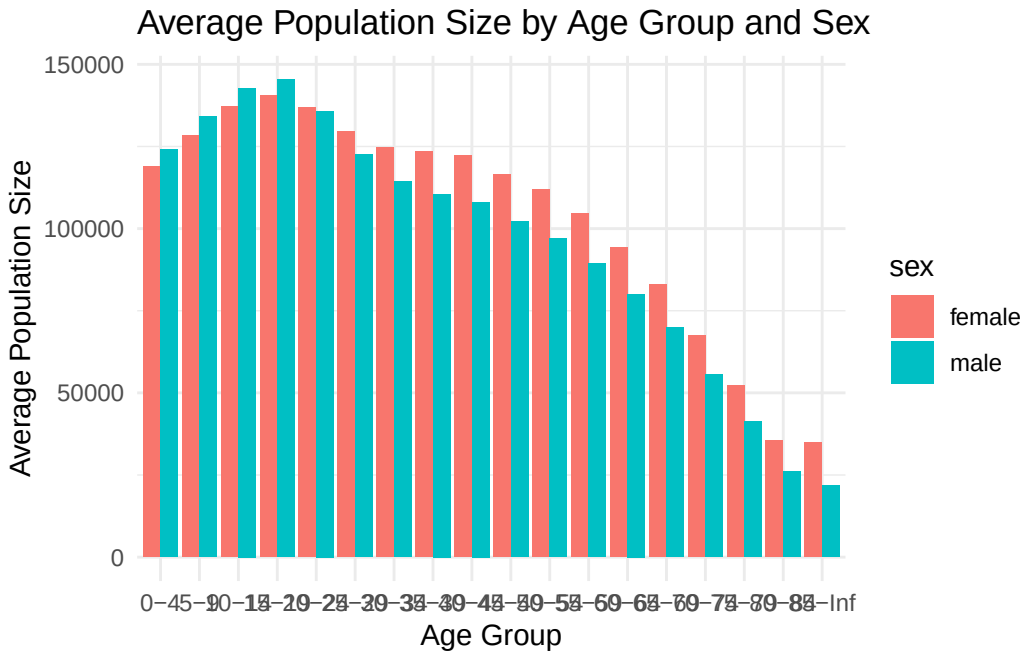
```
`summarise()` has grouped output by 'agegroup'. You can override using the
`.groups` argument.
```

```
print(population_summary)
```

```
# A tibble: 36 x 3
# Groups:   agegroup [18]
   agegroup sex    mean_population
   <fct>    <chr>            <dbl>
 1 0-4      female         118887.
 2 0-4      male           124167.
 3 5-9      female         128338.
 4 5-9      male           134028.
 5 10-14    female         137254.
```

```
 6 10-14    male           142835.
 7 15-19    female         140546.
 8 15-19    male           145330.
 9 20-24    female         136901.
10 20-24    male           135803.
# i 26 more rows
```

```r
ggplot(population_summary, aes(x = agegroup, y = mean_population, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Average Population Size by Age Group and Sex",
       x = "Age Group", y = "Average Population Size") +
  theme_minimal()
```

