DDS Expenditure Equity Analysis

Exploratory, permutation, and regression analysis of California DDS service expenditures

Aryan Bhojani

```r
library(tidyverse)
library(broom)
library(hexbin)
```

Project overview

This project analyzes a de-identified sample of California Department of Developmental Services (DDS) client records to evaluate whether service expenditures differ systematically by ethnicity and gender, and to understand how age structure can drive misleading aggregate comparisons (a Simpson's paradox style pattern).

The workflow has three parts:

1) Exploratory comparisons of expenditures across ethnicity, age, and gender
2) A permutation test within a fixed age cohort comparing minority vs non-minority expenditures
3) A regression model of log expenditures adjusting for age cohort, gender, and ethnicity

All results are descriptive associations based on a de-identified administrative dataset.

```r
dds <- read_csv("data/california-dds2.csv", show_col_types = FALSE)

names(dds) <- names(dds) |>
  stringr::str_replace_all("\u00A0", " ") |>
  stringr::str_squish()

glimpse(dds)
```

```
Rows: 1,000
Columns: 6
$ Id           <dbl> 10210, 10409, 10486, 10538, 10568, 10690, 10711, 10778, 1~
$ Age_cohort   <chr> "13 to 17", "22 to 50", "0 to 5", "18 to 21", "13 to 17",~
$ Age          <dbl> 17, 37, 3, 19, 13, 15, 13, 17, 14, 13, 13, 14, 15, 17, 20~
$ Gender       <chr> "Female", "Male", "Male", "Female", "Male", "Female", "Fe~
$ Expenditures <dbl> 2113, 41924, 1454, 6400, 4412, 4566, 3915, 3873, 5021, 28~
$ Ethnicity    <chr> "White not Hispanic", "White not Hispanic", "Hispanic", "~
```

```r
cohort_levels <- c("0 to 5", "6 to 12", "13 to 17", "18 to 21", "22 to 50", "51+")

dds_cat <- dds |>
  filter(!is.na(Age_cohort)) |>
  filter(Age_cohort %in% cohort_levels) |>
  mutate(across(c(Age_cohort, Ethnicity, Gender), as.factor)) |>
  mutate(Age_cohort = factor(Age_cohort, levels = cohort_levels))
```

Exploratory analysis

Median expenditures by ethnicity

I start by comparing median annual expenditures across ethnic groups and recording sample size per group.

```r
median_exp_by_eth <- dds_cat |>
  group_by(Ethnicity) |>
  summarise(
    n = n(),
    median_expenditure = median(Expenditures, na.rm = TRUE),
    mean_expenditure = mean(Expenditures, na.rm = TRUE),
    .groups = "drop"
  ) |>
  arrange(desc(median_expenditure))

median_exp_by_eth
```

```
# A tibble: 8 x 4
  Ethnicity              n median_expenditure mean_expenditure
  <fct>              <int>              <dbl>            <dbl>
1 Native Hawaiian        2              39103            39103
```

```
2 American Indian       2              15966.          15966.
3 White not Hispanic   335             10068           19186.
4 Asian               116              8204           14332.
5 Black                52              5454.          16425.
6 Hispanic            359              3745            8957.
7 Other                 2              3316.           3316.
8 Multi Race           26              2622            4457.
```

Figure 1: Median expenditures by ethnicity (log scale)

This plot orders ethnicities from highest to lowest median expenditure and uses a log scale to display the wide spread.

```
median_exp_by_eth_plot <- median_exp_by_eth |>
  mutate(Ethnicity = reorder(Ethnicity, -median_expenditure))

fig_1 <- ggplot(median_exp_by_eth_plot, aes(x = Ethnicity, y = median_expenditure)) +
  geom_line(aes(group = 1)) +
  geom_point() +
  scale_y_log10() +
  labs(
    x = NULL,
    y = "Median expenditure (log scale)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

fig_1
```

