JF_assignment1

Jeremy Fields

2026-01-22

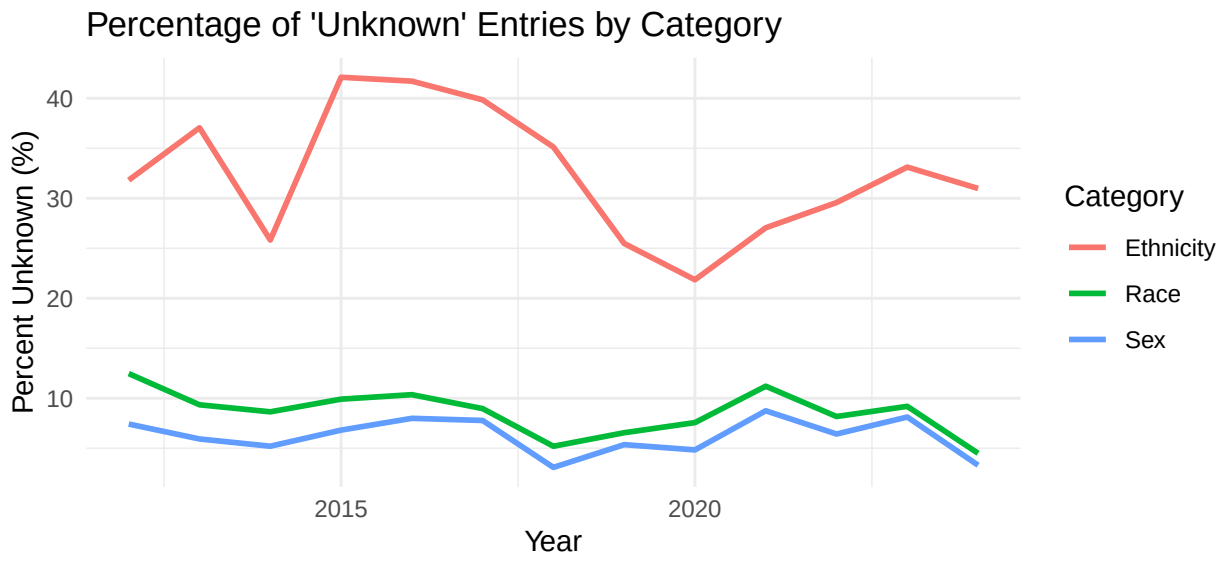
Analysis of Data Validity Trends (2012-2024)

The Marijuana Arrests dataset contains significant validity issues, primarily consisting of inconsistent "Unknown" markers (NAs, blanks, and "U") and trailing whitespace typos ("W" vs "W"). After standardizing these values, several patterns emerge regarding data collection quality over time.

Key Findings:

- **General Trend:** While total arrest volume dropped significantly after 2014, the quality of data entry generally improved for **Sex** and **Race**.
- **Race & Sex:** These categories saw a reduction in the proportion of unknown entries, suggesting better data intake procedures in recent years.
- **Ethnicity:** This remains the least valid category. Despite the decrease in total arrests, roughly 30% of records continue to be marked as "Unknown," a trend that has remained stubbornly high since 2012.
- **Independent Research:** The significant drop in total marijuana-related arrest records after 2014 corresponds to the fact that Initiative 71 became effective on February 26, 2025. Initiative 71 legalized small quantities of marijuana for adults 21+ years old.
- **Data Validity:** While the total number of arrests decreased over time (due to legalization), the validity only improved for RACE and SEX, while ETHNICITY data remained "invalid" / "unknown" at a high rate.

Source: Marijuana arrest data. (n.d.). Mpdc.



Appendix: Code

```r
# Load data
m_arrests_dc <- read.csv("~/INFO526 - Data Viz/Marijuana_Arrests.csv",
                         comment.char="#")

# Import libraries
library(dplyr)
library(ggplot2)

# Cleaning logic
m_cleaned <- m_arrests_dc %>%
  mutate(
    # Clean RACE: Trim spaces and group unknowns
    Race_Clean = trimws(RACE),
    Race_Status = case_when(
      is.na(Race_Clean) | Race_Clean == "" | Race_Clean == "U" ~ "Unknown",
      TRUE ~ "Valid"
    ),

    # Clean ETHNICITY: Trim spaces and group unknowns
    Eth_Clean = trimws(ETHNICITY),
    Eth_Status = case_when(
      is.na(Eth_Clean) | Eth_Clean == "" | Eth_Clean == "U" ~ "Unknown",
      TRUE ~ "Valid"
    ),

    # Clean SEX: Group unknowns
    Sex_Status = case_when(
      is.na(SEX) | SEX == "" | SEX == "U" ~ "Unknown",
      TRUE ~ "Valid"
    )
  )

# Analyze unknown vs valid trends over time
table(m_cleaned$YEAR, m_cleaned$Race_Status)
table(m_cleaned$YEAR, m_cleaned$Eth_Status)
table(m_cleaned$YEAR, m_cleaned$Sex_Status)


# Validate ratio of valid / unknown data is changing over time to ensure data quality is also improving
validity_trends <- m_cleaned %>%
  group_by(YEAR) %>%
  summarize(
    Total = n(),
    Race_Unknown_Pct = mean(Race_Status == "Unknown") * 100,
    Eth_Unknown_Pct = mean(Eth_Status == "Unknown") * 100,
    Sex_Unknown_Pct = mean(Sex_Status == "Unknown") * 100
  )
```