1.3.4: Missing Data and Sampling Weights

(In Person)

COMING SUMMER 2025

Module "1.3.4: Missing Data and Sampling Weights" will be posted prior to the In-Person Workshops in Summer 2025.

Session Objectives

1. Identify and summarize missing data.
2. Learn methods to handle missing data according to variable type.
3. Use a survey sampling weight to generate more representative descriptive and inferential statistical values (brief intro)
4. Discuss potential bias when removing missing observations without careful examination.

[to be removed……..]

Key points:

1. R packages that support missing data examination
2. Mean/median imputation for continuous variables
3. What to do with missing observations for categorical variables
4. Ways to examine potential differences between complete and missing observations in association between certain independent and dependent variables
    - What to do if such association significantly differs between complete and missing observations
5. R packages for complex survey data (e.g., survey package)
    - R codes to generate weighted descriptive statistics and contingency tables, as well as to develop weighted linear models

---

0. Prework - Before You Begin

Install Packages

Before you begin, please go ahead and install the following packages - these are all on CRAN, so you can install them using the RStudio Menu "Tools/Install" Packages interface:

- `VIM` on CRAN and `VIM` package website
- mice
- mi
- VIM
-
- `palmerpenguins` on CRAN

See Module 1.3.1 on Installing Packages

See additional resources below…

add to prework?

Begin with a NEW RStudio Project

Let's begin with a new RStudio Project.

---

1. Identify and summarize missing data.

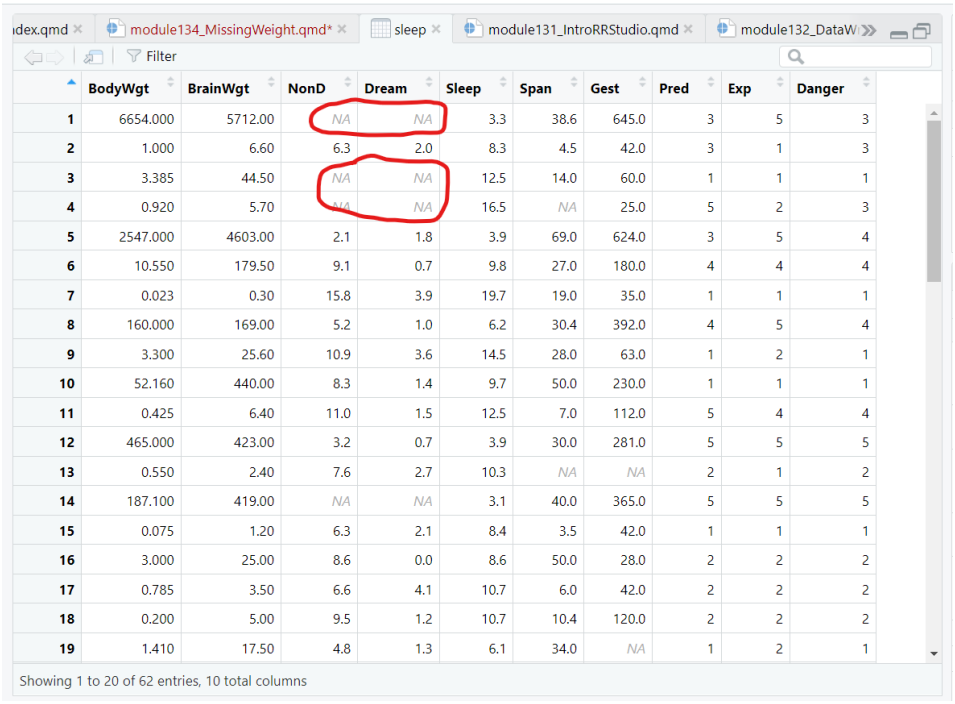
Find Missing Data in Your Dataset.

One simple way to find missing data is to open it in the Data Viewer window and sort the data.

