2025-CAH

Josephine McKelvy

2025-06-30

This R markdown file includes steps to clean and analyze fictional experimental data as part of a hiring exercise for the research analyst/behavioral researcher role at Duke University's Center for Advanced Hindsight (CAH).

Getting Started

```
packages <- c("DataExplorer",
              "dplyr",
              "ggmosaic",
              "ggplot2",
              "knitr",
              "readr",
              "sjPlot",
              "stats",
              "summarytools",
              "tinytex",
              "vcd",
              "waffle")

# Install packages that are not already installed:
installed_packages <- packages %in% rownames(installed.packages())
if (any(installed_packages == FALSE)) {
  install.packages(packages[!installed_packages])
}

# Load libraries:
lapply(packages, library, character.only = TRUE)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: ggplot2
```

```
## Loading required package: grid
```

```
##
## Attaching package: 'vcd'

## The following objects are masked from 'package:ggmosaic':
##
##     mosaic, spine

## [[1]]
## [1] "DataExplorer" "stats"        "graphics"     "grDevices"    "utils"
## [6] "datasets"     "methods"      "base"
##
## [[2]]
## [1] "dplyr"        "DataExplorer" "stats"        "graphics"
"grDevices"
## [6] "utils"        "datasets"     "methods"      "base"
##
## [[3]]
##  [1] "ggmosaic"     "ggplot2"      "dplyr"        "DataExplorer" "stats"
##  [6] "graphics"     "grDevices"    "utils"        "datasets"     "methods"
## [11] "base"
##
## [[4]]
##  [1] "ggmosaic"     "ggplot2"      "dplyr"        "DataExplorer" "stats"
##  [6] "graphics"     "grDevices"    "utils"        "datasets"     "methods"
## [11] "base"
##
## [[5]]
##  [1] "knitr"        "ggmosaic"     "ggplot2"      "dplyr"
"DataExplorer"
##  [6] "stats"        "graphics"     "grDevices"    "utils"
"datasets"
## [11] "methods"      "base"
##
## [[6]]
##  [1] "readr"        "knitr"        "ggmosaic"     "ggplot2"      "dplyr"
##  [6] "DataExplorer" "stats"        "graphics"     "grDevices"    "utils"
## [11] "datasets"     "methods"      "base"
##
## [[7]]
##  [1] "sjPlot"       "readr"        "knitr"        "ggmosaic"     "ggplot2"
##  [6] "dplyr"        "DataExplorer" "stats"        "graphics"
"grDevices"
## [11] "utils"        "datasets"     "methods"      "base"
##
## [[8]]
##  [1] "sjPlot"       "readr"        "knitr"        "ggmosaic"     "ggplot2"
##  [6] "dplyr"        "DataExplorer" "stats"        "graphics"
"grDevices"
## [11] "utils"        "datasets"     "methods"      "base"
##
```

```
## [[9]]
##  [1] "summarytools" "sjPlot"       "readr"        "knitr"
"ggmosaic"
##  [6] "ggplot2"      "dplyr"        "DataExplorer" "stats"
"graphics"
## [11] "grDevices"    "utils"        "datasets"     "methods"      "base"
##
## [[10]]
##  [1] "tinytex"      "summarytools" "sjPlot"       "readr"        "knitr"
##  [6] "ggmosaic"     "ggplot2"      "dplyr"        "DataExplorer" "stats"
## [11] "graphics"     "grDevices"    "utils"        "datasets"     "methods"
## [16] "base"
##
## [[11]]
##  [1] "vcd"          "grid"         "tinytex"      "summarytools" "sjPlot"
##  [6] "readr"        "knitr"        "ggmosaic"     "ggplot2"      "dplyr"
## [11] "DataExplorer" "stats"        "graphics"     "grDevices"    "utils"
## [16] "datasets"     "methods"      "base"
##
## [[12]]
##  [1] "waffle"       "vcd"          "grid"         "tinytex"
"summarytools"
##  [6] "sjPlot"       "readr"        "knitr"        "ggmosaic"     "ggplot2"
## [11] "dplyr"        "DataExplorer" "stats"        "graphics"
"grDevices"
## [16] "utils"        "datasets"     "methods"      "base"
```

