Analysis

Neena Varanasi

2025-09-02

```
# read in the data
dat <- 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.
```

Statement About the Data

The date with the most pollution is 2002-07-07 with a pollution value of 92.4.

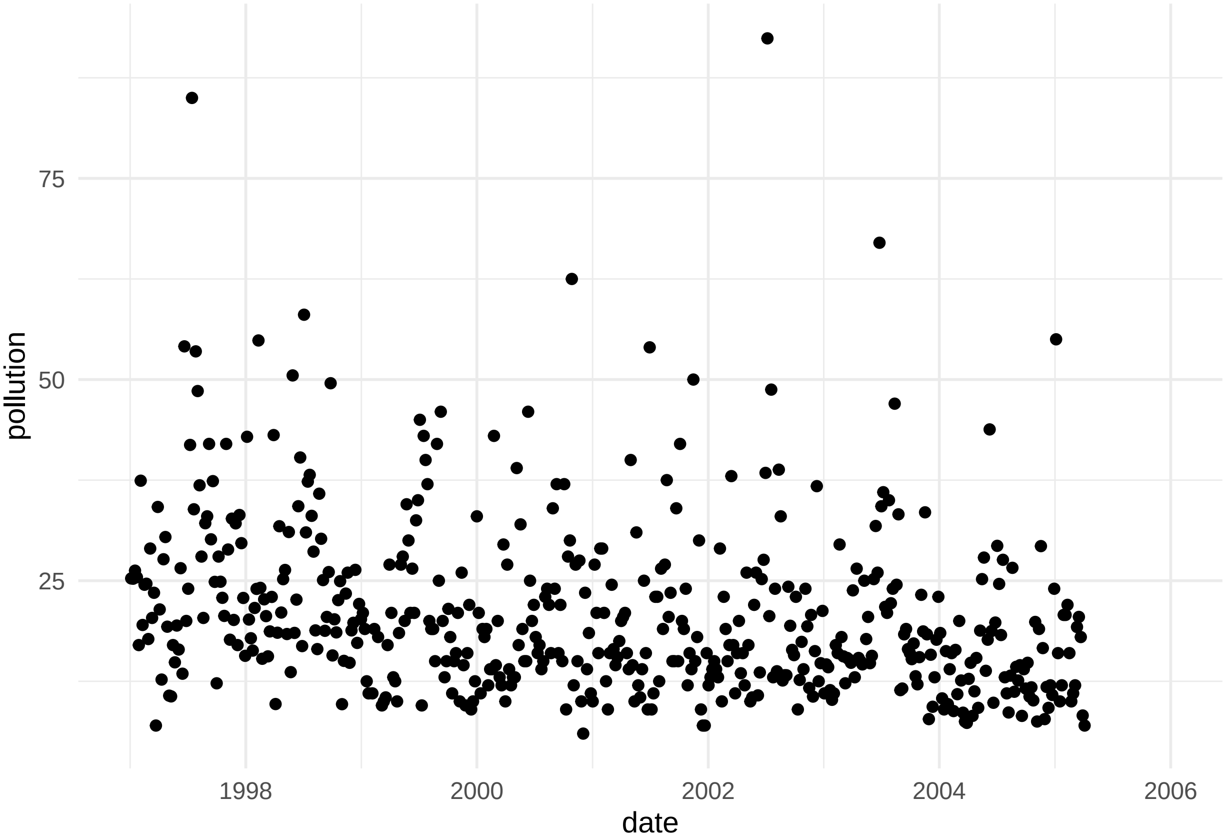
```
dat %>%
  slice_max(pollution, n = 1, with_ties = TRUE)
```

```
## # A tibble: 1 x 3
##   date       death pollution
##   <date>     <dbl>     <dbl>
## 1 2002-07-07   171      92.4
```