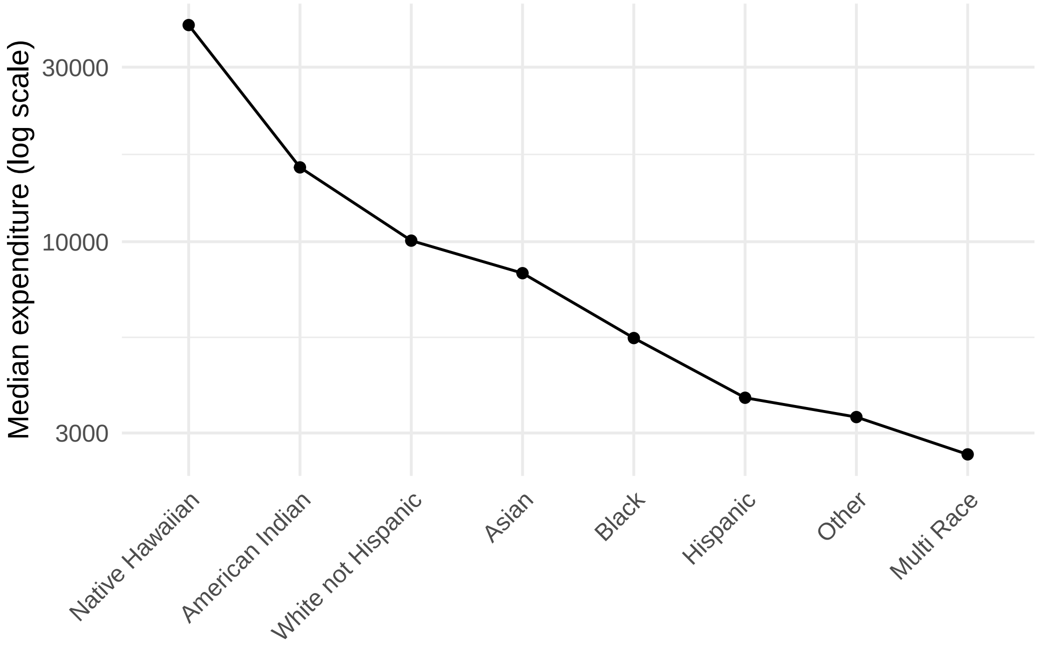


Figure 2: Expenditures vs age (hexbin, log scale)

This heatmap shows the overall relationship between age and expenditures. The log scale helps visualize extreme right skew.

```r
fig_2 <- ggplot(dds, aes(x = Age, y = Expenditures)) +
  geom_hex() +
  scale_y_log10() +
  labs(
    x = "Age (years)",
    y = "Expenditures (log scale)",
    fill = "Count"
  ) +
  theme_minimal()

fig_2
```