Merging, Cleaning, Transforming Data

```
# Set working directory to load and save files:
setwd("~/2025-CAH")

# Load data sets:
Data_set_A <- readr::read_csv("Data set A.csv")

## Rows: 2004 Columns: 3
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): condition, income_level
## dbl (1): identifier
##
## 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.

Data_set_B <- readr::read_csv("Data set B.csv")

## Rows: 1504 Columns: 2
## ── Column specification
```

```
──────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): increased_contribution
## dbl (1): identifier
##
## 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.

# Review the data sets (AKA summarize the dataframes):
summarytools::dfSummary(Data_set_A)

## Data Frame Summary
## Data_set_A
## Dimensions: 2004 x 3
## Duplicates: 1002
##
## ------------------------------------------------------------------------
---------------------------------------
## No   Variable       Stats / Values               Freqs (% of Valid)
Graph                 Valid     Missing
## ---- -------------- --------------------------- --------------------- --
------------------ --------- ---------
## 1    identifier     Mean (sd) : 1500.9 (289.6)   1001 distinct values   :
: : : : : : : : :   1003      1001
##      [numeric]      min < med < max:                                    :
: : : : : : : : :   (50.0%)   (50.0%)
##                     1000 < 1501 < 2000                                 :
: : : : : : : : :
##                     IQR (CV) : 501 (0.2)                               :
: : : : : : : : :
##                                                                        :
: : : : : : : : :
##
## 2    condition      1. control                   509 (50.8%)
IIIIIIIIII            1001      1003
##      [character]    2. recommendation            492 (49.2%)
IIIIIIIII             (50.0%)   (50.0%)
##
## 3    income_level   1. LMI                       516 (51.4%)
IIIIIIIIII            1003      1001
##      [character]    2. non-LMI                   487 (48.6%)
IIIIIIIII             (50.0%)   (50.0%)
## ------------------------------------------------------------------------
---------------------------------------

#There are 1002 duplicates and 1001 distinct identifiers (thus up to 1001
real cases, ranging from 1000 to 2000 as IDs) making up the 2004
observations; not sure what that remaining blank row is.
```

```
summarytools::dfSummary(Data_set_B)
```

```
## Data Frame Summary
## Data_set_B
## Dimensions: 1504 x 2
## Duplicates: 751
##
## ------------------------------------------------------------------------
----------------------------------------------
## No   Variable                 Stats / Values               Freqs (% of
Valid)    Graph                 Valid     Missing
## ---- ----------------------- --------------------------- ---------------
------ -------------------- --------- ---------
## 1    identifier               Mean (sd) : 1989.6 (236.8)   752 distinct
values       :     . : : : :   752       752
##      [numeric]                min < med < max:
: : : : : : : :   (50.0%)   (50.0%)
##                               1457 < 2005.5 < 2381
. : : : : : : : :
##                               IQR (CV) : 411.8 (0.1)
: : : : : : : : :
##
. : : : : : : : : :
##
## 2    increased_contribution   1. 1                        748 (99.5%)
IIIIIIIIIIIIIIIII   752       752
##      [character]              2. closed                     4 ( 0.5%)
(50.0%)   (50.0%)