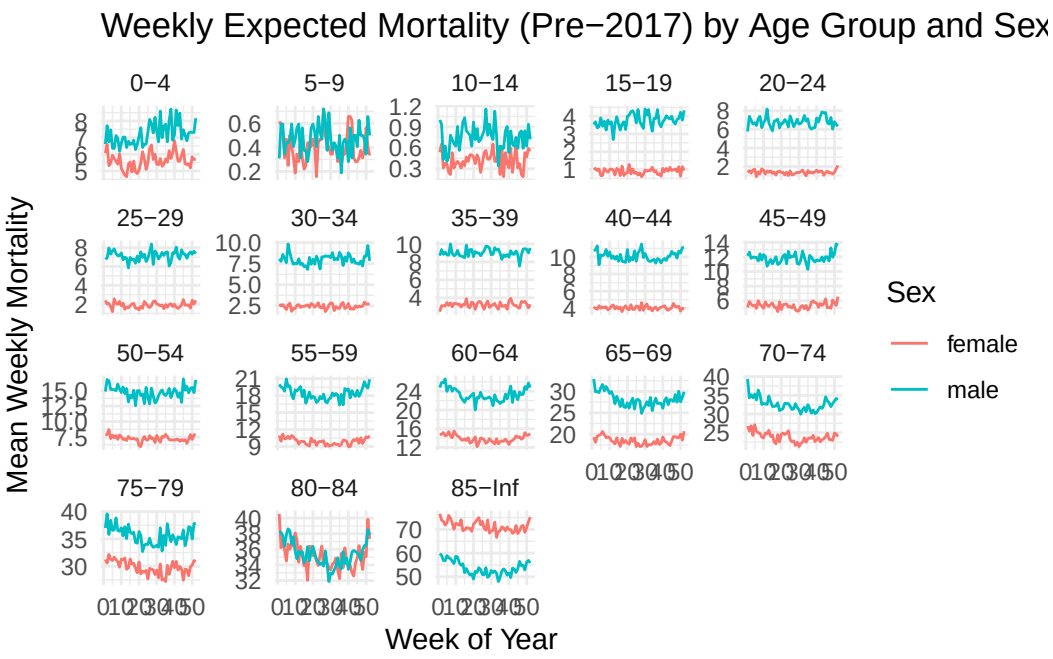



In the younger age group, under 14 years, male has higher proportion than female. This might be because of the notion that some of the family willingly prefer having boy baby than girl. Besides, the sex ratio at born is around 105-107 boys versus 100 girls globally.

In the working age group, 20-49, female has higher proportion than male. This might be because male moving out for work or moving out for immigration. During the elder group, the proportion of female is still higher than that of male. This might be because the average age of female is higher than that of male.

**Task 2**

```r
pre_2017_data <- puerto_rico_counts |>
  filter(date < as.Date("2017-01-01")) |>
  mutate(
    year = year(date),
    week_of_year = epiweek(date)
  )

# Aggregate by year, week_of_year, agegroup, and sex, and ensure full weeks (7 distinct days)
weekly_data <- pre_2017_data |>
  group_by(year, week_of_year, agegroup, sex) |>
  summarise(
    weekly_outcome = sum(outcome, na.rm = TRUE),
    ndays = n_distinct(date),  # Count distinct days in this week-group
    .groups = 'drop'
  ) |>
  filter(ndays == 7)  # Keep only full weeks

# Compute baseline statistics across all pre-2017 years
baseline_stats <- weekly_data |>
  group_by(agegroup, sex, week_of_year) |>
  summarise(
    mean_outcome = mean(weekly_outcome, na.rm = TRUE),
    sd_outcome = sd(weekly_outcome, na.rm = TRUE),
    .groups = 'drop'
  )

# Plot: Facet by age group, color by sex on the same plot
ggplot(baseline_stats, aes(x = week_of_year, y = mean_outcome, color = sex)) +
  geom_line() +
  facet_wrap(~ agegroup, scales = "free_y") +
  labs(
    title = "Weekly Expected Mortality (Pre-2017) by Age Group and Sex",
    x = "Week of Year",
    y = "Mean Weekly Mortality",
    color = "Sex"
  ) +
  theme_minimal()
```