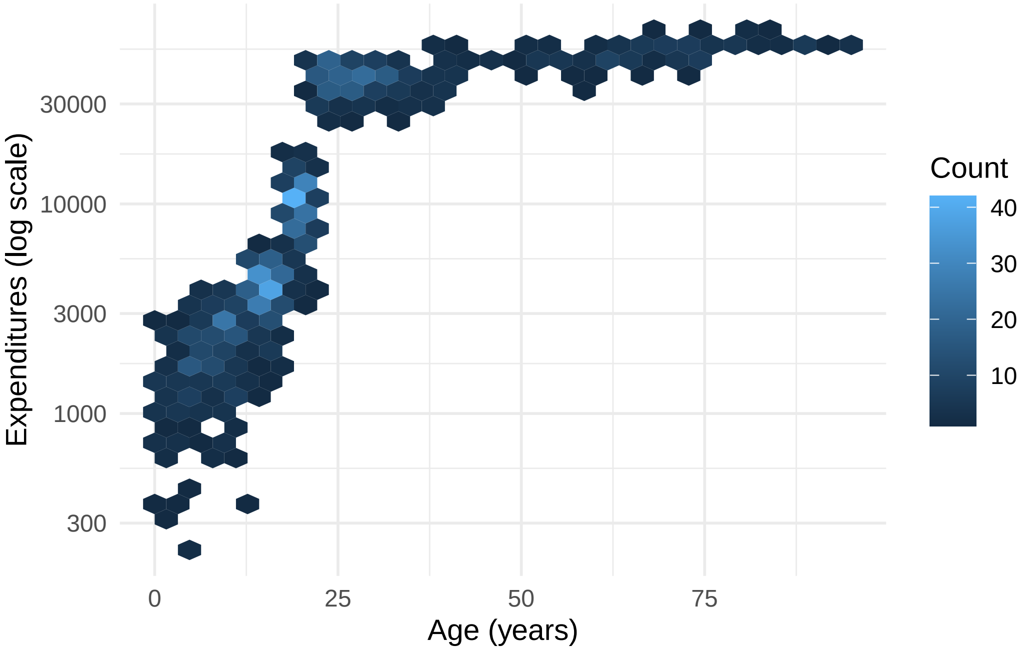


Figure 3: Sample size by age cohort and ethnicity

Different ethnic groups have different age composition in the sample, which can confound aggregate expenditure comparisons.

```r
samp_sizes <- dds_cat |>
  group_by(Age_cohort, Ethnicity) |>
  summarise(n = n(), .groups = "drop")

fig_3 <- ggplot(samp_sizes, aes(x = Age_cohort, y = n, color = Ethnicity, group = Ethnicity)
  geom_line() +
  geom_point() +
  labs(
    x = "Age cohort",
    y = "Sample size"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

fig_3
```

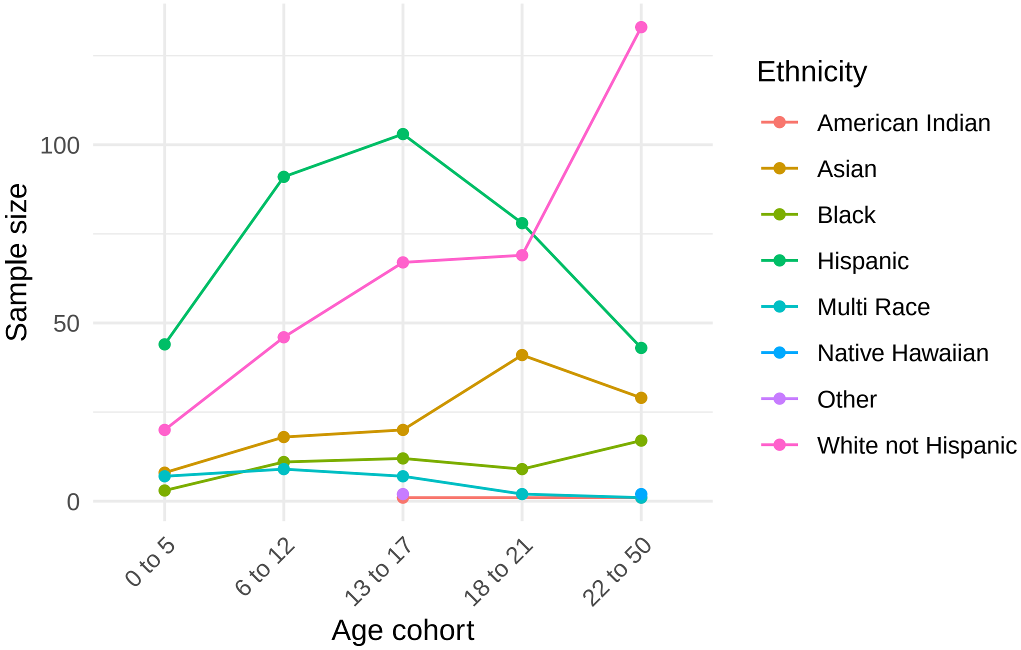
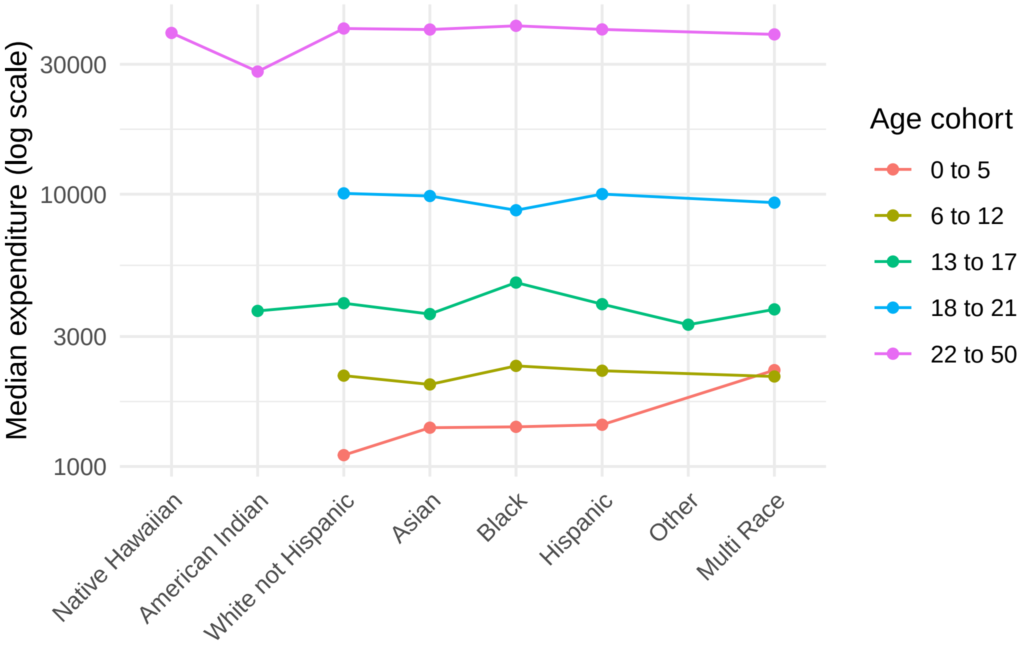


