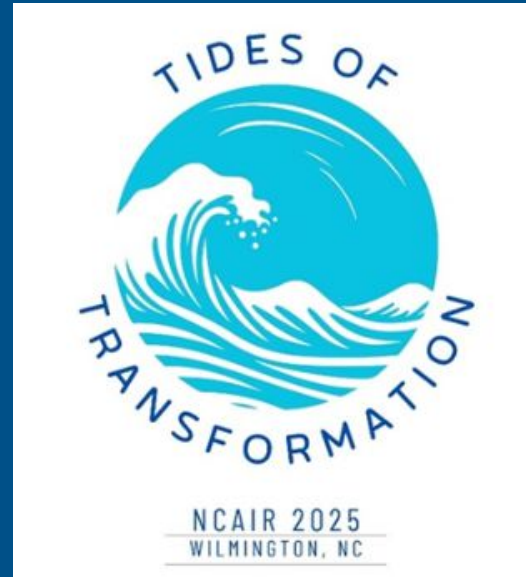


# Putting the *R* in Institutional *Research*

*Dr. Thomas Kirnbauer & Dr. Kyle Fassett*  
*NCAIR 2025*  
*March 10, 2025*



# Two IR Departments, One Common Tool



**Private Institution**

**Students: 2,000**

**IR Staff: 2 FTE**

**Primary Responsibilities:**

**Compliance Reporting,**

**Enrollment Projections,**

**Data Governance**



**Public Institution**

**Students: 32,400**

**IR Staff: ~14 FTE**

**Primary Responsibilities:**

**Accreditation, Analytics,**

**Assessment, Survey**

**Research, Reporting**



# Session Goals

- Attendees should be able to:
  - articulate the benefits of using R for Institutional Research work.
  - describe practical examples of how adopting R has improved upon legacy processes for reporting.
  - understand how R can be leveraged in different IR office settings (at both small and large institutions).

# A brief look at R...

Syntax,  
Libraries,  
#Comments

View Data,  
Connections,  
Values, etc.

Output  
(shown: SkimR)

View R Docs,  
Files, Graphs

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for loading libraries, reading a CSV file, and using `skimr` for data summarization. Comments are used to explain each step.
- Console:** Shows the output of the `skimr::skim(file)` command, providing a summary of the data file.
- Environment:** Shows the loaded data frame, `file`, with 1009 observations and 6 variables.
- Documentation:** Displays the R documentation for the `read_delim` function.

**Console Output (SkimR Summary):**

```
R 4.3.3 - C:/Users/fhkimbaun/Downloads/NCAR 2025/Example 1/ #> skimr::skim(file) # a neat function for viewing descriptive stats of the file
--- Data Summary ---
Name          values
Number of rows 1009
Number of columns 6

Column type frequency:
character      3
numeric        3

Group variables: None

--- Variable types: character ---
skin_variable n_missing complete_rate min max empty n_unique whitespace
1 First Name    0          1 3 14    0 579    0
2 Sex           0          1 4 6    0 2      0
3 Type          0          1 8 10   0 4      0

--- Variable type: numeric ---
skin_variable n_missing complete_rate mean sd p0 p25 p50 p75 p100 hist
1 Term code    0          1 54722726. 262135985. 100124054 31908293 546662374 774681334 99966232
2 Banner ID    0          1 54722726. 262135985. 100124054 31908293 546662374 774681334 99966232
3 Paid_Amount  833      0.174 5151. 2053. 745 3720 6195 6895 6895
```

**R Documentation (read\_delim):**

**Read a delimited file (including CSV and TSV) into a tibble**

**Description**

`read_csv()` and `read_tsv()` are special cases of the more general `read_delim()`. They're useful for reading the most common types of flat file data, comma separated values and tab separated values, respectively. `read_csv2()` uses ; for the field separator and , for the decimal point. This format is common in some European countries.