## ------------------------------------------------------------------------
----------------------------------------------
```

```
#There are 751 duplicates and 752 distinct identifiers (thus up to 752 real cases, ranging from 1457 to 2381 as IDs) making up the 1504 observations; not sure what that remaining blank row is.

# Merge data frames by ID variable, then clean that data set to reduce duplication of efforts:
contiguous <- merge(Data_set_A, Data_set_B, by="identifier")
summarytools::dfSummary(contiguous)
```

```
## Data Frame Summary
## contiguous
## Dimensions: 753125 x 4
## Duplicates: 752753
##
## ------------------------------------------------------------------------
----------------------------------------------
## No   Variable                 Stats / Values               Freqs (% of
Valid)    Graph                 Valid    Missing
```

```
## ---- ----------------------- --------------------------- ---------------
------ -------------------- -------- ----------
## 1    identifier               Mean (sd) : 1783.7 (129.4)   371 distinct
values         . : .   .   .   373      752752
##      [numeric]                min < med < max:
: : : : : : :   (0.0%)   (100.0%)
##                               1457 < 1779 < 2000
. : : : : : : :
##                               IQR (CV) : 210 (0.1)
: : : : : : : :
##
.   : : : : : : : :
##
## 2    condition                1. control                   159 (42.9%)
IIIIIIII              371      752754
##      [character]              2. recommendation            212 (57.1%)
IIIIIIIIII            (0.0%)   (100.0%)
##
## 3    income_level             1. LMI                       191 (51.2%)
IIIIIIIII             373      752752
##      [character]              2. non-LMI                   182 (48.8%)
IIIIIIII              (0.0%)   (100.0%)
##
## 4    increased_contribution   1. 1                         369 (98.9%)
IIIIIIIIIIIIIIIII     373      752752
##      [character]              2. closed                      4 ( 1.1%)
(0.0%)   (100.0%)
## ------------------------------------------------------------------------
-----------------------------------------------
```

```r
# There are multiple duplicates and 371 distinct identifiers common to both
data sets (thus up to 371 cases with conditions and outcomes, ranging from
1457 to 2000 as IDs) for the analytic sample.

# Remove empty rows and duplicate cases. Then summarize this dataframe again:
contiguous <- dplyr::distinct(contiguous)

# There are still 3 rows with missing conditions.
# Keep cases where condition is not missing:
contiguous <- filter(contiguous,!is.na(condition))

# There are 3 character (or raw text) columns (that should be factor
variables) and an identifier that is a double (or real number) data type.
# Save categorical variables as factor with new names for the categories:
contiguous <- contiguous %>%
  mutate(
    condition = factor(condition,
                       levels = c("recommendation","control"), #reorder
factors for the contingency table
                       labels = c("recommendation","informational")),
```

```
    income = factor(income_level,
                            levels = c("non-LMI","LMI")),
    outcome = factor(increased_contribution,
                                    levels = c("1","closed"), #reorder
factors for the contingency table
                                    labels = c("contributed","closed"))
  )

# Keep the renamed variables:
contiguous <-  contiguous %>%
  select(identifier, condition, income, outcome)
```