For example, load the `VIM` package and take a look at the `sleep` dataset provided within this package.

```
library(VIM)
data("sleep")
```

Click on the sleep dataset to open it in the data viewer:



idex.qmd × | module134_MissingWeight.qmd* × | sleep × | module131_IntroRRStudio.qmd × | module132_DataW

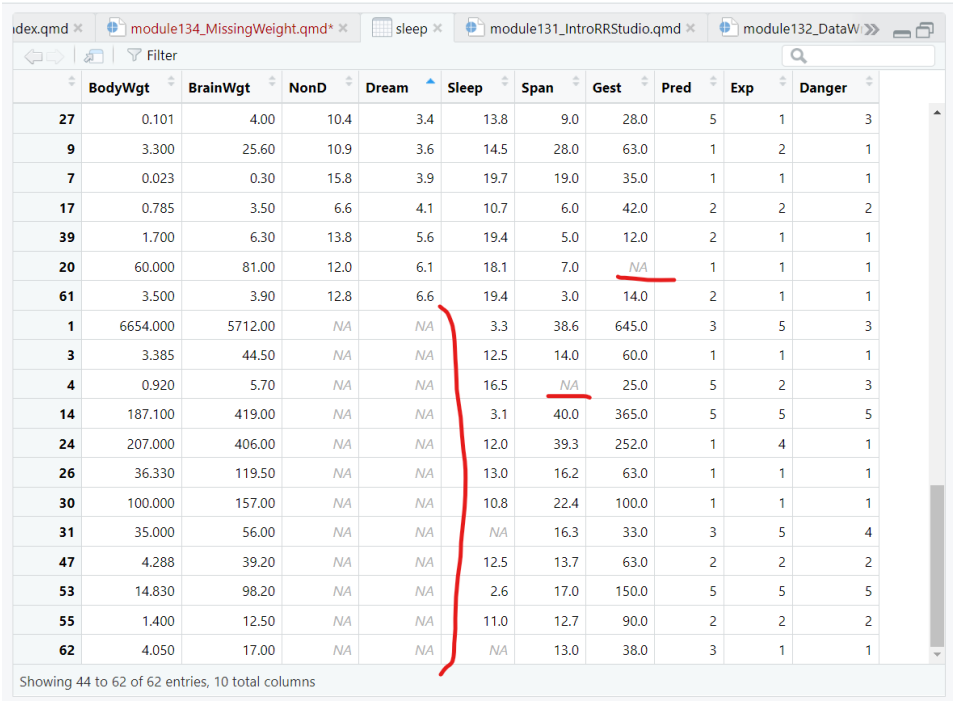
Filter

	BodyWgt	BrainWgt	NonD	Dream	Sleep	Span	Gest	Pred	Exp	Danger
1	6654.000	5712.00	NA	NA	3.3	38.6	645.0	3	5	3
2	1.000	6.60	6.3	2.0	8.3	4.5	42.0	3	1	3
3	3.385	44.50	NA	NA	12.5	14.0	60.0	1	1	1
4	0.920	5.70	NA	NA	16.5	NA	25.0	5	2	3
5	2547.000	4603.00	2.1	1.8	3.9	69.0	624.0	3	5	4
6	10.550	179.50	9.1	0.7	9.8	27.0	180.0	4	4	4
7	0.023	0.30	15.8	3.9	19.7	19.0	35.0	1	1	1
8	160.000	169.00	5.2	1.0	6.2	30.4	392.0	4	5	4
9	3.300	25.60	10.9	3.6	14.5	28.0	63.0	1	2	1
10	52.160	440.00	8.3	1.4	9.7	50.0	230.0	1	1	1
11	0.425	6.40	11.0	1.5	12.5	7.0	112.0	5	4	4
12	465.000	423.00	3.2	0.7	3.9	30.0	281.0	5	5	5
13	0.550	2.40	7.6	2.7	10.3	NA	NA	2	1	2
14	187.100	419.00	NA	NA	3.1	40.0	365.0	5	5	5
15	0.075	1.20	6.3	2.1	8.4	3.5	42.0	1	1	1
16	3.000	25.00	8.6	0.0	8.6	50.0	28.0	2	2	2
17	0.785	3.50	6.6	4.1	10.7	6.0	42.0	2	2	2
18	0.200	5.00	9.5	1.2	10.7	10.4	120.0	2	2	2
19	1.410	17.50	4.8	1.3	6.1	34.0	NA	1	2	1

Showing 1 to 20 of 62 entries, 10 total columns

Notice the light grey `NA`s shown for the missing data spots in this dataset.