**Usage**

```
read_delim(
  file,
  delim = NULL,
  quote = "\"",
  escape_backslash = FALSE,
  escape_double = TRUE,
  col_names = TRUE,
  col_types = NULL,
  col_select = NULL,
  id = NULL,
  locale = default_locale(),
  na = c("", "NA"),
```

Note: We will not be showing R Syntax during the presentation, but it will be made available after the session

# Benefits of Using R

- Open Source – Free with a large community
- Reproducibility – Readable syntax for consistent data management and reporting
- Flexibility – Ability to ingest, or export to, other data formats (e.g., SQL, Excel, SPSS)
- Complexity – Easily handle large datasets for analysis

**Practical  
examples to  
demonstrate  
why we love  
using R in IR**

*Ex. 1: Census Reporting &  
Cohort Tracking*

*Ex. 2: Survey Codebooks &  
Reporting*

# Example 1: *Census Enrollment & Cohort Tracking*

# Ex 1: Legacy process for calculating retention/graduation

Legacy process relied heavily on VLOOKUPS. The process was...

- Inefficient/slow to do calculations
- No documentation – difficult to replicate with staff turnover
- Tedious to calculate values across cohorts (1-YR Retention, 4-YR Grad, 6-YR grad, etc.)
- Incompatible with BI Tools

*Note: Data included are “dummy data” to illustrate the analysis and insights for this conference session*

The screenshot shows an Excel interface with a VLOOKUP formula in cell B896: `=VLOOKUP(B896,CurrentStudents_2022.csv!$B:$B,1,FALSE)`. A red arrow points from the formula bar to the formula cell. Below the formula bar, two data tables are visible. The first table, 'CurrentStudents\_2022.csv', has columns B through G. The second table, 'CurrentStudents', has columns A through E. A red arrow points from the formula bar to the first table.

B	C	D	E	F	G
Banner ID	FIRST_NAME	SEX	STYPE_DESC	1-Year Retention	
680921343	Macy	F	Entering First Year Student	TRUE	
308476159	Evelyn	F	Entering First Year Student	TRUE	
226655087	Daniel	M	Entering First Year Student	TRUE	
742924252	Tyler	M	Entering First Year Student	TRUE	
940774838	Eric	M	Entering First Year Student	TRUE	
270170693	Jackson	M	Entering First Year Student	TRUE	
356001624	Lacy	F	Entering First Year Student	TRUE	
771286341	Mary Gray	F	Entering First Year Student	TRUE	
779031255	Lucas	M	Entering First Year Student	TRUE	
764933967	William	M	Entering First Year Student	TRUE	
261295740	Sophia	F	Entering First Year Student	TRUE	
243041000	Elizabeth	F	Entering First Year Student	TRUE	
549010327	Zoe	F	Entering First Year Student	TRUE	
129241420	Lucy	F	Entering First Year Student	TRUE	
828519243	Shea	F	Entering First Year Student	TRUE	
299665959	Austin	M	Entering First Year Student	TRUE	
337240787	Isabelle	F	Entering First Year Student	TRUE	
540732356	John	M	Entering First Year Student	TRUE	
709650717	Grant	M	Entering First Year Student	TRUE	
438783487	Michaela	F	Entering First Year Student	TRUE	
635431866	Jalysa	F	Entering First Year Student	TRUE	
168390670	Elsah	F	Entering First Year Student	TRUE	
432693264	Nathan	M	Entering First Year Student	TRUE	
368972669	Jackson	M	Entering First Year Student	TRUE	
575309615	Michael	M	Entering First Year Student	TRUE	
351426471	Williams	M	Entering First Year Student	TRUE	
990745423	Juliette	F	Entering First Year Student	TRUE	

A	B	C	D	E
Term Code	Banner ID	First Name	Sex	Type
202201	214351435	Sydney	Female	On leave
202201	924945532	Charles	Male	Continuing
202201	125909549	Goziem	Male	Continuing
202201	325110611	Coy	Male	Returning
202201	937528162	Rory	Female	Continuing
202201	527846684	Hannah	Female	Continuing
202201	382767258	Max	Male	Continuing
202201	671325186	Chloe	Female	Continuing
202201	991153245	Nathan	Male	Continuing
202201	125195410	Lauren	Female	On leave
202201	767535937	Hampton	Male	Returning
202201	780747682	Sarah	Female	Returning
202201	721386828	Rhoda	Female	Returning
202201	431039520	Spencer	Male	Continuing
202201	691329711	Dawei	Male	Continuing
202201	141434392	Kathryn	Female	Returning
202201	196500014	Raosaam	Male	Continuing
202201	558955510	Brad	Male	Continuing
202201	751587323	Isabel	Female	Continuing
202201	998340797	Cadie	Female	On leave
202201	576852055	Mary	Female	Continuing
202201	831877387	Erini	Female	Off Campu
202201	424908909	Jordan	Female	Returning
202201	760099979	Chase	Male	Continuing
202201	785135591	Oguzhan	Male	Returning
202201	833125622	Christopher	Male	Continuing
202201	45088011	Duncan	Male	Continuing