Figure 4: Median expenditures by ethnicity within each age cohort

This plot compares ethnicities within age cohorts (color), reducing confounding due to different age composition.

```
eth_order <- dds_cat |>
  group_by(Ethnicity) |>
  summarise(med = median(Expenditures, na.rm = TRUE), .groups = "drop") |>
  arrange(desc(med)) |>
  pull(Ethnicity)

dds_cat_ordered <- dds_cat |>
  mutate(Ethnicity = factor(Ethnicity, levels = eth_order))

med_by_eth_age <- dds_cat_ordered |>
  group_by(Ethnicity, Age_cohort) |>
  summarise(median_expenditure = median(Expenditures, na.rm = TRUE), .groups = "drop")

fig_4 <- ggplot(med_by_eth_age, aes(x = Ethnicity, y = median_expenditure, color = Age_cohor
  geom_line() +
  geom_point() +
  scale_y_log10() +
  labs(
    x = NULL,
    y = "Median expenditure (log scale)",
    color = "Age cohort"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

fig_4
```



Statistical inference: permutation test within a fixed age cohort

To reduce confounding, I focus on a single age cohort (13 to 17) and test whether mean expenditures differ between:

- Non-minority: White not Hispanic
- Minority: all other ethnicities

This is a permutation test: shuffle the group labels many times to approximate the null distribution of the mean difference.

Create minority indicator and subset to age 13 to 17

```r
dds_perm <- dds_cat |>
  mutate(
    Minority = if_else(Ethnicity == "White not Hispanic", "Non-minority", "Minority"),
    Minority = factor(Minority, levels = c("Non-minority", "Minority"))
  ) |>
  filter(Age_cohort == "13 to 17")

dds_perm |>
  count(Minority)
```

```
# A tibble: 2 x 2
  Minority         n
  <fct>        <int>
1 Non-minority    67
2 Minority       145
```

Observed mean difference function (base R)

```r
compute_mean_diff <- function(dat) {
  is_minority <- dat$Minority == "Minority"
  mean_minority    <- mean(dat$Expenditures[is_minority],  na.rm = TRUE)
  mean_nonminority <- mean(dat$Expenditures[!is_minority], na.rm = TRUE)
  mean_minority - mean_nonminority
}

obs_diff <- compute_mean_diff(dds_perm)
obs_diff
```

```
[1] 26.69007
```