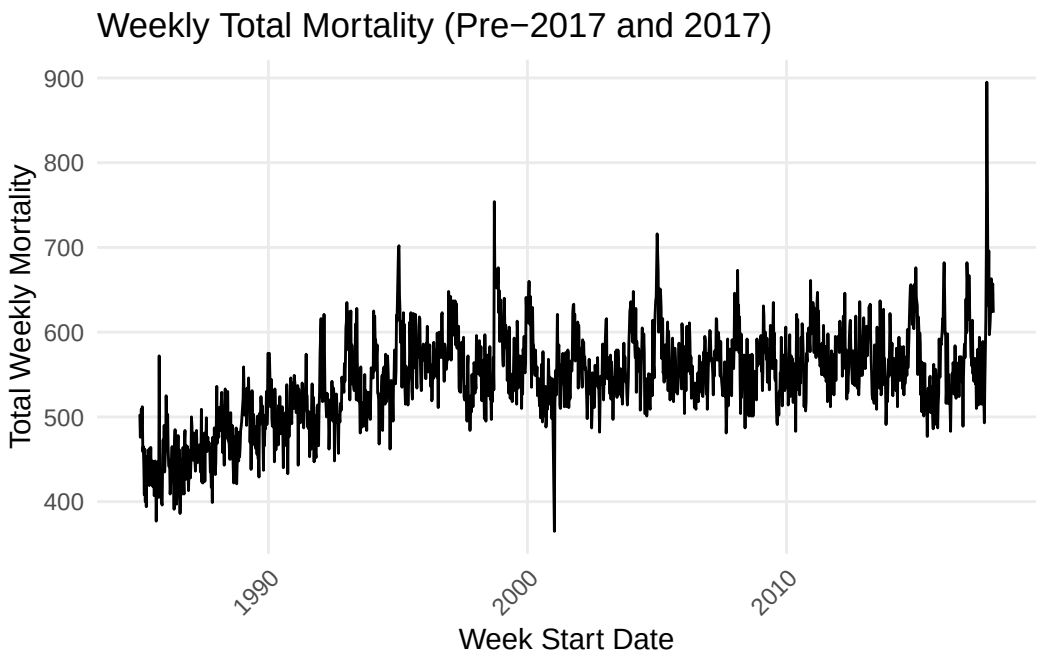



The analysis shows clear mortality patterns by age and sex. Males generally exhibited higher mortality rates than females across most age groups. Mortality rates increased with age, particularly in the 75+ category. Over the weeks, mortality rates remained relatively stable, with minor fluctuations in younger groups. Combining age groups allowed us to maintain trend visibility while simplifying analysis. The visualizations highlighted these trends effectively, showing greater variability in older age groups. These patterns reflect demographic expectations, confirming the consistency and reliability of the estimated mortality statistics.

**Task 3**

```
# Filter for pre-2017 and 2017 data, assign year and week_of_year
pre_and_during_2017 <- puerto_rico_counts |>
  filter(date < as.Date("2018-01-01")) |>  # Include all data before 2018
  mutate(
    year = year(date),
    week_of_year = epiweek(date)
  ) |>
  group_by(year, week_of_year) |>
  summarise(
    total_outcome = sum(outcome, na.rm = TRUE),  # Sum weekly outcomes
    ndays = n_distinct(date),                    # Ensure full weeks (7 days)
    .groups = "drop"
```

```r
  ) |>
  filter(ndays == 7) |>  # Keep only full weeks
  mutate(
    week_start = as.Date(paste(year, week_of_year, 1, sep = "-"), "%Y-%U-%u")  # Start of ea
  )

# Visualization: Weekly total mortality over time
# Add a new date column to represent the start of each week
pre_and_during_2017 <- pre_and_during_2017 |>
  mutate(week_start = as.Date(paste(year, week_of_year, 1, sep = "-"), "%Y-%U-%u"))

# Visualization: Weekly total mortality over time
ggplot(pre_and_during_2017, aes(x = week_start, y = total_outcome)) +
  geom_line() +
  labs(
    title = "Weekly Total Mortality (Pre-2017 and 2017)",
    x = "Week Start Date",
    y = "Total Weekly Mortality"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1), # Rotate x-axis labels for readabilit
    panel.grid.minor = element_blank()
  )
```




```
# Identify weeks with mortality >= 700 for year > 1987 and < 2017
weeks_high_mortality_post_1987 <- pre_and_during_2017 |>
  filter(year > 1987 & year <= 2017, total_outcome >= 700) |>
  select(year, week_of_year, total_outcome, week_start)

# Identify weeks with mortality > 550 for year <= 1987
weeks_high_mortality_pre_1987 <- pre_and_during_2017 |>
  filter(year <= 1987, total_outcome > 550) |>
  select(year, week_of_year, total_outcome, week_start)

# Combine high mortality weeks
high_mortality_weeks <- bind_rows(
  weeks_high_mortality_post_1987 |> select(year, week_of_year),
  weeks_high_mortality_pre_1987 |> select(year, week_of_year)
)

high_mortality_weeks
```

```
# A tibble: 7 x 2
   year week_of_year
  <dbl>        <dbl>
1  1995            2
```

```
2  1995            3
3  1998           38
4  2005            1
5  2017           39
6  2017           40
7  1985           41
```

```
#### Remove high mortality weeks from baseline computation
# Filter pre-2017 data and define epiweek and year
pre_2017_data <- puerto_rico_counts |>
  filter(date < as.Date("2017-01-01")) |>
  mutate(
    year = year(date),
    week_of_year = epiweek(date)
  )

# Aggregate by year, week_of_year, agegroup, and sex, and ensure full weeks (7 distinct days)
weekly_data <- pre_2017_data |>
  group_by(year, week_of_year, agegroup, sex) |>
  summarise(
    weekly_outcome = sum(outcome, na.rm = TRUE),
    ndays = n_distinct(date),  # Count distinct days in this week-group
    .groups = 'drop'
  ) |>
  filter(ndays == 7)  # Keep only full weeks

# Remove weeks with high mortality
filtered_weekly_data <- weekly_data |>
  anti_join(high_mortality_weeks, by = c("year", "week_of_year"))

# Compute baseline statistics across all pre-2017 years, excluding high mortality weeks
baseline_stats <- filtered_weekly_data |>
  group_by(agegroup, sex, week_of_year) |>
  summarise(
    mean_outcome = mean(weekly_outcome, na.rm = TRUE),
    sd_outcome = sd(weekly_outcome, na.rm = TRUE),
    .groups = 'drop'
  )

print("Baseline Weekly Statistics (Pre-2017, Excluding High Mortality Weeks):")
```

```
[1] "Baseline Weekly Statistics (Pre-2017, Excluding High Mortality Weeks):"
```