# Ex 1: Modernizing the process

**Goal:** Create a systemized, efficient approach to calculate enrollment, retention and graduation rates, across multiple cohorts

Challenges with a *messy data* across 20+ term files:

- Different column names (ID vs Banner\_ID)
- Different values (Female vs. F)
- Different formatting (\$3,000 vs 3000.00)
- Different missing values (NA, " ", 0)

Table: Fall 2019	
Column	Example
FILE_TERM	201901
ID	107XXXXXX
Sex	Female, Male
Paid	\$6195.00, NA

Table: Fall 2020	
Column	Example
Term	202001
ID	107XXXXXX
SEX	F, M
PAID AMT	\$6,345, \$0

Table: Fall 2021	
Column	Example
Term_Code	202101
Banner_ID	107XXXXXX
SEX	F, M
PAID AMOUNT	6495, NA

Table: Fall 2022	
Column	Example
Term Code	202201
Banner ID	107XXXXXX
Gender	F, M
PAID AMOUNT	6895, \$0

# Ex 1: Master Enrollment Table

## Process includes:

- Ingesting data in bulk
- Dynamically renaming columns
- Bind Rows (i.e., Data Union)
- Data Cleaning on existing info
- Creating/Joining new data



Table: Fall 2019	Table: Fall 2020	Table: Fall 2021	Table: Fall 2022
FILE_TERM	Term	Term Code	Term Code
ID	ID	Banner_ID	Banner ID
Sex	SEX	SEX	Gender



Cohort Term	Student ID	Sex	Student Type	Paid Amount
2019	XXX102	Female	First-Year	\$6,195
2020	XXX102	Female	Continuing	
2021	XXX107	Male	Transfer	\$6,495

# Ex 1: Cohort Tracking Output

## Process Includes:

- Leveraging master enrollment file
- Dynamically calculating ret/grad rates
  - Lots of data joins!

## Benefits:

- Reporting data across cohorts
- 0/1 values make it easy for calculating sums and averages
- Import into BI Tool (i.e., PowerBI)

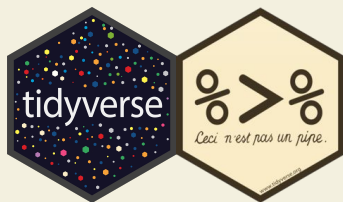


Table: Fall 2019	Table: Fall 2020	Table: Fall 2021	Table: Fall 2022
FILE_TERM	Term	Term Code	Term Code
ID	ID	Banner_ID	Banner ID
S Type	S Type	STYPE_DESC	Type



Cohort Term	Student ID	Type	1-Yr Retention	2-Yr Retention
2019	XXX119	First-Year	1	0
2020	XXX120	First-Year	1	1
2021	XXX121	Transfer	1	0

## Example 2: *Survey Codebooks & Reporting*

# Example 2: Survey Codebooks & Reporting

1. Generate graphs to see data across an entire dataframe
2. Create a codebook with variable, label, data type, value labels & descriptives into a concise format & export to excel, csv, etc.

Benefit: Consistent format allows combining codebooks to create a repository that can be built-out in Tableau

1

c82cop1

Do you feel you cope well as caregiver?

