CEL Data Assessment

Rhea Mendiratta

2025-03-14

Setting a working directory

```
setwd("/Users/rhea/Library/Mobile Documents/com~apple~CloudDocs/1. UChic/Jobs/CEL Data Assessment 2025")
```

Installing libraries I'll need

```
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   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(lubridate)
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:data.table':
##
##     yearmon, yearqtr
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
library(broom)
```

Reading data

```
program_cases <- read.csv("program_cases.csv")
str(program_cases)
```

```
## 'data.frame':    5000 obs. of  7 variables:
##  $ person_id           : int  5 16 47 49 57 59 65 71 72 74 ...
##  $ program_date        : chr  "2020-10-20" "2020-07-05" "2020-12-02" "2020-02-26" ...
##  $ treatment           : int  0 0 0 1 1 1 0 1 1 1 ...
##  $ age_at_enrollment   : int  49 50 55 35 32 35 50 26 27 36 ...
##  $ arrest_after_program: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ gender              : chr  "F" "F" "F" "F" ...
##  $ race                : chr  "BLACK" "BLACK" "BLACK" "BLACK" ...
```

Part A: Data Management

Question 1 Recoding the "race" variable into the following two variables: 1. A binary variable (hispanic_ethnicity) where "1" denotes that the individual's original race category includes "Hispanic" and "0" denotes otherwise. 2. A new race variable (race_only) with no "Hispanic" distinction

```
program_cases <- program_cases %>%
  mutate(
    hispanic_ethnicity = case_when(
      grepl("HISPANIC", race) ~1,
      TRUE ~ 0
    ),
    race_only = case_when(
   grepl("HISPANIC", race) ~ gsub(" HISPANIC", "", race),
   TRUE ~ race
    )
  )
```

Calculating the number of arrestees in the full sample of Hispanic ethnicity

```
hispanic_count <- sum(program_cases$hispanic_ethnicity)
print(hispanic_count)
```

```
## [1] 848
```

Question 2 Reading data

```
arrests <- read.csv("arrests.csv")
crimes <- read.csv("crimes.csv")
```

Merging the arrests dataset with the crimes dataset using "case_number" as the matching variable

```
merged_data <- arrests %>%
  inner_join(crimes, by = "case_number")
str(merged_data)
```

```
## 'data.frame':    38128 obs. of  22 variables:
##  $ case_number         : chr  "JA173553" "JB512615" "JA146950" "JA413888" ...
##  $ arrest_date         : chr  "2017-03-03T18:20:00Z" "2018-11-12T06:55:00Z" "2017-02-09T04:10:00Z" "
##  $ charge_1_statute    : chr  "720 ILCS 570.0/407-B-1" "720 ILCS 550.0/5-D" "720 ILCS 5.0/21-3-A-2" "
##  $ charge_1_description: chr  "MFG/DEL HEROIN/SCH/PUB HS/PK" "CANNABIS - MFG/DEL - 30-500 GRMS" "CRI
##  $ charge_1_type       : chr  "F" "F" "M" "M" ...
##  $ charge_2_statute    : chr  NA NA NA NA ...
##  $ charge_2_description: chr  NA NA NA NA ...
##  $ charge_2_type       : chr  NA NA NA NA ...
##  $ charge_3_statute    : chr  NA NA NA NA ...
##  $ charge_3_description: chr  NA NA NA NA ...
##  $ charge_3_type       : chr  NA NA NA NA ...
##  $ charge_4_statute    : chr  NA NA NA NA ...
##  $ charge_4_description: chr  NA NA NA NA ...
##  $ charge_4_type       : chr  NA NA NA NA ...
##  $ person_id           : int  64697 30661 79927 58241 91982 21350 17196 78943 39851 49108 ...
##  $ id                  : int  10940485 11504740 10843587 11073595 10843681 11278861 11742947 1120546
##  $ date                : chr  "2017-03-03T18:15:00Z" "2018-11-12T06:52:00Z" "2017-02-09T03:57:00Z" "
##  $ iucr                : chr  "2014" "1822" "1330" "0470" ...
##  $ primary_type        : chr  "NARCOTICS" "NARCOTICS" "CRIMINAL TRESPASS" "PUBLIC PEACE VIOLATION" .
##  $ fbi_code            : chr  "18" "18" "26" "24" ...
##  $ fbi_description     : chr  "Drug Abuse" "Drug Abuse" "Misc Non-Index Offense" "Disorderly Conduct
##  $ address             : chr  "0000X S WHIPPLE ST, CHICAGO IL" "006XX E 79TH ST, CHICAGO IL" "070XX
```

