

NA: When Missing Data is Valuable



Caitlin Hudon | Cascadia R
@beeonaposity

Missing data 🙅

Obscures underlying relationships in data

Can cover existing relationships or skew relationships

Can be dangerous when mishandled or ignored

Must be acknowledged before modeling, so good to address ASAP



Missing data can be valuable (!) ✦

1. When you can fix it
2. When the fact that data is missing tells you something important
3. When it tells you where something in your data pipeline is broken / can be improved





1. Look for the underlying reason for missing values

Things to consider ☐

- ☐ Is there a discernible pattern to the values that are present vs. missing? (MAR vs. MNAR)
- ☐ Is there an underlying explanation for missing values?
- ☐ How is the data collected?
- ☐ Should you analyze the data in groups that have similar amounts of data available?



Options for handling missings

- **Imputation** -- mean, median, zero, predicted values
- Add new **indicator variables** for whether values exist
- Focus only on **complete cases** (case-wise deletion)
- Create a new **unknown category** for categorical vars
- **Remove variables** with too many missing values



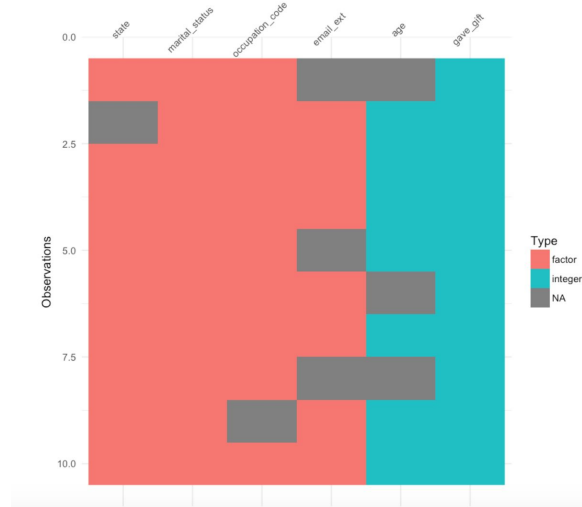
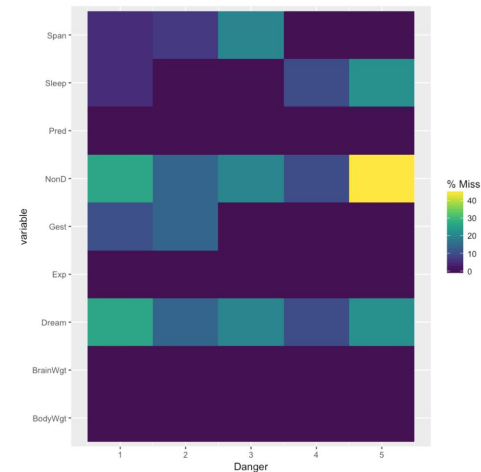
R Packages to Help


Exploring missing values:

→ visdat, naniar, dplyr (+ more!)

Handling missing values:

→ mice, Amelia, dplyr, forcats





**2. The presence or absence
of data can be just as
valuable as the data itself**

Missings I've flagged and loved

No email / phone?

Missing contact info flag

No first gift amount?

Didn't do "x" flag

No interactions?

Never used flag



Cascadia R
@beeonaposity

Possible underlying reasons for missings

Affinity / trust

email; phone number

Self-censoring

income

Lack of engagement

interaction / touchpoint

Data collection

1/NA; optional inputs



**Missing data is an
opportunity to improve
your data pipeline**

Adding value back to your data pipeline

- ❑ **Add in** any newly-created variables
- ❑ **Fix issues** w/ data collection mechanisms
- ❑ Think about **new fields** to collect
- ❑ **Communicate** feedback and ideas

Communicate findings with others

Relevance

- Data is perfectly relevant and usable as-is to make important decisions.
- This data is useful to add color to arguments or decisions but isn't a single definitive source of truth in itself.
- We don't recommend using this data to make important decisions.

Trustworthiness

- We trust this data, including the source and the way it's captured, and feel comfortable using it to make important decisions.
- We have some reservations with this data -- could be based on the way it's created, accessed, or an unexplained weirdness.
- We don't trust this data due to the way it is collected or stored in current state.

Repeatability

- The process to get this metric is fully automated or automatable.
- The process to get this data is standardized, but not fully automatable (involves a manual download or have to go through a point person).
- Data does not live in a database; the process to get this data is very manual and cannot be automated in current state.





To get value out of missing data:

Look for underlying reasons.

Fix or flag what you can.

Communicate your findings.

Improve your pipeline.

Thank you!



Caitlin Hudon



@beeonaposity



caitlinhudon.com

Slides available at <https://github.com/caitlinhudon/cascadiaRconf>