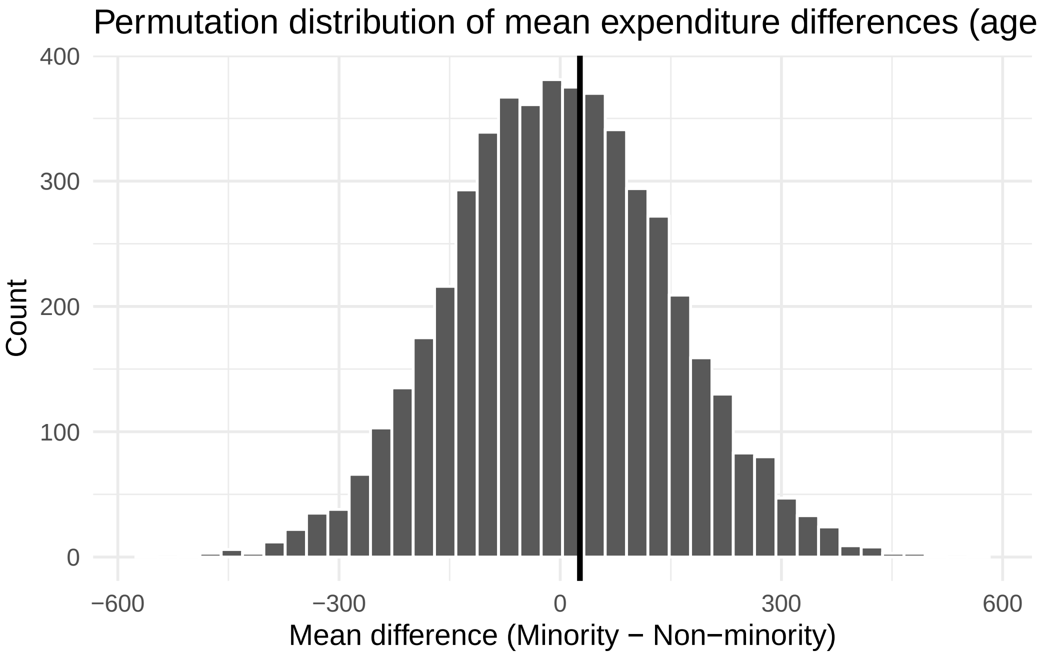
Permutation distribution and p-value

```r
set.seed(2004)

n_perm <- 5000
```

```r
perm_diffs <- numeric(n_perm)

for (b in 1:n_perm) {
  perm_dat <- dds_perm
  perm_dat$Minority <- sample(perm_dat$Minority)
  perm_diffs[b] <- compute_mean_diff(perm_dat)
}

p_val <- mean(abs(perm_diffs) >= abs(obs_diff))
p_val
```

```
[1] 0.8638
```

```r
perm_df <- tibble(diff = perm_diffs)

fig_perm <- ggplot(perm_df, aes(x = diff)) +
  geom_histogram(bins = 40, color = "white") +
  geom_vline(xintercept = obs_diff, linewidth = 1) +
  labs(
    x = "Mean difference (Minority - Non-minority)",
    y = "Count",
    title = "Permutation distribution of mean expenditure differences (age 13 to 17)"
  ) +
  theme_minimal()

fig_perm
```



Regression analysis: adjusting for age cohort, gender, and ethnicity

To quantify differences while adjusting for confounding, I fit a linear model for log expenditures:

log(Expenditures) = f(Age cohort) + f(Gender) + f(Ethnicity) + error

This is an association model (not causal). Coefficients represent differences relative to reference categories.

Prepare reference categories

- Age cohort baseline: 0 to 5
- Ethnicity baseline: White not Hispanic

```
dds_reg <- dds_cat |>
  mutate(
    Ethnicity = relevel(Ethnicity, ref = "White not Hispanic")
  ) |>
  filter(Expenditures > 0)

levels(dds_reg$Age_cohort)
```

```
[1] "0 to 5"   "6 to 12"  "13 to 17" "18 to 21" "22 to 50" "51+"
```

```r
levels(dds_reg$Ethnicity)[1:5]
```

```
[1] "White not Hispanic" "American Indian"    "Asian"
[4] "Black"              "Hispanic"
```