No. of rows in the resulting dataset

```
merged_rows <- nrow(merged_data)
print(merged_rows)
```

```
## [1] 38128
```

Question 3 Filtering the merged_data to include only crimes that occurred in Chicago using the "address" variable

```
chicago_merged_data <- merged_data %>%
  filter(grepl("CHICAGO", address))
```

Rows in the new dataset

```r
chicago_merged_rows <- nrow(chicago_merged_data)
print(chicago_merged_rows)
```

```
## [1] 35797
```

Part B: Variable Creation

Question 4 Defining the statutes and creating indicator variables by checking all "charge_x_statute" columns for the specified statutes.

```r
homicide_statutes <- c("720 ILCS 5.0/9-1-A-1", "720 ILCS 5.0/9-1-A-2", "720 ILCS 5.0/9-1-A-3")
weapons_statutes <- c("720 ILCS 5.0/24-1.6-A-1", "720 ILCS 5.0/24-1.1-A")

chicago_merged_data <- chicago_merged_data %>%
  mutate(
    homicide_charge = as.integer(if_any(matches("charge_\\d+_statute"), ~ .x %in% homicide_statutes)),
    weapons_charge = as.integer(if_any(matches("charge_\\d+_statute"), ~ .x %in% weapons_statutes))
  )
```

Number of arrests with a homicide charge

```r
homicide_number <- sum(chicago_merged_data$homicide_charge, na.rm = TRUE)
print(homicide_number)
```

```
## [1] 228
```

```r
#verifying that my code for weapons_charge worked
weapons_number <- sum(chicago_merged_data$weapons_charge, na.rm = TRUE)
print(weapons_number)
```

```
## [1] 2973
```

Question 5 Creating a separate dataset containing only post-program arrests by using person_id to find each individual's program date. Then, getting all unique person_ids with a post-program arrest to create a fresh indicator variable. Then, merging this data with the program_cases dataset and setting our indicator to 0 for anyone who doesn't appear in the post-program arrest dataset.

```r
recreated_arrest_data <- chicago_merged_data %>%
  left_join(program_cases %>% select(person_id, program_date), by = "person_id") %>%
  filter(arrest_date > program_date) %>%
  distinct(person_id) %>%
  mutate(recreated_arrest_after_program = 1)

program_cases <- program_cases %>%
  left_join(recreated_arrest_data%>% select(person_id, recreated_arrest_after_program),
            by = "person_id") %>%
  mutate(recreated_arrest_after_program = ifelse(is.na(recreated_arrest_after_program), 0, 1))
```

Comparing the original arrest_after_program variable with the recreated_arrest_after_program variable

```r
match_percentage <- mean(program_cases$arrest_after_program == program_cases$recreated_arrest_after_prog
print(match_percentage)
```

```
## [1] 100
```