Exploratory Data Analysis

- https://bookdown.org/lyzhang10/lzhang_r_tips_book/preface.html
- https://geanders.github.io/RProgrammingForResearch/exploring-data-1.html

```
contiguousDE <- dummify(contiguous)
DataExplorer::create_report(contiguousDE) # creates report.html output

##
##
## processing file: report.rmd

##   |                                                      |
|   0%  |                                                     |.
|   2%                                                  |
|..                                                 |   5% [global_options]
|                                                       |...
|   7%                                                  |
|....                                               |  10% [introduce]
|                                                       |....
|  12%                                                  |
|.....                                              |  14% [plot_intro]

##   |                                                       |......
|  17%                                                  |
|.......                                            |  19% [data_structure]
|                                                       |........
|  21%                                                  |
|.........                                          |  24% [missing_profile]

##   |                                                       |..........
|  26%                                                  |
|...........                                        |  29% [univariate_distribution_header]
|                                                       |...........
|  31%                                                  |
|............                                       |  33% [plot_histogram]

##   |                                                       |.............
|  36%                                                  |
```

```
|.............                         |  38% [plot_density]
|                                              |.............
|  40%                                  |
|.............                         |  43% [plot_frequency_bar]
|                                              |.............
|  45%                                  |
|.............                         |  48% [plot_response_bar]
|                                              |.............
|  50%                                  |
|.............                         |  52% [plot_with_bar]
|                                              |.............
|  55%                                  |
|.............                         |  57% [plot_normal_qq]

##   |                                                |.............
|  60%                                  |
|.............                         |  62% [plot_response_qq]
|                                              |.............
|  64%                                  |
|.............                         |  67% [plot_by_qq]
|                                              |.............
|  69%                                  |
|.............                         |  71% [correlation_analysis]

##   |
|.............                         |  74%
|                                              |.............
|  76% [principal_component_analysis]

##   |
|.............                         |  79%
|                                              |.............
|  81% [bivariate_distribution_header]  |
|.............                         |  83%
|
|.............                         |  86% [plot_response_boxplot]
|
|.............                         |  88%
|
|.............                         |  90% [plot_by_boxplot]
|
|.............                         |  93%
|
|.............                         |  95% [plot_response_scatterplot]
|
|.............                         |  98%
|
|.............                         | 100% [plot_by_scatterplot]

## output file: C:/Users/josep/Documents/2025-CAH/report.knit.md
```

```
## "C:/Program Files/RStudio/resources/app/bin/quarto/bin/tools/pandoc" +RTS
-K512m -RTS "C:\Users\josep\Documents\2025-CAH\report.knit.md" --to html4 --
from markdown+autolink_bare_uris+tex_math_single_backslash --output
pandoc453853b241b5.html --lua-filter "C:\Users\josep\AppData\Local\R\win-
library\4.5\rmarkdown\rmarkdown\lua\pagebreak.lua" --lua-filter
"C:\Users\josep\AppData\Local\R\win-
library\4.5\rmarkdown\rmarkdown\lua\latex-div.lua" --lua-filter
"C:\Users\josep\AppData\Local\R\win-
library\4.5\rmarkdown\rmarkdown\lua\table-classes.lua" --embed-resources --
standalone --variable bs3=TRUE --section-divs --table-of-contents --toc-depth
6 --template "C:\Users\josep\AppData\Local\R\win-
library\4.5\rmarkdown\rmd\h\default.html" --no-highlight --variable
highlightjs=1 --variable theme=yeti --mathjax --variable "mathjax-
url=https://mathjax.rstudio.com/latest/MathJax.js?config=TeX-AMS-
MML_HTMLorMML" --include-in-header
"C:\Users\josep\AppData\Local\Temp\RtmpC287D7\rmarkdown-str45382468410d.html"

##
## Output created: report.html

# weak correlations between contribution, condition, and income level
```

Stacked Bar Chart of condition

- https://www.cedricscherer.com/2021/07/05/a-quick-how-to-on-labelling-bar-graphs-in-ggplot2/#dataviz

```r
# Create the values to graph:
condition_ct  <- table(contiguous$condition) #counts for each condition
condition_pct <- prop.table(condition_ct)    #percentages for each condition
condition_tab <- data.frame(condition_ct,condition_pct)
condition_tab <- condition_tab %>%
  rename(condition = Var1, count = Freq, percent = Freq.1) %>%
  select(condition, count, percent) %>%
  arrange(percent)

# Plot:
bar_plot <- ggplot(condition_tab, aes(x = percent, y = condition)) %>% +
  geom_col(fill = "#156082", width = .75, show.legend = FALSE) +
  geom_text(aes(label = paste(round(percent*100),"%")), position =
position_dodge(width = 1), hjust = -0.5) +
  theme_void() +
  theme(axis.ticks = element_blank(),
        axis.text.x = element_blank(),
        axis.text.y = element_text(size = 14, hjust = 1)) +
  labs(title = "Figure 1. Stacked bar chart of conditions")

# Save plot:
ggsave(bar_plot, filename = "plot-bar.png", height = 4, width = 12)
```

