2025-CAH

Josephine McKelvy, PhD

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

# 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%) IIIIIIIIIIIIIIIIIII 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%) IIIIIIIIIII (0.0%) (100.0%)   
##   
## 3 income\_level 1. LMI 191 (51.2%) IIIIIIIIII 373 752752   
## [character] 2. non-LMI 182 (48.8%) IIIIIIIII (0.0%) (100.0%)   
##   
## 4 increased\_contribution 1. 1 369 (98.9%) IIIIIIIIIIIIIIIIIII 373 752752   
## [character] 2. closed 4 ( 1.1%) (0.0%) (100.0%)   
## --------------------------------------------------------------------------------------------------------------------------

# 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 <- 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 <- subset(contiguous, select = c(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 pandocad08670bd22.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\RtmpW2ZvsO\rmarkdown-strad08765c211d.html"

##   
## Output created: report.html

# weak correlations between contribution, condition, and income level  
  
# Create a mosaic plot (or percent stacked bar chart) of the contingency table:  
mosaic <- ggplot(data = contiguous) +  
 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(panel.background = element\_blank(),  
 axis.text.y = element\_blank(),  
 axis.ticks.y = element\_blank())+  
 labs(title = "Figure 1. Mosaic plot of income and increased contributions, by condition")  
  
# Almost all workers increased contributions, regardless of condition.  
# More non-LMI workers received the recommendation email.  
# More LMI workers received the informational (control) email.  
  
# Save plot:  
ggsave(mosaic, filename = "mosaic.png")

## 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.

# 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.

# Descriptive Statistics: Return the frequencies/counts of the categorical outcome by condition:  
sjPlot::sjt.xtab(contiguous$outcome,  
 contiguous$condition,  
 title = "Table 1. Contingency Table of TSP contributions by email condition",  
 file = "condition.doc")

Table 1. Contingency Table of TSP contributions by email condition

outcome

condition

Total

recommendation

control

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

# At least 2 cells have frequencies that are less than 5, so a Fisher's exact test may be more appropriate than chi-square. (Neither were statistically significant.) Overall, just about everyone increased contributions, regardless of condition.  
  
ctab <- table(contiguous$condition, contiguous$outcome)  
stats::fisher.test(ctab)

##   
## 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.   
  
# Other variables  
sjPlot::sjt.xtab(contiguous$outcome,  
 contiguous$income,  
 title = "Table 2. Contingency Table of TSP contributions by income level",  
 file = "income.doc")

Table 2. Contingency Table of TSP 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

# At least 2 cells have frequencies that are less than 5, so a Fisher's exact test may be more appropriate than chi-square. (Neither were statistically significant.) Overall, those who increased contributions were evenly distributed across income level.  
  
itab <- table(contiguous$income, contiguous$outcome)  
stats::fisher.test(itab)

##   
## 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 were not categorized as low-to-moderate income were % more likely (than those whose household income is below 80% of the Area Median Income) to increase their TSP contribution (OR = infinity,95% CI 0.64-infinity), if these results were statistically significant.

# Conclusion

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