NYPD CCRB Allegations Analysis

Kanit Mann

Interview Question: In the dataset, City Civilian Complaint Review Board (CCRB) Allegations, what does a 0 mean in the column "officer days on force at incident"? How might your interpretation of this data change if you thought the 0 was a data artifact versus if you thought it was real data?

Answer: In the "CCRB Allegations Against Police" dataset, the variable "officer days on force at incident" records the number of days an officer had served at the time of the incident. Upon exploration, I found that the value 0 appears in this field for a large number of records (see Table 1).

Table 1: Distribution of Officer Days on Force at Incident

Days_on_Force	Count	Percent
0	140942	35.8
1–30	4	0.0
31–180	60	0.0
181+	232519	59.1
NA	19942	5.1

The 0 in the "officer days on force at incident" could mean a couple of things in the data, depending on how we interpret it.

1. **0 as a Data Artifact:**

- The first case, which is most likely one, is that 0 acts as a Data Artifact, which means it is an error or missing data placeholder (e.g., value not recorded, default 0), and hence needs to be excluded or flagged when performing any analysis on the data to avoid bias, or inaccuracy.

2. **0 as Real Value(First Day):**

- 0 could also be reflecting a true first-day scenario, which means the incident literally occurred on officer's first day of duty. This could highlight important issues in early training or onboarding, although such cases are likely rare.

In summary, upon careful examination, the value "0" likely reflects a data artifact in most cases, but might occasionally might represent a real scenario. The distinction is critical, treating artifact zeros as real experience could bias findings about officer tenure and incident outcomes. What is intruiging is, that we have N/A available too, which could have been used as a placeholder at all positions. Another reason why I believe 0 is a placehold is because of the uneven spread after Day 0. If the officers was actually abusing their power, it should have been reflected for days following day 0 as well, which is clearly not the case.

My Code:

```
#
library(readr)

# Reading Data and specifying all columns as characters (avoids factor problems)
officer_data <- read_csv("./CCB allegations.csv", col_types = cols(.default = "c"))

# Extract and clean numeric officer days
raw_col <- officer_data[["Officer Days On Force At Incident"]]
cleaned <- gsub(",", "", trimws(raw_col))
days_numeric <- as.numeric(cleaned)

# Bin
bins <- cut(
  days_numeric,
  breaks = c(-Inf, 0, 30, 180, Inf),
  labels = c("0", "1-30", "31-180", "181+"),
  right = TRUE
)

# Table and percent
tab <- as.data.frame(table(bins, useNA = "ifany"))
tab$Percent <- round(100 * tab$Freq / sum(tab$Freq), 1)
names(tab) <- c("Days_on_Force", "Count", "Percent")

library(knitr)
knitr::kable(tab, caption = "Distribution of Officer Days on Force at Incident")

# Declaration: I have used AI assistance to understand and troubleshoot few issues,
#such as what is chunks, syntax for data extraction, how to create chunks, and
#hiding/showing code outputs as this is my first time working on Rmarkdown, but
#all the code and content that is submissed is understood, and written by me.
```