Part C

Question 6 Creating a balance table comparing demographics between the treatment and control groups

```r
#finding all unique values in the race_only variable
unique(program_cases$race_only)
```

```
## [1] "BLACK"                 "WHITE"                 "OTHER/UNKNOWN/REFUSED"
```

```r
balance_table <- program_cases %>%
  group_by(treatment) %>%
  summarise(
    n = n(),
    mean_age = mean(age_at_enrollment, na.rm = TRUE),
    prop_male = mean(gender=="M", na.rm = TRUE),
    prop_hispanic = mean(hispanic_ethnicity, na.rm = TRUE),
    prop_white = mean(race_only == "WHITE", na.rm = TRUE),
    prop_black = mean(race_only == "BLACK", na.rm = TRUE),
    prop_other_race = mean(race_only == "OTHER/UNKNOWN/REFUSED", na.rm = TRUE)
  )

print(balance_table)
```

```
## # A tibble: 2 x 8
##   treatment     n mean_age prop_male prop_hispanic prop_white prop_black
##       <int> <int>    <dbl>     <dbl>         <dbl>      <dbl>      <dbl>
## 1         0  2500     49.5     0.48          0.165      0.258      0.733
## 2         1  2500     29.4     0.475         0.174      0.265      0.731
## # i 1 more variable: prop_other_race <dbl>
```

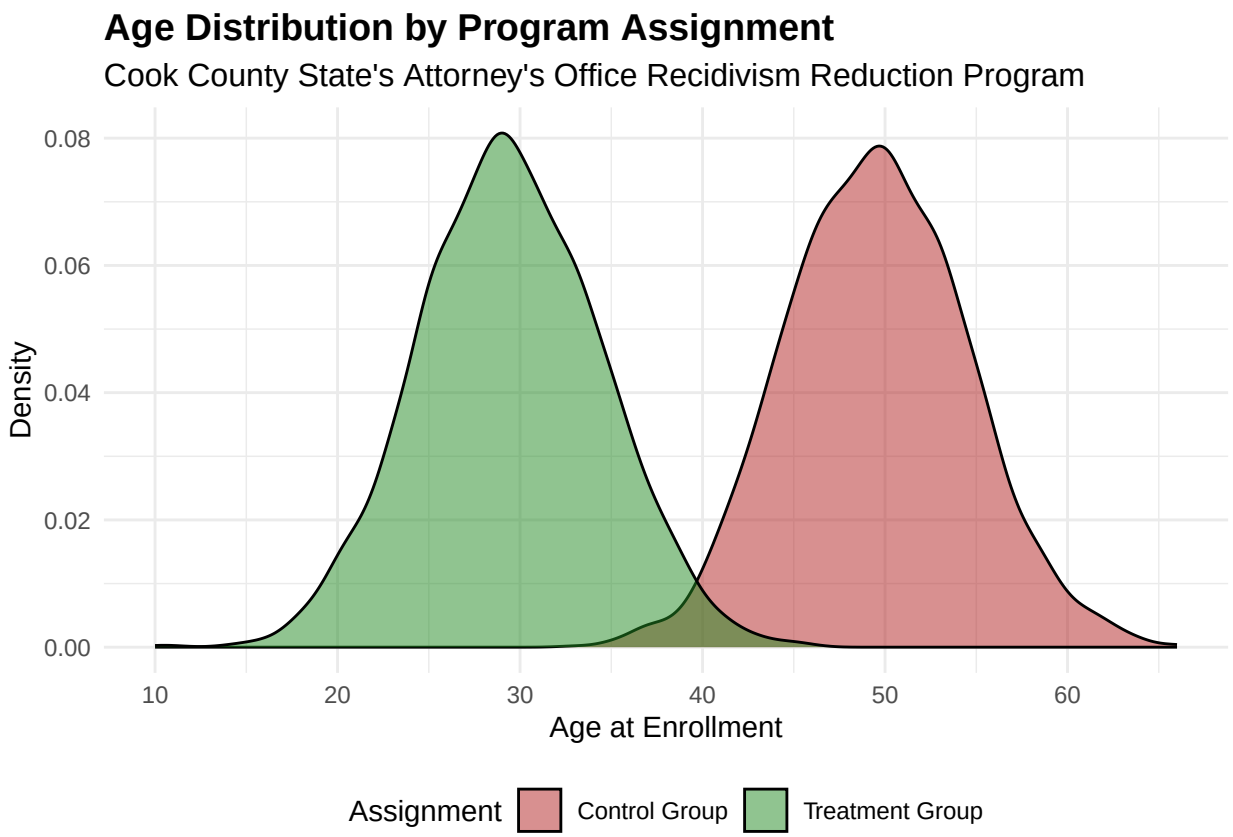
Conducting Statistical Tests (t-test for age and chi-square for others) and compiling p-values

```r
age_test <- t.test(age_at_enrollment ~ treatment, data = program_cases)
gender_test <- chisq.test(table(program_cases$gender, program_cases$treatment))
hispanic_test <- chisq.test(table(program_cases$hispanic_ethnicity, program_cases$treatment))
race_test <- chisq.test(table(program_cases$race_only, program_cases$treatment))

balance_p_values <- data.frame(
  Variable = c("Age", "Gender", "Hispanic", "Race"),
  P_Value = c(age_test$p.value, gender_test$p.value, hispanic_test$p.value, race_test$p.value)
)

print(balance_p_values)
```

```
##   Variable   P_Value
## 1      Age 0.0000000
## 2   Gender 0.8849579
## 3 Hispanic 0.4287301
## 4     Race 0.1283638
```