If we click on the column for the `Dream` variable and sort these values, notice that the `NA`s all now show up at the bottom of the viewer window. It does not matter if you sort ascending or descending, the `NA`s are always at the bottom of the viewer.



idex.qmd | module134_MissingWeight.qmd* | sleep | module131_IntroRRStudio.qmd | module132_DataW

Filter

	BodyWgt	BrainWgt	NonD	Dream	Sleep	Span	Gest	Pred	Exp	Danger
27	0.101	4.00	10.4	3.4	13.8	9.0	28.0	5	1	3
9	3.300	25.60	10.9	3.6	14.5	28.0	63.0	1	2	1
7	0.023	0.30	15.8	3.9	19.7	19.0	35.0	1	1	1
17	0.785	3.50	6.6	4.1	10.7	6.0	42.0	2	2	2
39	1.700	6.30	13.8	5.6	19.4	5.0	12.0	2	1	1
20	60.000	81.00	12.0	6.1	18.1	7.0	NA	1	1	1
61	3.500	3.90	12.8	6.6	19.4	3.0	14.0	2	1	1
1	6654.000	5712.00	NA	NA	3.3	38.6	645.0	3	5	3
3	3.385	44.50	NA	NA	12.5	14.0	60.0	1	1	1
4	0.920	5.70	NA	NA	16.5	NA	25.0	5	2	3
14	187.100	419.00	NA	NA	3.1	40.0	365.0	5	5	5
24	207.000	406.00	NA	NA	12.0	39.3	252.0	1	4	1
26	36.330	119.50	NA	NA	13.0	16.2	63.0	1	1	1
30	100.000	157.00	NA	NA	10.8	22.4	100.0	1	1	1
31	35.000	56.00	NA	NA	NA	16.3	33.0	3	5	4
47	4.288	39.20	NA	NA	12.5	13.7	63.0	2	2	2
53	14.830	98.20	NA	NA	2.6	17.0	150.0	5	5	5
55	1.400	12.50	NA	NA	11.0	12.7	90.0	2	2	2
62	4.050	17.00	NA	NA	NA	13.0	38.0	3	1	1

Showing 44 to 62 of 62 entries, 10 total columns

This method is ok for a small dataset with not too many variables or rows of data. But let's look at other ways to summarize the amounts of missing data in your dataset.

Describe Missing Data.

As we saw back in Module 1.3.2, Section 5, we can use the `summary()` function to get some basic statistics for each variable in the dataset, including the number of `NA`s.

```
summary(sleep)
```

```
    BodyWgt            BrainWgt             NonD            Dream      
 Min.   :    0.005   Min.   :    0.14   Min.   : 2.100   Min.   :0.000  
 1st Qu.:    0.600   1st Qu.:    4.25   1st Qu.: 6.250   1st Qu.:0.900  
 Median :    3.342   Median :   17.25   Median : 8.350   Median :1.800  
 Mean   :  198.790   Mean   :  283.13   Mean   : 8.673   Mean   :1.972  
 3rd Qu.:   48.203   3rd Qu.:  166.00   3rd Qu.:11.000   3rd Qu.:2.550  
 Max.   :6654.000   Max.   :5712.00   Max.   :17.900   Max.   :6.600  
                                       NA's   :14       NA's   :12     
     Sleep            Span             Gest            Pred      
 Min.   : 2.60   Min.   :  2.000   Min.   : 12.00   Min.   :1.000  
 1st Qu.: 8.05   1st Qu.:  6.625   1st Qu.: 35.75   1st Qu.:2.000  
 Median :10.45   Median : 15.100   Median : 79.00   Median :3.000  
```

```
Mean   :10.53   Mean   : 19.878   Mean   :142.35   Mean   :2.871
3rd Qu.:13.20   3rd Qu.: 27.750   3rd Qu.:207.50   3rd Qu.:4.000
Max.   :19.90   Max.   :100.000   Max.   :645.00   Max.   :5.000
NA's   :4       NA's   :4         NA's   :4
      Exp             Danger
Min.   :1.000   Min.   :1.000
1st Qu.:1.000   1st Qu.:1.000
Median :2.000   Median :2.000
Mean   :2.419   Mean   :2.613
3rd Qu.:4.000   3rd Qu.:4.000
Max.   :5.000   Max.   :5.000
```

Another helpful package is the `skimr` package which has the `skim()` function which provides a count of the amount of missing data and the proportion of complete data for that variable.