Donut chart of income

(not sure why geom_rect didn't work but geom_col did): - https://r-charts.com/part-whole/donut-chart-ggplot2/ - https://rfortherestofus.com/2022/09/how-to-make-a-donut-chart-in-ggplot

```
# Create the values to graph:
# https://r-graph-gallery.com/128-ring-or-donut-plot.html
income_ct  <-  table(contiguous$income) #counts for each income level
income_pct <-  prop.table(income_ct)    #percentages for each income level
income_ymax <- cumsum(income_pct)       #cumulative percentages of each
income level
income_ymin <- c(0, head(income_ymax, n = -1))

donut_tab <- data.frame(income_ct, income_pct,income_ymax,income_ymin)
donut_tab <- donut_tab %>%
  rename(income = Var1, count = Freq, percent = Freq.1) %>%
  select(income, count, percent, income_ymax, income_ymin)

# Plot:
hsize <- 4 # 1=small hole size & thick donut, 10=large hole size & thin donut

donut_tab <- donut_tab %>%
  mutate(x = hsize)

donut_plot <- ggplot(donut_tab, aes(x = hsize, y = percent, fill = income)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label = paste(round(percent*100), "% ", income)), position =
position_stack(vjust = 0.5)) +
  coord_polar(theta = "y",      # A donut chart is a bar chart with polar
coordinates.
              direction = -1) + # Set the direction to -1 so the filled in
part starts at the top and goes clockwise.
  xlim(c(0.2, hsize + 0.5)) +
  scale_fill_manual(values = c("LMI"="#156082","non-LMI"="#E7EAF3")) +
  theme_void() # Removes grid lines and background +
  labs(title = "Figure 2. Donut chart of income levels")
```

```
## $title
## [1] "Figure 2. Donut chart of income levels"
##
## attr(,"class")
## [1] "labels"
```

```
# Save plot:
ggsave(donut_plot, filename = "plot-donut.png")
```

```
## Saving 5 x 4 in image
```

Waffle Chart of action

- https://r-charts.com/part-whole/waffle-chart-ggplot2/#geom-waffle

```
# Create the values to graph:
waffle_ct  <-  table(contiguous$outcome)   #counts for each outcome
waffle_pct <-  prop.table(waffle_ct) * 100 #percentages for each outcome to
standardize to a 10x10 grid plot

# Plot:
waffle_plot <- waffle(waffle_pct, rows = 10,
       colors = c("contributed"="#156082","closed"="#E7EAF3"),
       legend_pos = "none")

# Save plot:
ggsave(waffle_plot, filename = "plot-waffle.png")

## Saving 5 x 4 in image
```

Mosaic Plot of the action by condition and income

- https://rstudio-pubs-static.s3.amazonaws.com/584765_5ab02919bd374db7ad7c58f20a11e86f.html
- https://cran.r-project.org/web/packages/ggmosaic/vignettes/ggmosaic.html

```
# Optional: Create values for labels:
mosaic_ct <- table(contiguous$income, contiguous$condition,
contiguous$outcome)
mosaic_pct <- prop.table(mosaic_ct)
mosaic_tab <- data.frame(mosaic_ct,mosaic_pct)
mosaic_tab <- mosaic_tab %>%
  rename(income = Var1,
         condition = Var2,
         outcome = Var3,
         conditionXincome = Freq) %>%
  select(income, condition, outcome, conditionXincome)

# Create a mosaic plot (or percent stacked bar chart) of the contingency
table:
ggsave(filename = "plot-mosaic_1.png",
       mosaic_1 <- ggplot(data = contiguous) +
         ggmosaic::geom_mosaic(aes(x = product(outcome), fill = income), #
normally, fill is your outcome variable
                   divider = c("vspine","hbar"), # vspine keeps the column
widths constant & hbar lets the heights vary
                   offset = 0.02) +
         facet_grid(~condition) +
         scale_fill_manual(values = c("LMI"="#156082","non-LMI"="#E7EAF3")) +
         #theme_void() +
         theme(panel.background = element_blank(),
               axis.text.y = element_blank(),
```

```
              axis.ticks.y = element_blank()) +
        labs(title = "Figure 4. Mosaic plot of income and increased
contributions, by condition"))

## Saving 5 x 4 in image

## Warning: The `scale_name` argument of `continuous_scale()` is deprecated
as of ggplot2
## 3.5.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## Warning: The `trans` argument of `continuous_scale()` is deprecated as of
ggplot2 3.5.0.
## ℹ Please use the `transform` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## Warning: `unite_()` was deprecated in tidyr 1.2.0.
## ℹ Please use `unite()` instead.
## ℹ The deprecated feature was likely used in the ggmosaic package.
##   Please report the issue at <https://github.com/haleyjeppson/ggmosaic>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

# Almost all workers increased contributions, regardless of condition.
# More non-LMI workers received the recommendation email.
# More LMI workers received the informational (control) email.

# Create and save a mosaic plot with Pearson residuals:
jpeg(filename = "plot-mosaic_2.png")
mosaic_2 <- vcd::mosaic(~ outcome + condition + income,
                        data = contiguous,
                        main = "Retirement contributions by condition and
income",
                        shade = TRUE, legend = TRUE)
# Statistically (but not substantively) significant correlations
```