Fit model and extract coefficients

```r
expenditures_lm <- lm(
  log(Expenditures) ~ Age_cohort + Gender + Ethnicity,
  data = dds_reg
)

coef_tbl <- tidy(expenditures_lm)
coef_tbl
```

```
# A tibble: 13 x 5
   term                     estimate std.error statistic   p.value
   <chr>                       <dbl>     <dbl>     <dbl>     <dbl>
 1 (Intercept)                7.13      0.0440  162.     0
 2 Age_cohort6 to 12          0.490     0.0461   10.6    5.90e- 25
 3 Age_cohort13 to 17         1.10      0.0450   24.5    1.85e-101
 4 Age_cohort18 to 21         2.02      0.0457   44.3    1.95e-226
 5 Age_cohort22 to 50         3.47      0.0458   75.7    0
 6 GenderMale                -0.0370    0.0230   -1.61   1.08e-  1
 7 EthnicityAmerican Indian  -0.166     0.244    -0.679  4.97e-  1
 8 EthnicityAsian            -0.0283    0.0374   -0.757  4.49e-  1
 9 EthnicityBlack             0.0409    0.0513    0.798  4.25e-  1
10 EthnicityHispanic          0.0350    0.0274    1.28   2.01e-  1
11 EthnicityMulti Race        0.0381    0.0713    0.534  5.93e-  1
12 EthnicityNative Hawaiian  -0.0120    0.244    -0.0490 9.61e-  1
13 EthnicityOther            -0.193     0.245    -0.788  4.31e-  1
```

Bootstrap standard error for the 6 to 12 coefficient

This estimates uncertainty for the age 6 to 12 indicator by resampling individuals with replacement and refitting the model.

```
set.seed(2004)

n_boot <- 2000
boot_beta <- numeric(n_boot)

for (i in 1:n_boot) {
  idx <- sample(seq_len(nrow(dds_reg)), size = nrow(dds_reg), replace = TRUE)
  boot_dat <- dds_reg[idx, ]
  fit_boot <- lm(log(Expenditures) ~ Age_cohort + Gender + Ethnicity, data = boot_dat)
  boot_beta[i] <- coef(fit_boot)["Age_cohort6 to 12"]
}

boot_se <- sd(boot_beta)
boot_se
```

```
[1] 0.06968902
```

Discussion of results

Summary and interpretation

What the raw comparisons show

- Median expenditures vary across ethnic groups in this sample. The highest-median group is Native Hawaiian (median about \$39103, n = 2), while the lowest-median group is Multi Race (median about \$2622, n = 26).
- Important caveat: some ethnicity categories have very small sample sizes, so their medians can be unstable and should not be overinterpreted.
- In this dataset, there are 4 ethnicity categories with fewer than 30 observations.
- Because expenditures are highly right-skewed, medians and log-scale plots are more informative than raw means for comparing typical spending.

Age is strongly related to expenditures

- Expenditures differ sharply by age cohort. The cohort with the lowest median is 0 to 5 (median about \$1380, n = 82), and the cohort with the highest median is 22 to 50 (median about \$40456, n = 226).
- The age vs expenditure hex plot shows a dense band of low expenditures with a long upper tail at many ages, consistent with extreme right skew.

Why aggregate ethnicity comparisons can be misleading

- The sample size plot by age cohort and ethnicity shows that ethnic groups have different age composition.
- If some ethnic groups contain more children (lower typical spending), their overall median can look smaller even if within-cohort spending is similar.
- The within-cohort median plot reduces this confounding and helps separate age structure from ethnicity differences.

Permutation test within age 13 to 17

- Observed mean difference (Minority - Non-minority) in age 13 to 17 is about $26.7.
- Two-sided permutation p-value with 5000 shuffles is about 0.864.
- A large p-value means the observed difference is not unusual under random relabeling, so this subgroup test does not provide strong evidence of a systematic minority vs non-minority mean difference in this cohort.

Regression results adjusting for age cohort, gender, and ethnicity

- The age 6 to 12 coefficient in the log-expenditure model is about 0.49 with model-based SE about 0.046 and bootstrap SE about 0.07.
- After adjusting for age cohort, gender, and ethnicity, age cohort effects remain large, matching the exploratory finding that age is a major driver of spending differences in this sample.

Overall conclusion

The dataset shows apparent differences in raw expenditures across ethnic groups, but strong age effects and different age composition across ethnicities can explain much of the aggregate disparity. The within-cohort visualization, the permutation test in a fixed age cohort, and the adjusted regression model all point to age cohort as the dominant predictor of spending in this sample. Evidence for systematic ethnicity or gender differences is weaker after controlling for age. These results are descriptive associations from de-identified administrative records, not causal estimates.