> **Rmarkdown**
>
> When "knitting" to HTML the code below creates the summary table with the miniture histograms. However, when "knitting" to PDF (using the default portrait layout)m the histograms get cutoff on the page. Additional LaTex customization is needed to change the layout to landscape to be able to see the histograms.

```
library(skimr)
skim(sleep)
```

Table 1: Data summary

Name	sleep
Number of rows	62
Number of columns	10
_____	
Column type frequency:	
numeric	10
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
BodyWgt	0	1.00	198.79	899.16	0.00	0.60	3.34	48.20	6654.0	
BrainWgt	0	1.00	283.13	930.28	0.14	4.25	17.25	166.00	5712.0	
NonD	14	0.77	8.67	3.67	2.10	6.25	8.35	11.00	17.9	
Dream	12	0.81	1.97	1.44	0.00	0.90	1.80	2.55	6.6	
Sleep	4	0.94	10.53	4.61	2.60	8.05	10.45	13.20	19.9	
Span	4	0.94	19.88	18.21	2.00	6.62	15.10	27.75	100.0	
Gest	4	0.94	142.35	146.81	12.00	35.75	79.00	207.50	645.0	
Pred	0	1.00	2.87	1.48	1.00	2.00	3.00	4.00	5.0	
Exp	0	1.00	2.42	1.60	1.00	1.00	2.00	4.00	5.0	
Danger	0	1.00	2.61	1.44	1.00	1.00	2.00	4.00	5.0	

better printing of the output object? SKIP

try datasummary_skim from modelsummary

```
#library(kableExtra)
library(modelsummary)

# datasummary_skim(sleep,
#                  output = "kableExtra") %>%
#   kable_styling(latex_options = c("HOLD_position"),
#                 full_width = TRUE,
#                 font_size = 10)

library(tinytable)
datasummary_skim(sleep) %>%
  theme_tt("placement", latex_float = "H")
```

	Unique	Missing Pct.	Mean	SD	Min	Median	Max	Histogram
BodyWgt	60	0	198.8	899.2	0.0	3.3	6654.0	
BrainWgt	59	0	283.1	930.3	0.1	17.2	5712.0	
NonD	40	23	8.7	3.7	2.1	8.4	17.9	
Dream	31	19	2.0	1.4	0.0	1.8	6.6	
Sleep	45	6	10.5	4.6	2.6	10.4	19.9	
Span	48	6	19.9	18.2	2.0	15.1	100.0	
Gest	50	6	142.4	146.8	12.0	79.0	645.0	
Pred	5	0	2.9	1.5	1.0	3.0	5.0	
Exp	5	0	2.4	1.6	1.0	2.0	5.0	
Danger	5	0	2.6	1.4	1.0	2.0	5.0	

Try It On Your Own

Try running `summary()` or `skim()` on the `penguins` dataset from the `palmerpenguins` package. Notice the summaries for the numeric and the factor type variables.

```
library(palmerpenguins)
summary(penguins)
```

```
      species          island    bill_length_mm  bill_depth_mm
 Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
 Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
 Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
                                 Mean   :43.92   Mean   :17.15
                                 3rd Qu.:48.50   3rd Qu.:18.70
                                 Max.   :59.60   Max.   :21.50
                                 NA's   :2       NA's   :2
 flipper_length_mm  body_mass_g       sex           year
 Min.   :172.0     Min.   :2700   female:165   Min.   :2007
 1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
 Median :197.0     Median :4050   NA's  : 11   Median :2008
 Mean   :200.9     Mean   :4202                Mean   :2008
 3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
 Max.   :231.0     Max.   :6300                Max.   :2009
 NA's   :2         NA's   :2
```

```
skim(penguins)
```

Table 3: Data summary

Name	penguins
Number of rows	344
Number of columns	8
_____	
Column type frequency:	
factor	3
numeric	5
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
species	0	1.00	FALSE	3	Ade: 152, Gen: 124, Chi: 68
island	0	1.00	FALSE	3	Bis: 168, Dre: 124, Tor: 52
sex	11	0.97	FALSE	2	mal: 168, fem: 165

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
bill_length_mm	2	0.99	43.92	5.46	32.1	39.23	44.45	48.5	59.6	
bill_depth_mm	2	0.99	17.15	1.97	13.1	15.60	17.30	18.7	21.5	
flipper_length_mm	2	0.99	200.92	14.06	172.0	190.00	197.00	213.0	231.0	
body_mass_g	2	0.99	4201.75	801.95	2700.0	3550.00	4050.00	4750.0	6300.0	
year	0	1.00	2008.03	0.82	2007.0	2007.00	2008.00	2009.0	2009.0	