Statistical Analysis

- Null Hypothesis: Treated participants (receiving a recommendation email that leverages peer information) are just as likely as control participants (receiving a generic informational email) to increase their retirement contributions.

```
# Create cross-tabs of the frequencies/counts for the categorical outcome by
condition:
```

```r
ctab <- table(contiguous$condition, contiguous$outcome)
summary.table(ctab) # Chisq = 0.08518, df = 1, p-value = 0.7704

## Number of cases in table: 369
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 0.08518, df = 1, p-value = 0.7704
##  Chi-squared approximation may be incorrect

# At least 2 cells have frequencies that are less than 5, so a Fisher's exact
test may be more appropriate than chi-square.
stats::fisher.test(ctab) # p-value = 1

##
##  Fisher's Exact Test for Count Data
##
## data:  ctab
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.09608662 18.65096428
## sample estimates:
## odds ratio
##   1.338652

# Workers who received the recommendation email were 34% more likely (than
those who received the informational email) to increase their TSP
contribution (OR = 1.338,95% CI 0.09-18.65), if these results were
statistically significant. Overall, just about everyone increased
contributions, regardless of condition.

# Export the cross-tab:
sjPlot::sjt.xtab(contiguous$outcome, contiguous$condition,
                 title = "Table 1. Contingency table of contributions by
email condition",
                 file = "table-condition.doc")
```

Table 1. Contingency table of contributions by email condition

outcome

condition

Total

recommendation

informational

contributed

209

156

365

closed

2

2

4

Total

211

158

369

χ2=0.000 · df=1 · &phi=0.015 · Fisher’s p=1.000

```
# Explore other variables:
itab <- table(contiguous$income, contiguous$outcome)
summary.table(itab) # Chisq = 3.936, df = 1, p-value = 0.04727

## Number of cases in table: 369
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 3.936, df = 1, p-value = 0.04727
##  Chi-squared approximation may be incorrect

# statistically significant but at least 2 cells have frequencies that are
less than 5.
stats::fisher.test(itab) # p-value = 0.12

##
##  Fisher's Exact Test for Count Data
##
## data:  itab
## p-value = 0.1231
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.6465504       Inf
## sample estimates:
## odds ratio
##        Inf

# Workers who closed (their accounts?) were categorized as low-to-moderate
income.

# Export the cross-tab:
```

```
sjPlot::sjt.xtab(contiguous$outcome, contiguous$income,
                 title = "Table 2. Contingency table of contributions by
income level",
                 file = "table-income.doc")
```

Table 2. Contingency table of contributions by income level

outcome

income

Total

non-LMI

LMI

contributed

182

183

365

closed

0

4

4

Total

182

187

369

χ2=2.194 · df=1 · &phi=0.103 · Fisher's p=0.123

Conclusion

- On average, receiving an email prompted federal workers to increase their retirement contributions. Those who closed (their accounts?) were all employees with low-to-moderate income. There was no relationship between the type of email that employees received and their subsequent action.