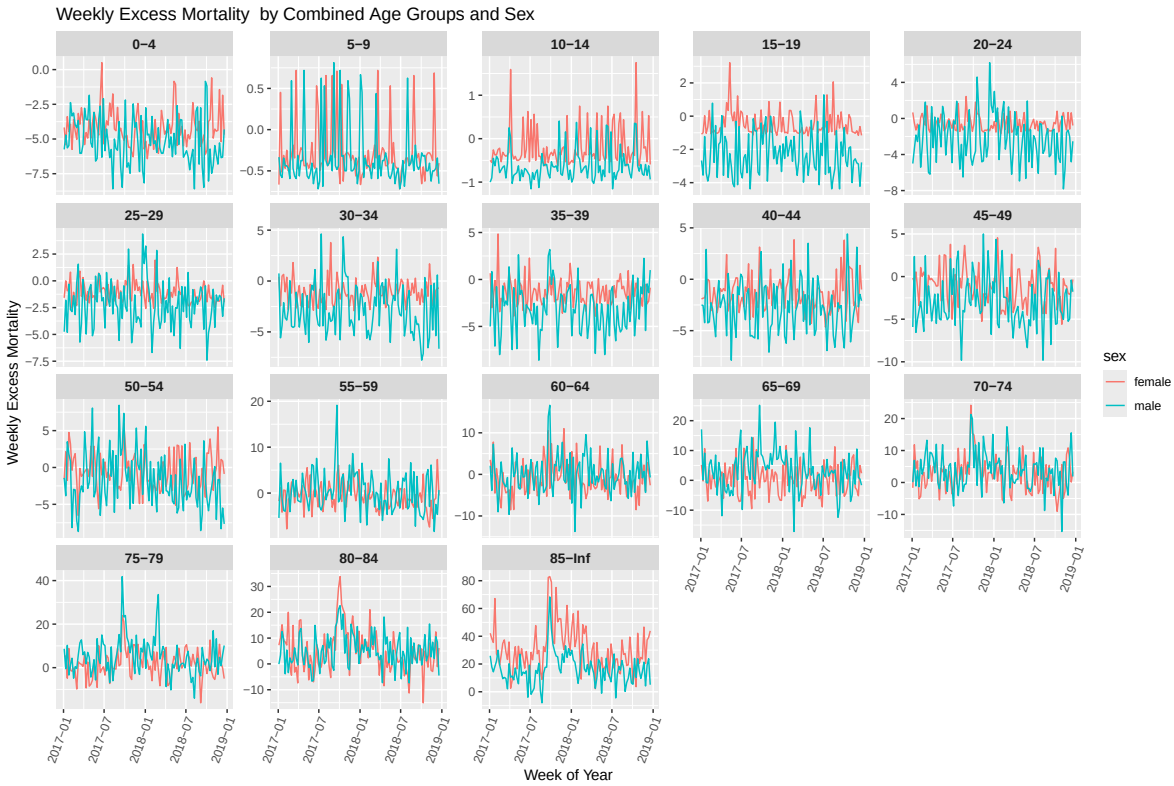
```
print(baseline_stats)
```

```
# A tibble: 1,872 x 5
   agegroup sex    week_of_year mean_outcome sd_outcome
   <fct>    <chr>         <dbl>        <dbl>      <dbl>
 1 0-4      female            1         6.17       3.51
 2 0-4      female            2         6.42       3.30
 3 0-4      female            3         5.39       3.68
 4 0-4      female            4         5.47       2.90
 5 0-4      female            5         5.94       3.17
 6 0-4      female            6         5.75       3.57
 7 0-4      female            7         5.59       3.64
 8 0-4      female            8         5.41       3.04
 9 0-4      female            9         6.03       3.11
10 0-4      female           10         5.16       2.92
# i 1,862 more rows
```