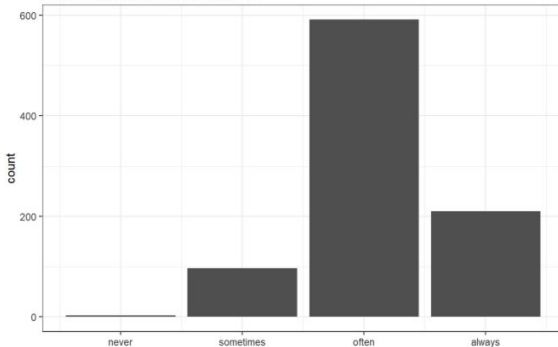
Distribution

Summary statistics

Value labels

c82cop1

Do you feel you cope well as caregiver?



2

name	label	data_type	value_labels	n_missing	complete_rate	min	median	max	mean	sd	hist
c12hour	Average number of hours of care per week	numeric		6	0.99	4	20	168	42	51	
e15relat	Relationship to elder	numeric	1. spouse/partner,2. child,3. sibling,4. daughter or son -in-law	7	0.99	1	2	8	2.9	2.1	
e16sex	Elder's gender	numeric	1. male,2. female	7	0.99	1	2	2	1.7	0.47	
e17age	Elder's age	numeric		17	0.98	65	79	103	79	8.1	
e42dep	Elder's dependency	numeric	1. independent,2. slightly dependent,3. moderately dependent	7	0.99	1	3	4	2.9	0.94	
c82cop1	Do you feel you cope well as caregiver?	numeric	1. never,2. sometimes,3. often,4. always	7	0.99	1	3	4	3.1	0.58	

