NY Data Analysis

Vijetha Ramdas

2025-09-05

Analysis

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)

data <- read_csv('ny_pollution.csv.gz')
```

```
## Rows: 3287 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (2): death, pollution
## date (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data$date <- ymd(data$date)
data$year <- year(data$date)

data = data |>
  filter(year %in% c(1997, 2004))
data |>
  group_by(year) |>
  summarise(mean_pollution = mean(pollution, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##    year mean_pollution
##   <dbl>          <dbl>
## 1  1997           26.9
## 2  2004           15.2
```

Statement

The mean pollution in New York in 2004 was lower than the mean pollution in New York in 1997.

Supporting Premise

The first 3 quartiles of 2004 data are below the second quartile of the 1997 data based on the histograms below. Additionally, the right tail of the 2004 distribution is within the range of the 1997 data. This evidence supports our statement that the mean pollution in New York in 2004 was lower than the mean pollution in New York in 1997.

```
data$year_fact <- as.factor(data$year)

ggplot(data, aes(x = pollution)) +
  geom_boxplot() +
  facet_wrap(~ year_fact, ncol = 1)
```

```
## Warning: Removed 609 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```