Supporting Premise

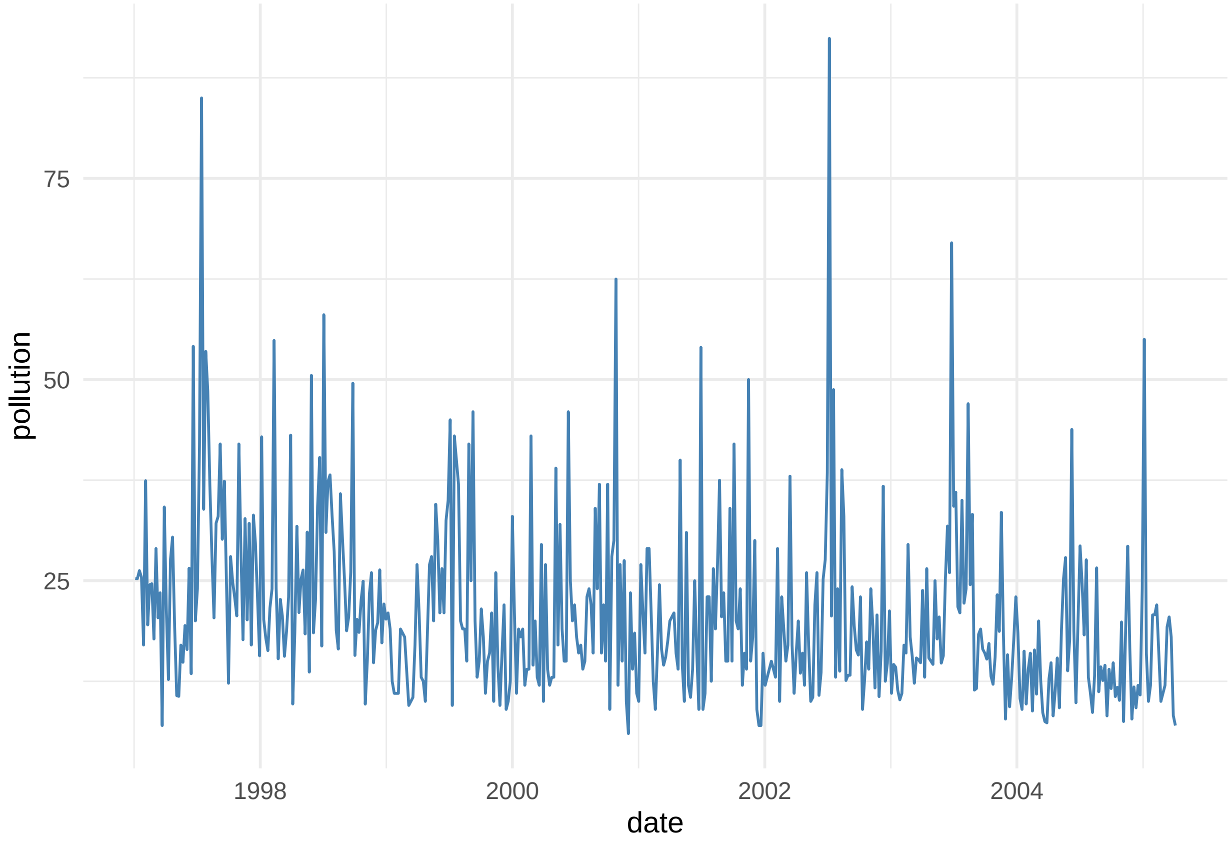
Scatter plot showing the maximum pollution value was in the middle of 2002.

```
ggplot(dat, aes(x = date, y = pollution)) +
  geom_point() +
  theme_minimal()
```



Line graph showing a maximum pollution in the middle of 2002.

```
dat %>%
  filter(!is.na(pollution)) %>%
  ggplot(aes(x = date, y = pollution)) +
  geom_line(color = "steelblue") +
  theme_minimal()
```



Analysis

The output from the supporting premise allows us to visually see that the maximum pollution level from 1997-2005 in Detroit was in July 2002.

The two alternate views of the scatter plot and line plot confirm these findings. However, it is also important to note that the data set contained many NAs for pollution. Therefore, these supporting visuals only show the maximum pollution levels from the data given.