Visualize Missing Data.

discuss ways to identify and quantify missing data

look at visualization methods - looking for patterns - again how to quantify

---

2. Learn methods to handle missing data according to variable type.

discuss pairwise versus listwise and discuss impacts on modeling especially for stepwise variable selection - always check the final N for each model

show correlations pairwise and listwise

add details on modeling adjustments - covariate predicted missingness

options on imputation - brief intro

---

3. Use a survey sampling weight to generate more representative descriptive and inferential statistical values (brief intro)

introduction to survey weights

show how this impacts the amounts of missing data

---

4. Discuss potential bias when removing missing observations without careful examination.

talk about assumptions for missing data - MCAR, MAR and NMAR (or MNAR)

add more examples here

also for publication - running models for comparison - sensitivity tests - model for all complete data - models based on pairwise selections - n changes - models before and after covariate adjustments - models before and after imputation

---

R Code For This Module

- `module_134.R`

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

———. 2025a. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://modelsummary.com.

———. 2025b. *Tinytable: Simple and Configurable Tables in HTML, LaTeX, Markdown, Word, PNG, PDF, and Typst Formats*. https://vincentarelbundock.github.io/tinytable/.

Horst, Allison, Alison Hill, and Kristen Gorman. 2022. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://allisonhorst.github.io/palmerpenguins/.

Ihaka, Ross, Paul Murrell, Kurt Hornik, Jason C. Fisher, Reto Stauffer, Claus O. Wilke, Claire D. McWhite, and Achim Zeileis. 2023. *Colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes*. https://colorspace.R-Forge.R-project.org/.

Kowarik, Alexander, and Matthias Templ. 2016. "Imputation with the R Package VIM." *Journal of Statistical Software* 74 (7): 1–16. https://doi.org/10.18637/jss.v074.i07.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Stauffer, Reto, Georg J. Mayr, Markus Dabernig, and Achim Zeileis. 2009. "Somewhere over the Rainbow: How to Make Effective Use of Colors in Meteorological Visualizations." *Bulletin of the American Meteorological Society* 96 (2): 203–16. https://doi.org/10.1175/BAMS-D-13-00155.1.

Templ, Matthias, Alexander Kowarik, Andreas Alfons, Gregor de Cillia, and Wolfgang Rannetbauer. 2022. *VIM: Visualization and Imputation of Missing Values*. https://github.com/statistikat/VIM.

Waring, Elin, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu, and Shannon Ellis. 2022. *Skimr: Compact and Flexible Summaries of Data*. https://docs.ropensci.org/skimr/.

Zeileis, Achim, Jason C. Fisher, Kurt Hornik, Ross Ihaka, Claire D. McWhite, Paul Murrell, Reto Stauffer, and Claus O. Wilke. 2020. "colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes." *Journal of Statistical Software* 96 (1): 1–49. https://doi.org/10.18637/jss.v096.i01.

Zeileis, Achim, Kurt Hornik, and Paul Murrell. 2009. "Escaping RGBland: Selecting Colors for Statistical Graphics." *Computational Statistics & Data Analysis* 53 (9): 3259–70. https://doi.org/10.1016/j.csda.2008.11.033.

Other Helpful Resources

Other Helpful Resources

Missing Data Resources

- CRAN Task View for Missing Data
- R-miss-tastic Website
- Flexible Imputation of Missing Data (online book for 2nd edition) by Stef van Buuren
- more ...
- https://www.datawim.com/post/missing-data-visualization-in-r/
- https://libguides.princeton.edu/R-Missingdata
- https://cran.r-project.org/web/packages/mice/index.html
- https://cran.r-project.org/web/views/MissingData.html
- https://rmisstastic.netlify.app/
- https://rmisstastic.netlify.app/tutorials/josse_tierney_bookdown_user2018tutorial_2018
- https://modelsummary.com/vignettes/datasummary.html
- https://dabblingwithdata.amedcalf.com/2018/01/02/my-favourite-r-package-for-summarising-data/
- https://cran.r-project.org/web/packages/summarytools/vignettes/introduction.html
- https://cran.r-project.org/web/packages/skimr/vignettes/skimr.html