Question 7 Creating an age-distribution plot by treatment status

```
ggplot(program_cases, aes(x = age_at_enrollment, fill = factor(treatment))) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("firebrick", "forestgreen"),
                    labels = c("Control Group", "Treatment Group"),
                    name = "Assignment") +
  labs(
    title = "Age Distribution by Program Assignment",
    subtitle = "Cook County State's Attorney's Office Recidivism Reduction Program",
    x = "Age at Enrollment",
    y = "Density"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 12),
    legend.position = "bottom"
  )
```



Question 8 Using a linear regression model to evaluate the relationship between program assignment and re-arrest

```
linear_model <- lm(arrest_after_program ~ treatment, data = program_cases)
summary(linear_model)
```

```
##
## Call:
## lm(formula = arrest_after_program ~ treatment, data = program_cases)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3516 -0.3516 -0.3472  0.6484  0.6528
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.351600   0.009537  36.865   <2e-16 ***
## treatment   -0.004400   0.013488  -0.326    0.744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4769 on 4998 degrees of freedom
## Multiple R-squared:  2.129e-05,  Adjusted R-squared:  -0.0001788
## F-statistic: 0.1064 on 1 and 4998 DF,  p-value: 0.7443
```

```
adjusted_model <- lm(arrest_after_program ~ treatment + age_at_enrollment +
                       gender + race_only + hispanic_ethnicity,
                       data = program_cases)
summary(adjusted_model)
```

```
##
## Call:
## lm(formula = arrest_after_program ~ treatment + age_at_enrollment +
##     gender + race_only + hispanic_ethnicity, data = program_cases)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5212 -0.3625 -0.3039  0.6110  0.8213
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.884545   0.068230  12.964  < 2e-16 ***
## treatment                       -0.220825   0.030394  -7.265 4.29e-13 ***
## age_at_enrollment               -0.010754   0.001359  -7.912 3.10e-15 ***
## genderM                         -0.005874   0.013546  -0.434    0.665
## genderOTHER/REFUSED              0.064086   0.049571   1.293    0.196
## race_onlyOTHER/UNKNOWN/REFUSED -0.113746   0.085558  -1.329    0.184
## race_onlyWHITE                  -0.001090   0.022294  -0.049    0.961
## hispanic_ethnicity               0.014145   0.026095   0.542    0.588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4739 on 4992 degrees of freedom
```

```
## Multiple R-squared:  0.01347,    Adjusted R-squared:  0.01208
## F-statistic: 9.735 on 7 and 4992 DF,  p-value: 4.266e-12
```