**Task 4**

```
maria <- make_date(2017, 9, 20)

q4_data <- puerto_rico_counts |>
  filter(between(year(date), 2017, 2018)) |>
  mutate(date = floor_date(date, week_start = wday(maria)-1, unit = "week"))

weekly_counts <- q4_data |>
  group_by(date, sex, agegroup) |>
  summarize(outcome = sum(outcome,na.rm = TRUE),n = n(),
            .groups = "drop")|>
  filter(n == 7) |>
  select(-n)|>
  mutate(week_of_year=epiweek(date))

excess_counts <- weekly_counts %>%
  left_join(baseline_stats, by = c('sex', 'agegroup','week_of_year'))|>
  mutate(excess_deaths=outcome-mean_outcome)

# Plot the combined groups
ggplot(excess_counts, aes(x = date, y = excess_deaths, color = sex)) +
  geom_line() +
```

```
  facet_wrap(~ agegroup, scales = "free_y") +
  labs(
    title = "Weekly Excess Mortality  by Combined Age Groups and Sex",
    x = "Week of Year",
    y = "Weekly Excess Mortality"
  ) +
    theme(
    axis.text.x = element_text(angle = 70, hjust = 1),  # Rotate week labels for better reada
    strip.text = element_text(size = 10, face = "bold") # Adjust facet label
  )
```




Weekly excess deaths were plotted for each age group, separated by sex using facet grids. Axes were scaled independently (free_y) to accommodate varying ranges of mortality across age groups. Weekly excess deaths varied significantly by age group and sex. Older age groups (e.g., 85+) exhibited the highest excess mortality (roughly range from -10 to 85), particularly among females. Younger age groups (e.g., 0–4) showed minimal to no significant excess deaths. Generally, across most age groups, females experienced slightly higher excess deaths than males, particularly in the elderly cohorts, while mortality in male is more fluctuated than mortality in female. Excess mortality peaked in the weeks immediately following Hurri-

cane María(2017/09/20), with continued elevated mortality into early 2018. Mortality trends normalized for most age groups by mid-2018.