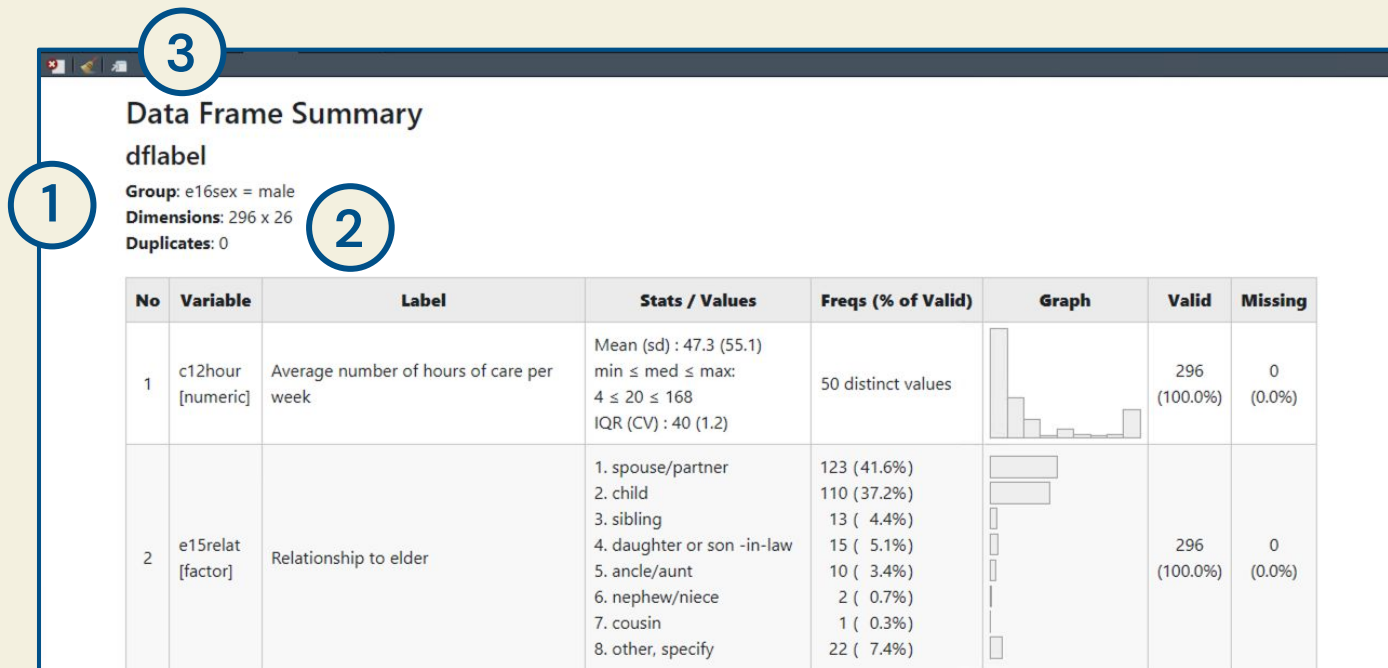


codebook 0.9.5

# Example 2: Survey Codebooks & Reporting

1. Add grouping variables
2. Choose variables or the whole df
3. Exported to word, excel, pdf

**Benefit:** quick to see what's going on or little time for a task.



# Example 2: Survey Codebooks & Reporting

1. Customize headers & footnotes
2. Two grouping variables
3. Descriptives & statistical analyses

**Benefit:** answers more specific client questions.

Table of Items by Sex & Education								
Characteristic	Male				Female			
	low level of education N = 41 <sup>1</sup>	intermediate level of education N = 113 <sup>1</sup>	high level of education N = 47 <sup>1</sup>	p-value <sup>2</sup>	low level of education N = 138 <sup>1</sup>	intermediate level of education N = 393 <sup>1</sup>	high level of education N = 138 <sup>1</sup>	p-value <sup>2</sup>
Average number of hours of care per week	41 (44)	33 (47)	39 (55)	0.6	53 (55)	44 (53)	35 (42)	0.033
Do you find caregiving too demanding?				0.14				0.047
Never	13 (32%)	30 (27%)	9 (19%)		26 (19%)	73 (19%)	20 (18%)	
Sometimes	20 (49%)	70 (62%)	26 (55%)		81 (59%)	253 (64%)	61 (56%)	
Often	5 (12%)	12 (11%)	10 (21%)		19 (14%)	51 (13%)	25 (23%)	
Always	3 (7.3%)	1 (0.9%)	2 (4.3%)		12 (8.7%)	16 (4.1%)	3 (2.8%)	

<sup>1</sup> Mean (SD); n (%)

<sup>2</sup> One-way analysis of means; Pearson's Chi-squared test

-Data: sjlabelled::efc

-Missing data listwise removed



# Perspectives on *R* Journey



# Reflecting on our R Journey

- Increased efficiency in daily workflow. Working smarter, not harder!
- Improved collaboration with colleagues
- Building robust documentation for projects (Syntax & Codebooks)
- Potential for advanced statistical analysis to further our work
- Remembering Roche's Maxim: "Data should be transformed as far upstream as possible, and as far downstream as necessary."

# Challenges

- Colleagues may not use R and institutions (or other departments) may prefer another tool or be tool agnostic
- Training staff on using R – There can be a steep initial learning curve coming from Excel or SPSS (point-and-click)
- Replacing legacy processes can be more tedious than anticipated. It's usually not ideal to replicate an old process.

# Free Resources

- RforIR.com: \*\*\* Highly Recommended \*\*\* IR-specific resources from colleagues at Furman University.
- RStudio: Beginners resources
- R for Data Science: The seminal resource for learning R, written by the creator of R
- R for Excel Users: A useful aid for converting Excel users
- Data Transformation Cheat Sheet: Print it out & Hang it!

# Questions?

- presentation, data, & code available on:  
[github.com/kfassett/NCAIR\\_2025](https://github.com/kfassett/NCAIR_2025)

Thomas Kirnbauer, PhD  
Director of IR  
Davidson College  
[thkirnbauer\[at\]davidson.edu](mailto:thkirnbauer[at]davidson.edu)

Kyle Fassett, PhD  
Senior Research Associate  
UNC-Chapel Hill  
[kfassett\[at\]unc.edu](mailto:kfassett[at]unc.edu)

# Bonus Slide – Conversation Prompts

1. Do you have any notable case examples of how you use R at your institution?
2. Do you have any concerns or barriers preventing you from using R at your institution?
3. How can institutional research pros advocate for improving data processes at your institution?

# Bonus Slide – Our Favorite R functions!

- `SkimR::skim()` – Provides descriptives of all variables (missing, mean, std dev, quartiles, histogram)
- `Janitor::clean_names()` – Cleans column names
- `Clipr::write_clip()` – Copies your dataframe to your clipboard
- `dbplyr::show_query()` – Converts syntax into SQL query
- `styler::style_file()` – Cleans script to make code format consistent