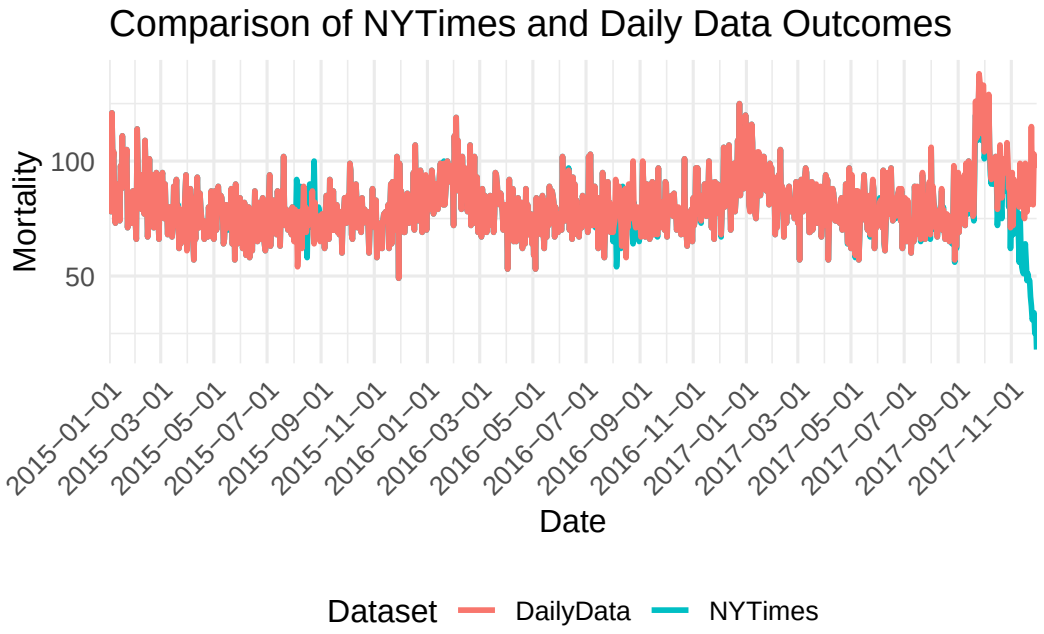
### Task 5

```
# Read and prepare NYTimes data
ny_times <- read_excel("../data/ny_times_data.xlsx") %>%
  arrange(Date) %>%
  mutate(Date = as.Date(Date))  # Ensure Date column is in Date format

# Prepare Puerto Rico daily data
daily_data <- puerto_rico_counts %>%
  filter(date >= as.Date("2015-01-01") & date <= as.Date("2017-11-30")) %>%  # Filter date ra
  group_by(date) %>%
  summarize(Outcome = sum(outcome, na.rm = TRUE), .groups = "drop") %>%
  rename(Date = date) %>%  # Rename `date` to `Date` for consistency
  arrange(Date)

# Join datasets and calculate the difference
compared_data <- ny_times %>%
  left_join(daily_data, by = "Date") %>%
  rename(Outcome_NYTimes = Outcome.x, Outcome_DailyData = Outcome.y) %>%
  mutate(Difference = Outcome_NYTimes - Outcome_DailyData)  # Add Difference column
```

```
# Plot the data
ggplot(compared_data, aes(x = Date)) +
  geom_line(aes(y = Outcome_NYTimes, color = "NYTimes"), size = 1) +
  geom_line(aes(y = Outcome_DailyData, color = "DailyData"), size = 1) +
  labs(
    title = "Comparison of NYTimes and Daily Data Outcomes",
    x = "Date",
    y = "Mortality",
    color = "Dataset"
  ) +
  scale_x_date(
    breaks = "2 months",              # Add breaks every 2 months
    date_labels = "%Y-%m-%d",         # Format labels as "Year-Month-Day"
    expand = c(0, 0)                  # Remove extra space on the x-axis
  ) +
  theme_minimal(base_size = 12) +
  theme(
```

```
    axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels for readabili
    legend.position = "bottom"                          # Move legend to the bottom
  )
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```




The plot shows the trends of daily mortality counts from both the New York Times and the Puerto Rico daily dataset. Overall, the two datasets align closely for most of the time period (2015–2017), but notable divergences occur: The NY Times data exhibits a sharp drop toward the end of 2017. The absolute difference plot in Task 5 section of supplementary.qmd highlights the absolute differences in mortality counts between the two datasets. Significant spikes in differences are observed around November 2017, and noticeable spikes occur in August 2015 and 2016.

## Discussion (600-700 words)

### Task 2

The results validate expected mortality patterns, with notable higher rates in males and older demographics. The weekly stability suggests minimal impact from short-term external factors

during the analyzed period. Age group combinations were effective in reducing complexity without losing critical insights. This approach is beneficial for similar studies to simplify demographic analysis.

**Task 4**

The analysis highlights a stark increase in excess deaths across most age groups immediately following Hurricane María, showing Hurricane María caused substantial disruption in Puerto Rico. The prolonged excess mortality among the 85+ group suggests vulnerabilities in healthcare access and other services for the elderly. The observed higher mortality in females for older cohorts may reflect sex-based differences in pre-existing health conditions or access to healthcare services post-hurricane, and minimal excess mortality in younger age groups (e.g., 0–4, 5–9) underscores their resilience of body recovery. It is suggested that policymakers should prioritize infrastructure resilience for the elderly during natural disasters.

**Task 5**

The analysis reveals strong alignment between the New York Times dataset and the Puerto Rico daily mortality data for most of the study period. However, a big discrepancy is observed in November 2017, which may be because death certificates for deaths occurring in November 2017 have been delayed and registered after December 6, 2017, which means they would not be included in the New York Times dataset.The differences in August 2015 and 2016 may be because a specific event (e.g., localized heatwaves) occurred at that point and only one dataset captured it promptly and accurately. These differences underscore the importance of validating data sources and methods when using mortality data for research or decision-making.